## Wprowadzenie

**O czym jest ten wykład?**
Dowiesz się z niego, jak zbudowane są najważniejsze rodzaje układów scalonych, i w jaki sposób się je projektuje. Ten wykład przygotowuje nie tylko do korzystania z gotowych układów scalonych, które można kupić w sklepie, ale do tworzenia własnych.

**A po co uczyć się, jak projektuje się układy scalone?**
Bez układów scalonych nie można dziś wyobrazić sobie żadnego wyrobu elektronicznego. Oprócz standardowych układów, zwanych katalogowymi, praktycznie w każdym urządzeniu elektronicznym znajdziemy dziś układy scalone zwane specjalizowanymi – zaprojektowane i wytwarzane specjalnie do tego urządzenia. Takich układów nigdzie nie kupimy gotowych. Ich zaprojektowanie jest zadaniem konstruktora urządzenia. Znajomość zasad projektowania specjalizowanych układów scalonych zalicza się dziś do podstawowych umiejętności inżyniera elektronika. Nawet konstruktor, który sam nie projektuje układów scalonych, powinien wiedzieć, jak to się robi, aby móc porozumieć się z projektantem, który zaprojektuje układy specjalizowane do konstruowanego przez niego urządzenia.

**Ale czy to się w Polsce do czegoś przydaje? Przecież nie mamy fabryk układów scalonych!**
Rzeczywiście, w Polsce jest obecnie tylko jedna linia produkcyjna w Instytucie Technologii Elektronowej w Warszawie. Służy ona głównie do celów doświadczalnych i do wytwarzania niewielkich ilości nietypowych wyrobów półprzewodnikowych. Mimo to polski inżynier ma równie dobry i łatwy dostęp do możliwości wytwarzania specjalizowanych układów scalonych, jak inżynier we Francji, Niemczech, Japonii czy też USA. Więcej na ten temat dowiesz się z tego wykładu.

**A co z kosztami?**
W tym wykładzie zapoznasz się także z aspektami ekonomicznymi zastosowań mikroelektroniki. Zobaczysz, że zastosowanie specjalizowanych układów scalonych może być ekonomicznie uzasadnione przy każdej skali produkcji - nawet wtedy, gdy potrzebne jest zaledwie kilkadziesiąt egzemplarzy układu.

**A czy projektowanie układów scalonych nie jest niezwykle trudne, czy nie wymaga wysoko wyspecjalizowanego wykształcenia?**
Nie. Istnieje dziś szereg metod projektowania i opartych na nich systemów wspomagania komputerowego, które upraszczają projektowanie do takiego stopnia, że nie wymaga ono ani bardzo głębokiej znajomości technologii półprzewodnikowych, ani żadnej innej “wiedzy tajemnej”. Przekonasz się o tym! Oczywiście jak w każdej dziedzinie techniki, tak i w tej istnieją projekty łatwe i trudne, i są projektanci o różnym poziomie umiejętności, ale – jak zobaczysz – i Ty możesz stać się projektantem specjalizowanych układów scalonych, a wtedy Twoje możliwości jako konstruktora sprzętu elektronicznego ograniczać będzie tylko Twoja pomysłowość i wyobraźnia.

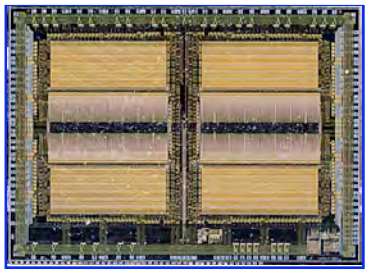
**A skąd wziąć oprogramowanie komputerowe potrzebne do projektowania?**
Proste oprogramowanie, które pozwoli Ci zobaczyć, jak to się robi, jest dołączone na tej płycie. Otrzymujesz je do własnego użytku dzięki uprzejmości prof. Etienne Sicarda z Politechniki w Tuluzie. Działa ono w środowisku PC/Windows. Jeśli używasz komputera firmy Apple z procesorem Intela i systemem Mac OS X, możesz również bez problemów używać tego oprogramowania, wykorzystując możliwość zainstalowania systemu Windows. Jeśli nie wiesz, jak to zrobić, wyślij e-mail do autora podręcznika. Istnieje również oprogramowanie do projektowania działające w środowisku Mac OS X, opracowane na Politechnice Warszawskiej. Nie jest ono wykorzystywane w tym wykładzie, ale możesz - jeśli chcesz - zapoznać się z nim.

**Co trzeba umieć, by bez kłopotów studiować ten przedmiot?**
Przedmiot "Układy scalone" prezentuje głównie praktyczne zagadnienia projektowe mikroelektroniki. Aby dobrze sobie z nim poradzić, trzeba znać przynajmniej w podstawowym zakresie zasady działania przyrządów półprzewodnikowych, ich charakterystyki i parametry. Trzeba też orientować się w zasadach działania podstawowych układów elektronicznych, cyfrowych i analogowych, oraz w podstawach teorii układów logicznych.

# Wykład 1: Rola mikroelektroniki w technikach informacyjnych

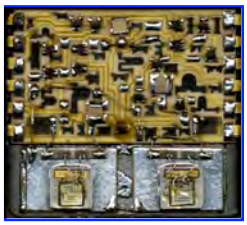
## Wstęp

Mikroelektroniką nazywamy ogólnie dziedzinę techniki, która umożliwia realizację urządzeń i systemów informacyjnych przy użyciu układów scalonych. Są to przy dzisiejszym stanie technologii niemal wyłącznie **monolityczne półprzewodnikowe układy scalone**. Nazywamy je **monolitycznymi** i **półprzewodnikowymi**, ponieważ są to struktury wykonane w całości wewnątrz i na powierzchni płytki półprzewodnika, którym jest (z rzadkimi wyjątkami) krzem. Wszystkie elementy monolitycznego układu scalonego, a także połączenia elektryczne między tymi elementami, wytwarzane są równocześnie, w wyniku wykonania szeregu operacji technologicznych tworzących wewnątrz oraz na powierzchni płytki półprzewodnikowej obszary odpowiednio domieszkowane, a także obszary przewodzące i dielektryczne.



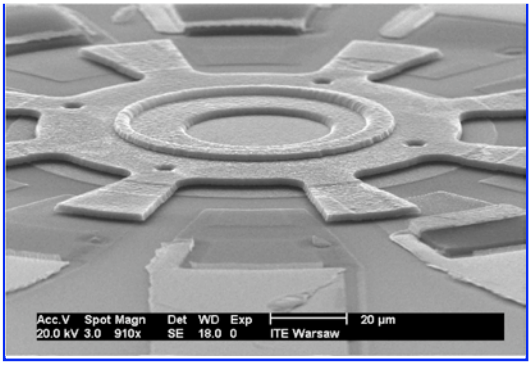
Rys. 1.1. Mikrofotografia monolitycznego układu scalonego

Istnieją także układy scalone zwane **hybrydowymi**. Takie układy powstają w inny sposób. Na podłożu ceramicznym (lub niekiedy szklanym) wytwarzane są obszary dielektryczne i przewodzące, zaś półprzewodnikowe elementy czynne (np. tranzystory) dołączane są jako elementy zewnętrzne, wyprodukowane osobno i zamknięte we własnych obudowach. Układy hybrydowe są dziś bardzo rzadko stosowane. Nie będą one omawiane w tym wykładzie.



Rys. 1.2. Mikrofotografia hybrydowego układu scalonego

Procesy technologiczne stosowane w mikroelektronice, takie jak nakładanie warstw różnego rodzaju, fotolitografia, utlenianie, selektywne trawienie, mogą służyć do wytwarzania nie tylko struktur układów scalonych, ale także różnego rodzaju miniaturowych mechanizmów, czujników, elementów optycznych itp. Dzięki temu możliwe staje się scalenie nie tylko układów elektronicznych, ale wytwarzanie struktur, w których wraz z układem elektronicznym scalone są czujniki mechaniczne, chemiczne czy też optyczne, lub różnorodne mechanizmy wykonawcze, jak na przykład mikroskopijnej wielkości silniki. Struktury takie zwane są w skrócie **mikrosystemami scalonymi**. Jest to nowa, bardzo dynamicznie rozwijająca się dziedzina techniki, mająca już obecnie liczne praktyczne zastosowania (na przykład samochodowe poduszki powietrzne wyzwalane są przy użyciu scalonych krzemowych mikroczujników przyspieszenia). Łączy ona mikroelektronikę z mechaniką precyzyjną, optyką, chemią i biochemią. W tym wykładzie nie ma miejsca na omawianie technologii i projektowania mikrosystemów scalonych, ale wspominamy tu o nich, bowiem jest to naturalne rozszerzenie klasycznej mikroelektroniki o szybko rosnącym znaczeniu i licznych obszarach zastosowań.



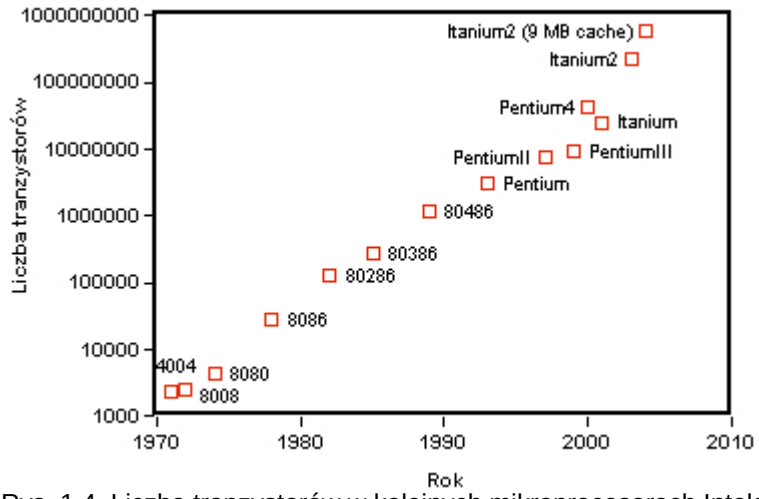
Rys. 1.3. Mikrofotografia mikromechanizmu krzemowego

Pierwszy wykład stanowi wprowadzenie do przedmiotu: ukazuje rolę mikroelektroniki we współczesnych technikach informacyjnych. Dowiesz się, dlaczego stała się tak powszechnie stosowanym sposobem realizacji urządzeń i systemów elektronicznych.

## 1.1. Do czego potrzebna jest mikroelektronika

Dążenie do miniaturyzacji sprzętu elektronicznego doprowadziło w latach sześćdziesiątych XX wieku do powstania pierwszych półprzewodnikowych układów scalonych. Narodziła się mikroelektronika. Początkowo układy scalone były prymitywne technicznie, projektowanie ich odbywało się w dużym stopniu metodą prób i błędów, a produkcja była bardzo kosztowna. Mikroelektronikę w tych czasach uważano za niszową technologię do nielicznych zastosowań. Uważano, że tylko konieczność zapewnienia małych wymiarów i ciężaru sprzętu uzasadnia użycie w nim układów scalonych.

Pomimo tego mikroelektronika zaczęła rozwijać się niezwykle szybko. Już w 1965 roku Gordon Moore, późniejszy założyciel firmy Intel, przewidział że liczba tranzystorów w cyfrowych układach scalonych będzie rosła wykładniczo, podwajając się mniej więcej co dwa lata. Prognoza ta sprawdziła się, i jest dziś powszechnie znana jako "prawo Moore'a". Ilustruje je (na przykładzie mikroprocesorów firmy Intel) rys. 1.4.




Rys. 1.4. Liczba tranzystorów w kolejnych mikroprocesorach Intela

Równocześnie ze wzrostem liczby tranzystorów i stopniem złożoności układów wykładniczo wzrastały także ich możliwości takie, jak częstotliwość taktowania czy też pojemność pamięci. Było to możliwe między innymi dzięki systematycznemu zmniejszaniu wymiarów tranzystorów. We wczesnych latach siedemdziesiątych ubiegłego wieku minimalna długość kanału tranzystora MOS wynosiła 10 mikrometrów, a dziś (rok 2010) w produkcji są układy, w których minimalna długość kanału tranzystora wynosi 32 nanometry, czyli o prawie trzy rzędy wielkości mniej. Nie ma w historii ludzkości żadnego innego przykładu dziedziny techniki, której rozwój odbywałby się równie dynamicznie.

Małe wymiary i ciężar układów scalonych są również dziś ich ważną zaletą. Istnieją tysiące zastosowań, w których właśnie małe wymiary i ciężar układów scalonych są niezastąpione. Wszyscy znamy wiele z tych zastosowań: elektroniczny zegarek, przenośny radiomagnetofon, kieszonkowy odtwarzacz plików mp3, telefon komórkowy, palmtop, karta kredytowa z mikroprocesorem, wszczepialny stymulator serca, komputer sterujący silnikiem samochodu, elektroniczny system sterowania lotem samolotu i wiele, wiele innych. Wszystkie te urządzenia i systemy nie dałyby się zrealizować, gdyby nie miniaturyzacja możliwa dzięki układom scalonym. Ale dziś nie miniaturyzacja, lecz dwie inne zalety układów scalonych są uważane za najważniejsze: **wysoka niezawodność** i **niski koszt**. One właśnie stały się głównymi czynnikami napędowymi rozwoju mikroelektroniki.

Wzrost niezawodności systemów elektronicznych, jaki umożliwiła mikroelektronika, można dobrze zilustrować na przykładzie niezawodności komputerów. W latach siedemdziesiątych XX wieku średni czas bezawaryjnej pracy komputera wynosił typowo kilka godzin. Dziś średni czas bezawaryjnej pracy mikroprocesora, będącego pod względem mocy obliczeniowej odpowiednikiem komputera z lat siedemdziesiątych, wynosi kilkadziesiąt lat. Oznacza to wzrost niezawodności, mierzony czasem bezawaryjnej pracy, w stosunku 1:10 000, czyli o **4 rzędy wielkości**. Przy takim poziomie niezawodności, z jakim mieliśmy do czynienia przed trzydziestu laty, nie byłoby żadnych szans realizacji takich systemów, jak na przykład obejmujący cały świat system telefonii tradycyjnej czy też komórkowej. Po prostu w tak wielkich systemach liczba uszkodzonych podzespołów i bloków byłaby w każdej chwili na tyle duża, że systemy te jako całość praktycznie nie nadawałyby się do użytku.

Komputery są także dobrym przykładem spadku kosztu urządzeń elektronicznych, jaki zawdzięczamy mikroelektronice. Koszt procesora komputera z połowy lat siedemdziesiątych XX wieku zawierał się w przedziale 10 000 $ - 1 000 000 $, a dziś typowy koszt mikroprocesora zawiera się między 1 $ a 100 $. Nastąpiła więc redukcja kosztu w stosunku 1:10 000, czyli także o **4 rzędy wielkości**. Większość istniejących dziś i powszechnie używanych urządzeń elektronicznych mogłaby teoretycznie być produkowana już przed trzydziestu laty, lecz byłyby one wówczas tak kosztowne, że ich produkcja nie miałaby ekonomicznego sensu, ponieważ nie znalazłyby nabywców lub użytkowników.

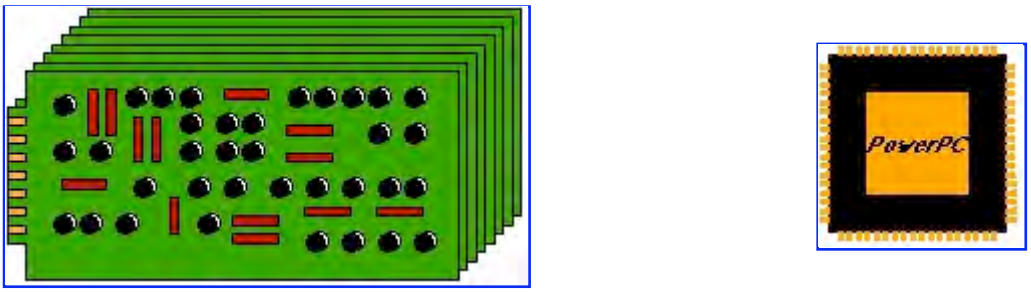
To właśnie mikroelektronika zapewniła zarówno ten ogromny wzrost niezawodności, jak i ogromny spadek kosztu urządzeń i systemów elektronicznych. Można śmiało powiedzieć, że mikroelektronice zawdzięczamy współczesny kształt naszej cywilizacji.

Czy ten trwający już ponad 40 lat rozwój mikroelektroniki nie będzie miał końca? Oczywiście nie. Istnieją fizyczne granice, wynikające z fundamentalnych praw przyrody. Będzie o nich mowa w końcowej częsci tych wykładów. Ale nic nie wskazuje na to, aby mikroelektronika miała tracić swoje gospodarcze i cywilizacyjne znaczenie nawet wtedy, gdy rozwój według "prawa Moore'a" nie będzie już możliwy.

## 1.2. Niezawodność i koszt systemów elektronicznych

Zarówno niezawodność systemu elektronicznego, jak i jego koszt zależą od wielu czynników (będzie o nich mowa w następnych wykładach). Jest jednak pewna wspólna, a zarazem bardzo prosta przyczyna decydująca zarówno o wzroście niezawodności, jak i o spadku kosztu, jakie zawdzięczamy mikroelektronice. Aby poznać tę przyczynę, przyjrzyjmy się dokładniej komputerowi z lat siedemdziesiątych i współczesnemu mikroprocesorowi.

Procesor komputera z lat siedemdziesiątych zbudowany był z setek tysięcy indywidualnych elementów: głównie tranzystorów, a także diod, rezystorów, kondensatorów (lub ewentualnie układów scalonych małej skali integracji, z których każdy zawierał kilkanaście elementów). Wszystkie te elementy były zmontowane na dziesiątkach, a nawet setkach pakietów drukowanych, które z kolei połączone były ze sobą przy pomocy odpowiednich złącz i okablowania. Dzisiejszy mikroprocesor, zawierający wewnątrz podobną (lub nawet znacznie większą) liczbę elementów, jest wykonany w płytce krzemowej, w której wszystkie elementy i połączenia między nimi stanowią jednolitą strukturę.




Rys. 1.5 Pakiety drukowane, z których zbudowane były dawniej procesory komputerów, zostały zastąpione mikroprocesorem

I na tym właśnie polega zasadnicza różnica. Komputer z lat siedemdziesiątych wymagał wielkiej liczby indywidualnie wykonywanych połączeń między elementami. W przypadku mikroprocesora takich połączeń jest zaledwie kilkadziesiąt lub kilkaset. Jest ich tyle, ile jest zewnętrznych wyprowadzeń elektrycznych łączących mikroprocesor z pozostałymi elementami urządzenia lub systemu. Porównajmy dla przykładu hipotetyczny procesor komputera zbudowanego w latach siedemdziesiątych z miliona indywidualnych tranzystorów. Każdy z tych tranzystorów ma trzy wyprowadzenia, wymaga więc najpierw wykonania trzech połączeń wewnątrz obudowy, łączących strukturę półprzewodnikową z zewnętrznymi wyprowadzeniami, a następnie trzech punktów lutowniczych na pakiecie drukowanym. W sumie mamy 3 miliony indywidualnie wykonanych połączeń (nie licząc połączeń między pakietami – milion tranzystorów nie zmieściłby się oczywiście na jednym pakiecie). Współczesny mikroprocesor zawierający wewnątrz milion tranzystorów ma kilkaset wyprowadzeń i tyle jest do wykonania indywidualnych połączeń. Porównując liczby indywidualnych wyprowadzeń w procesorze komputera z lat siedemdziesiątych i we współczesnym mikroprocesorze odnajdujemy znany już nam stosunek 1:10 000, czyli **4 rzędy wielkości**.

Nie jest to przypadek. Badania niezawodności urządzeń i systemów elektronicznych pokazują, że w typowym, poprawnie skonstruowanym i eksploatowanym urządzeniu uszkodzeniom nie ulegają elementy, lecz połączenia między nimi. Stąd bierze się poprawa niezawodności. Zastąpienie kilku milionów indywidualnie wykonanych połączeń między elementami przez kilkaset wyprowadzeń układu scalonego daje **proporcjonalne do zmniejszenia liczby indywidualnych połączeń zmniejszenie częstości występowania uszkodzeń**.

Podobnie wygląda zagadnienie kosztu. Przyjrzyjmy się temu bliżej. Układy scalone produkowane są na płytkach krzemowych o średnicach od 6 do 12 cali. Na takiej płytce mieści się od kilkuset dużych do kilkuset tysięcy małych układów scalonych. Koszt wytworzenia takiej płytki z układami waha się w granicach od kilku do kilkudziesięciu tysięcy dolarów, co oznacza, że typowy koszt wyprodukowania jednej struktury krzemowej układu scalonego może zawierać się w granicach od ułamka centa (dla małych, prostych układów) do kilku dolarów. Droższe są tylko struktury układów bardzo dużych, produkowanych przy zastosowaniu najbardziej zaawansowanych i najdroższych technologii. Po wyprodukowaniu struktury trzeba ją zmontować w obudowie i przetestować. Ten etap produkcji układu często kosztuje znacznie więcej, niż sama struktura krzemowa. Przykładowo, koszt zmontowania układu w obudowie ceramicznej typu PGA (takiej, w jakich m.in. montuje się układy mikroprocesorów) wynosi orientacyjnie od 20 do 150 dolarów, przy czym koszt ten jest z grubsza proporcjonalny do liczby wyprowadzeń układu (na ten koszt składa się zarówno cena samej obudowy, jak i koszt zmontowania układu – oba składniki są proporcjonalne do liczby wyprowadzeń). W przypadku elementów dyskretnych (pojedynczych tranzystorów) oraz małych układów scalonych, gdy sama struktura półprzewodnikowa kosztuje ułamek centa, koszt gotowego elementu jest praktycznie równy kosztowi obudowy i montażu! I teraz widać, że zmontowanie w obudowach kilku milionów struktur indywidualnych tranzystorów, a następnie wmontowanie ich w pakiety drukowane, musi kosztować o rzędy wielkości więcej, niż zmontowanie w obudowie i wmontowanie w pakiet pojedynczego układu scalonego. **Redukcja kosztu jest z grubsza proporcjonalna do**

**redukcji liczby indywidualnych wyprowadzeń.**

! Bardzo ważny wniosek, jaki z tego wynika, jest następujący: **zwiększenie stopnia scalenia (czyli zastąpienie dużej liczby prostych układów scalonych przez mniejszą liczbę układów bardziej złożonych) zawsze prowadzi do poprawy niezawodności i obniżki kosztu urządzenia lub systemu. To jest właśnie główna przyczyna stymulująca rozwój mikroelektroniki.**

## 1.3. Szybkość działania systemów elektronicznych

W wielu zastosowaniach duże znaczenie ma także szybkość działania urządzenia lub systemu. Zagadnieniem szybkości działania układów zajmiemy się szczegółowo w dalszych wykładach. Teraz wystarczy zauważyć, że elementy czynne (tranzystory) we współczesnych układach scalonych działają tak szybko, iż bardzo często nie one ograniczają szybkość działania układu. Szybkość działania układu jako całości jest coraz częściej ograniczona szybkością, z jaką propagowane są sygnały w połączeniach między elementami.

Jak wiemy, w próżni fala elektromagnetyczna przebywa w ciągu jednej sekundy około 300 000 km. To dużo, ale to oznacza, że w ciągu jednej nanosekundy (1 ns = $10^{-9}$ s) fala elektromagnetyczna może przebyć drogę najwyżej 30 centymetrów! Najszybsze dziś krzemowe układy cyfrowe pracują z częstotliwością zegara ponad 3 GHz, a więc w czasie trwania jednego taktu zegara sygnał może przebyć drogę rzędu zaledwie 10 cm. To jest tylko o rząd wielkości więcej, niż wynosi długość najdłuższych połączeń we współczesnych dużych układach scalonych! Jest oczywiste, że osiągnięcie tak dużych (a w przyszłości jeszcze większych) częstotliwości pracy układów cyfrowych jest możliwe tylko dzięki miniaturyzacji, jaką umożliwia mikroelektronika.

Problem czasu propagacji sygnałów występuje oczywiście nie tylko wewnątrz układów scalonych, lecz w jeszcze większym stopniu w przypadku połączeń pomiędzy układami scalonymi współpracującymi w urządzeniu, na przykład w przypadku płyt głównych komputerów, gdzie mamy mikroprocesor, układy pamięci, układy wejścia-wyjścia, układy obsługujące pamięci masowe (dysk, CD-ROM) itp. Wszystkie te układy są zmontowane na płycie drukowanej, a połączenia między nimi mają długości rzędu kilku do kilkudziesięciu centymetrów. Toteż występuje obecnie dążenie do tego, aby w jednym układzie scalonym scalać całe urządzenie lub system. Taki system zintegrowany (w jęz. angielskim **System on Chip**, w skrócie **SoC**) może zawierać w jednym układzie scalonym jeden lub więcej mikroprocesorów, pamięć, układy wejścia-wyjścia i wszystko inne, co potrzeba w kompletnym systemie. Zapewnia to nie tylko większą szybkość działania, ale także niższy koszt i lepszą niezawodność.

**! Biorąc pod uwagę wszystkie trzy czynniki: niezawodność, koszt i szybkość, można stwierdzić, że nie da się dziś wyprodukować żadnego wyrobu elektronicznego, który miałby szanse konkurencyjne na rynku, bez optymalnego wykorzystania w nim możliwości stwarzanych przez współczesną mikroelektronikę.**

## 1.4. Rola specjalizowanych układów scalonych

W początkowym okresie rozwoju mikroelektroniki układy scalone były projektowane wyłącznie przez ich producentów – firmy półprzewodnikowe. Konstruktorzy sprzętu mogli jedynie wybierać z katalogów firm półprzewodnikowych najbardziej odpowiadające im układy. Szybko okazało się, że z punktu widzenia zastosowań jest to duża niedogodność, ograniczająca kreatywność konstruktorów sprzętu. Nieco upraszczając, można dać taki przykład: jeśli kilku różnych producentów telewizorów skonstruowało swoje telewizory przy użyciu tego samego zestawu układów scalonych firmy XXX, to telewizory te z punktu widzenia funkcji i parametrów technicznych były takie same lub prawie takie same. Funkcje i parametry techniczne były bowiem określone przez taką, a nie inną konstrukcję użytych układów scalonych. Fakt ten jest dobrze znany tym, którzy interesują się komputerami PC. Jeśli wiadomo, że w płycie głównej firmy AAA użyto **zestawu układów** (zwanego z angielska **chipset**) BBB, to mniej więcej wiadomo, jakie są możliwości i parametry tej płyty.

A przecież na konkurencyjnym rynku producenci starają się, by ich wyroby wyraźnie odróżniały się od wyrobów konkurentów, by oferowały użytkownikowi nowe funkcje i lepsze parametry użytkowe! Odpowiedzią na to dążenie są specjalizowane układy scalone, powszechnie znane jako układy **ASIC** (jest to skrót od angielskiego terminu **Application-Specific Integrated Circuits)**. Są to układy przeznaczone do konkretnego wyrobu, a nie do sprzedaży na rynku dowolnemu odbiorcy. Układy ASIC znajdziemy dziś praktycznie w każdym urządzeniu elektronicznym, nie znajdziemy ich natomiast w handlowych katalogach.

Bardzo ważnym aspektem układów ASIC jest **ochrona własności intelektualnej**. Urządzenie zbudowane z elementów katalogowych, które każdy może kupić, jest łatwe do skopiowania przez konkurenta (w Polsce w ostatnich latach było sporo takich przypadków). Ochrona patentowa nie zawsze jest możliwa. Urządzenie z układem specjalizowanym jest praktycznie nie do skopiowania. Sam układ można wprawdzie "rozszyfrować" i zaprojektować podobny, ale rozszyfrowywanie układu scalonego jest tak pracochłonne i kosztowne, że w praktyce nieopłacalne. Ponadto projekt układu specjalizowanego można zarejestrować w urzędzie patentowym (procedura jest bardzo prosta i niewiele kosztuje). Podlega on wówczas ochronie przed bezprawnym skopiowaniem.

Są dwa sposoby uzyskania układu ASIC o potrzebnej nam funkcji. Pierwszy z nich polega na użyciu **układu programowalnego**. Istnieją dziś układy scalone zawierające w sobie zbiór bramek i bardziej złożonych bloków logicznych oraz programowalną sieć połączeń. Układy te, znane najczęściej pod ogólną nazwą **FPGA** (skrót od angielskiej nazwy **Field Programmable Logic Array**), bezpośrednio po wyprodukowaniu nie wykonują żadnej konkretnej funkcji. Funkcję tę definiuje użytkownik poprzez zaprogramowanie układu. Zaprogramowanie układu określa jego schemat logiczny, a tym samym wykonywaną funkcję.

Drugi sposób uzyskania układu ASIC polega na zaprojektowaniu go od początku i zleceniu produkcji firmie półprzewodnikowej. Istnieje wielu producentów specjalizujących się w produkcji układów scalonych na zamówienie, według projektu klienta. Takie firmy są na Dalekim Wschodzie, w USA, a także w Europie, m.in. w Austrii, Belgii, Francji, Niemczech. Każdy ze sposobów uzyskiwania układów ASIC ma swoje wady i zalety i swój zakres przydatności. Układy programowalne są dziś tak ważne i powszechnie stosowane, że jest o nich mowa w odrębnym wykładzie. W naszym wykładzie skoncentrujemy się na drugim sposobie tworzenia specjalizowanych układów scalonych, tj. na samodzielnym ich projektowaniu.

## 1.5. Mikroelektronika na świecie i w Polsce

We współczesnej mikroelektronice dominują pod względem wielkości produkcji i sprzedaży firmy z USA, Korei, Europy i Japonii. Do największych producentów zaliczane są firmy (dane z roku 2008):

- Intel (USA)
- Samsung (Korea Płd.)
- Toshiba (Japonia)
- Texas (USA)
- STMicroelectronics (Francja i Włochy)
- Renesas (Hitachi i Mitsubishi - Japonia)
- Sony (Japonia)
- Hynix (Korea Płd.)
- Infineon (Niemcy)
- AMD (USA)
- NEC (Japonia)
- Micron (USA)
- NXP (Holandia)

Wśród tych firm Europę reprezentują: Infineon (dawniej dział mikroelektroniki firmy Siemiens), STMicroelectronics i NXP (dawniej dział mikroelektroniki firmy Philips). Są to firmy europejskie, jednak w dobie globalizacji ich zakłady produkcyjne i biura projektowe rozproszone są po całym świecie (podobnie jak pozostałych firm).

Wszystkie wymienione wyżej firmy zajmują się przede wszystkim produkcją masową wyrobów katalogowych (np. Intel - mikroprocesory, Samsung, Hynix - pamięci półprzewodnikowe). Żadna z nich nie przyjmuje zamówień na produkcję specjalizowanych układów scalonych w krótkich seriach, chociaż niektóre - np. NXP - zajmują się produkcją układów specjalizowanych na własne potrzeby (w przypadku NXP - na potrzeby firmy Philips). Wszystkie te firmy dysponują najnowocześniejszymi technologiami i liniami produkcyjnymi, ponieważ w produkcji wielkoseryjnej wyrobów katalogowych tylko takie wyposażenie pozwala produkować wyroby o konkurencyjnej jakości i cenie. Wszystkie te firmy same projektują wszystkie lub znaczną większość produkowanych wyrobów.

Oprócz nich działa z powodzeniem wiele firm mniejszych. Niektóre z nich specjalizują się w określonych rodzajach technologii i wyrobów (np. układy dla elektroniki motoryzacyjnej), inne nastawione są na przyjmowanie zamówień na układy specjalizowane. Wśród tych ostatnich warto wymienić: TSMC (Taiwan Semiconductor Manufacturing Company), UMC (United Microelectronics Corporation), AMS (austriamicrosystems GMBH) i AMIS (AMI Semiconductor). Dwie pierwsze mieszczą się na Tajwanie i zajmują się głównie masową produkcją cyfrowych układów specjalizowanych w technologiach CMOS według projektów dostarczanych przez klientów. Przyjmują także mniejsze zamówienia wykonywane w technice płytek wieloprojektowych (co to znaczy? - dowiesz się z następnego wykładu). AMS jest stosunkowo niewielką, ale bardzo solidną firmą europejską zajmującą się produkcją układów specjalizowanych różnych rodzajów, rownież w krótkich seriach, oferuje wiele zróżnicowanych procesów technologicznych. AMI Semiconductor jest firmą amerykańską, mającą jednak jeden z dwóch głównych zakładów produkcyjnych w Belgii. Ma podobny profil działalności, jak AMS. TSMC, UMC, AMS i AMIS są to producenci łatwo dostępni dla polskich firm (w jaki sposób? - dowiesz się z następnego wykładu).

Nowym zjawiskiem na rynku dostawców układów scalonych są firmy, które w ogóle nie posiadają własnych linii produkcyjnych, a całą produkcję zlecają innym (najczęściej firmom takim, jak TSMC, UMC czy AMIS). Są to firmy specjalizujące się w projektowaniu i dostarczaniu bardzo zaawansowanych układów do specyficznych zastosowań. Przykłady to firma Broadcom, specjalizująca się w układach do szybkiej transmisji danych cyfrowych (w tym do transmisji bezprzewodowej, jak Bluetooth czy też sieci WiFi), czy też dobrze znana wielbicielom gier komputerowych firma NVidia, która dostarcza najbardziej zaawansowane układy do kart graficznych komputerów. Firmy takie określane są mianem "Fabless" - po polsku "bez fabryki". Ich udział w rynku szybko rośnie.

Osobną grupą producentów są wytwórcy układów programowalnych (FPGA). Królują tu firmy amerykańskie, dwie najważniejsze to Altera i Xilinx. Nie będziemy poświęcać im uwagi, ponieważ układom programowalnym poświęcona jest duża część innego przedmiotu Twoich studiów.

W ostatnim czasie bardzo gwałtownie rozwija się produkcja w Chinach. Ekonomiści przewidują, że w niedługim czasie blisko 30% całej światowej produkcji sprzętu elektronicznego dostarczać będą wytwórnie zlokalizowane w Chinach, i towarzyszyć temu będzie odpowiednio duży procent udziału w światowym rynku mikroelektronicznym.

Wielkość światowego rynku mikroelektronicznego oceniana jest w końcu pierwszej dekady XXI wieku na około 280 miliardów USD, co w skali gospodarki światowej nie jest wielką liczbą, ale mikroelektronika zasila swymi wyrobami cały przemysł sprzętu elektronicznego, którego wielkość oceniana jest na 1200 miliardów USD, i wiele innych gałęzi przemysłu. Jeśli doliczyć do tego usługi, które nie istniałyby bez sprzętu elektronicznego, to można

ocenić, że od mikroelektroniki zależy bezpośrednio lub pośrednio 30% całej światowej gospodarki.

W Polsce nie ma obecnie żadnego producenta układów scalonych. W latach 1975 - 1990 istniało Naukowo-Produkcyjne Centrum Półprzewodników CEMI zrzeszające kilka zakładów produkcyjnych (w tym fabrykę TEWA w Warszawie) oraz instytucji badawczo-rozwojowych i projektowych. CEMI było przez pewien czas znaczącym w naszej części Europy producentem układów scalonych specjalizującym się głównie w układach do sprzętu powszechnego użytku. Strategia CEMI (wymuszona przez system gospodarki planowej) polegała na masowej produkcji układów katalogowych będących funkcjonalnymi odpowiednikami wyrobów innych producentów. W latach osiemdziesiątych zaprzestano w CEMI jakichkolwiek poważniejszych inwestycji. Po otwarciu polskiego rynku w roku 1990 produkcja CEMI straciła sens ekonomiczny, ponieważ katalogowe wyroby CEMI, w dużej części przestarzałe i produkowane na bardzo przestarzałych liniach produkcyjnych, nie miały żadnych szans konkurowania z takimi samymi wyrobami z innych firm. CEMI nie otrzymało żadnego wsparcia i mimo zainteresowania kilku potencjalnych zagranicznych inwestorów uległo likwidacji. Ośrodkiem badawczo-rozwojowym w dziedzinie mikroelektroniki, który pozostał po likwidacji CEMI, jest Instytut Technologii Elektronowej w Warszawie (ITE). Dysponuje on niewielką linią produkcyjną w Piasecznie pod Warszawą. Służy ona głównie celom badań i małoseryjnej produkcji w dziedzinie czujników i mikromechanizmów krzemowych (w tej dziedzinie ITE osiągnął znaczny dorobek), na linii tej można też wytwarzać układy scalone CMOS. Instytut przyjmuje także zamówienia na projekty układów specjalizowanych, które mogą być następnie wytwarzane w wytwórniach zagranicznych.

Doświadczalna linia produkcyjna Instytutu Technologii Elektronowej w Piasecznie pod Warszawą: po lewej pracownia fotolitografii, po prawej piece, w których wykonuje się operacje wysokotemperaturowe.

## Dodatek: ochrona własności intelektualnej w przypadku projektów układow scalonych

Ochrona własności intelektualnej w przypadku projektów układów scalonych ma specyficzną formę. Ochronie podlega wyłącznie projekt topografii układu, tj. jego struktury fizycznej: rozmieszczenia elementów i połączeń między nimi. Ustawa określa topografię układu scalonego prawniczym językiem w następujący, nieco zawiły sposób:

**"Oryginalne rozwiązanie polegające na wyrażonym w dowolny sposób przestrzennym rozplanowaniu elementów, z których co najmniej jeden jest aktywny oraz połączeń między nimi, przy czym pojęcie "układ scalony" oznacza warstwowy wytwór przestrzenny, utworzony w celu spełniania funkcji elektronicznych z elementów materiału półprzewodnikowego tworzącego ciągłą warstwę oraz ich połączeń przewodzących i obszarów izolujących, które są ze sobą nierozdzielnie sprzężone."**

Nie podlega ochronie funkcja układu, jego schemat logiczny ani elektryczny. W praktyce oznacza to, że zaprojektowanie i produkcja układu scalonego, którego funkcja, schemat i parametry są takie same, jak innego, już istniejącego ukladu, jest **całkowicie legalne**, pod warunkiem, ż e topografia nowego układu jest zaprojektowana od początku i nie stanowi kopii topografii układu już istniejącego. Całkowicie legalne jest także wykonywanie kopii topografii już istniejącego układu w postaci na przykład zdjęć fotograficznych, analiza tej topografii, odczytanie z niej schematu elektrycznego układu itp. - pod warunkiem, że kopia topografii nie posłuży bezpośrednio do produkcji kopii istniejącego układu. Wolno nawet zaprojektować nowy układ wzorując się na topografii już istniejącego układu (na przykład rozmieszczając elementy układu w podobny sposób, co w istniejącym układzie), byle tylko nowy projekt był zrobiony samodzielnie od początku, i nie był ani w całości, ani nawet we fragmentach wierną kopią topografii układu już istniejącego.

Taki sposób określenia zakresu ochrony ma na celu pogodzenie dwóch sprzecznych celów. Z jednej strony chodzi o to, by twórca projektu topografii mógł osiągnąć spodziewane korzyści z wykonanej pracy, nie będąc narażonym na nieuczciwą konkurencję. Projektowanie topografii jest najbardziej pracochłonnym i najtrudniejszym, a więc najbardziej kosztownym etapem całego procesu projektowania układu scalonego. Nieuczciwy konkurent, kopiując topografię układu i produkując przy jej użyciu wierną kopię układu, mógłby sprzedawać ten układ po znacznie obniżonej cenie, bo nie poniósł wysokich kosztów projektowania. Dlatego właśnie projekt topografii podlega ochronie. Z drugiej strony, zakres ochrony jest określony w taki sposób, by nie hamować postępu technicznego, który w dużej mierze polega na studiowaniu wcześniejszych rozwiązań technicznych i ulepszaniu ich.

Topografia układu scalonego podlega ochronie, jeśli została zarejestrowana w Urzędzie Patentowym. Rejestrować można wyłącznie topografie oryginalne, tj. nie będące kopiami innych topografii. Urząd Patentowy nie prowadzi jednak żadnych badań oryginalności i rejestruje każdą zgłoszoną topografię. Jeżeli zarejestrowana jest topografia, która nie jest oryginalna (tj. jest kopią innej topografii), to twórca oryginalnej topografii ma prawo zażądać unieważnienia rejestracji. Procedura rejestracji jest prosta, szybka i nie pociąga dużych kosztów.

W Polsce okres ochrony zarejestrowanej topografii układu scalonego wynosi 10 lat.

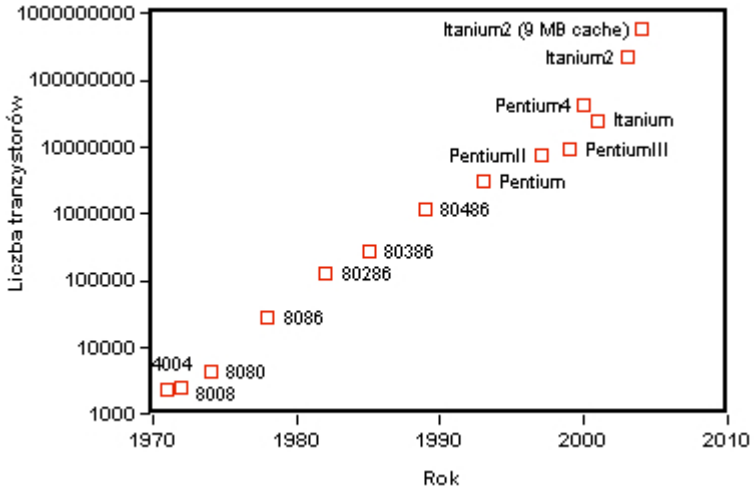
Szczegóły zawiera ustawa "Prawo własności przemysłowej" z dnia 30 czerwca 2000 r, z późniejszymi zmianami (tekst jednolity: Dz. U. z 2003 r. Nr 119, poz. 1117). Więcej można znaleźć na stronach internetowych Urzędu Patentowego: http://www.uprp.pl/Polski

Ochrona topografii układów scalonych jest niezależna od innych form ochrony własności intelektualnej. Przykładowo, jeśli schemat elektryczny układu zawiera oryginalne rozwiązanie mające charakter wynalazku, to można takie rozwiązanie opatentować na ogólnych zasadach.

# ZADANIA DO WYKŁADU 1

**Zadanie 1**

Oszacuj na podstawie rys. 1.4 (powtórzonego poniżej) dokładność prognozy Moore'a. Czy rzeczywiście liczba tranzystorów w mikroprocesorach Intela podwajała się średnio co dwa lata?




Rys. 1.4. Liczba tranzystorów w kolejnych procesorach Intela

## Bibliografia

Poniższe podręczniki pomogą przypomnieć podstawowe zagadnienia z przyrządów półprzewodnikowych, układów elektronicznych i technologii stosowanych w mikroelektronice.

[1] A. Filipkowski, "*Układy elektroniczne analogowe i cyfrowe*", WNT Warszawa 2006
[2] W. Marciniak, "*Przyrządy półprzewodnikowe i układy scalone*", WNT Warszawa 1987
[3] Z. Nosal, J. Baranowski, "*Układy elektroniczne cz. 1; układy analogowe liniowe*", WNT Warszawa 2003
[4] J. Kalisz, "*Podstawy elektroniki cyfrowej*", WKiŁ Warszawa 2002
[5] P Horowitz, W. Hill, "*Sztuka elektroniki*", WKiŁ Warszawa 2003
[5] R. A. Colclaser, D. A. Neamen, C. F. Hawkins, "*Electronic circuit analysis, basic principles*", J.Wiley&Sons 1984
[6] R. Geiger, P. Allen, N. Strader, "*VLSI Design Techniques for Analog and Digital Circuits"*, McGraw-Hill, Inc. 1990

Bardzo obszerną wiedzę teoretyczną i praktyczną z dziedziny elektroniki zawiera książka

[7] P. Horowitz, W.Hill, "Sztuka elektroniki", WKiŁ Warszawa**,** wyd. 9, 2009

# Wykład 2: Aspekty ekonomiczne mikroelektroniki

## Wstęp

W tym wykładzie zapoznasz się dokładniej z ekonomicznymi aspektami mikroelektroniki. Są one ważne z dwóch powodów. Po pierwsze, trzeba się w nich orientować, by stosować w sprzęcie układy scalone w optymalny sposób. Po drugie, wiele rozwiązań technicznych spotykanych w mikroelektronice podyktowanych zostało właśnie przez względy ekonomiczne.

Dowiesz się, od czego zależy koszt układu scalonego i co z tego wynika dla konstruktora sprzętu. Dowiesz się, jak wybrać najwłaściwszą technologię produkcji do układu, który zamierzasz zaprojektować. Dowiesz się także, kto i gdzie wykona Ci prototyp układu, jak się to załatwia i ile to kosztuje. Dowiesz się, gdzie i jak będziesz mógł zamówić większą serię układów, gdy prototyp spełni Twoje oczekiwania. A co najważniejsze – zobaczysz, że to wszystko leży w zasięgu ręki. Dostęp do najnowocześniejszych technologii mikroelektronicznych jest w Polsce równie łatwy, jak w krajach najbardziej zaawansowanych technologicznie.

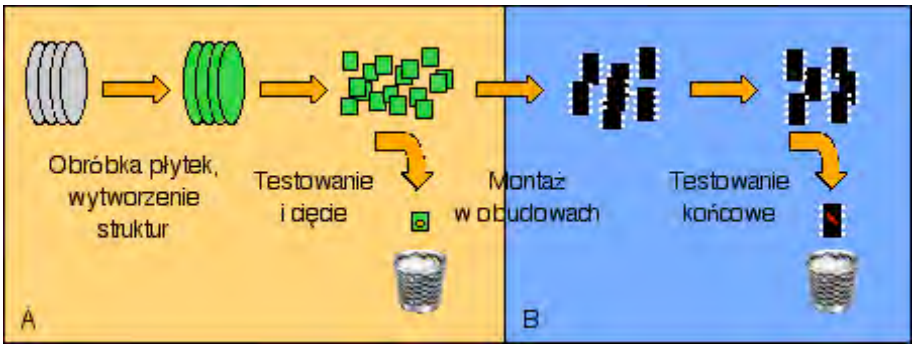
## 2.1. Czynniki określające koszt układu scalonego

Całkowity koszt jednego egzemplarza układu scalonego jest sumą dwóch składników:

- $K_{prod}$ - koszt wytworzenia jednego egzemplarza układu, jego montażu i testowania.
- $K_{proj}$ - część kosztu zaprojektowania układu i przygotowania jego produkcji przypadająca na jeden wyprodukowany egzemplarz układu

Teraz zajmiemy się pierwszym składnikiem $K_{prod}$. Drugi składnik $K_{proj}$ będzie omawiany dalej.

Proces wytworzenia gotowego układu scalonego można podzielić na dwie fazy, zilustrowane symbolicznie na rysunku 2.1.




Rys. 2.1: Wytwarzanie układu scalonego: faza A - wytworzenie struktur, testy ostrzowe i cięcie, faza B - montaż i testy końcowe

W pierwszej fazie (A na rys. 2.1) płytki półprzewodnikowe przechodzą proces obróbki, w wyniku którego powstają na nich struktury układów scalonych. Wszystkie wytworzone struktury są kolejno poddawane testom zwanym **testami ostrzowymi**. Kontakt z elektrycznymi wyprowadzeniami każdej struktury na płytce zapewniają sprężyste ostrza. Automatyczny tester doprowadza do wejść sygnały testowe i bada prawidłowość sygnałów wyjściowych układu. Układy niesprawne są znakowane farbą, co pozwala je później odsortować i odrzucić. Dopiero po wykonaniu testów ostrzowych płytka jest cięta na poszczególne struktury. Struktury zakwalifikowane jako sprawne przechodzą do drugiej fazy (B na rys. 2.1). Struktury są montowane w obudowach, a po montażu następują **testy końcowe**. W tych testach niektóre układy okazują się niesprawne mimo przejścia przez testy ostrzowe. Zmontowane układy mogą okazać się niesprawne z dwóch powodów. Po pierwsze, testy ostrzowe nie zawsze pozwalają zbadać dostatecznie dokładnie działanie układu. Po drugie, układ może ulec uszkodzeniu w trakcie montażu lub też montaż może być wykonany wadliwie.

Załóżmy, że partia produkcyjna układów scalonych składa się z $L_p$ płytek o powierzchni $A_p$ każda. Załóżmy, że na tych płytkach wytwarzany jest układ scalony o powierzchni $A_u$. Oznacza to, że (w uproszczeniu) całkowita liczba struktur $N_u$, które zostaną wyprodukowane, wynosi

$$N_u = L_p \frac{A_p}{A_u} \tag{2.1}$$

Z tych struktur część odpadnie w testach ostrzowych, pozostanie $N_{so}$ struktur, które zakwalifikowane zostały jako sprawne w testach ostrzowych. Stosunek $N_{so}/N_u$ nazwiemy **uzyskiem produkcyjnym** $u_p$. Dla dobrze zaprojektowanych układów produkowanych w dojrzałym procesie produkcyjnym powinien on być bliski 1, ale w praktyce nigdy tej wartości nie osiąga.

$$u_p = \frac{N_{so}}{N_u} \tag{2.2}$$

Struktury zakwalifikowane jako sprawne, w liczbie równej $N_{so}$, zostaną zmontowane, po czym po testach końcowych pozostanie z nich $N_{su}$ gotowych, sprawnych układów. Stosunek $N_{su}/N_{so}$ nazwiemy **uzyskiem montażu** $u_m$. Ostatecznie, z początkowej liczby $N_u$ pozostanie $N_{su}$ sprawnych układów. Stosunek $N_{su}/N_u$

nazwiemy **uzyskiem ostatecznym** $u$. Jest on iloczynem uzysków produkcyjnego i montażowego i w praktyce zawsze jest mniejszy od jedności.

$$u = \frac{N_{su}}{N_u} = u_p u_m \tag{2.3}$$

Niech koszt wykonania wszystkich operacji technologicznych wykonywanych w fazie A, dla jednej partii produkcyjnej, wynosi $K_A$. Koszt ten jest dla danej technologii praktycznie stały, nie zależy od liczby płytek w partii produkcyjnej ani od liczby i rodzaju układów wykonywanych na tych płytkach. Dalsze operacje - montaż i testy końcowe - są wykonywane na każdej strukturze osobno. Dlatego ich koszt jest wprost propocjonalnych do liczby tych struktur. Niech dla jednej struktury wynosi on $k_s$. Koszt ten jest w pierwszym przybliżeniu proporcjonalny do liczby wyprowadzeń układu scalonego, bowiem od tej liczby zależy zarówno koszt obudowy, jak i pracochłonność montażu oraz testowania.

Możemy już teraz policzyć, ile kosztuje wyprodukowanie jednego sprawnego układu scalonego. Koszt $k_s$ należy pomnożyć przez liczbę układów poddanych montażowi i testom końcowym $N_{so}$, zsumować z kosztem $K_A$, a następnie sumę tę podzielić przez liczbę gotowych sprawnych układów $N_{su}$. Otrzymujemy następującą zależność określającą całkowity koszt $K_{prod}$ wytworzenia jednego sprawnego układu

$$K_{prod} = \frac{1}{N_{su}}\left(k_s N_{so} + K_A\right) = \frac{1}{u_m}\left(k_s + \frac{K_A}{u_p}\frac{A_u}{L_p A_p}\right) \tag{2.4}$$

Ze wzoru (2.4) widać, że

- **koszt wytworzenia jednego egzemplarza układu scalonego składa się z dwóch składników: jeden z nich ($k_s$) jest w przybliżeniu proporcjonalny do liczby zewnętrznych wyprowadzeń układu, drugi do jego powierzchni $A_u$, przy czym proporcje tych składników mogą być różne,**
- **układ jest tym droższy, im niższe są uzyski: produkcyjny $u_p$ i montażu $u_m$.**

Jak zobaczymy dalej, te stwierdzenia mają szereg konsekwencji ważnych dla projektanta układów.

Rozważymy jeszcze, co się stanie w przypadku rezygnacji z testów ostrzowych (jak zobaczymy dalej, w pewnych przypadkach przy produkcji układów specjalizowanych testów ostrzowych nie wykonuje się). W takim przypadku wszystkie wyprodukowane struktury są montowane w obudowach, a w testach końcowych odrzucane są zarówno te, które od początku były wadliwe, jak i te, które zostały uszkodzone przy montażu lub źle zmontowane. Jak łatwo sprawdzić, wzór (2.4) uzyskuje wówczas postać

$$K_{prod} = \frac{1}{u_m u_p}\left(k_s + K_A \frac{A_u}{L_p A_p}\right) \tag{2.5}$$

Gdyby uzysk produkcyjny $u_p$ był równy jedności, koszt pozostałby bez zmian (co jest oczywiste). Ale jeśli ten uzysk jest znacznie mniejszy od jedności, to koszt określony wzorem (2.5) jest wyraźnie wyższy od kosztu danego wzorem (2.4). Przyczyna jest jasna: jeśli montowane są wszystkie struktury, również te, które nie działają prawidłowo (byłyby odrzucone w testach ostrzowych), to ponosi się niepotrzebnie dodatkowy koszt montażu i testowania tych wadliwych struktur.

## 2.2. Defekty, rozrzuty produkcyjne i uzysk

Dlaczego pewna część wyprodukowanych układów scalonych okazuje się być wadliwa (nie spełnia wymagań technicznych) i w testach ostrzowych lub końcowych zostaje odrzucona?

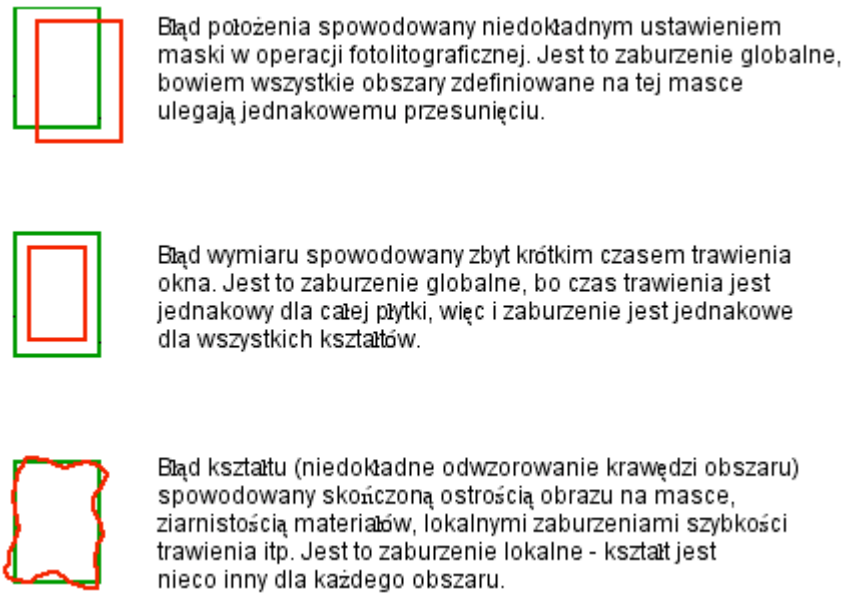
Wyprodukowany układ może w ogóle nie działać (nie wykonywać swej funkcji). O takim układzie mówimy, że wystąpiło w nim **uszkodzenie katastroficzne**. Może być i tak, że układ działa, ale jego parametry nie mieszczą się w dopuszczalnych granicach zwanych **granicami tolerancji**. O takim układzie mówimy, że wystąpiło w nim **uszkodzenie parametryczne**.

Najczęstszą przyczyną uszkodzeń katastroficznych są **defekty** zwane **strukturalnymi**, zmieniające w istotny sposób fizyczną strukturę układu i jego elektryczny schemat. Przykłady takich uszkodzeń to zwarcie ścieżek połączeń, przerwa w takiej ścieżce, "dziura" w tlenku bramkowym tranzystora MOS powodująca zwarcie bramki do położa itp. Te defekty powstają zwykle na skutek zanieczyszczeń pyłkami kurzu podczas wykonywania operacji fotolitograficznych. Konstruktor układu nie ma w praktyce istotnego wpływu na występowanie defektów strukturalnych. Nie będziemy ich więc omawiali w tym wykładzie.

Przyczyną uszkodzeń parametrycznych są zwykle nadmierne **rozrzuty producyjne**. Te omówimy dokładniej, bowiem ich występowanie ma bezpośredni i silny wpływ na to, jak projektuje się układy scalone.

W mikroelektronice - jak w każdej innej dziedzinie techniki - we wszystkich operacjach technologicznych występują nieuchronnie zaburzenia powodujące, że wyniki operacji nigdy nie są dokładnie zgodne z zamierzonymi. Zaburzenia te nazywamy **rozrzutami produkcyjnymi**. Procesy produkcyjne mikroelektroniki są niezwykle subtelne, toteż nawet znikomo małe zaburzenia tych procesów prowadzą do dużych rozrzutów parametrów i charakterystyk elementów układów scalonych. Nie należą do rzadkości rozrzuty na poziomie na przykład +/- 50% wartości nominalnej danego parametru - rzecz nie do pomyślenia w innych dziedzinach techniki, np. w mechanice. Sztuka projektowania układów scalonych polega między innymi na tym, by z elementów o bardzo dużych rozrzutach parametrów zbudować układ, który nie tylko będzie działał, ale którego parametry użytkowe będą utrzymane w wąskich granicach tolerancji.

Jest to możliwe, jeśli poznamy dokładniej naturę rozrzutów produkcyjnych. Można je podzielić na dwa rodzaje: **rozrzuty globalne** i **rozrzuty lokalne**. Rozrzuty globalne to takie, które jednakowo oddziałują na wszystkie elementy w układzie scalonym, zaś rozrzuty lokalne to takie, które dla każdego elementu mają inną wielkość. Innymi słowy, gdyby istniały tylko rozrzuty globalne, to w danym układzie elementy identycznie zaprojektowane miałyby zawsze identyczne parametry i charakterystyki (które jednak miałyby różne wartości w układach pochodzących z różnych płytek i różnych serii produkcyjnych). Natomiast **rozrzuty lokalne powodują istnienie różnic pomiędzy elementami, które powinny być identyczne.** Różnicę między rozrzutami globalnymi i lokalnymi wygodnie jest pokazać na przykładzie fotolitografii. Fotolitografia jest to proces technologiczny (będzie on bardziej szczegółowo omówiony w wykładzie 3) służący do odwzorowania na płytce półprzewodnikowej kształtów obszarów tworzących elementy układu scalonego oraz połączenia między tymi elementami. Obszary uzyskane nigdy nie są identyczne z zaprojektowanymi. Trzy rodzaje zaburzeń prowadzących do powstawania różnic między kształtem zaprojektowanym, a uzyskanym, pokazuje rys. 2.2.




Rys.2.2. Trzy rodzaje zaburzeń obserwowanych w procesach fotolitografii.

Zielony kontur: kształt obszaru według projektu, czerwony kontur: kształt rzeczywiście otrzymany

Podział rozrzutów na lokalne i globalne nie dotyczy wyłącznie fotolitografii. Praktycznie wszystkie rozrzuty produkcyjne mają składową globalną i składową lokalną. Przykładowo, podczas procesów wysokotemperaturowych kilkadziesiąt płytek znajduje się w piecu w różnych miejscach, każda z nich w nieco innej temperaturze. Jest to rozrzut globalny. Różnice temperatury występują jednak także w obrębie każdej płytki. Jest to rozrzut lokalny.

! **Cechą charakterystyczną mikroelektroniki jest to, że chociaż rozrzuty globalne są bardzo duże, to równocześnie rozrzuty lokalne są małe. Innymi słowy, *nie można* liczyć na to, że wyprodukowane elementy będą miały parametry zawsze bardzo bliskie nominalnym, ale *można* liczyć na to, że para elementów zaprojektowanych jako identyczne i znajdujących się tuż obok siebie w tym samym układzie scalonym będzie miała prawie identyczne parametry. Tę własność powszechnie wykorzystuje się w projektowaniu układów scalonych, a zwłaszcza układów analogowych.**

## 2.3. Pracochłonność i koszt projektu układu

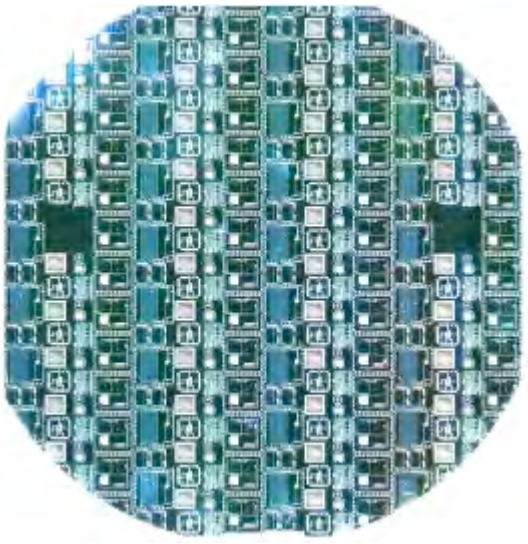
Do kosztu wytworzenia układu należy doliczyć koszt $K_{proj}$ zaprojektowania układu i przygotowania produkcji. Jeśli całkowita liczba potrzebnych układów (wielkość serii produkcyjnej) wynosi $S$, to całkowity koszt jednego egzemplarza układu $K_{calk}$ wynosi

$$K_{calk} = K_{prod} + \frac{K_{proj}}{S} \tag{2.6}$$

Koszt zaprojektowania układu jest proporcjonalny do potrzebnych do tego nakładów pracy. Te mogą być bardzo wysokie. Kiedyś, gdy układy scalone liczyły nie więcej niż kilkadziesiąt elementów, cały proces projektowania był wykonywany przez człowieka (ale przy wykorzystaniu wspomagających programów komputerowych): projektant opracowywał schemat układu (logiczny i/lub elektryczny), wykonywał wszystkie obliczenia projektowe, projektował poszczególne elementy, rozmieszczał je w układzie i projektował połączenia. Ten sposób projektowania nazywany jest **projektowaniem w stylu *full custom*** (brak udanego polskiego terminu). Ma on i dziś pewien obszar zastosowań, o czym będzie mowa w wykładach 5 i 6. Pracochłonność tego sposobu projektowania można w dużym przybliżeniu oszacować następująco: doświadczony projektant mający do dyspozycji współczesne oprogramowanie wspomagające zużywa, średnio biorąc, jedną godzinę pracy na jeden tranzystor. Czy to dużo? Jeżeli układ liczy **1000 tranzystorów**, to pracochłonność projektu wynosi około **7 osobo-miesięcy**. Ale dla mikroprocesora mającego **40 milionów tranzystorów** (np. klasy Pentium 4) otrzymujemy pracochłonność rzędu **20 000 osobo-lat**! Do zagadnienia procochłonności projektowania powrócimy w wykładach 5 i 6. W wykładzie 6 poznamy sposoby wspomaganego i zautomatyzowanego projektowania, które pozwalają zmniejszyć pracochłonność o wiele rzędów wielkości. Niemniej, projekt dużego i złożonego układu może być kosztowny. Na szczęście typowe układy ASIC są daleko mniej złożone, niż najnowocześniejsze mikroprocesory. Przy wykorzystaniu metod projektowania omawianych w wykładzie 6 wykonanie projektu cyfrowego układu specjalizowanego zajmuje projektantowi zwykle od kilku tygodni do kilku miesięcy. Redukuje to koszt wykonania projektu do rozsądnego poziomu.

Koszt przygotowania produkcji to w praktyce koszt wykonania kompletu masek, które służą w procesach fotolitografii do określenia kształtów, wymiarów i położenia wszystkich elementów i połączeń w układzie. Dla starszych, mniej zaawansowanych technologii (gdzie minimalny wymiar w układzie jest rzędu 1 mikrometra lub nieco mniej) koszt wykonania takiego kompletu wynosi kilkanaście do kilkudziesięciu tysięcy dolarów. Dla najnowocześniejszych procesów technologicznych (minimalny wymiar na poziomie 90 nanometrów i poniżej) potrzebne są maski o znacznie bardziej skomplikowanej strukturze (t.zw. maski z kontrastem fazowym). Koszt wykonania kompletu takich masek może sięgać nawet miliona dolarów. Jest oczywiste, że gdyby dla każdego układu specjalizowanego wykonywany był komplet masek za milion dolarów, to układy specjalizowane miałyby sens jedynie przy bardzo dużych seriach produkcyjnych, rzędu co najmniej setek tysięcy sztuk.

Istnieje jednak rozwiązanie tego problemu, w postaci **płytek wieloprojektowych** (**Multi-Project Wafer**, w skrócie **MPW** ). Płytką wieloprojektową nazywamy płytkę, na której wytwarza się równocześnie wiele różnych układów według różnych projektów. Całkowity koszt wykonania kompletu masek rozkłada się wtedy na wiele projektów.



Rys. 2.3. Płytka wieloprojektowa. Na płytce znajduje się 7 różnych układów scalonych CMOS. Zdjęcie płytki wyprodukowanej na linii produkcyjnej Instytutu Technologii Elektronowej w Warszawie.

W jednej partii produkcyjnej wytwarzane są w ten sposób układy dla bardzo wielu różnych odbiorców. Dzięki takiej organizacji produkcji, oraz zautomatyzowanym metodom projektowania, układy scalone mogą być opłacalne również wtedy, gdy potrzeba ich niewiele - kilkaset lub kilka tysięcy egzemplarzy, a nawet pojedyncze sztuki.

Producenci układów na ogół nie chcą się zajmować kontaktami z setkami drobnych klientów zamawiających niewielkie ilości układów specjalizowanych. Dlatego powstały instytucje pośredniczące. Mają one zawarte umowy z producentami, zbierają projekty od klientów, sprawdzają czy projekty te są prawidłowe (w sensie zgodności z wymaganiami producenta), łączą zebrane projekty ze sobą i wysyłają do producenta. Producent wykonuje maski wieloprojektowe, wytwarza płytki i odsyła. Instytucja pośrednicząca organizuje montaż: wysyła płytki do firmy, która tnie płytki na poszczególne struktury i montuje w obudowach. Po 2 - 3 miesiącach od wysłania projektu klient otrzymuje zamówione układy. Zamawia się zazwyczaj najpierw prototyp układu (10 - 20 egzemplarzy). Gdy prototypowe układy spełniają wymagania, zamawia się potrzebną ilość układów. Jeśli jest ona niewielka, zostanie wyprodukowana w technice płytek wieloprojektowych. Przy dużej serii (zwykle od kilkudziesięciu tysięcy układów wzwyż) producent może uznać, że celowe jest wykonanie indywidualnego kompletu masek dla zamówionego układu i przeznaczenie na ten układ całych płytek lub nawet partii produkcyjnych.

W przypadku układów produkowanych metodą płytek wieloprojektowych nieco odmiennie, niż przy zwykłej produkcji seryjnej, wygląda testowanie. Płytki wieloprojektowe nie są testowane ostrzowo. Testowane są dopiero gotowe, zmontowane w obudowach układy. Oznacza to, że produkcja płytek wieloprojektowych ma sens tylko wtedy, gdy uzysk produkcyjny $u_p$ jest bliski 1 - porównaj wzór (2.5).

W Europie najważniejszą instytucją pośredniczącą między klientami zamawiającymi układy specjalizowane, a ich producentami jest EUROPRACTICE. Praktyczne informacje na ten temat znajdziesz dalej.

## 2.4. Wybór optymalnej technologii

Możemy teraz rozważyć następujący problem. Mamy do zaprojektowania i wykonania układ specjalizowany określonego rodzaju. Do dyspozycji mamy kilka różnych technologii wytwarzania układów jednego lub różnych producentów. Którą z tych technologii wybrać? Zakładamy przy tym, że wszystkie rozważane technologie zapewniają spełnienie wymagań technicznych dla naszego układu. Wyboru należy dokonać na podstawie analizy ekonomicznej - która z dostępnych technologii zapewni nam najniższy koszt układu?

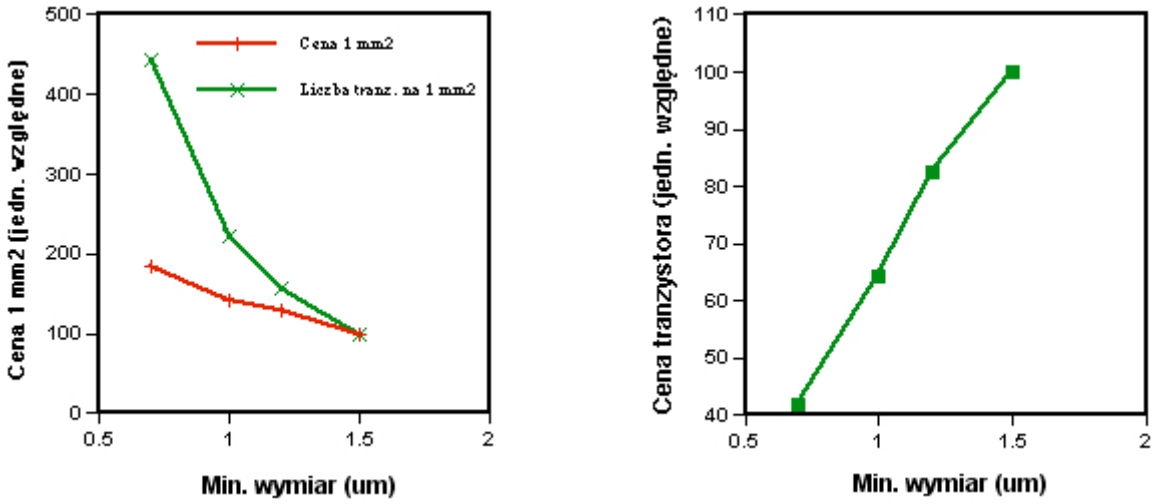
Jak już wiemy, na koszt jednego egzemplarza układu składa się koszt wytworzenia struktur oraz koszt montażu. Gdy zamawiamy układy specjalizowane, oba składniki liczone są odrębnie. Cena za wyprodukowanie struktur obliczana jest proporcjonalnie do powierzchni układu. Dla przykładu, jeden z europejskich producentów przy zamawianiu prototypów układów CMOS stosuje następujące ceny (podane są ceny dla trzech generacji technologii, z malejącą minimalną długością kanałów tranzystorów, od 0,7 mikrometra do 0,35 mikrometra):

Tablica 4.1. Przykładowe ceny 1 $mm^2$ układu CMOS w trzech różnych technologiach

| Technologia | Cena za 1 $mm^2$ układu w EURO |
|---|---|
| CMOS 0.7 | 280 |
| CMOS 0.5 | 390 |
| CMOS 0.35 | 590 |

Za tę cenę otrzymuje się 20 struktur nieobudowanych i nie testowanych.

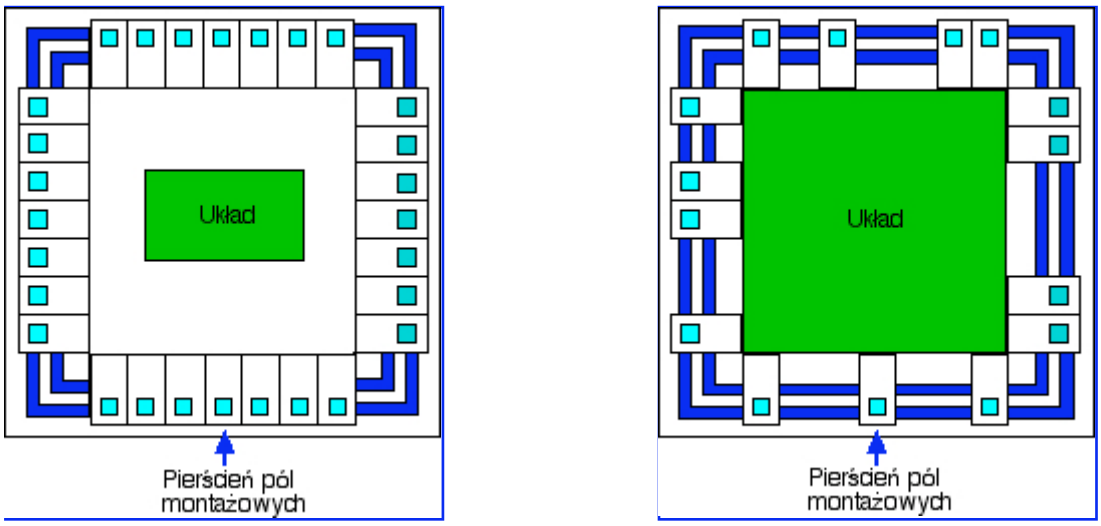
Jeżeli każda z tych trzech technologii zapewnia wymagane parametry potrzebnego nam układu, to na pierwszy rzut oka wydawałoby się, że należy wybrać technologię najtańszą (CMOS 0.7). Może się jednak okazać, że w rzeczywistości układ wykonany w tej technologii będzie najdroższy. Rzecz w tym, że liczba elementów, jakie można zmieścić na 1 $mm^2$ układu, ze zmniejszaniem się minimalnego wymiaru rośnie szybciej, niż cena 1 $mm^2$. Innymi słowy, cena **jednego milimetra kwadratowego** jest najniższa dla technologii CMOS 0.7, ale cena za **jeden tranzystor** w układzie może być najniższa dla technologii CMOS 0.35. Ilustruje to przykład dla czterech generacji technologii innego producenta:




Rys. 2.4. Cena jednego $mm^2$ układu i średnia liczba tranzystorów na 1 $mm^2$ dla czterech generacji technologii CMOS
oraz wynikający z tego średni koszt jednego tranzystora (w jednostkach względnych).

Należy więc ustalić, jaką powierzchnię zajmie układ zaprojektowany w każdej z dostępnych technologii i dopiero wtedy można zdecydować, która technologia da nam układ najtańszy. Trzeba rozważać układ kompletny, tj. wnętrze wraz z pierścieniem pól montażowych, do których dołączane są zewnętrzne wyprowadzenia (więcej o technologii montażu w wykładzie 3). Może się okazać, że jeśli nasz układ jest prosty i ma niewiele elementów, to jego powierzchnia będzie określona przez pierścień pól montażowych, a nie przez sam układ. Pola montażowe (ilustracja dalej) muszą mieć określone wymiary (zwykle 100 x 100 mikrometrów) niezależnie od technologii i muszą znajdować się w określonych odstępach, aby dołączenie wyprowadzeń było możliwe. Jeśli więc układ jest prosty i ma niewiele elementów, a równocześnie ma dużą liczbę zewnętrznych wyprowadzeń, to całkowita

powierzchnia może być określona przez pierścień pól montażowych, a wewnątrz tego pierścienia pozostaje wolna, nie wykorzystana powierzchnia. Taką sytuację ilustruje rys. 2.5a. Układ o dużej liczbie elementów i małej liczbie wyprowadzeń jest pokazany na rys. 2.5b. W tym przypadku wnętrze pierścienia pól montażowych jest całkowicie wypełnione. O całkowitej powierzchni decyduje powierzchnia zajęta przez układ.



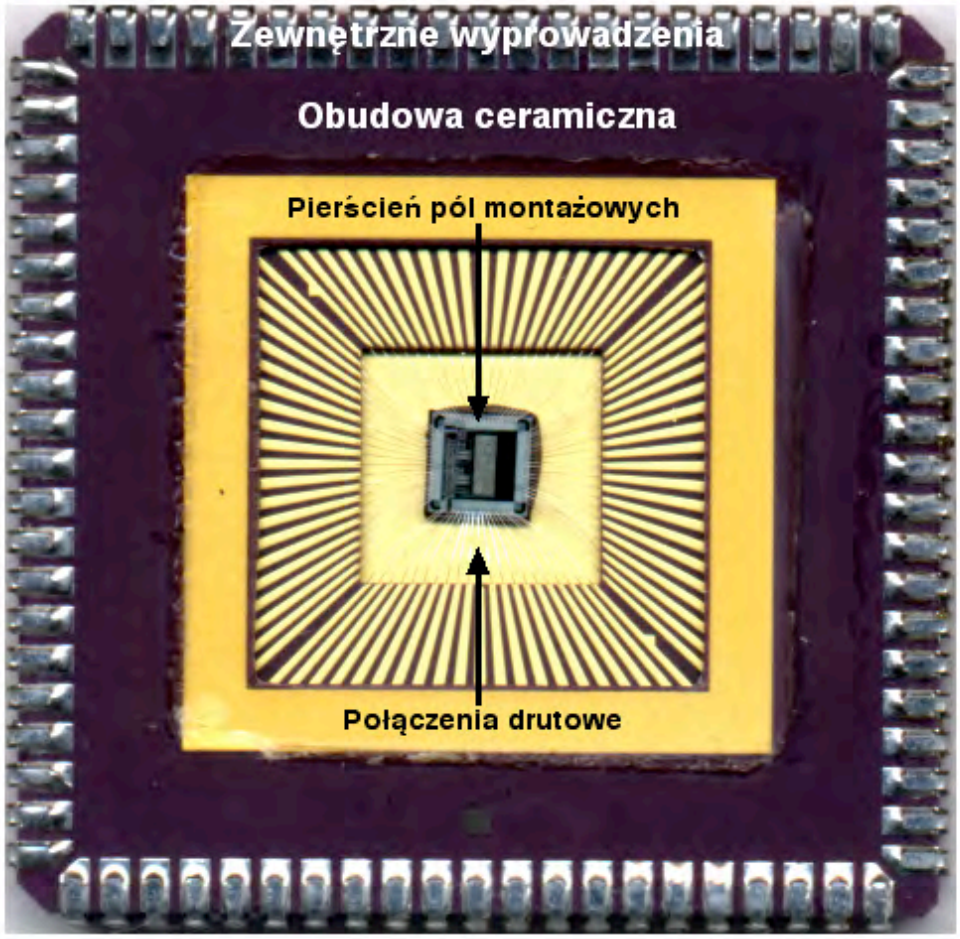

a) b)

Rys. 2.5. Mały układ z dużą liczbą wyprowadzeń (a) i duży układ z małą liczbą wyprowadzeń (b)

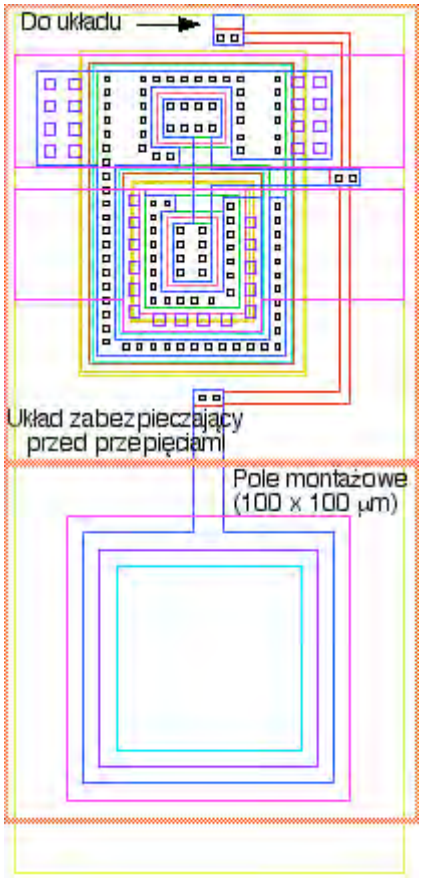
A oto wnioski:

- **Im bardziej zaawansowana technologia, tym więcej kosztuje jednostka powierzchni układu.**
- **Im bardziej zaawansowana technologia, tym mniej kosztuje jeden element układu.**
- **Układy duże i złożone z reguły opłaca się wytwarzać w najbardziej zaawansowanych technologiach.**
- **Układy proste i małe mogą być tańsze, gdy są wytwarzane w mniej zaawansowanych technologiach (czy tak jest, zależy od liczby wyprowadzeń).**

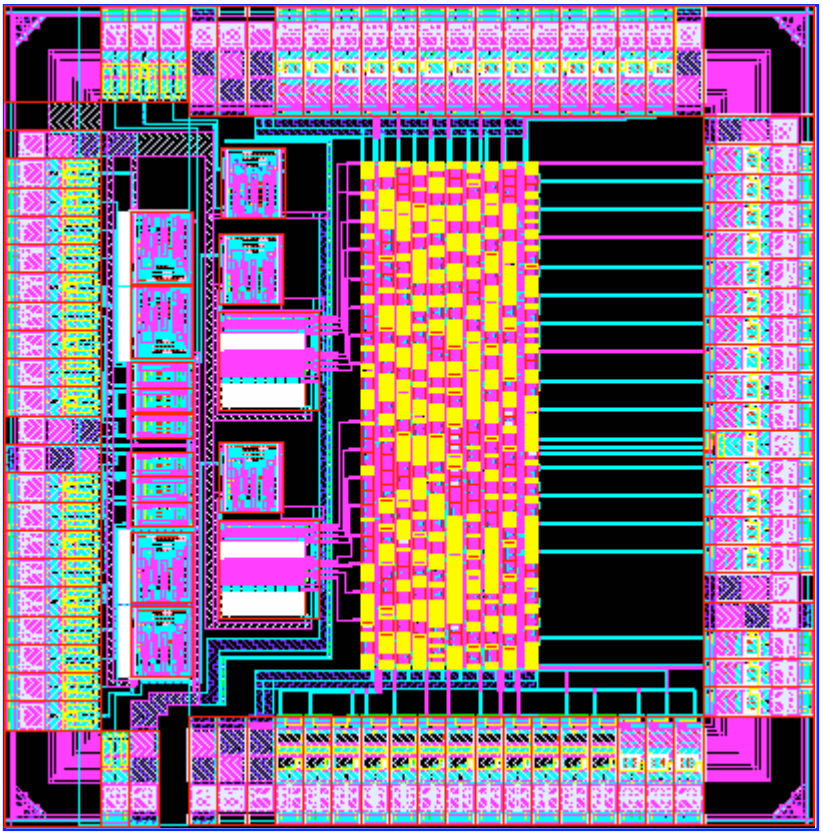
Szacowanie kosztu układu wymaga umiejętności przewidywania, jaką zajmie on powierzchnię, zanim jeszcze został zaprojektowany. Można takie oszacowanie zrobić jedynie w grubym przybliżeniu. Mogą się tu przydać dane statystyczne o przeciętnej powierzchni potrzebnej na jeden tranzystor w danej technologii. Powierzchnia ta zależy jednak silnie od rodzaju projektowanego ukladu (np. w układach analogowych tranzystory są znacznie większe, niż w cyfrowych) oraz od sposobu (stylu) projektowania. Toteż danych takich nie podają producenci układów. Średnią powierzchnię przypadającą na jeden tranzystor można oszacować samemu po próbnym zaprojektowaniu kilku małych fragmentów układu.



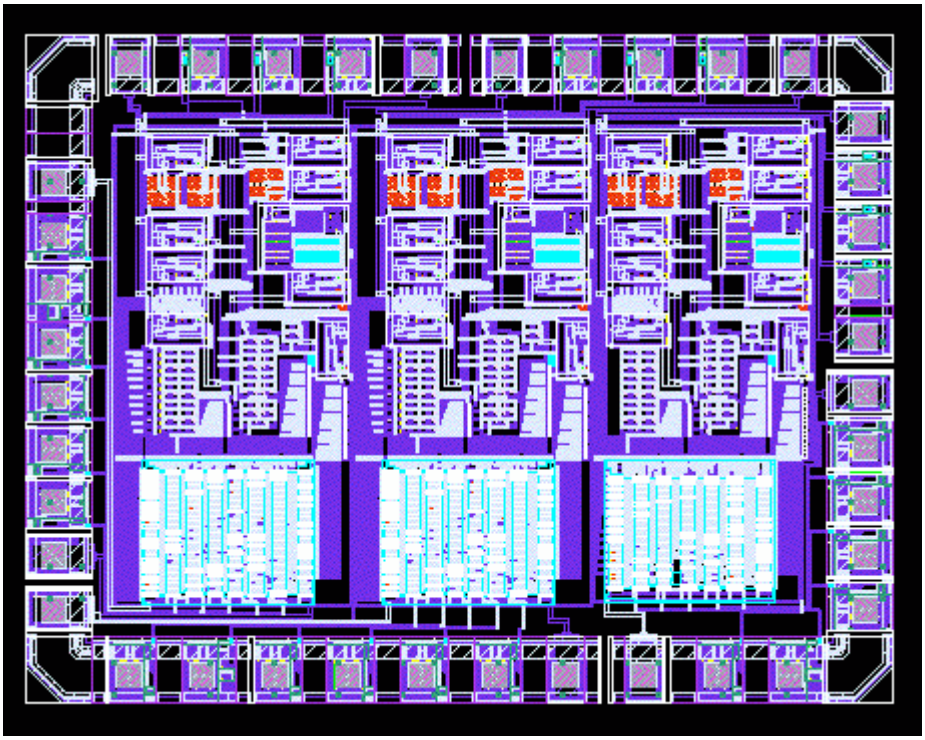

Widok układu scalonego zmontowanego w obudowie.
Na zdjęciu układ analogowo-cyfrowy CMOS o nazwie "ROSETTABIS" zaprojektowany w Instytucie Mikroelektroniki i Optoelektroniki PW przy współpracy z Centrum Badań Kosmicznych PAN, do lądownika eksperymentu ROSETTA Europejskiej Agencji Badań Kosmicznych.




Topografia (obraz masek) pola montażowego wraz z układem zabezpieczającym wnętrze układu scalonego przed przepięciami (technologia CMOS ITE3M, Instytut Technologii Elektronowej w Warszawie). Z reguły pola montażowe są częścią składową gotowych bloków, których kompletne projekty dostarcza producent układów. Bloki te zawierają oprócz samych pól montażowych także układy buforów wejścia lub wyjścia oraz układy zabezpieczające wnętrze układu przed przepięciami, jakie mogą się pojawić na wyprowadzeniach układu przy niewłaściwym obchodzeniu się z nim.



Widok projektu topografii układu, w którym układ nie wypełnia w całości obszaru wewnątrz pierścienia pól montażowych. Całkowitą powierzchnię określa ten pierścień.
(Układ analogowo-cyfrowy CMOS "ROSETTABIS" zaprojektowany w Instytucie Mikroelektroniki i Optoelektroniki PW przy współpracy z Cetrum Badań Kosmicznych PAN. Układ przeznaczony do lądownika misji "ROSETTA" Europejskiej Agencji Badań Kosmicznych.)



Widok projektu topografii układu, w którym układ wypełnia w całości obszar wewnątrz pierścienia pól montażowych i określa powierzchnię całkowitą.
(Układ analogowo-cyfrowy CMOS "CHAOTIC" zaprojektowany w Instytucie Mikroelektroniki i Optoelektroniki PW przy współpracy z Instytutem Systemów Elektronicznych PW. Układ służy do eksperymentów z szyfrowaniem sygnałów telekomunikacyjnych przy wykorzystaniu t.zw. sygnałow chaotycznych.)

## 2.5. Wykonywanie prototypów i produkcja seryjna specjalizowanych układów scalonych

Gdy potrzebny jest specjalizowany układ scalony, postępujemy następująco:

**1. Określamy sposób realizacji układu**

Jeżeli układ jest cyfrowy, można rozważyć użycie układu programowalnego (FPGA). Może to być najlepsze rozwiązanie, jeśli potrzeba niewiele egzemplarzy układu, lub jeżeli zależy nam na bardzo szybkim uzyskaniu działającego układu. W polskich warunkach argumentem na rzecz układu programowalnego może być też łatwa dostępność oprogramowania służącego do projektowania i sprzętu służącego do programowania takich układów. Jednak układy programowalne są rozwiązaniem niekonkurencyjnym cenowo w przypadku sprzętu produkowanego w dużych seriach, a ich parametry techniczne (przede wszystkim szybkość) nie dorównują układom specjalizowanym zaprojektowanym do bezpośredniej realizacji w krzemie, o jakich mówimy w tym wykładzie.

**2. Wybieramy technologię dla projektowanego układu**

Najpierw ze wszystkich dostępnych technologii wybieramy te, które pozwalają osiągnąć cel techniczny. Jeśli takich technologii jest kilka, kierujemy się względami ekonomicznymi, o czym była już mowa. Aby móc wybierać technologie, trzeba oczywiście wiedzieć, do jakich technologii możemy mieć dostęp i znać ich dane techniczne oraz ceny. Te informacje uzyskujemy bądź bezpośrednio od producenta, bądź (najczęściej) od instytucji pośredniczącej. Uzyskanie dostępu do danych technicznych charakteryzujących daną technologię wymaga podpisania umowy o poufności (w języku ang. **Non-Disclosure Agreement**), ponieważ zdecydowana większość producentów traktuje tego rodzaju dane jako poufne. Gdy korzystamy z instytucji pośredniczącej, umowę taką zwykle podpisujemy z tą instytucją. Dla polskich instytucji i firm nie ma tu żadnych barier ani ograniczeń, chociaż oczywiście każdy producent może zastrzec sobie prawo odmowy przekazania informacji poufnych.

**3. Decydujemy, kto będzie projektował układ**

Jeżeli dysponujemy odpowiednimi umiejętnościami i oprogramowaniem komputerowym, możemy to robić sami. Na świecie bardzo wiele firm produkujących sprzęt elektroniczny projektuje samodzielnie układy specjalizowane do tego sprzętu. W polskich warunkach pewną barierą może być koszt potrzebnego do tego oprogramowania. Projekt można zlecić wyspecjalizowanej firmie projektowej (w Polsce istnieje kilka takich firm, a oprócz nich projekty na zamówienie wykonuje także Instytut Technologii Elektronowej w Warszawie). Jeżeli nie planujemy częstego projektowania nowych układów, zlecenie projektu wyspecjalizowanej firmie może być najlepszym rozwiązaniem (ale i w tym przypadku potrzebna jest elementarna znajomość zagadnień projektowania, by mieć "wspólny język" z projektantami z firmy projektowej).

Dalsze kroki dotyczą przypadku, gdy projektujemy układ samodzielnie.

**4. Wybieramy sposób, w jaki zostanie zaprojektowany układ**

Istnieje szereg sposobów uproszczonego i zautomatyzowanego projektowania układów. Będą one omawiane dalej. Każdy z tych sposobów (zwanych także **stylami projektowania**) ma swoje zalety i wady. Wybór może zależeć także od tego, jakim oprogramowaniem komputerowym dysponujemy.

**5. Zamawiamy wykonanie prototypu układu**

W tym celu gotowy projekt wysyłamy do instytucji pośredniczącej lub bezpośrednio do producenta. Projekt wysyła się w postaci elektronicznej przez Internet. Równocześnie wysyła się zamówienie (dokument "papierowy"). W tym momencie określony jest przybliżony koszt prototypu. Jako prototyp otrzymuje się kilkadziesiąt (zwykle od 10 do 50) egzemplarzy układu w obudowach wybranego rodzaju, nie testowanych. Za dodatkową opłatą można zamówić testowanie - trzeba wówczas równoczesnie z projektem układu przesłać program testowania. Można także za dodatkową opłatą otrzymać większą liczbę egzemplarzy układu, zamówić montaż w nietypowych obudowach itp. Instytucja lub firma, do której wysłany został projekt, weryfikuje jego formalną poprawność. Oznacza to zbadanie, czy projekt sporządzono zgodnie z wymaganiami producenta, przede wszystkim z t.zw. **geometrycznymi regułami projektowania** (w języku ang. **Design Rules** - będzie o nich mowa w następnych wykładach). Nie jest weryfikowana funkcjonalna poprawność projektu. Za to, czy schemat układu jest poprawny i czy układ będzie poprawnie wykonywał swą funkcję, odpowiada wyłącznie projektant. W razie stwierdzenia, że wymagania producenta nie są spełnione, odsyłana jest lista stwierdzonych błędów (naruszeń reguł) w projekcie. Istnieje możliwość dostarczenia poprawionej wersji projektu, a specjaliści producenta lub instytucji pośredniczącej zwykle pomagają w poprawieniu błędów. Na wyraźne życzenie projektanta może być przyjęty do produkcji układ naruszający reguły projektowania, jednak wówczas producent ostrzega, że układ najprawdopodobniej nie będzie działał prawidłowo.

**6. Otrzymujemy prototypowy układ i badamy jego działanie**

Prototypowe egzemplarze układu otrzymujemy zwykle po 2 - 3 miesiącach od wysłania projektu. W

praktyce najniższa opłata za wykonanie prototypowych egzemplarzy niezbyt złożonego układu może wynosić około 1000 EURO (za całość zamówienia, czyli kilkadziesiąt egzemplarzy), ale układy duże, wykonane w zaawansowanych, kosztownych technologiach mogą być oczywiście wielokrotnie droższe.

**7. Jeśli układ prototypowy spełnia oczekiwania, a potrzeba więcej egzemplarzy, zamawiamy krótką serię lub produkcję masową**

Można złożyć takie zamówienie bezpośrednio u producenta (zwykle tylko wtedy, gdy dotyczy ono produkcji masowej) lub w instytucji pośredniczącej. Warunki (ceny, terminy) są w każdym przypadku indywidualnie negocjowane. Można uzyskać wcześniej (nawet jeszcze przed zamówieniem prototypu) wstępne oszacowanie kosztu układu w produkcji seryjnej, jeśli znane są podstawowe dane: technologia produkcji, rodzaj układu i jego wielkość, typ obudowy, wielkość serii itp. Pozwala to oszacować ekonomiczną celowość całego przedsięwzięcia. Cena jednego egzemplarza układu w produkcji seryjnej może się wahać w bardzo szerokich granicach, od ułamka EURO do kilkudziesięciu EURO, zależnie od rodzaju układu, a przede wszystkim od długości serii produkcyjnej.

A gdzie można zamawiać układy?

W Europie najważniejszą instytucją pośredniczącą jest konsorcjum EUROPRACTICE powołane i częściowo finansowane przez Unię Europejską. Szczegółowe informacje najprościej uzyskać przez Internet (http://www.europractice.com/). Można tam znaleźć wszystkie potrzebne informacje o zakresie dostępnych usług, oferowanych technologiach, cenach, procedurze uzyskiwania dostępu do danych technicznych, sposobie przesyłania projektów itp. EUROPRACTICE nie tylko pośredniczy między klientami, a producentami układów, ale także dostarcza (dla instytucji akademickich i badawczych) oprogramowanie do projektowania układów po bardzo niskich cenach i świadczy szereg innych usług.

Inną instytucją pośredniczącą jest CMP Service przy Narodowym Instytucie Politechnicznym w Grenoble (http://cmp.imag.fr/). Zakres usług i ich koszt jest podobny jak w EUROPRACTICE. CMP Service nastawia się głównie na obsługę klientów francuskich (dla których, dzięki finansowaniu przez rząd Francji, ma szczególnie korzystne warunki), ale jest to również, jak EUROPRACTICE, instytucja dostępna dla wszystkich.

W Polsce usługi w zakresie wykonywania prototypów i produkcji małoseryjnej układów CMOS świadczy (wykorzystując własną linię produkcyjną) Instytut Technologii Elektronowej w Warszawie. W Instytucie można też zamówić usługę "pod klucz", tj. zlecić wszystko: wykonanie projektu, zamówienie i dostarczenie prototypów, zamówienie produkcji seryjnej i dostawy układów.

Podkreślmy: w opisanym wyżej sposobie postępowania przy projektowaniu i zamawianiu specjalizowanych układów scalonych role pełnione przez projektanta i przez producenta są ściśle rozdzielone. Projektanta nie interesuje, jak realizowane są operacje technologiczne, w wyniku których powstaje struktura scalona (choć oczywiście nie zaszkodzi, by miał pewną ogólną wiedzę na ten temat). Zadaniem projektanta jest dostarczenie producentowi projektu topografii układu, czyli mówiąc w uproszczeniu - rysunku wszystkich masek fotolitograficznych, które określają strukturę każdego elementu, jego położenie oraz sieć połączeń między elementami. Projekt topografii musi spełniać wymagania producenta zwane geometrycznymi regułami projektowania. Producent sprawdza dostarczony projekt wyłącznie pod kątem zgodności z tymi regułami. Nie interesuje go ani funkcja układu, ani jego schemat logiczny i elektryczny. Innymi słowy:

- projektant odpowiada za prawidłowość zdefiniowania funkcji układu, zaprojektowania jego schematu logicznego i elektrycznego oraz struktury fizycznej (czyli topografii), natomiast nie ma wpływu na proces wytwarzania układu (nie może na przykład zażądać wykonania tranzystorów o innych niż typowe parametrach),
- producent odpowiada za prawidłowość wykonania wszystkich operacji technologicznych i wytworzenie struktur półprzewodnikowych, które spełniają normy przewidziane dla danego procesu technologicznego, natomiast nie bierze na siebie żadnej odpowiedzialności za to, czy wytworzone według dostarczonych z zewnątrz projektów układy będą działały zgodnie z intencjami ich projektantów,
- testowanie układu należy do projektanta (lub użytkownika), a nie producenta.

Ten sposób organizacji projektowania i produkcji układów specjalizowanych (zwany w literaturze anglojęzycznej "fabless design") umożliwił powstanie firm wykonujących na zamówienie projekty układów, a nie dysponujących żadną własną bazą produkcyjną. Firmy takie istnieją także w Polsce.

# PRZYKŁADY I ZADANIA DO WYKŁADU 2

**Przykład 1**

W pewnej technologii CMOS koszt jednej partii produkcyjnej liczącej 80 płytek wynosi 800 000 EURO. Płytki mają średnicę 20 cm, a wytwarzany na nich jest układ scalony średniej wielkości: wymiary 2 x 4 $mm^2$. Układ jest montowany w taniej obudowie z tworzywa sztucznego, koszt montażu (wraz z testowaniem końcowym) jednego układu wynosi 1 EURO. Ile wynosi koszt wyprodukowania jednego sprawnego układu, jeżeli uzysk produkcyjny wynosi: (a) 95%, (b) 50%, (c) 5% ? Dla uproszczenia zakładamy uzysk montażu równy 100%.

**Rozwiązanie**

Posłużymy się wzorem (2.4). W tym wzorze$K_A$ = 800 000 EURO, $L_p$ = 80, $A_p$ = 314 $cm^2$, $A_u$ = 0,08 $cm^2$, $k_s$ = 1 EURO, $u_m$ = 1, $u_p$ równy (a) 0,95 lub (b) 0,5 lub (c) 0,05. Dla przypadku (a) otrzymujemy (w zaokrągleniu) koszt $K_{prod}$ = 3,7 EURO, dla przypadku (b) $K_{prod}$ = 6,1 EURO, dla przypadku (c) $K_{prod}$ = 52 EURO. Jak widać, niski uzysk produkcyjny drastycznie zwiększa koszt układu!

---

**Przykład 2**

Jak zmieni się koszt układu produkowanego w technologii omawianej w przykładzie 1, jeśli montaż odbywa się w obudowie ceramicznej. Koszt obudowy, montażu i testów końcowych wynosi 18 EURO. Inne dane jak w przykładzie 1.

**Rozwiązanie**

Dla danych z przykładu 2 otrzymujemy następujące wyniki: (a) 20,7 EURO, (b) 23,1 EURO, (c) 68,9 EURO. Uzysk nadal ma wpływ, ale w dwóch pierwszych przypadkach dominuje koszt obudowy i montażu.

---

**Przykład 3 - zadanie do samodzielnych obliczeń**

Ile wynosi koszt układu produkowanego w technologii omawianej w przykładzie 1, jeśli produkowany układ jest duży: 8 x 10 $mm^2$? Rozważ obie wersje montażu (jak w przykładzie 1 i jak w przykładzie 2) oraz wszystkie 3 przypadki wartości uzysku produkcyjnego. Inne dane jak w przykładzie 1.

---

**Przykład 4**

Należy wybrać najtańszą technologię w celu wykonania prototypu cyfrowego układu CMOS, dla którego spodziewana liczba tranzystorów wynosi 40 000, i oszacować przybliżony koszt wykonania prototypu. Do dyspozycji są trzy technologie:

Tabl. 2.2. Dane trzech technologii CMOS

| Technologia | Cena za 1 $mm^2$ układu w EURO | Średnia powierzchnia przypadająca na 1 tranzystor w układzie cyfrowym ( w mikrometrach kw.) |
|---|---|---|
| CMOS 0.7 | 280 | 80 |
| CMOS 0.5 | 390 | 50 |
| CMOS 0.35 | 590 | 20 |

Układ będzie miał 20 wyprowadzeń, po 5 na każdym boku płytki. Każde wyprowadzenie zajmuje w układzie obszar prostokątny o wymiarach 200 x 500 mikrometrów niezależnie od technologii (w tym prostokącie mieści się pole montażowe oraz bufory wejścia lub wyjścia oraz elementy zabezpieczające układ przed przepięciami). Układ będzie zmontowany w najtańszej możliwej obudowie ceramicznej (typ DIL24, 24 wyprowadzenia), koszt obudowy

i montażu 1 egzemplarza prototypu wynosi 25 EURO. Zmontowane ma być 20 sztuk prototypowych układów.

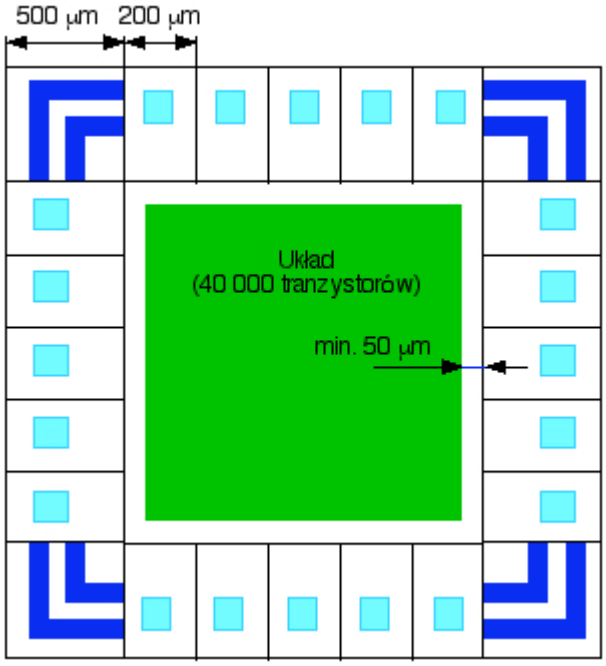
**Rozwiązanie**

Najpierw oszacujemy powierzchnię, jaką zajmie układ zawierający 40 000 tranzystorów, bez wyprowadzeń. Mnożąc liczbę tranzystorów przez powierzchnię przypadającą średnio na jeden tranzystor otrzymujemy

Tabl. 2.3. Szacunkowa powierzchnia i koszt układu (bez wyprowadzeń) dla trzech technologii CMOS

| Technologia | Powierzchnia układu bez wyprowadzeń (w $mm^2$) | Koszt (dla powierzchni układu bez wyprowadzeń) w EURO | Wymiary w mikrometrach (kwadrat) |
|---|---|---|---|
| CMOS 0.7 | 3,2 | 896 | 1790 x 1790 |
| CMOS 0.5 | 2 | 780 | 1410 x 1410 |
| CMOS 0.35 | 0,8 | 472 | 890 x 890 |

Jak widać, zgodnie z wcześniejszymi rozważaniami najtaniej wypada układ wykonany w technologii najbardziej zaawansowanej. Musimy jednak dodatkowo uwzględnić powierzchnię, jaka potrzebna jest na wyprowadzenia. Założymy, że topografia układu będzie wyglądała jak na rys. niżej:




Rys. 2.5. Szkic topografii układu do przykładu 4 (rysunek schematyczny, bez zachowania skali i proporcji wymiarów)

Jak widać, najmniejsza możliwa powierzchnia układu wraz z pierścieniem pól montażowych wynosi 2 x 2 = 4 $mm^2$. Ale układ musi zmieścić się we wnętrzu pierścienia, z pozostawieniem min. 50 mikrometrów z każdej strony dla poprowadzenia połączeń między układem, a wyprowadzeniami. Przy topografii jak na rys. 2.5 wnętrze pierścienia ma wymiary 1000 x 1000 mikrometrów, i układ zrealizowany w technologii CMOS 0.35 mieści się. Układy zrealizowane w pozostałych dwóch technologiach nie mieszczą się, konieczne więc jest luźniejsze rozmieszczenie wyprowadzeń w celu powiększenia wnętrza pierścienia (porównaj rys. 2.4 (a) i (b)). Można łatwo obliczyć, że dla układu zrealizowanego w technologii CMOS 0.5 minimalne wymiary zewnętrzne pierścienia wynoszą 2,51 x 2,51 = 6,3 $mm^2$, a dla technologii CMOS 0.7 te wymiary wynoszą 2,89 x 2,89 = 8,35 $mm^2$ . Ostatecznie otrzymujemy następujące powierzchnie całkowite układów i koszty:

Tabl. 2.4. Całkowity koszt układu dla trzech technologii CMOS

| Technologia | Powierzchnia układu całkowita w $mm^2$ | Koszt (dla powierzchni całkowitej) w EURO | Koszt wraz z ceną montażu 20 szt. układów |
|---|---|---|---|
| CMOS 0.7 | 8,35 | 2338 | 2838 |
| CMOS 0.5 | 6,3 | 2457 | 2957 |
| CMOS 0.35 | 4 | 2360 | 2860 |

Jak widać, ostatecznie najtańszy jest układ wykonany w technologii CMOS 0.7, ale różnice są w naszym przykładzie bardzo małe. Biorąc pod uwagę szacunkowy charakter obliczeń (na przykład bardzo przybliżone obliczenia powierzchni układu na podstawie średniej powierzchni przypadającej na jeden tranzystor) możemy uznać, że w tym przypadku różnice te są bez znaczenia, i przy wyborze technologii kierować się innymi kryteriami (na przykład układ zrealizowany w technologii CMOS 0.35 będzie działał najszybciej).

Przykład 4 ilustruje sposób wyboru najtańszej technologii i daje dość realistyczne wyniki dla niezbyt dużych układów. Nie należy jednak wyciągać zeń wniosku, że zawsze różnice kosztu będą równie małe. Zazwyczaj dla małych i prostych układów wyraźnie najtańsze są technologie najmniej zaawansowane, a dla układów bardzo dużych przeciwnie - technologie najbardziej zaawansowane.

## Bibliografia

[1] W. Marciniak,"*Przyrządy półprzewodnikowe i układy scalone*", WNT Warszawa 1987 (i następne wydania)
[2] B. T. Preas, M. J. Lorenzetti ed., "*Physical design automation of VLSI systems*", The Benjamin/Cummings Publishing Company, Inc., 1988
[3] R. L. Geiger, P. E. Allen, N. R. Strader, "*VLSI design techniques for analog and digital circuits*", McGraw-Hill, Inc. 1994

# Wykład 3: Wytwarzanie układów scalonych

## Wstęp

W wykładzie 3 przypomniane i rozszerzone zostaną podstawowe wiadomości o technologiach wytwarzania półprzewodnikowych monolitycznych układów scalonych. Szczegółowa znajomość technologii nie jest potrzebna projektantowi i użytkownikowi układów scalonych, jednak podstawy trzeba znać, by rozumieć, jak wyglądają i działają elementy półprzewodnikowe w układach scalonych i co z tego wynika dla projektanta tych układów.

Materiał wykładu 3 jest traktowany jako powtórzenie. Materiał ten - przynajmniej w ogólnym zarysie - powinien być znany z wcześniejszych przedmiotów. Wykład 3 ma pomóc powtórzyć ten materiał i uporządkować go, a także w pewnym stopniu rozszerzyć. Aby w pełni rozumieć materiał wykładu, konieczna jest znajomość podstaw fizyki półprzewodników i elementów półprzewodnikowych - pojęć przewodnika, dielektryka, półprzewodnika, monokryształu, domieszki donorowej i akceptorowej, mechanizmu przewodzenia prądu w półprzewodniku oraz budowy i działania podstawowych elementów czynnych - tranzystora unipolarnego z izolowaną bramką (zwanego w skrócie tranzystorem MOS) oraz tranzystora bipolarnego.

W dalszych wykładach będziemy mówić głównie o półprzewodnikowych układach scalonych CMOS. Są one podstawą współczesnej mikroelektroniki. W tych układach podstawowymi elementami czynnymi są dwa rodzaje tranzystorów MOS - tranzystory z kanałem typu *n* i tranzystory z kanałem typu *p*. Wspominać także będziemy o układach bipolarnych, które służą dziś tylko do wytwarzania niektórych typów układów analogowych. Podstawowymi elementami tych układów są tranzystory bipolarne. Istnieją również technologie BiCMOS - można w nich wytwarzać układy, w których występują zarówno tranzystory MOS, jak i bipolarne o bardzo dobrych parametrach. Procesy te są skomplikowane i kosztowne, ich zastosowania ograniczone do układów, które muszą spełniać pewne szczególne wymagania (układy dużej mocy, układy mikrofalowe). Dlatego w wykładzie 3 bardziej szczegółowo omówione jest wytwarzanie układów CMOS, nieco bardziej pobieżnie - układów bipolarnych, a inne rodzaje układów są tylko wspomniane.

W ostatniej części wykładu 3 po omówieniu sposobów montażu układów scalonych w obudowach zawarte są także ważne uwagi praktyczne dotyczące obchodzenia się z układami scalonymi.

## 3.1. Materiały w mikroelektronice

Aby produkować półprzewodnikowe układy scalone, zarówno układy CMOS, jak i bipolarne, musimy dysponować odpowiednim podłożem półprzewodnikowym i móc wykonywać w nim i na nim:

- warstwy domieszkowane o różnych typach przewodnictwa (*p* lub *n*) i różnych koncentracjach domieszek,
- warstwy dielektryczne,
- warstwy przewodzące,

oraz móc nadawać warstwom domieszkowanym, dielektrycznym i przewodzącym wymagane kształty i wymiary. Służą do tego procesy fotolitografii.

**Podłoże i warstwy domieszkowane**

Podstawowym materiałem, z którego wytwarzane są półprzewodnikowe układy scalone, jest **krzem** (**Si**). Jest to jeden z najbardziej rozpowszechnionych w przyrodzie pierwiastków. Do wytwarzania wyrobów półprzewodnikowych potrzebny jest krzem w postaci monokrystalicznych płytek, płytki te mają grubość około 1 mm i średnicę sięgającą 30 cm. Płytki te powstają w wyniku cięcia monokryształów krzemu mających postać walców o średnicy równej średnicy płytek i długości sięgającej kilkudziesięciu centymetrów.

Warto wiedzieć, że powszechnie dziś stosowana metoda wytwarzania monokryształów krzemu została opracowana na wiele lat przed powstaniem przemysłu półprzewodnikowego przez Jana Czochralskiego - pracującego do końca pierwszej wojny światowej na uczelniach niemieckich, a w latach międzywojennych profesora Politechniki Warszawskiej. Po drugiej wojnie światowej prof. Czochralski został z przyczyn politycznych odsunięty od pracy naukowej.

W najbardziej zaawansowanych technologiach układów scalonych, zwłaszcza układów BiCMOS, występuje **krzemogerman** - materiał zawierający w sieci krystalicznej krzemu pewną domieszkę atomów **germanu** (**Ge**). Krzemogerman cechuje się węższą przerwą zabronioną niż czysty krzem i większą ruchliwością nośników. Pozwala to uzyskać lepsze parametry niektórych elementów. Krzemogerman nie występuje jako materiał na płytki podłożowe, powstaje przez wprowadzenie do podłoża krzemowego atomów germanu.

W latach osiemdziesiątych i na początku lat dziewięćdziesiątych XX w. za materiał przyszłości uważano **arsenek galu** (**GaAs**).Materiał ten cechuje bardzo duża ruchliwość elektronów, ma on też szereg innych interesujących właściwości. Służy do wytwarzania układów pracujących przy częstotliwościach mikrofalowych. Jednak technologie wykorzystujące arsenek galu są trudne i kosztowne. Dziś w zakresach częstotliwości, w których do niedawna wykorzystywano wyłącznie arsenek galu, z powodzeniem działają układy krzemowe, toteż arsenek galu ze względów ekonomicznych wychodzi z użycia.

Pod pojęciem warstw domieszkowanych rozumiane są warstwy w płytce podłożowej, w których występują celowo wprowadzone domieszki donorowe lub akceptorowe. Domieszkami donorowymi są pierwiastki z piątej grupy układu okresowego, przede wszystkim **fosfor** (**P**) i **arsen** (**As**), zaś jako domieszka akceptorowa wykorzystywany jest zwykle **bor** (**B**), chociaż akceptorami są także inne pierwiastki z trzeciej grupy układu okresowego (np. **aluminium** (**glin**, **Al**)). Sposoby uzyskiwania warstw domieszkowanych będą omówione dalej.

**Warstwy dielektryczne**

Najważniejszym dielektrykiem wykorzystywanym w mikroelektronice jest **dwutlenek krzemu** (**$SiO_2$**). Materiał ten jest jednym z najlepszych znanych dielektryków z punktu widzenia parametrów elektrycznych, a przy tym materiałem wygodnym technologicznie. Powstaje on w wyniku utleniania krzemu (płytki podłożowej) lub jest nakładany z zewnątrz, powstając w wyniku reakcji chemicznej zachodzącej w środowisku gazowym. Na powierzchni krzemu tworzy nie tylko warstwę dielektryczną o dobrych właściwościach, ale także skutecznie chroni krzem przed szkodliwymi oddziaływaniami otaczającego środowiska (utlenianie, korozja, zanieczyszczenia). Łatwo poddaje się procesom fotolitografii. Jest bardzo dobrym podłożem pod warstwy aluminium, z których wykonywane są ścieżki połączeń w układach scalonych - aluminium wykazuje bardzo dobrą przyczepność do dwutlenku krzemu.

Mimo bardzo korzystnych właściwości dwutlenku krzemu coraz częściej stosowane są także inne dielektryki. Jednym z nich jest **azotek krzemu** (**$Si_3N_4$**). Do niedawna występował on wyłącznie jako materiał pomocniczy w niektórych procesach technologicznych. Obecnie atomami azotu wzbogaca się warstwy $SiO_2$, taka warstwa ma nieco wyższą przenikalność dielektryczną od czystego dwutlenku krzemu. Jeszcze wyższe wartości przenikalności dielektrycznej uzyskuje dwutlenek krzemu wzbogacony atomami **hafnu** (**Hf**). Owe wyższe wartości przenikalności dielektrycznej są korzystne w warstwach dielektrycznych izolujących bramki tranzystorów MOS (będzie jeszcze o tym mowa w końcowych wykładach). W mikroelektronice wykorzystywane są także dielektryczne warstwy organiczne (poliimidowe).

**Warstwy przewodzące**

Do niedawna podstawowym materiałem przewodzącym, z którego wykonywane były ścieżki połączeń w układach scalonych, było **aluminium**. Jest to jeden z najlepszych znanych przewodników, wytwarzanie warstw aluminium jest bardzo łatwe, podobnie jak trawienie w procesach fotolitografii. Aluminium jest materiałem trwałym, nie ulegającym korozji. Ma jednak też istotne wady. Przede wszystkim atomy aluminium są w krzemie domieszką akceptorową, toteż krzem z domieszką atomów aluminium może stać się półprzewodnikiem typu *p*. Aluminiowy kontakt do krzemu typu *n* może utworzyć złącze *pn*, czyli diodę, zamiast dobrego kontaktu elektrycznego o liniowej charakterystyce prądowo-napięciowej i małej rezystancji. Aby tego uniknąć, obszary typu *n*, do których wykonuje się kontakty aluminiowe, muszą wykazywać na powierzchni wysoką koncentrację domieszki, co może wymagać dodatkowego procesu domieszkowania.

Obecnie w większości zaawansowanych procesów technologicznych aluminium nie kontaktuje się bezpośrednio z krzemem. Pomiędzy krzemem, a ścieżką aluminiową występuje inny metal, np. **wolfram** (**W**). Obok aluminium ścieżki przewodzące coraz częściej wykonuje się z **miedzi** (**Cu**). Miedź jest lepszym przewodnikiem od aluminium. Technologia połączeń miedzianych jest jednak bardzo skomplikowana, bowiem miedzi nie można nakładać bezpośrednio na warstwę dwutlenku krzemu w taki sposób, jak aluminium. Poza tym miedź nie może przedostać się do podłoża krzemowego, ponieważ atomy miedzi w krzemie stanowią bardzo skuteczne centra rekombinacji, skracając o rzędy wielkości czas życia nośników.

Niezastąpionym we współczesnej mikroelektronice materiałem przewodzącym jest **krzem polikrystaliczny** zwany potocznie **polikrzemem**. Struktura polikrystaliczna powstaje przy osadzaniu warstwy krzemu na podłożu, które nie jest monokryształem, na przykład na warstwie dwutlenku krzemu. Krzem polikrystaliczny silnie domieszkowany domieszką donorową (a więc typu *n*) jest stosunkowo dobrym przewodnikiem, ale jego głównym zastosowaniem nie są ścieżki przewodzące, lecz obszary bramek tranzystorów MOS.

Przy montażu układów scalonych w obudowach powszechnie wykorzystuje się **złoto** (**Au**). Bywają nim pokrywane metalowe podstawki, do których techniką lutowania mocowane są płytki układów scalonych, a ponadto złoto jest najpowszechniej stosowanym materiałem, z którego wykonuje się połączenia drutowe między układem scalonym, a zewnętrznymi wyprowadzeniami.

## 3.2. Podstawowe operacje technologiczne

Aby móc wytworzyć dowolną strukturę układu scalonego, trzeba mieć możliwość wytworzenia obszarów domieszkowanych w półprzewodniku (z nich budowane są struktury tranzystorów i innych elementów), obszarów dielektrycznych (służą one do izolowania elementów oraz ścież ek przewodzących, które nie powinny być połączone elektrycznie) oraz obszarów przewodzących (które tworzą sięć połączeń elektrycznych między elementami układu). Zatem podstawowe operacje technologiczne w mikroelektronice to **operacje wytwarzania obszarów domieszkowanych, dielektrycznych i przewodzących**. Obszary te muszą mieć właściwe kształty, wymiary i położenie w układzie. Uzyskuje się to przy zastosowaniu **fotolitografii** oraz **operacji selektywnego trawienia**.

**Wytwarzanie obszarów domieszkowanych**

W przemyśle półprzewodnikowym stosowane są dziś trzy sposoby wytwarzania warstw półprzewodnika domieszkowanych domieszkami donorowymi lub akceptorowymi: epitaksja, dyfuzja oraz implantacja jonów.

**Epitaksja** polega na nakładaniu na podłoże warstwy półprzewodnika tego samego rodzaju, ale różniącego się domieszkowaniem (inny niż w podłożu typ przewodnictwa i/lub koncentracja domieszki). Warstwa epitaksjalna nałożona na monokrystaliczne podłoże stanowi przedłużenie jego monokrystalicznej struktury. Jeśli powierzchnia płytki podłożowej nie jest płaska i zawiera wgłębienia lub wypukłości, to są one powtarzane na powierzchni warstwy epitaksjalnej. Typowe grubości warstw epitaksjalnych: od ułamka mikrometra do kilkunastu mikrometrów. Krzemowe warstwy epitaksjalne osadzane są na monokrystalicznym krzemie w procesie, w którym rozkład związku chemicznego krzemu w wysokiej temperaturze w fazie gazowej uwalnia atomy krzemu (ang. **Vapor-Phase Epitaxy**, **VPE**).

**Dyfuzja** polega na wprowadzaniu domieszki do wnętrza półprzewodnika z zewnętrznego źródł a domieszki znajdującego się w kontakcie z płytką półprzewodnikową. Wymaga wysokiej temperatury (powyżej $800^{o}$ C). Dyfuzję wykonuje się zazwyczaj w dwóch procesach zwanych *predyfuzją* i *redyfuzją*. Podczas predyfuzji płytka podłożowa jest umieszczona w piecu w atmosferze zawierającej atomy domieszki i zarazem utleniającej. Na powierzchni tworzy się t.zw. szkliwo domieszkowane - mieszanka dwutlenku krzemu i tlenku domieszki (np. $P_2O_5$). Szkliwo to stanowi źródło, z którego następuje dyfuzja domieszki do wnętrza płytki. Proces predyfuzji powoduje wytworzenie silnie domieszkowanej, ale na ogół płytkiej warstwy przy powierzchni płytki. Po predyfuzji szkliwo domieszkowane jest usuwane chemicznie, a płytka ponownie umieszczona w piecu w atmosferze utleniającej, ale już nie zawierającej domieszki. Na powierzchni tworzy się warstwa tlenkowa ($SiO_2$), która w znacznym stopniu zabezpiecza przed ucieczką domieszki na zewnątrz. Z płytkiej, ale silnie domieszkowanej warstwy wytworzonej podczas predyfuzji domieszka dyfunduje w głąb płytki. Zasięg domieszki wzrasta, a jej koncentracja na powierzchni maleje. Dobierając odpowiednio temperaturę i czas predyfuzji i redyfuzji można otrzymać w sposób powtarzalny wymagany rozkład domieszki, a w tym rozkłady charakteryzujące się niską koncentracją na powierzchni, co trudno byłoby uzyskać w innym sposób. Dyfuzja nie nadaje się jednak do wytwarzania warstw, które są bardzo płytkie (znacznie poniżej 1 mikrometra) i równocześnie słabo domieszkowane. Takie warstwy można wykonywać wykorzystując implantację jonów - proces, który we współczesnej mikroelektronice niemal całowicie zastąpił dyfuzję.

Więcej o dyfuzji - patrz dodatek 1

Warto dodać, że dyfuzja atomów domieszki zachodzi zawsze, gdy płytka półprzewodnikowa zawierająca domieszkę o nierównomiernym rozkładzie koncentracji poddawana jest procesom wysokotemperaturowym, na przykład wygrzewaniu po implantacji jonów. Proces dyfuzji prowadzi zawsze do wyrównywania różnic koncentracji, toteż utrudnia uzyskanie w elementach gotowego układu scalonego rozkładów koncentracji domieszek o dużym gradiencie. Z tego powodu we współczesnej mikroelektronice dąży się do tego, aby procesów wysokotemperaturowych było jak najmniej, a ich temperatura możliwie niska.

**Implantacja jonów** polega na "wstrzeliwaniu" w płytkę podłożową jonów domieszki, którym nadana została duża energia kinetyczna przez rozpędzenie w silnym polu elektrycznym. Implantacja jonów odbywa się w pró żni w temperaturze otoczenia. Jony wytracają energię w zderzeniach z atomami półprzewodnika i lokują się w jego sieci krystalicznej. Bombardowanie płytki podłożowej jonami o wysokiej energii powoduje powstawanie defektów sieci monokryształu. Po implantacji stosuje się wygrzewanie płytki w temperaturze kilkuset stopni C. Powoduje to odbudowę regularności sieci krystalicznej oraz zajmowanie przez atomy domieszki pozycji w węzłach sieci krystalicznej (atomy domieszek donorowych i akceptorowych zachowują się jak donory lub akceptory tylko wtedy, gdy zastępują atomy półprzewodnika w sieci, a nie wtedy, gdy są ulokowane między tymi węzłami). Procesem implantacji można precyzyjnie sterować: zasięg jonów zależy od ich energii (czyli od napięcia przyspieszającego), a dawkę (czyli liczbę implantowanych jonów) można ustalić całkując w czasie prąd jonów płynący między ich źródłem, a płytką podłożową. Dzięki łatwej i precyzyjnej regulacji zasięgu i dawki jonów implantacja jonów pozwala wykonywać warstwy płytkie i słabo domieszkowane. Jest we współczesnej mikroelektronice najczęściej stosowanym sposobem wprowadzania domieszek do wnętrza płytek półprzewodnikowych.

Więcej o implantacji jonów - patrz dodatek 2

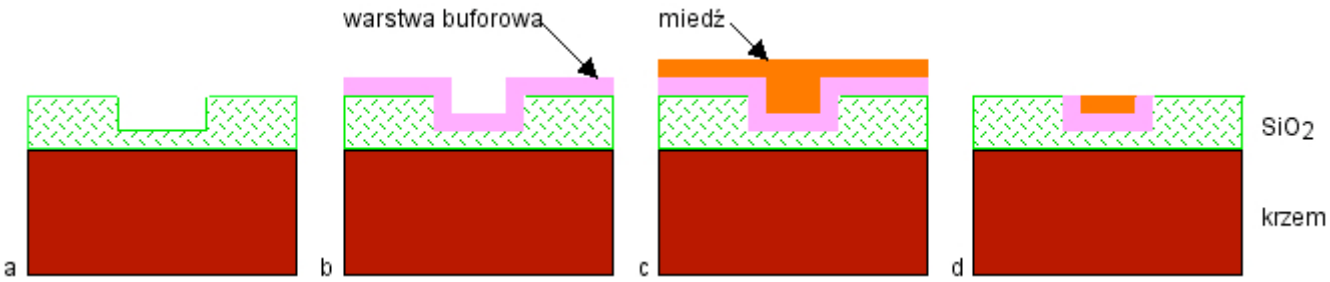
**Wytwarzanie warstw dielektrycznych**

Dominujące w mikroelektronice warstwy $SiO_2$ można wytwarzać przez **utlenianie krzemu**, co wykonuje się umieszczając płytkę krzemową w atmosferze utleniającej w wysokiej temperaturze. Jest to jednak możliwe tylko w miejscach, w których powierzchnia krzemu jest odsłonięta.W każdym innym przypadku warstwę $SiO_2$ uzyskuje się przez **osadzanie $SiO_2$** z par powstających w wyniku reakcji chemicznej w fazie gazowej (ang. **Chemical Vapor Deposition, CVD**). Utlenianie krzemu daje warstwy o najlepszych właściwościach elektrycznych, dlatego w ten sposób wytwarza się warstwy dielektryku bramkowego w tranzystorach MOS. Inne rodzaje warstw dielektrycznych uzyskuje się przez osadzanie różnymi metodami, np. $Si_3N_4$ podobnie jak SiO2 metodą CVD.

Warto dodać, że we współczesnej mikroelektronice warstwa dwutlenku krzemu pod bramką tranzystora MOS jest niezwykle cienka - jej grubość wynosi kilka nanometrów, co oznacza zaledwie około 10 warstw atomowych. Aby zapewnić mały rozrzut produkcyjny parametrów tranzystorów, trzeba zapewnić jednakową (kilkuatomową!) grubość tej warstwy na całej płytce podłożowej o średnicy 30 cm. Pokazuje to, jak niezwykła jest precyzja procesów technologicznych współczesnej mikroelektroniki.

**Wytwarzanie warstw przewodzących**

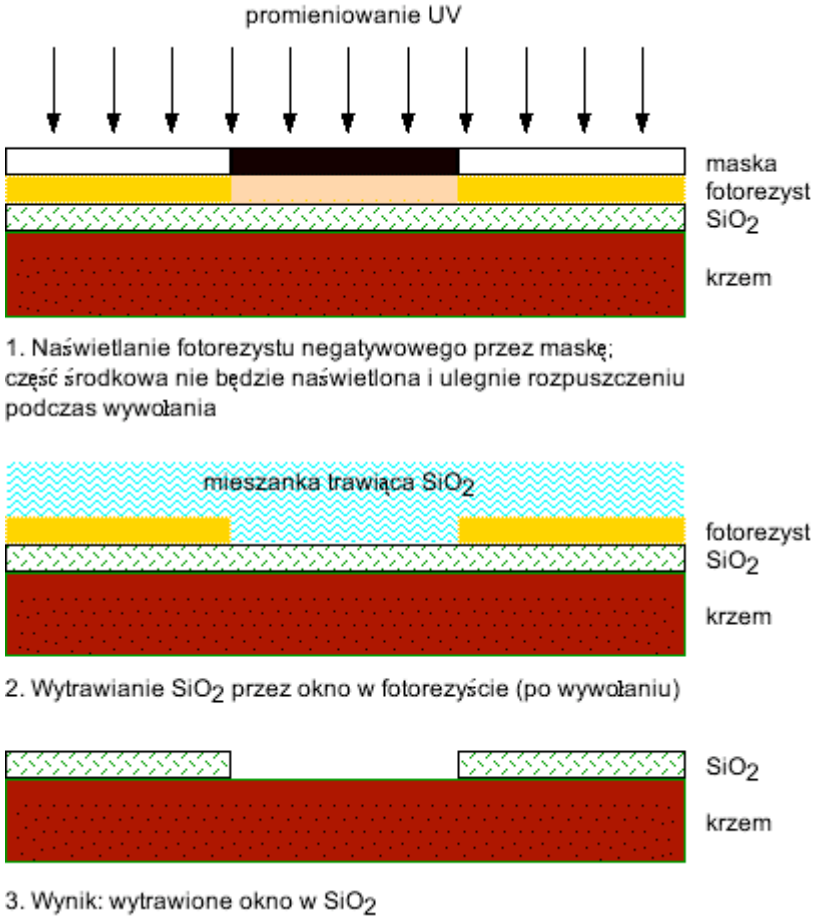
Najpowszechniej do niedawna stosowane warstwy aluminium są otrzymywane w prosty sposób przez **naparowanie w próżni** ze źródła par aluminium. Równie ważne warstwy polikrzemu są osadzane w wyniku wydzielania się atomów krzemu przez **rozkład silanu** ($SiH_4$) w wysokiej temperaturze. Wytwarzanie warstw przewodzących z miedzi jest procesem daleko bardziej skomplikowanym. Proces ten, zwany damasceńskim, składa się z kilku etapów (rys. 3.1). Najpierw przy użyciu fotolitografii i trawienia wykonuje się rowki w warstwie $SiO_2$ - w nich będą ścieżki przewodzące (Rys. 3.1a). Następnie na płytce osadzana jest warstwa buforowa, ma ona dobrą przyczepność do $SiO_2$, nie przepuszcza atomów miedzi w kierunku płytki podłożowej i jest przewodząca (rys. 3.1b). Płytka pokrywana jest elektrolitycznie warstwą miedzi (rys. 3.1c). W ostatnim kroku wykonywane jest mechaniczno-chemiczne polerowanie płytki (ang. **Chemical-Mechanical Polishing**, **CMP**), po którym miedź pozostaje tylko w głębi rowków.




Rys. 3.1. Proces wytwarzania miedzianej ścieżki przewodzącej

**Fotolitografia**

Służy do nadawania obszarom domieszkowanym, dielektrycznym i przewodzącym wymaganych położ eń, kształtów i wymiarów. Wykorzystywana jest tu wrażliwość niektórych związków chemicznych na promieniowanie elektromagnetyczne. Istnieją substancje, które pod wpływem tego promieniowania (w mikroelektronice - ultrafioletu) ulegają utwardzeniu - tracą rozpuszczalność w określonych rozpuszczalnikach. Takie substancje mogą być użyte jako *fotorezysty negatywowe*. Inne substancje pod wpływem promieniowania stają się łatwo rozpuszczalne. Można ich użyć jako fotorezystów *pozytywowych*. Rysunek 3.2 pokazuje zasadę fotolitografii na przykładzie wykonywania okna w warstwie $SiO_2$. Płytka jest pokrywana warstwą fotorezystu negatywowego, następnie naświetlana przez maskę mającą obszary przezroczyste i nieprzezroczyste, naświetlona płytka jest następnie zanurzona w rozpuszczalniku rozpuszczającym nienaświetlony fotorezyst, po czym mieszanka trawiąca $SiO_2$, a nie naruszająca fotorezystu wytrawia okno w warstwie $SiO_2$. W podobny sposób wytrawia się w warstwie aluminium ścieżki przewodzące. W przypadku procesu implantacji jonów okna w warstwie fotorezystu określają obszary, do których wprowadzone będą jony domieszki - warstwa fotorezystu o dostatecznej grubości zatrzymuje jony nie dopuszczając ich do powierzchni półprzewodnika.




Rys. 3.2. Proces wytwarzania okna w warstwie $SiO_2$ przy pomocy fotolitografii

Rysunek 3.2 pokazuje najprostszy proces fotolitograficzny, w którym płytka jest naświetlana bezpośrednio przez maskę (fotolitografia kontaktowa). We współczesnej mikroelektronice stosowana jest niemal wyłącznie fotolitografia projekcyjna - obraz maski jest wyświetlany na płytkę przez obiektyw, na podobnej zasadzie, jak w zwykłym rzutniku do przezroczy lub powiększalniku fotograficznym. Urządzenie do naświetlania zwane jest potocznie steperem, ponieważ naświetla nie całą płytkę równocześnie, lecz kolejne układy na płytce ("step and repeat" - "zrób krok i powtórz").

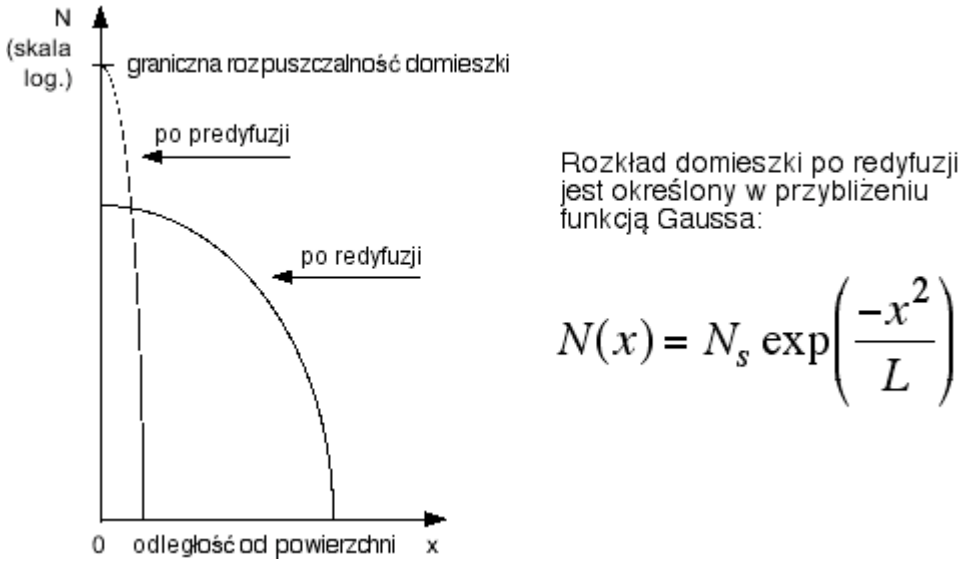
Warto wiedzieć, że urządzenia do fotolitografii, które umożliwiają fotolitografię o zdolności rozdzielczej na poziomie nanometrów, osiągają szczyty istniejących możliwości technologicznych w zakresie optyki i mechaniki precyzyjnej, i co za tym idzie - są niezwykle kosztowne. Ich koszt stanowi znaczną część kosztu wyposażenia technologicznego współczesnych linii produkcyjnych.

## 3.2. Dodatek 1: Więcej o dyfuzji

Dwa główne mechanizmy dyfuzji w ciele stałym to dyfuzja międzywęzłowa i dyfuzja podstawieniowa. Ta ostatnia jest mechanizmem dyfuzji donorów i akceptorów w krzemie. Szybkość, z jaką zachodzi dyfuzja, zależy bardzo silnie od temperatury. W przypadku dyfuzji podstawieniowej jej szybkość jest także funkcją gęstości defektów krystalograficznych, które ułatwiają przemieszczanie się atomów. W monokrysztale bez defektów dyfuzja podstawieniowa zachodzi bardzo powoli.

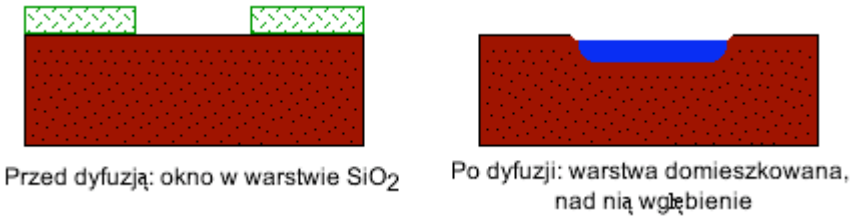

Rys. 3.3. Mechanizmy dyfuzji

Domieszkowanie przy użyciu dyfuzji odbywa się w dwóch etapach: redyfuzji i predyfuzji. Podczas predyfuzji powstaje warstwa bardzo silnie domieszkowana i na ogół płytka. Koncetracja domieszki na powierzchni określona jest przez graniczną rozpuszczalność domieszki w półprzewodniku. Wielkość ta jest charakterystyczna dla danej pary domieszka-półprzewodnik i bardzo słabo zależy od warunków prowadzenia dyfuzji, toteż wynik procesu predyfuzji jest łatwy do kontroli i powtarzalny. Czas i temperatura predyfuzji decydują o dawce (czyli liczbie atomów ) wprowadzonej domieszki. W czasie redyfuzji wprowadzona wcześniej domieszka rozpływa się - wędruje w głąb płytki półprzewodnikowej. Ponieważ dawka nie ulega zmianie, koncentracja przy powierzchni maleje, a zasięg domieszki wzrasta.




Rys. 3.4. Rozkłady domieszek otrzymane w wyniku predyfuzji i redyfuzji

Aby wprowadzić domieszkę do ściśle określonego obszaru, stosuje się maskowanie warstwą $SiO_2$. Dyfuzja wykonywana jest przez okna w tej warstwie, których kształty i wymiary są określone przy użyciu fotolitografii. Predyfuzja i redyfuzja odbywają się w wysokiej temperaturze w atmosferze utleniającej. W związku z tym tam, gdzie były dyfundowane domieszki, pojawia się wgłębienie, bo część krzemu uległa utlenieniu.



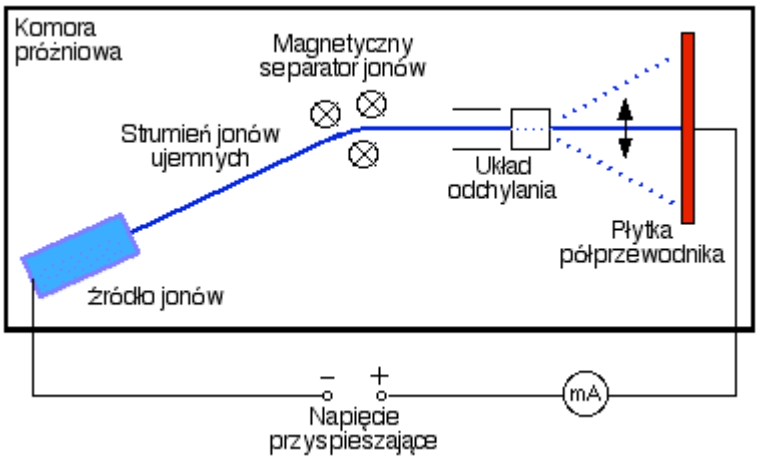

Rys. 3.5. Dyfuzja selektywna - do obszaru określonego przez okno w warstwie $SiO_2$

Regulując czasy i temperatury predyfuzji i redyfuzji można dość dokładnie kontrolować końcowy rozkład domieszek. Jednak w całym cyklu produkcyjnym występuje wiele operacji wysokotemperaturowych, a w każdej z

nich zachodzi proces redyfuzji wcześniej wprowadzonych domieszek. Dlatego zgranie warunków wszystkich operacji w całym procesie produkcji układów jest trudne, wymaga symulacji komputerowych i eksperymentów. Raz ustawiony proces nie jest już później zmieniany. Konstruktor nie może wymagać zmiany warunków procesu, musi dostosować projekt do istniejącego procesu produkcyjnego.

## 3.2. Dodatek 2: Więcej o implantacji jonów

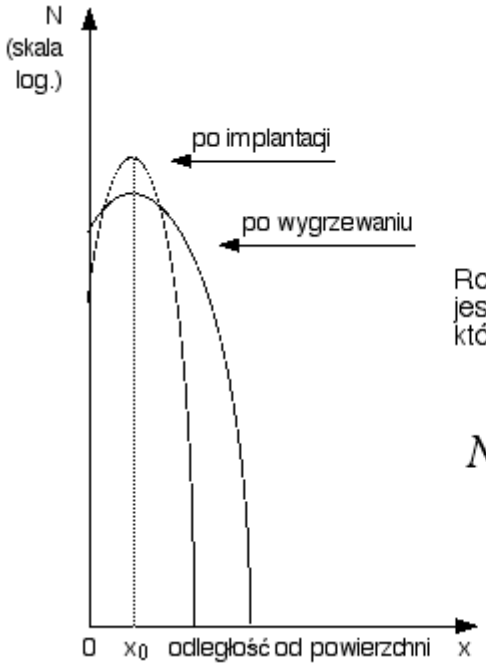
Implantacja jonów odbywa się w implantatorze, którego budowa przypomina mały akcelerator cząstek elementarnych. Przyspieszane są jednak nie cząstki, lecz jony pierwiastka domieszkującego półprzewodnik. Uproszczony schemat implantatora jonów:




Rys. 3.6. Zasada działania implantatora jonów

Jony ze źródła jonów są w poprzecznym polu magnetycznym "oczyszczane" (jony różnych pierwiastków mają różną masę, toteż odchylane są pod różnymi kątami) i przyspieszane w silnym polu elektrycznym. Układ elektrostatycznego odchylania odchyla strumień jonów i "przemiata" nim po całej płytce półprzewodnikowej. Jony przyspieszone silnym polem elektrycznym uderzają w płytkę z dużą energią kinetyczną i dzięki temu przemieszczają się w głąb, wytracając stopniowo energię w zderzeniach z atomami sieci krystalicznej półprzewodnika. Równocześnie ich ładunek ulega zobojętnieniu, np. jony ujemne oddają elektron, który przez zewnętrzny obwód odpływa do źródła napięcia przyspieszającego.

W zderzeniach jonów z atomami półprzewodnika część tych atomów zostaje przemieszczona, regularna budowa monokryształu zostaje w mniejszym lub większym stopniu zaburzona - powstają defekty sieci krystalicznej. Implantowane atomy domieszki zajmują przypadkowe położenia, wiele z nich ląduje w położeniach międzywęzłowych, gdzie nie wykazują właściwości domieszki donorowej czy też akceptorowej. Po implantacji poddaje się płytkę wygrzewaniu, co powoduje odbudowę regularnej sieci monokryształu, zaś atomy domieszki lokują się w węzłach sieci krystalicznej. Równocześnie rozkład domieszki ulega redyfuzji.




Rozkład domieszki po implantacji i wygrzewaniu jest określony w przybliżeniu funkcją Gaussa, której maksimum znajduje się na głębokości $x_0$:

$$N(x) = N_m \exp\left[\frac{-(x-x_0)^2}{L}\right]$$

Rys. 3.7. Rozkłady domieszek po procesie implantacji jonów i wygrzewania

Maksimum rozkładu znajduje się na głębokości zależnej od energii implantowanych jonów, czyli od napięcia

przyspieszającego. Dawka, czyli liczba atomów domieszki przypadająca na 1 $cm^2$ powierzchni płytki, jest regulowana przez pomiar i całkowanie w czasie prądu płynącego w obwodzie implantatora.

Implantację można wykonać poprzez bardzo cienką warstwę $SiO_2$. Wówczas maksimum rozkładu domieszek może znaleźć się bliżej powierzchni, a nawet wewnątrz warstwy tlenku. Implantacja przez tlenek zmniejsza energię jonów bombardujących monokryształ i tym samym zmniejsza liczbę powstających defektów.

Implantację, podobnie jak dyfuzję, wykonuje się zazwyczaj przez okno wykonane techniką fotolitografii. Materiałem maskującym może być fotorezyst, ponieważ implantację wykonuje się (w przeciwieństwie do dyfuzji) w temperaturze otoczenia, w której warstwa fotorezystu nie ulega zniszczeniu.

## 3.3. Procesy wytwarzania układów CMOS

W układach scalonych CMOS elementami czynnymi są tranzystory unipolarne z izolowaną bramką: n - kanałowe (dalej nazywane także w skrócie tranzystorami nMOS) i p - kanałowe (dalej nazywane także w skrócie tranzystorami pMOS). Tranzystory obu typów przewodnictwa są wzbogacane, tj. kanał nie istnieje dla zerowej polaryzacji bramki, i aby go utworzyć trzeba spolaryzować bramkę względem źródła napięciem większym (co do wartości bezwzględnej) od napięcia progowego. Napięcie progowe tranzystorów nMOS jest dodatnie, tranzystorów pMOS - ujemne. Typowa wartość wynosi około 0.7 - 0.9 V dla układów o napięciu zasilania 5 V, mniej dla układów przeznaczonych do pracy przy niższych napięciach zasilania. W najnowocześniejszych układach z wymiarami elementów rzędu 100 nm i mniej typowe napięcie zasilania wynosi 1 V, a napięcia progowe mają wartości około 0.3 V.

Dla układów o dwóch różnych typach kanału potrzebne są obszary podłoża o dwóch typach przewodnictwa. Obecnie produkuje się niemal wyłącznie układy, w których podłożem jest płytka półprzewodnikowa typu *p* (jest to podłoże dla tranzystorów nMOS), a w niej wytwarza się wyspy o przewodnictwie typu *n* (są one podłożem dla tranzystorów pMOS). Wyspy typu *n* dla zapewnienia wzajemnej izolacji **muszą być spolaryzowane zaporowo** (czyli napięciem dodatnim) względem podłoża typu *p*. Dlatego obszary wysp względem podłoża są spolaryzowane dodatnim napięciem zasilania układu. W normalnych warunkach polaryzacji wszystkie obszary tranzystora MOS (źródło, dren i kanał) są spolaryzowane zaporowo względem podłoża, na którym są wykonane, toteż w jednym obszarze podłoża (lub wyspy w przypadku tranzystorów pMOS) można umieścić wiele tranzystorów, i będą one wzajemnie od siebie odizolowane.

Rysunek 3.8 wraz z komentarzami pokazuje główne etapy wytwarzania struktury układu CMOS. Pokazany jest proces zwany **LOCOS** (ang. **LOCal Oxidation of Silicon**). Lokalnie wytwarzany tlenek polowy oddziela obszary aktywne, w których wykonywane są tranzystory. W najbardziej zaawansowanych technologiach stosowany jest nieco inny sposób produkcji układów CMOS, w którym zamiast obszarów tlenku polowego wytwarza się metodą fotolitografii i trawienia rowki wypełniane następnie dielektrykiem ($SiO_2$). Technologie te znane są jako **STI** (ang. **Shallow Trench Isolation**). Występuje w nich wiele innych różnic w stosunku do klasycznej technologii LOCOS, które jednak z punktu widzenia projektowania układu nie są najistotniejsze i nie będą tu omawiane.

Płytka podłożowa - ma zwykle grubość około 1 mm i średnicę do 30 cm.

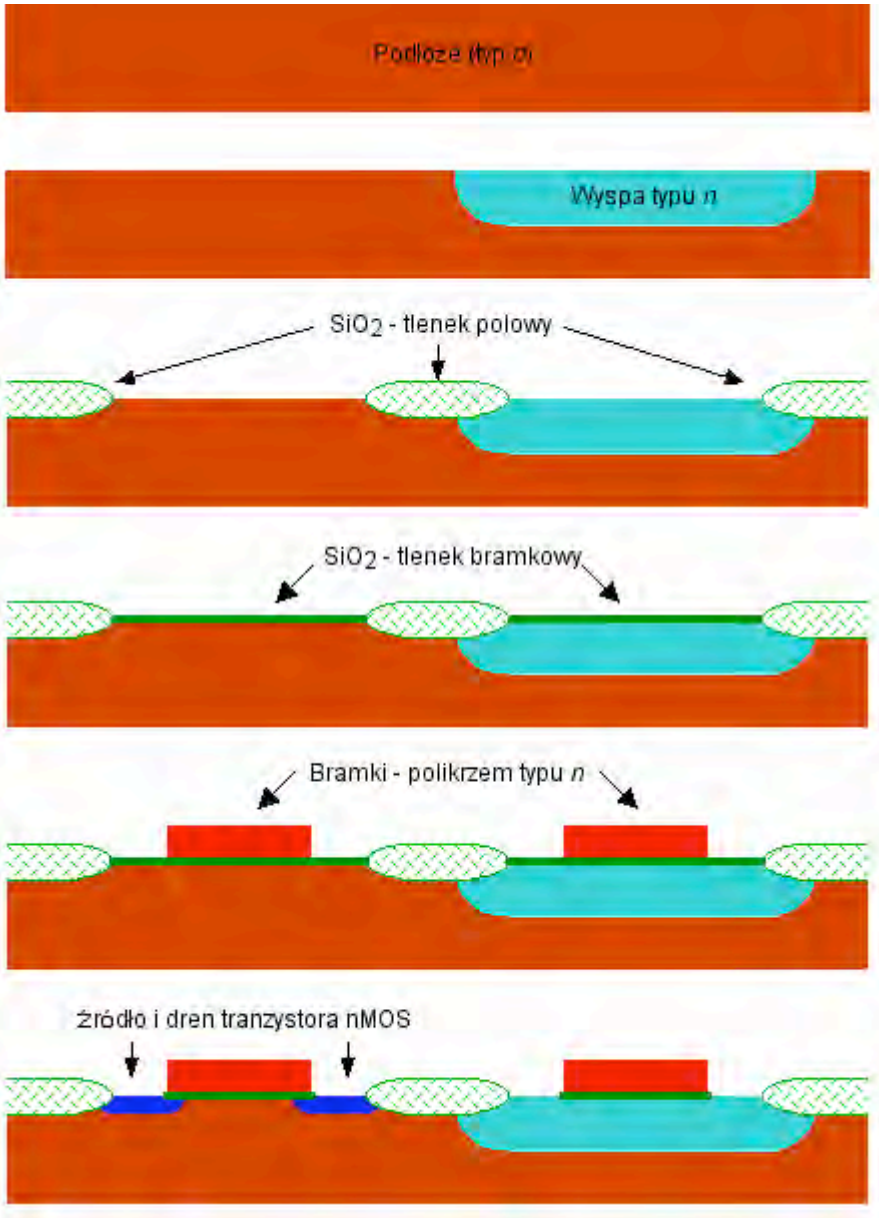
Wyspa typu n powstaje w wyniku procesu fotolitografii i następującego po nim procesu domieszkowania (implantacji jonów donorowych).

Obszary grubego (0.5 - 1 um) tlenku zwanego polowym, pomiędzy nimi obszary zwane aktywnymi. Powstawanie: płytka jest pokrywana azotkiem krzemu ($Si_3N_4$), który następnie jest usuwany w procesie fotolitografii znad obszarów, gdzie będzie tlenek polowy. Następnie płytka jest utleniania. Tlenek $SiO_2$ powstaje tam, gdzie usunięto azotek. Na koniec azotek jest usuwany chemicznie, pozostaje tlenek i odsłonięte obszary aktywne.

Tlenek bramkowy: powstaje w wyniku utlenienia odsłoniętej powierzchni krzemu w obszarach aktywnych. Jest bardzo cienki (w najbardziej zaawansowanych technologiach 2 - 3 nm).

Bramki powstają przez osadzenie warstwy polikrzemu domieszkowanego atomami donorowymi, oraz procesu fotolitografii.

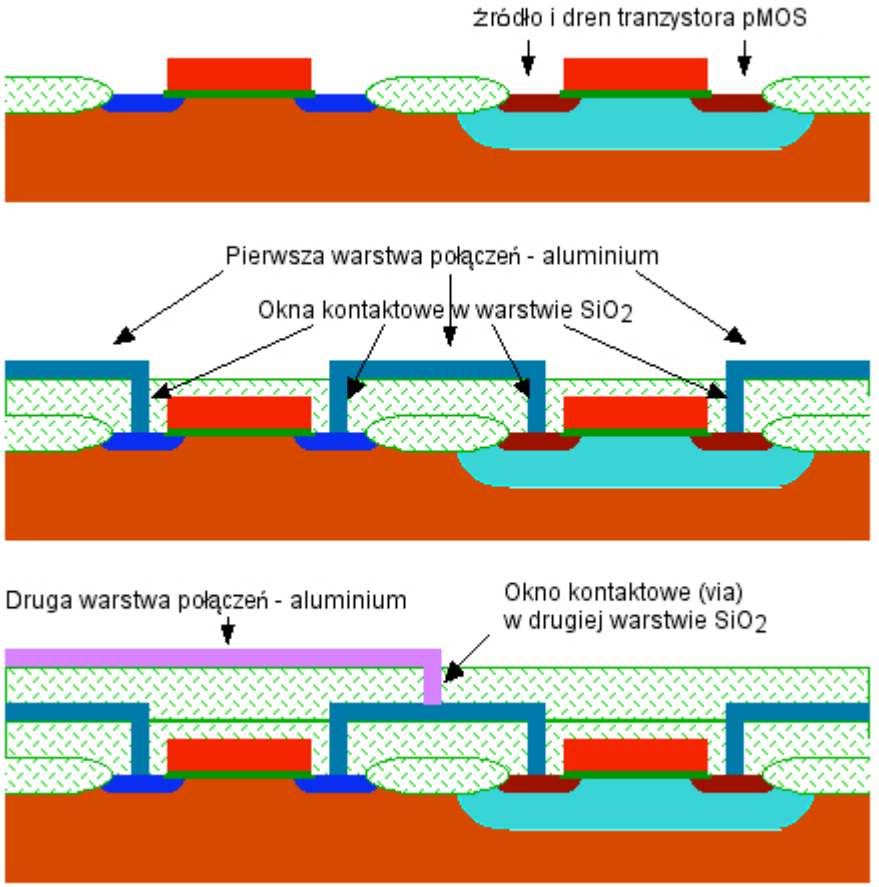
Źródła i dreny tranzystorów nMOS powstają w wyniku implantacji jonów donorowych. Przed tym procesem wykonywana jest fotolitografiia, której celem jest zasłonięcie obszarow aktywnych na wyspach. Tam będą tranzystory pMOS. W procesie implantacji obszary źródeł i drenów powstają tam, gdzie nie ma tlenku polowego ani polikrzemu - warstwy te są na tyle grube, że nie przepuszczają jonów domieszki.



Źródła i dreny tranzystorów pMOS powstają w wyniku implantacji jonów akceptorowych. Przed tym procesem wykonywana jest fotolitografiia, której celem jest zasłonięcie już wykonanych tranzystorów nMOS. W procesie implantacji obszary źródeł i drenów powstają tam, gdzie nie ma tlenku polowego ani polikrzemu - warstwy te są na tyle grube, że nie przepuszczają jonów domieszki.

Aby wykonać połączenia elektryczne, pokrywa się płytkę dielektrykiem, po czym wykonuje się fotolitografię i wytrawienie okien kontaktowych w tym dielektryku. Następnie płytkę pokrywa się warstwą metalu i wykonuje kolejną fotolitografię, w wyniku której powstają ścieżki połączeń. Na przekroju nie pokazano połączeń do bramek tranzystorów. W tym przekroju nie są one widoczne, ponieważ nie wykonuje się ich nad kanałami tranzystorów.

Kolejny poziom połączeń elektrycznych wykonuje się pokrywając poprzednie połączenia drugą warstwą dielektryka, następnie wykonuje się w niej okna kontaktowe (zwane potocznie via) do ścieżek połączeń pierwszej warstwy, osadza kolejną warstwę metalu i wykonuje kolejną fotolitografię, w wyniku której powstają ścieżki połączeń drugiego poziomu.



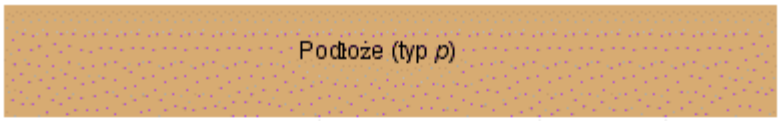

Rys. 3.8. Najważniejsze etapy wytwarzania struktury układu CMOS. Pokazany jest przekrój przez dwa tranzystory - nMOS i pMOS. Pokazano tylko dwa poziomy ścieżek połączeń, ale we współczesnych układach może być ich więcej, nawet ponad 10.

## 3.4. Procesy wytwarzania układów bipolarnych

W bipolarnych układach scalonych najważniejszymi elementami czynnymi są tranzystory bipolarne npn. Układy te dają jednak możliwość wykonania także innych elementów czynnych (tranzystorów bipolarnych pnp), a także biernych (rezystory, diody). Technologia bipolarna jest historycznie najstarszą, ale wciąż stosowaną technologią wytwarzania półprzewodnikowych układów scalonych, używaną głównie do wytwarzania układów analogowych. W tej technologii wzajemna izolacja elementów jest zapewniona przez umieszczenie ich na wyspach typu n otoczonych obszarami typu p, przy czym wyspy muszą być spolaryzowane zaporowo, czyli napięciem dodatnim względem otaczających je obszarów. W układach bipolarnych wyspa jest częścią składową struktury tranzystora npn (kolektorem), toteż na wyspie można umieścić wspólnie dwa lub więcej tranzystorów npn tylko wtedy, gdy ich kolektory są w schemacie układu połączone ze sobą.

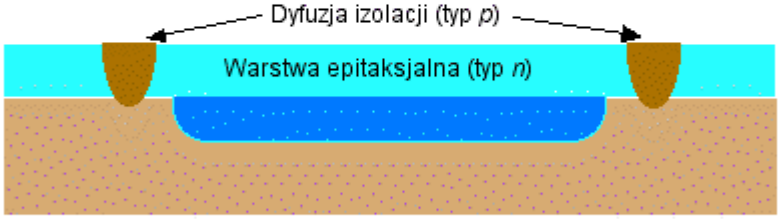
Rysunek 3.9 pokazuje kolejne etapy wytwarzania struktury układu bipolarnego. Jest to wspomniana wyżej klasyczna technologia ze złączową izolacją elementów. Współczesne układy bipolarne są często wytwarzane przy zastosowaniu innych, bardziej zaawansowanych technologii, które jednak nie są tu omawiane.

Płytka podłożowa typu *p*.

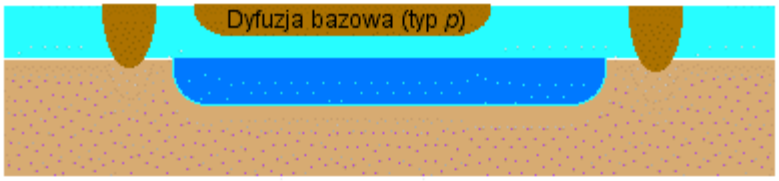



Warstwa zagrzebana typu *n* powstaje w wyniku procesu fotolitografii i następującego po nim procesu domieszkowania (implantacji lub dyfuzji jonów donorowych). Warstwa zagrzebana jest obszarem silnie domieszkowanym, chodzi o uzyskanie możliwie jak najmniejszej rezystancji tej warstwy.

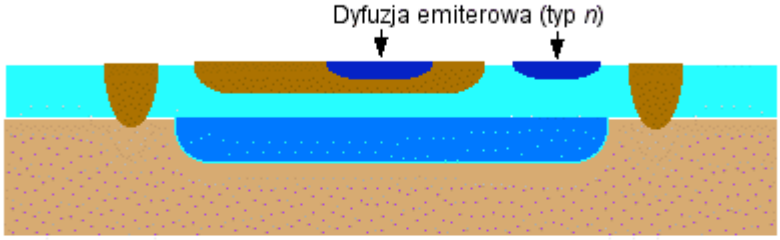

Po wytworzeniu warstwy zagrzebanej na płytce osadzana jest warstwa epitaksjalna. Jest ona, podobnie jak warstwa zagrzebana, typu *n*, ale o znacznie niższej koncentracji domieszki. Następuje kolejny proces fotolitografii i dyfuzja (lub implantacja) obszarów dyfuzji izolacji. Są to obszary typu p, sięgają płytki podłożowej. Obszary dyfuzji izolacji dzielą warstwę epitaksjalną na wyspy typu *n* otoczone obszarami typu *p*.



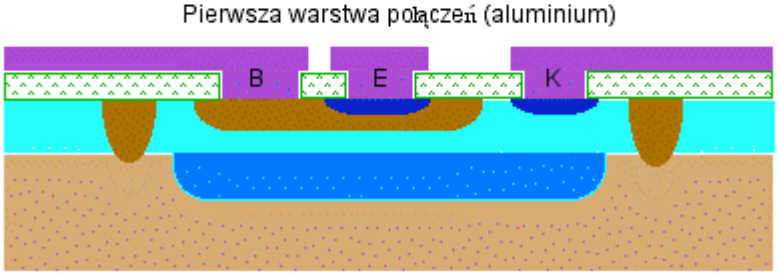

Po kolejnej fotolitografii następuje dyfuzja (lub implantacja) obszarów bazowych. Będą one stanowić bazy tranzystorów bipolarnych npn.




W obszarach bazowych (po kolejnej fotolitografii) wytwarzane są silnie domieszkowane obszary typu n (dyfuzja lub implentacja emiterowa). Będą to emitery tranzystorów bipolarnych npn. W tym samym procesie domieszkowane są też obszary warstwy zagrzebanej, do których wykonane będą kontakty.

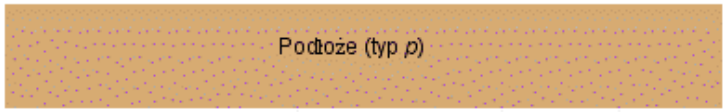



Połączenia wykonywane są na warstwie $SiO_2$, w której wykonuje się okna kontaktowe, następnie osadza się warstwę aluminium i po fotolitografii uzyskuje ścież ki połączeń. Podobnie jak w układach CMOS, warstw połączeń może być więcej niż jedna.



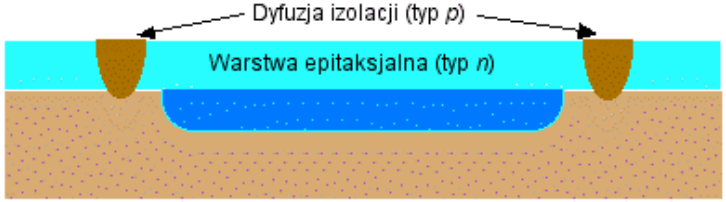

Rys. 3.9. Najważniejsze etapy wytwarzania struktury układu bipolarnego (technologia złączowej izolacji

Płytka podłożowa typu *p*.



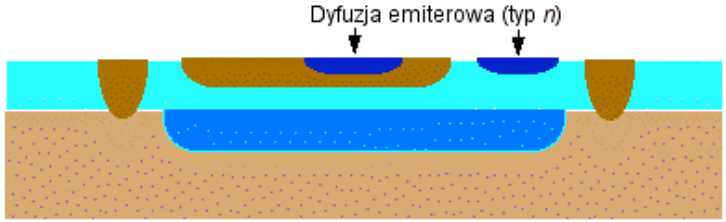

Warstwa zagrzebana typu *n* powstaje w wyniku procesu fotolitografii i następującego po nim procesu domieszkowania (implantacji lub dyfuzji jonów donorowych). Warstwa zagrzebana jest obszarem silnie domieszkowanym, chodzi o uzyskanie możliwie jak najmniejszej rezystancji tej warstwy.


Po wytworzeniu warstwy zagrzebanej na płytce osadzana jest warstwa epitaksjalna. Jest ona, podobnie jak warstwa zagrzebana, typu *n*, ale o znacznie niższej koncentracji domieszki. Następuje kolejny proces fotolitografii i dyfuzja (lub implantacja) obszarów dyfuzji izolacji. Są to obszary typu p, sięgają płytki podłożowej. Obszary dyfuzji izolacji dzielą warstwę epitaksjalną na wyspy typu *n* otoczone obszarami typu *p*.

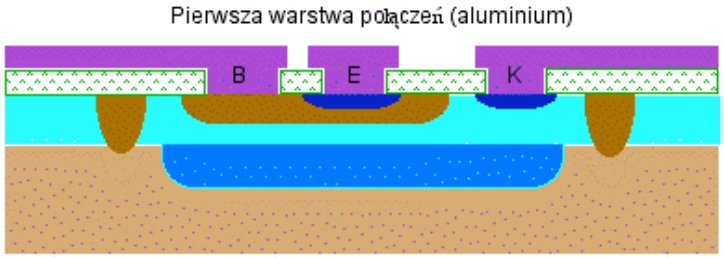



Po kolejnej fotolitografii następuje dyfuzja (lub implantacja) obszarów bazowych. Będą one stanowić bazy tranzystorów bipolarnych npn.


W obszarach bazowych (po kolejnej fotolitografii) wytwarzane są silnie domieszkowane obszary typu n (dyfuzja lub implentacja emiterowa). Będą to emitery tranzystorów bipolarnych npn. W tym samym procesie domieszkowane są też obszary warstwy zagrzebanej, do których wykonane będą kontakty.




Połączenia wykonywane są na warstwie $SiO_2$, w której wykonuje się okna kontaktowe, następnie osadza się warstwę aluminium i po fotolitografii uzyskuje ścieżki połączeń. Podobnie jak w układach CMOS, warstw połączeń może być więcej niż jedna.




Rys. 3.9. Najważniejsze etapy wytwarzania struktury układu bipolarnego (technologia złączowej izolacji elementów). Pokazany jest przekrój przez bipolarny tranzystor npn.

## 3.5. Montaż i obudowy układów scalonych

Jak wspomniano w wykładzie 2, układy scalone po wyprodukowaniu i testach ostrzowych są montowane w obudowach. Obudowy umożliwiają elektryczne i mechaniczne połączenie układu scalonego z urządzeniem, w którym układ ma działać, zapewniają ochronę układu przed uszkodzeniami mechanicznymi i szkodliwymi wpływami środowiska (np. wilgocią) oraz umożliwiają odprowadzenie ciepła wydzielającego się w czasie pracy układu. Zdecydowana większość obudów przystosowana jest do montażu na stałe na płytkach drukowanych przez lutowanie. Wiele z tych obudów umożliwia jednak także umieszczanie w podstawkach, co umożliwia łatwy demontaż i wymianę układu.

Same obudowy wykonywane są z tworzywa sztucznego lub z ceramiki. Obudowy z tworzyw są najtańsze w produkcji wielkoseryjnej, stosuje się je więc do montażu układów katalogowych produkowanych masowo. Istotną wadą obudów z tworzyw sztucznych jest znaczna różnica współczynnika rozszerzalności cieplnej tworzywa i płytki półprzewodnikowej. Ogranicza to zakres temperatur, w jakich mogą pracować układy zamknięte w takich obudowach. Obudowy te nie zapewniają także idealnej szczelności, zwłaszcza gdy poddawane są częstym zmianom temperatury w szerokim zakresie.

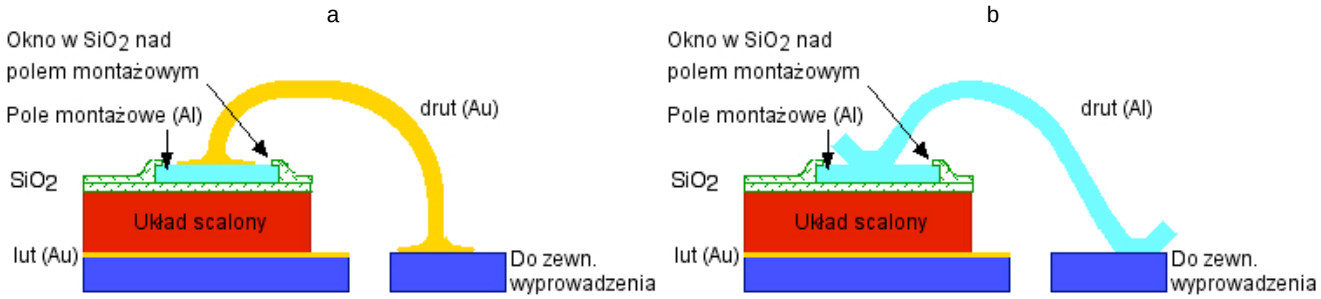
Obudowy ceramiczne są znacznie droższe, ale mają wiele zalet. Znacznie lepiej chronią układ przed szkodliwymi wpływami zewnętrznymi, umożliwiają pracę układu w szerszym zakresie temperatur oraz lepsze odprowadzanie ciepła. Dlatego układy zamknięte w takich obudowach cechują się zwykle wyższą niezawodnością. Obudowy ceramiczne są też dogodniejsze, gdy trzeba zmontować niewielką liczbę egzemplarzy układu, toteż w przypadku zamawiania prototypów układów specjalizowanych (ASIC) z reguły otrzymuje się je w obudowach ceramicznych.

W przypadku montażu w obudowie z tworzywa układ jest najpierw mocowany (zwykle lutowany lutem złotym) do przeznaczonego na to pola na metalowej kształtce zwanej ażurem. Następnie wykonywane są połączenia drutowe pomiędzy polami montażowymi układu, a paskami metalu, które w gotowym układzie będą służyć jako zewnętrzne wyprowadzenia elektryczne. Po wykonaniu połączeń układ jest zalewany tworzywem w formie o odpowiednim kształcie. Po zastygnięciu tworzywa zbędne fragmenty ażuru są odcinane, a wyprowadzenia zaginane tak, by mogły służyć do połączenia przez lutowanie z punktami lutowniczymi na płytce drukowanej.

Obudowa ceramiczna składa się z dwóch części: podstawy i pokrywki. Podstawa jest niemal kompletną obudową. Zawiera wnękę, w której umieszczony będzie układ, oraz komplet odpowiednio ukształtowanych zewnętrznych wyprowadzeń. Układ jest mocowany (lutowany lutem złotym lub klejony) do przeznaczonego na to pola we wnęce, a następnie wykonywane są drutowe połączenia między polami montażowymi układu, a wyprowadzeniami. Zmontowany w obudowie układ jest zamykany trwale i szczelnie metalową lub ceramiczną pokrywką.

Warto wiedzieć, że w przypadku zamawiania prototypów można na życzenie otrzymać kilka egzemplarzy układu, w których pokrywka nie jest trwale przymocowana, co pozwala obejrzeć układ pod mikroskopem zwykłym lub elektronowym, a niekiedy nawet "naprawić" układ, w którym wystąpił błąd w projekcie, na przykład przeciąć laserem błędnie wykonane połączenie. Układy nie zamknięte pokrywką służą oczywiście tylko do testów i badań, a nie do normalnej eksploatacji.

Połączenia między układem, a zewnętrznymi wyprowadzeniami wykonuje się drutem złotym metodą termokompresji lub niekiedy drutem aluminiowym metodą ultrakompresji. Metoda termokompresji polega na tym, że drut z uformowaną na końcu kulką jest specjalnym narzędziem dociskany do miejsca, w którym ma nastąpić elektryczne połączenie, a wszystko to dzieje się w podwyższonej temperaturze. Pod wpływem nacisku i wysokiej temperatury kulka ulega deformacji i zarazem trwale łączy się z metalem, do którego jest dociskana. W metodzie ultrakompresji zamiast wysokiej temperatury stosuje się drgania ultradźwiękowe, a drut jest dociskany narzędziem o kształcie klina. Oba sposoby wykonania połączeń pokazuje rys. 3.10.




Rys. 3.10. Połączenie termokompresyjne (a) i ultrakompresyjne (b)

Przykłady układów zmontowanych w obudowach - dodatek 3

Zarówno obudowy z tworzyw, jak i ceramiczne mogą być przeznaczone do montażu zwanego przewlekanym (wyprowadzenia przechodzą na wylot przez otwory w płytce drukowanej i są lutowane do ścieżek znajdujących się po przeciwnej stronie, niż układ) lub do montażu powierzchniowego (wyprowadzenia są lutowane do ścieżek po tej samej stronie, po której znajduje się układ, nie przechodzą przez otwory w płytce drukowanej). Wiele rodzajów obudów umożliwia także umieszczenie układu w podstawce, co pozwala na łatwy demontaż i wymianę układu. Ten sposób jest szczególnie godny polecenia, gdy wykonuje się prototypowe urządzenie, lub gdy przewiduje się możliwość wymiany układu na inny przez użytkownika (przykład: płyta główna komputera, w której można użyć kilku różnych typów lub wersji procesora). Natomiast układy trwale wlutowane w płytkę drukowaną są trudne do wylutowania i wymiany bez uszkodzenia płytki. Dotyczy to zwłaszcza precyzyjnych wielowarstwowych płytek drukowanych, na których zminiaturyzowane elementy są zmontowane metodą montażu powierzchniowego. Obecnie powszechnie stosowane są luty bezołowiowe, których temperatura topnienia jest wyższa, niż dawniej używanych lutów ołowiowo-cynowych. Demontaż układu wlutowanego lutem bezołowiowym jest szczególnie trudny.

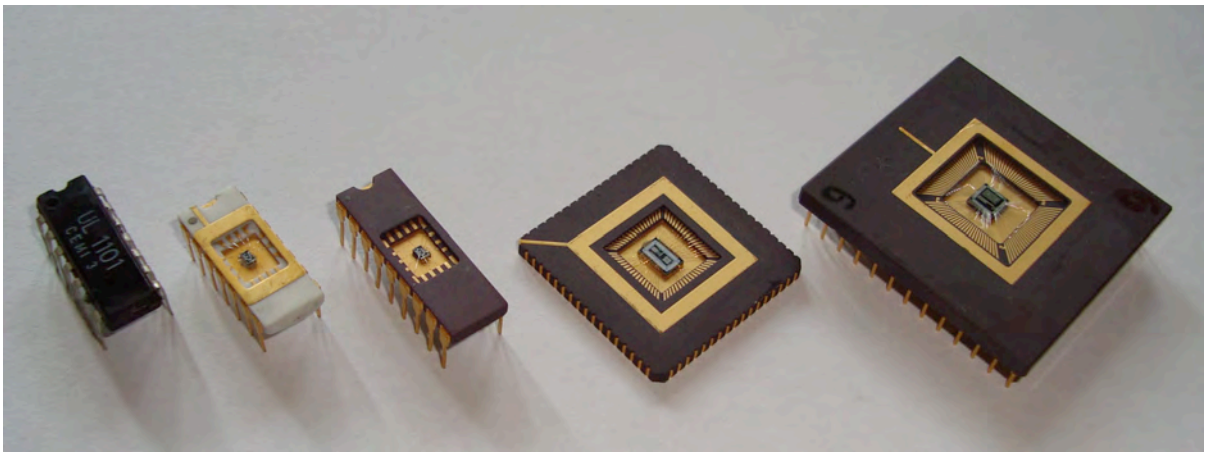
Stosowane bywa także montowanie układu bez obudowy. Ten sposób jest wykorzystywany wtedy, gdy nawet najmniejsze z istniejących obudów zajęłyby zbyt dużo miejsca (np. w zegarkach, kartach płatniczych itp.) lub gdy tradycyjne połączenia drutowe między układem i wyprowadzeniami byłyby zbyt długie (układy mikrofalowe). Układ przyklejony bezpośrednio do płytki drukowanej lub podłoża ceramicznego jest po wykonaniu drutowych połączeń zabezpieczany kroplą tworzywa sztucznego.

Rysunki i dane techniczne wielu typów obudów można znaleźć m.in. na stronach serwisu MPW EUROPRACTICE.

Przy obchodzeniu się z układami scalonymi trzeba pamiętać, że są one wrażliwe na wyładowania elektrostatyczne. Dotyczy to zwłaszcza wejść układów CMOS wykonanych w najnowocześniejszych technologiach. Napięcie przebicia bramki tranzystora MOS, w którym grubość dielektryka pod bramką wynosi kilka nanometrów, ma wartość niewiele wyższą, niż maksymalne dopuszczalne napięcie zasilania układu. Jest to wartość na poziomie 2 - 5 V. Tymczasem nawet bardzo niewielki ładunek elektrostatyczny, jaki może powstać na przykład w wyniku tarcia, prowadzi do powstania napięć większych o rzędy wielkości (ciało człowieka spacerującego po podłodze wyłożonej wykładziną dywanową z tworzywa sztucznego w pomieszczeniu o bardzo niskiej wilgotności powietrza może łatwo naładować się do napięcia rzędu kilku kV). Układy scalone zawierają na wejściach i wyjściach specjalne bufory zabezpieczające do pewnego stopnia przed uszkodzeniami spowodowanymi wyładowaniami elektrostatycznymi. Mimo to **obchodzenie się z układami scalonymi wymaga szczególnej ostrożności**. Producenci dostarczają je w specjalnych pojemnikach z tworzyw przewodzących prąd elektryczny. Wyjmowanie układów z takich pojemników i ich montaż powinny odbywać się na odpowiednio zabezpieczonym stanowisku pracy - na uziemionej płycie metalowej, uziemione powinny być też narzędzia, a także człowiek manipulujący układami (służą do tego specjalne opaski na ręce) - ogólnie wszystko, z czym układ może się zetknąć.

Szczegółowe omówienie technologii montażu sprzętu elektronicznego wykracza poza zakres przedmiotu "Układy scalone". Producenci układów podają w swoich danych technicznych między innymi zalecane techniki i wymagane warunki montażu (jak np. maksymalna temperatura i czas lutowania) układów w urządzeniach, tak by nie spowodować uszkodzenia obudowy ani umieszczonego w niej układu.

## 3.5. Dodatek 3: Przyklady układów scalonych w obudowach



Przykłady układów w obudowach.
Pierwsza z lewej: obudowa typu DIL14 z tworzywa, do montażu przewlekanego.
Druga: obudowa typu DIL14 ceramiczna, z pokrywką metalową, do montażu przewlekanego.
Trzecia: obudowa typu DIL18 ceramiczna, z pokrywką ceramiczną, do montażu przewlekanego.
Czwarta: obudowa typu PLCC68 ceramiczna, z pokrywką ceramiczną, do montażu w podstawce.
Piąta: obudowa typu PGA100 ceramiczna, do montażu w podstawce lub przewlekanego.
Wszystkie obudowy ceramiczne mają zdjęte pokrywki dla pokazania wnętrza obudowy ze zmontowanym układem.

## Bibliografia

[1] Romuald B. Beck, "*Technologia krzemowa*", PWN Warszawa 1991

# Wykład 4: Przegląd elementów układów scalonych

## Wstęp

W wykładzie 4 omawiane są elementy, z których buduje się układy scalone. Są to przede wszystkim tranzystory: MOS i bipolarne. W układach scalonych (głównie analogowych) używane są też elementy bierne: diody, rezystory, kondensatory, a ostatnio w układach pracujących w zakresie mikrofal także indukcyjności: cewki i transformatory. Tych ostatnich nie będziemy omawiać, bo tematyka układów mikrofalowych wykracza poza zakres przedmiotu "Układy scalone", natomiast pozostałe elementy, a zwłaszcza tranzystory MOS, będą omówione na tyle szczegółowo, aby można było wykorzystać tę wiedzę w dalszych wykładach, tak by można było rozważać nie tylko budowę, ale także podstawowe parametry elektryczne bramek cyfrowych i prostych układów analogowych.

Materiał wykładu 4 jest w dużej części powtórzeniem z innych przedmiotów. Przyjęte zostało założenie, że zasady działania omawianych elementów i zachodzące w nich zjawiska fizyczne zostały poznane wcześniej. Dlatego nie ma tu wyprowadzeń podstawowych wzorów opisujących charakterystyki elementów.

W wykładzie 4 jest też mowa o elementach zwanych pasożytniczymi. Mogą one mieć znaczny wpływ na działanie układów.

Wykład 4 jest dość długi. Warto dobrze przyswoić jego materiał, ponieważ będą do niego liczne odwołania w całej dalszej części podręcznika.

## 4.1. Tranzystory MOS, ich charakterystyki i parametry

**Charakterystyki prądowo-napięciowe**

Przypomnimy teraz charakterystyki tranzystora MOS i ich opis. W układach CMOS wykorzystywane są tranzystory p-kanałowe (w skrócie pMOS) i tranzystory n-kanałowe (w skrócie nMOS). Oba rodzaje tranzystorów są typu wzbogacanego, czyli do bramki trzeba przyłożyć napięcie, aby utworzył się kanał między źródłem i drenem i tranzystor zaczął przewodzić. W przypadku tranzystorów nMOS jest to napięcie dodatnie względem źródła, a w przypadku tranzystorów pMOS - ujemne. Gdy analizujemy działanie tranzystora MOS w bramkach logicznych, można w uproszczeniu powiedzieć, że tranzystor nMOS jest włączany napięciem dodatnim, a pMOS - ujemnym, przy czym w każdym przypadku napięcie to powinno być większe co do wartości bezwzględnej od napięcia progowego tranzystora $U_T$.

W rozważaniach dotyczących bramek logicznych i układów analogowych potrzebna nam będzie przede wszystkim znajomość opisu charakterystyk prądowo-napięciowych prądu drenu $I_D$ w funkcji napięcia dren-źródło $U_{DS}$ i napięcia bramki $U_{GS}$. Można je w najprostszy sposób opisać wzorami:
- w zakresie zwanym **zakresem podprogowym**:

$$I_D = 0 \qquad \text{dla} \qquad U_{GS} < U_T \tag{4.1}$$

- w zakresie zwanym **zakresem liniowym** (choć charakterystyki w tym zakresie wcale nie są liniowe!):

$$I_D = \mu C_{ox} \frac{W}{L}\left[(U_{GS} - U_T)U_{DS} - \frac{U_{DS}^2}{2}\right] \qquad \text{dla} \qquad \begin{array}{l} U_{GS} \geq U_T \text{ i} \\ U_{DS} \leq U_{DSsat} \end{array} \tag{4.2}$$

gdzie $U_{DSsat}$ nazywane jest **napięciem nasycenia**,

$$U_{DSsat} = U_{GS} - U_T \tag{4.3}$$

- oraz w zakresie zwanym **zakresem nasycenia**:

$$I_D = \mu C_{ox} \frac{W}{L} \frac{(U_{GS} - U_T)^2}{2} \qquad \text{dla} \qquad \begin{array}{l} U_{GS} \geq U_T \text{ i} \\ U_{DS} \geq U_{DSsat} \end{array} \tag{4.4}$$

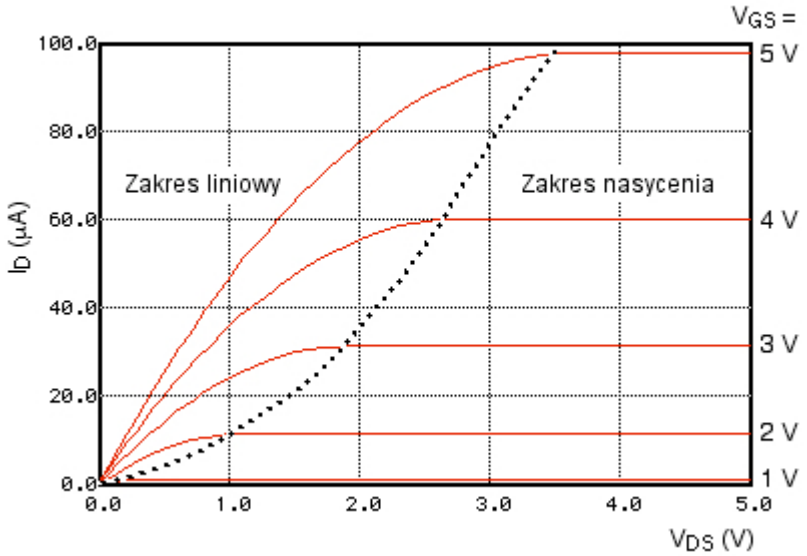
W tych wzorach $W$ jest szerokością kanału, $L$ - jego długością, $\mu$ jest ruchliwością nośników ładunku w kanale, $C_{ox}$ - pojemnością tlenku bramkowego na jednostkę powierzchni.

! **Długość i szerokość kanału są to jedyne dwie wielkości, które może zmieniać konstruktor projektując układy z tranzystorami MOS. Napięcie progowe $U_T$, podobnie jak ruchliwość nośników ładunku w kanale $\mu$ oraz jednostkowa pojemność tlenku bramkowego $C_{ox}$, są określone przez proces technologiczny, w którym wytwarzane są układy. Projektant nie ma możliwości zmiany tych wielkości.**

Wartości $U_T$, $\mu$ oraz $C_{ox}$ podaje producent układów.

! **Ważna uwaga: wzory (4.1) - (4.4) opisują charakterystyki tranzystora nMOS. W przypadku tranzystora pMOS można stosować te same wzory podstawiając do nich wartość bezwzględną napięcia progowego (które - jak wiemy - jest dla tranzystorów pMOS ujemne).**

Przykładowa rodzina charakterystyk opisana zależnościami (4.1) - (4.4) wygląda tak:




Rys. 4.1. Rodzina charakterystyk tranzystora nMOS według wzorów (4.1) - (4.4). Linia przerywana oddziela zakres liniowy od zakresu nasycenia.

Wzory (4.1) - (4.4) stanowią podstawę najprostszego modelu matematycznego tranzystora MOS używanego w symulatorach układów elektronicznych, zwanego modelem poziomu 1 ("Level 1"). (To określenie pochodzi z najstarszych wersji symulatora SPICE, w których dostępne były trzy modele tranzystora MOS o różnym stopniu komplikacji i różnej dokładności rozróżniane wartością parametru o nazwie "level" i wartościach równych 1, 2 i 3).

Wzory (4.1) - (4.4) opisują charakterystyki tranzystora w bardzo uproszczony sposób. Nie są w nich uwzględnione liczne zjawiska fizyczne wpływające bardzo poważnie na kształt charakterystyk takich tranzystorów MOS, jakie występują we współczesnych układach CMOS. Wzory te będziemy jednak stosować do prostych obliczeń ilustrujących działanie bramek logicznych i ukłdów analogowych, ponieważ umożliwiają one łatwe wyprowadzenie podstawowych zależności ilustrujących jakościowo właściwości tych bramek i układów. Niemniej trzeba pamiętać, że do symulacji elektrycznej w praktycznych pracach projektowych wzory (4.1) - (4.4) i oparty na nich model "level 1" w żadnym przypadku nie wystarczają, a otrzymane przy ich użyciu wyniki będą z reguły bardzo odległe od rzeczywistości. Modelami stosowanymi dziś najczęściej w projektowaniu układów CMOS są modele zwane BSIM3, BSIM4 i PSP. Występują one w różnych symulatorach z różnymi wartościami parametru "level". Są to modele skomplikowane od strony matematycznej, ale dobrze oddające charakterystyki tranzystorów o bardzo małych długościach kanału (zobacz porównanie z modelem "level1" - dodatek 1). Producenci układów scalonych podają wartości parametrów tych modeli dla typowych struktur tranzystorów wytwarzanych w dostępnych u nich procesach technologicznych.

Dokładność wzorów (4.1) - (4.4) można nieco poprawić uwzględniając dwa ważne zjawiska występujące w tranzystorach MOS: zależność napięcia progowego $U_T$ od napięcia polaryzacji podłoża względem źródła $U_{BS}$ ("**efekt polaryzacji podłoża**") i zjawisko zależności elektrycznej długości kanału od napięcia dren-źródło $U_{DS}$ ("**efekt skracania kanału**").

Zależność napięcia progowego od napięcia polaryzacji podłoża można w przybliżony sposób opisać wzorem

$$U_T = U_{T0} + \gamma\left(\sqrt{|2\phi_F| - U_{BS}} - \sqrt{|2\phi_F|}\right) \tag{4.5}$$

gdzie $U_{T0}$ jest napięciem progowym dla napięcia polaryzacji podłoża równego zeru, zaś $\phi_F$ i $\gamma$ są parametrami modelu. $\phi_F$ ma sens potencjału Fermiego w podłożu tranzystora. $\gamma$ zależy od domieszkowania podłoża tranzystora $N_B$ i w przybliżeniu wyraża się wzorem

$$\gamma = \frac{\sqrt{2q\varepsilon_{Si}N_B}}{C_{ox}} \tag{4.6}$$

w którym $q$ jest ładunkiem elementarnym, zaś $\varepsilon_{Si}$ jest przenikalnością dielektryczną krzemu. W praktyce wartości $\phi_F$ i $\gamma$ są dobierane tak, by uzyskać najlepszą zgodność charakterystyk opisanych wzorami (4.1) - (4.6) i rzeczywistych. Wartości te podaje producent układów.

Efekt skracania kanału polega na tym, że rzeczywista "elektryczna" długość kanału jest mniejsza od długości projektowej $L$, z dwóch powodów: po pierwsze, obszary domieszkowane źródła i drenu zachodzą pod obszar bramki na pewną odległość $\Delta L$, a po drugie warstwa zaporowa złącza drenu wnika na pewną odległość w obszar kanału efektywnie skracając go. Ten drugi efekt powoduje, że długość kanału maleje ze wzrostem napięcia $U_{DS}$, zaś prąd drenu wzrasta. W rezultacie w zakresie nasycenia prąd drenu nie jest stały (jak wynikałoby ze wzoru (4.4) i rysunku 4.1), lecz wzrasta. Oba te efekty łącznie można uwzględnić opisując długość kanału $L$ we wzorach (4.2) i (4.3) zależnością:

$$\frac{1}{L} = \frac{1}{L_t - 2\Delta L}\left(1 + \lambda U_{DS}\right) \tag{4.7}$$

W tym wzorze $\lambda$ jest parametrem empirycznym, wyznaczanym tak, by charakterystyki tranzystora w zakresie nasycenia miały nachylenie zgodne z rzeczywiście obserwowanym. Wartość tego parametru podaje producent układów.

Wszytkie podane wyżej wzory można stosować zarówno dla tranzystorów nMOS, jak i dla pMOS. Dla tranzystorów nMOS w normalnych warunkach pracy napięcia $U_{DS}$ i $U_{GS}$ są dodatnie, podobnie jak napięcie progowe. Napięcie polaryzacji podłoża $U_{BS}$ jest ujemne, gdy podłoże jest spolaryzowane względem źródła zaporowo (taka polaryzacja jest typowa i dopuszczalna). Prąd drenu uważamy za dodatni. Dla tranzystorów pMOS będziemy także przyjmować, podobnie jak w większości podręczników, że napięcia $U_{DS}$ i $U_{GS}$ są dodatnie, napięcie polaryzacji podłoża $U_{BS}$ jest ujemne, gdy podłoże jest spolaryzowane względem źródła zaporowo, i prąd drenu jest dodatni. Napięcie progowe tranzystorów pMOS jest ujemne, toteż w przypadku tych tranzystorów we wzorach będzie podstawiana wartość bezwzględna tego napięcia.

! **W dalszej części wykładu będziemy najczęściej przy wyprowadzaniu wzorów pomijali zarówno wpływ napięcia polaryzacji podłoża na napięcie progowe, jak i wpływ napięcia drenu na długość kanału tranzystora. Jest to równoznaczne z założeniem, że parametry $\gamma$, $\lambda$ oraz $\Delta L$ mają wartości równe zeru. Dzięki temu uzyskamy proste i łatwe do interpretacji zależności, trzeba jednak pamiętać, że w większości przypadków będą one dawać ilościowo wyniki dalekie od rzeczywistości.**

Wzór (4.1) przewiduje, że dla napięcia bramki mniejszego od progowego prąd drenu jest dokładnie równy zeru. Tak jednak w rzeczywistości nie jest. Gdy napięcie bramki staje się mniejsze od progowego, prąd drenu nie spada dokładnie do zera, lecz wykazuje zależność od napięcia bramki o charakterze wykładniczym. Prąd ten, zwany prądem podprogowym, dla napięć wyraźnie mniejszych od progowego można przybliżyć wyrażeniem

$$I_{Dp} = I_t \frac{W}{L} \exp\left[\frac{q\left(U_{GS} - U_T\right)}{nkT}\right]\left[1 - \exp\left(-\frac{qU_{DS}}{kT}\right)\right] \tag{4.8}$$

w którym $I_t$ oraz $n$ są parametrami zależnymi od konstrukcji tranzystora. Prąd podprogowy nie ma bezpośredniego wpływu na działanie większości typów bramek logicznych, jednak nie jest całkiem bez znaczenia, bowiem jego obecność zwiększa prąd, jaki pobierają ze źródła zasilania układy logiczne. Praca tranzystorów w zakresie podprogowym, czyli gdy prąd drenu jest wykładniczą funkcją napięcia bramki, bywa stosowana w niektórych układach analogowych.

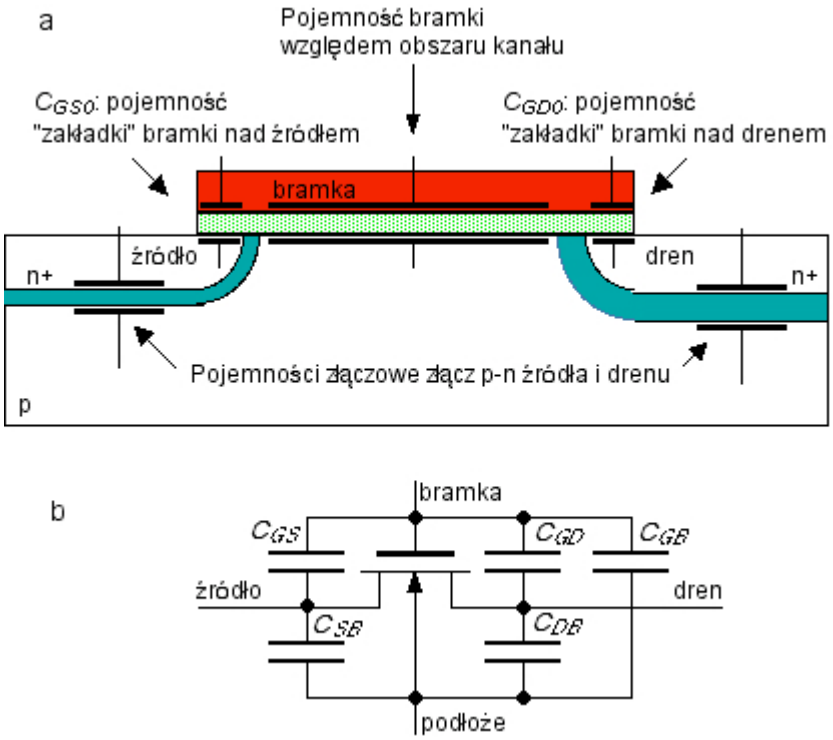
Charakterystyki prądowo-napięciowe złącz p-n źródło-podłoże i dren-podłoże opisywane są z dostateczną dla naszych celów dokładnością wzorem

$$I = I_s\left[\exp\left(\frac{qU}{n_i kT}\right) - 1\right] \tag{4.9}$$

gdzie $I_s$ jest prądem nasycenia złącza, $U$ - napięciem polaryzującym (ujemnym w przypadku polaryzacji zaporowej, dodatnim w przeciwnym razie), a $n_j$ - współczynnikiem o wartości zawartej zwykle między 1 i 2.

**Pojemności**

Do projektowania układów cyfrowych i niektórych rodzajów układów analogowych potrzebna jest, oprócz znajomości charakterystyk prądowo-napięciowych, także znajomość pojemności występujących w tranzystorze MOS. Są to pojemności typu metal-dielektryk-półprzewodnik związane z bramką tranzystora, oraz pojemności złączowe związane ze złączami p-n obszarów źródła oraz drenu. Wszystkie te pojemności ilustruje rys. 4.2.




Rys. 4.2. Pojemności w strukturze tranzystora MOS: (a) w strukturze fizycznej, (b) reprezentacja w schemacie

Jak widać, możemy wyróżnić trzy rodzaje pojemności:

- pojemności "zakładek" bramki nad źródłem i drenem $C_{GS0}$ i $C_{GD0}$,
- pojemności złącz p-n źródła i drenu,
- pojemność bramki względem obszaru kanału.

Pojemności $C_{GS0}$ i $C_{GD0}$ można uważać za niezależne od napięć polaryzujących tranzystor. Są one proporcjonalne do szerokości kanału tranzystora i dlatego w modelach tranzystorów są wyrażane jako pojemność na jednostkę długości (a nie powierzchni).

Pojemności złącz p-n źródła i drenu $C_j$ są opisywane najprostszą zależnością

$$C_j = \frac{C_{j0}}{\left[1 - \left(\frac{U}{U_D}\right)\right]^m} \tag{4.10}$$

w której $U$ jest napięciem polaryzującym złącze (ujemnym w przypadku polaryzacji zaporowej, dodatnim w przeciwnym przypadku), $U_D$ jest napięciem dyfuzyjnym, $C_{j0}$ jest pojemnością złącza niespolaryzowanego, zaś $m$ jest wykładnikiem o wartości zależnej od rozkładu domieszek w złączu. W bardziej dokładnych obliczeniach pojemności złączowe są rozdzielane na pojemności dna złącza (które jest złączem płaskim) i pojemności

obszarów bocznych (które nie są płaskie). Obie składowe całkowitej pojemności są opisywane tym samym wzorem (4.10), ale wartości parametrów $C_{j0}$, $U_D$ i $m$ są różne ze względu na różnice w kształcie obszarów płaskich i bocznych oraz różnice w rokładach domieszek. Parametr $C_{j0}$ jest zwykle podawany na jednostkę powierzchni złącza w przypadku obszarów dna i na jednostkę obwodu (czyli długości) w przypadku obszarów bocznych.

Z pojemnością bramki względem obszaru kanału sytuacja jest bardziej skomplikowana. Pojemność ta musi być dla celów symulacji układów elektronicznych „rozdzielona“ na składowe: pojemność bramka-dren, pojemność bramka-źródło i pojemność bramka-podłoże. Sposób tego podziału zależeć musi od napięć polaryzujących. Przykładowo, w zakresie głęboko podprogowym, gdy kanał nie istnieje, uzasadnione jest utożsamienie całej pojemności bramki z pojemnością bramka-podłoże. Gdy kanał istnieje, ekranuje on elektrostatycznie bramkę od podłoża. Wówczas mówienie o pojemności bramka-podłoże traci sens, a pojemność bramki względem kanału musi być w jakiejś proporcji podzielona na dwie: bramka-źródło i bramka-dren. W prostym modelu („level 1“) przyjęto dość arbitralnie następujące założenia:

- całkowita pojemność bramki jest równa $C_g = WLC_{ox}$
- w zakresie podprogowym pojemność ta jest równa pojemności bramka-podłoże $C_{GB}$; pojemności bramka-dren i bramka-źródło są równe zeru,
- w zakresie liniowym pojemność ta jest dzielona po połowie między pojemność bramka-dren $C_{GD}$ i pojemność bramka-źródło $C_{GS}$, zaś pojemność bramka-podłoże jest równa zeru,
- w zakresie nasycenia pojemność ta jest przypisywana pojemności bramka-źródło $C_{GS}$, przy czym przyjmuje się, że jest zmniejszona do wartości równej $2WLC_{ox}/3$; pojemności bramka-dren i bramka-podłoże są równe zeru.

Ponieważ w rzeczywistości pojemności nie zmieniają się w sposób skokowy, w modelu "level 1" wartości pojemności obliczane według powyższych założeń są w zakresach pośrednich pomiędzy podprogowym a liniowym, czy też liniowym a nasycenia „sklejane“ przy pomocy odpowiednio dobranych krzywych przejściowych. Taki model pojemności cechuje prostota, ale niestety ma on fundamentalną wadę: można pokazać, iż nie spełnia zasady zachowania ładunku. Dlatego wyniki symulacji układów, w których istotna jest zmiana ładunku w funkcji czasu, należy z góry traktować jako mało dokładne. Zaawansowane modele (jak np. wspomniane wyżej modele BSIM3, BSIM4 i PSP) używają innych, bardziej złożonych metod obliczania pojemności, w których zasada zachowania ładunku nie jest naruszona. Do naszych rozważań jednak proste modele opisane wyżej będą wystarczające.

W rozważaniach układowych wszystkie pojemności są reprezentowane przez pojemności bramka-źródło $C_{GS}$, bramka-dren $C_{GD}$, bramka-podłoże $C_{GB}$, źródło-podłoże $C_{SB}$ i dren-podłoże $C_{DB}$ (patrz rys. 4.2b ). Pojemności te mają następujące składowe:

- pojemność $C_{GS}$ jest sumą pojemności "zakładki" $C_{GS0}$ i uzależnionej od napięcia części pojemności bramki $C_g$
- pojemność $C_{GD}$ jest sumą pojemności "zakładki" $C_{GD0}$ i uzależnionej od napięcia części pojemności bramki $C_g$
- pojemność $C_{GB}$ jest uzależnioną od napięcia częścią pojemności bramki $C_g$, ma wartość równą zeru gdy tranzystor przewodzi (istnieje kanał między źródłem i drenem)
- pojemność $C_{SB}$ jest równa pojemności złączowej źródła
- pojemność $C_{DB}$ jest równa pojemności złączowej drenu.

**Parametry małosygnałowe**

W przypadku układów analogowych charakterystyki prądowo-napięciowe i pojemności także są ważne, ale oprócz nich musimy też operować parametrami małosygnałowymi tranzystora, ponieważ układy analogowe bardzo często przeznaczone są do pracy z sygnałami o małej amplitudzie.

Dla elementów aktywnych wygodnie jest posługiwać się trzema parametrami małosygnałowymi: konduktancją wejściową $g_{we}$, konduktacją wyjściową $g_{wy}$ oraz transkonduktancją $g_m$. Pierwszy z tych parametrów nie ma zastosowania w przypadku tranzystorów MOS, ponieważ prąd wejściowy tych tranzystorów jest równy zeru, zatem $g_{we}=0$. Pozostałe dwa parametry otrzymamy różniczkując zależności opisujące charakterystyki prądowo-

napięciowe.

Konduktancja wyjściowa $g_{wy}$ jest zdefiniowana następująco:

$$g_{wy} = g_{ds} = \frac{\partial I_D}{\partial U_{DS}} \tag{4.11}$$

Dla zakresu liniowego różniczkując wzór (4.2) otrzymujemy

$$g_{ds} = K\left(U_{GS} - U_T - U_{DS}\right) \tag{4.12}$$

w którym $K$ oznacza współczynnik przewodności tranzystora

$$K = \mu C_{ox} \frac{W}{L} \tag{4.13}$$

Dla zakresu nasycenia musimy zróżniczkować wzór (4.4) w połączeniu z (4.7); otrzymujemy

$$g_{ds} = \lambda I_D \tag{4.14}$$

Transkonduktacja $g_m$ jest zdefiniowana jako

$$g_m = \frac{\partial I_D}{\partial U_{GS}} \tag{4.15}$$

Dla zakresu liniowego, różniczkując wzór (4.2), otrzymujemy

$$g_m = KU_{DS} \tag{4.16}$$

zaś dla zakresu nasycenia, różniczkując wzór (4.4), otrzymujemy

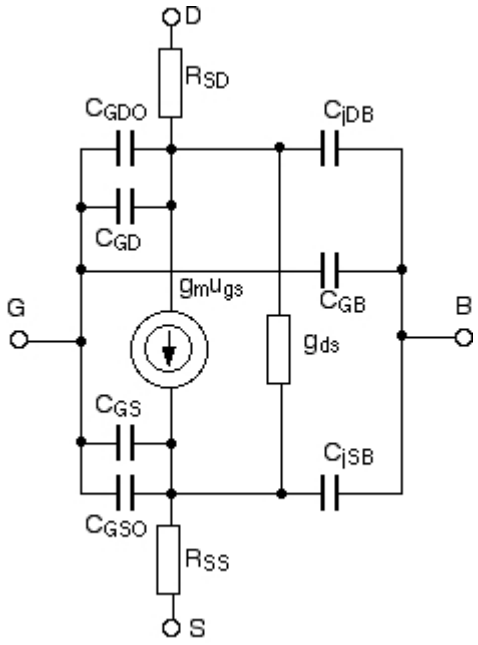
$$g_m = K\left(U_{GS} - U_T\right) = \sqrt{2KI_D} = \frac{2I_D}{U_{GS} - U_T} \tag{4.17}$$

(wszystkie trzy postacie wzoru na $g_m$ są równoważne). Przydatna bywa także wartość $g_m$ w zakresie podprogowym, którą otrzymujemy różniczkując wzór (4.8)

$$g_m = \frac{I_D}{n\frac{kT}{q}} \tag{4.18}$$

Transkonduktancja jest w układach analogowych ważnym parametrem, ponieważ od jej wartości zależy wzmocnienie, jakiego może dostarczyć tranzystor.

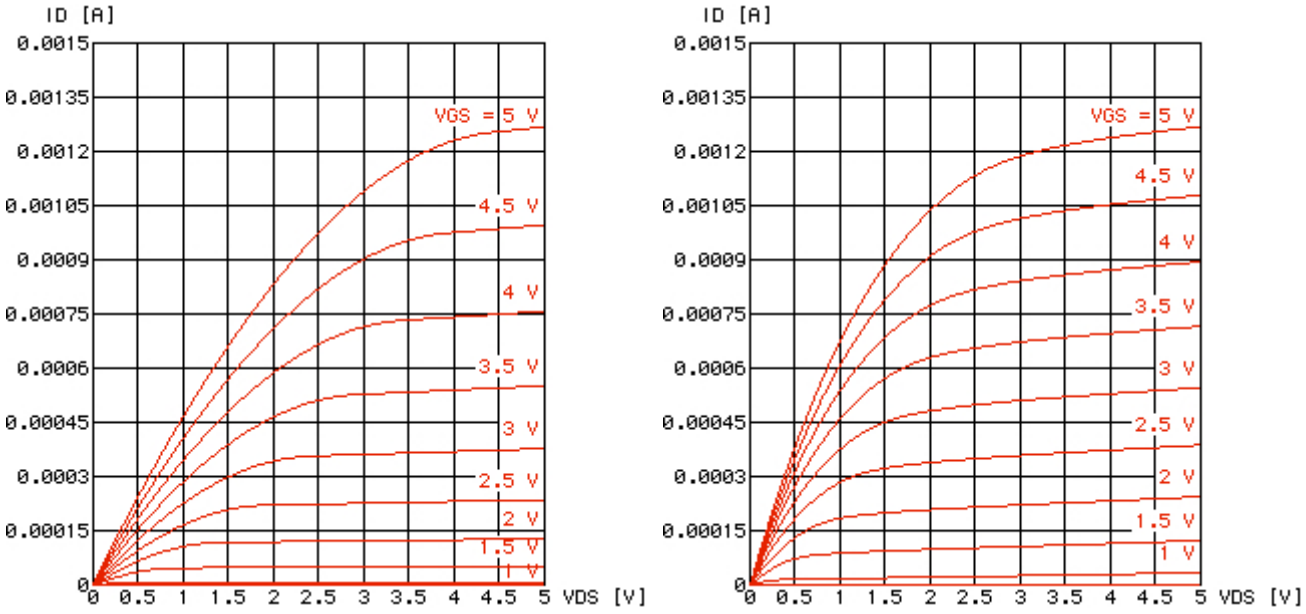
Te zależności opisują właściwości tranzystora dla sygnałów o małej amplitudzie i małej częstotliwości. Analiza dla dużych częstotliwości wymaga uwzględnienia pojemności tranzystora. Najwygodniej wówczas posługiwać się małosygnałowym schematem zastępczym tranzystora, który wygląda następująco:




Rys. 4.3. Małosygnałowy schemat zastępczy tranzystora MOS.
Uwzględniono w nim także rezystancje rozproszone źródła $R_{SS}$ i drenu $R_{SD}$ oraz pojemności złączowe źródła $C_{jSB}$ i drenu $C_{jDB}$.

Małosygnałowy schemat zastępczy tranzystora stosujemy wtedy, gdy rozważamy działanie układu z tym tranzystorem dla sygnałow zmiennych o małej amplitudzie (co w praktyce oznacza amplitudy rzędu co najwyżej miliwoltów). Wszystkie elementy w takim schemacie są liniowe. Ich parametry (np. pojemność kondensatorów, konduktancja wyjściowa itp.) zależą od punktu pracy tranzystora, tj. od wartości składowych stałych napięć polaryzujących. Występujące w schemacie źródł o prądowe wymuszające prąd $g_m u_{gs}$ reprezentuje efekt sterowania w tranzystorze, tj. składową zmienną prądu wyjściowego powstającą na skutek istnienia składowej zmiennej napięcia wejściowego bramka-źródło o amplitudzie $u_{gs}$.

## 4.1. Dodatek 1: Porownanie modeli tranzystora MOS



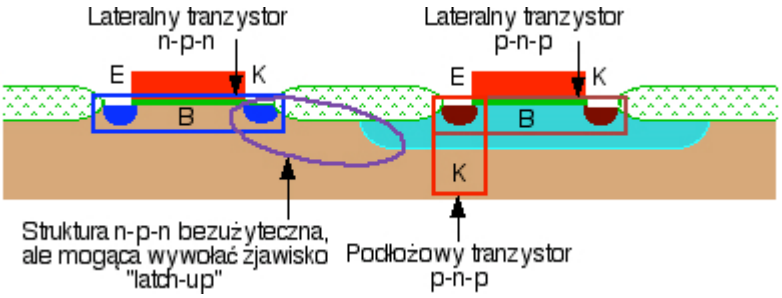

Porównanie charakterystyk tego samego tranzystora opisanych modelem "level 1" (po lewej) i modelem BSIM3 (po prawej).

## 4.2. Tranzystory bipolarne, ich charakterystyki i parametry

### Struktury i charakterystyki prądowo-napięciowe

Ponieważ we współczesnej mikroelektronice królują układy CMOS, zajmiemy się przede wszystkim tranzystorami bipolarnymi w tych układach. W układzie scalonym CMOS tranzystory bipolarne występują zwykle tylko jako elementy pasożytnicze (dokładniej o elementach pasożytniczych będzie mowa dalej w tym wykładzie), ale niektóre z nich mogą być wykorzystane jako aktywne elementy w układzie. Istnieją także technologie BiCMOS, w których na równi można się posługiwać tranzystorami MOS i bipolarnymi (nie mówimy o nich w tym wykładzie). Jak zobaczymy, w układach analogowych w wielu przypadkach tranzystory bipolarne są korzystniejsze od tranzystorów MOS. Warto więc przypomnieć sobie także ich charakterystyki i parametry.

W typowej strukturze układu CMOS można wyróżnić struktury bipolarne n-p-n i p-n-p pokazane na rys. 4.4.

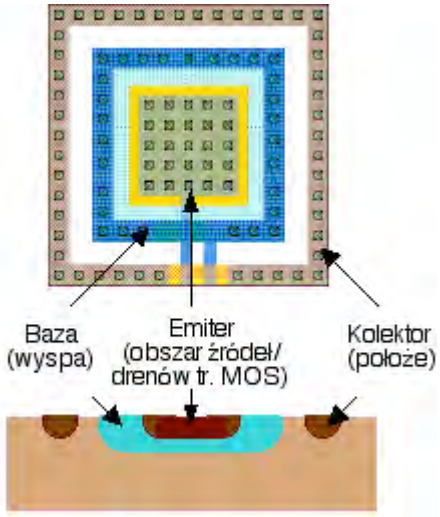



Rys. 4.4. Struktury bipolarne występujące w układach CMOS

Widzimy tu lateralny tranzystor n-p-n równoległy do tranzystora nMOS, lateralny tranzystor p-n-p równoległy do tranzystora pMOS, tranzystor p-n-p utworzony przez obszary źródła/drenu tranzystora pMOS oraz obszary wyspy i podłoża i wreszcie strukturę n-p-n znajdującą się między obszarem źródła/drenu tranzystora nMOS i wyspy typu n. W normalnych warunkach pracy układów CMOS wszystkie te struktury tranzystorowe mają złącza p-n spolaryzowane zaporowo, są więc nieaktywne.

Zwrócimy uwagę najpierw na strukturę n-p-n znajdującą się między obszarem źródła/drenu tranzystora nMOS i wyspą typu n. Struktura ta, zaznaczona na rys. 4.4 jako bezużyteczna, może w pewnych sytuacjach poważnie zakłócić działanie układu (niezależnie od tego, czy jest to układ cyfrowy, czy analogowy). Zauważmy, że wraz z obszarem źródł a/drenu tranzystora pMOS tworzy ona układ czterech obszarów: n(źródło/dren nMOS)-p(podłoże)-n(wyspa)-p(źródło/dren pMOS). Taka struktura czterowarstwowa jest znana pod nazwą tyrystora. Struktura tyrystorowa z wszystkimi trzema złączami p-n spolaryzowanymi zaporowo znajduje się w stanie, który nazywamy stanem blokowania i nie przewodzi prądu. Ma ona jednak tę własność, że jeśli jedno ze skrajnych złącz p-n (lub oba) przejdzie w stan przewodzenia, to może nastąpić przejście całej struktury w stan przewodzenia (mechanizmu fizycznego powodującego to zjawisko nie będziemy tu omawiać). Obszary skrajne (źródło/dren tranzystora nMOS i źródło/dren tranzystora pMOS) zostają wówczas praktycznie zwarte, i układ przestaje działać prawidłowo. Aby powrócił do normalnego stanu, trzeba wyłączyć i ponownie włączyć napięcie zasilania. To zjawisko nosi nazwę "zatrzaskiwania się" układów CMOS ("latch-up" w jęz. angielskim). Może ono wystąpić w przypadku, gdy jedno ze skrajnych złącz struktury tyrystorowej zostanie spolaryzowane dostatecznie silnie w kierunku przewodzenia. W prawidłowo skonstruowanych układach CMOS w stanie ustalonym taka sytuacja nigdy nie zachodzi, ale może się zdarzyć w stanie przejściowym, głównie wtedy, gdy przez podłoże lub wyspę przepływają znaczące prądy, i wobec tego obszary te nie są ekwipotencjalne. Dlatego ważne jest dołączanie do wyspy, a także do podłoża, kontaktów ustalających potencjały tych obszarów (zero dla podłoża, $U_{DD}$ dla wyspy). Producenci układów określają reguły polaryzacji podłoża i wyspy gwarantujące uniknięcie omawianego zjawiska.

Dla nas jednak bardziej interesująca z punktu widzenia zastosowań w układach analogowych jest struktura zwana podłożowym tranzystorem p-n-p, jaka istnieje pomiędzy obszarami źródł a/drenu tranzystora pMOS oraz obszarami wyspy i podłoża. Ten układ obszarów p-n-p bywa wykorzystywany jako aktywny tranzystor bipolany. Emiterem jest obszar implantacji typu p w wyspie, bazą obszar wyspy, a kolektorem - podłoże. Aby był to użyteczny tranzystor bipolarny, trzeba tym obszarom nadać odpowiednie kształty i wymiary. Przykład budowy takiego tranzystora - widok z góry i odpowiadający mu przekrój pokazuje rys. 4.5.




Rys. 4.5. Bipolarny tranzystor podłożowy p-n-p: widok z góry i przekrój. Przekrój w uproszczeniu: nie pokazano obszarów grubego tlenku oraz kontaktów i metalizacji

Charakterystyki i parametry tranzystora bipolarnego omówimy w uproszczeniu, tylko w takim zakresie, jaki będzie potrzebny w dalszych wykładach. Najprostszym opisem charakterystyk prądowo-napięciowych tranzystora bipolarnego jest model Ebersa-Molla. W tym modelu ogólna zależność prądu kolektora od napięć emiter-baza $U_{BE}$ i kolektor-baza $U_{CB}$ dana jest wzorem

$$I_C = I_{ES0}\left[\exp\left(\frac{qU_{BE}}{kT}\right)-1\right] - I_{CS0}\left[\exp\left(\frac{qU_{CB}}{kT}\right)-1\right] \tag{4.19}$$

W tym i następnych wzorach będziemy przyjmować następującą konwencję: napięcia polaryzujące złącza mają znak dodatni przy polaryzacji w kierunku przewodzenia, i ujemny przy polaryzacji w kierunku zaporowym.

W układach analogowych tranzystory bipolarne pracują prawie zawsze w zakresie napięć zwanym polaryzacją normalną: złącze kolektor-baza jest polaryzowane zaporowo, a złącze emiter-baza w kierunku przewodzenia. Dla takich warunków polaryzacji model Ebersa-Molla można poważnie uprościć. Dla $U_{BE}>>kT/q$ (ten warunek jest zawsze spełniony w typowych warunkach polaryzacji krzemowego tranzystora bipolarnego) oraz dla $U_{CB}<=0$ pierwszy składnik we wzorze (4.19) ma wartość o wiele rzędów wielkości większą od drugiego, i równocześnie $\exp(qU_{BE}/kT)>>1$, co pozwala sprowadzić wzór (4.19) do prostej i bardzo użytecznej postaci

$$I_C = I_{ES0}\exp\left(\frac{qU_{BE}}{kT}\right) \tag{4.20}$$

Ta zależność jest podstawą wielu rozwiązań układowych w analogowych układach bipolarnych. Wzór ten opisuje rzeczywistą charakterystykę tranzystora z dużą dokładnością w szerokim zakresie prądów kolektora (kilka dekad). Charakterystyka ta jest przewidywalna i powtarzalna. Odstępstwa obserwowane są dopiero w zakresie dużych gęstości prądu kolektora. W tym zakresie prąd kolektora rośnie z napięciem $U_{BE}$ wolniej, niż przewiduje wzór (4.20).

Współczynnik $I_{ES0}$ jest wprost proporcjonalny do powierzchni złącza emiter-baza

$$I_{ES0} = J_{ES0}A_E \tag{4.21}$$

zaś gęstość tego prądu, oznaczona $J_{ES0}$, zależy od elektrycznej grubości bazy tranzystora - jest do niej w przybliżeniu odwrotnie proporcjonalna. W tym ukryta jest zależność prądu kolektora od napięcia kolektor-baza $U_{CB}$, które nie występuje jawnie w zależności (4.20). Gdy napięcie $U_{CB}$ (polaryzujące złącze kolektorowe w kierunku zaporowym) wzrasta, elektryczna grubość bazy maleje (bo wzrasta głębokość wnikania warstwy zaporowej złącza kolektor-baza) i współczynnik $J_{ES0}$ wzrasta.

Współczynnik $J_{ES0}$ jest też silnie zależny od temperatury, o czym będzie mowa dalej.

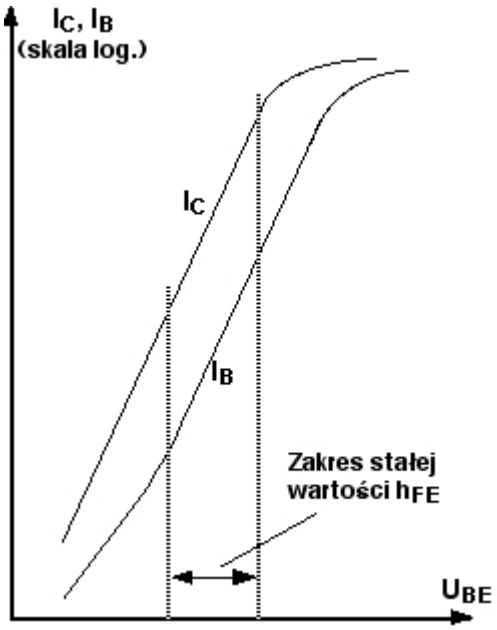
Wyznaczając z (4.20) napięcie $U_{BE}$ otrzymujemy inną często wykorzystywaną zależność

$$U_{BE} = \frac{kT}{q}\ln\left(\frac{I_C}{I_{ES0}}\right) = \frac{kT}{q}\ln\left(\frac{I_C}{J_{ES0}A_E}\right) \tag{4.22}$$

Do opisu działania tranzystora bipolarnego potrzebna jest jeszcze zależność określająca prąd bazy. Nie będzie nam tu potrzebna pełna zależność wynikająca z modelu Ebersa-Molla, wystarczy powszechnie stosowane uproszczenie definiujące tak zwany stałoprądowy współczynnik wzmocnienia prądowego tranzystora w układzie wspólnego emitera. Jest on z przyczyn historycznych często oznaczany symbolem $h_{FE}$. Współczynnik ten jest to stosunek składowych stałych prądu kolektora i prądu bazy:

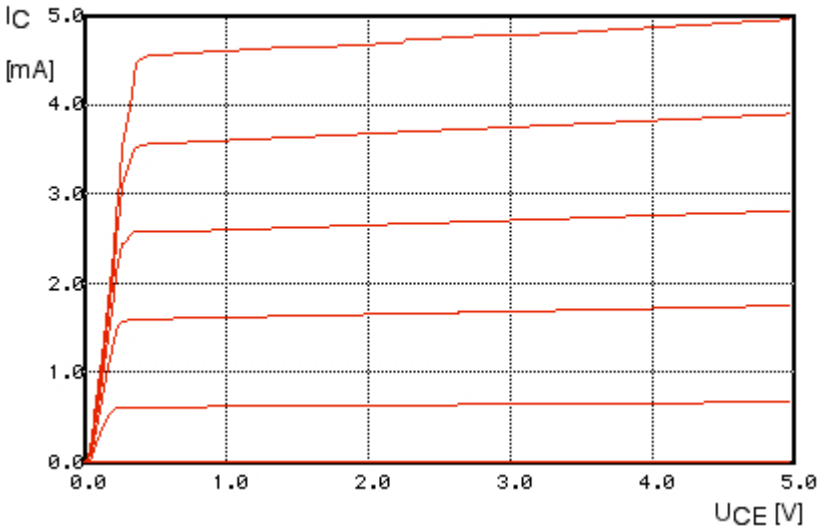
$$h_{FE} = \frac{I_C}{I_B} \tag{4.23}$$

Tak zdefiniowany współczynnik jest użyteczny tylko w zakresie polaryzacji normalnej. Przy polaryzacji normalnej w dość szerokim zakresie prądów (kilka dekad) dla większości tranzystorów bipolarnych obserwuje się, że prądy kolektora i bazy są wprost proporcjonalne, a ich iloraz, czyli $h_{FE}$, ma wartość praktyczne stałą. Odstępstwa są obserwowane w zakresie prądów bardzo dużych i bardzo małych. Najlepiej można to zilustrować na wykresie w skali półlogarytmicznej




Rys. 4.6. Zależności prądu kolektora i prądu bazy od napięcia $U_{BE}$ w tranzystorze bipolarnym, w skali półlogarytmicznej

Typową zależność prądu kolektora od napięcia  dla różnych wartości prądu bazy ilustruje rys. 4.7.



Rys. 4.7. Przykładowa rodzina charakterystyk wyjściowych $I_C(U_{CE})$ dla różnych wartości prądu bazy tranzystora bipolarnego

**Parametry małosygnałowe**

Parametry małosygnałowe tranzystora bipolarnego zdefiniujemy dla pracy tranzystora w układzie wspólnego emitera. Parametrami, które będą nam potrzebne, są: transkonduktancja $g_m$

$$g_m = \frac{\partial I_C}{\partial U_{BE}} = \frac{qI_C}{kT} \tag{4.24}$$

małosygnałowy współczynnik wzmocnienia prądowego $\beta$

$$\beta = \frac{\partial I_C}{\partial I_B} \tag{4.25}$$

który w zakresie prądów kolektora, w którym współczynnik $h_{FE}$ ma wartość niezależną od prądu, jest w przybliżeniu równy temu współczynnikowi

$$\beta \approx h_{FE} \tag{4.26}$$

konduktancja wejściowa $g_{be}$

$$g_{be} = \frac{\partial I_B}{\partial U_{BE}} = \frac{g_m}{\beta} \tag{4.27}$$

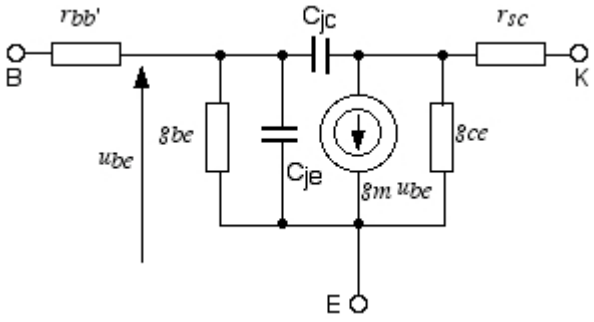
oraz konduktancja wyjściowa $g_{bc}$

$$g_{ce} = \frac{\partial I_C}{\partial U_{CE}} \tag{4.28}$$

która jest miarą nachylenia charakterystyki $I_C$=f($U_{BE}$). Wprowadzając dodatkowy parametr $V_A$ zwany napięciem Early'ego można zapisać (4.28) w postaci

$$g_{ce} = \frac{I_C}{V_A} \tag{4.29}$$

Podobnie jak dla tranzystora MOS, także dla tranzystora bipolarnego wymienione wyżej parametry małosygnałowe opisują właściwości tranzystora dla sygnałów o małej amplitudzie i małej częstotliwości. Opis pracy tranzystora przy wielkich częstotliwościach wymaga użycia wzorów zawierających czynniki zależne od częstotliwości oraz uwzględnienia pojemności. Omawianie układów wielkiej częstotliwości i mikrofalowych wykracza poza zakres tego wykładu, jednak dla kompletności wykładu przytoczymy tu małosygnałowy schemat zastępczy tranzystora




Rys. 4.8. Małosygnałowy schemat zastępczy tranzystora bipolarnego.
Uwzględniono w nim rezystancje rozproszone bazy i kolektora oraz pojemności złączowe

Producenci układów podają wartości parametrów modeli tranzystorów bipolarnych dla swoich technologii. Zwykle parametry te odnoszą się do tranzystora o ściśle określonej budowie i wymiarach, np. dla tranzystora p-n-p podłożowego takiego, jak na rys. 4.5.

**Porównanie z tranzystorami MOS**

Na koniec warto porównać tranzystor bipolarny z unipolarnym. Porównamy wartość transkonduktancji - parametru, od którego w stopniach wzmacniających zależy wartość wzmocnienia napięciowego. Stosunek tych

transkonduktancji wynosi (patrz wzory (4.17) i (4.24)

$$\frac{g_{mBIP}}{g_{mMOS}} = \frac{\frac{qI_C}{kT}}{\frac{2I_D}{U_{GS}-U_T}} = \frac{I_C}{I_D}\frac{U_{GS}-U_T}{2\frac{kT}{q}} \tag{4.30}$$

Dla jednakowych wartości prądów - drenu w tranzystorze MOS i kolektora w tranzystorze bipolarnym - transkonduktancja tranzystora bipolarnego jest kilkadziesiąt razy wyższa, bowiem $kT/q$ jest to napięcie mające wartość około 26 mV (przy temperaturze otoczenia), podczas gdy różnica napięć $U_{GS}$-$U_T$ w typowych warunkach pracy tranzystora MOS wynosi kilka woltów. Dużo wyższa transkonduktancja czyni tranzystor bipolarny elementem korzystniejszym w układach analogowych. Są też dalsze cechy tego elementu dające mu przewagę. W tablicy poniżej pokazane są główne różnice między tranzystorami MOS i bipolarnymi.

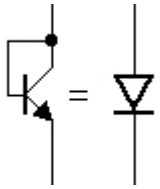
| Właściwość | Tranzystor MOS | Tranzystor bipolarny |
|---|---|---|
| Transkonduktancja | mała | duża |
| Max. częstotliwość pracy | do kilku GHz | do kilkudziesięciu GHz |
| Konduktancja wejściowa | praktycznie równa zeru | duża |
| Poziom szumów | | mniejszy niż w tranzystorze MOS |
| Rozrzuty produkcyjne | | mniejsze niż w tranzystorze MOS |

Tranzystory bipolarne ukazują swoją przewagę głównie w układach analogowych. Przez wiele lat układy analogowe były wykonywane wyłącznie przy użyciu tranzystorów bipolarnych. Dziś popularność technologii CMOS spowodowała, że układy analogowe CMOS są stosowane na równi z układami bipolarnymi.

## 4.3. Elementy bierne: diody, rezystory, pojemności, indukcyjności

Układy cyfrowe CMOS zawierają wyłącznie tranzystory MOS (wyjątkiem są pamięci dynamiczne DRAM, będzie o nich mowa w dalszych wykładach), natomiast w układach analogowych, zarówno MOS, jak i bipolarnych, same tranzystory zwykle nie wystarczają. Używa się również elementów biernych: diod, rezystorów, kondensatorów, a w układach mikrofalowych także indukcyjności. Ponieważ zajmujemy się głównie układami CMOS, omówione będą sposoby wykonywania elementów biernych typowe dla technologii CMOS.

Jako diody wykorzystywane są zwykle struktury tranzystorów bipolarnych w połączeniu zwanym diodowym - kolektor zwarty z bazą. W układach CMOS mogą to być na przykład tranzystory takie, jak pokazany na rys. 4.5.



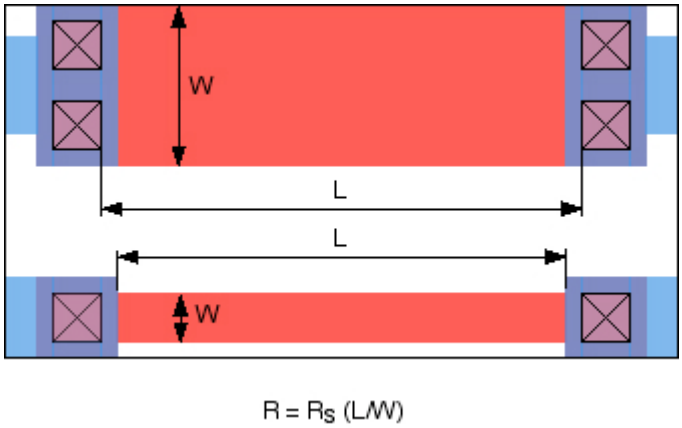
Rys.4.9. Tranzystor bipolarny w połączeniu diodowym

Prąd takiej diody jest sumą prądu kolektora i prądu bazy tranzystora, z tym że prąd bazy jest $h_{FE}$ - razy mniejszy, a ponieważ $h_{FE}$ ma zwykle wartość rzędu 50 - 200, prąd bazy jest w pierwszym przybliżeniu do pominięcia. Zatem charakterystyka prądowo-napięciowa takiej diody wynika bezpośrednio z charakterystyki $I_C$=f($U_{BE}$) tranzystora, jest więc opisana wzorem 4.20, który w odniesieniu do diody otrzymuje postać:

$$I = I_{ES0} \exp\left(\frac{qU}{kT}\right) \tag{4.31}$$

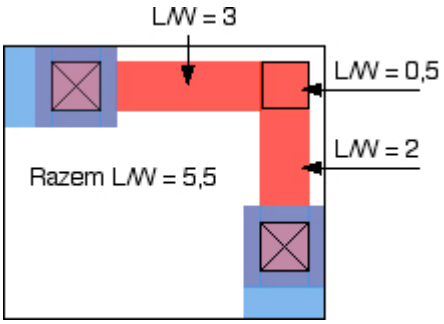
Rezystory w układach CMOS są wykonywane jako ścieżki polikrzemowe. Możliwy do uzyskania zakres rezystancji jest ograniczony. Typowa rezystancja warstwowa polikrzemu jest rzędu 10 - 30 Ω/□ (rezystancja warstwowa - definicja w dodatku 2). Gdyby ze ścieżki polikrzemowej o rezystancji warstwowej 20 Ω wykonać rezystor 20 kΩ, to miałby on stosunek długości do szerokości $L/W$ równy 1000. Przy szerokości ścieżki równej 1 μm długość wynosiłaby 1000 μm, a powierzchnia 1000 $\mu m^2$. Porównajmy to z powierzchnią zajmowaną przez tranzystor MOS - jest ona rzędu kilkudziesięciu $\mu m^2$. Zatem jeden rezystor o stosunkowo dużej rezystancji zająłby tyle miejsca, co kilkadziesiąt tranzystorów. Jest to więc element kosztowny (przypomnijmy sobie, że koszt układu jest proporcjonalny do jego powierzchni - wykład 2). W niektórych technologiach CMOS przeznaczonych specjalnie do układów analogowych wykonywane są dwie warstwy polikrzemu. Pierwsza warstwa służy do wykonania bramek tranzystorów i jest taka sama, jak w technologiach układów cyfrowych. Druga warstwa jest słabiej domieszkowana i może mieć rezystancję warstwową rzędu kilku kΩ. Służy ona do wykonywania rezystorów. Można wówczas wykonać rezystory o dużej rezystancji przy umiarkowanej długości i powierzchni. Jednak w wielu przypadkach bardziej ekonomiczne jest użycie jako dużej rezystancji odpowiednio ukształtowanego ($W/L$ < 1) i spolaryzowanego tranzystora MOS.

Projektowanie rezystora o prostym kształcie nie jest trudne. Rys. 4.10 pokazuje typowe kształty i przybliżoną wartość rezystancji.




Rys. 4.10. Rezystory polikrzemowe. Czerwony polikrzem, niebieski metal 1

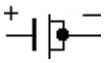
Jeśli ścieżka rezystora zawiera zagięcia pod kątem prostym, to każde takie zagięcie traktuje się jako fragment ścieżki o stosunku $L/W$ równym 0,5 - patrz rys. 4.11.




Rys. 4.11. Rezystor z zagięciem

Projektując rezystory, a zwłaszcza rezystory o małych rezystancjach, nie wolno zapominać o rezystancjach kontaktów, które mogą wynosić nawet kilkadziesiąt omów. Wartość rezystancji kontaktów podaje producent układów.

W układach analogowych stosuje się też niekiedy kondensatory. Jako kondensator może być wykorzystana pojemność bramka-kanał tranzystora MOS. Źródło z drenem zwiera się, a na bramce musi panować napięcie znacznie powyżej progowego, tak by kanał istniał i miał możliwie wysoką przewodność. Korzystniejszy jest tranzystor nMOS ze względu na większą przewodność kanału typu n.




Rys. 4.12. Tranzystor nMOS jako kondensator

Użycie struktury tranzystorowej jest uzasadnione tym, że tlenek bramkowy jest w układzie dielektrykiem o najmniejszej grubości, a więc i największej pojemności jednostkowej. Kondesator taki będzie więc zajmował najmniejszą powierzchnię.

Jako kondensatory mogą być też wykorzystane pojemności złączowe złącz p-n.

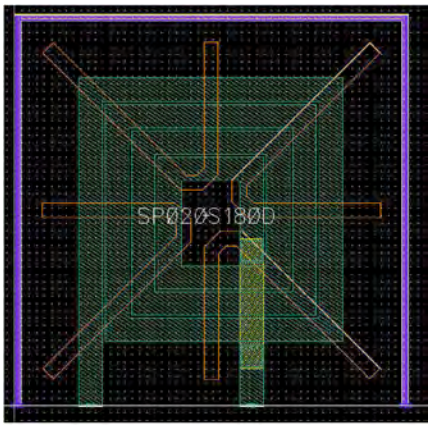
Istnieją technologie CMOS, w których do wykonywania kondensatorów przeznaczone są dwie warstwy polikrzemu (obie silnie domieszkowane, o niskiej rezystywności). Druga warstwa polikrzemu leży na bardzo cienkiej warstwie dielektrycznej $SiO_2$ (grubość podobna, jak dla tlenku bramkowego). To rozwiązanie daje możliwość wykonania kondensatorów o najlepszych właściwościach. Jednak dodatkowe warstwy podnoszą koszt układu.

Projektowanie kondensatorów sprowadza się do określenia powierzchni okładek $A$ potrzebnej dla uzyskania wymaganej pojemności, zgodnie ze wzorem

$$C = \frac{\varepsilon_{ox}}{t_{ox}} A \tag{4.32}$$

w którym $t_{ox}$ jest grubością, a $\varepsilon_{ox}$ przenikalnością dielektryczną dielektryka. Dane te podaje producent układów.

Indukcyjności do niedawna były uważane za elementy niemożliwe do wykonania w układach scalonych. Ten stan rzeczy zaczął ulegać zmianom, gdy częstotliwości pracy układów CMOS sięgnęły kilku GHz. Przy tych częstotliwościach potrzebne są indukcyjności bardzo małe, o wartościach rzędu kilku nH. Można je wykonać jako płaskie spirale w warstwach metalu. Dobroć takich indukcyjności jest niewielka. Wynika to z sąsiedztwa przewodzącego podłoża krzemowego. Cewka indukuje w nim prądy wirowe, które są przyczyną strat energii pola magnetycznego cewki. Można temu w pewnym stopniu zapobiegać wprowadzając w podłożu pod cewką obszary utrudniające przepływ prądów wirowych. Jednym ze sposobów jest wprowadzenie do podłoża, do wąskich prostokątnych obszarów skierowanych prostopadle do kierunku prądów wirowych, domieszki przeciwnego typu niż podłoże. Tworzą się w ten sposób złącza pn skutecznie przecinające drogę prądów wirowych. Przykładową topografię takiej cewki pokazuje rys. 4.13.

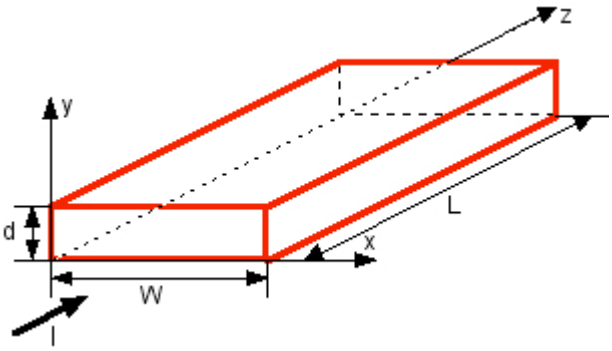



Rys. 4.13. Topografia cewki w układzie scalonym CMOS

Szczegóły projektowania cewek nie będą omawiane, bowiem tematyka układów pracujących na częstotliwościach mikrofalowych wykracza poza zakres przedmiotu "Układy scalone". W praktyce służą do tego specjalne programy generujące topografie cewek o zadanej indukcyjności. Producenci układów dostarczają też biblioteczne elementy indukcyjne o znanych, zmierzonych parametrach.

## 4.3. Dodatek 2: Definicja rezystancji warstwowej

Rezystancja warstwowa (zwana także rezystancją powierzchniową lub gwarowo rezystancją "na kwadrat") jest pojęciem wygodnym do charakteryzowania rezystancji obszarów, które są niejednorodne w kierunku prostopadłym do kierunku przepływu prądu, na przykład z powodu nierównomiernego rozkładu domieszek. Taki obszar pokazany jest na rysunku poniżej. Może to być na przykład ścieżka polikrzemu, w którym koncentracja domieszki maleje w kierunku od powierzchni w głąb.




Ilustracja do definicji rezystancji warstwowej: prostopadłościan o wymiarach $W * L * d$ ma rezystywność zmieniającą się wzdłuż osi y, prąd $I$ płynie równolegle do osi z.

Rozważmy przepływ prądu przez prostopadłościan pokazany na rysunku. Jego rezystywność zmienia się wzdłuż osi y, zaś prąd płynie w kierunku osi z. Konduktancję tego prostopadłościanu dla prądu $I$ można obliczyć całkując konduktywność $\sigma$ w granicach od 0 do $d$. Konduktancja warstwy o nieskończenie małej grubości $dy$ jest równa

$$G(y) = \frac{W}{L}\sigma(y)dy$$

zatem rezystancja prostopadłościanu dla prądu $I$ wynosi

$$R = \frac{1}{G} = \frac{1}{\frac{W}{L}\int_0^d \sigma(y)dy}$$

co można zapisać w postaci

$$R = R_s \frac{L}{W}$$

gdzie $R_s$ jest rezystancją warstwową:

$$R_s = \frac{1}{\int_0^d \sigma(y)dy}$$

Jest ona potocznie nazywana rezystancją "na kwadrat", ponieważ jest to rezystancja obszaru o długości $L$ równej szerokości $W$, czyli - patrząc z góry - kwadratu. Mianem rezystancji warstwowej jest om, ale dla zaznaczenia charakteru tej wielkości używa się często oznaczenia $\Omega/\square$ ("om na kwadrat").

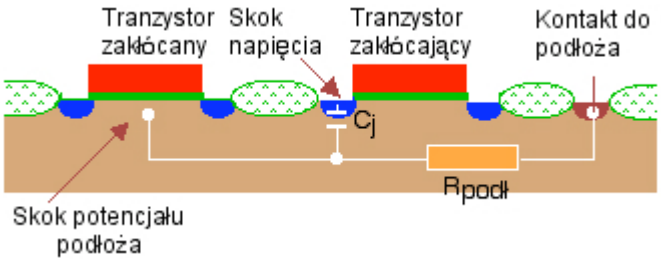
.

## 4.4. Elementy pasożytnicze: pojęcie i przykłady

Działanie układu scalonego i jego parametry zależą nie tylko od właściwości elementów czynnych i biernych, z jakich zbudował układ jego projektant, ale także od nieuchronnie występujących w układzie efektów zwanych pasożytniczymi. Każdy obszar półprzewodnikowy i przewodzący ma pewną rezystancję, na której powstaje w przypadku przepływu prądu spadek napięcia. Pomiędzy obszarami przewodzącymi rozdzielonymi dielektrykiem występują pojemności, które wprowadzają sprzężenia dla sygnałów zmiennych między węzłami elektrycznymi układu. Tego rodzaju oddziaływania uwzględnia się mówiąc, że w układzie scalonym występują elementy pasożytnicze - rezystory, kondensatory - i w miarę możliwości uwzględniając je w schematach układów. Oprócz biernych elementów pasożytniczych występują też elementy czynne, na przykład struktury tranzystorów bipolarnych nieuchronnie związane z tranzystorami MOS (jak na rys. 4.4).

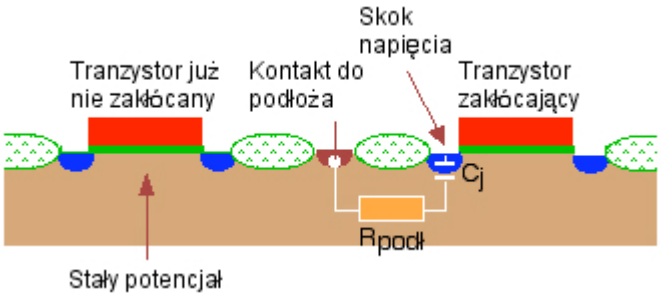
Dobrze zaprojektowane układy cyfrowe są stosunkowo mało wrażliwe na obecność elementów pasożytniczych. Wyjątkiem są rezystancje i pojemności długich połączeń, które mogą znacznie ograniczyć szybkość działania dużego układu cyfrowego. Zagadnienie to będzie omawiane w jednym z dalszych wykładów. Układy analogowe są znacznie bardziej wrażliwe na efekty pasożytnicze.

Wpływ takich elementów pasożytniczych, jak na przykład rezystancja ścieżki lub jej pojemność do podłoża lub do innej ścieżki, jest stosunkowo łatwy do uwzględnienia przez wprowadzenie tego elementu do schematu układu i wykonanie odpowiednich obliczeń lub symulacji. Istnieją jednak także oddziaływania pasożytnicze, których uwzględnienie jest znacznie trudniejsze. Należą do tej grupy sprzężenia przez podłoże - specyficzny mechanizm zakłócający działanie układów scalonych, który wynika stąd, że podłoże jest wspólne dla wielu elementów i jest obszarem przewodzącym o dość znacznej rezystywności.




Rys. 4.14. Mechanizm sprzężenia przez podłoże.
Rezystor symbolizuje rezystancję rozproszoną między drenem tranzystora zakłócającego, a kontaktem uziemiającym podłoże.

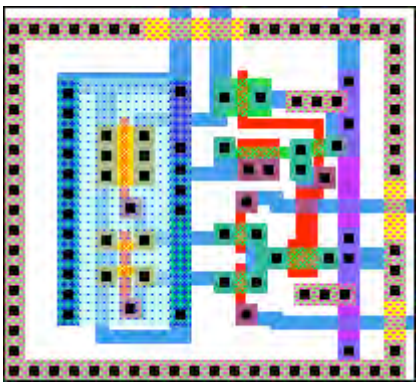
Mechanizm sprzężenia przez podłoże można opisać następująco (rys. 4.14). Jeśli na obszarze drenu tranzystora zakłócającego pojawi się skok napięcia, to zmieni się ładunek zgromadzony w pojemności złącza dren-podłoże. Spowoduje to przepływ impulsu prądu przez podłoże, a ponieważ podłoże jest obszarem o dość znacznej rezystywności, wystąpić musi spadek napięcia. Potencjał podłoża w okolicy tranzystora zakłócanego zmieni się, a tym samym zmieni się napięcie $U_{BS}$ tego tranzystora. To, jak wiemy, powoduje zmianę napięcia progowego tranzystora (patrz wzór (4.5)). Zmiana napięcia progowego wywoła zmianę wartości prądu drenu. W ten sposób zakłócenia przenoszą się przez podłoże (podłoże w tym przypadku oznacza także obszar wyspy, w którym może wystąpić to samo zjawisko). Zjawisko to jest bardzo trudne do uwzględnienia i ilościowej analizy ze względu na trójwymiarowy rozpływ prądu w podłożu uzależniony od rozmieszczenia elementów układu. Zjawisko to nie daje się modelować prostym schematem zastępczym z elementami o stałych skupionych. Wiadomo natomiast, jak należy projektować układ, by temu zjawisku zapobiec lub znacznie je ograniczyć. Pokazuje to rys. 4.15.




Rys. 4.15. Sposób zredukowania sprzężenia przez podłoże przez uziemienie podłoża pomiędzy tranzystorami

Efekt redukcji sprzężenia uzyskujemy dzięki uziemieniu podłoża pomiędzy tranzystorami. Dzięki temu zmiana potencjału podłoża w pobliżu tranzystora, który w poprzedniej konfiguracji był "odbiornikiem" zakłaceń, teraz nie zachodzi. Potencjał ten jest określony przez potencjał kontaktu do podłoża.

W układach cyfrowych, które są mało wrażliwe na zakłócenia, wystarcza uziemianie podłoża przez możliwie jak najgęściej rozmieszczone kontakty. Na ogół wystarcza umieszczanie kontaktów do podłoża nie rzadziej, niż co 50 ... 100 µm. W przypadku układów analogowych może być celowe otoczenie tranzystorów zakłócających lub całych bloków zakłócających pierścieniami kontaktów zwanymi pierścieniami ochronnymi (rys. 4.16). W przypadku podłoża są to kontakty uziemiające, a w przypadku obszaru wyspy - kontakty połączone z plusem zasilania $U_{DD}$. Jeżeli w tym samym układzie występują zarówno bloki cyfrowe, jak i analogowe, otoczenie bloków cyfrowych pierścieniami ochronnymi jest z reguły niezbędne. W układach cyfrowych CMOS występują bowiem skoki napięcia od zera do napięcia zasilania $U_{DD}$ , toteż układy te są źródłem zakłoceń o dużej amplitudzie.



Rys. 4.16. Mały blok otoczony pierścieniem kontaktów uziemiających podłoże

Specyficznym dla układów scalonych rodzajem sprzężeń pasożytniczych są sprzężenia elektryczno-cieplne. Charakterystyki i parametry elementów półprzewodnikowych dość silnie zależą od temperatury. W układach, w których w czasie pracy wydziela się znaczna moc, temperatura elementów obciążonych dużą mocą wzrasta, niekiedy bardzo znacznie. Pozostałe elementy układu są więc podgrzewane, ich parametry ulegają zmianom. W ten sposób powstaje sprzężenie między elementami grzejącymi, a podgrzewanymi. Ten rodzaj sprzężeń jest także bardzo trudny do analizy i uwzględnienia przy projektowaniu układów, ponieważ wymaga równoczesnej analizy elektrycznej oraz cieplnej, a ta ostatnia wymaga symulacji trójwymiarowego rozpływu ciepła w płytce układu scalonego z uwzględnieniem odpływu do otoczenia (obudowy układu i ewentualnie radiatora, na którym układ jest umieszczony). Do tych zagadnień powrócimy w jednym z dalszych wykładów.

# ZADANIA DO WYKŁADU 4

**Zadanie 1**

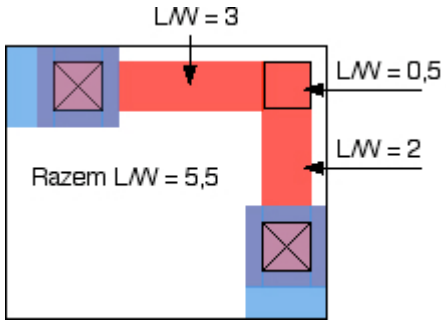
Dana jest technologia CMOS, w której $U_{Tn}$ = 0,75 V, $U_{Tp}$ = - 0,85 V, $\mu_n C_{ox}$ = 80 μA/V$^2$, $\mu_p C_{ox}$ = 27 μA/V$^2$. Oblicz wartość transkonduktancji oraz prądu drenu tranzystorów nMOS i pMOS dla $U_{GS}$=5 V i $U_{DS}$=5V, dla dwóch przypadków:
(a) L=0,7 μm, W = 1 μm.
(b) L=3 μm, W = 12 μm.

**Zadanie 2**

Oblicz wartość transkonduktancji tranzystorów bipolarnych, dla których prąd kolektora ma tę samą wartość, co prąd drenu tranzystorów nMOS i pMOS dla przypadków (a) i (b) z zadania 1.

**Zadanie 3**

Oblicz rezystancję rezystora pokazanego na rys. 4.11 (powtórzonego niżej), jeśli wiadomo, że rezystancja warstwowa polikrzemu wynosi 25 Ω/□, a rezystancja kontaktu jest równa 15 Ω .




**Zadanie 4**

Oblicz, jaką co najmniej powierzchnię musiałby zająć rezystor o rezystancji 10 kΩ wykonany z polikrzemu o danych jak w zadaniu 3, jeśli minimalna szerokość ścieżki rezystora wynosi 1 μm.

**Zadanie 5**

Oblicz, jaką powierzchnię musiałaby mieć bramka tranzystora MOS, aby mogła pełnić rolę kondensatora o pojemności 1 pF, jeśli grubość tlenku bramkowego wynosi 6 nm. Względna przenikalność dielektryczna $SiO_2$ wynosi 3,82.

## Bibliografia

[1] W. Marciniak, "*Przyrządy półprzewodnikowe i układy scalone*", WNT Warszawa 1987
[2] W. Marciniak "*Przyrządy półprzewodnikowe MOS*". WNT Warszawa 1991

# Wykład 5: Projektowanie układów scalonych

## Wstęp

Wykład 5 przedstawia specyfikę projektowania układów elektronicznych przeznaczonych do realizacji w postaci układów scalonych. Opowiada o przebiegu procesu projektowania. Jest w nim mowa o najpoważniejszych problemach, jakie występują przy projektowaniu, oraz odpowiedź na pytanie: jak sobie z tymi problemami radzimy. Zawiera szereg praktycznych wskazówek. Przedstawia narzędzia wspomagania komputerowego, jakich używa projektant. Przygotowuje też do następnego wykładu, w którym omawiane będą metody projektowania uproszczonego i zautomatyzowanego.

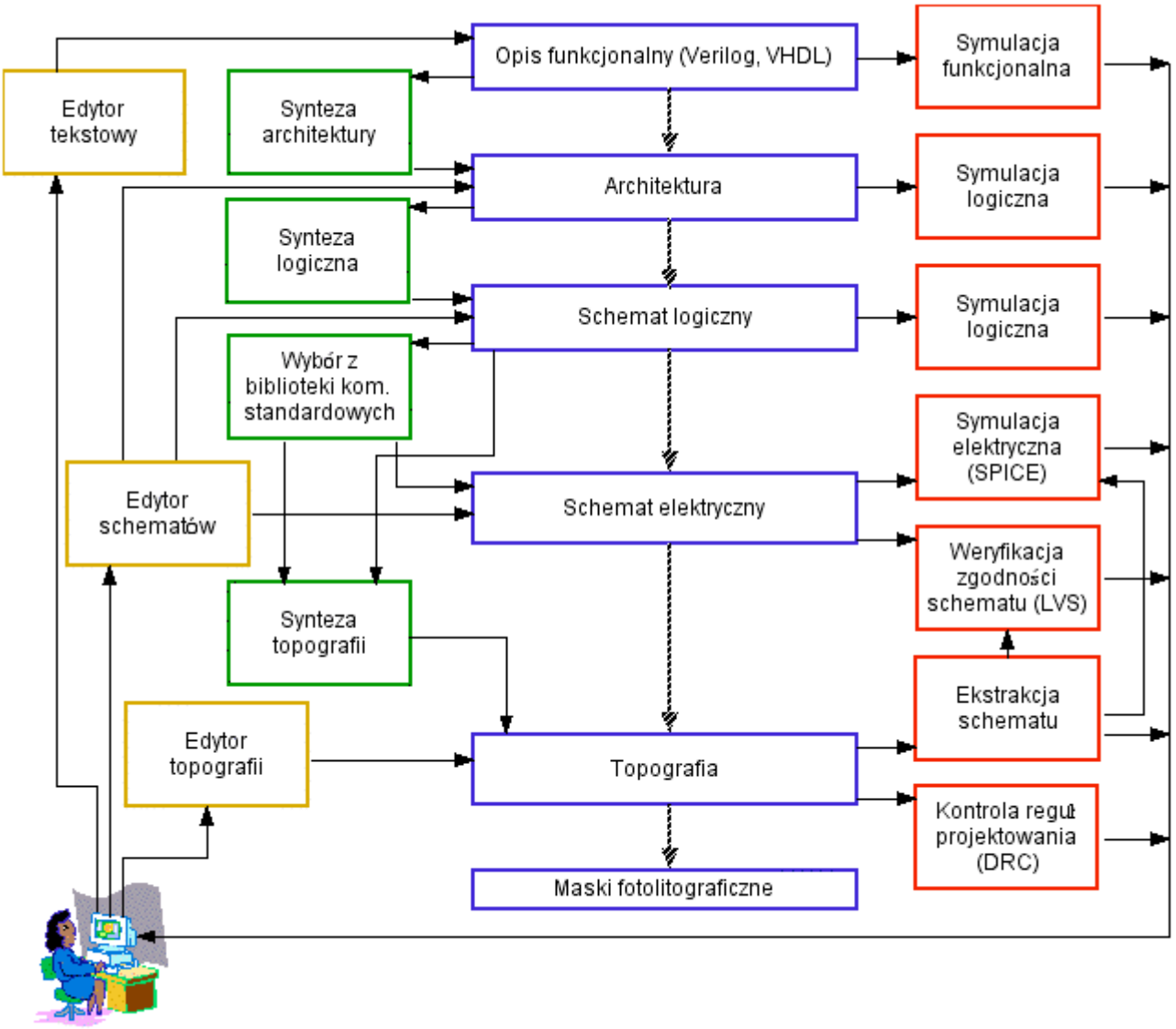
## 5.1. Proces projektowania układu scalonego

Projektowanie układu scalonego jest procesem składającym się z wielu etapów. Dla układu cyfrowego wyglądają one następująco:

- określenie **funkcji układu**,
- projekt **architektury układu** (czyli schematu złożonego z bloków funkcjonalnych),
- projekt **logiczny** (czyli schematu złożonego z bramek logicznych),
- projekt **elektryczny** (czyli schematu złożonego z tranzystorów),
- projekt struktury fizycznej układu (**topografii układu**, czyli projektu elementów, ich rozmieszczenia oraz połączeń między nimi).

W przypadku układów analogowych nie ma oczywiście etapu tworzenia projektu logicznego.

Na każdym etapie projektowania dostępne są programy komputerowe wspomagające proces projektowania oraz programy służące do weryfikacji projektu na danym etapie. Wszystkie te programy tworzą zintegrowany system i są powiązane ze sobą poprzez wspólne formaty danych. Istotną częścią składową systemu projektowania są specjalizowane programy do wprowadzania danych: edytory graficzne do graficznego definiowania schematów elektrycznych i logicznych oraz do "rysowania" topografii układu. Potrzebny jest też edytor tekstów do wprowadzania opisów w językach opisu sprzętu i ewentualnie innych informacji tekstowych. Oprogramowanie do projektowania układów scalonych będzie w szczegółach omawiane nieco dalej.




Rys. 5.1. Schemat typowego systemu projektowania i przebieg procesu projektowania.

Rysunek 5.1 przedstawia ogólny schemat procesu projektowania oraz oprogramowania tworzącego typowy system projektowania. W niebieskich ramkach opisy układu na kolejnych poziomach abstrakcji i etapach projektowania. W czerwonych ramkach programy komputerowe służące do weryfikacji projektu (będą one omawiane dalej). W zielonych ramkach programy wykonujące kolejne etapy automatycznej syntezy projektu (przy wykorzystaniu biblioteki komórek standardowych - ten sposób projektowania będzie omawiany dalej). W żółtych

ramkach programy komputerowe służące projektantowi do wprowadzania informacji (one także będą omawiane dalej). Cienkie linie pokazują kierunki przepływu informacji. Omówimy teraz krótko poszczególne etapy powstawania projektu.

**Definicja funkcji układu**
Na tym etapie określa się funkcję wykonywaną przez układ. Jak uczy wieloletnie doświadczenie, na tym, pozornie oczywistym, etapie projektowania popełniane są bardzo często błędy prowadzące do niepowodzenia całego procesu projektowania. Funkcja układu może być zdefiniowana w sposób niekompletny lub wewnętrznie sprzeczny. Zdarza się też, że definicja funkcji układu jest kompletna i spójna, ale nie oddaje dokładnie intencji projektanta lub użytkownika układu, który zamawia projekt i będzie potem wykorzystywać układ. Aby uniknąć takich sytuacji, opracowano specjalne języki, zwane **językami opisu sprzętu**. Miały one początkowo jedynie uporządkować i sformalizować proces definiowania funkcji układu, pomóc w sprawdzeniu jego kompletności i wewnętrznej spójności. Obecnie pełnią bardzo ważną rolę jako języki wejściowe w systemach zautomatyzowanego projektowania. Będzie o nich mowa w dalszej części wykładu.

**Projekt architektury układu**
Na tym etapie tworzy się projekt układu w postaci powiązanych ze sobą bloków funkcjonalnych o zdefiniowanych funkcjach, oraz definiuje się powiązania między blokami i przepływy danych między nimi. Projekt ten można wykonać "ręcznie" i wprowadzić w postaci graficznej lub w postaci opisu tekstowego w języku opisu sprzętu (ta druga możliwość nie jest pokazana na rys. 5.1.). Projekt architektury układu w przypadku układów dużych o złożonych funkcjach ma z reguły budowę hierarchiczną. Cały układ zbudowany jest z niewielkiej liczby dużych bloków funkcjonalnych, każdy z tych bloków ma z kolei swoją wewnętrzną architekturę złożoną z mniejszych bloków o prostszych funkcjach itp. Takich poziomów hierarchii może być wiele. Hierarchiczna budowa układu umożliwia zapanowanie nad całym projektem przez podzielenie go na niewielkie fragmenty, które można projektować mniej lub bardziej niezależnie. Bez takiej strategii projektowania wykonanie dużego i złożonego projektu byłoby praktycznie niemożliwe - projektowanie układu mającego tysiące, a tym bardziej miliony elementów jako jednolitej całości jest niewykonalne.

**Projekt schematu logicznego układu**
Projekt schematu logicznego polega na "wypełnieniu" bloków funkcjonalnych schematami zbudowanymi z bramek logicznych lub prostych typowych bloków. Przejście od architektury do schematu logicznego nie jest jednak jednoznaczne, każdą funkcję logiczną można zrealizować na wiele różnych sposobów. Tę fazę projektu można wykonać "ręcznie" i opracowany schemat wprowadzić w postaci graficznej lub w postaci tekstowego opisu w języku opisu sprzętu (ta druga możliwość nie jest pokazana na rys. 5.1.). Projekt schematu logicznego może być też wykonany automatycznie na podstawie projektu architektury układu.

**Projekt schematu elektrycznego układu**
Ten etap polega na określeniu schematow elektrycznych bramek i bloków występujących w schemacie logicznym. Jeżeli wykorzystuje się typowe, standardowe sposoby realizacji bramek kombinacyjnych, przerzutników itp., to ich schematy elektryczne są znane, a przejście od schematu logicznego do elektrycznego jest niemal automatyczne (i może być zautomatyzowane). W tym etapie określa się jednak także parametry elementów (w przypadku układu CMOS oznacza to określenie wymiarów kanałów tranzystorów MOS). Jeżeli układ ma spełniać wysokie wymagania techniczne (np. szybkość działania), to określenie najlepszych wymiarów tranzystorów jest niełatwe (i nie poddaje się łatwo automatyzacji). W większości przypadków korzysta się jednak z biblioteki standardowych bramek wcześniej zaprojektowanych, sprawdzonych i scharakteryzowanych dla danej technologii produkcji układów. Taką bibliotekę dostarcza producent układów.

Niekiedy wyśrubowane wymagania techniczne zmuszają do użycia niestandardowych rozwiązań bramek logicznych (np. układy zwane logiką dynamiczną; będzie o nich mowa w dalszych wykładach). Wtedy przekształcanie schematu logicznego w elektryczny jest twórczym, i często niełatwym, zadaniem dla projektanta.

W przypadku projektowania układu analogowego projekt schematu elektrycznego rozpoczyna cały proces projektowania. Nie ma, jak dotąd, ogólnych metod syntezy układów analogowych na podstawie ich opisów funkcjonalnych. Schemat układu analogowego musi być więc opracowany przez projektanta. Istnieją dziś rozszerzenia języków opisu sprzętu pozwalające opisywać nie tylko układy cyfrowe, ale i analogowe.

**Projekt topografii układu**
To jest etap, w którym powstaje projekt fizycznej struktury układu. Przy projektowaniu *"ręcznym"* wszystkie elementy są "rysowane" przy pomocy specjalizowanego edytora graficznego przez projektanta, który również określa ich położenie i "rysuje" schemat połączeń (jest to sposób projektowania topografii wspomniany w wykładzie 2 jako projektowanie w stylu *full custom*). Ten sposób projektowania daje największe możliwości optymalizacji topografii układu pod względem zajmowanej powierzchni, długości połączeń i innych kryteriów. Jest jednak nadzwyczaj prachłonny i podatny na omyłki. W praktyce jest stosowany w projektowaniu układów analogowych, a w przypadku układów cyfrowych wykorzystuje się go do projektowania bramek i małych bloków, które następnie są wykorzystywane jako komórki standardowe (będzie o tym mowa dalej). Projektowanie w taki sposób nie poddaje się automatyzacji, istnieją natomiast inne metody projektowania struktury fizycznej układu umożliwiające automatyczną syntezę topografii. Będzie o nich mowa w następnym wykładzie.

Końcowym rezultatem procesu projektowania jest opis masek fotolitograficznych definiujących strukturę fizyczną układu. Opis ten otrzymuje się w jednym z dwóch standardowych języków: **CIF** lub **GDSII**. Opis w języku CIF ma postać pliku tekstowego możliwego do analizy i interpretacji przez człowieka, natomiast opis w języku GDSII jest plikiem binarnym. Język GDSII jest powszechnie stosowany w przemyśle. Język CIF wychodzi obecnie z praktycznego użycia, bowiem wszystkie obiekty geometryczne są w nim opisywane na siatce o kroku równym 100 nanometrów, co nie wystarcza do opisu kształtów, jakie występują w układach o minimalnych wymiarach rzędu 100 nm i poniżej.

## 5.2. Zarys problemów projektowania

Trzy specyficzne cechy odróżniają projektowanie specjalizowanych układów scalonych od projektowania układów elektronicznych budowanych z indywidualnych, dyskretnych elementów:

1. Przy projektowaniu układów scalonych generowana i przetwarzana jest olbrzymia ilość informacji.
2. Przy przetwarzaniu olbrzymich ilości informacji nieuniknione są błędy, a nawet najdrobniejsza omyłka często prowadzi do tego, że układ nie działa.
3. Elementy układów scalonych nie występują poza tymi układami, nie można więc wiarygodnie sprawdzić działania zaprojektowanego fragmentu układu budując jego odpowiednik z elementów dyskretnych i wykonując pomiary.

Z tych trzech powodów nie można sobie dziś wyobrazić projektowania układów scalonych przy użyciu kartki i ołówka, bez wspomagania odpowiednim oprogramowaniem komputerowym.

Oszacujmy najpierw ilość wytwarzanej i przetwarzanej informacji. Zrobimy to w dość mechaniczny, uproszczony sposób, ale wynik da nam pojęcie o rozmiarach problemu. Załóżmy, że na początku mamy do czynienia z opisem funkcji układu w języku opisu sprzętu. Taki opis nawet dla bardzo złożonego układu to kilkaset do kilku tysięcy linii kodu. Traktując ten kod jako zwykły tekst możemy oszacować jego objętość na kilkanaście do kilkudziesięciu kB. Ostatecznym wynikiem procesu projektowania jest opis masek układu scalonego. Załóżmy, że układ liczy milion tranzystorów MOS (największe projektowane obecnie układy liczą kilkaset milionów elementów). Aby opisać strukturę tranzystora MOS w układzie CMOS w najprostszy sposób, trzeba zdefiniować położenie i wymiary 6 prostokątów na 4 różnych maskach. Kompletny opis prostokąta (o bokach równoległych do osi układu współrzednych) wymaga podania 4 liczb (np. współrzędnych lewego dolnego i prawego górnego wierzchołka). A więc jeden tranzystor wymaga opisu złożonego z 24 liczb. Milion tranzystorów to $24x10^6$ liczb. Jeśli są to liczby 32-bitowe, otrzymujemy objętość opisu rzędu 100 MB (w rzeczywistości dużo więcej, bo nie są uwzględnione w naszym rachunku połączenia). Ten sposób szacowania objętości informacyjnej projektu, choć naiwny, daje pojęcie o tym, jaką ogromną ilość informacji trzeba wygenerować i przetwarzać w procesie projektowania. Wiąże się to bezpośrednio z pracochłonnością procesu projektowania, o czym była już mowa w wykładzie 2.

Z olbrzymią ilością przetwarzanej informacji wiąże się bezpośrednio problem omyłek i błędów. Statystyki zebrane w różnych obszarach działalności człowieka pokazują, że w swoich działaniach człowiek przy starannej pracy popełnia średnio około 2% pomyłek. Oznacza to, że gdyby człowiek "ręcznie" zaprojektował układ złożony z 1000 tranzystorów, to w przypadku około 20 z nich mielibyśmy do czynienia z omyłkami, np. błędnie doprowadzonymi połączeniami, pomyłkami w wymiarach itp. Zwykle nawet jedna taka omyłka prowadzi do układu, który nie działa lub działa wadliwie. Doświadczenie pokazuje, że nawet w najprostszych układach liczących kilkadziesiąt elementów człowiek nie jest w stanie dostrzec wszystkich popełnionych błędów. Toteż użycie komputerowych narzędzi do weryfikacji poprawności projektu jest konieczne nawet dla bardzo prostych układów.

W procesie weryfikacji poprawności projektu układu scalonego możemy wyróżnić **weryfikację formalną** i **weryfikację funkcjonalną**.

Weryfikacja formalna projektu układu polega między innymi na sprawdzeniu, czy:

- projekt masek układu, opisujący jego topografię, odpowiada zadanemu schematowi elektrycznemu układu,
- projekt masek nie narusza **geometrycznych reguł projektowania** określonych przez producenta, jak np. minimalne dopuszczalne wymiary obszarów, odstępy między nimi itp.

Pierwszy rodzaj weryfikacji określany jest skrótem **LVS** (od angielskiego **Layout versus Schematic**), drugi - skrótem **DRC** (od angielskiego **Design Rule Checking**).

Weryfikacja formalna nie zapewnia, że zaprojektowany układ będzie poprawnie działał, bo przecież można sobie wyobrazić, że projekt spełnia wszystkie reguły projektowania, a odczytany z masek schemat jest zgodny z założonym, tylko że ten schemat był od początku błędny. Konieczna jest więc weryfikacja funkcjonalna, której istotą jest zbadanie przy pomocy symulacji komputerowych, jak będzie działał zaprojektowany układ.

Weryfikacja funkcjonalna wykonywana jest metodami symulacyjnymi, ponieważ, jak wspomniano wcześniej, nie da się zbudować prototypu układu z poszczególnych elementów i poddać go pomiarom. Symulacja działania układu może być wykonywana na kilku poziomach abstrakcji. Wyróżniamy następujące rodzaje symulacji:

- symulacja elektryczna,
- symulacja logiczna,
- symulacja funkcjonalna.

**Symulacja elektryczna** polega na rozwiązywaniu układów równań opisujących sieć elektryczną układu. Napięcia

i prądy w sieci opisane są równaniami teorii obwodów. Elementy są reprezentowane przez modele matematyczne. Taki model to równanie lub układ równań opisujących, jakie są zależności między napięciami na zewnętrznych wyprowadzeniach elementu, a prądami płynącymi przez element. Najprostszy model to model rezystora o stałej rezystancji - jest nim po prostu prawo Ohma: $I = U/R$. $R$ jest w tym modelu **parametrem modelu elementu**, czyli wielkością zmienną o wartości specyficznej dla danego konkretnego rezystora. Modele elementów półprzewodnikowych (diod, tranzystorów MOS, tranzystorów bipolarnych) są dużo bardziej skomplikowane - przypomnijmy na przykład model Ebersa-Molla tranzystora bipolarnego. Najprostsze modele tranzystora MOS i tranzystora bipolarnego były omówione w poprzednim wykładzie.

Historycznie najstarszym symulatorem układów elektronicznych, który został powszechnie zaakceptowany jako użyteczne narzędzie, jest symulator **Spice** opracowany na Uniwersytecie Kalifornijskim w Berkeley. Od niego wywodzi się bezpośrednio lub pośrednio olbrzymia liczba istniejących dziś symulatorów, samą symulację elektryczną nazywa się często "symulacją typu Spice", a symulatory określa się nazwą "symulator klasy Spice" (nawet jeśli jest to symulator nie wykorzystujący kodu źródłowego oryginalnego symulatora Spice).

Symulacja elektryczna jest niezastąpionym narzędziem weryfikacji przy projektowaniu układów scalonych. Umożliwia zbadanie prądów i napięć w układzie w stanie ustalonym, poznanie charakterystyk amplitudowych i fazowych układu w funkcji częstotliwości dla sygnałów zmiennych o małej amplitudzie, określenie przebiegów napięć i prądów w funkcji czasu dla sygnałów o dowolnej amplitudzie i zmienności w czasie, a także zbadanie bardziej subtelnych właściwości układu takich, jak poziom szumów własnych lub zniekształceń nieliniowych. Elegancka grafika współczesnych symulatorów sprawia, że na ekranie komputera oglądamy wyniki symulacji w postaci przebiegów do złudzenia przypominających przebiegi na ekranie oscyloskopu podłączonego do realnego układu. Może to powodować nadmierne zaufanie do wiarygodności wyników symulacji. Tymczasem trzeba pamiętać, że:

> **!** **Symulacja elektryczna układu przy użyciu symulatora klasy Spice jest jedynie numerycznym rozwiązywaniem równań algebraicznych i różniczkowych, i niczym więcej. Jeśli równania nie opisują dobrze symulowanego układu lub jego elementów, to wyniki symulacji mogą daleko odbiegać od działania rzeczywistego układu. Symulacja nie zastępuje zrozumienia działania układu. Jeżeli wyniki symulacji nie są zgodne z oczekiwanymi, należy je traktować z wielką ostrożnością i koniecznie ustalić przyczynę. Ślepa wiara w wyniki symulacji elektrycznej doprowadziła do niejednej klęski projektowej.**

Typowe przyczyny rozbieżności między wynikami symulacji, a dzialaniem rzeczywistego układu są następujące:

- omyłki w symulowanym schemacie elektrycznym,
- modele elementów (zwłaszcza tranzystorów) nie oddające dostatecznie dokładnie ich rzeczywistych właściwości,
- źle określone parametry elementów układu,
- brak odwzorowania w schemacie niektórych zjawisk występujących w rzeczywistym układzie (np. sprzężeń między elementami poprzez podłoże układu scalonego).

Omyłki w symulowanym schemacie to sprawa trywialna, a jednak zdarzają się one dość często i nie zawsze są łatwe do zauważenia. Dlatego, jeśli wyniki symulacji nie są zgodne z oczekiwanymi, należy przede wszystkim sprawdzić, czy poprawnie zdefiniowano symulowany schemat (a w tym także wartości parametrów elementów, jednostki, wartości i znaki napięć zasilania itp.).

Modele elementów (w postaci odpowiednich układów równań) są zwykle wbudowane w symulator. W przypadku tranzystorów, a zwłaszcza tranzystorów MOS, mamy zwykle do wyboru kilka różnych modeli. Producent układów scalonych w swej dokumentacji określa, jakich modeli należy używać i podaje dla nich wartości parametrów. Niekiedy ten sam element może być opisany kilkoma różnymi modelami do różnych zastosowań, należy to sprawdzać w dokumentacji producenta. Nawet bardzo dokładny model da złe wyniki przy niewłaściwych wartościach parametrów.

Brak odwzorowania w układzie niektórych zjawisk dotyczy przede wszystkim oddziaływań zwanych pasożytniczymi, takich jak pojemności między ścieżkami połączeń w układzie, rezystancje rozproszone obszarów elementów, rezystancje i indukcyjności ścieżek połączeń, działanie pasożytniczych elementów aktywnych. Dopóki dysponujemy jedynie schematem układu, a jego fizyczna struktura nie jest jeszcze zaprojektowana, nie da się dobrze przewidzieć tych wszystkich oddziaływań. Zależą one bowiem od konkretnych kształtów, wymiarów i rozmieszczenia elementów i połączeń między nimi. Dlatego w projektowaniu układów scalonych symulację elektryczną wykonuje się zwykle dwukrotnie. Najpierw symuluje się projektowany układ przed zaprojektowaniem jego struktury fizycznej, dla sprawdzenia poprawności projektu i zgrubnego oszacowania parametrów układu. Po zaprojektowaniu struktury fizycznej ukladu (kształtów i wymiarów elementów, ich rozmieszczenia i połączeń) projekt tej struktury poddaje się **ekstrakcji** - specjalny program komputerowy zwany ekstraktorem odtwarza schemat elektryczny układu na podstawie projektu jego struktury fizycznej (w praktyce na podstawie projektu masek produkcyjnych). Dobre ekstraktory znajdują i umieszczają w schemacie także wiele rodzajów elementów

pasożytniczych i określają ich parametry. Można teraz powtorzyć symulację elektryczną, a jej wyniki będą bliższe rzeczywistemu działaniu projektowanego układu.

Istotną wadą symulacji elektrycznej jest jej duża złożoność obliczeniowa. Nawet przy użyciu bardzo wydajnych komputerów symulację elektryczną można w praktyce wykonywać dla układów mających co najwyżej kilkaset do kilku tysięcy elementów czynnych. Zatem jest ona w pełni przydatna (a zarazem niezbędna) dla układów analogowych, które są zwykle niewielkie. W przypadku układów cyfrowych nie do pomyślenia jest symulacja elektryczna całego układu liczącego setki tysięcy lub miliony tranzystorów. Dlatego symulację elektryczną wykonuje się dla bramek lub stosunkowo niewielkich bloków funkcjonalnych. Mamy jednak w przypadku układów cyfrowych inne możliwości symulacji: symulację logiczną i symulację funkcjonalną.

W **symulacji logicznej** układ jest reprezentowany przez sieć połączonych ze sobą abstrakcyjnych obiektów - bramek logicznych. W sieci tej nie występują napięcia i prądy, lecz abstrakcyjne sygnały - stany logiczne "0" i "1", a także inne stany, np. stan nieokreślony oznaczany zwykle symbolem "X". W symulacji logicznej także występują modele elementów (w tym przypadku bramek). Takimi modelami są np. tablice prawdy dla bramek kombinacyjnych lub opis przejść między stanami dla przerzutników. Symulator logiczny określa zmiany stanów w układzie logicznym w funkcji czasu dla zadanej sekwencji zmian stanów wejść układu. Taka symulacja daje pojęcie o tym, czy od strony logicznej układ jest zaprojektowany poprawnie. W symulacji mogą być uwzględniane zależności czasowe wynikające z opóźnień wnoszonych przez bramki, tj. skończonego czasu reakcji bramek na zmianę stanów wejść. Pozwala to zaobserwować takie zjawiska, jak hazardy i wyścigi. Jednak modelowanie zjawisk opóźnień czasowych w symulacji logicznej jest bardzo uproszczone, dalekie od realizmu fizycznego. Dlatego wyniki symulacji logicznej nie mogą być podstawą do wyciągania wniosków na przykład o maksymalnej częstotliwości pracy układu. Tę pozwala oszacować tylko symulacja elektryczna.

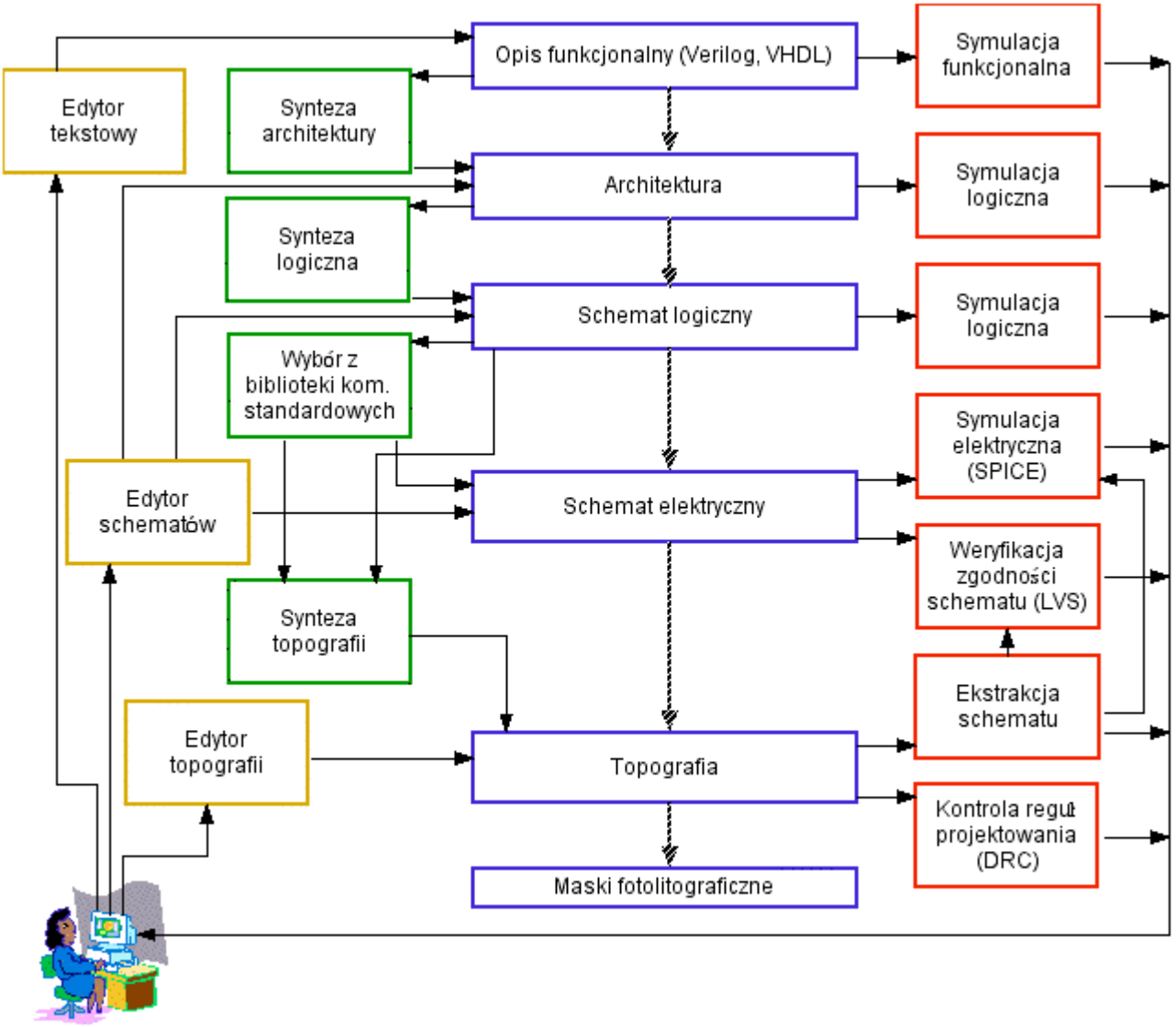
**Symulacja funkcjonalna** występuje na jeszcze wyższym poziomie abstrakcji. Schemat logiczny układu nie jest jeszcze znany, natomiast zdefiniowana jest funkcja układu (czyli opis jego działania, zwany także **opisem behawioralnym**). Opis taki formułuje się w jednym z języków opisu sprzętu (ang. **Hardware Description Language**, **HDL**). Dwa najbardziej popularne języki tego typu to **Verilog** i **VHDL**. Języki te początkowo służyły wyłącznie do opisu układów cyfrowych, ale obecnie mają rozszerzenia pozwalające także na opisywanie układów analogowych. Opis funkcji układu w takim języku jest formalnie podobny do opisu algorytmu w typowym języku programowania, np. w języku C (zresztą składnia języka Verilog jest wzorowana na składni języka C). Dysponując odpowiednim symulatorem można opis behawioralny układu w języku VHDL lub Verilog poddać symulacji. Jej celem jest sprawdzenie, czy funkcja układu jest zdefiniowana w sposób prawidłowy, kompletny i zgodny z intencją projektanta. Istnieją obecnie komputerowe systemy projektowania, które umożliwiają dokonanie syntezy układu cyfrowego na podstawie jego opisu behawioralnego. Jednak nawet jeśli nie dysponujemy takim systemem lub nie zamierzamy go użyć, opisanie funkcji układu w Verilogu lub VHDL i wykonanie symulacji funkcjonalnej jest bardzo pożądane, bo pozwala uniknąć projektowania układu o funkcji zdefiniowanej błędnie, w sposób niekompletny lub wewnętrznie sprzeczny. Doświadczenie pokazuje, że takie przypadki zdarzają się częściej, niż można by się spodziewać.

Jednym ze skutecznych sposobów unikania błędów w projektowaniu jest wykonywanie projektu w sposób zautomatyzowany, przez odpowiednie programy komputerowe (co jest określane angielskim terminem "**correctness by construction** "). Będzie o tym mowa dalej. Nawet tak wykonany projekt poddaje się jednak wszystkim weryfikacjom.

**! Każdy projekt, niezależnie od tego, w jaki sposób i przy pomocy jakich narzędzi wspomagających został wykonany, musi być poddany zarówno weryfikacji formalnej, jak i weryfikacji funkcjonalnej.**

## 5.3. Etapy projektowania i narzędzia wspomagające

Spójrzmy jeszcze raz na rysunek 5.1 pokazujący proces projektowania oraz wspomagające ten proces oprogramowanie. W czerwonych ramkach programy komputerowe służące do weryfikacji projektu, w zielonych ramkach programy wykonujące kolejne etapy automatycznej syntezy projektu, w żółtych ramkach programy komputerowe służące projektantowi do wprowadzania informacji. Cienkie linie pokazują kierunki przepływu informacji.




Rys. 5.1 (powtórzony). Schemat typowego systemu projektowania i przebieg procesu projektowania.

Absolutnie minimalny zestaw programów umożliwiający zaprojektowanie układu scalonego składa się z następujących programów:

Programy do wprowadzania danych i tworzenia projektu:

- Edytor tekstowy
- Edytor topografii

Programy do weryfikacji:

- Program do kontroli reguł projektowania (DRC)
- Ekstraktor schematu elektrycznego
- Symulator układów elektronicznych (typu SPICE)

Funkcje tych programów mogą być połączone, na przykład edytor topografii ma często wbudowany moduł

umożliwiający kontrolę reguł projektowania, co daje możliwość kontroli na bieżąco, w trakcie tworzenia topografii. Dzięki temu unika się pracochłonnych rozległych przeróbek projektu, jakie mogłyby być konieczne, gdyby naruszenia reguł projektowania zostały wykryte dopiero po zakończeniu projektowania topografii.

Wymienione wyżej programy umożliwiają wprowadzenie opisu schematu elektrycznego w postaci tekstowej (np. w języku wejściowym programu SPICE), wykonanie symulacji, zaprojektowanie topografii (tj. "narysowanie" jej na ekranie komputera w edytorze topografii), skontrolowanie, czy topografia ta spełnia reguły projektowania, odtworzenie (wyekstrahowanie) z topografii schematu elektrycznego, oraz poddanie tego schematu symulacji dla sprawdzenia, czy układ wyprodukowany na podstawie zaprojektowanej topografii będzie działał zgodnie z intencją projektanta. Może to być zarówno układ analogowy, jak i cyfrowy. Dodatkowym ułatwieniem przy projektowaniu moźe być graficzny edytor schematów, czyli program pozwalający "narysować" na ekranie komputera schemat układu zamiast opisywać go tekstowo. Dużą pomocą przy weryfikacji jest program do sprawdzenia zgodności schematów elektrycznych (Layout versus Schematic - LVS). Pozwala on skontrolować, czy schemat układu opisany tekstowo lub wprowadzony graficznie zgadza się ze schematem odczytanym z topografii przez ekstraktor.

Taki zestaw programów umożliwia projektowanie w stylu *full custom*. W tym przypadku programy komputerowe wspomagają pracę czlowieka przy projektowaniu i umożliwiają weryfikację projektu, ale tej pracy nie automatyzują. Jak już wiemy, taki sposób projektowania jest niezwykle pracochłonny. W praktyce możliwe jest projektowanie w ten sposób układów mających do kilkuset elementów. Ten sposób projektowania nie ogranicza jednak w niczym swobody projektanta, dając mu największe możliwości optymalizacji projektu pod każdym względem. Od projektanta wymaga zarówno znajomości zagadnień układowych, jak i zrozumienia struktury fizycznej układu i działania jego elementów.

W przypadku układów cyfrowych mamy jednak także możliwość daleko idącej automatyzacji procesu projektowania. Mamy do dyspozycji następujące programy:

Programy automatyzujące projektowanie

- Program syntezy logicznej
- Programy automatycznej syntezy struktury fizycznej układu

Programy do weryfikacji

- Symulator funkcjonalny
- Symulatory logiczne

Układ cyfrowy może być opisany w języku opisu sprzętu (o tych językach będzie mowa dalej) i poddany symulacji funkcjonalnej w celu sprawdzenia, czy opis jest poprawny, a układ działa zgodnie z intencją projektanta. Służą do tego programy umożliwiające symulację funkcjonalną. Prawidłowy opis funkcjonalny może być poddany syntezie logicznej, której ostatecznym wynikiem jest schemat logiczny układu złożony z bramek oraz ewnetualnie większych bloków takich, jak sumatory, układy mnożące czy bloki pamięci. Taki schemat może być w mniejszym lub większym stopniu ulepszany lub uzupełniany przez projektanta, na przykład przy wykorzystaniu edytora schematów. Gotowy do dalszych prac projektowych schemat logiczny poddaje się weryfikcji przy użyciu symulatora logicznego. Jest kilka rodzajów takich symulatorów. Schemat logiczny może być reprezentowany przy pomocy typowych bramek cyfrowych, które wykonują określoną funkcję logiczną i dodatkowo mogą być scharakteryzowane pod względem parametrów elektrycznych, np. czasów propagacji sygnału (będzie o nich mowa w dalszych wykładach) czy też poboru mocy. Te dane mają oczywiście charakter przybliżony, bo na etapie symulacji logicznej, gdy fizyczna realizacja układu nie jest jeszcze znana, dokładne wartości parametrów elektrycznych nie są określone. Inny rodzaj symulatora logicznego to symulator, w którym elementami są pojedyncze tranzystory, ale traktowane są one jako idealne przełączniki, które przewodzą prąd lub nie. W wielu symulatorach mogą występować zarówno bramki, jak i pojedyczne tranzystory. Istnieją także symulatory umożliwiające symulację mieszaną - część układu może być symulowana logicznie, a część elektrycznie. Są one przydatne przy projektowaniu układów mieszanych, analogowo-cyfrowych.

Gdy schemat logiczny jest gotowy i zweryfikowany przy pomocy symulatorów, można projektować fizyczną strukturę układu. Wiele profesjonalnych systemów projektowania daje tu różne możliwości. Można wygenerować plik danych do zaprogramowania układu programowalnego (FPGA). Możliwe jest wykonanie automatycznej syntezy fizycznej struktury układu przy wykorzystaniu matryc bramkowych lub komórek standardowych. Będzie to omawiane dokładniej w następnym wykładzie.

W skład profesjonalnych systemów projektowania wchodzą też zwykle programy do rozwiązywania pewnych specjalnych problemów występujących zwłaszcza przy projektowaniu układów przeznaczonych do wytwarzania w najbardziej zaawansowanych technologiach lub układów, którym stawia się szczególnie wysokie wymagania. Przykłady takich programów to:

- Programy do szacowania poboru mocy
- Programy do projektowania i analizy sieci połączeń zasilania bloków układu
- Programy do projektowania sieci połączeń rozprowadzających sygnały zegarowe
- Programy do generacji testów dla układów cyfrowych

Te programy nie będą dalej omawiane, ale poruszone będą w dalszych wykładach problemy, do rozwiązywania których te programy służą.

Odrębną kategorię systemów oprogramowania do projektowania układów scalonych stanowią systemy projektowania układów mikrofalowych. Krzemowe układy scalone CMOS z powodzeniem mogą pracować do częstotliwości rzędu 5 - 6 GHz, a układy bipolarne są przydatne nawet przy częstotliwościach rzędu kilkudziesięciu GHz. Te układy i sposoby ich projektowania wykraczają poza zakres przedmiotu "Układy scalone".

Większość programów do projektowania i weryfikacji układów scalonych wymaga danych dotyczących konkretnej technologii, w której układy będą produkowane. Przykładowo, program do kontroli reguł projektowania musi mieć dany zestaw tych reguł, program do symulacji elektrycznej wymaga parametrów modeli elementów, itp. Tego rodzaju dane dostarczane są przez producentów w postaci zestawów plików zwanych plikami technologicznymi (w jęz. angielskim zwanych "design kit", co można przetłumaczyć jako "pakiet projektowy").

## 5.4. Języki opisu sprzętu

Języki opisu sprzętu były początkowo pomyślane jako narzędzie do definiowania projektów w sformalizowany sposób, aby ułatwić uniknięcie luk, omyłek i niespójności. Obecnie dają one znacznie większe możliwości. Projekt układu opisany w jednym z tych języków może być poddany symulacji w celu sprawdzenia, czy układ będzie działał zgodnie z intencją projektanta. Możliwa jest także (choć nie w każdym przypadku) automatyczna synteza układu - do poziomu projektu logicznego, a nawet do poziomu topografii. W ten sposób języki opisu sprzętu stały się językami wejściowymi systemów automatycznego projektowania układów scalonych. Dwa standardowe, powszechnie używane języki opisu sprzętu to **Verilog** i **VHDL**. Każdy z nich może być używany w powiązaniu z oprogramowaniem do projektowania układów scalonych pochodzącym z wielu różnych firm. Istnieją też języki opisu sprzętu związane z konkretnym systemem projektowania. Z jednym z takich języków, o nazwie **AHDL**, zapoznasz się w ramach przedmiotu "Systemy cyfrowe". Język AHDL jest związany z systemem projektowania programowalnych układów cyfrowych firmy Altera.

Języki opisu sprzętu pozwalają opisywać układy na kilku poziomach abstrakcji. Możliwy jest **opis funkcjonalny** (zwany też **behawioralnym**); taki opis pozwala zdefiniować funkcję układu (przez opis algorytmu realizowanego przez układ) bez określania, w jaki sposób ta funkcja ma być zrealizowana. Możliwy jest **opis architektury**, gdzie definiuje się bloki, z jakich składa się układ, określa ich działanie, a w tym sygnały wejściowe i wyjściowe przesyłane miedzy blokami, ale nie określa się, jak bloki mają być zrealizowane. Możliwy jest także **opis strukturalny**, w którym definiowany jest konkretny schemat logiczny bloku lub całego układu. Opisy te można łączyć w jednym projekcie, i najczęściej w opisach bardziej złożonych układów znajdziemy fragmenty opisane na każdym z tych poziomow abstrakcji. Przykładowo, układ może być opisany przez zdefiniowanie jego architektury, a w tym opisie jedne bloki mogą być opisane przez zdefiniowanie ich funkcji (czyli opis funkcjonalny), a inne przez podanie konkretnego schematu logicznego (czyli opis strukturalny).

Te trzy poziomy abstrakcji potrzebne są po pierwsze dlatego, że nie każdy opis funkcjonalny może służyć do automatycznej syntezy układu. Mówimy, że opis funkcjonalny jest syntezowalny, jeśli automatyczna synteza jest możliwa. Jednak w pewnych przypadkach automatyczna synteza nie jest możliwa bez znajomości intencji projektanta. Przykładowo, instrukcja zsumowania dwóch liczb może być jednoznacznie przekształcona na schemat logiczny sumatora tylko wtedy, gdy znany jest sposób reprezentacji tych liczb oraz rodzaj sumatora, jaki ma być zastosowany. Jak widzimy, czysty opis funkcjonalny zwykle nie wystarcza, konieczne jest zdefiniowanie pewnych szczegółów pożądanej implementacji funkcji w postaci ukladu scalonego. Niemniej, opis funkcjonalny jest bardzo wygodny jako punkt startowy projektu, ponieważ można poddać go symulacji funkcjonalnej nawet jeśli nie jest syntezowalny. W ten sposób można uniknąć błędów na etapie definiowania funkcji układu. Ponadto opis funkcjonalny jest znacznie krótszy, prostszy i możliwy do interpretacji przez człowieka, podczas gdy opis architektury, a tym bardziej opis strukturalny jest dla człowieka bardzo trudny do analizy i interpretacji. Nawet nie znając żadnego z wymienionych języków nie trudno wyobrazić sobie, że gdy widzimy instrukcję postaci Z = X+Y, to intuicyjnie rozumiemy, o co chodzi, zaś gdy widzimy opis schematu logicznego sumatora złożonego z kilkuset bramek logicznych (czyli kilkaset linii kodu opisu strukturalnego), to bez dodatkowych wyjaśnień nie domyślimy się, co ów układ robi (chyba że sami go zaprojektowaliśmy), i nie sprawdzimy również bez użycia symulatora, czy jest to schemat poprawny.

W dodatku 1 zobaczysz, jak wygląda przykładowy prosty opis funkcjonalny w języku Verilog.

Po drugie, opis architektury oraz opis strukturalny są potrzebne nie tylko dlatego, że nie każdy opis funkcjonalny jest syntezowalny, ale także dlatego, że wyniki automatycznej syntezy układu mogą być dalekie od optymalnych. Dlatego najczęściej spotyka się opisy, w których projektant definiuje architekturę układu, część bloków opisuje funkcjonalnie, a niektóre opisuje strukturalnie, ponieważ wie, jaki schemat logiczny będzie dla realizacji funkcji tych bloków najlepszy.

Języki opisu sprzętu były początkowo przeznaczone tylko do opisywania układów cyfrowych. Obecnie istnieją rozszerzenia (Verilog-A, VHDL-AMS) umożliwiające także opisywanie układów analogowych. Opisy układów analogowych nie są syntezowalne, ponieważ nie udało się, jak dotąd, zautomatyzować projektowania układów analogowych. Możliwa jest natomiast ich symulacja. Możliwa jest też symulacja układów mieszanych, analogowo-cyfrowych.

Opis funkcjonalny układu w języku opisu sprzętu jest na pierwszy rzut oka podobny do opisu algorytmu w tradycyjnym języku programowania (zwłaszcza że składnia języków opisu sprzętu pod wieloma względami wywodzi się z języków programowania, np. dla języka Verilog wzorem składni był język C, dla VHDL - język ADA). Istnieje jednak kilka zasadniczych różnic. Po pierwsze, operacje arytmetyczne i logiczne w językach programowania mają sens operacji na abstrakcyjnych obiektach, jakimi są liczby, zmienne czy też wartości logiczne. Podobnie zapisane operacje w językach opisu sprzętu oznaczają działanie konkretnych fizycznych obiektów - bramek, przerzutników, rejestrów itp. Dlatego instrukcja postaci Z = X+Y w języku opisu sprzętu nie jest wystarczająca. Aby opis był kompletny, muszą być znane takie informacje, jak format liczb, liczba bitów, ich kolejność (bit najmniej znaczący na początku czy na końcu?), a także co się ma stać po wykonaniu operacji - czy wynik ma być zapamiętany, czy gdzieś przesłany, jesli tak, to gdzie? Po drugie, języki opisu sprzętu umożliwiają

zdefiniowanie zależności czasowych między operacjami, w tym także operacji współbieżnych, tj. wykonywanych równocześnie i niezależnie.

Na koniec warto dodać, że języki opisu sprzętu mogą być używane do opisu układów cyfrowych, analogowych i mieszanych realizowanych w dowolny sposób, nie tylko jako układy specjalizowane (ASIC), którym poświęcony jest przedmiot "Układy scalone".

## 5.4. Dodatek 1: Przykładowy opis funkcjonalny w języku Verilog

Oto prosty przykład opisu funkcjonalnego. Opisywany blok cyfrowy o nazwie "arp" ma realizować następujący algorytm. Na wejściu dane jest czterobitowe słowo A oraz jeden bit oznaczony PM. Po każdej zmianie wartości A lub PM należy sprawdzić, czy nowa wartość PM jest jedynką i czy nowa wartość A nie jest równa maksymalnej możliwej, tj. czterem jedynkom. Jeśli oba warunki są spełnione, to wyjściowe czterobitowe słowo Q otrzymuje wartość o 1 większą od A, w przeciwnym razie Q otrzymuje wartość równą A. Wartość Q jest zapamiętywana w czterobitowym rejestrze.

A oto opis:

```
module arp (A, PM, Q);
input [3:0]    A;
input    PM;
output [3:0]    Q;
reg [3:0]    Q;
    always @(A  or  PM)
       if (PM   &&   A < 4'b1111)
          Q = A + 1;
       else
          Q = A;
endmodule
```

## Bibliografia

[1] W. Marciniak, "*Przyrządy półprzewodnikowe i układy scalone*", WNT Warszawa 1987
[2] J.Porębski, P.Korohoda, "*SPICE program analizy nieliniowej układów elektronicznych*", WNT Warszawa 1992
[3] T. Łuba, B. Zbierzchowski, "*Komputerowe projektowanie układów cyfrowych*", WKiŁ Warszawa 2000
(Książka omawia projektowanie układów cyfrowych przy wykorzystaniu programowalnych układów scalonych)
[4] J. Ogrodzki, "*Komputerowa analiza układów elektronicznych*", PWN Warszawa 1994
(Książka omawia matematyczne podstawy i algorytmy symulacji elektrycznej)
[5] S. M. Rubin, "*Computer Aids for VLSI Design*", Addison-Wesley Publishing Co., Inc. 1987
(Książka dostępna w Internecie, wraz z bezpłatnym oprogramowaniem do projektowania układów scalonych - system "Electric")

# Wykład 6: Metody i style projektowania

## Wstęp

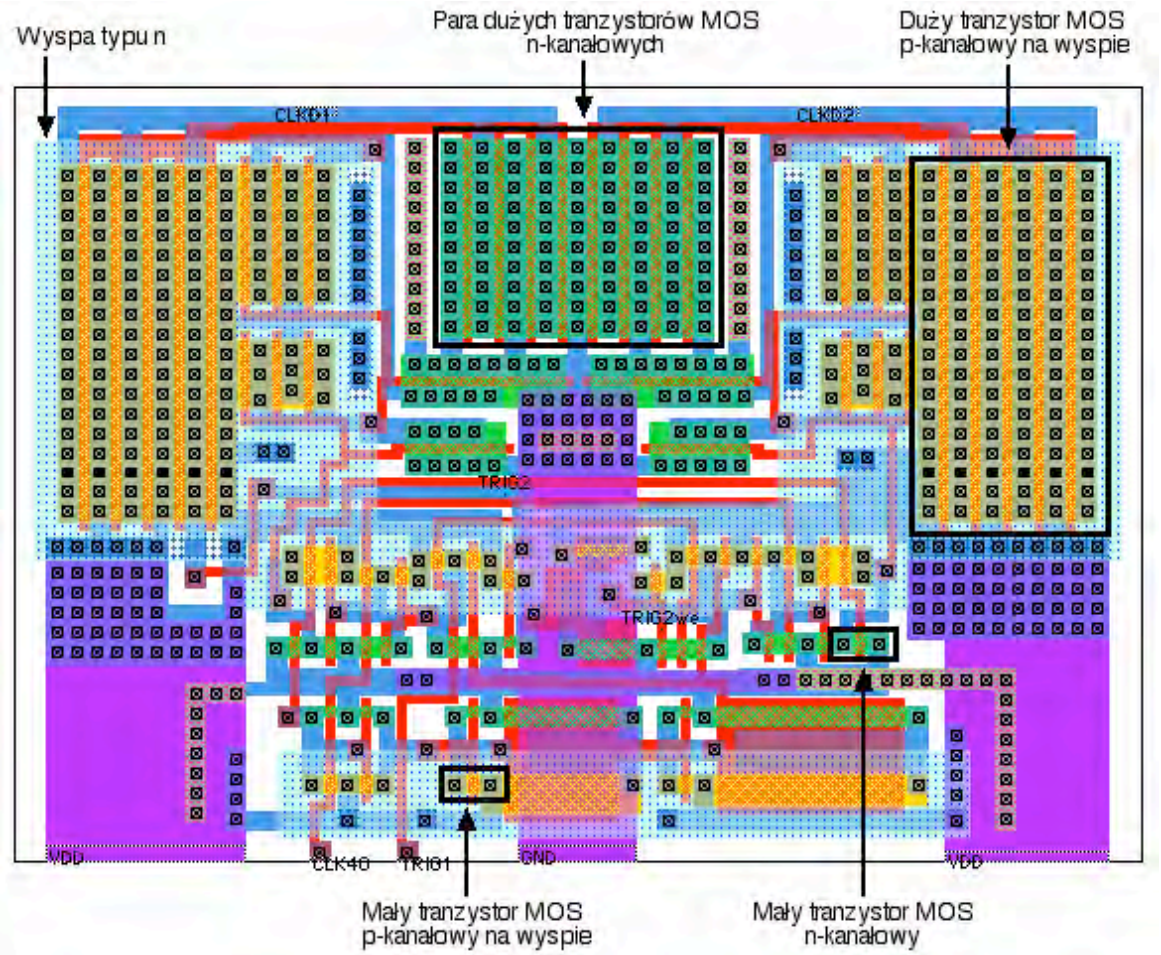
Wykład 6 omawia najważniejsze sposoby zmniejszania pracochłonności projektowania układów scalonych i redukowania ryzyka popełnienia błędów. W pierwszej części wykładu omawiane jest projektowanie w stylu *full custom* oraz uproszczenia tego sposobu projektowania. Podstawowe umiejętności projektowania w tym stylu są potrzebne każdemu, kto zamierza zajmować się projektowaniem, chociażby po to, by mieć świadomość praktycznych problemów występujących przy projektowaniu fizycznej struktury układu. Druga część wykładu przedstawia najpowszechniej stosowane metody projektowania zautomatyzowanego, a trzecia omawia zakresy ich przydatności.

Praktyczne ćwiczenia nauczą Cię podstawowych umiejętności projektowania topografii w stylu *full custom*. Przerób uważnie i starannie te ćwiczenia, bowiem nabyte w ten sposób umiejętności będą niezbędne do wykonania projektów, które są podstawą zaliczenia przedmiotu. Wprowadzenie do ćwiczeń znajdziesz także w postaci prezentacji wideo.

## 6.1. Projektowanie w stylu *full custom*

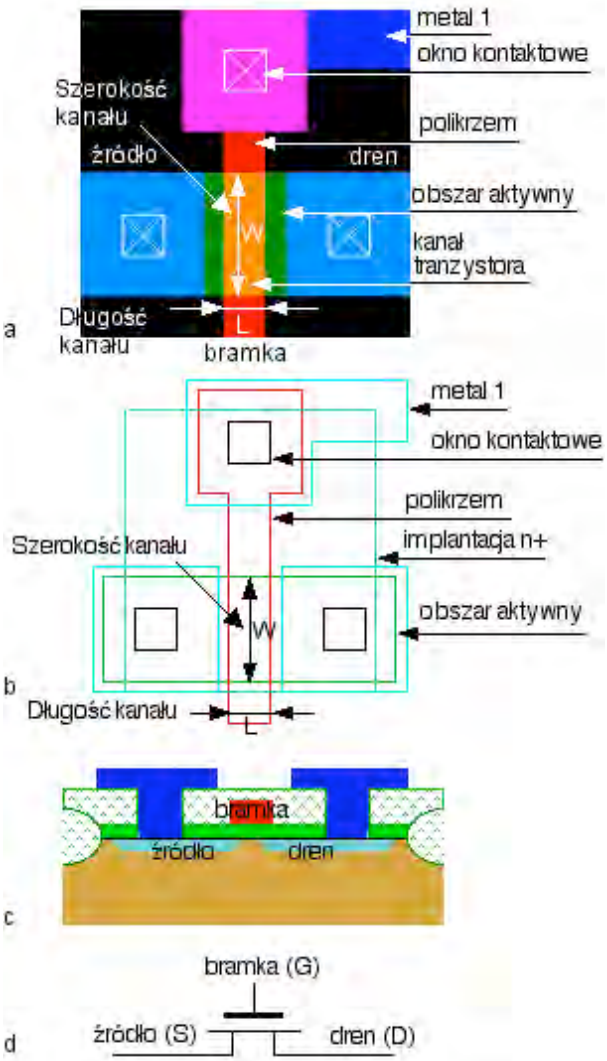
Zanim zaczniesz czytać, przypomnij sobie jak powstają układy scalone CMOS (wykład 3)!

Projektowanie w stylu *full custom* polega na "narysowaniu" przez projektanta całej topografii układu. Używany jest do tego specjalizowany edytor graficzny zwany edytorem topografii. Taki edytor jest niezbędnym składnikiem każdego komputerowego systemu projektowania układów scalonych. Przykład fragmentu topografii układu zaprojektowanej w stylu full custom pokazuje rys. 6.1.




Rys. 6.1. Mały blok funkcjonalny (generator dwóch przesuniętych w czasie sygnałów zegarowych) zaprojektowany w stylu full custom (projekt wykonany edytorem topografii "Uncle" wchodzącym w skład systemu IMiOCAD opracowanego na Politechnice Warszawskiej)

Zasadniczo rysowane są wszystkie maski fotolitograficzne, które posłużą do produkcji układu, ale możliwe są tu różne uproszczenia. Omówimy je na najprostszym przykładzie: projektu tranzystora MOS. Rys. 6.2 pokazuje symbol tranzystora MOS n-kanałowego, przekrój przez jego strukturę oraz maski fotolitograficzne określające tę strukturę. Widoczny jest też obraz topografii tego tranzystora w programie "Microwind", który służy do ćwiczeń w projektowaniu w ramach przedmiotu "Układy scalone".

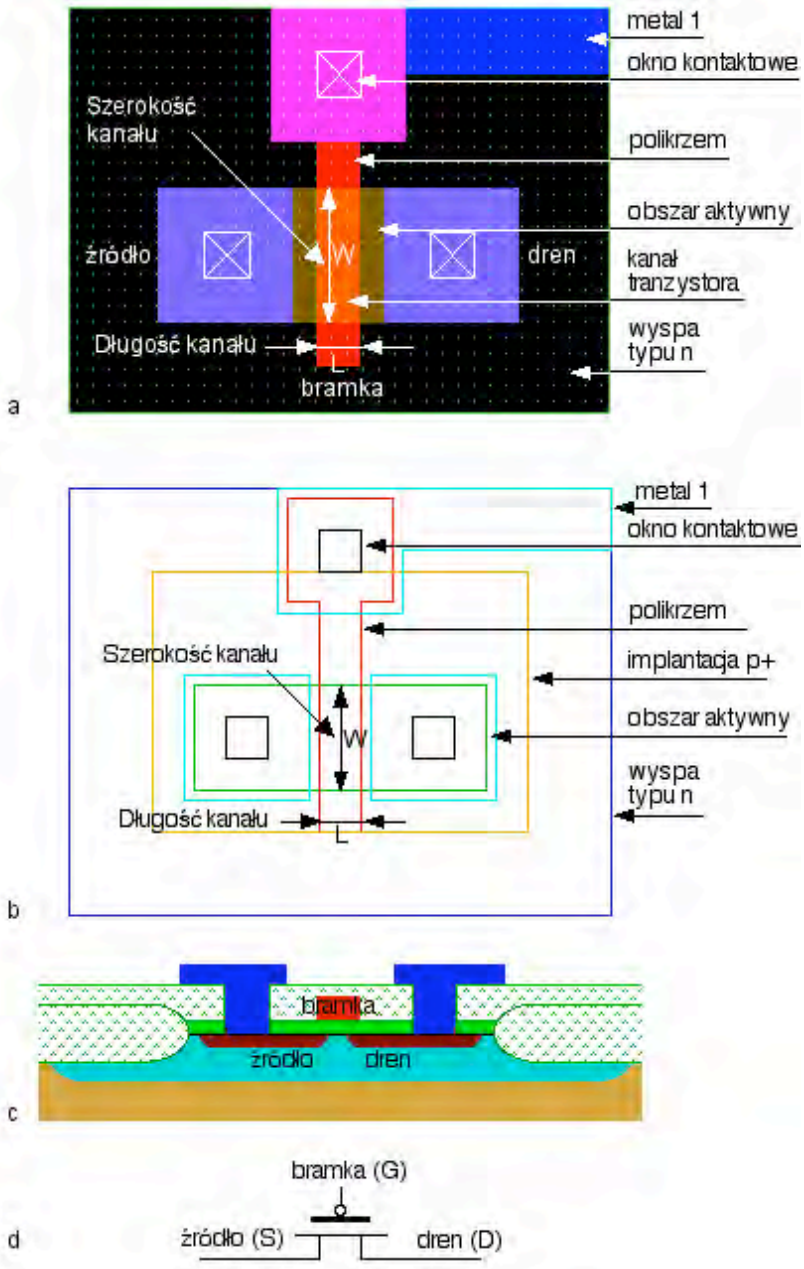



Rys. 6.2. Widok topografii tranzystora nMOS w programie "Microwind" (a), kontury masek fotolitograficznych (b), przekrój przez strukturę tranzystora (c) oraz symbol w schematach elektrycznych (d)

Na rys. 6.2b widoczne są następujące maski:

- Maska obszaru aktywnego (zwana czasami z przyczyn historycznych maską dyfuzji) - kolor jasnozielony. Ta maska określa szerokość kanału tranzystora.
- Maska polikrzemu - kolor czerwony. Ta maska określa długość kanału tranzystora.
- Maska implantacji typu n - kolor ciemnozielony.
- Maska kontaktów - kolor czarny (trzy obszary).
- Maska metalu 1 - kolor jasnobłękitny (trzy obszary).

W przypadku tranzystora p-kanałowego wygląd topografii i zestaw masek jest podobny - rys. 6.3.




Rys. 6.3. Widok topografii tranzystora pMOS w programie "Microwind" (a), kontury masek fotolitograficznych (b), przekrój przez strukturę tranzystora (c) oraz symbol w schematach elektrycznych (d)

Na rys. 6.3 widoczne są następujące maski:

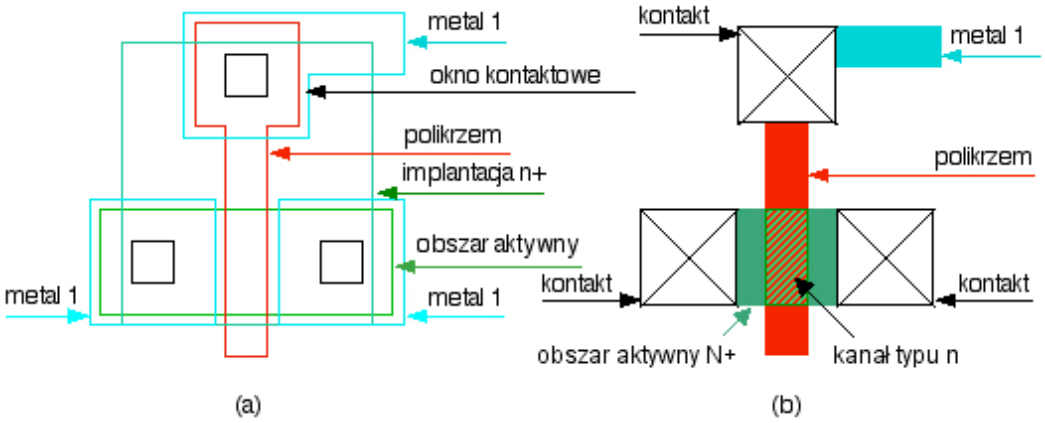
- Maska wyspy typu n - kolor ciemnoniebieski (część krawędzi tej maski jest zasłonięta przez krawędź maski metalu 1).
- Maska obszaru aktywnego (zwana czasami z przyczyn historycznych maską dyfuzji) - kolor jasnozielony. Ta maska określa szerokość kanału tranzystora.
- Maska polikrzemu - kolor czerwony. Ta maska określa długość kanału tranzystora.
- Maska implantacji typu p - kolor ciemnożółty.
- Maska kontaktów - kolor czarny (trzy obszary).
- Maska metalu 1 - kolor jasnobłękitny (trzy obszary).

Dla wszystkich masek producenci określają **geometryczne reguły projektowania**. Dotyczą one minimalnych szerokości i odstępów obszarów na tej samej masce, minimalnych odstępów obszarów na różnych maskach, minimalnych zakładek gdy obszary powinny się częściowo lub całkowicie nakładać itp. Dla najbardziej zaawansowanych technologii pełny zestaw reguł może zawierać nawet kilkaset reguł. Rysowanie topografii układu z zachowaniem tych wszystkich reguł jest bardzo pracochłonne i uciążliwe.

Dla uproszczenia i zmniejszenia pracochłonności stosowane bywają dwa uproszczone sposoby reprezentacji topografii: topografia wyrażona w postaci zwanej półsymboliczną przy pomocy **warstw abstrakcyjnych** oraz

topografia w postaci symbolicznej wyrażona poprzez **schemat kreskowy**.

**Reprezentacją półsymboliczną** topografii nazywamy taką reprezentację, w której obok obiektów geometrycznych na maskach reprezentujących wprost warstwy fizyczne występujące w strukturze układu (np. polikrzem, metal) występują także obiekty o złożonej strukturze, takie jak obszar aktywny N+, obszar aktywny P+, kanał tranzystora. Obiekty te występują na warstwach abstrakcyjnych. Warstwami tymi przy projektowaniu posługujemy się tak samo, jak maskami, jednak warstwy abstrakcyjne nie są równoważne maskom. Obiekty geometryczne na warstwach abstracyjnych odpowiadają pewnym kombinacjom obiektów na dwóch lub więcej maskach. Przykładowo: warstwa "obszar aktywny N+" oznacza obszar aktywny, do którego wykonano implantację donorów. Każdy obiekt geometryczny na tej warstwie oznacza więc obszar wspólny obiektów na dwóch maskach: na masce obszaru aktywnego i na masce implantacji typu *n*. Z kolei kanał tranzystora nMOS powstaje tam, gdzie obszar polikrzemu przecina obszar aktywny typu N+. Zatem obiekt geometryczny na warstwie abstracyjnej "kanał tranzystora nMOS" oznacza obszar wspólny obiektów na trzech maskach: obszaru aktywnego, implantacji typu *n* i polikrzemu. Gdy topografia układu jest już zaprojektowana, edytor topografii przekształca obiekty geometryczne na warstwach abstrakcyjnych na odpowiednie obiekty na maskach. Rys. 6.4. ilustruje ideę warstw abstrakcyjnych.




Rys. 6.4. Maski tranzystora nMOS (a) i odpowiadająca im topografia zbudowana z warstw abstrakcyjnych (b)

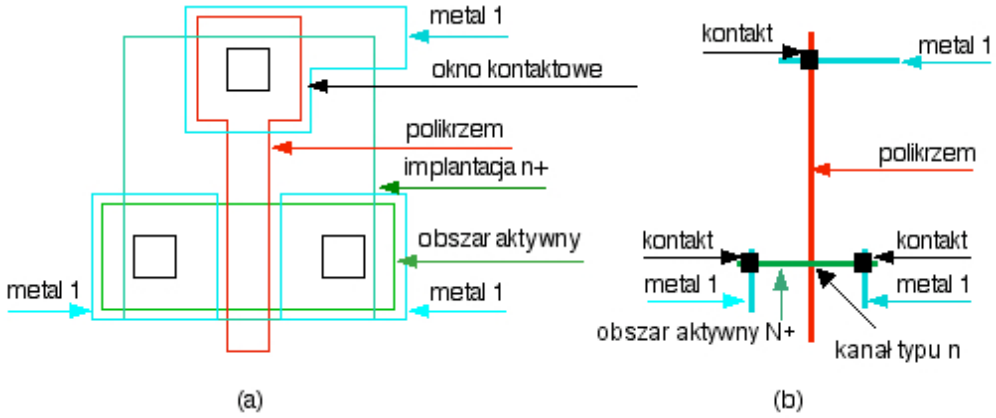
Zastosowanie umiejętnie zdefiniowanych warstw abstrakcyjnych przynosi kilka korzyści:

- Topografia układu staje się bardziej przejrzysta i czytelna.
- Interesujące dla konstruktora obiekty (przede wszystkim tranzystory) rysuje się wprost, a nie jako kombinacje kształtów na kilku maskach.
- Znaczną część reguł projektowania można ukryć w algorytmach przekształcania warstw abstrakcyjnych na maski, przez co zbiór reguł projektowania ulega dużemu uproszczeniu.

Ponadto dla warstw abstrakcyjnych często definiuje się uproszczone **skalowalne reguły projektowania**. Są to reguły zdefiniowane w umownych jednostkach oznaczanych zwykle symbolem **lambda**, a nie w mikrometrach. Są one tak obmyślone, że mogą być stosowane do wielu różnych technologii, pod warunkiem ustalenia właściwego dla danej technologii przeliczenia lambdy na mikrometry. Reguły skalowalne są proste i jest ich niewiele. Jednak po to, aby można je było dostosowywać do różnych technologii bez zmian, a jedynie przez dobór wielkości lambdy, mają one zwykle minimalne wymiary, odstępy itp. określone z pewnym zapasem. Zatem stosując tego rodzaju reguły otrzymujemy topografię układu nieco mniej gęsto "upakowaną", zajmującą nieco większą powierzchnię, niż w przypadku stosowania reguł ściśle dostosowanych do jednej konkretnej technologii. Nie ma to jednak w praktyce na ogół większego znaczenia. W ć wiczeniach towarzyszących temu wykładowi stosowane będą reguły projektowania (znajdziesz je w dodatku 1) zgodne z regułami zdefiniowanymi dla programu "Microwind".

W typowych systemach projektowania (również w przypadku programu "Microwind") reguły projektowania są definiowane w **pliku technologicznym**, a nie w samym programie. Umożliwia to stosowanie systemu projektowania do różnych technologii. Plik technologiczny definiuje także warstwy abstrakcyjne (ich nazwy, wygląd - kolor, deseń - na ekranie komputera) oraz algorytmy przekształcania obiektów na warstwach abstrakcyjnych na obiekty na maskach fotolitograficznych.

**Reprezentacją symboliczną** topografii w postaci schematu kreskowego nazywamy taką reprezentację, w której wszystkie obiekty takie, jak obszary aktywne, obszary polikrzemu i ścieżki połączeń są przedstawione w postaci linii prostych. Taka reprezentacja abstrahuje od rzeczywistych wymiarów obiektów i określa jedynie ich wzajemne położenia oraz połączenia elektryczne między nimi, do czego służą symbolicznie przedstawione kontakty. Ideę schematu kreskowego ilustruje rys. 6.5.




Rys. 6.5. Maski tranzystora nMOS (a) i odpowiadająca im topografia w postaci schematu kreskowego (b)

Istnieją edytory topografii, które pozwalają narysować topografię w postaci schematu kreskowego, a następnie potrafią przetworzyć ten schemat na rzeczywistą topografię (tj. komplet masek) przy zastosowaniu reguł projektowania danej technologii. Jednak otrzymane w ten sposób topografie są dalekie od doskonałości. Metoda schematów kreskowych miała przed laty okres pewnej popularności, ale utraciła tę popularność, gdy pojawiły się systemy projektowania umożliwiające automatyczną syntezę topografii przy zastosowaniu komórek standardowych (będzie o tym mowa dalej). Dziś można polecić metodę schematów kreskowych jako sposób wstępnego rozplanowania rozmieszczenia elementów. Nawet jeśli nie mamy edytora topografii przystosowanego do rysowania schematów kreskowych, naszkicowanie takiego schematu na papierze pozwala wstępnie ocenić, jak wzajemnie rozmieścić elementy i połączenia w projektowanym fragmencie układu.

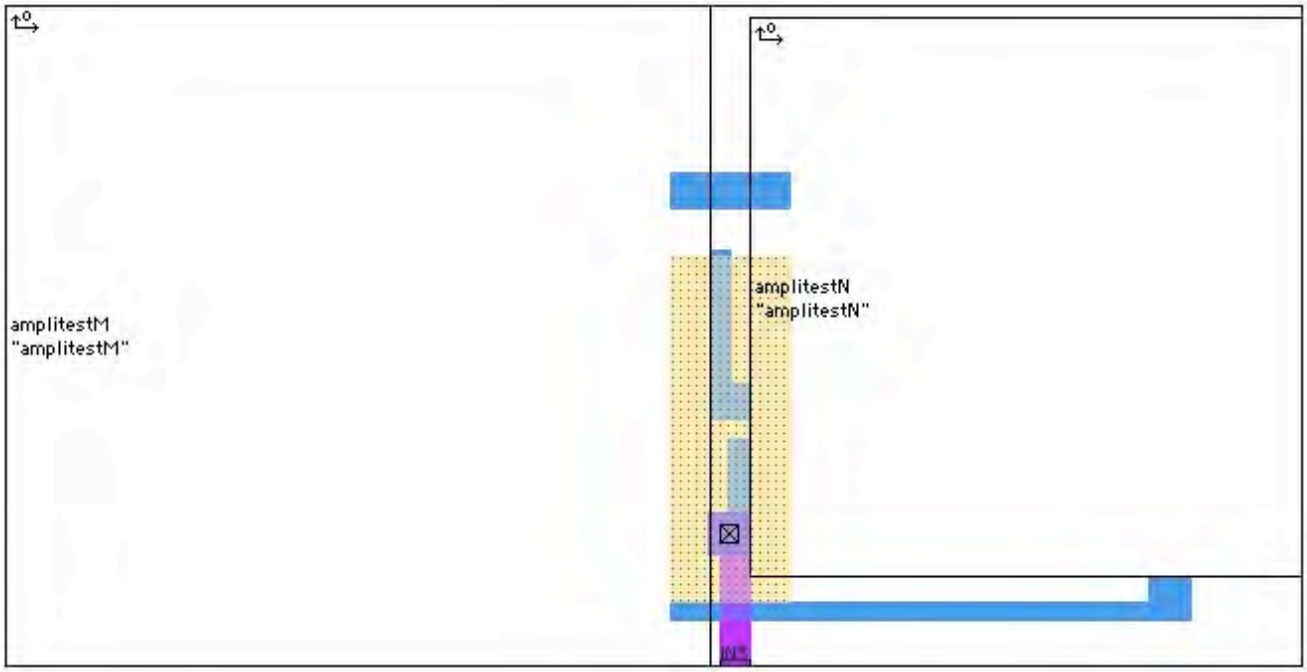
Mimo wielu lat intensywnych prac projektowania topografii w stylu *full custom* nie udało się jak dotąd skutecznie zautomatyzować, Dlatego tym sposobem projektowania posługujemy się tylko wtedy, gdy jest to niezbędne. Jest kilka takich przypadków:

- projektowanie układów analogowych,
- projektowanie topografii komórek standardowych (będzie o nich mowa dalej),
- prowadzenie połączeń między wnętrzem układu, a pierścieniem pól montażowych.

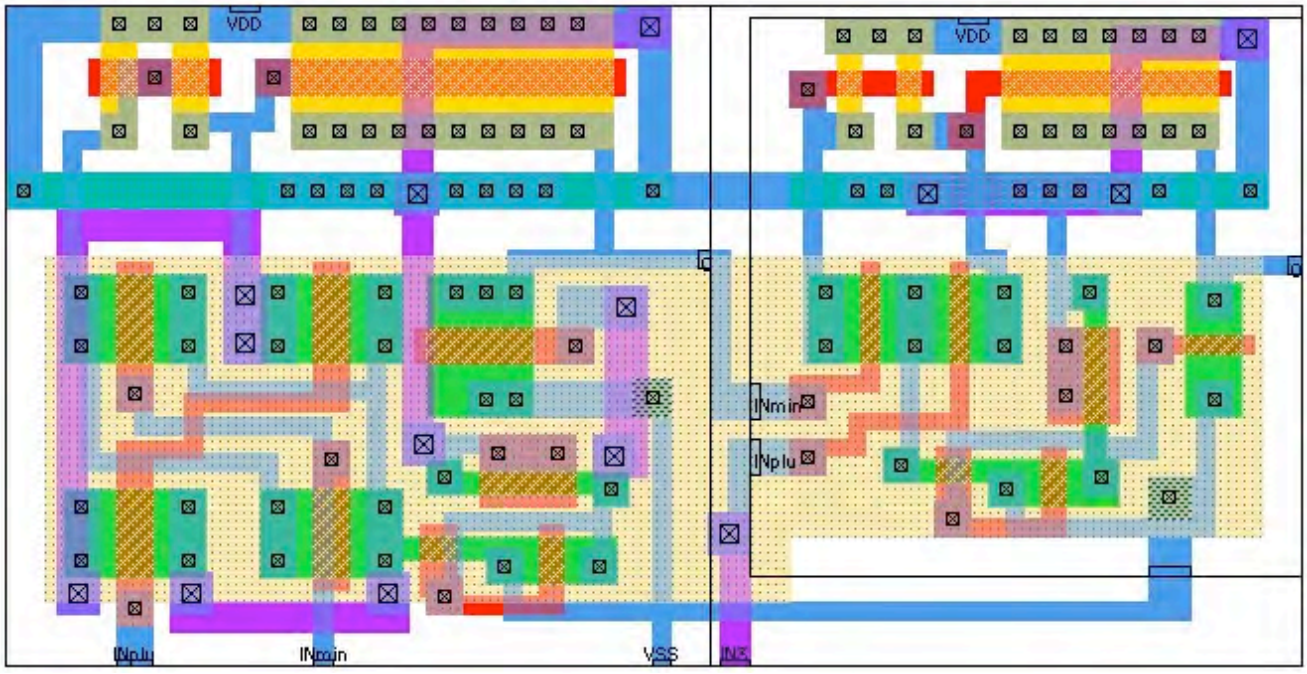
Zasadą powszechnie stosowaną w projektowaniu dużych układów jest **projektowanie hierarchiczne**. Z reguły każdy większy układ, a w tym praktycznie wszystkie układy cyfrowe, składa się z pewnej liczby bloków funkcjonalnych, każdy z tych bloków z kolei można podzielić na mniejsze i prostsze bloki funkcjonalne itd., aż dojdziemy do poziomu najprostszych bramek. Ta naturalna hierarchia jest widoczna nie tylko na poziomie architektury i schematu logicznego układu, ale znajduje także bezpośrednie odbicie w projekcie topografii. Każdy profesjonalny edytor topografii umożliwia projektowanie hierarchiczne, które polega na tym, że w projekcie topografii można używać wcześniej zaprojektowanych bloków jako obiektów wstawianych w całości i reprezentowanych przez prostokątne obrysy. W projekcie zorganizowanym hierarchicznie dopiero po zakończeniu dokonuje się **spłaszczenia hierarchii**, tj. topografię zorganizowaną hierarchicznie przekształca się na topografię, w której nie ma już hierarchicznie zorganizowanych bloków, lecz jedynie tranzystory i połączenia. Taka topografia jest potrzebna do wykonania końcowej kontroli reguł projektowania oraz do wytworzenia masek produkcyjnych.

Hierarchiczna organizacja projektu topografii nie tylko pozwala uzyskać bardziej przejrzysty i łatwiejszy do opanowania projekt dużego układu, ale także znacznie zmniejszyć jego pracochłonność, jeśli ten sam blok powtarza się w projekcie w wielu miejscach. Projektuje się go wówczas tylko raz, a następnie wstawia wszędzie gdzie trzeba. Jest to regułą w układach cyfrowych.

Rys. 6.6 i 6.7 przedstawiają prosty przykład topografii hierarchicznej. W jej skład wchodzą dwa kompletne bloki oraz połączenia między nimi.




Rys. 6.6. Topografia hierarchiczna z dwoma blokami oraz połączeniami




Rys. 6.7. Topografia hierarchiczna po spłaszczeniu

Ćwiczenia do tego wykładu nauczą Cię podstaw projektowania topografii układu w stylu *full custom* przy zastosowaniu reprezentacji półsymbolicznej.

## 6.1. Dodatek 1: Przykładowe reguły projektowania dla technologii CMOS

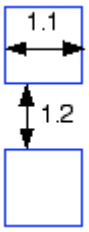
Technologia CMOS 0,8 μm, wartość lambda = 0,4 μm

Wszystkie reguły są wyrażone w umownych jednostkach lambda (powierzchnie w lambda$^2$). Zbiór reguł projektowania może być też stosowany do innych technologii CMOS pod warunkiem właściwego określenia wymiaru jednostki lambda w mikrometrach. Zazwyczaj lambda = 0,5 * $L_{min}$, gdzie $L_{min}$ jest minimalną długością bramki tranzystora MOS w danej technologii. Wartość odstępu równa 0 oznacza, że krawędzie obszarów mogą się stykać, ale nie przecinać.

Reguły mogą być stosowane w projektach wykonywanych przy użyciu programu "Microwind". Niektóre reguły nieistotne z punktu widzenia tego wykładu pominięto.

1. Reguły dla wyspy typu n

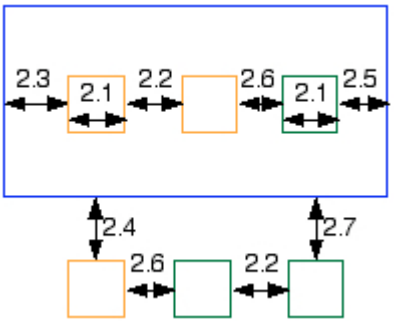
| Nr reguły | Opis | Wartość |
|---|---|---|
| 1.1 | Min. szerokość | 13 |
| 1.2 | Min. odstęp | 18 |
| 1.3 | Min. powierzchnia | 144 |




---

2. Reguły dla obszarów aktywnych (zwanych także obszarami dyfuzji) typu n (oznaczonych N+) i typu p (oznaczonych P+)

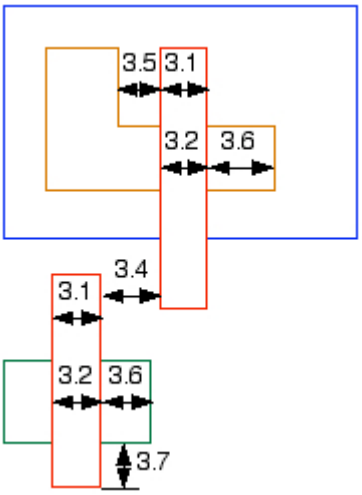
| Nr reguły | Opis | Wartość |
|---|---|---|
| 2.1 | Min. szerokość | 5 |
| 2.2 | Min. odstęp | 6 |
| 2.3 | Min. odstęp między obszarem aktywnym P+ na wyspie, a krawędzią wyspy | 8 |
| 2.4 | Min. odstęp między obszarem aktywnym N+ poza wyspą, a krawędzią wyspy | 6 |
| 2.5 | Min. odstęp między obszarem aktywnym N+ na wyspie, a krawędzią wyspy | 2 |
| 2.6 | Min. odstęp między obszarem aktywnym N+, a obszarem aktywnym P+ | 0 |
| 2.7 | Min. odstęp między obszarem aktywnym P+ poza wyspą, a krawędzią wyspy | 6 |
| 2.10 | Min. powierzchnia | 24 |




---

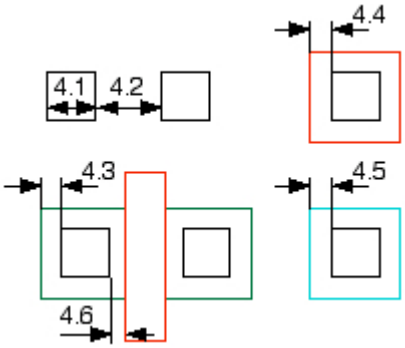
3. Reguły dla obszarów polikrzemu

| Nr reguły | Opis | Wartość |
|---|---|---|
| 3.1 | Min. szerokość | 2 |
| 3.2 | Min. szerokość nad obszarem aktywnym (dł. bramki tranzystora) | 2 |
| 3.4 | Min. odstęp | 3 |
| 3.5 | Min. odstęp polikrzemu od obszaru aktywnego | 2 |
| 3.6 | Min. zakładka obszaru aktywnego w stosunku do bramki tranzystora | 4 |
| 3.7 | Min. zakładka obszaru polikrzemu w stosunku do bramki tranzystora | 3 |
| 3.10 | Min. powierzchnia | 8 |



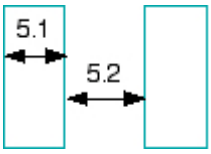

4. Reguły dla kontaktów

| Nr reguły | Opis | Wartość |
|---|---|---|
| 4.1 | Wymagany wymiar (uwaga: kontakt musi być kwadratem o podanej długości boku, inne kształty i wymiary nie są dozwolone) | 2 |
| 4.2 | Min. odstęp | 3 |
| 4.3 | Zakładka obszaru aktywnego wokół kontaktu | 2 |
| 4.4 | Zakładka polikrzemu wokół kontaktu | 2 |
| 4.5 | Zakładka metalu 1 wokół kontaktu | 2 |
| 4.6 | Min. odstęp od bramki tranzystora | 3 |



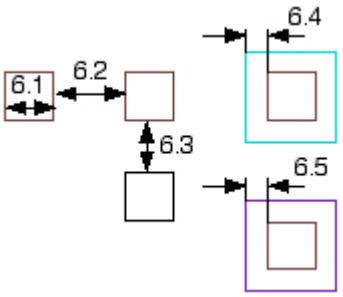

5. Reguły dla pierwszej warstwy metalu (metal 1)

| Nr reguły | Opis | Wartość |
|---|---|---|
| 5.1 | Min. szerokość | 3 |
| 5.2 | Min. odstęp | 3 |
| 5.3 | Min. powierzchnia | 16 |



6. Reguły dla kontaktów metal 1 - metal 2 (zwanych via)

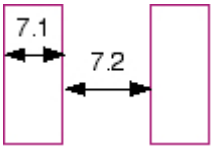
| Nr reguły | Opis | Wartość |
|---|---|---|
| 6.1 | Wymagany wymiar (uwaga: via musi być kwadratem o podanej długości boku, inne kształty i wymiary nie są dozwolone) | 3 |
| 6.2 | Min. odstęp | 3 |
| 6.3 | Min. odstęp od kontaktu (uwaga: kontakt i via mogą się nakładać) | 0 |
| 6.4 | Min. zakładka metalu 1 wokół via | 2 |
| 6.5 | Min. zakładka metalu 2 wokół via | 2 |




---

7. Reguły dla metalu 2

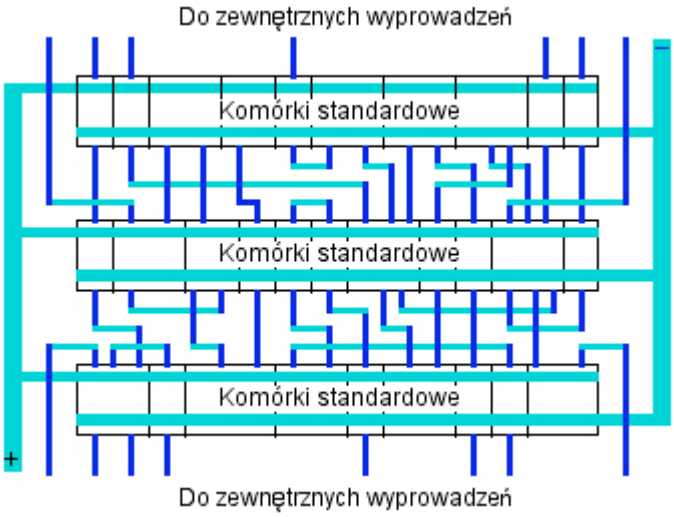
| Nr reguły | Opis | Wartość |
|---|---|---|
| 7.1 | Min. szerokość | 3 |
| 7.2 | Min. odstęp | 3 |
| 7.3 | Min. powierzchnia | 16 |



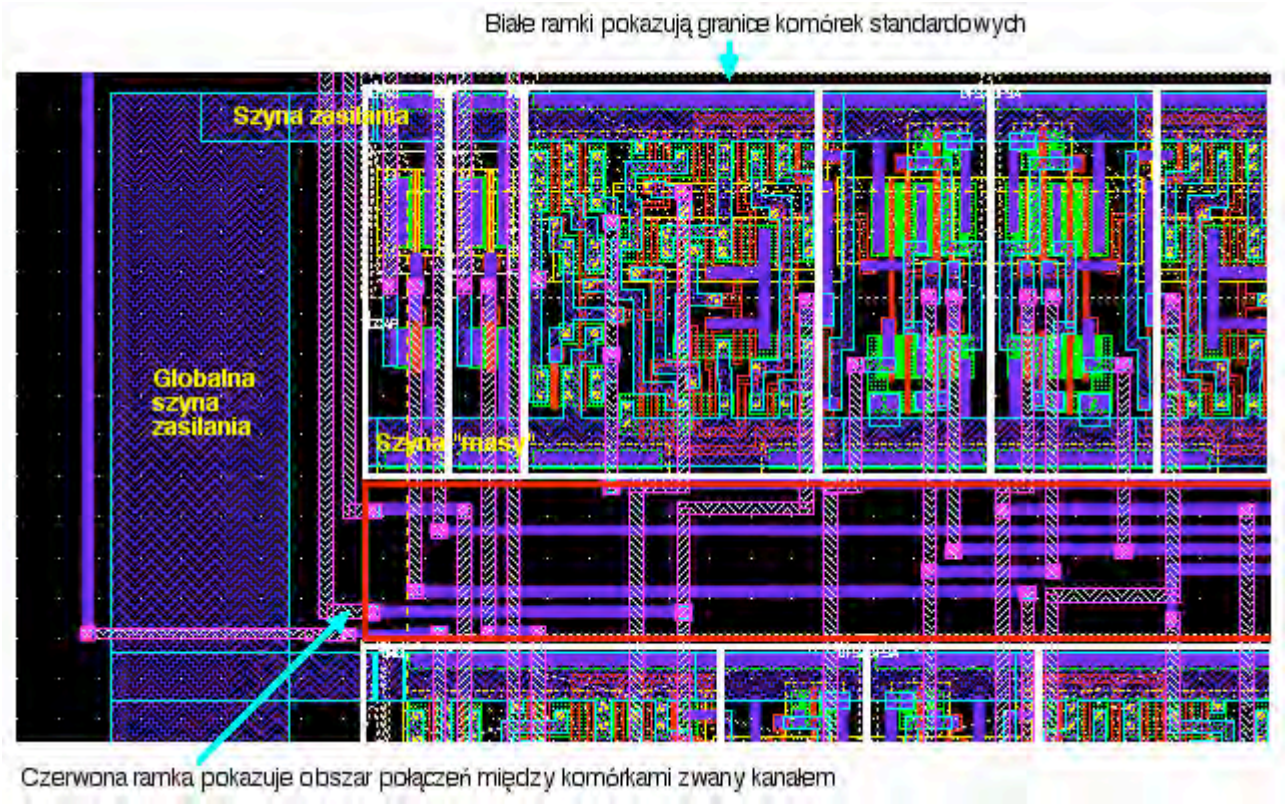
## 6.2. Style projektowania uproszczonego i zautomatyzowanego

Najbardziej pracochłonnym etapem projektowania układów scalonych jest projektowanie ich topografii. Ten etap jest też najtrudniejszy do zautomatyzowania. Podejmowane od wielu lat próby automatyzacji sposobu projektowania *full custom* nie dały jak dotąd wyników przydatnych dla praktyki inżynierskiej. Gdy zarówno rozmieszczenie elementów, jak i trasy połączeń między nimi są zupełnie dowolne, to żaden znany algorytm komputerowy - choć opracowano ich wiele - nie potrafi zastąpić intuicji i doświadczenia projektanta. Automatyzacja staje się możliwa, jeśli zrezygnujemy z całkowitej dowolności w rozmieszczeniu elementów i połączeń między nimi, i narzucimy pewne sztywne ograniczenia. Umożliwia to uproszczenie procesu projektowania i pozwala ten proces zautomatyzować. Omówimy teraz najważniejsze metody projektowania uproszczonego i zautomatyzowanego. Pozwalają one poważnie zmniejszyć pracochłonność projektu oraz ryzyko popełnienia błędów. Często w odniesieniu do nich używa się terminu "**style projektowania**".

Oczywistym sposobem zmniejszenia pracochłonności projektowania jest budowanie topografii układu nie z pojedynczych tranzystorów, lecz z wcześniej zaprojektowanych bloków o sprawdzonym działaniu, które można używać wielokrotnie w różnych projektach. Ta koncepcja doprowadziła do powstania stylu projektowania zwanego stylem **komórek standardowych** (ang. *standard cells* ). Komórka standardowa jest to blok funkcjonalny (cyfrowy lub analogowy) do wielokrotnego wykorzystania, którego topografia została zaprojektowana w specjalny sposób. Topografia każdej komórki mieści się w prostokącie, którego jeden wymiar (wysokość) jest standardowy - dla wszystkich komórek taki sam. Drugi wymiar jest dla różnych komórek różny, zależny od schematu elektrycznego komórki, liczby i wielkości elementów. Standardowe jest również rozmieszczenie doprowadzeń zasilania i masy. Są one realizowane jako szyny przechodzące przez całą szerokość komórki w jej dolnej i górnej części. Jeśli ustawić dwie komórki obok siebie w taki sposób, by ich pionowe krawędzie zetknęły się, szyny zasilania i masy zetkną się i połączą elektrycznie. Dzięki temu długi rząd komórek stykających się pionowymi krawędziami (które mają jednakową długość) ma gotowe połączenia masy i zasilania. Wejścia i wyjścia wyprowadzone są w kierunku poziomych krawędzi. Połączenia sygnałowe między komórkami ustawionymi w rzędy prowadzone są w obszarach pomiędzy rzędami, zwanych kanałami. Zazwyczaj w kanałach wszystkie odcinki pionowe połączeń prowadzone są w jednej warstwie metalu, a poziome - w drugiej. Jeśli dana technologia dysponuje większą liczbą warstw połączeń (co w technologiach najbardziej zaawansowanych jest regułą), to połączenia mogą być także prowadzone ponad obszarami komórek, co znacznie zmniejsza powierzchnię układu. Rys. 6.8 pokazuje schematycznie topografię układu zbudowanego z komórek standardowych, a rys. 6.9 przedstawia fragment projektu topografii rzeczywistego układu.




Rys. 6.8. Zasada budowy topografii układu z komórek standardowych




Rys. 6.9. Fragment topografii układu zbudowanego z komórek standardowych

Projektowanie topografii zbudowanej z komórek standardowych może odbywać się całkowicie automatycznie, i to jest wielka zaleta tego stylu projektowania. Projekt powstaje w trzech etapach:

- przydział komórek do rzędów,
- ustawienie kolejności komórek w rzędach,
- zaprojektowanie połączeń.

Każdy z tych trzech etapów można dość łatwo zalgorytmizować. Istnieje wiele systemów projektowania umożliwiających automatyczne zaprojektowanie topografii w stylu komórek standardowych, jeśli dany jest schemat logiczny układu. Schemat taki nie może być oczywiście zupełnie dowolny. Musi on być zbudowany wyłącznie z bramek logicznych, dla których istnieją odpowiednie komórki standardowe. Biblioteki komórek standardowych dostarcza każdy producent układów cyfrowych CMOS. Można oczywiście także zaprojektować samemu takie komórki, ale jest to pracochłonne, ponieważ odbywać się musi w stylu *full custom*. Ponadto komórki muszą być scharakteryzowane, tzn. każda z nich musi mieć znane parametry elektryczne, takie jak czasy opóźniena sygnału, pojemności wejściowe, pobór prądu itd.

Automatyczna synteza topografii układu w stylu komórek standardowych przy równoczesnym wykorzystaniu istniejących dziś metod automatycznej syntezy układów logicznych umożliwia zautomatyzowanie całego procesu projektowania. Najpierw na podstawie opisu behawioralnego układu wyrażonego w języku opisu sprzętu (Verilog, VHDL) wykonywana jest (w kilku etapach) automatyczna synteza schematu logicznego. Następnie schemat ten służy do automatycznego zaprojektowania topografii w stylu komórek standardowych. W ten sposób można uzyskać projekt układu przy minimalnym udziale człowieka. Oznacza to nie tylko olbrzymie zmniejszenie nakładów pracy, ale i wyeliminowanie możliwości popełnienia błędów, zgodnie z zasadą *correctness by construction*. Dzięki możliwości pełnej automatyzacji projektowania większość cyfrowych układów scalonych jest obecnie projektowania w całości lub w znacznej części w stylu komórek standardowych. Przy automatycznej syntezie układu punkt ciężkości pracy projektanta przenosi się na sporządzenie możliwie najlepszego opisu behawioralnego w języku opisu sprzętu. Tę samą funkcję układu można opisać na wiele sposobów. Jak już wiemy, nie każdy opis nadaje się do automatycznej syntezy układu nawet jeśli jest całkowicie prawidłowy. Ponadto jakość zaprojektowanego automatycznie układu w bardzo dużym stopniu zależy od umiejętnego sformułowania opisu jego funkcji. Jest tu pewna analogia z tradycyjnym programowaniem. Wiadomo, że jeden i ten sam algorytm można opisać w języku programowania na wiele sposobów, dających w rezultacie programy niekiedy bardzo różniące się pod względem sprawności obliczeniowej, dokładności itp. Niestety, w tym wykładzie nie ma miejsca na rozwijanie tego skądinąd bardzo ciekawego i ważnego tematu.

W bibliotekach komórek standardowych istnieją też komórki analogowe, toteż ten styl projektowania umożliwia zaprojektowanie niektórych układów analogowych lub mieszanych (analogowo-cyfrowych). Jednak układy analogowe są tak dalece zróżnicowane pod względem funkcji i wymagań technicznych, że w bardzo wielu

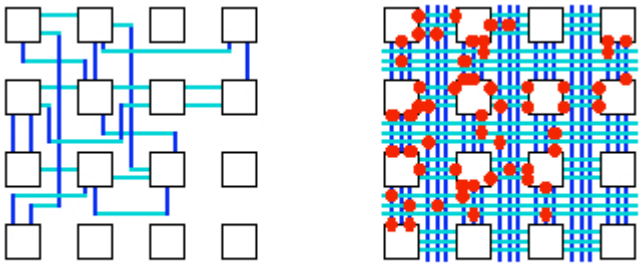
przypadkach nowy układ analogowy trzeba projektować całkowicie od początku, bowiem istniejące komórki nie nadają się do wykorzystania.

W porównaniu ze stylem *full custom* styl komórek standardowych ma też pewne ograniczenia i wady. Skończony zbiór komórek w bibliotece nie zawsze umożliwia pełną optymalizację projektu logicznego. Komórki projektuje się w taki sposób, aby były możliwie uniwersalne, tj. aby można było z nich zestawiać jak z klocków układy o dowolnych schematach logicznych. Takie uniwersalne komórki nie są zaprojektowane optymalnie z punktu widzenia szybkości działania, powierzchni czy też poboru mocy. Toteż spotyka się projekty, w których niewielkie, a krytyczne dla parametrów układu fragmenty są starannie zoptymalizowane i zaprojektowane w stylu *full custom*. Projekt w stylu komórek standardowych ma też z reguły większą powierzchnię, niż dobrze zrobiony projekt w stylu *full custom*. Doświadczenie pokazuje, że typowy projekt zbudowany z komórek standardowych da nam układ działający wolniej o kilkadziesiąt procent oraz zajmujący o kilkadziesiąt procent większą powierzchnię, niż ten sam układ zaprojektowany w stylu *full custom*. Jeśli jednak pamiętamy o olbrzymiej pracochłonności projektów w stylu *full custom*, staje się jasne dlaczego styl komórek standardowych stał się dominujący w projektowaniu układów cyfrowych.

Jeszcze dalej idącym uproszczeniem jest zastosowanie **matryc bramkowych** (ang. *gate arrays*). Matrycą bramkową nazywamy układ w postaci matrycy regularnie rozmieszczonych komórek (jednakowych lub różnych), połączonych siecią połączeń w taki sposób, by uzyskać zadany schemat. Zaletą matryc bramkowych jest to, że projektuje się tylko połączenia. Nie projektuje się rozmieszczenia komórek w matrycy. Oznacza to dalsze zmniejszenie pracochłonności i ryzyka popełnienia omyłek. Co więcej, układy matryc bramkowych można produkować bez połączeń i takie półfabrykaty przechowywać. Gdy powstanie projekt konkretnego układu, wykonuje się odpowiednie do tego projektu maski ścieżek metalu i kontaktów, i wykonuje operacje wytwarzania połączeń na wcześniej wyprodukowanych matrycach. Dzięki temu bardzo skraca się nie tylko czas potrzebny na zaprojektowanie układu, ale i czas oczekiwania na prototypowe układy oraz koszt prototypów. Podobnie jak w przypadku komórek standardowych, układ przeznaczony do wykonania przy użyciu matrycy bramkowej może być zaprojektowany całkowicie automatycznie na podstawie opisu behawioralnego.

Matryce bramkowe miały okres sporej popularności, ale zostały w znacznym stopniu wyparte przez **układy programowalne** , które były już wspominane wcześniej. Układ programowalny to taki rodzaj matrycy bramkowej, w której sieć połączeń określa użytkownik przez zaprogramowanie układu. Matryca zawiera nie tylko komórki, ale i wstępnie poprowadzoną sieć połączeń, która jednak wymaga wykonania kontaktów. Kontakty te służą do połączenia wejść i wyjść komórek ze ścieżkami oraz ścieżek między sobą. W ten sposób powstaje gotowy układ. Kontakty są programowalne, tj. powstają w drodze elektrycznego zaprogramowania układu. Istnieje kilka sposobów wytworzenia programowalnego kontaktu, będzie o tym mowa w jednym z dalszych wykładów.

Rys. 6.10 pokazuje schematycznie oba rodzaje matryc bramkowych.




Rys. 6.10. Zasada budowy matryc bramkowych: (a) zwykłej, (b) programowalnej.

Czerwone punkty oznaczają kontakty, które zostały zaprogramowane w matrycy programowalnej. Tam, gdzie nie ma takiego punktu, nie ma połączenia elektrycznego.

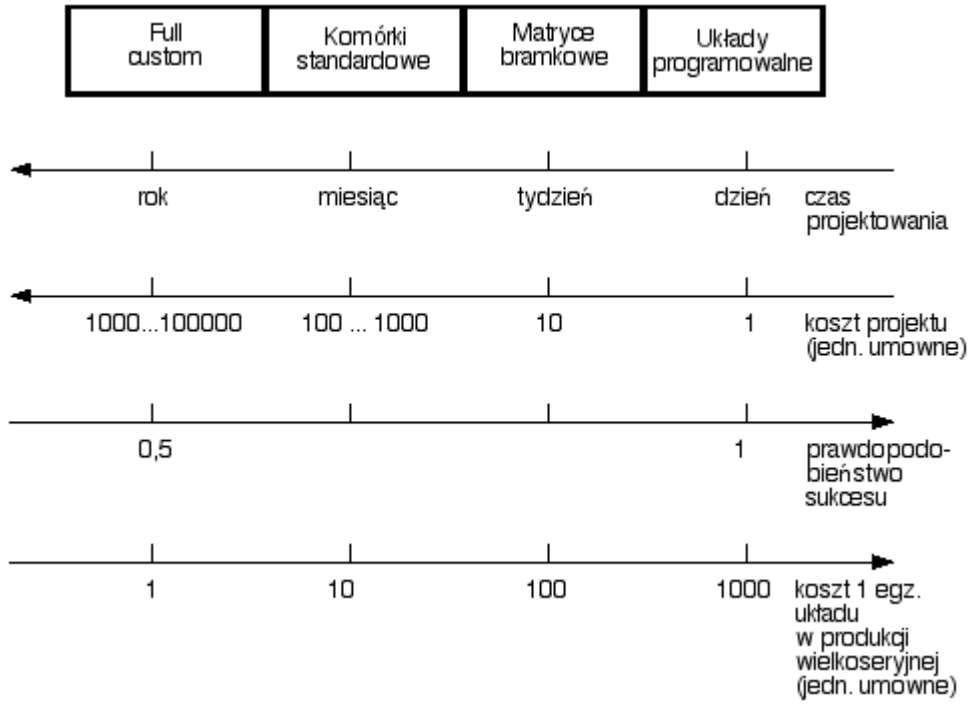
Programowalne matryce bramkowe znane powszechnie jako układy FPGA (od angielskiego "Field Programmable Logic Array") zdobyły w ostatnich latach ogromną popularność, ponieważ przy ich użyciu działające układy otrzymuje się niemal natychmiast. Opis behawioralny układu poddawany jest syntezie logicznej, po czym generowane są dane dla programatora, który programuje połączenia w matrycy podając odpowiedni ciąg sygnałów na specjalne wejścia programujące układu FPGA.

W nowoczesnych układach FPGA programowalne mogą być nie tylko połączenia, ale i funkcje logiczne samych komórek w matrycy, co daje tym układom bardzo dużą elastyczność. Oprócz matryc prostych komórek najbardziej zaawansowane układy FPGA zawierają także większe bloki funkcjonalne, aż do rdzeni mikroprocesorów i bloków pamięci włącznie. Produkowane są także programowalne układy analogowo-cyfrowe, zawierające matryce bramek cyfrowych i bardziej złożone bloki, a równocześnie typowe bloki analogowe, jak np. wzmacniacze operacyjne, oraz przetworniki analogowo-cyfrowe i cyfrowo-analogowe. Umożliwia to łatwe realizacje układów

typu "System on chip", stąd ten rodzaj układów programowalnych znany jest pod nazwą "**PSoC**[1]" ("Programmable system on chip").

Wadą wszelkich matryc bramkowych i układów programowalnych jest "marnotrawstwo krzemu" - zazwyczaj wykorzystane jest od 30% do 80% bramek w matrycy. Toteż realizacja układów w postaci matryc bramkowych jest stosunkowo kosztowna. Dotyczy to zwłaszcza matryc programowalnych FPGA, w których część powierzchni zajmują połączenia (nigdy nie wykorzystane w 100%) oraz układy pomocnicze umożliwiające programowanie. Układy FPGA są doskonałym rozwiązaniem, gdy trzeba szybko zbudować prototyp projektowanego urządzenia, lub gdy potrzebne są pojedyncze egzemplarze układów, lub gdy koszt układów FPGA jest znikomym procentem kosztu całego urządzenia.

Jak widzimy, można dobrać najwłaściwszy styl projektowania do każdego zadania technicznego. Rys. 6.11 stanowi podsumowanie - klasyfikuje omówione wyżej style projektowania z kilku punktów widzenia: czasu projektowania (rozumianego jako czas, jaki musi upłynąć od rozpoczęcia projektowania do otrzymania pierwszego egzemplarza układu), kosztu projektu (proporcjonalnego do pracochłonności), prawdopodobieństwa sukcesu (rozumianego jako prawdopodobieństwo, że pierwsze wyprodukowane egzemplarze układu będą spełniać wszystkie wymagania techniczne) oraz kosztu jednego egzemplarza układu przy produkcji wielkoseryjnej. Wartości liczbowe na osiach należy oczywiście traktować jedynie jako czysto orientacyjne, służące do porównań z dokładnością do rzędu wielkości.




Rys. 6.11. Porównanie stylów projektowania

Biorąc pod uwagę zalety i wady wszystkich omówionych stylów projektowania można wskazać następujące ich główne obszary zastosowań:

- **styl *full custom*: układy analogowe i bloki analogowe układów analogowo-cyfrowych, topografie komórek standardowych, krytyczne z punktu widzenia parametrów fragmenty układów cyfrowych, niezbyt wielkie układy przeznaczone do produkcji masowej,**
- **styl komórek standardowych: większość układów cyfrowych przeznaczonych do produkcji w małych, średnich i długich seriach, niektóre układy analogowo-cyfrowe,**
- **styl matryc bramkowych: dziś coraz rzadziej stosowany**
- **układy programowalne (FPGA): układy cyfrowe potrzebne w niewielkiej liczbie egzemplarzy, prototypy do szybkiego sprawdzenia konstrukcji urządzenia, układy cyfrowe potrzebne w dłuższej serii wtedy, gdy ich koszt jest bez znaczenia.**

Postępy technologii umożliwiają wytwarzanie coraz większych i bardziej złożonych układów, i nawet wykorzystanie wymienionych wyżej sposobów projektowania uproszczonego i zautomatyzowanego nie wystarcza,

gdy układ ma wiele milionów elementów. Jednym z rozwiązań problemu jest budowa wielkich systemów scalonych z wykorzystaniem gotowych dużych i złożonych bloków funkcjonalnych takich, jak rdzenie mikroprocesorów, standardowe układy peryferyjne i komunikacyjne itp. Rozwinął się rynek projektów takich bloków, można takie projekty zamawiać, kupować i sprzedawać. Taki produkt może występować w dwóch postaciach: syntezowalnego kodu w języku opisu sprzętu (można go wtedy użyć w układach wytwarzanych w różnych technologiach) lub gotowego projektu topografii dla konkretnej technologii. Jest to rynek myśli technicznej w czystej postaci, a produkty występujące na tym rynku noszą ogólną nazwę **bloków IP**, od angielskiego terminu **intellectual property** - własność intelektualna. Istnieją firmy (również w Polsce), których jedynym produktem są bloki IP.

---

[1]PSoC - nazwa zastrzeżona firmy Cypress Semiconductor

# ĆWICZENIE 1 DO WYKŁADU 6

**Cel ćwiczenia**

W tym ćwiczeniu zapoznasz się z programem "Microwind" służącym do nauki projektowania układów scalonych, oraz zaczniesz opanowywać podstawy projektowania w stylu *full custom* projektując pojedynczy tranzystor.

**1. Instalacja programu**

Przed rozpoczęciem ćwiczenia musisz zainstalować program "Microwind" na swoim komputerze. Instalacja jest bardzo prosta - skopiuj cały katalog "Microwind2" do katalogu "Program Files". W katalogu "Microwind2" znajdziesz plik "Microwind2.exe" - to jest właśnie program. Możesz dla wygody zrobic skrót i umieścić go na pulpicie lub dodać program do menu startowego. W katalogu "Microwind2" znajdziesz także kilkaset plików. Są to m.in. pliki technologiczne definiujące różne technologie dla programu "Microwind2" (rozszerzenie ".rul"), pliki z przykładowymi projektami (rozszerzenie ".msk") i inne, oraz katalog "Html".

**2. Obsługa programu**

W kolejnych ćwiczeniach otrzymasz szczegółowe wskazówki, jak wykonywać poszczególne czynności. Nie będą jednak omówione wszystkie możliwości i opcje programu.

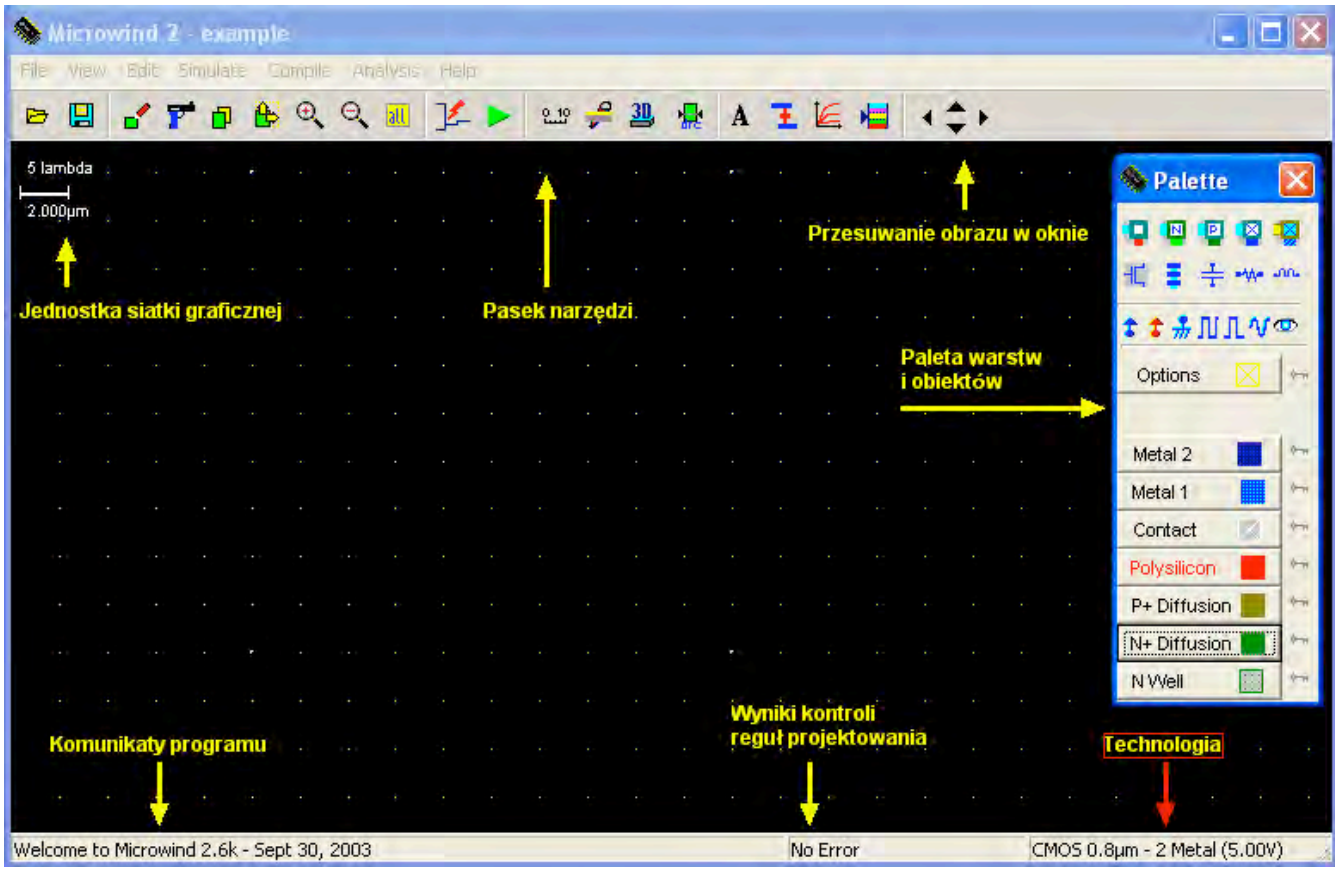
W katalogu "Microwind2\Html" znajdziesz zwięzłą instrukcję obsługi programu (w języku angielskim) w postaci stron w języku HTML do przeglądania dowolną przeglądarką. Aby je przejrzeć, otwórz plik "index.htm".

Na płycie znajdziesz też książkę: Etienne Sicard, "Microwind and Dsch User's Manual" w postaci pliku PDF. Jeśli dostatecznie dobrze znasz język angielski, przejrzyj ją choćby pobieżnie. W książce opisane są nie tylko programy "Microwind" i "Dsch" (ten ostatni poznasz nieco później), ale i zawarta jest wielka liczba dobrze opisanych przykładów stanowiących w sumie znakomite wprowadzenie do mikroelektroniki.

Możesz również zapoznać się z wprowadzeniem do ćwiczenia w wersji wideo, gdzie wykładowca opowie Ci i pokaże na ekranie, jak wykonywać poszczególne czynności. Wprowadzenie to znajdziesz w sekcji "Wykłady VIDEO".

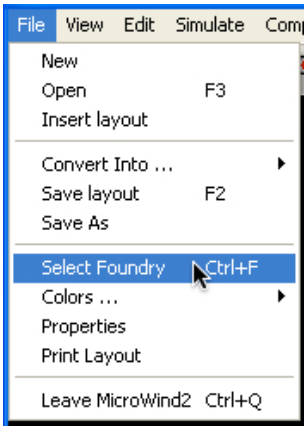
**3. Przebieg ćwiczenia**

Uruchom program "Microwind2". Zobaczysz okno jak niżej. Na ilustracji kolorem żółtym opisano najważniejsze widoczne obiekty.

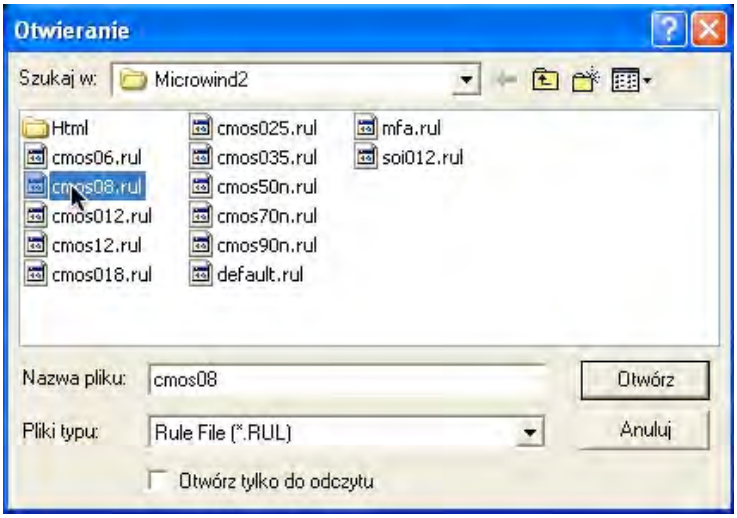



W tym i następnych ćwiczeniach będziemy wykorzystywać prostą technologię CMOS z minimalną długością bramki 0,8 mikrometra. W prawym dolnym rogu widnieje nazwa wczytanej przez program technologii. Przed rozpoczęciem projektowania zawsze sprawdź, czy jest to technologia CMOS 0.8 µm. Jeśli nie, trzeba najpierw wczytać odpowiedni plik technologiczny.

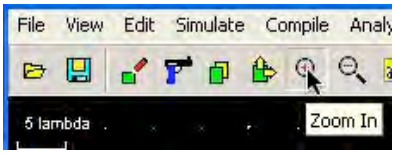
Z menu "File" wybierz "Select Foundry". Otrzymasz na ekranie typowe okno wyboru plików.




Otwórz plik "cmos08.rul".




Dla wygody rysowania zmień skalę w oknie wybierając z paska narzędzi ikonę powiększenia.

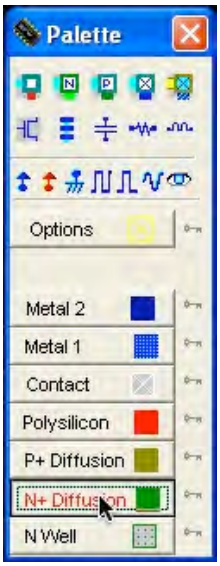



Zapoznaj się z najważniejszymi regułami projektowania - wybierz "Design Rules" z menu "Help". Otrzymasz na ekranie tabelę z podstawowymi regułami projektowania. Są w niej też inne dane, na razie nieistotne.

Możesz także otworzyć stronę zawierającą wszystkie potrzebne reguły.

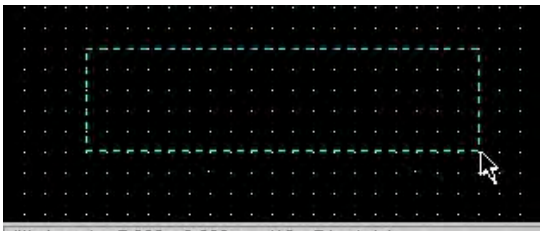

Zaczynasz rysowanie n-kanałowego tranzystora MOS.

Wybierz z palety "N+ diffusion". Jest to warstwa abstrakcyjna oznaczająca obszar aktywny typu n. Nazwa warstwy, która jest w danej chwili wybrana, jest w palecie oznaczona kolorem czerwonym.



Narysujesz teraz prostokąt na warstwie "N+ diffusion".

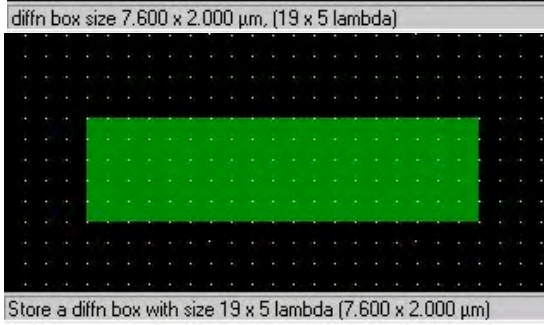
Wybierz ikonę rysowania prostokąta z paska narzędzi.

Ustaw kursor w pobliżu wybranego węzła siatki, naciśnij lewy klawisz myszki i ciągnij aż do otrzymania prostokąta o potrzebnych wymiarach, następnie puść klawisz.

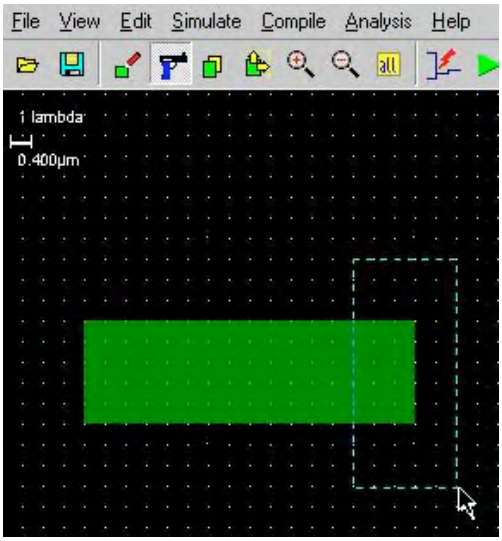
Prostokąt o krawędziach zaznaczonych przerywaną kreską, który pokazuje prostokąt do narysowania, będziemy nazywali **selektorem**.

Zauważ, że nawet jeśli nie trafiasz dokładnie w węzły siatki, narysowany będzie prostokąt o wymiarach będących całkowitą wielokrotnością jednostki lambda. Staraj się narysować prostokąt o szerokości 5 lambda i długości zbliżonej do pokazanej na ilustracji z prawej strony. Jeśli Ci się nie uda, możesz wybrać "Undo" z menu "Edit" i zacząć jeszcze raz.




Jeśli narysujesz prostokąt zbyt długi lub szeroki, możesz usunąć jego fragment.

Wybierz ikonę usuwania (pistolet) z paska narzędzi.

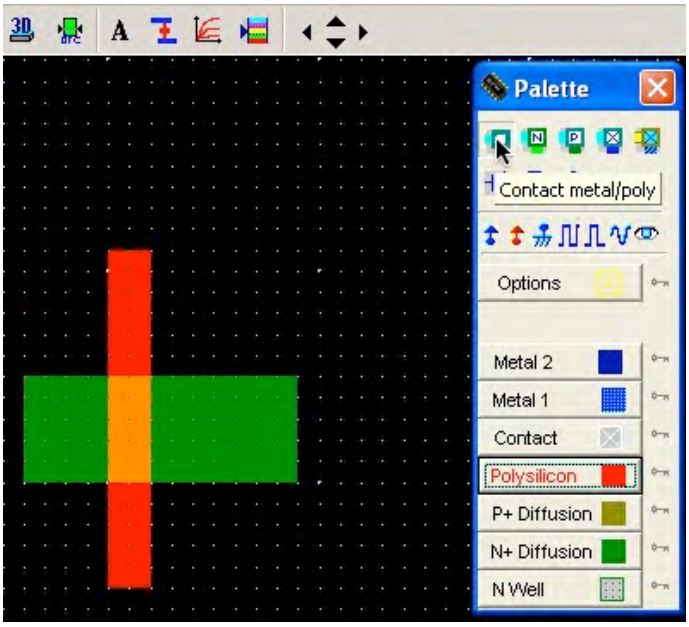

Następnie postępuj tak jak przy rysowaniu.

Gdy puścisz klawisz myszki, wnętrze selektora zostanie wymazane.

Narysujesz teraz prostokąt na warstwie "Polysilicon" (polikrzem), a potem dorysujesz do niego kontakt do warstwy "Metal 1".

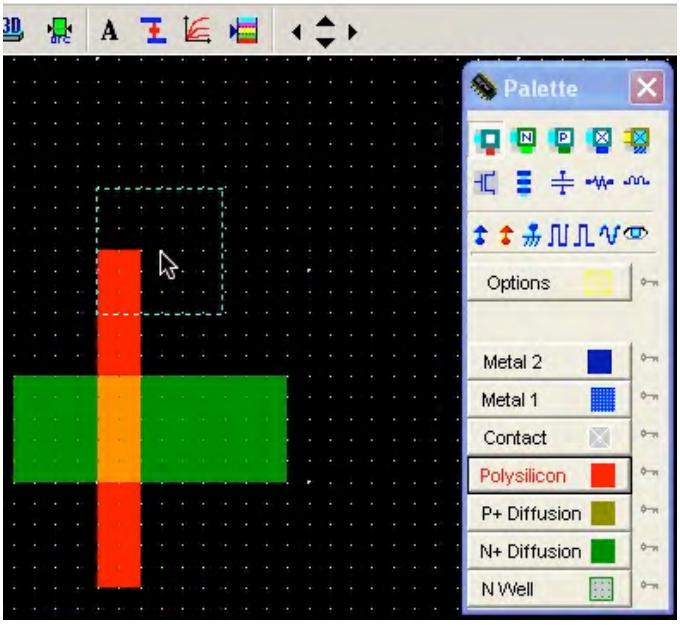
Wybierz ikonę rysowania prostokąta z paska narzędzi. Następnie wybierz "Polysilicon" z palety i narysuj pionowy pasek polikrzemu o szerokości 2 lambda przecinający obszar aktywny.

Następnie wybierz z palety obiekt "Contact metal/poly".

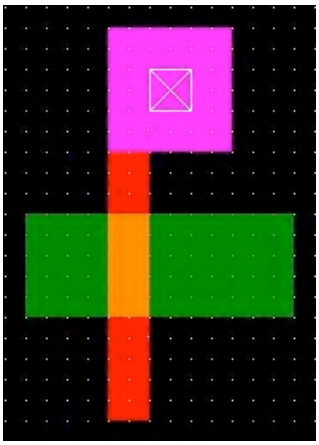


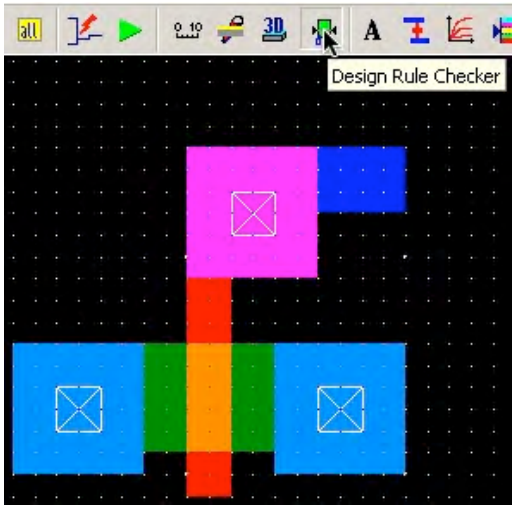

Kontakt jest obiektem zdefiniowanym w pliku technologicznym, ma wymagany kształt kwadratu i zawiera wszystkie potrzebne warstwy: polikrzem, okno kontaktowe i metal 1.

Ciągnąc myszką kwadrat selektora umieść kontakt tak jak na ilustracji i puść klawisz myszki.




Obok wynik tej operacji.






Projektuj dalej tranzystor.

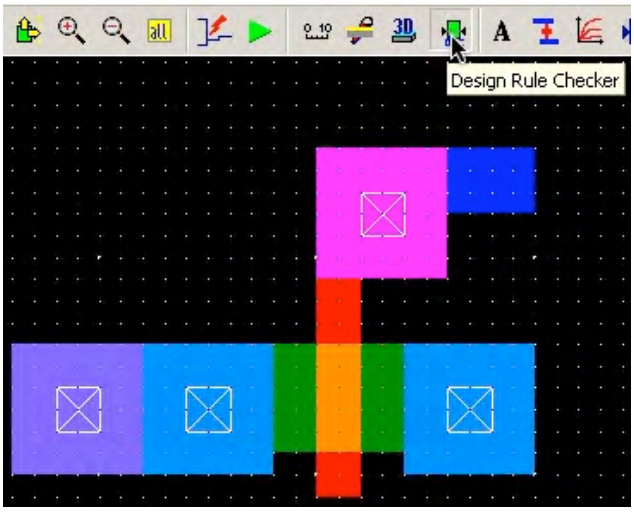
Postępując tak, jak poprzednio, dodaj kontakty do obszaru aktywnego (czyli źródła i drenu tranzystora) oraz pasek metalu 1 do kontaktu bramki. Pasek polikrzemu możesz skrócić, wystarczy że wystaje o 2 lambda poza obszar kanału tranzystora. Następnie wybierz z paska narzędzi ikonę kontroli reguł projektowania.

Jeśli wykonany projekt topografii wygląda tak, jak obok, otrzymasz komunikat o braku błędów.


Twój tranzystor jest w zasadzie gotowy, ale trzeba jeszcze coś dodać. W układach CMOS podłoże musi być dokładnie uziemione (dlaczego? - to było opisane w wykładzie 4, zobacz rys. 4.13 - 4.15 i tekst do niech się odnoszący). Aby móc uziemić podłoże, trzeba wykonać do niego kontakt. Podłoże jest półprzewodnikiem typu p, należy użyć kontaktu między metalem, a obszarem typu p.

Wybierz z palety obiekt "Contact P+diff/Metal1" i postępując tak, jak przy umieszczaniu poprzednich kontaktów, umieść kontakt tak, by sąsiadował z obszarem źródła tranzystora.

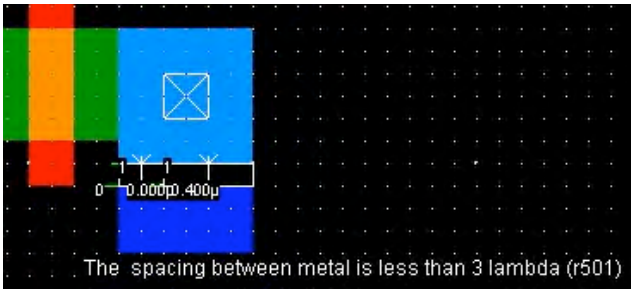
Kontakt zawiera wszytkie potrzebne warstwy, w tym warstwę metalu 1. Umieszczenie go tak, by sąsiadował z obszarem źródła tranzystora, oznacza że w gotowym układzie źródło będzie elektrycznie połączone z kontaktem, a ponieważ kontakt będzie uziemiony, tj. połączony z "minusem" zasilania, to uziemione będzie też źródło tranzystora. Oczywiście nie zawsze tak musi być, w układzie zawierającym wiele tranzystorów nMOS tylko niektóre będą miały źródła połączone z minusem zasilania. Ale w każdym układzie musi być przynajmniej jeden uziemiony kontakt do podłoża; w każdym większym układzie takich kontaktów musi być wiele - patrz wykład 4.




Jeśli wszystko zostało wykonane poprawnie, twój tranzystor wraz z kontaktem do podłoża powinien wyglądać jak obok.

Wykonaj jeszcze raz kontrolę reguł projektowania.

Powinien ukazać się kontakt o braku błędów.




Teraz zobacz, co się stanie, jeśli w projekcie będzie błąd.

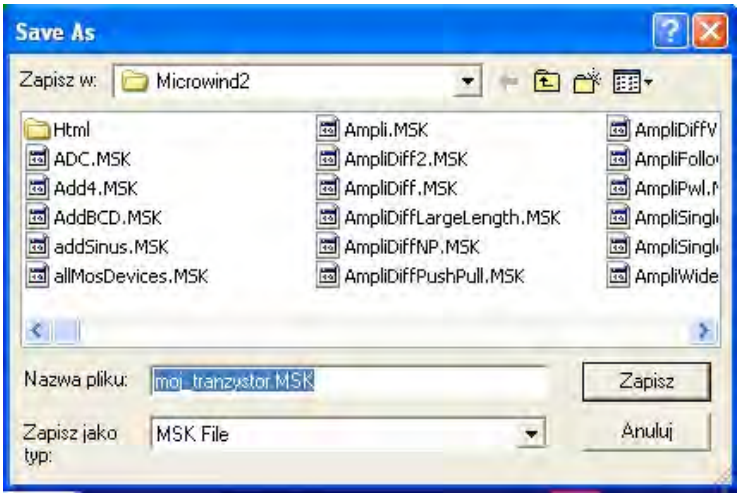
Dorysuj pasek metalu 1 w odległości 1 lambda od innego obszaru metalu 1, a następnie wybierz z paska narzędzi ikonę kontroli reguł projektowania.

Obok widzisz wynik: komunikat o błędzie.

Przed dalszymi czynnościami usuń dorysowany pasek tak, aby pozostał prawidłowy projekt.

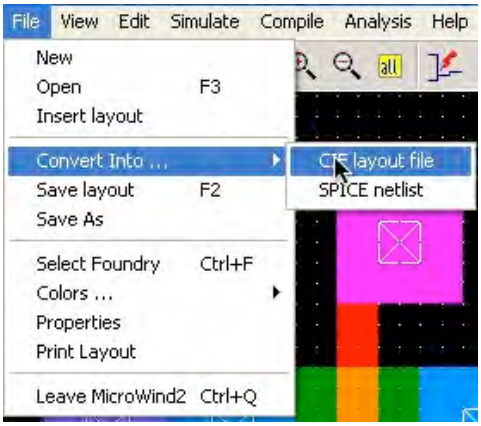
Jeśli wszystko jest w porządku, zapisz swój pierwszy projekt na dysku. Będzie potrzebny w dalszych ćwiczeniach!

Wybierz "Save As" z menu "File" i zapisz projekt pod nową nazwą, np. "moj_tranzystor".



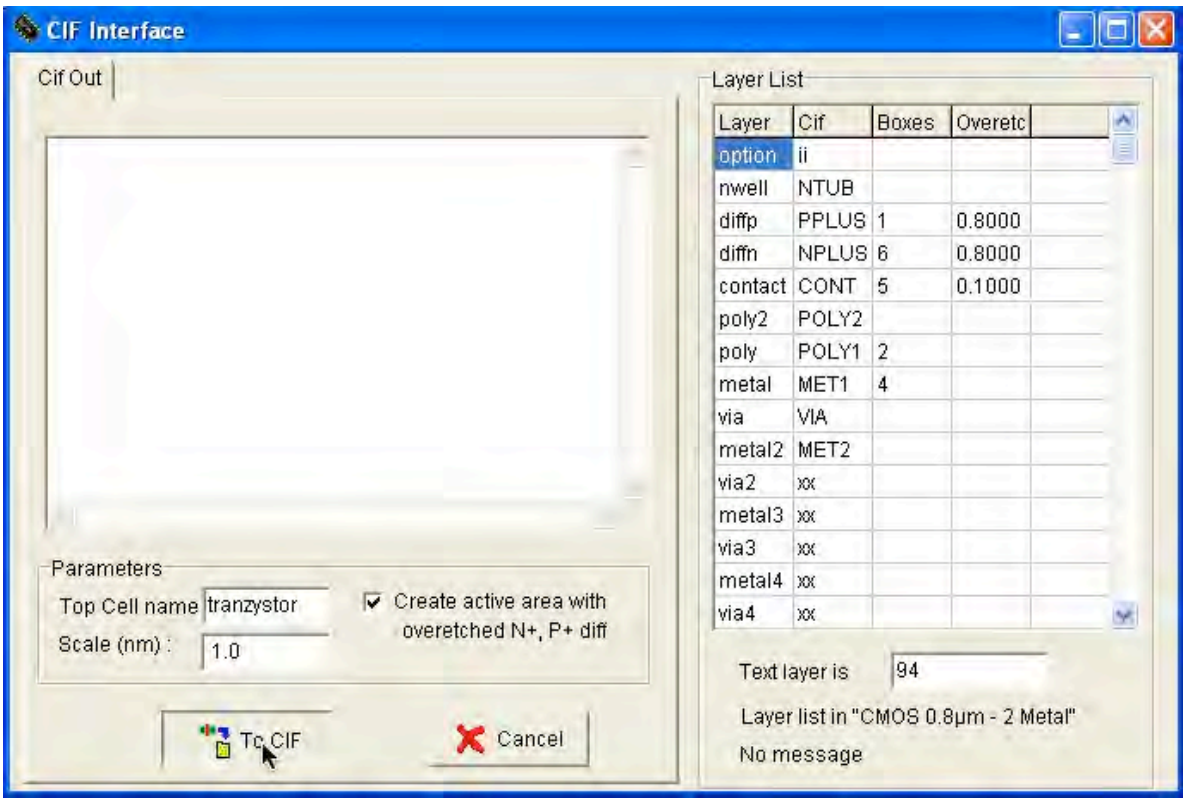

Zapisany uprzednio plik (z rozszerzeniem "msk") zawiera opis Twojego projektu w wewnętrznym formacie programu "Microwind". Teraz możesz jeszcze zapisać projekt w standardowym formacie języka CIF i zobaczyć, jak taki zapis wygląda.

Wybierz "Convert Into..." -> "CIF layout file" z menu "File". Otrzymasz na ekranie tablicę, w której nie musisz zmieniać niczego poza dopisaniem nazwy zaprojektowanej komórki (topcell).

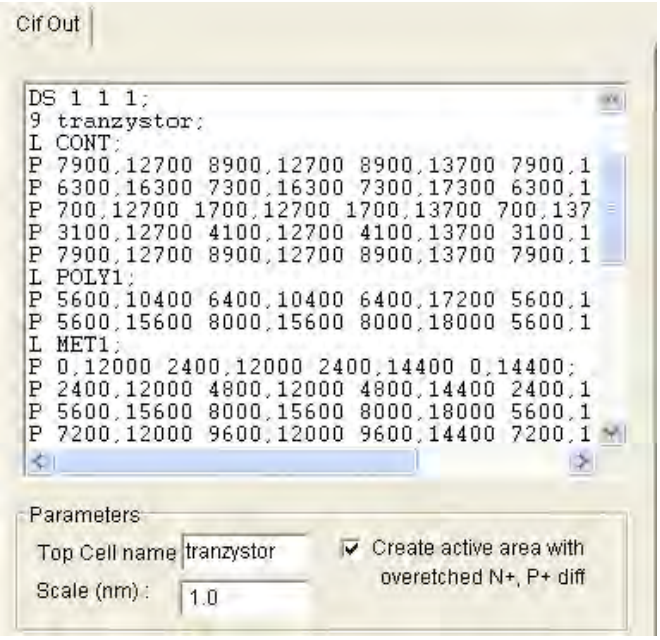



Z prawej strony tablica pokazuje jakie maski zostaną zapisane w pliku w formacie CIF. Podczas zapisu dokonywana będzie konwersja z warstw abstrakcyjnych połączona w przypadku niektórych masek ze zmianami wymiarów. Są to nieistotne dla nas w tym momencie szczegóły technologiczne.

Wpisz nazwę komórki (np. "tranzystor") w polu "Top Cell name" i kliknij "To CIF".




Projektowanie zakończone! Możesz teraz jeszcze obejrzeć zawartość przykładowego pliku CIF (Twój może w szczegółach wyglądać nieco inaczej).




# ĆWICZENIE 2 DO WYKŁADU 6

**Cel ćwiczenia**

Samodzielny trening

**Przebieg ćwiczenia (do samodzielnego wykonania)**

Narysuj topografię tranzystora MOS p-kanałowego analogiczną do narysowanej w ćwiczeniu 1 topografii tranzystora MOS n-kanałowego. Zachowaj te same wymiary kanału tranzystora. Różnice będą następujące:

- Tranzystor p-kanałowy musi być na wyspie typu n ("N Well").
- Zamiast obszaru aktywnego typu n należy użyć obszaru aktywnego typu p ("P+ diffusion").
- Wyspa oprócz tranzystora musi zawierać kontakt, który w gotowym układzie będzie podączony do "plusa" zasilania. Wyspa jest obszarem typu n, więc użyj obiektu "Contact N+diff/metal1".

Nie zapomnij po zakończeniu rysowania sprawdzić, czy spełnione są reguły projektowania!

Zapisz Twój pierwszy samodzielny projekt na dysk. Będzie potrzebny w dalszych ćwiczeniach!

## Bibliografia

[1] E. Sicard, "*Microwind & Dsch User's Manual*", National Institute of Applied Sciences INSA, Toulouse, 2003
(Jest to podręcznik do oprogramowania wykorzystywanego w tym wykładzie, załączony na płycie w postaci pliku PDF)
[2] S. M. Rubin, "*Computer Aids for VLSI Design*", Addison-Wesley Publishing Co., Inc. 1987
(Książka dostępna w Internecie, wraz z bezpłatnym oprogramowaniem do projektowania układów scalonych - system "Electric")
[3] B. Preas, M. Lorenzetti, "*Physical Design Automation of VLSI Systems"*, Benjamin/Cummings Publishing Co., Inc. 1988
(Książka omawia metody i algorytmy automatyzacji projektowania układów scalonych)

# Wykład 7: Statyczne bramki kombinacyjne CMOS

## Wstęp

Wykład 7 zaczyna się od omówienia podstawowych wymagań dla bramek logicznych. Następnie omawiana jest najprostsza bramka - inwerter. Omówienie jej parametrów (napięcie przełączania, marginesy zakłóceń, czasy przełączania, pobór mocy) przygotowuje do projektowania. Znając zasady projektowania inwertera można je następnie uogólnić na bramki wykonujące bardziej złożone funkcje kombinacyjne.

Materiał wykładu 7 jest bardzo ważny, ponieważ kombinacyjne bloki logiczne realizowane w technologii CMOS budowane są w zdecydowanej większości przypadków właśnie z bramek statycznych, takich jak omawiane w tym wykładzie. Wykład jest dłuższy od poprzednich i zawiera bardzo obszerny materiał. Poświęć mu dostatecznie dużo czasu!

Ważnym uzupełnieniem wykładu są zadania rachunkowe i ćwiczenia. Zadania pokażą Ci, jak wykonywać proste obliczenia projektowe. Praktyczne ćwiczenia nauczą Cię, jak zaprojektować topografię prostej bramki i wykonać jej symulację elektryczną. Wykonanie ćwiczeń będzie łatwiejsze, gdy obejrzysz prezentacje wideo.

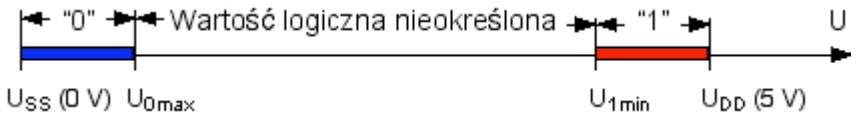
## 7.1. Statyczne bramki logiczne CMOS - podstawy

Omówimy teraz podstawowe wymagania, jakie powinny spełniać bramki logiczne, oraz główne parametry charakteryzujące te bramki. W tym wykładzie mówimy o bramkach kombinacyjnych. Bramki realizujące funkcje pamięciowe (przerzutniki, rejestry, komórki pamięci) będą omawiane dalej.

Zdefiniujemy na początek bramkę statyczną. Takie właśnie bramki są stosowane w zdecydowanej większości układów cyfrowych CMOS do budowy bloków logiki kombinacyjnej.

! **Bramka statyczna jest to bramka mająca tę własność, że jak długo włączone jest napięcie zasilania, a stany logiczne na wejściach nie ulegają zmianie, to i stany logiczne na wyjściach nie zmieniają się.**

W większości układów cyfrowych wartości logiczne zera i jedynki są reprezentowane przez dwa różne napięcia zwane **poziomami logicznymi** zera i jedynki (w dalszej części wykładu wartości logiczne zera i jedynki będą oznaczane symbolami "0" i "1" w cudzysłowach, by nie myliły się ze zwykłymi liczbami). W układach CMOS powszechnie przyjmuje się, że wartość logiczna "0" jest reprezentowana przez napięcie równe zeru, a wartość logiczna "1" jest reprezentowana przez napięcie równe napięciu zasilania układu (w dalszej części wykładu napięcie zasilania będzie oznaczane symbolem $U_{DD}$, a napięcie równe zeru będzie także niekiedy oznaczane symbolem $U_{SS}$). Jednak taka definicja nie wystarcza. Z różnych powodów logiczne "0" może być reprezentowane w układzie przez napięcie bliskie zeru, ale nieco od zera wyższe, zaś logiczna "1" może być reprezentowana przez napięcie nieco niższe od napięcia zasilania. Dlatego definiuje się zakresy wartości napięć, w których mieszczą się napięcia reprezentujące logiczne "0" i logiczną "1".




Rys. 7.1. Definicja poziomów logicznych (przykład dla napięcia zasilania równego 5 V)

Omówimy teraz najważniejsze cechy i właściwości bramek cyfrowych. Są to:

- zdolność do regeneracji poziomów logicznych,
- zdolność do tłumienia zakłóceń,
- szybkość działania,
- pobór mocy,
- kierunkowość.

Zdolność bramek logicznych do **regeneracji** poziomów logicznych wiąże się z definicją tych poziomów. Zdolność tę można określić następująco.

! **Mówimy, że bramka logiczna regeneruje poziom logiczny "0", jeżeli dla dowolnej kombinacji stanów logicznych na wejściu reprezentowanych przez napięcia *na granicach* odpowiednich zakresów (tj. $U_{0max}$ dla stanów "0" i $U_{1min}$ dla stanów "1"), dla których na wyjściu bramki mamy stan "0", stan ten jest reprezentowany przez napięcie *wewnątrz* zakresu zera, tj. napięcie U spełniające warunek $U < U_{0max}$. Podobnie, mówimy, że bramka logiczna regeneruje poziom logiczny "1", jeżeli dla dowolnej kombinacji stanów logicznych na wejściu reprezentowanych przez napięcia *na granicach* odpowiednich zakresów (tj. $U_{0max}$ dla stanów "0" i $U_{1min}$ dla stanów "1"), dla których na wyjściu bramki mamy stan "1", stan ten jest reprezentowany przez napięcie *wewnątrz* zakresu jedynki, tj. napięcie U spełniające warunek $U > U_{1min}$.**

Intuicyjnie zdolność do regeneracji można rozumieć jako zdolność do "poprawiania" napięć reprezentujących stany logiczne, tak aby napięcia te nie mogły się znaleźć poza dopuszczalnymi przedziałami.

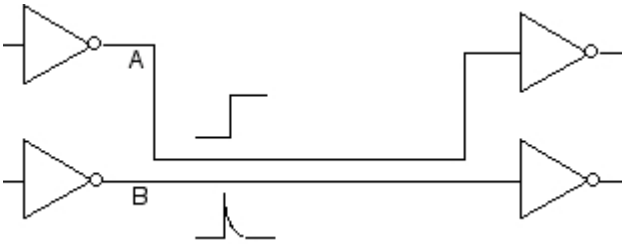
Warto dodać, że istnieją bramki nie mające zdolności regeneracji poziomów logicznych, a nawet takie, które powodują degradację tych poziomów, tj. dają na wyjściu napięcia spoza dopuszczalnego zakresu. Jednak w każdym złożonym układzie przynajmniej część bramek musi mieć zdolność regenerowania poziomów logicznych. W dalszej części wykładu będzie mowa o tym, jak sobie radzimy, gdy trzeba zastosować bramki nie mające tej zdolności.

Wiemy już więc, że bramki powinny mieć zdolność regeneracji poziomów logicznych. Drugim ważnym wymaganiem jest zdolność bramek do **tłumienia zakłóceń**.

Zakłócenia w układach cyfrowych mają zwykle charakter impulsów nakładających się na prawidłowy poziom napięcia reprezentujący "0" lub "1". Impulsy takie mogą pochodzić z kilku źródeł:

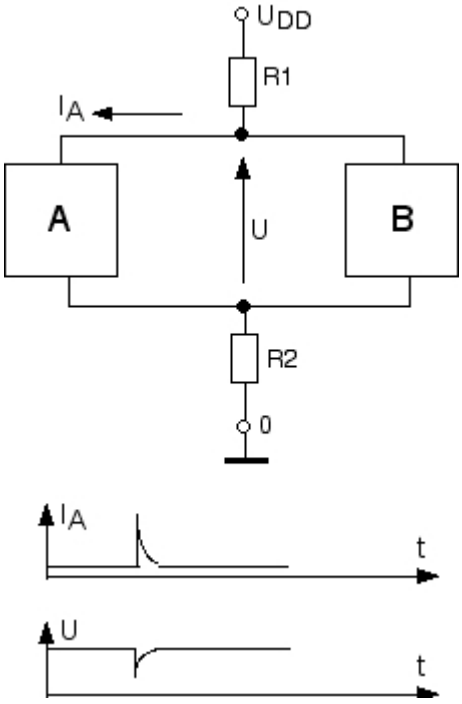
- pasożytnicze sprzężenia pomiędzy połączeniami (pojemnościowe, indukcyjne),
- sprzężenia poprzez wspólne zasilanie,
- zewnętrzne pola elektromagnetyczne,
- promieniowanie jonizujące.

Wpływ sprzężenia między połączeniami ilustruje symbolicznie rys. 7.2. Jeśli na jednej z dwóch położonych blisko siebie ścieżek pojawia się skokowa zmiana stanu logicznego, to na drugiej pojawia się krótki impuls wynikający z istnienia pojemności pomiędzy tymi ścieżkami (indukcyjność wzajemna też istnieje, ale ma zwykle drugorzędne znaczenie).




Rys. 7.2. Zakłócenie wywołane sprzężeniem pojemnościowym: zmiana napięcia na ścieżce A indukuje impuls w ścieżce B

Sprzężenie poprzez wspólne zasilanie jest wywołane skończoną, niezerową rezystancją połączeń między blokami logicznymi, a ich zasilaniem. Jak zobaczymy dalej, bramki CMOS pobierają prąd w postaci krótkich, stromych impulsów. Rys. 7.3 pokazuje, jak pobór prądu o takim charakterze wywołuje powstawanie i przenikanie zakłóceń impulsowych. Impuls prądu $I_A$ pobieranego przez blok logiczny A związany jest ze spadkiem napięcia na rezystancjach R1 i R2, i co za tym idzie wywołuje chwilowe zmniejszenie napięcia zasilającego U. Ten spadek napięcia o charakterze krótkiego impulsu jest widoczny nie tylko dla bloku A, ale i dla bloku B, ponieważ ich napięcie zasilania U jest wspólne.




Rys. 7.3. Zakłócenie wywołane chwilowym spadkiem napięcia zasilania spowodowanym impulsowym poborem prądu

Zewnętrzne pola elektromagnetyczne mogą indukować zakłócenia w układach, ale ich wpływ na działanie układów cyfrowych na ogół nie jest zauważalny, chyba że są to bardzo silne pola.

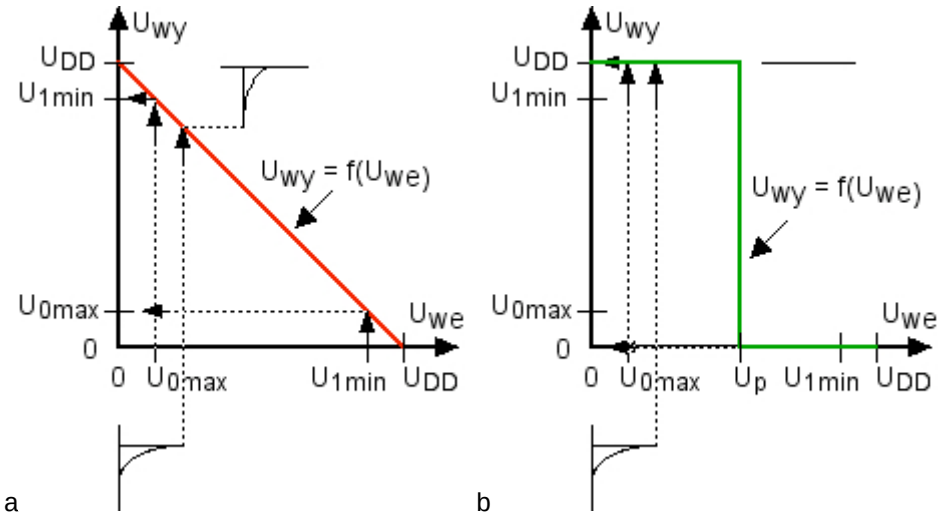
Promieniowanie jonizujące wywołuje w półprzewodniku generację par elektron-dziura, co powoduje przepływ

impulsów zwiększonego prądu wstecznego w spolaryzowanych zaporowo złączach p-n, np. złączach źródeł i drenów tranzystorów. Może to zakłócać przede wszystkim działanie układów zwanych dynamicznymi, w których informacja jest reprezentowana przez ładunek zgromadzony w pojemności (o takich układach będzie mowa w dalszych wykładach).

Nawet najstaranniejsze zaprojektowanie i wykonanie układu oraz prawidłowa jego eksploatacja nie umożliwiają całkowitej eliminacji zakłóceń, toteż zdolność do tłumienia zakłóceń jest równie ważną cechą bramek logicznych, jak zdolność do regeneracji poziomów logicznych.

**! Zdolność bramki do tłumienia zakłóceń polega na tym, że impuls zakłócający ma na wyjściu bramki mniejszą amplitudę, niż na wejściu.**

Zastanówmy się teraz, jak pogodzić regenerację poziomów logicznych z tłumieniem zakłóceń. Nietrudno pokazać, że te wymagania są do pewnego stopnia sprzeczne. Pokażemy to na hipotetycznym przykładzie "liniowego inwertera". Inwerter jest to najprostsza bramka logiczna wykonująca funkcję negacji NOT: "0" na wejściu daje "1" na wyjściu, i odwrotnie. Taką funkcję mógłby na przykład wykonywać układ elektroniczny o liniowej charakterystyce przejściowej $U_{wy} = U_{DD} - U_{we}$ pokazanej na rys. 7.4a. Jednak taki układ ani nie regeneruje poziomów logicznych ($U_{1min}$ na wejściu daje $U_{0max}$ na wyjściu, i odwrotnie), ani nie tłumi zakłóceń (impuls zakłócający na tle zera logicznego na wejściu pojawia się na tle jedynki logicznej na wyjściu z taką samą amplitudą). Wniosek z tego jest taki, że *bramka logiczna mająca równocześnie zdolność regeneracji poziomów logicznych i tłumienia zakłoceń nie może być układem liniowym*.



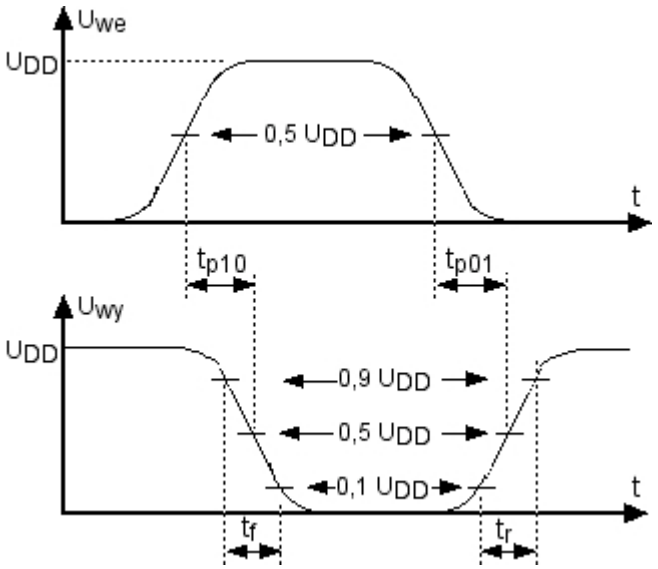

Rys. 7.4. Inwerter o liniowej charakterystyce przejściowej (a) i o idealnej charakterystyce przejściowej (b)

Charakterystyka przejściowa "inwertera idealnego", łączącego zdolność do regeneracji poziomów logicznych i do tłumienia zakłóceń, pokazana jest na rys. 7.4b. Widać, że $U_{1min}$ na wejściu daje 0 na wyjściu, a $U_{0max}$ na wejściu daje $U_{DD}$ na wyjściu. Widać także, że impuls zakłócający w ogóle nie pojawia się na wyjściu, ale *pod warunkiem, że jego amplituda nie przekracza napięcia przełączania inwertera* $U_p$.

Charakterystyki rzeczywistych inwerterów CMOS nie są idealne, ale są znacznie bliższe charakterystyce idealnej, niż charakterystyce liniowej. Będzie o tym mowa dalej.

Kolejnym wymaganiem dla bramek jest możliwie krótki **czas propagacji sygnału**. Od niego w dużym stopniu zależy, jak szybko będzie działał układ cyfrowy

Po zmianie stanów na wejściu zmiana stanów na wyjściu bramki cyfrowej nie następuje natychmiast. Miarami szybkości działania bramki są: **czasy narastania** $t_r$ i **opadania** $t_f$ sygnału na wyjściu oraz **czasy propagacji sygnału** $t_{p10}$ i $t_{p01}$. Są one zdefiniowane na rys. 7.5 dla najprostszego przypadku inwertera. Czasy narastania i opadania są zdefiniowane w odniesieniu do punktów na osi czasu, w których napięcie osiąga wartość 0,1 i 0,9 poziomu logicznego jedynki. Czasy propagacji są zdefiniowane w odniesieniu do punktów na osi czasu, w których napięcie ma wartość połowy poziomu logicznego jedynki.




Rys. 7.5. Czasy narastania, opadania i propagacji sygnału na przykładzie przebiegów na wejściu i wyjściu inwertera

Często definiuje się dla bramki jeden czas propagacji sygnału $t_p$, który jest średnią czasów $t_{p10}$ i $t_{p01}$: $t_p = (t_{p10} + t_{p01})/2$.

Od bramek cyfrowych oczekujemy także, że będą wykonywać swe funkcje przy możliwie jak najmniejszym poborze mocy.

**Pobór mocy** jest obecnie bardzo krytycznym parametrem. Pojedyncza bramka CMOS pobiera bardzo mało prądu. Jak zobaczymy dalej, w pierwszym przybliżeniu można założyć, że pobór prądu występuje tylko w chwilach przełączania (czyli zmiany stanów logicznych) i ma charakter krótkich, stromych impulsów. Jednak współczesne duże układy cyfrowe zawierają dziesiątki milionów bramek, co daje w sumie w chwilach przełączania impulsy prądowe sięgające wielu amperów i uśredniony w czasie pobór mocy rzędu dziesiątków, a nawet setek watów. Jest to bardzo poważny problem techniczny, tym bardziej, że owa moc wydziela się na bardzo małej powierzchni rzędu co najwyżej kilku $cm^2$. Temperatura, w której mogą pracować monolityczne krzemowe układy scalone, wynosi co najwyżej około 180 $^oC$. Odprowadzanie ciepła musi być bardzo intensywne, aby przy kilkudziesięciu W mocy wydzielających się na bardzo małej powierzchni temperatura nie przekroczyła maksymalnej dopuszczalnej wartości. Dlatego zminimalizowanie poboru mocy przez układ cyfrowy jest obecnie jednym z często spotykanych zadań dla projektanta. Mały pobór mocy jest także, z oczywistych względów, wymagany w przypadku układów do sprzętu przenośnego, o zasilaniu bateryjnym. O tym, od czego zależy pobór mocy w przypadku bramek CMOS, będzie mowa dalej.

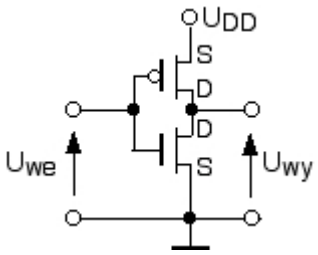
Jeszcze jedną cechą bramek cyfrowych, która ma w niektórych przypadkach znaczenie, jest **kierunkowość**.

Kierunkowość jest właściwością bramek polegającą na tym, że sygnały propagują się tylko w jedną stronę - od wejść do wyjść. Mówiąc precyzyjniej, stany wejść określają stany wyjść, natomiast stany wyjść nie mają wpływu na stany wejść. Jak zobaczymy, nie wszystkie bramki mają tę właściwość.

## 7.2. Inwerter CMOS

Jak wiemy, inwerterem nazywana jest bramka wykonująca funkcję negacji. Omówimy teraz budowę inwertera CMOS i jego parametry: napięcie przełączania, marginesy zakłóceń, czasy przełączania i pobór mocy.

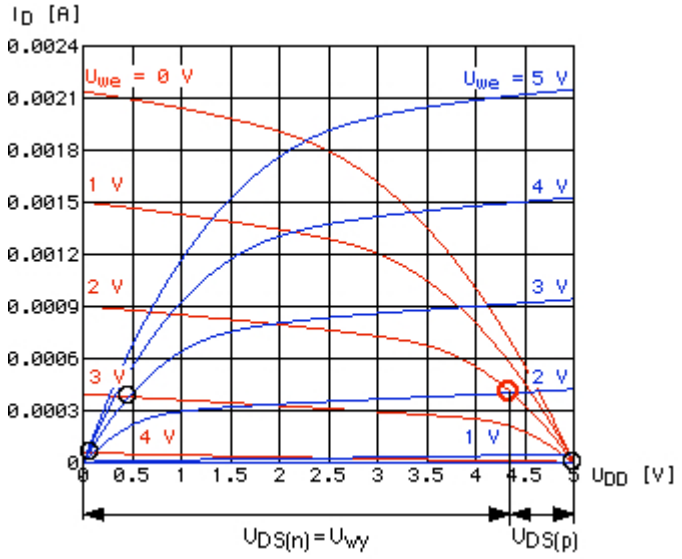
Inwerter CMOS zbudowany jest z dwóch połączonych szeregowo tranzystorów: nMOS i pMOS (rys. 7.6).




Rys. 7.6. Inwerter CMOS

Dreny obu tranzystorów są połączone z wyjściem. Bramki obu tranzystorów są połączone ze sobą i sterowane przez sygnał wejściowy. Zasada działania jest bardzo prosta. Gdy wejście jest w stanie "0", czyli napięcie wejściowe jest równe lub bliskie zeru, tranzystor nMOS jest wyłączony (nie przewodzi), zaś tranzystor pMOS jest włączony (przewodzi). Wyjście jest połączone przez tranzystor pMOS ze źródłem zasilania, napięcie na wyjściu jest równe $U_{DD}$, czyli wyjście jest w stanie "1". Odwrotna sytuacja powstaje, gdy wejście jest w stanie "1", czyli napięcie wejściowe jest równe lub bliskie $U_{DD}$. Włączony jest wówczas tranzystor nMOS, zaś tranzystor pMOS jest wyłączony. Wyjście jest uziemione przez tranzystor nMOS, a więc napięcie na nim jest równe zeru, co oznacza logiczne "0".

Dla określenia stałoprądowej charakterystyki przejściowej zauważmy, że oba prądy drenu są sobie równe, napięcie $U_{DS(n)}$ tranzystora nMOS jest równe napięciu wyjściowemu $U_{wy}$, napięcie $U_{DS(p)}$ tranzystora pMOS jest równe $U_{DD}$ - $U_{wy}$. Napięcie $U_{GS(n)}$ tranzystora nMOS jest równe napięciu wejściowemu $U_{we}$, zaś napięcie $U_{GS(p)}$ tranzystora pMOS jest równe $U_{DD}$ - $U_{we}$. Mimo prostoty wzorów (4.1) - (4.4) nie jest możliwe podanie opisu charakterystyki przejściowej inwertera w postaci wzoru. Można natomiast skonstruować tę charakterystykę graficznie. Sposób pokazany jest na rys. 7.7. Rysunek ten ilustruje graficzny sposób rozwiązania układu rownań: $I_{D(n)} = I_{D(p)}$, $U_{DS(n)} + U_{DS(p)} = U_{DD}$, $U_{GS(n)} + U_{GS(p)} = U_{DD}$.

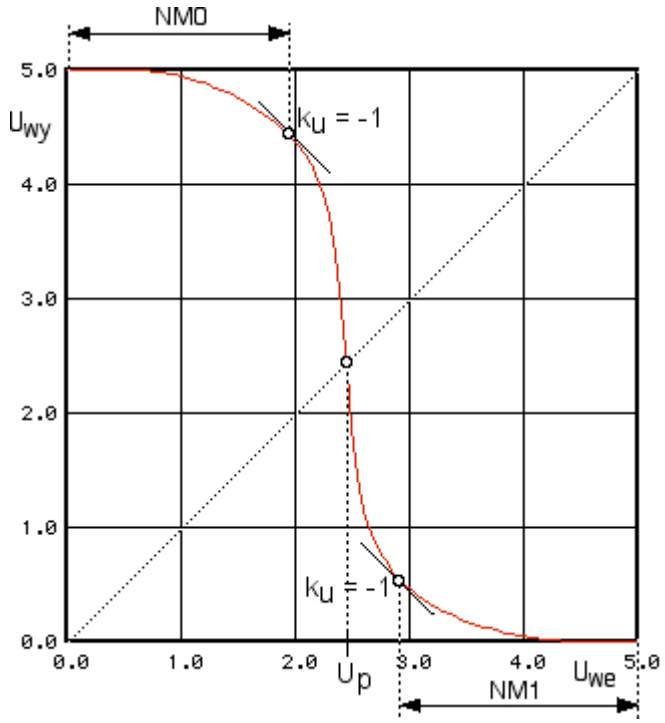



Rys. 7.7. Graficzny sposób wyznaczenia punkt po punkcie charakterystyki przejściowej inwertera

W polu charakterystyk $I_D(U_{DS})$ wrysowujemy w tej samej skali charakterystyki tranzystora nMOS (niebieskie) i tranzystora pMOS (czerwone). Dla tych ostatnich początkiem układu współrzędnych jest punkt ($U_{DD}$,0), zaś oś napięć ma przeciwny zwrot. Kółkami zaznaczone są punkty przecięcia charakterystyk tranzystora nMOS i pMOS odpowiadających temu samemu napięciu wejściowemu (dla punktu zaznaczonego na czerwono na rys 7.7 jest to 2 V). Przecięcie charakterystyk oznacza, że właśnie w tym punkcie spełnione jest równanie $I_{D(n)} = I_{D(p)}$ dla danego

napięcia wejściowego. Równocześnie dzięki umieszczeniu początku układu współrzędnych dla charakterystyk tranzystora pMOS w punkcie ($U_{DD}$,0) suma napięć $U_{DS(n)}$ i $U_{DS(p)}$ jest równa $U_{DD}$. Wartość $U_{DS(n)}$ jest równa napięciu wyjściowemu $U_{wy}$ dla danego napięcia wejściowego (dla punktu zaznaczonego na czerwono na rys. 7.7 jest to około 4,4 V). Powtarzając tę konstrukcję dla kolejnych napięć wejściowych można odczytać odpowiadające im napięcia wyjściowe, a następnie sporządzić wykres charakterystyki przejściowej.

Chociaż 4 punkty zaznaczone na rys. 7.7 nie pozwalają na uzyskanie dokładnego rysunku charakterystyki przejściowej, widać że w zakresie przełączania charakterystyka ta jest bardzo stroma. Przy zmianie napięcia wejściowego od 2 V do 3 V napięcie wyjściowe zmienia się od ok. 4,4 V do około 0,45 V. Typowy kształt charakterystyki przejściowej inwertera pokazuje rys. 7.8. Jest to charakterystyka otrzymana przy użyciu symulatora SPICE dla tych samych tranzystorów, których charakterystyki $I_D(U_{DS})$ pokazano na rys.7.7.




Rys. 7.8. Charakterystyka przejściowa inwertera CMOS, z zaznaczonym napięciem przełączania Up i marginesami zakłóceń

Na charakterystyce przejściowej inwertera zaznaczono trzy charakterystyczne punkty. Dwa z nich są to punkty, w których wzmocnienie napięciowe $k_u$ (tj. pochodna $dU_{wy}/dU_{we}$) ma wartość bezwzględną równą 1. Te punkty wyznaczają maksymalne wartości amplitudy zakłóceń, dla których zakłócenia są tłumione. Odpowiednie wartości amplitud są oznaczone symbolami NM0 - od strony "0", i NM1 - od strony "1" (symbol NM pochodzi od angielskiego terminu "noise margin"). Wartości NM0 i NM1 nazywamy **marginesami zakłóceń**. Trzeci punkt to umowne **napięcie przełączania** inwertera $U_p$. Jest to punkt przecięcia charakterystyki z prostą o równaniu $U_{we} = U_{wy}$.

Ze wszystkich znanych sposobów realizacji inwertera inwerter CMOS ma charakterystykę przejściową najbardziej zbliżoną do idealnej. Przy odpowiednim doborze wymiarów kanałów tranzystorów można uzyskać charakterystykę symetryczną o napięciu przełączania równym połowie napięcia zasilania. Taka charakterystyka oznacza także maksymalną odporność na zakłócenia - oba marginesy zakłóceń są wówczas jednakowe. W obszarze przejściowym charakterystyka jest bardzo stroma, co oznacza, że punkty na charakterystyce wyznaczające marginesy zakłóceń leżą niezbyt daleko od punktu wyznaczającego napięcie przełączania.

W obszarze przejściowym oba tranzystory pracują w zakresie nasycenia. Wartość napięcia przełączania można więc oszacować przyrównując oba prądy drenu wyznaczane z zależności 4.4 :

$$K_n\left(U_p - U_{Tn}\right)^2 = K_p\left(U_{DD} - U_p - \left|U_{Tp}\right|\right)^2 \qquad (7.1)$$

gdzie oznaczono

$$K = \mu C_{ox} \frac{W}{L} \tag{7.2}$$

($K$ będzie nazywany **współczynnikiem przewodności tranzystora**) oraz przypisano indeks "*n*" wielkościom odnoszącym się do tranzystora nMOS, zaś indeks "*p* " wielkościom odnoszącym się do tranzystora pMOS. Rozwiązując równanie (7.1) otrzymujemy:

$$U_p = \frac{U_{Tn} + \sqrt{r}\left(U_{DD} - \left|U_{Tp}\right|\right)}{1 + \sqrt{r}} \tag{7.3}$$

gdzie

$$r = \frac{K_p}{K_n} = \frac{\mu_p}{\mu_n} \frac{\left(\frac{W}{L}\right)_p}{\left(\frac{W}{L}\right)_n} \tag{7.4}$$

Ze wzoru (7.3) można wyznaczyć wartość $r$ dla wymaganej wartości napięcia przełączania $U_p$:

$$r = \left[\frac{U_p - U_{Tn}}{\left(U_{DD} - \left|U_{Tp}\right|\right) - U_p}\right]^2 \tag{7.5}$$

Nietrudno sprawdzić, że dla uzyskania charakterystyki symetrycznej, przy założeniu $U_{Tn} = |U_{Tp}|$ (które jest często, choć nie zawsze, spełnione), musimy uzyskać $r = 1$. Oznacza to warunek

$$\frac{\left(\frac{W}{L}\right)_p}{\left(\frac{W}{L}\right)_n} = \frac{\mu_n}{\mu_p} \tag{7.6}$$

Sens fizyczny tego warunku jest bardzo prosty: dla uzyskania pełnej symetrii różnicę ruchliwości nośników w kanałach tranzystorów należy skompensować różnicą szerokości $W$ tych kanałów (przy założeniu jednakowych długości $L$).

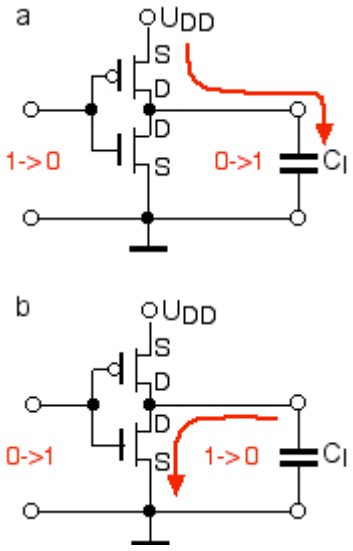
**! Ponieważ ruchliwość elektronów $\mu_n$ jest 2 ... 2,5 raza większa od ruchliwości dziur $\mu_p$, dla uzyskania symetrii kanał tranzystora pMOS powinien być 2 ... 2,5 raza szerszy od kanału tranzystora nMOS.**

Jeżeli zależy nam na dokładnej symetrii, to dobranie tej szerokości wymaga wykonania kilku symulacji elektrycznych z zastosowaniem dokładnych modeli tranzystorów, bowiem zależności, z których korzystamy, są przybliżone, a ponadto warunek równości napięć progowych $U_{Tn} = |U_{Tp}|$ nie zawsze jest spełniony.

**! Zauważmy, że charakterystyka przejściowa inwertera zależy od ilorazu stosunków $W/L$ obu tranzystorów, a nie od ich bezwzględnych wartości. Jak zobaczymy dalej, tę samą własność mają wszystkie statyczne bramki kombinacyjne CMOS. To oznacza, że jeśli proporcjonalnie skalujemy wszystkie wymiary tranzystorów, to takie właściwości bramek, jak poziomy logiczne, napięcia przełączania i marginesy zakłóceń pozostają bez zmian (pod dodatkowym warunkiem, że bez zmian pozostają także napięcia progowe i napięcie zasilania).**

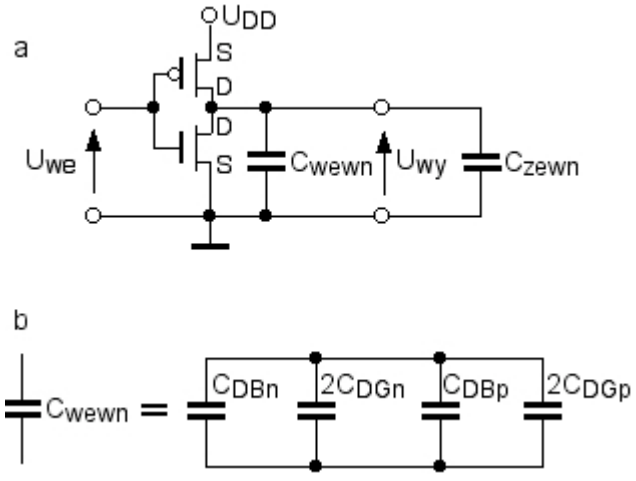
Zajmiemy się teraz czasami przełączania. Terminem "**czasy przełączania**" będziemy ogólnie określać czasy narastania i opadania sygnału na wyjściu oraz czasy propagacji sygnału. Czasy propagacji sygnału decydują o szybkości działania bramki w układzie. Czasy narastania i opadania sygnału sterującego bramkę mają wpływ na jej czasy propagacji, więc także są istotne. Dobre oszacowanie szybkości działania układu z bramkami CMOS jest możliwe tylko przy pomocy symulacji elektrycznej. Tu jednak wyprowadzimy kilka prostych wzorów dla ogólnej orientacji i zgrubnych oszacowań.

Zarówno czasy propagacji, jak i czasy narastania i opadania są uwarunkowane szybkością ładowania lub rozładowywania pojemności, jaką obciążony jest inwerter - patrz rys. 7.9. Gdy stan na wyjściu zmienia się z "0" na "1", pojemność obciążająca $C_l$ ładuje się w wyniku przepływu prądu ze źródła zasilania przez tranzystor pMOS. Gdy stan na wyjściu zmienia się z "1" na "0", pojemność obciążająca $C_l$ rozładowuje się w wyniku przepływu prądu przez tranzystor nMOS.




Rys. 7.9. Ładowanie (a) i rozładowywanie (b) pojemności obciążającej przy przełączaniu inwertera

Na pojemność obciążającą $C_l$ składają się wewnętrzne pojemności samego inwertera oraz pojemności zewnętrzne w stosunku do niego, takie jak suma pojemności wejściowych innych bramek obciążających inwerter i połączeń prowadzących do tych bramek - rys. 7.10.

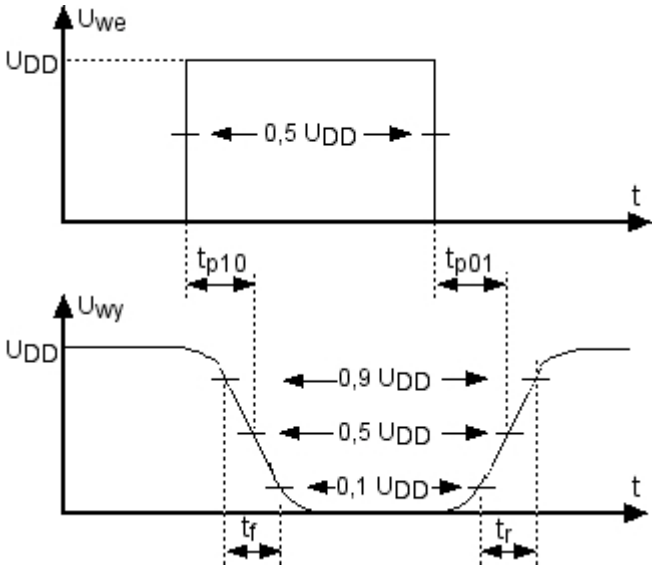



Rys. 7.10. Inwerter obciążony pojemnością wewnętrzną i zewnętrzną (a) oraz składniki pojemności wewnętrznej (b)

Pojemności wewnętrzne obciążające inwerter to suma pojemności złączowych drenów tranzystorów nMOS i pMOS oraz suma podwojonych pojemności dren-bramka tych tranzystorów. Dlaczego podwojonych? Otóż pojemności te są w rzeczywistości włączone między węzeł wyjściowy, a wejściowy. W procesie zmiany stanów logicznych napięcie na wejściu zmienia się od zera do $U_{DD}$ (lub odwrotnie), a napięcie na wyjściu zmienia się w przeciwnym kierunku: od $U_{DD}$ do zera (lub odwrotnie). W rezultacie napięcie na pojemnościach włączonych między wyjściem, a wejściem zmienia się o 2 $U_{DD}$. Pojemności te w schemacie z rys. 7.10 są przeniesione na

wyjście, gdzie przy zmianie stanów logicznych napięcie zmienia się tylko o wartość równą $U_{DD}$. Zatem aby pojemności po przeniesieniu na wyjście gromadziły taki sam ładunek, jak w miejscu, w którym rzeczywiście występują, ich wartości trzeba podwoić. Zabieg polegający na przeniesieniu tych pojemności na wyjście bardzo ułatwia oszacowanie czasów przełączania.

Założymy dla uproszczenia, że inwerter CMOS jest sterowany sygnałem o kształcie idealnego impulsu prostokątnego, tj. o czasach narastania i opadania równych zeru. Wówczas czas propagacji liczyć należy od chwili, w której nastąpiła skokowa zmiana napięcia na wejściu do chwili, w której napięcie wyjściowe osiągnęło (malejąc lub rosnąc, zależnie od kierunku zmiany) wartość równą 0,5 $U_{DD}$ (patrz rys. 7.11, który jest zmodyfikowaną wersją rys. 7.5).




Rys. 7.11. Czasy propagacji przy wyidealizowanym sygnale wejściowym

Dla oszacowania czasów $t_{p10}$ i $t_{p01}$ założymy, że w czasie ładowania pojemności obciążającej $C_l$ płynie przez nią prąd ładowania o stałej wartości równej prądowi nasycenia tranzystora pMOS, a podczas rozładowania pojemność ta rozładowuje się prądem o stałej wartości równej prądowi nasycenia tranzystora nMOS. Nie jest to bardzo złe przybliżenie, bowiem symulacje pokazują, że tranzystory pozostają w stanie nasycenia przez większą część czasu ładowania lub rozładowywania pojemności obciążającej. Początkową wartością napięcia przy ładowaniu jest 0, końcową (dla oszacowania czasu $t_{p01}$) 0,5 $U_{DD}$. Początkową wartością napięcia przy rozładowywaniu jest $U_{DD}$, końcową (dla oszacowania czasu $t_{p10}$) 0,5 $U_{DD}$. Przy tych założeniach czasy propagacji sygnału można oszacować przy pomocy wzorów

$$t_{p10} = \frac{C_l U_{DD}}{K_n \left(U_{DD} - U_{Tn}\right)^2} = \frac{C_l U_{DD}}{\mu_n C_{ox} \left(U_{DD} - U_{Tn}\right)^2 \left(\frac{L}{W}\right)_n} \tag{7.7}$$

$$t_{p01} = \frac{C_l U_{DD}}{K_p \left(U_{DD} - \left|U_{Tp}\right|\right)^2} = \frac{C_l U_{DD}}{\mu_p C_{ox} \left(U_{DD} - \left|U_{Tp}\right|\right)^2 \left(\frac{L}{W}\right)_p} \tag{7.8}$$

Jak widać, inwerter działa tym szybciej, im mniejsza jest pojemność obciążająca, im większa jest szerokość kanału, im mniejsza jest długość kanału i im większe jest napięcie zasilania układu. Na te wielkości ma wpływ konstruktor układu.

Z punktu widzenia szybkości działania układu logicznego korzystne jest na ogół, by czasy propagacji $t_{p10}$ i $t_{p01}$ miały zbliżone wartości. Ze wzorów (7.7) i (7.8) widać, że jednakowe czasy propagacji $t_{p10}$ i $t_{p01}$ osiąga się przy jednakowych wartościach napięć progowych i jednakowych wartościach współczynników $K_n$ i $K_p$. Są to te same warunki, które zapewniają symetryczną charakterystykę przejściową inwertera.

---

Na koniec zajmiemy się oszacowaniem poboru mocy. Jest to bardzo ważny problem, w dzisiejszym stanie

technologii CMOS nie można już dalej zwiększać szybkości układów CMOS, bo na przeszkodzie stoi wzrost poboru mocy. Dlatego temu problemowi poświęcona jest osobny wykład 14. Na razie tylko podstawowe informacje i szacunkowe wzory.

Prąd, jaki pobiera ze źródła zasilania statyczny inwerter CMOS, ma dwie składowe: statyczną i dynamiczną. **Składowa statyczna** to prąd, jaki płynie w stanie ustalonym, gdy stany logiczne nie zmieniają się. Prąd ten ma małą wartość, bowiem zarówno w stanie "0" na wejściu, jak i w stanie "1" jeden z połączonych szeregowo tranzystorów - nMOS lub pMOS - jest wyłączony, nie przewodzi. Statyczny prąd ma kilka składników, z których najistotniejszy jest zwykle prąd podprogowy tego z tranzystorów MOS, który jest w danej chwili wyłączony. Jeżeli sumę wszystkich prądów składających się na prąd statyczny nazwiemy prądem statycznego upływu $I_{stat}$, to moc statyczna $P_{stat}$ pobierana przez inwerter wynosi

$$P_{stat} = I_{stat} U_{DD} \tag{7.9}$$

Moc statyczna była do niedawna uważana za całkowicie pomijalną. W najnowocześniejszych technologiach tak już nie jest, a dlaczego - o tym będzie mowa w wykładzie 14.

**Składowa dynamiczna** poboru prądu pojawia się, gdy zmieniają się stany logiczne. Jest to prąd, który płynie tylko w czasie zmiany stanu logicznego. Ma on dwa składniki. Pierwszy z nich związany jest z ładowaniem i rozładowywaniem pojemności obciążającej. Drugi płynie w czasie przełączania z tego powodu, że istnieje taki zakres napięć wejściowych, dla których oba tranzystory inwertera równocześnie przewodzą, a zatem podczas zmiany napięcia na wejściu przez krótki czas prąd może płynąć bezpośrednio ze źródła zasilania do masy.

Przy każdej zmianie stanu powodującej naładowanie pojemności obciążającej $C_l$ do napięcia $U_{DD}$ ze źródła zasilania wypływa energia o wartości $E_c = C_l U_{DD}^2$. W każdym cyklu zmiany stanów na wyjściu "0"->"1"->"0" następuje jedno naładowanie i jedno rozładowanie. Można pokazać, że energia $E_c$ ulega rozproszeniu w połowie w tranzystorze pMOS (podczas ładowania) i w połowie w tranzystorze nMOS (podczas rozładowania). Jeżeli w ciągu sekundy cykli ładowanie-rozładowanie jest *f*, to moc $P_c$ pobierana ze źródła zasilania wynosi

$$P_c = C_l U_{DD}^2 f \tag{7.10}$$

Do niedawna był to w układach CMOS główny składnik pobieranej mocy. Moc $P_c$ jest proporcjonalna do częstotliwości, z jaką przełączają bramki (czyli - z grubsza - do częstotliwości zegara, jakim taktowany jest układ), do pojemności obciążającej bramki oraz do kwadratu napięcia zasilającego.

Ostatnim omawianym prądem jest prąd, który płynie bezpośrednio przez tranzystory w okresie, gdy w czasie przełączania oba jednocześnie przewodzą. Gdyby czasy narastania i opadania sygnału na wejściu były równe zeru, pobór mocy związany z tym prądem także byłby równy zeru, bo odcinek czasu, w którym tranzystory równocześnie przewodzą, byłby nieskończenie krótki. Przy różnych od zera czasach $t_r$ i $t_f$ pobór mocy $P_j$ można w przybliżeniu oszacować tak:

$$P_j = I_{max} U_{DD} \frac{t_r + t_f}{2} f \tag{7.11}$$

gdzie $I_{max}$ jest szczytową wartością prądu płynącego w czasie przełączania przez równocześnie przewodzące tranzystory. Z punktu widzenia poboru mocy korzystne jest więc, by sygnały wejściowe miały jak najkrótsze czasy narastania i opadania.

Łączny pobór mocy jest sumą mocy określonych wzorami (7.9) - (7.1), przy czym zazwyczaj dominuje moc związana z ładowaniem-rozładowywaniem pojemności $P_c$.

Co z tego wynika dla projektanta? Jak dobrać wymiary tranzystorów w inwerterze? Co do długości kanału, w układach cyfrowych regułą jest stosowanie najmniejszej długości, na jaką pozwala proces technologiczny. Wynika to stąd, że im krótszy kanał, tym krótsze czasy propagacji sygnałów. Szerokości kanałów można dobierać ze względu na kilka kryteriów:

- maksymalna szybkość działania,

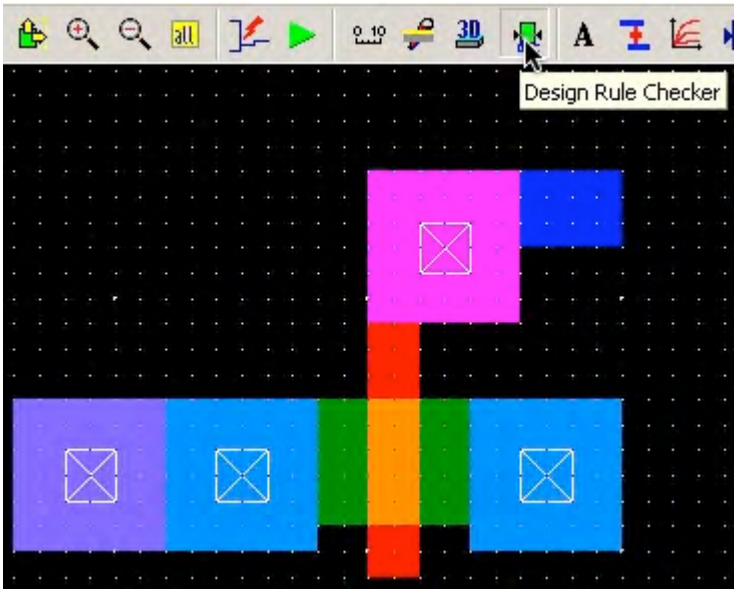
- minimalny pobór mocy,
- minimalna powierzchnia.

Dla uzyskania maksymalnej szybkości działania, czyli możliwie krótkich czasów propagacji, trzeba przede wszystkim ustalić, jaki charakter ma pojemność obciążająca. Dwa skrajne przypadki to: (a) dominująca pojemność zewnętrzna $C_{zewn}$ >> $C_{wewn}$, i (b) dominująca pojemność wewnętrzna $C_{zewn}$ << $C_{wewn}$. Jeśli dominuje pojemność zewnętrzna, której wartość nie zależy od wymiarów tranzystorów w inwerterze, to im większa szerokość kanałów tranzystorów, tym krótsze czasy propagacji (patrz wzory (7.7) i (7.8)). Jeśli natomiast dominuje pojemność wewnętrzna, to poszerzanie kanałów nie jest celowe. Ze wzrostem szerokości kanałów tranzystorów proporcjonalnie rosną powierzchnie złącz drenów, czyli pojemności $C_{DB}$, a także powierzchnie bramek i zakładek bramek nad drenami, czyli w sumie pojemności $C_{DG}$. Ze wzrostem szerokości kanałów rośnie więc proporcjonalnie pojemność $C_l$, a czasy propagacji pozostają bez zmiany. Większe szerokości kanałów są w tej sytuacji wręcz szkodliwe, bo wzrasta niepotrzebnie pobór mocy. Zarówno w przypadku (a), jak i w przypadku (b), zachowywany jest zwykle stosunek $W_p/W_l$ rzędu 2 ... 2,5 dla zapewnienia maksymalnej odporności na zakłócenia i jednakowych czasów propagacji $t_{p10}$ i $t_{p01}$.

Minimalny pobór mocy układu uzyskuje się przy minimalnej sumie wszystkich pojemności obciążających bramki, zgodnie ze wzorem (7.10). Reguła jest więc prosta: wszystkie wymiary tranzystorów (bramek, ale także obszarów źródeł i drenów) powinny być jak najmniejsze. Możliwe są tu dwa przypadki. W pierwszym przypadku minimalną długość i szerokość kanału mają tranzystory nMOS, natomiast tranzystory pMOS mają kanały szersze 2 ... 2,5 raza dla zapewnienia maksymalnej odporności na zakłócenia. W drugim przypadku również tranzystory pMOS mają oba wymiary minimalne. Charakterystyki przejściowe nie są wówczas symetryczne. Napięcie przełączania jest mniejsze od $0{,}5U_{DD}$, a margines zakłóceń od strony zera logicznego ulega zmniejszeniu. W wielu przypadkach margines ten pozostaje jednak wystarczająco duży, zwłaszcza w przypadku układów zasilanych napięciem 5V.

Minimalną powierzchnię układu również uzyskamy zmniejszając do wartości minimalnych dopuszczalnych wymiary kanałów tranzystorów (a także obszarów źródeł i drenów). Jest to więc przypadek już omówiony. Warto jednak w tym miejscu dodać, że we współczesnych technologiach CMOS o powierzchni układu nie decydują wymiary tranzystorów, lecz kontaktów i ścież ek połączeń (porównaj w dodatku 1 topografię tranzystora zaprojektowanego w ramach ćwiczenia 1 do wykładu 6!). Może się więc okazać, że zmniejszanie do minimum wymiarów tranzystorów nie jest celowe, ponieważ nie przynosi istotnego zmniejszenia powierzchni układu.

## 7.2. Dodatek 1: Tranzystor nMOS zaprojektowany w ćwiczeniu 1 do wykładu 6

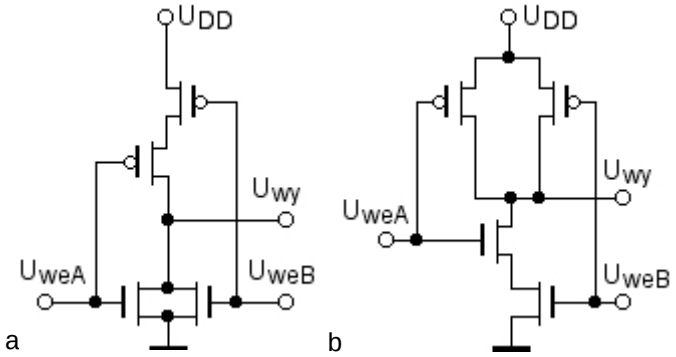



Tranzystor nMOS zaprojektowany w ćwiczeniu 1, wykład 6. Powierzchnia kanału jest bardzo mała w porównaniu z całkowitą powierzchnią zajmowaną przez tranzystor łącznie z niezbędnymi kontaktami.

## 7.3. Bramki NOR i NAND

Sam inwerter oczywiście nie wystarcza do budowy układów cyfrowych. Omówiony był dość szczegółowo dlatego, że jego właściwości łatwo uogólnić na bramki wielowejściowe. Zajmiemy się teraz bramkami **NOR** i **NAND**.

Bramki wykonujące funkcje NOR i NAND tworzy się przez równoległe i szeregowe połączenia tranzystorów. Rysunki poniżej przedstawiają bramki dwuwejściowe.




Rys. 7.12. Dwuwejściowe bramki NOR2 (a) i NAND2 (b)

Działanie bramek łatwo sprawdzić.

W przypadku bramki NOR stan "1" na dowolnym z wejść, A lub B, włącza odpowiedni tranzystor nMOS i wyłącza odpowiedni tranzystor pMOS. W tym stanie wyjście jest uziemione i panuje na nim stan "0". Tylko w przypadku zer na obu wejściach oba tranzystory nMOS są wyłączone, a oba tranzystory pMOS - włączone. Wyjście jest wówczas połączone ze źródłem zasilania i panuje na nim stan "1". Jest to właśnie funkcja NOR. Zauważmy, że przy żadnej kombinacji stanów na wejściu nie ma w stanie ustalonym przepływu prądu ze źródła zasilania, bowiem gdy któryś z połączonych równolegle tranzystorów nMOS (lub oba) jest włączony, to któryś z połączonych szeregowo tranzystorów pMOS (lub oba) jest wyłączony.

Działanie bramki NAND jest podobne: stan "0" na dowolnym z wejść, A lub B, włącza odpowiedni tranzystor pMOS i wyłącza odpowiedni tranzystor nMOS. Wyjście jest wówczas połączone ze źródłem zasilania i panuje na nim stan "1". Tylko w przypadku jedynek na obu wejściach oba tranzystory nMOS są włączone, a oba tranzystory pMOS - wyłączone. W tym stanie wyjście jest uziemione i panuje na nim stan "0". Jest to właśnie funkcja NAND. Również i w przypadku tej bramki przy żadnej kombinacji stanów na wejściu nie ma w stanie ustalonym przepływu prądu ze źródła zasilania, bowiem gdy któryś z połączonych równolegle tranzystorów pMOS (lub oba) jest włączony, to któryś z połączonych szeregowo tranzystorów nMOS jest wyłączony (lub oba).

W podobny sposób można zbudować bramki NOR i NAND z większą liczbą wejść.

Nasza znajomość parametrów i charakterystyk inwertera da się łatwo uogólnić na bramki wielowejściowe NOR i NAND. Rozważmy na początek charakterystyki przejściowe i napięcie przełączania. Rozróżnić trzeba dwa przypadki: przełączania równoczesnego obu wejść i przełączania tylko jednego wejścia.

Rozważymy najpierw przełączanie równoczesne obu wejść. Gdy na oba wejścia doprowadzone są sygnały identycznie zmieniające się w czasie, wejścia te można potraktować jako zwarte. Bramka NOR staje się wówczas równoważna inwerterowi, w którym rolę tranzystora nMOS pełnią dwa tranzystory połączone równolegle, a rolę tranzystora pMOS - dwa tranzystory połączone szeregowo. Do obliczenia napięcia przełączania można użyć zależności (7.3), w której wartość $r$ daną wzorem (7.4) zastąpimy zmodyfikowaną wartością $r'$:

$$r' = \frac{\mu_p}{\mu_n}\frac{\left(\frac{W}{2L}\right)_p}{\left(\frac{2W}{L}\right)_n} = \frac{\mu_p}{\mu_n}\frac{1}{4}\frac{\left(\frac{W}{L}\right)_p}{\left(\frac{W}{L}\right)_n} = \frac{1}{4}r \qquad (7.12)$$

Otrzymujemy wówczas (w uogólnieniu dla bramki NOR o N wejściach)

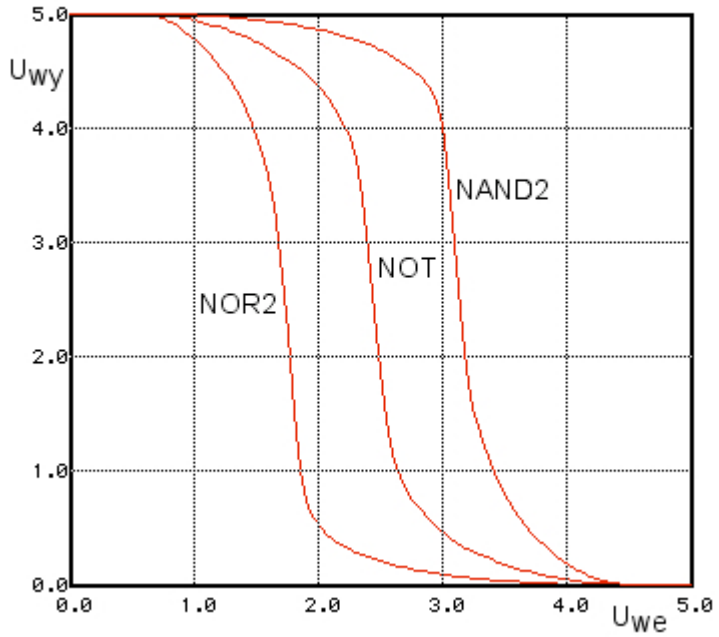
$$U_p = \frac{U_{Tn} + \frac{1}{N}\sqrt{r}\left(U_{DD} - \left|U_{Tp}\right|\right)}{1 + \frac{1}{N}\sqrt{r}} \tag{7.13}$$

Analogicznie rozumując dla bramki NAND o N wejściach otrzymujemy

$$U_p = \frac{U_{Tn} + N\sqrt{r}\left(U_{DD} - \left|U_{Tp}\right|\right)}{1 + N\sqrt{r}} \tag{7.14}$$

W obu wzorach, (7.13) i (7.14), *r* dane jest wzorem (7.4).

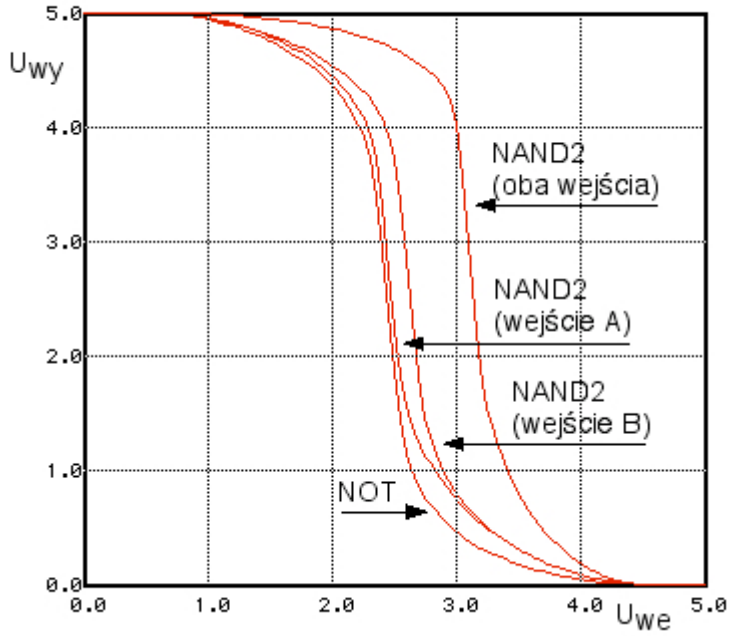
Z zależności (7.13) i (7.14) widać, że jeśli do budowy bramek NOR i NAND użyjemy takich samych tranzystorów, jak dla inwertera, to przy jednoczesnym przełączaniu wejść otrzymamy wartość napięcia przełączania różną od napięcia przełączania inwertera. Różnica jest tym większa, im większa jest liczba wejść bramki. Na rys. 7.13 pokazano charakterystyki przejściowe inwertera oraz bramek dwuwejściowych NOR i NAND zbudowanych z takich samych tranzystorów jak inwerter:




Rys. 7.13. Charakterystyki przejściowe jednoczesnego przełączania obu wejść dla dwuwejściowych bramek NOR i NAND zbudowanych z takich samych tranzystorów, jak inwerter, w porównaniu z charakterystyką tego inwertera.

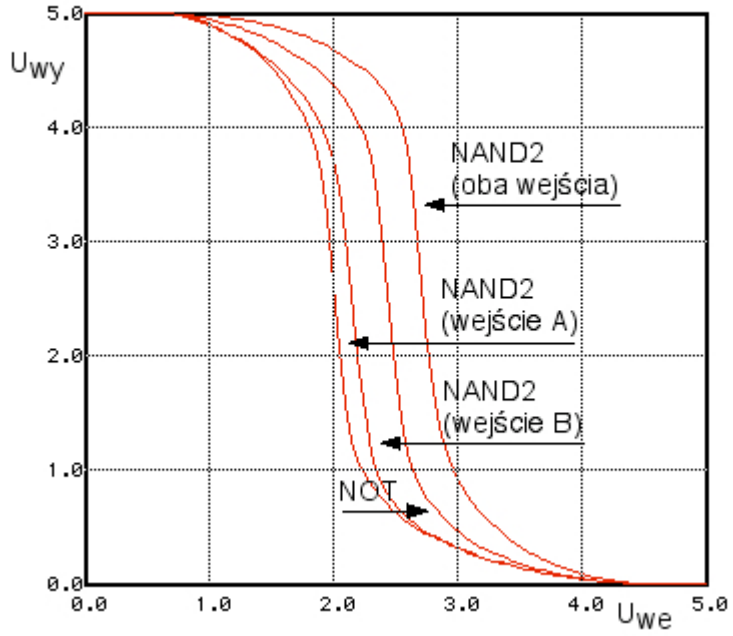
Zmiana napięcia przełączania jest niekorzystna, bowiem zmniejsza odporność bramki na zakłócenia od strony zera (NOR) lub od strony jedynki (NAND).

Przy przełączaniu tylko na jednym wejściu przesunięcie charakterystyki także występuje, ale jest niewielkie. Przykładowo, dla bramki NAND odpowiednie charakterystyki wyglądają następująco:




Rys. 7.14. Charakterystyki przejściowe jednoczesnego przełączania obu wejść oraz każdego wejścia z osobna dla dwuwejściowej bramki NAND zbudowanej z takich samych tranzystorów, jak inwerter, w porównaniu z charakterystyką tego inwertera.

Bardziej korzystne charakterystyki można uzyskać poszerzając kanały tranzystorów połączonych szeregowo. Stosowana jest tu prosta reguła: kanały te poszerza się tylokrotnie, ile tranzystorów jest połączonych szeregowo. Wówczas stosunek $W/L$ dla całego łańcucha połączonych szeregowo tranzystorów jest taki sam, jak dla pojedynczego tranzystora przed poszerzeniem. W rezultacie charakterystyki przełączania dla poszczególnych wejść nieco się pogarszają, za to poprawia się charakterystyka jednoczesnego przełączania. Rys. 7.15 pokazuje takie charakterystyki dla bramki NAND, w której dwukrotnie zwiększono szerokość kanałów tranzystorów nMOS.




Rys. 7.15. Charakterystyki przejściowe jednoczesnego przełączania obu wejść oraz każdego wejścia z osobna dla dwuwejściowej bramki NAND, w której kanały tranzystorów nMOS poszerzono dwukrotnie w stosunku do tranzystora nMOS inwertera, w porównaniu z charakterystyką tego inwertera.

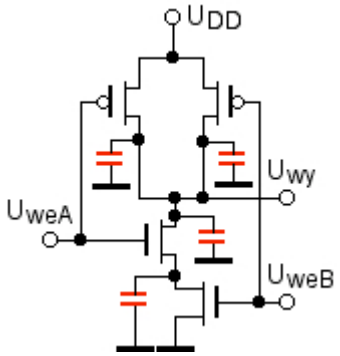
**Jak widać, zwymiarowanie tranzystorów w taki sposób, że tranzystory w połączeniu równoległym pozostają bez zmiany, a w połączeniu szeregowym są poszerzone tylokrotnie, ile jest ich w łańcuchu, daje w rezultacie charakterystyki przełączania bliskie symetrycznej charakterystyce inwertera, zapewniające dostateczną odporność bramki na zakłócenia przy wszystkich kombinacjach stanów wejść. Dotyczy to zarówno bramek NOR, jak i NAND.**

Jednak przesunięcia charakterystyk w stosunku do symetrycznej charakterystyki inwertera, widoczne na rys. 7.15, rosną przy wzroście liczby wejść. Dlatego statyczne bramki kombinacyjne CMOS nie mogą mieć dowolnie dużej liczby wejść.

**W praktyce nie stosuje się bramek statycznych NOR i NAND o liczbie wejść większej niż 4.**

Czasy propagacji sygnału w bramkach NOR i NAND są określone przez ten sam mechanizm, co w inwerterze - ładowanie pojemności obciążającej poprzez tranzystory pMOS, rozładowywanie poprzez tranzystory nMOS. Do oszacowania czasów propagacji można użyć zależności (7.7) i (7.8), w których trzeba podstawić wartości $L/WN$ połączonych równolegle (i równocześnie włączonych) tranzystorów, to szerokość pojedynczego tranzystora należy pomnożyć przez $N$. Czas propagacji wówczas maleje $N$-krotnie, co jest z reguły korzystne. Gdy $N$ tranzystorów połączonych jest szeregowo, wówczas przez $N$ należy pomnożyć długość kanału pojedynczego tranzystora $L$. W tym, bardzo niekorzystnym, przypadku czas propagacji rośnie $N$-krotnie. Nadmiernemu wydłużeniu czasu propagacji przeciwdziała reguła poszerzania kanałów tranzystorów połączonych szeregowo, o której była mowa wyżej. Jeżeli w szeregowym połączeniu $N$ tranzystorów kanały są poszerzone $N$-krotnie, to *w pierwszym przybliżeniu* $N$ tranzystorów jest skompensowany $N$-krotnym poszerzeniem ich kanałów.

W rzeczywistości jednak większa liczba tranzystorów w bramce wydłuża czasy propagacji także dlatego, że suma pojemności ładowanych i rozładowywanych jest większa. W połączeniu równoległym sumują się pojemności złączowe drenów wszystkich tranzystorów. W połączeniu szeregowym dochodzą dodatkowe pojemności związane z węzłami wewnętrznymi w łańcuchu połączonych szeregowo tranzystorów. Pojemności te przedstawia rys. 7.16 na przykładzie bramki NAND.




Rys. 7.16. Pojemności w bramce NAND, które ulegają ładowaniu i rozładowywaniu przy zmianach stanów logicznych

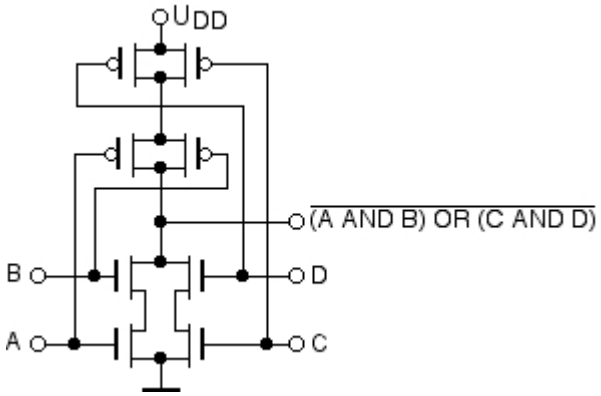
Większa suma pojemności w bramkach wielowejściowych nie pozwala uzyskać krótkich czasów propagacji. Jest to drugi powód, dla którego nie używa się bramek o dowolnej liczbie wejść. Powtórzmy więc jeszcze raz:

**W praktyce nie stosuje się bramek statycznych NOR i NAND o liczbie wejść większej niż 4.**

Podsumujmy: projektowanie bramek NOR i NAND w najprostszym przypadku odbywa się następująco. Dla tranzystorów połączonych równolegle zachowuje się te same wymiary, które określone zostały dla inwertera. Dla tranzystorów połączonych szeregowo zwiększa się szerokość kanału tylokrotnie, ile jest tranzystorów w szeregowym łancuchu. Jeżeli bramka jest obciążona pojemnością zewnętrzną znacznie większą od sumy pojemności wewnętrznych, to dla skrócenia czasów propagacji można poszerzyć kanały tranzystorów. Poszerza się wtedy wszystkie tranzystory w bramce w tej samej proporcji.

## 7.4. Bramki złożone

Przy pomocy połączeń równoległych i szeregowych można zbudować bramki wykonujące funkcje bardziej złożone, niż NOR i NAND. W tym celu zauważmy, że szeregowo łączone tranzystory nMOS można uznać za realizację funkcji AND, a równolegle łączone tranzystory nMOS za realizację funkcji OR. Zatem łącząc równolegle dwa łańcuchy tranzystorów połączonych szeregowo otrzymamy funkcję (A AND B) OR (C AND D). Dla zbudowania kompletnej bramki dodajemy tranzystory pMOS w następujący sposób: każdemu połączeniu szeregowemu tranzystorów nMOS odpowiada połączenie równoległe pMOS, i odwrotnie. Zatem łączymy szeregowo dwie pary równolegle połączonych tranzystorów pMOS. Otrzymujemy w rezultacie schemat jak na rys. 7.17. Bramka realizuje funkcję NOT((A AND B) OR (C AND D)). Tak zbudowane bramki nazywane bywają bramkami **AND-OR-INVERT** (w skrócie **AOI**).



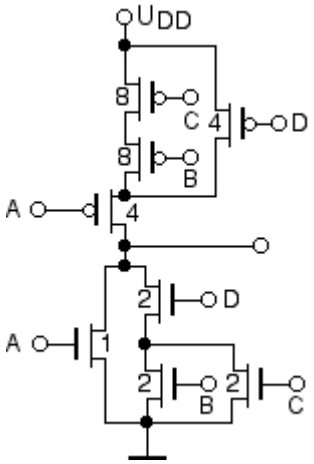

Rys. 7.17. Przykład bramki AND-OR-INVERT

Zamieniając połączenia szeregowe na równoległe, a równoległe na szeregowe otrzymamy bramkę realizującą funkcję NOT((A OR B) AND (C OR D)). Bramki o takiej strukturze nazywane bywają bramkami **OR-AND-INVERT** (w skrócie **OAI**).

Bramki AOI oraz OAI mogą mieć różne liczby wejść, także nieparzyste, i mogą mieć więcej wejść niż 4. Nie należy jedynie budować bramek, w których byłyby łańcuchy szeregowo łączonych tranzystorów o długości większej niż 4.

Poprawnie skonstruowane bramki AOI i OAI mają tę samą cenną właściwość, co inwerter oraz bramki NOR i NAND: statyczny pobór prądu jest bardzo mały, bowiem w stanie ustalonym dla żadnej kombinacji stanów na wejściach nie ma możliwości przepływu prądu ze źródła zasilania. Podobnie jak w bramkach poprzednio omawianych, rownież w bramkach AOI i OAI znaczący pobór prądu występuje jedynie przy zmianach stanów logicznych.

Wymiarowanie tranzystorów w bramkach AOI i OAI polega, tak jak i w przypadku bramek NOR i NAND, na poszerzaniu tranzystorów w połączeniach szeregowych. Wykonuje się to przez znajdowanie w schemacie bramki łańcuchów tranzystorów i poszerzanie ich kanałów odpowiednio do ich liczby w łańcuchu. Rys. 7.18 pokazuje przykład.




Rys. 7.18. Przykład wymiarowania tranzystorów. Dla przejrzystości schematu połączenia bramek tranzystorów z wejściami zaznaczono tylko literami. Liczby oznaczają szerokość kanałów tranzystorów względem pewnej szerokości jednostkowej.

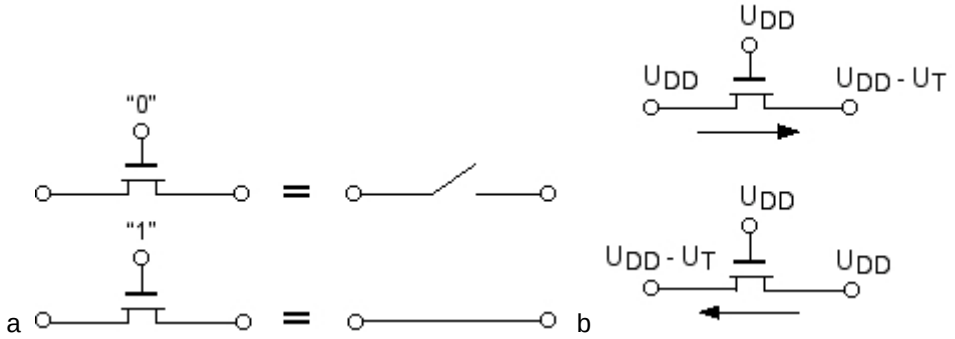
W przykładzie z rys. 7.18 przyjęto założenie, że stosunek ruchliwości nośników $\mu_n/\mu_p$ wynosi 2, czyli w przypadku inwertera tranzystor pMOS powinien mieć kanał 2 razy szerszy od tranzystora nMOS. Znajdujemy najpierw łańcuchy tranzystorów nMOS. Są dwa takie łańcuchy: BD i CD. W obu kanały poszerzamy dwukrotnie. Dla określenia szerokości kanałów tranzystorów pMOS zaczynamy od najkrótszego łańcucha szeregowego: AD. Kanały poszerzamy dwukrotnie, a ponieważ są one i tak 2 razy szersze od kanałów tranzystorów nMOS, otrzymujemy wymiary podane na rys. 7.18. Dla określenia wymiarów tranzystorów B i C zauważmy, że stanowią one połączenie szeregowe z tranzystorem A, który ma już nadany wymiar. Kanały tranzystorów B i C musimy poszerzyć 4 razy, a wtedy ich szeregowe połączenie będzie równoważne tranzystorowi D. Ponieważ tranzystory pMOS są i tak 2 razy szersze od kanałów tranzystorów nMOS, otrzymujemy ostatecznie wymiary podane na rys. 7.18.

Tak nadane wymiary należy traktować jako pierwsze przybliżenie. W przypadku bramek złożonych należy zawsze wykonać symulacje, by sprawdzić, czy charakterystyki przejściowe są do zaakceptowania i czy dostatecznie krótkie są czasy przełączania. Symulacji takich będzie wiele, bo należy sprawdzić wszystkie możliwe kombinacje zmian stanów na wejściach.

Zaletą bramek złożonych jest możliwość realizacji w jednej bramce funkcji bardziej skomplikowanych, niż NOR i NAND. Umożliwia to uproszczenie schematu i zmniejszenie liczby bramek w układzie, co na ogół poprawia także szybkość działania układu. Jednak większość systemów automatycznej syntezy logicznej nie uwzględnia możliwości wykonywania bramek złożonych. Konsekwencja tego jest taka, że i w bibliotekach komórek standardowych takich bramek może nie być. Wówczas pozostaje projektowanie "ręczne" zarówno schematu logicznego, jak i jego fizycznej realizacji.

## 7.5. Bramki transmisyjne i trójstanowe

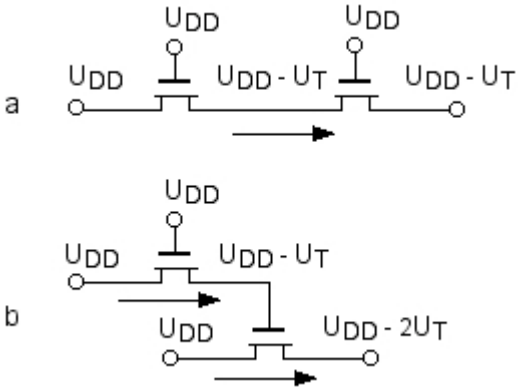
**Bramki transmisyjne** działają inaczej, niż bramki dotąd omawiane. Są one odpowiednikami sterowanego wyłącznika. W zależności od stanu logicznego sygnału sterującego bramka transmisyjna przepuszcza sygnał z wejścia na wyjście lub nie. Najprostszą bramką transmisyjną jest pojedynczy tranzystor nMOS (rys. 7.19).




Rys. 7.19. Tranzystor nMOS jako najprostsza bramka transmisyjna: (a) zasada działania, (b) zjawisko degradacji jedynki logicznej

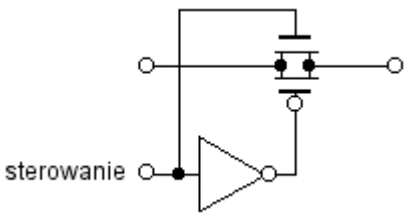
Bramka transmisyjna nie wykazuje kierunkowości - jeśli jest otwarta, może transmitować sygnał w obu kierunkach. Może to być wygodne w niektórych zastosowaniach, ale najczęściej jest to właściwość kłopotliwa.

Bramka transmisyjna w postaci pojedynczego tranzystora nMOS ma ponadto istotną wadę: wprowadza degradację poziomu jedynki logicznej (patrz rys. 7.19b). Degradacja polega na zmniejszeniu napięcia jedynki o wartość równą w przybliżeniu napięciu progowemu tranzystora $U_T$, a bierze się stąd, że aby tranzystor przewodził, musi istnieć między bramką i źródłem różnica napięć równa co najmniej $U_T$. Dlatego nie można sterować bramki transmisyjnej sygnałem już zdegradowanym, np. pochodzącym z wyjścia innej bramki transmisyjnej. Można natomiast łączyć bramki transmisyjne szeregowo - patrz rys. 7.20.




Rys. 7.20. Dozwolone (a) i niedozwolone (b) łączenie bramek transmisyjnych degradujących jedynkę logiczną

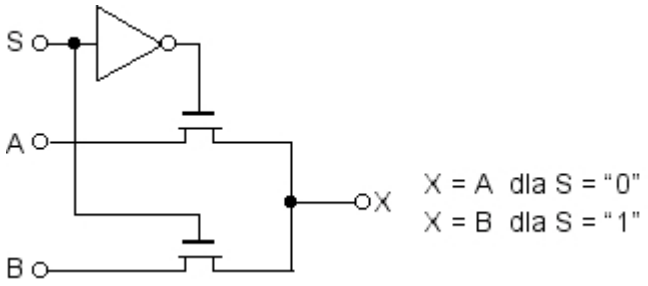
Degradację jedynki można usunąć stosując **pełną bramkę transmisyjną CMOS** złożoną z pary tranzystorów - nMOS i pMOS - połączonych równolegle i sterowanych przeciwnymi stanami logicznymi. Tranzystor pMOS zapewnia transmisję jedynki logicznej bez degradacji, a tranzystor nMOS podobnie zapewnia transmisję zera. Jednak cała bramka łącznie z niezbędnym do jej sterowania inwerterem ma aż 4 tranzystory (rys. 7.21).




Rys. 7.21. Pełna bramka transmisyjna CMOS

Żadna bramka transmisyjna nie ma właściwości regeneracji poziomów logicznych. Niekiedy na wyjściu bramki transmisyjnej, zwłaszcza bramki w postaci pojedynczego tranzystora nMOS, stosowany jest inwerter pełniący rolę bufora regenerującego poziomy logiczne.

Bramki transmisyjne pozwalają prosto realizować niektóre funkcje kombinacyjne, np. układy multiplekserów i demultiplekserów. Przykład pokazuje rys. 7.22.




Rys.7.22. Multiplekser na bramkach transmisyjnych

Układ z rys. 7.22 pokazuje prostotę realizacji multipleksera przy użyciu bramek transmisyjnych, ale zarazem pozwala też zilustrować pewne niebezpieczeństwa takiej realizacji. Układ działa prawidłowo pod warunkiem, że nie dochodzi do sytuacji równoczesnego włączenia obu bramek transmisyjnych. Gdyby obie bramki były włączone równocześnie, a stany wejść A i B byłyby różne, na wyjściu powstałby niedopuszczalny konflikt. W dodatku ze względu na dwukierunkową transmisję sygnału przez bramkę transmisyjną sygnał z wejścia A oddziaływałby bezpośrednio na stan wejścia B, i odwrotnie. Skutki tej niedozwolonej sytuacji nie są możliwe do przewidzenia bez znajomości szczegółów budowy bramek dostarczających sygnały na wejścia A i B. W stanie ustalonym do równoczesnego włączenia obu bramek oczywiście dojść nie może, natomiast może się to zdarzyć podczas zmiany stanu logicznego wejścia sterującego. Wyobraźmy sobie, że początkowy stan wejścia sterującego S to "0". Włączona jest wówczas górna bramka transmisyjna, zaś dolna - wyłączona. Gdy stan wejścia S zmienia się na "1", dolna bramka zostaje włączona, a górna wyłączona, ale górna bramka wyłączona jest z opóźnieniem wynikającym z niezerowego czasu propagacji sygnału w inwerterze. W rezultacie może pojawić się taki odcinek czasu, w którym obie bramki transmisyjne przewodzą równocześnie. Nie musi, ale może spowodować to błędne działanie układu z multiplekserem - zależy to od tego, jak zbudowany jest cały układ.

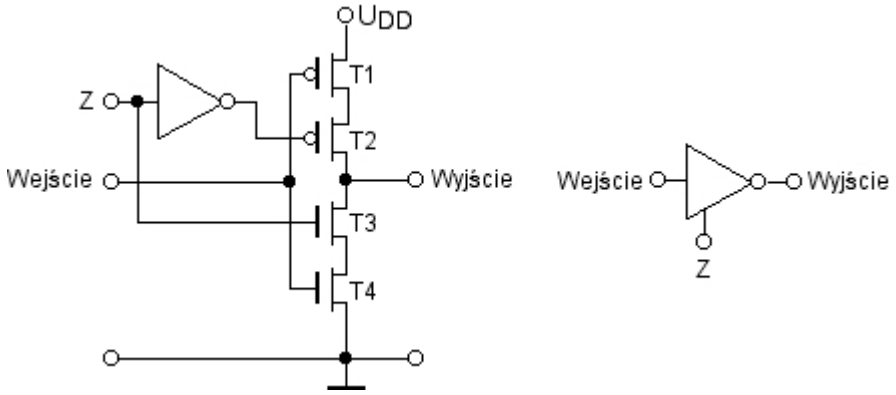
Jak widać z tego przykładu, logika kombinacyjna budowana przy użyciu bramek transmisyjnych wymaga ostrożności, starannego przemyślenia działania układu zarówno w stanach ustalonych, jak i w stanach przejściowych oraz symulacji elektrycznej dla wychwycenia ewentualnych sytuacji błędnych i niedopuszczalnych. Niemniej bramki transmisyjne są dość powszechnie używane, są one niezbędne w niektórych rodzajach bramek dynamicznych, rejestrów itp. Będzie o tych układach mowa w dalszych wykładach.

Wymiarowanie tranzystorów w bramkach transmisyjnych jest bardzo proste. Wymiary dobierane są tak, by uzyskać jak najmniejsze opóźnienia sygnału wynikające z tego, że tranzystory bramki transmisyjnej wnoszą w tor sygnału pewną nieliniową rezystancję i pewną pojemność. Na ogół używa się tranzystorów o minimalnych dopuszczalnych wymiarach. Poszerzanie kanałów ponad minimalną szerokość nie jest celowe, wraz z szerokością kanału rosną bowiem proporcjonalnie pojemności złącz źródła i drenu oraz pojemność $C_{DG}$, zatem nie uzyskuje się skrócenia czasu propagacji sygnału. Jedynym wyjątkiem jest sytuacja, gdy wyjście bramki transmisyjnej jest obciążone dużą pojemnością, znacznie przekraczającą pojemności wewnętrzne tranzystorów bramki. Opóźnienie wnoszone przez bramkę można z grubsza utożsamić z jej stałą czasową $R_t C_l$, gdzie $R_t$ jest rezystancją wnoszoną przez bramkę, a $C_l$ - pojemnością na wyjściu bramki. Rezystancję wnoszoną przez pełną dwutranzystorową bramkę CMOS można w pierwszym przybliżeniu oszacować z bardzo prostej zależności

$$R_t = \frac{1}{K_n\left(U_{DD} - U_{Tn}\right) + K_p\left(U_{DD} - \left|U_{Tp}\right|\right)} \tag{7.15}$$

Zależność ta może służyć tylko do przybliżonych szacunków stałej czasowej, bowiem rezystancja wnoszona przez bramkę jest w rzeczywistości nieliniowa.

Specjalnym rodzajem bramek statycznych są **bramki trójstanowe**. Są to bramki, których wyjście może być w stanie zera, jedynki lub wysokiej impedancji. W tym ostatnim przypadku wyjście bramki może być uważane za odłączone od układu. Pozwala to na przykład dołączyć do tego samego węzła elektrycznego wyjścia kilku bramek, z których w każdym momencie wszystkie z wyjątkiem jednej są w stanie wysokiej impedancji. Najprostszą bramką trójstanową jest inwerter trójstanowy. Jego schemat pokazany jest na rys. 7.23. Można go traktować jako zwykły inwerter skojarzony z bramką transmisyjną.




Rys. 7.23. Inwerter trójstanowy: schemat i symbol, jaki będzie używany w tym wykładzie

Tranzystory T1 i T4 tworzą zwykły inwerter. Tranzystory T2 i T3 są sterowane sygnałem z dodatkowego wejścia Z w taki sposób, że albo oba są włączone (dla Z = "1") albo oba są wyłączone (dla Z = "0"). W pierwszym przypadku inwerter działa w zwykły sposób. W drugim przypadku w węźle wyjściowym panuje stan wysokiej impedancji, ponieważ żaden z tranzystorów T2 i T3 nie przewodzi. Taki inwerter projektuje się tak samo, jak zwykły inwerter dwutranzystorowy, po czym przyjmuje się szerokość kanałów wszystkich tranzystorów dwukrotnie większą, niż w zwykłym inwerterze, zgodnie z zasadą poszerzania kanałów tranzystorów w połączeniach szeregowych.

Tak oto poznaliśmy podstawowe rodzaje statycznych bramek CMOS. W następnym wykładzie będzie mowa o bramkach zwanych dynamicznymi, a także o przerzutnikach - układach pozwalających zapamiętać pojedyncze bity.

# ZADANIA I ĆWICZENIA DO WYKŁADU 7

**Zadanie 1**

Dana jest technologia CMOS, w której $U_{Tn}$ = 0,75 V, $U_{Tp}$ = - 0,85 V, $\mu_n/\mu_p$ = 2,9, minimalna długość kanału tranzystora $L$ = 0,7 µm, minimalna szerokość kanału tranzystora $W$ = 1µm. Oblicz napięcie przełączania inwertera z tranzystorami o jednakowych, minimalnych wymiarach kanałów dla napięcia zasilania $U_{DD}$ = 5 V, a następnie dla obniżonego napięcia zasilania $U_{DD}$ = 1,5 V. Zastanów się nad sensem rozwiązania dla napięcia 1,5 V - zwróć uwagę na napięcia progowe tranzystorów!
Wskazówka: należy posłużyć się wzorem (7.3).

**Zadanie 2**

Oblicz szerokość kanału tranzystora pMOS inwertera, dla technologii o danych z zadania 1, przy której to szerokości otrzymuje się napięcie przełączania dokładnie równe 0,5 $U_{DD}$ (zwróć uwagę na niejednakowe napięcia progowe!). Wykonaj obliczenia dla napięcia zasilania $U_{DD}$ = 5 V, a następnie dla obniżonego napięcia zasilania $U_{DD}$ = 1,5 V. Zastanów się nad sensem rozwiązania dla napięcia 1,5 V - zwróć uwagę na napięcia progowe tranzystorów!

**Zadanie 3**

Oszacuj czasy propagacji sygnałów dla inwertera o minimalnych wymiarach tranzystorów (dane jak w zad. 1) i inwertera o szerokości kanału tranzystora pMOS obliczonej w zadaniu 2, dla napięcia zasilania $U_{DD}$ = 5 V. Przyjmij $\mu_n C_{ox}$ = 80 µA/V$^2$, $\mu_p C_{ox}$ = 27 µA/V$^2$. Załóż, że oba inwertery są obciążone całkowitą pojemnością równą 50 fF.
Wskazówka: należy posłużyć się wzorami (7.7) i (7.8).

**Zadanie 4**

Oblicz napięcie przełączania przy równoczesnym przełączaniu wszystkich wejść dla bramek NOR i NAND o liczbie wejść 2, 4 i 6, zakładając że tranzystory nMOS mają minimalne wymiary, a tranzystory pMOS mają minimalną długość kanału i szerokość obliczoną w zadaniu 2 dla napięcia zasilania $U_{DD}$ = 5 V.
Wskazówka: należy posłużyć się wzorami (7.13) i (7.14).

**Zadanie 5**

Oblicz zastępczą rezystancję pełnej bramki transmisyjnej CMOS dla tranzystorów o minimalnych wymiarach i danych jak w zadaniach wyżej.
Wskazówka: należy posłużyć się wzorem (7.15).

# ĆWICZENIE 1 DO WYKŁADU 7

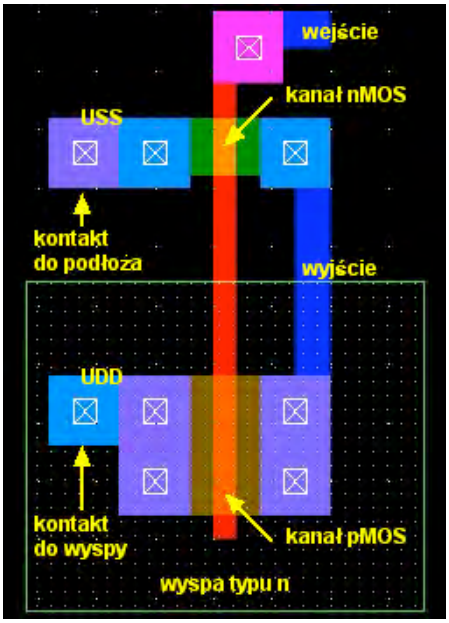
**Cel ćwiczenia**

W tym ćwiczeniu zaprojektujesz inwerter CMOS i wykonasz jego symulacje, posługując się programem "Microwind".

**Przebieg ćwiczenia**

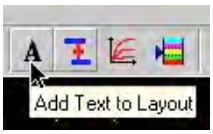
W sekcji "Wykłady VIDEO" znajdziesz prezentację wprowadzającą do ćwiczenia.

Uruchom program "Microwind2". Będziemy nadal wykorzystywać technologię CMOS 0,8 mikrometra . Musimy więc wczytać odpowiedni plik technologiczny.Z menu "File" wybierz "Select Foundry" iotwórz plik "cmos08.rul". Następnie wczytaj z dysku plik z projektem tranzystora nMOS, który był wykonany w ćwiczeniu 1 do wykładu 6. Dorysuj do tego tranzystora tranzystor pMOS tak, aby powstał kompletny inwerter (możesz wykorzystać tranzystor pMOS zaprojektowany w ćwiczeniu 2 do wykładu 6). Staraj się zaprojektować tranzystor pMOS o szerokości kanału 2 razy większej, niż szerokość kanału tranzystora nMOS. Twój projekt może wyglądać na przykład tak (kolorem żółtym oznaczono i opisano najważniejsze obszary):

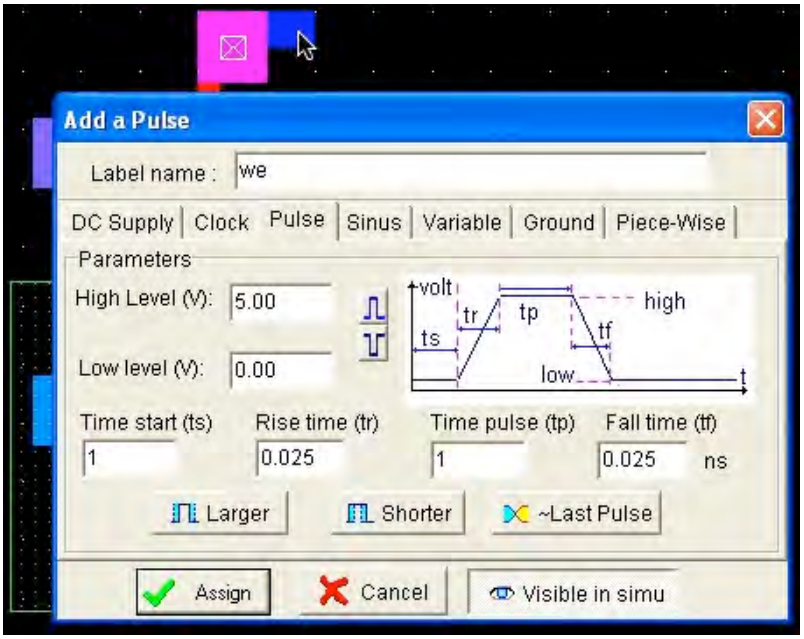



Teraz musisz przygotować projekt do symulacji. W tym celu trzeba oznaczyć "masę", zasilanie, wejście i wyjście i przypisać im odpowiednie sygnały elektryczne. Mówimy, że węzłom elektrycznym przypisane będą etykiety. Oto opis postępowania:

Wybierz z paska narzędzi dodawanie etykiet (literka A)

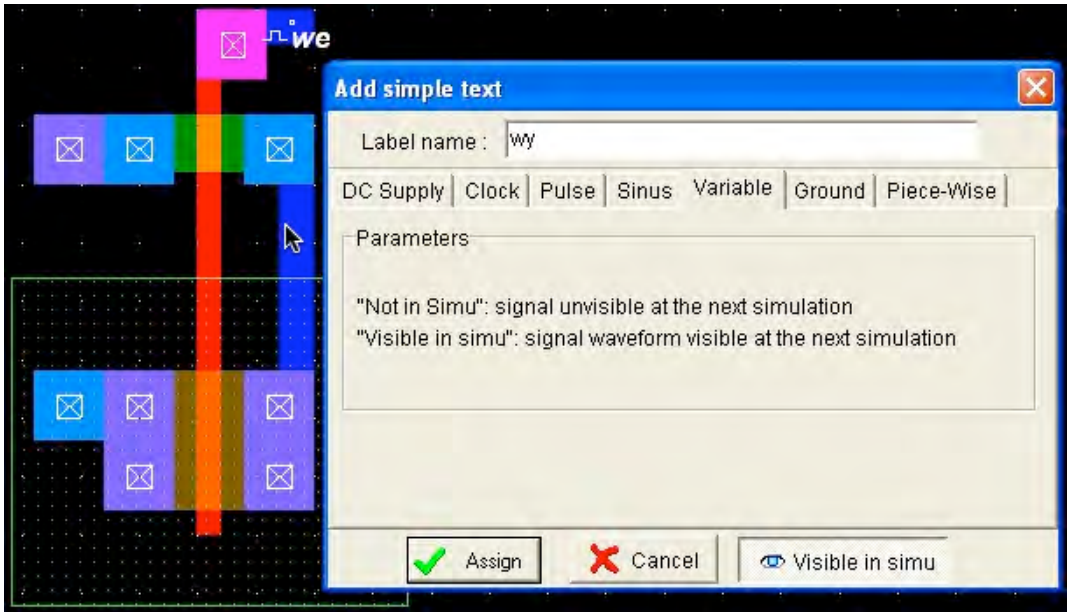



Teraz ustaw kursor tak, by wskazywał obszar metalu połączony z wejściem, i kliknij. W oknie, które się otworzy, wybierz zakładkę "Pulse". Jako tekst etykiety możesz wpisać dowolną nazwę, np "we". Przy symulacji w dziedzinie czasu do wejścia doprowadzony będzie sygnał o parametrach jak pokazano w oknie. Ustaw czas startu impulsu oraz jego długość na 1 ns, czas narastania oraz czas opadania impulsu na 0,025 ns. Kliknij także w przycisk "Not in simulation", aż zmieni się na "Visible in simu". Oznacza to, że ten sygnał elektryczny będzie pokazany na ekranie podczas symulacji. Niczego innego nie musisz zmieniać. Na koniec naciśnij "Assign".

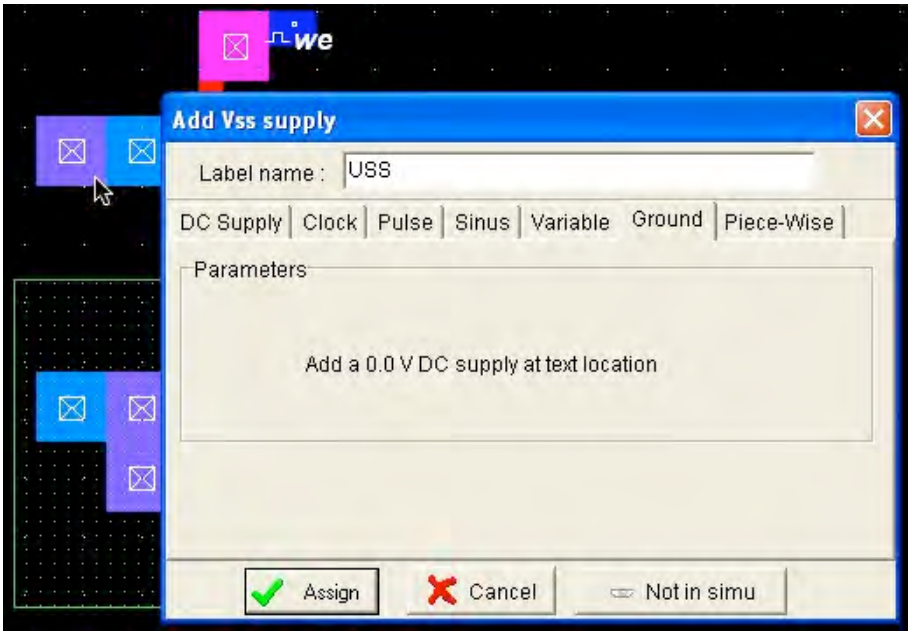



Teraz w ten sam sposób zaznaczysz etykietą wyjście, "masę" i zasilanie.

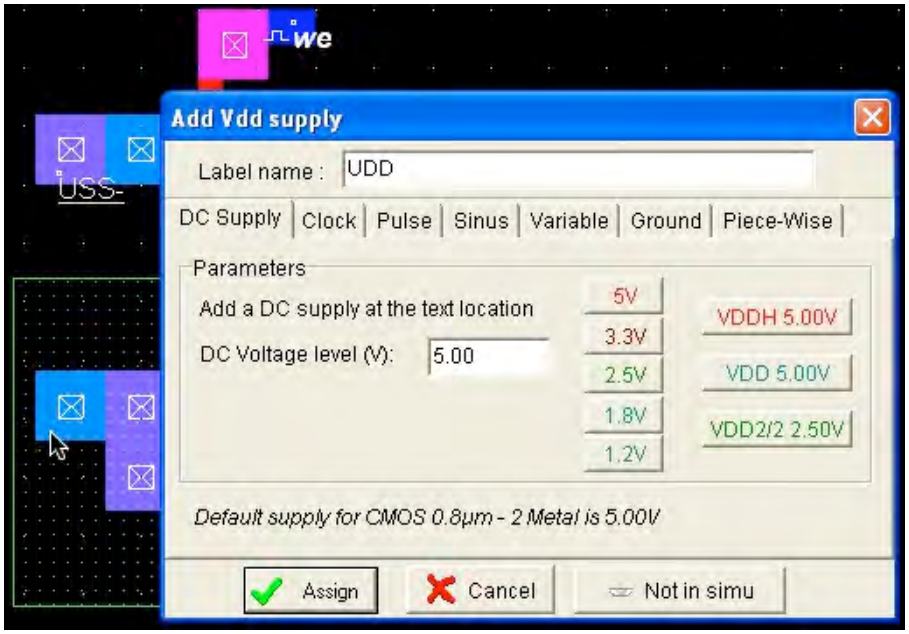
Dla wyjścia wybierz zakładkę "Variable". Oznacza ona, że do tego węzła elektrycznego nie doprowadza się z zewnątrz żadnego sygnału.



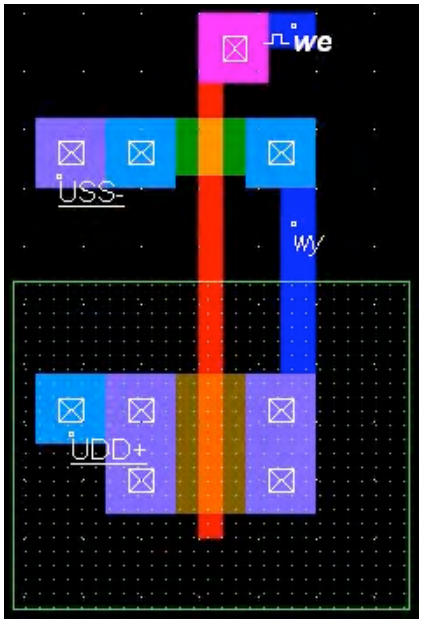

Dla węzła "masy" wybierz zakładkę "Ground".




Dla węzła zasilania wybierz zakładkę "DC Supply". Napięcie powinno wynosić 5 V.



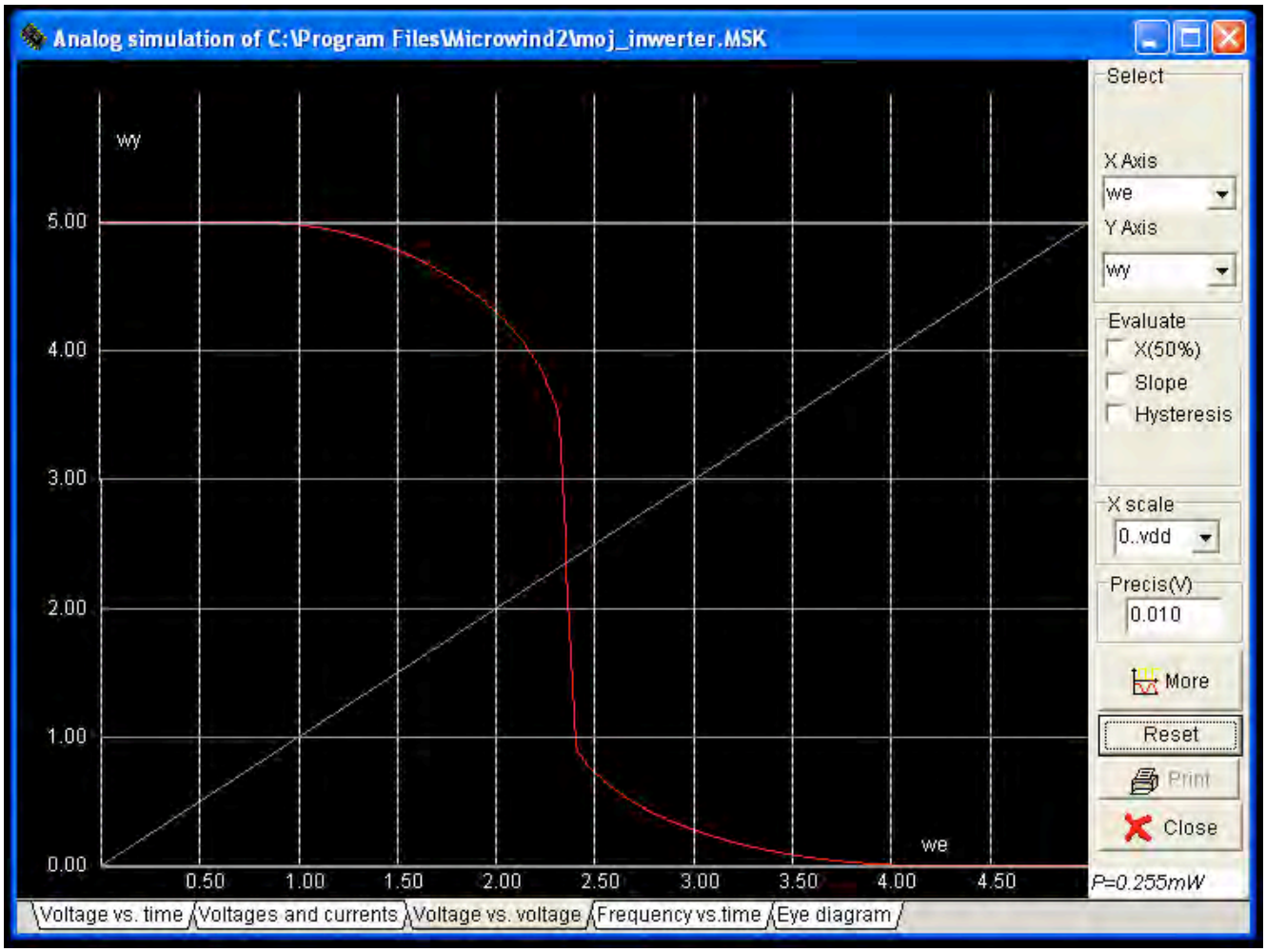

Po dodaniu wszystkich etykiet Twój projekt powinien wyglądać mniej więcej tak:




Nie zapomnij po zakończeniu rysowania sprawdzić, czy spełnione są reguły projektowania! W razie potrzeby zrób poprawki.

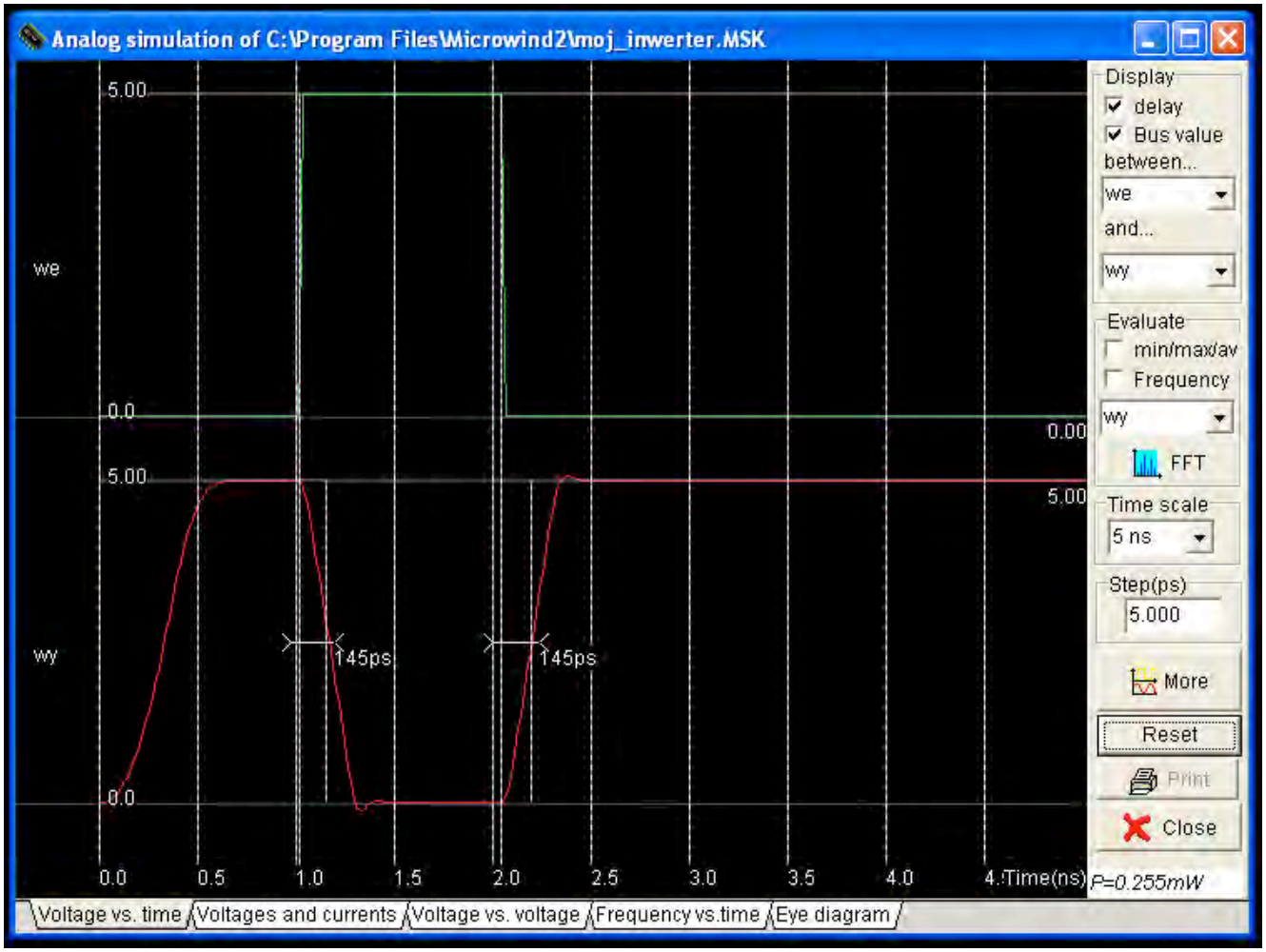

Teraz możesz przystąpić do symulacji.
Wybierz z paska narzędzi symbol symulatora

Otworzy się okno symulatora. Wybierz w nim zakładkę "Voltage vs. voltage". W menu "X Axis" powinna być widoczna nazwa węzła wejściowego, w menu "Y Axis" - wyjściowego. W razie potrzeby dokonaj odpowiednich zmian. Powinna zostać narysowana charakterystyka przejściowa zaprojektowanego inwertera. Jeśli rysunek nie narysował się lub wygląda nie tak, jak powinien, naciśnij "Reset". Oto wynik, jaki powinien być uzyskany, jeśli projekt jest prawidłowy, a tranzystor pMOS jest 2 razy szerszy od tranzystora nMOS:

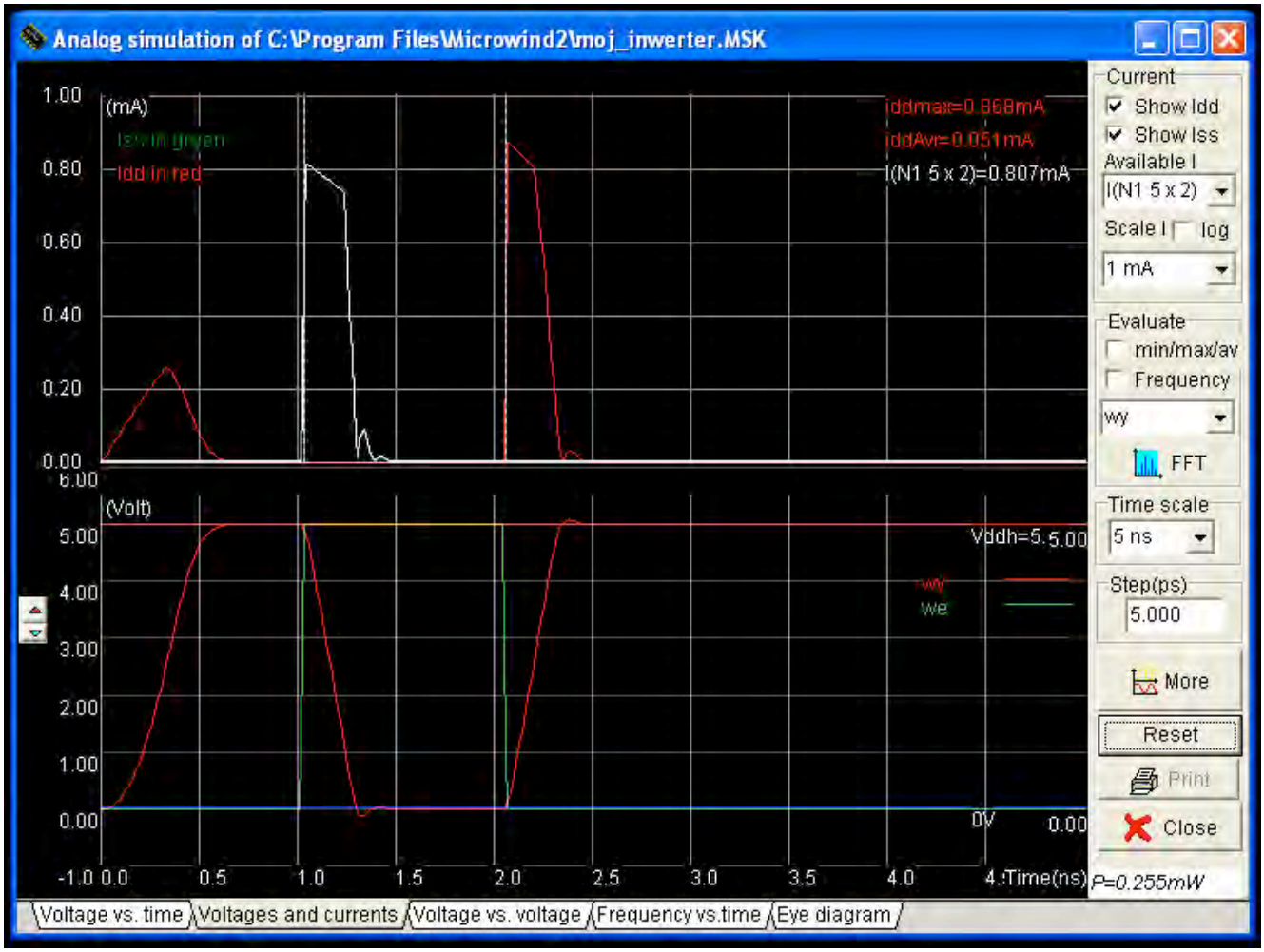


Zauważ, że napięcie przełączania inwertera wynosi około 2,4 V, jest bliskie połowie napięcia zasilania

Teraz wybierz zakładkę "Voltage vs. time" - symulacja w dziedzinie czasu. Powinny narysować się przebiegi napięcia wejściowego i wyjściowego w funkcji czasu. Jeśli rysunek nie wygląda tak jak niżej, naciśnij "Reset". Zauważ, że symulator wyznaczył czasy propagacji sygnału: 145 ps.




Oprócz napięć możesz obejrzeć także prądy w funkcji czasu. Naciśnij "Voltages and currents" i obserwuj wyniki symulacji.




Zauważ, jak dużą amplitudę mają impulsy prądu. W prawym górnym rogu okna możesz wybierać, które prądy chcesz oglądać.

Teraz już wiesz, jak będzie zachowywał się zaprojektowany przez Ciebie inwerter. Dla nabrania wprawy popróbuj różnych opcji - wyboru prądów, zmiany skali ("Time scale"), zmiany kroku symulacji ("Step (ps)") itd.

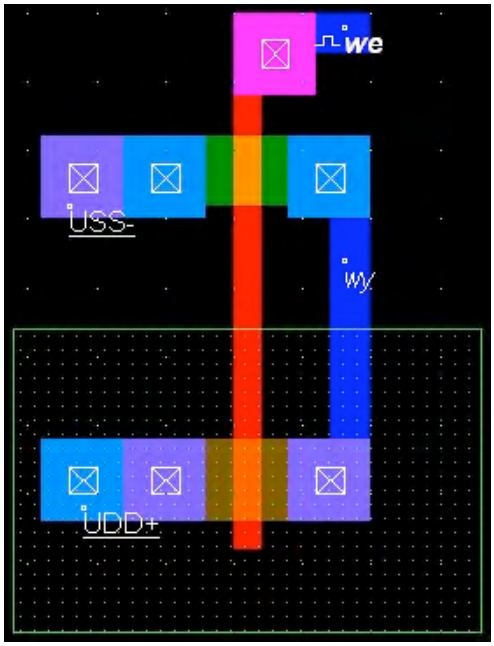
## ĆWICZENIE 2 DO WYKŁADU 7

**Cel ćwiczenia**

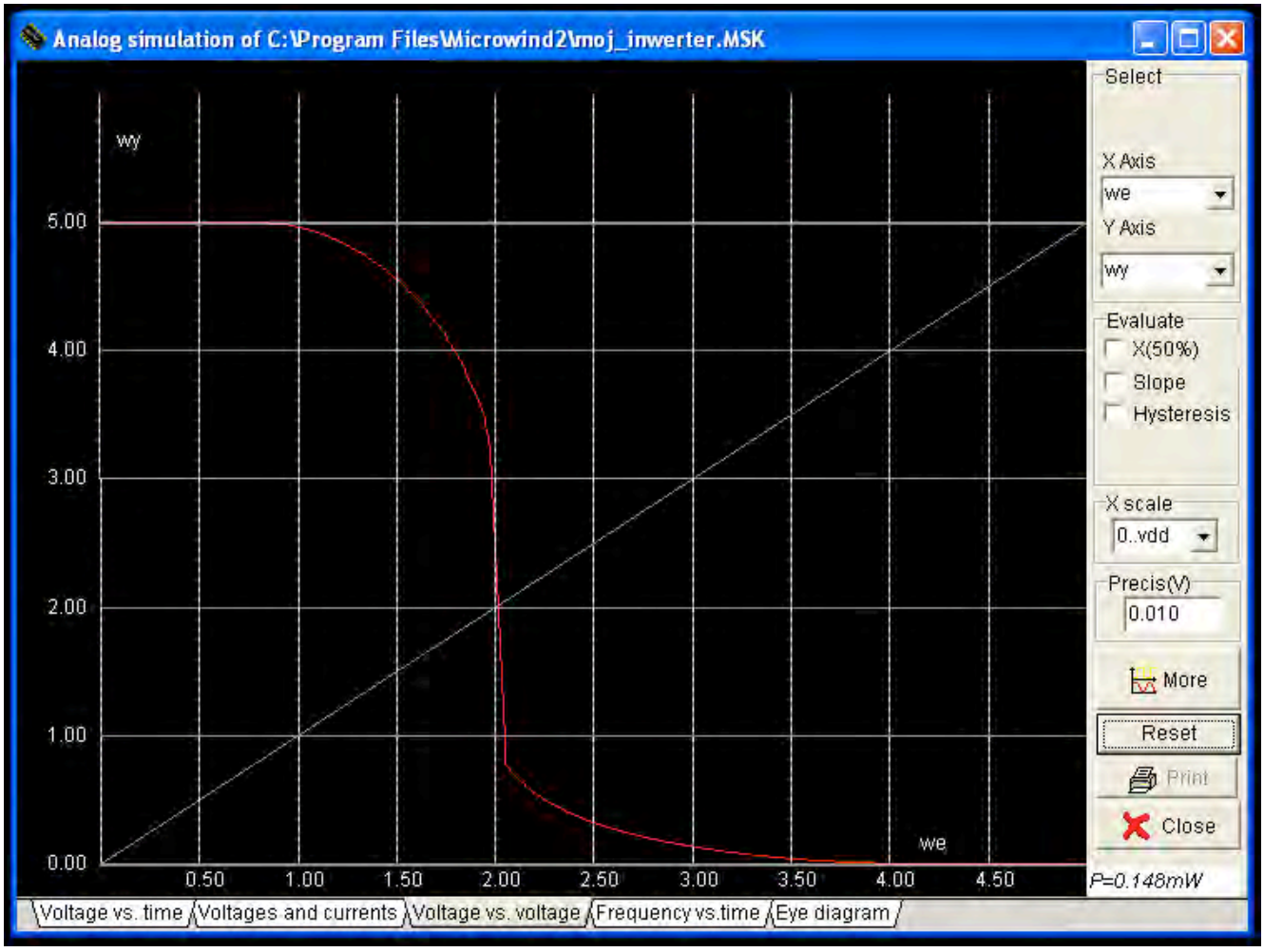
Samodzielny trening

**Przebieg ćwiczenia (do samodzielnego wykonania)**

Zamknij okno symulatora i przerób projekt topografii w taki sposób, by szerokości kanałów obu tranzystorów były jednakowe. Sprawdź, czy spełnione są reguły projektowania. Następnie powtórz symulacje. Zauważ różnice, w szczególności różnicę w charakterystyce przejściowej. Jeśli wszystko zostało wykonane prawidłowo, Twój projekt powinien wyglądać mniej więcej tak:




a wynik symulacji charakterystyki przejściowej - tak:




Zauważ, że charakterystyka przesunęła się. Napięcie przełączania inwertera wynosi teraz około 2 V.

Teraz już możesz próbować projektować większe układy! Niedługo przyjdzie na to czas.

## Bibliografia

[1] M. J. Patyra, "*Projektowanie układów MOS w technice VLSI*", Wydawnictwa Naukowo-Techniczne, 1993
[2] E. Sicard, "*Microwind & Dsch User's Manual*", National Institute of Applied Sciences INSA, Toulouse, 2003
(Jest to podręcznik do oprogramowania wykorzystywanego w tym wykładzie, załączony na płycie w postaci pliku PDF)
[3] J. M. Rabaey, "*Digital Integrated Circuits, A Design Perspective*", Prentice Hall, Inc. 1996
[4] J. P. Uyemura, "*CMOS Logic Circuit Design*", Kluwer Academic Publishers, 2001

# Wykład 8: Bramki dynamiczne i przerzutniki

## Wstęp

Bramki statyczne, omawiane w poprzednim wykładzie, nie są jedynym sposobem realizacji układów kombinacyjnych CMOS. W pewnych przypadkach stosuje się układy zwane dynamicznymi. Będą one omawiane w wykładzie 8. Terminem "układy dynamiczne" określana jest szeroka klasa układów, w których wartości logiczne - zera i jedynki - są reprezentowane przez ładunki gromadzone w pojemnościach. Jednak głównym obszarem zastosowań układów dynamicznych nie jest logika kombinacyjna, lecz układy sekwencyjne. Praktycznie każdy układ cyfrowy jest układem sekwencyjnym, tj. zawiera elementy pamięciowe: przerzutniki, rejestry, a nawet całe bloki pamięci. Wykład 8 zawiera więc także omówienie tych podstawowych elementów: przerzutników i rejestrów, zarówno statycznych, jak i dynamicznych.

Jedną z cech układów dynamicznych, a także wszystkich rodzajów układów pamięciowych jest konieczność ich taktowania. W wykładzie 8 będzie więc też mowa o zegarach - sygnałach taktujących, ich generacji i problemach związanych z ich rozprowadzeniem w układzie.

Budowa typowych komórek pamięci, organizacja pamięci, układy odczytu, zapisu itp. będą omawiane w następnym wykładzie.

Uzupełnieniem wykładu są ćwiczenia, w których nauczysz się posługiwać symulatorem logicznym. Wykonanie ćwiczeń będzie łatwiejsze, jeśli przedtem obejrzysz prezentację wideo.

## 8.1. Istota układów dynamicznych, bramki typu DOMINO

**Układami dynamicznymi** nazywamy takie układy, w których wartości logiczne są reprezentowane przez ładunek zgromadzony w pojemności. Z reguły przyjmowana jest konwencja: "0" - brak ładunku (pojemność nie naładowana), "1" - pojemność naładowana. Istnieje wiele rodzajów układów dynamicznych, w różny sposób wykorzystujących przechowywanie wartości logicznych jako ładunku. W niektórych z nich jest to jedynie potrzebne pomocniczo, w krótkich odcinkach czasu (np. podczas zmiany stanu). W innych stanowi podstawę działania - jak na przykład w komórkach pamięci dynamicznych RAM.

Układy dynamiczne, niezależnie od budowy i przeznaczenia, mają pewne wspólne cechy:

1. Pojemności, w których gromadzony jest ładunek, są zawsze związane z pewnymi upływnościami, takimi jak prądy wsteczne złącz p-n i prądy podprogowe tranzystorów MOS. W związku z tym czas przechowywania ładunku w pojemnościach jest ograniczony. Naładowana do napięcia $U_{DD}$ pojemność ulega stopniowemu rozładowaniu. Czas tego rozładowania może być mierzony milisekundami w temperaturze otoczenia, ale maleje do mikrosekund dla temperatur o kilkadziesiąt stopni wyższych, bo ze wzrostem temperatury bardzo szybko rosną prądy upływu. Wynika z tego, że wartość "1" zapisana jako ładunek w pojemności powinna być w krótkim czasie odczytana i wykorzystana, a jeżeli ma być przechowywana przez długi czas, to wymagać będzie okresowego odświeżania (czyli odczytu i ponownego zapisu tej samej wartości).

2. Układy dynamiczne wymagają taktowania, konieczny jest więc sygnał **zegara**. Większość układów dynamicznych wymaga przy tym dość precyzyjnego taktowania, co stwarza problemy z generacją sygnałów zegarowych i ich rozprowadzeniem w dużych układach.

3. W związku z tym, że czas przechowywania ładunku w pojemności jest ograniczony, układy dynamiczne nie mogą działać z dowolnie małą częstotliwością zegara. Zbyt mała częstotliwość powoduje błędy w działaniu układu.

4. W układach dynamicznych nie mamy do czynienia ze statycznymi charakterystykami przejściowymi, a odporność na zakłócenia jest definiowana inaczej, niż dla bramek statycznych.

5. Układy dynamiczne mogą wprowadzać degradację jedynki logicznej, zarówno ze względu na rozładowywanie pojemności w czasie, jak i ze względu na zjawisko **podziału ładunku**. Polega ono na tym, że przy odczycie pojemność, w której zgromadzony został ładunek reprezentujący stan "1", zostaje połączona równolegle z inną, często znacznie większą pojemnością. Zgodnie z prawami elektrostatyki napięcie na pojemności jest proporcjonalne do ładunku i odwrotnie proporcjonalne do pojemności

$$U = \frac{Q}{C} \tag{8.1}$$

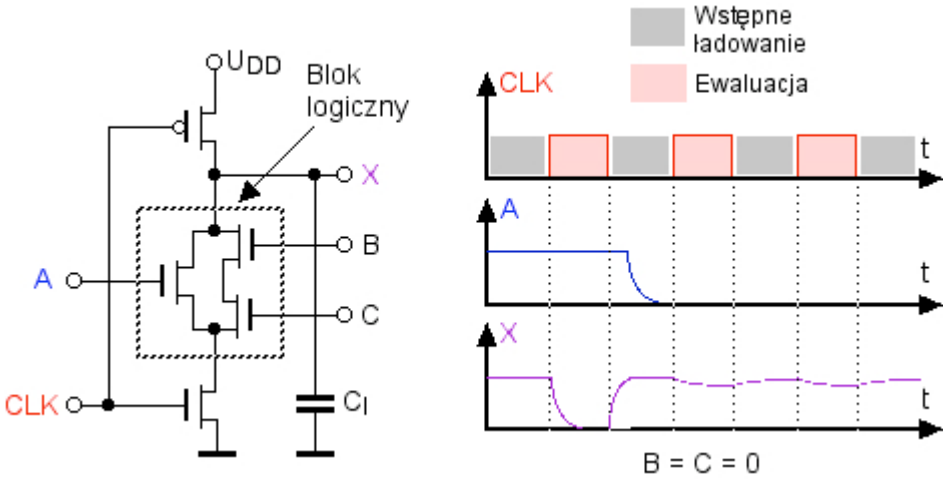
a więc jeśli pewna pojemność $C_1$ zostanie naładowana do napięcia $U_{DD}$ i zostanie w niej zgromadzony ładunek $Q=C_1U_{DD}$, a potem zostanie ona połączona równolegle z drugą pojemnością $C_2$, to napięcie spadnie do wartości

$$U = U_{DD}\frac{C_1}{C_1 + C_2} \tag{8.2}$$

Jeżeli $C_2 > C_1$, degradacja jest znaczna i wymaga zastosowania specjalnych układów regenerujących poziom jedynki. Jest to bardzo poważny problem na przykład w pamięciach dynamicznych RAM.

Wymienione wyżej cechy układów dynamicznych powodują, że projektowanie tych układów jest trudniejsze, niż bramek statycznych. Niemniej, układy dynamiczne mają szereg zalet i dlatego warto się z nimi zapoznać.

Dobrym przykładem układów dynamicznych są dość często stosowane w praktyce dynamiczne bramki kombinacyjne zwane bramkami typu **DOMINO**. Przykładowa bramka typu DOMINO pokazana jest na rys. 8.1.




Rys. 8.1. Schemat przykładowej bramki typu DOMINO i przebiegi napięć w funkcji czasu (przy złożeniu że na wejściach B i C są cały czas zera)

Bramka składa się z dwóch części: bloku logicznego, w którym odpowiednie połączenia szeregowe i równoległe tranzystorów nMOS określają wykonywaną funkcję kombinacyjną, oraz dwóch dodatkowych tranzystorów - dolnego nMOS i górnego pMOS, które są okresowo na zmianę włączane i wyłączane sygnałem zegara CLK. Działanie bramki odbywa się w dwóch fazach: **wstępnego ładowania** (gdy zegar CLK jest w stanie "0") i **ewaluacji** (gdy zegar CLK jest w stanie "1"). W fazie wstępnego ładowania włączony jest górny tranzystor pMOS, a dolny tranzystor nMOS jest wyłączony. Niezależnie od stanów wejść A, B i C pojemność $C_l$ obciążająca węzeł wyjściowy X ładuje się ze źródła zasilania do napięcia $U_{DD}$ reprezentującego "1". Gdy stan zegara zmienia się z "0" na "1", zaczyna się faza ewaluacji. Górny tranzystor pMOS zostaje wyłączony, a włącza się dolny tranzystor nMOS. Teraz stan na wyjściu zależy od stanów wejść A, B i C. Jeśli tranzystor A jest włączony, pojemność $C_l$ rozładowuje się. To samo dzieje się, gdy włączone są równocześnie tranzystory B i C. W tych przypadkach na wyjściu pojawia się (po upływie czasu potrzebnego na rozładowanie pojemności $C_l$ ) zero. Jeśli stany wejść są takie, że pojemność nie może się rozładować, na wyjściu pozostaje stan "1". Rys. 8.1 pokazuje działanie przykładowej bramki. Po pierwszej fazie wstępnego ładowania tranzystor A jest włączony, więc w fazie ewaluacji na wyjściu pojawia się "0". Po dwóch następnych fazach wstępnego ładowania wszystkie trzy tranzystory A, B i C są wyłączone, więc na wyjściu w fazach ewaluacji utrzymuje się "1".

Bramka omawianego typu ma kilka istotnych zalet:

- przy liczbie wejść większej od 2 mniej tranzystorów niż w bramkach statycznych,
- zawsze tylko jeden tranzystor pMOS,
- nie ma potrzeby utrzymywania określonej proporcji wymiarów tranzystorów nMOS i pMOS, w wielu przypadkach tranzystor pMOS może mieć mniejszą szerokość kanału, niż tranzystory pMOS w bramkach statycznych,
- możliwość zmniejszenia zajętej powierzchni w stosunku do bramek statycznych,
- tylko jeden tranzystor dołączony do każdego wejścia logicznego, stąd mniejsza pojemność wejściowa obciążająca poprzednią bramkę i potencjalnie większa szybkość działania
- pojemność $C_l$, niezbędna do działania bramki, nie wymaga wykonania w postaci odrębnego elementu - wystarczają "naturalne" pojemności dołączone do węzła X, takie jak pojemności złącz p-n drenów tranzystorów.

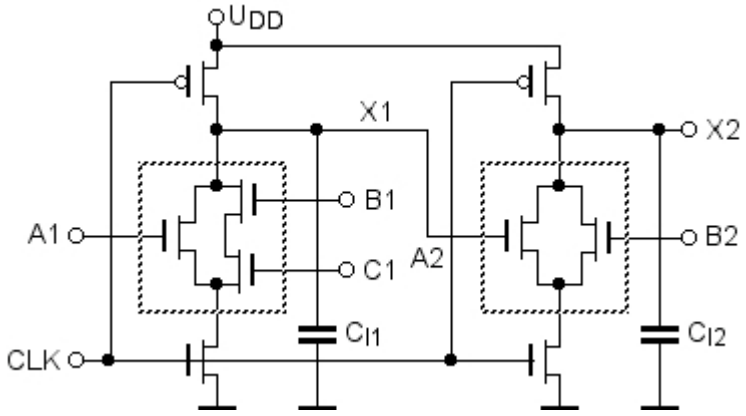
! **Przy umiejętnym zaprojektowaniu bramki dynamiczne są najszybciej działającymi bramkami kombinacyjnymi CMOS. Są więc często spotykane w kluczowych dla szybkości działania blokach układów takich, jak np. mikroprocesory.**

Projektowanie układów z bramkami dynamicznymi jest bardziej skomplikowane, niż w przypadku bramek statycznych. Wynika to przede wszystkim z konieczności ścisłego przestrzegania zależności czasowych między zegarem, a sygnałami logicznymi. I tak, sygnały na wejściach mogą się swobodnie zmieniać tylko podczas fazy wstępnego ładowania, natomiast muszą być stabilne w fazie ewaluacji. Sygnały wyjściowe mają prawidłową wartość tylko podczas fazy ewaluacji (po upływie czasu ewentualnego rozładowania pojemności $C_l$ ).

Dodatkowym utrudnieniem jest to, że

**! wyjść bramek dynamicznych takich, jak pokazano na rys. 8.1, nie wolno łączyć bezpośrednio z wejściami bramek tego samego rodzaju.**

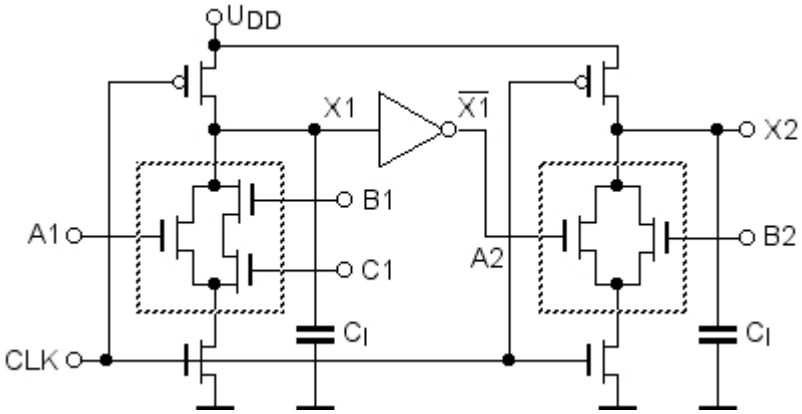
Skąd to ograniczenie? Bierze się ono stąd, że w układzie kombinacyjnym typu DOMINO *wszystkie bramki taktowane są tym samym sygnałem zegara*. Wobec tego *fazy ewaluacji zaczynają się we wszystkich bramkach w tym samym momencie*. Problem powstaje stąd, że *w chwili rozpoczęcia fazy ewaluacji wyjścia wszystkich bramek są w stanie "1"* . Spójrzmy na rys. 8.2, pokazujący dwie bramki typu DOMINO, przy czym druga z nich otrzymuje wejściowy sygnał logiczny wprost z wyjścia pierwszej.




Rys. 8.2. Bramka typu DOMINO sterująca drugą podobną bramkę.

Gdy zaczyna się faza ewaluacji, na wyjściu X1 mamy zawsze stan "1". Zatem w tym momencie tranzystor drugiej bramki połączony z jej wejściem A2 jest włączony i pojemność $C_{I2}$ zaczyna się rozładowywać. Dopiero po upływie pewnego czasu, potrzebnego na rozładowanie pojemności $C_{I1}$, stan X1 zmieni się, być może, na "0" (zależy to, oczywiście, od stanów logicznych na wejściach A1, B1 i C1). Ale w tym momencie pojemność $C_{I2}$ może być już rozładowana do tego stopnia, że na wyjściu X2 będziemy mieli "0" lub stan nieokreślony pomiędzy "0", a "1", a nie jedynkę. Zatem układ da nam wynik błędny.

Można tego uniknąć w prosty sposób rozdzielając bramki typu DOMINO statycznymi inwerterami - patrz rys. 8.3.




Rys. 8.3. Bramka typu DOMINO sterująca drugą podobną bramkę poprzez inwerter.

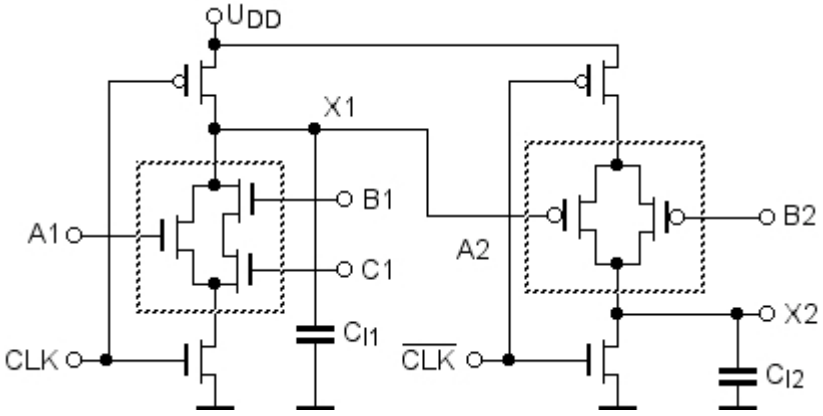
Inwerter powoduje, że na wejście A2 drugiej bramki na początku fazy ewaluacji podawane jest zawsze "0", a nie "1". Dopiero po ustaleniu się właściwego stanu wyjścia X1 stan wejścia A2 zmieni się, być może, na "1". Nie ma niebezpieczeństwa błędnych zadziałań, ponieważ w kolejnych stopniach układu bramki "czekają", aż w poprzednich stopniach ustalą się docelowe, prawidłowe stany.

Stąd zresztą pochodzi nazwa "domino". Każdy układ kombinacyjny zbudowany z bramek dynamicznych rozdzielonych inwerterami działa w taki sposób, że w czasie trwania fazy ewaluacji właściwe poziomy logiczne ustalają się stopniowo w kolejnych, coraz dalszych od wejściach bramkach. Twórcom tego sposobu realizacji układów kombinacyjnych skojarzyło się to z dziecięcą zabawą polegającą na pionowym ustawieniu jeden za drugim klocków domina, a następnie pchnięciu pierwszego klocka, który przewracając się popycha i przewraca drugi, ten przewraca następny i w końcu cały wąż klocków się kładzie.

Dodatkowe inwertery oczywicie wprowadzają pewne opóźnienia, ale nawet z nimi układ typu DOMINO może działać szybciej od układu z bramek statycznych. Pojawia się jednak pewne dodatkowe, często kłopotliwe ograniczenie - nie ma inwertera typu DOMINO, a bramki (z uwzględnieniem statycznych inwerterów) realizują tylko funkcje nie zawierające nigdzie negacji. Mówiąc ściśle, negacje (w postaci dodatkowych statycznych

inwerterów) są możliwe ne wejściu i na wyjściu bloku kombinacyjnego zbudowanego z bramek DOMINO, ale nie mogą występować wewnątrz tego bloku. Komplikuje to projekt logiczny.

Istnieją inne wersje układów dynamicznych, w których stosowanie inwerterów statycznych rozdzielających poszczególne stopnie nie jest potrzebne. Przykładem są układy dynamiczne, w których bloki logiczne budowane są na przemian z tranzystorów nMOS i pMOS. Bloki takie wymagają taktowania dwoma sygnałami zegarowymi. Sygnał dla bloków z tranzystorami pMOS jest negacją sygnału dla bloków z tranzystorami nMOS. Taki układ pokazuje rys. 8.4.




Rys. 8.4. Bramki dynamiczne nMOS-pMOS.

W układzie zbudowanym wg zasady pokazanej na rys. 8.4 fazy wstępnego ładowania i ewaluacji także następują równocześnie we wszystkich bramkach. Działanie bramek z tranzystorami pMOS jest dualne w stosunku do bramek z tranzystorami nMOS. W szczególności, wstępne ładowanie w tym przypadku oznacza rozładowanie pojemności $C_{l2}$ do zera, a w fazie ewaluacji pojemność ta może naładować się do napięcia $U_{DD}$ (czyli stanu "1"), jeśli umożliwią to stany logiczne na wejściach drugiej bramki. Ponieważ blok logiczny drugiej bramki zawiera tranzystory pMOS, które są włączane stanem "0", występowanie stanu "1" na wejściu A2 bezpośrednio po rozpoczęciu fazy ewaluacji nie niesie żadnego ryzyka błędnego zadziałania bramki. Wadą układu zbudowanego według zasady pokazanej na rys. 8.4 jest konieczność stosowania tranzystorów pMOS, co daje układ działający wolniej ze względu na mniejsze prądy drenu tranzystorów pMOS spowodowane mniejszą ruchliwością dziur, niż elektronów. Ze względu na to, że korzystniej jest łączyć tranzystory MOS równolegle, a unikać długich łańcuchów połączeń szeregowych, stopnie z tranzystorami nMOS powinny raczej realizować funkcje NOR, a stopnie z tranzystorami pMOS - funkcje NAND.

Dla uniknięcia nieporozumień trzeba zwrócić uwagę, że pod względem realizowanej funkcji logicznej układy z rys. 8.2, 8.3 i 8.4 nie są równoważne. Projekt logiczny układu przeznaczonego do realizacji w technice bramek dynamicznych musi być dostosowany do rodzaju tych bramek.

Wadą bramek dynamicznych jest mniejsza od bramek statycznych odporność na zakłócenia. Impuls zakłócający, którego amplituda jest większa od napięcia progowego tranzystorów w bloku logicznym, a czas trwania dostatecznie długi, może spowodować fałszywe rozładowanie pojemności $C_l$, a tym samym błąd w działaniu bramki.

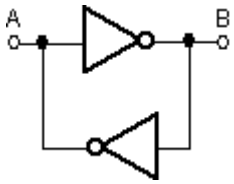
Inną słabą stroną bramek dynamicznych jest dodatkowy pobór mocy wynikający z obecności sygnału zegarowego. Bramki dynamiczne są taktowane nawet wtedy, gdy stany na ich wejściach nie zmieniają się. Oznacza to pobór mocy przez generator zegara.

Biorąc to wszystko pod uwagę można radzić stosowanie bramek dynamicznych tylko wtedy, gdy układ z bramkami statycznymi nie pozwala osiągnąć niezbędnej szybkości działania. Projektowanie układów z bramkami dynamicznymi jest znacznie trudniejsze, niż układów z bramkami statycznymi, i z reguły wymaga szczegółowych symulacji elektrycznych. Nie będziemy go tutaj szczegółowo omawiać. Obszerne informacje na ten temat zawiera literatura, zwłaszcza książka Uyemury (poz. [4] literatury).

## 8.2. Przerzutniki statyczne i dynamiczne

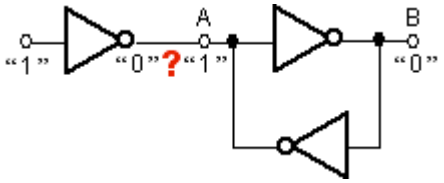
Zajmiemy się teraz elementarnymi układami pamięciowymi, jakimi są przerzutniki. Najprostszy przerzutnik dwustabilny jest układem statycznym. Takie przerzutniki używane są jako komórki pamięci w statycznych pamięciach RAM. Będzie o nich mowa w następnym wykładzie. Natomiast większość przerzutników używanych poza pamięciami należy do klasy układów dynamicznych.

Podstawowy **przerzutnik statyczny** powstaje przez połączenie dwóch inwerterów w taki sposób, że sygnał z wyjścia pierwszego inwertera jest podawany na wejście drugiego, a sygnał z wyjścia drugiego inwertera jest podawany na wejście pierwszego (rys. 8.5). Taki układ, jak łatwo się przekonać, ma dwa samopodtrzymujące się stany stabilne: gdy w węźle A jest stan "1", to w węźle B "0", i odwrotnie. Taki układ można więc użyć jako elementarną komórkę pamiętającą jeden bit informacji.



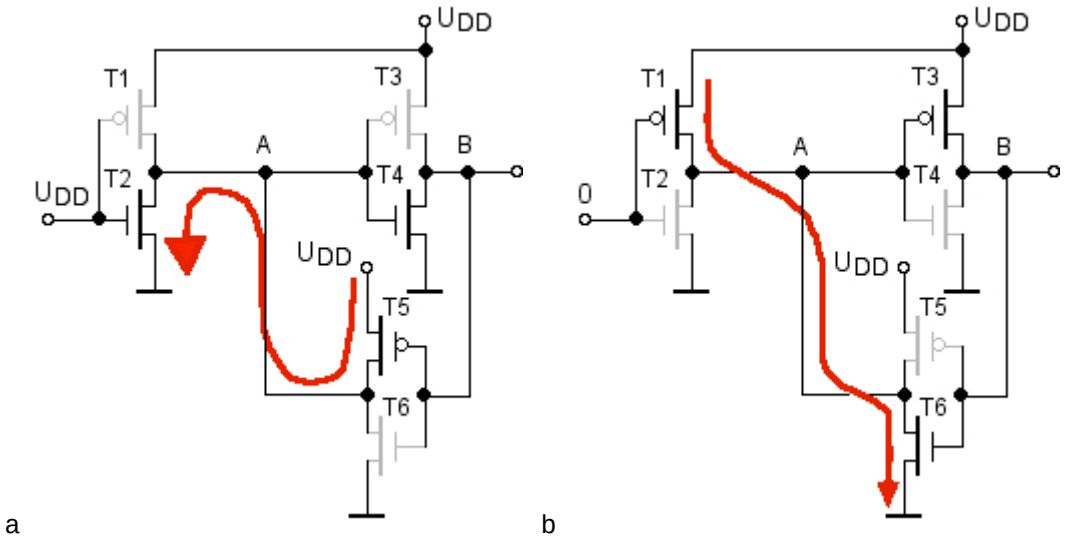

Rys. 8.5. Podstawowy przerzutnik statyczny

Aby przełączyć układ z jednego stabilnego stanu w drugi, trzeba wymusić na jednym z węzłów - A lub B - lub w obu równocześnie napięcie odpowiadające przeciwnemu stanowi. Wymaga to sterowania z odpowiednio zaprojektowanego bufora sterującego, którym w najprostszym przypadku może być inwerter. Załóżmy, że taki inwerter steruje węzłem A, w którym panuje stan "1", czyli napięcie równe napięciu zasilania układu $U_{DD}$. Załóżmy, że sterujący inwerter ma na wejściu stan "1", a więc na wyjściu powinien wymusić stan "0", czyli napięcie równe lub bliskie zeru. Powstaje wówczas sytuacja pokazana poniżej.




Rys. 8.6. Przełączanie przerzutnika przy pomocy inwertera

Analiza układu z rys. 8.6 jako układu logicznego nie pozwala stwierdzić, czy w węźle A ustali się "0", czy też "1". Trzeba zbadać wartość napięcia w węźle A. Schemat zastępczy układu (rys. 8.7.a), w którym pominięto tranzystory znajdujące się w stanie odcięcia, pokazuje że napięcie w weźle A określone jest przez dzielnik napięcia złożony z tranzystorów T5 i T2. Oba znajdują się w stanie przewodzenia.




Rys. 8.7. Przełączanie przerzutnika: (a) dzielnik napięcia T2 - T5, (b) dzielnik napięcia T6 - T1

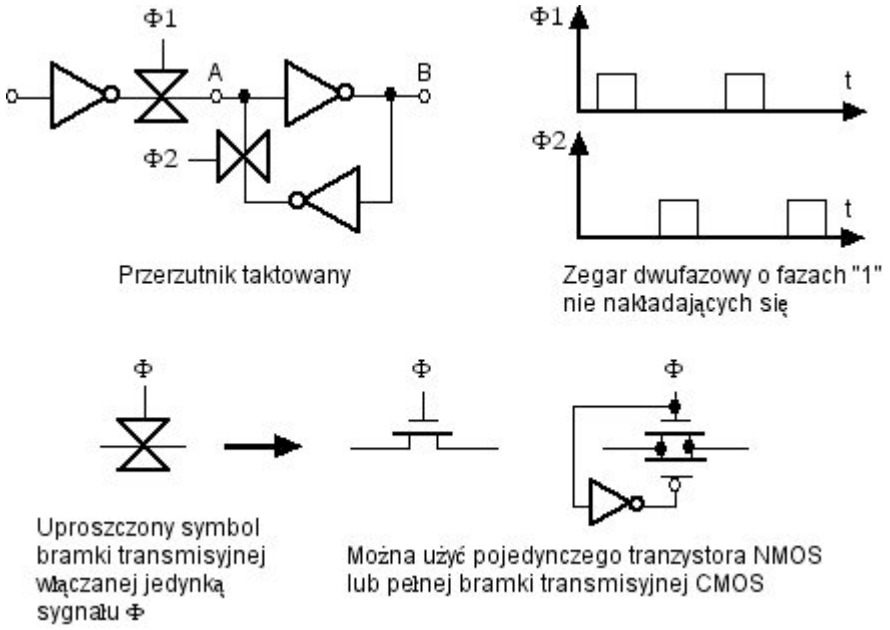
Jeśli tranzystor T2 będzie miał bardzo małą rezystancję w stosunku do tranzystora T5, to napięcie w węźle A spadnie do wartości bliskiej zeru. Wówczas ulegnie przełączeniu inwerter złożony z tranzystorów T3 i T4, w węźle B pojawi się "1", czyli napięcie równe $U_{DD}$, które przełączy inwerter złożony z tranzystorów T5 i T6. W ten sposób nastąpi zmiana stanu przerzutnika. Małą rezystancję tranzystora T2 osiąga się przez dobór odpowiednio dużej szerokości kanału.

Przy przełączaniu w drugą stronę (tj. zmianie stanu w weźle A z "0" na "1") dzielnik napięcia tworzą tranzystory T1 i T6 (rys. 8.7b). Tranzystor T1 musi mieć bardzo dużą szerokośc kanału, aby wymusić w węźle A napięcie bliskie $U_{DD}$.

Zatem zmiana stanu przerzutnika wymaga sterowania go z inwertera mającego tranzystory (T1 i T2) o szerokościach kanału znacznie większych, niż szerokości kanałow tranzystorów T5 i T6 w przerzutniku. Nie ma prostego wzoru pozwalającego obliczyć wymagane szerokości kanałów tranzystorów w inwerterze sterującym. Należy dobrać te szerokości korzystając z symulacji elektrycznej. Jako wartość startową można wybrać szerokości kanałów tranzystorów w inwerterze sterującym (T1, T2) 3 razy większe od szerokości kanałów tranzystorów w przerzutniku (T5, T6).

Omawiany przerzutnik znajduje zastosowanie głównie jako podstawowa komórka pamięci statycznych RAM, natomiast w innych zastosowaniach stosuje się różne wersje przerzutników dynamicznych.

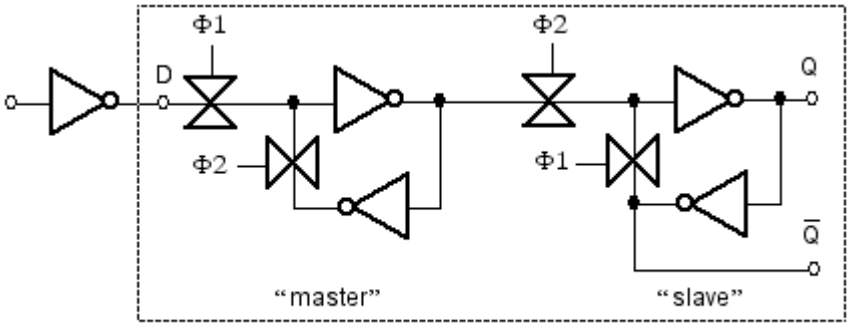
Dodając do podstawowego przerzutnika dwie bramki transmisyjne można zbudować przerzutnik, który będzie się przełączał przy sterowaniu z dowolnego inwertera (lub innej bramki), bez względu na wymiary tranzystorów. Idea polega na tym, że na czas przełączania przerywa się pętlę dodatniego sprzężenia zwrotnego występującą w przerzutniku. Układ zbudowany według tej idei pokazany jest na rys. 8.8.




Rys. 8.8. Przerzutnik dynamiczny wykorzystujący bramki transmisyjne

Gdy zegar Ø1 jest w stanie "1", na wejście A podawany jest sygnał z inwertera sterującego. W tym czasie Ø2 jest w stanie "0", a więc sygnał z dolnego inwertera w przerzutniku nie dociera do węzła A. Wobec tego poziomy logiczne w węzłach A i B ustalają się bez trudu. Po zmianie stanu Ø2 na "1", a Ø1 na "0" następuje utrwalenie stanów w węzłach A i B. *Układ działa prawidłowo, gdy fazy "1" obu zegarów nie nakładają się w czasie.* W tych odcinkach czasowych, w których oba sygnały zegarowe są w stanie "0", stany logiczne w węzłach A i B są podtrzymywane dzięki istniejącym w tych węzłach pojemnościom pasożytniczym. Układ ten należy więc do klasy układów dynamicznych. Zwany jest jednak także często pseudo-statycznym, bowiem gdy Ø2 jest w stanie "1", oba inwertery przerzutnika wzajemnie podtrzymują swoje stany tak samo, jak w omawianym poprzednio przerzutniku statycznym.

Z dwóch przerzutników taktowanych można łatwo zbudować **przerzutnik typu D**. Jest to najczęściej używany rodzaj przerzutnika w układach cyfrowych CMOS - w większości układów nie używa się w ogóle innych przerzutników. Przerzutnik typu D zapamiętuje sygnał wejściowy i opóźnia go o jeden takt zegara. Schemat takiego przerzutnika zwanego przerzutnikiem "master-slave" (dosłownie po polsku "pan"-"niewolnik") jest pokazany na rys. 8.9.

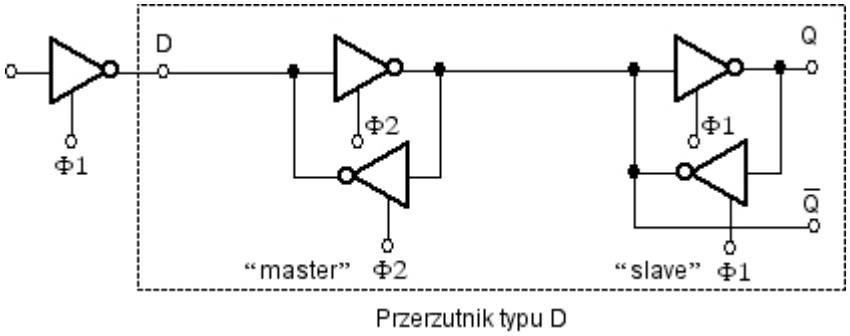



Rys. 8.9. Przerzutnik typu D

Gdy zegar Ø1 jest w stanie "1", następuje wpisanie stanu wejścia do pierwszego stopnia ("master"). Podczas jedynki zegara Ø2 następuje przepisanie do drugiego stopnia ("slave"). Tu również ważne jest, aby jedynki Ø1 i Ø2 nie nakładały się w czasie, bowiem równoczesne otwarcie wszystkich bramek transmisyjnych uniemożliwia prawidłowe działanie układu.

Łącząc w łańcuch przerzutniki typu D można zbudować szeregowy rejestr przesuwający. O rejestrach będzie mowa nieco dalej.

Dość często spotyka się przerzutniki takie, jak na rys. 8.8 i 8.9, w których jednak zastosowano zamiast bramek transmisyjnych inwertery trójstanowe. Takie przerzutniki działają dokładnie tak samo. Sygnały zegarowe Ø1 i Ø2 włączają lub wyłączają stan wysokiej impedancji. Schemat takiego przerzutnika pokazuje rys. 8.10.

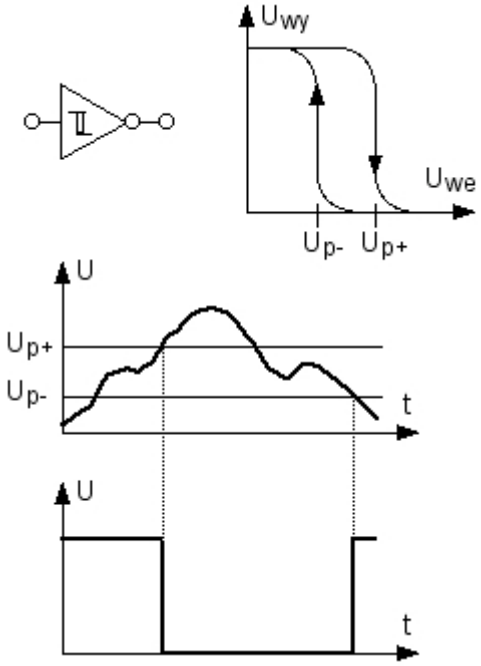



Rys. 8.10. Przerzutnik typu D z inwerterami trójstanowymi.

Przerzutnik z inwerterami trójstanowymi wymaga taktowania zegarem dwufazowym o fazach nie nakładających się, podobnie jak przerzutnik z bramkami transmisyjnymi.

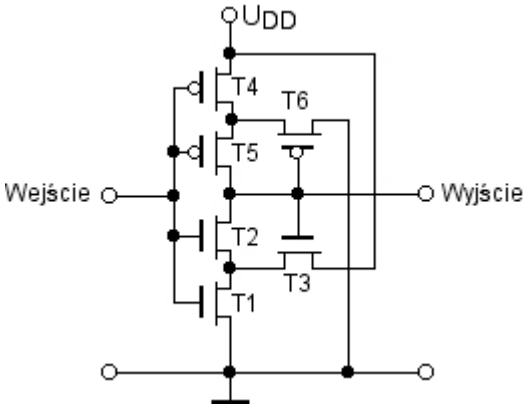
Jak zobaczymy dalej, zapewnienie właściwego taktowania zegarem dwufazowym o fazach nie nakładających się może być poważnym problemem technicznym w dużych układach.

Szczególnym rodzajem przerzutnika jest układ zwany **przerzutnikiem Schmitta**. Jest to **inwerter z histerezą** - charakterystyka przejściowa jest inna przy przełączaniu "0"->"1" niż przy przełączaniu "1"->"0". Taki układ bywa stosowany tam, gdzie trzeba przekształcić sygnał o niezbyt regularnym kształcie w ciąg prawidłowych zer i jedynek. Symbol, charakterystyki i typowe zastosowanie przerzutnika Schmitta ilustruje rys. 8.11.



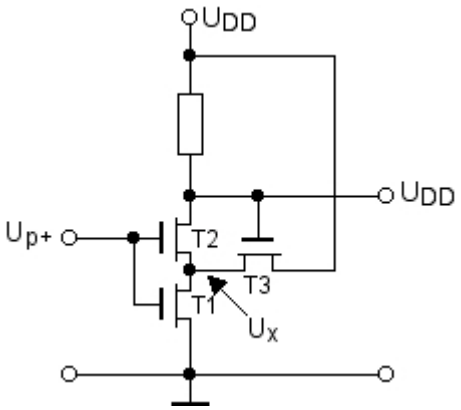

Rys. 8.11. Przerzutnik Schmitta: symbol, typowe charakterystyki oraz zastosowanie do kształtowania sygnału

Schemat przerzutnika Schmitta zrealizowanego jako bramka CMOS pokazuje rys. 8.12.




Rys. 8.12. Przerzutnik Schmitta

Zanalizujemy najpierw proces przełączania od stanu "0" do stanu "1" na wejściu, Pozwoli nam to także określić napięcie przełączania $U_{p+}$. Przyjmiemy dla uproszczenia, że napięcie $U_{p+}$ jest to graniczne napięcie, po przekroczeniu którego zaczyna się proces przełączania. Zatem gdy na wejściu panuje napięcie $U_{p+}$, to na wyjściu jest jeszcze stan "1", czyli napięcie $U_{DD}$ . W analizie pomocny będzie uproszczony schemat przerzutnika (rys. 8.13), w którym szeregowe połączenie przewodzących tranzystorów pMOS (T4 i T5) zastąpiono pewną nieistotną z punktu widzenia naszych rozważań rezystancją.



Rys. 8.13. Uproszczony schemat przerzutnika dla przełączania "0"->"1" na wejściu

Przełączenie, czyli spadek napięcia na wyjściu do wartości 0, zaczyna się w momencie gdy zaczyna przewodzić tranzystor T2. Zatem proces przełączania zaczyna się, gdy spełniony jest warunek

$$U_{p+} - U_x = U_{Tn} \tag{8.3}$$

W momencie początku przełączania tranzystor T1 jest już na pewno włączony, bowiem potencjał bramki obu tranzystorów, T1 i T2, jest taki sam, a potencjał źródła tranzystora T1 jest na pewno niższy, niż potencjał źródła tranzystora T2, a więc spełniony jest warunek $U_{GS1} > U_{Tn}$. Napięcie $U_x$ jest w momencie początku przełączania określone przez dzielnik napięcia, jaki tworzą dwa przewodzące tranzystory: T1 i T3 (T3 jest włączony, bo w chwili rozpoczęcia procesu przełączania na jego bramce jest jeszcze stan "1", czyli napięcie $U_{DD}$). Przyrównując prądy drenu tranzystorów T1 i T3 (prąd drenu tranzystora T2 na początku procesu przełączania jest do pominięcia) otrzymujemy

$$K_{n1}\left(U_{p+} - U_{Tn}\right)^2 = K_{n3}\left[\left(U_{DD} - U_x\right) - U_{Tn}\right]^2 \tag{8.4}$$

Łącząc to równanie z warunkiem (8.3) po przekształceniach otrzymujemy wyrażenie pozwalające oszacować napięcie $U_{p+}$:

$$U_{p+} = \frac{U_{DD} + \sqrt{\frac{K_{n1}}{K_{n3}}}\,U_{Tn}}{1 + \sqrt{\frac{K_{n1}}{K_{n3}}}} \tag{8.5}$$

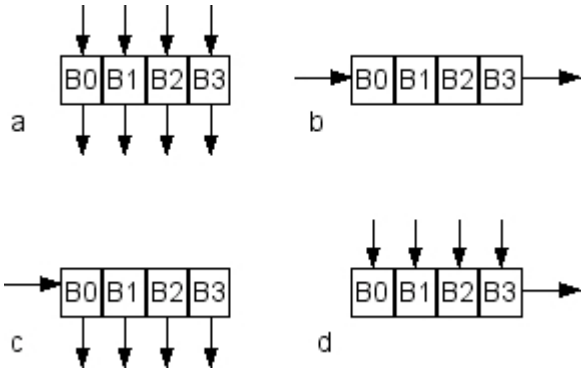
gdzie $K_n$ jest współczynnikiem przewodności tranzystora (wzór 4.13). Zależność (8.5) daje dość mało dokładne oszacowanie między innymi dlatego, że napięcia progowe tranzystorów T1, T2 i T3 nie są identyczne. Źródła tranzystorów T2 i T3 nie są połączone z podłożem układu (o potencjale równym 0), lecz z węzłem, w którym panuje wyższe napięcie $U_x$. To oznacza niezerowe napięcie polaryzacji podłoża $U_{BS}$ tych tranzystorów, co powoduje wzrost napięcia progowego zgodnie z zależnością 4.5. Zatem po wstępnym określeniu wymiarów tranzystorów należy wartość $U_{p+}$ uściślić przy pomocy symulacji i ewentualnie wprowadzić korektę wymiarów. Zauważmy, że tylko wymiary tranzystorów T1 i T3 mają wpływ na napięcie $U_{p+}$. Wymiary tranzystora T2 można przyjąć takie same, jak T1.

Rozumowanie analogiczne do przytoczonego wyżej pozwala określić napięcie przełączania $U_{p-}$. Wynosi ono

$$U_{p-} = \frac{\sqrt{\frac{K_{p4}}{K_{p6}}}\left(U_{DD} - \left|U_{Tp}\right|\right)}{1 + \sqrt{\frac{K_{p4}}{K_{p6}}}} \tag{8.6}$$

## 8.3. Rejestry

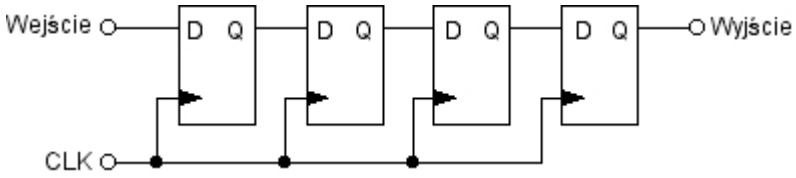
Rejestry służą do zapisu, przechowywania i odczytu grup bitów - typowo od kilku do kilkuset bitów. Istnieją rejestry równoległe i szeregowe. Do rejestrów równoległych zapisuje się i odczytuje wszystkie bity równocześnie. Rejestry szeregowe umożliwiają szeregowe wpisywanie kolejnych bitów, i podobnie odczyt. Istnieją też rejestry szeregowo-równoległe, umożliwiające np. szeregowe wpisywanie i równoległy odczyt, lub odwrotnie.




Rys. 8.14. Rodzaje rejestrów: (a) równoległy (wpis i odczyt równoległy), (b) szeregowy przesuwający (wpis i odczyt szeregowy), (c) szeregowo-równoległy (wpis szeregowy, odczyt równoległy) (d) równoległo-szeregowy (wpis równoległy, odczyt szeregowy)

Rejestry równoległe to zespoły przerzutników jednobitowych (statycznych lub dynamicznych). Bardziej interesujące są rejestry szeregowe. Zwane są one przesuwającymi, ponieważ przy szeregowym wpisywaniu i odczycie kolejne bity są "przesuwane" przy każdym takcie zegara w kierunku od wejścia do wyjścia.

Typowym rejestrem przesuwającym jest rejestr zbudowany z przerzutników typu D.

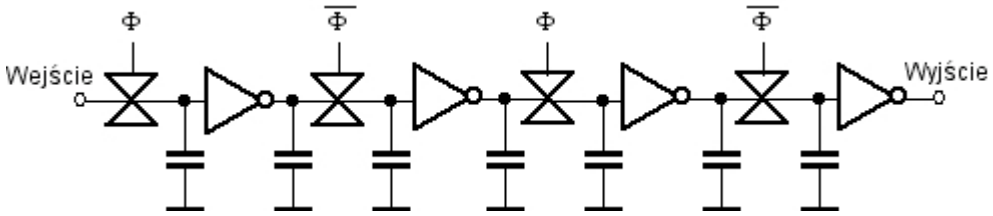



Rys. 8.15. Rejestr szeregowy przesuwający z przerzutników typu D

Każdy takt zegara CLK powoduje przesunięcie zapisanych w przerzutnikach bitów o jedną pozycję w prawo i wczytanie nowego bitu z wejścia.

Rejestr pokazany na rys. 8.15 jest taktowany pojedynczym sygnałem zegara. Tymczasem w przerzutnikach typu D takich, jak omawiane poprzednio (rys. 8.9 i 8.10) wymagany jest zegar dwufazowy o fazach "1" nie nakładających się. Możliwe są dwa rozwiązania: albo druga faza zegara jest generowana lokalnie w każdym przerzutniku (takie rozwiązanie jest zwykle stosowane w przerzutnikach wchodzących w skład bibliotek komórek standardowych), albo oba zegary są generowane centralnie i doprowadzane do każdego przerzutnika. Do tego zagadnienia będziemy jeszcze wracać.

Oprócz rejestrów złożonych z przerzutników pseudo-statycznych (takich, jak te z rys. 8.9 i 8.10) bywają też stosowane rejestry w pełni dynamiczne. Takie rejestry mają znacznie prostszą budowę. Pojedynczy stopień rejestru składa się z inwertera, bramki transmisyjnej oraz pojemności, która służy jako element pamięciowy:




Rys. 8.16. Dynamiczny rejestr szeregowy przesuwający

W rejestrze dynamicznym kolejne bramki transmisyjne (mogą to być pojedyncze tranzystory nMOS) są otwierane i zamykane na przemian sygnałem zegara i jego negacją. Otwarta bramka powoduje podanie sygnału z wyjścia poprzedniego inwertera na wejście następnego. Łatwo się przekonać, że w tej sytuacji jeden pełny takt zegara

powoduje przesunięcie o dwa stopnie.

Rejestr dynamiczny wymaga ciągłego taktowania, a częstotliwość zegara nie może być dowolnie mała, jest ograniczona od dołu zjawiskiem upływności rozładowującej kondensatory.

## 8.4. Zegary i taktowanie

**Zegar** jest niezbędny w każdym układzie sekwencyjnym, a także w układach kombinacyjnych z bramkami dynamicznymi. W praktyce zegar jest potrzebny w każdym układzie cyfrowym, ponieważ układy cyfrowe przeznaczone do realizacji jako scalone układy CMOS są układami synchronicznymi. Stosowane są dwa sposoby generacji sygnałów zegarowych:

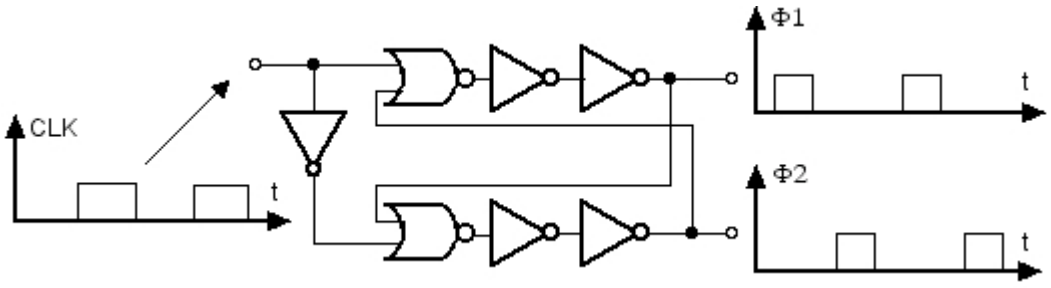
- zegar generowany w systemie poza układem,
- zegar generowany w układzie.

W pierwszym przypadku układ ma wejście dla zewnętrznego sygnału zegara. Sygnału tego zwykle nie dostarcza się bezpośrednio do bramek w układzie. Steruje on wewnętrznymi układami formującymi sygnały zegarowe i dostarczającymi je w wymagane miejsca.

Drugi przypadek zachodzi wtedy, gdy nie ma potrzeby synchronizowania działania układu z innymi układami, na przykład wtedy, gdy układ stanowi funkcjonalną całość. Przykłady takich układów: układ do zegarka elektronicznego, układ prostego kalkulatora itp. W praktyce nie ma potrzeby samodzielnego projektowania oscylatorów, które generują sygnał zegara. Producenci układów ASIC dostarczają gotowe projekty takich oscylatorów w bibliotekach komórek standardowych. Na ogół trzeba do takiego oscylatora dołączyć z zewnątrz rezonator kwarcowy, który ustala częstotliwość zegara.

W dużym i złożonym układzie może być potrzebne wiele różnych sygnałów zegarowych różniących się fazą, a nawet częstotliwością. Są one wytwarzane z głównego sygnału zegarowego, dostarcznego z zewnątrz lub generowanego w układzie. Sygnał o częstotliwości $n$-krotnie mniejszej od głównego, gdzie $n$ jest całkowitą wielokrotnością 2, uzyskuje się bez trudu przez użycie dzielnika częstotliwości. Sygnał o częstotliwości $n$-krotnie wyższej zsynchronizowany z sygnałem głównym jest trudniej uzyskać. Potrzeba takiego sygnału występuje z reguły tylko wtedy, gdy sygnał główny jest dostarczany z zewnątrz. Do synchronizacji wykorzystuje się układy **pętli fazowej** ("phase-locked loop", w skrócie **PLL**) dla synchronizacji wewnętrznego generatora sygnału o częstotliwości $nf$ z zewnętrznym sygnałem o częstotliwości $f$. Układy PLL są skomplikowane, a ich projektowanie trudne. Temat ten wykracza poza ramy naszego wykładu.

Częstym, a znanym nam już przypadkiem jest przypadek zegara dwufazowego o fazach "1" nie nakładających się. Takie dwa sygnały zegarowe można łatwo wygenerować przy użyciu układu pokazanego na rys. 8.17.

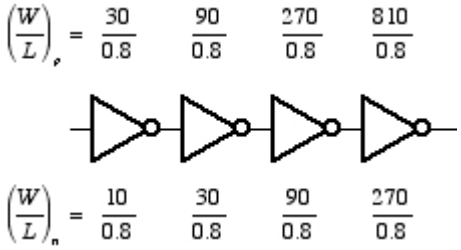



Rys.8.17. Układ generujący dwa sygnały zegarowe o fazach "1" nie nakładających się

Odstęp czasu między stanami "1" obu sygnałów jest regulowany opóźnieniami wprowadzanymi przez inwertery wprowadzone w pętlę sprzężenia zwrotnego. W razie potrzeby można wprowadzić zamiast dwóch - cztery inwertery lub większą (zawsze parzystą) ich liczbę.

Rozprowadzanie sygnału zegara w dużych układach jest poważnym problemem technicznym. W idealnym przypadku dla każdego taktowanego danym sygnałem zegarowym elementu układu (jak przerzutnik, bramka transmisyjna, rejestr itp.) stany jedynki i zera zegara powinny zaczynać się i kończyć w dokładnie tych samych momentach. Tylko wtedy układ jest rzeczywiście synchroniczny. W praktyce spełnienie tego warunku jest trudne. W dużych układach sygnały zegara są doprowadzone równocześnie do bardzo wielu wejść zegarowych w przerzutnikach, bramkach dynamicznych itp., a to oznacza, że układ, który jest źródłem sygnału zegara, jest obciążony bardzo dużą pojemnością. Dlatego stosuje się bufory, których zadaniem jest zapewnić jak najkrótsze czasy narastania i opadania impulsów sygnału zegarowego mimo dużej pojemności obciążającej. Takim buforem jest zwykle inwerter o stosunku $W/L$ tranzystorów wynoszącym kilkaset, a nawet kilka tysięcy. Tranzystory takie mają wiele równolegle połączonych kanałów. Przykład konstrukcji takich tranzystorów był pokazany na rys. 6.1 - są to tranzystory znajdujące się przy górnej krawędzi bloku pokazanego na tym rysunku (powtórzonym w dodatku 1). Inwerter z takimi tranzystorami nie może być sterowany bezpośrednio z inwertera lub bramki o "zwykłych" wymiarach tranzystorów, tj, o stosunku $W/L$ rzędu 1 ... 10, ponieważ bardzo duże tranzystory same stwarzają duże obciążenie pojemnościowe, zbut duże dla zwykłego inwertera lub bramki. Bufory zwykle buduje się jako

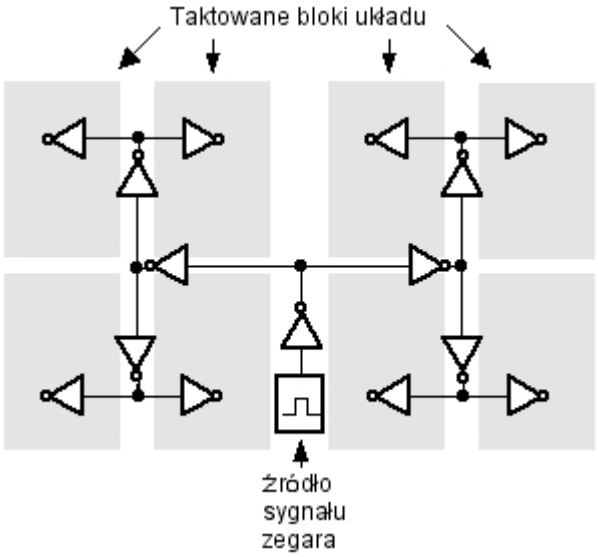
łańcuchy inwerterów o stopniowo rosnących stosunkach $W/L$ tranzystorów. Liczba buforów w takim łańcuchu może być nieparzysta lub parzysta, w zależności o tego, czy chcemy uzyskać na wyjściu sygnał zanegowany, czy też nie. Przykład tak zbudowanego bufora pokazuje rys. 8.18. Prostą praktyczną regułą jest trzykrotne powiększanie szerokości tranzystorów w kolejnych inwerterach bufora.




Rys. 8.18. Czterostopniowy bufor zegara, pokazane przykładowe wymiary tranzystorów w kolejnych stopniach

Na zasadzie pokazanej na rys. 8.18 buduje się nie tylko bufory dla sygnału zegara, ale także bufory dla wszystkich innych sygnałów, które trzeba dostarczyć do obciążenia o dużej pojemności wejściowej (a także małej rezystancji).

W dużych układach liczących miliony bramek całkowita pojemność wszystkich wejść zegarowych jest tak duża, że pojedynczy bufor musiałby mieć absurdalnie duży stosunek $W/L$ tranzystorów. W takim przypadku układ dzieli się na fragmenty mające mniej więcej jednakowe pojemności wejść zegara, i do każdego z tych fragmentów doprowadza się sygnał zegara przez system buforów tworzący strukturę drzewiastą. Pokazuje to rys. 8.19.




Rys. 8.19. Rozprowadzenie i buforowanie sygnału zegara przy zastosowaniu drzewa typu H

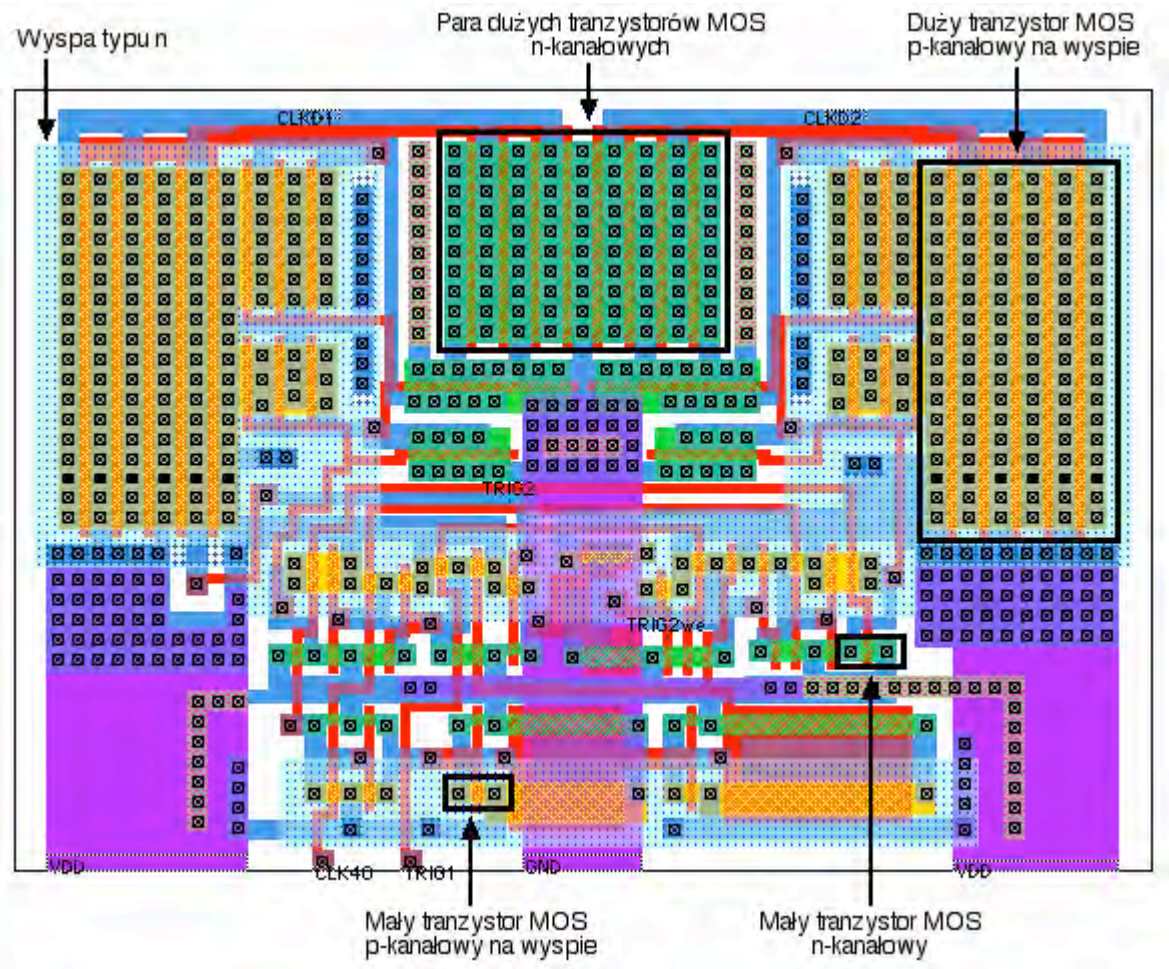
W strukturze pokazanej na rys. 8.19 (zwanej drzewem typu H) sygnał zegara na drodze do każdego bloku przechodzi przez tę samą liczbę identycznych i jednakowo obciążonych buforów. Jeżeli na dodatek uda się w topografii układu zachować geometryczną identyczność wszystkich symetrycznych segmentów drzewa, to także długość drogi dla zegara jest w sensie odległości geometrycznej taka sama dla każdego bloku. Jeżeli wszystkie taktowane bloki układu mają jednakowe pojemności obciążające bufory zegara, to uzyskuje się dokładnie jednakowy czas propagacji sygnałów zegarowych z ich źródła do wszystkich bloków układu (mogą jednak istnieć różnice wynikające z lokalnych zaburzeń procesu produkcyjnego - patrz wykład 2).

Szczególne trudności mogą powstać, jeśli w dużym układzie trzeba rozprowadzić dwa lub więcej sygnałow zegarowych o ściśle określonych zależnościach czasowych. Dotyczy to na przykład omawianego wcześniej zegara dwufazowego o fazach "1" nie nakładających się w czasie. Nawet niewielkie różnice w czasach propagacji tych dwóch sygnałów mogą spowodować, że w jakimś bloku układu jedynki należące do dwóch różnych sygnałów zegara nałożą się w czasie, co spowoduje błędne działanie bloku. Dlatego unika się rozprowadzania w dużych układach więcej niż jednego sygnału zegara. Znacznie bezpieczniej jest dodatkowe sygnały zegarowe generować lokalnie w blokach, w których są potrzebne.

Przy projektowaniu buforów zegara i sposobu rozprowadzenia zegara należy pamiętać, że celem *nie jest* uzyskanie jak najmniejszego opóźnienia między źródłem tego sygnału, a taktowanymi blokami. Celem jest uzyskanie jak najkrótszych czasów narastanie i opadania impulsów zegarowych oraz jednakowego czasu propagacji tych impulsów do wszystkich taktowanych bloków. Osiągnięcie tego celu zapewnia prawidłową,

synchroniczną pracę układu jako całości.

## 8.4. Dodatek 1: Duże tranzystory nMOS i pMOS w generatorze sygnałów zegarowych




Rys. 6.1.

# ĆWICZENIE 1 DO WYKŁADU 8

**Cel ćwiczenia**

W tym ćwiczeniu zapoznasz się z programem "Dsch". Jest to symulator logiczny. Przy jego pomocy można sprawdzić poprawność projektu układu logicznego. Tego rodzaju programy są powszechnie używane w projektowaniu cyfrowych układów scalonych. Celem ćwiczenia jest wykonanie symulacji układu generującego dwa sygnały zegara dwufazowego o fazach "1" nie nakładających się (rys. 8.17).

**1. Instalacja programu**

Przed rozpoczęciem ćwiczenia musisz zainstalować program "Dsch2" na swoim komputerze. Instalacja jest bardzo prosta - skopiuj cały katalog "Dsch2" do katalogu "Program Files". W katalogu "Dsch2" znajdziesz plik "Dsch2.exe" - to jest właśnie program. Możesz dla wygody zrobic skrót i umieścić go na pulpicie lub dodać program do menu startowego. W katalogu "Dsch2" znajdziesz także kilkaset plików. Są to m.in. pliki technologiczne definiujące różne technologie dla programu "Dsch2" (rozszerzenie ".tec"), pliki z przykładowymi projektami (rozszerzenie ".sch") i inne, oraz katalog "Html".

**2. Obsługa programu**

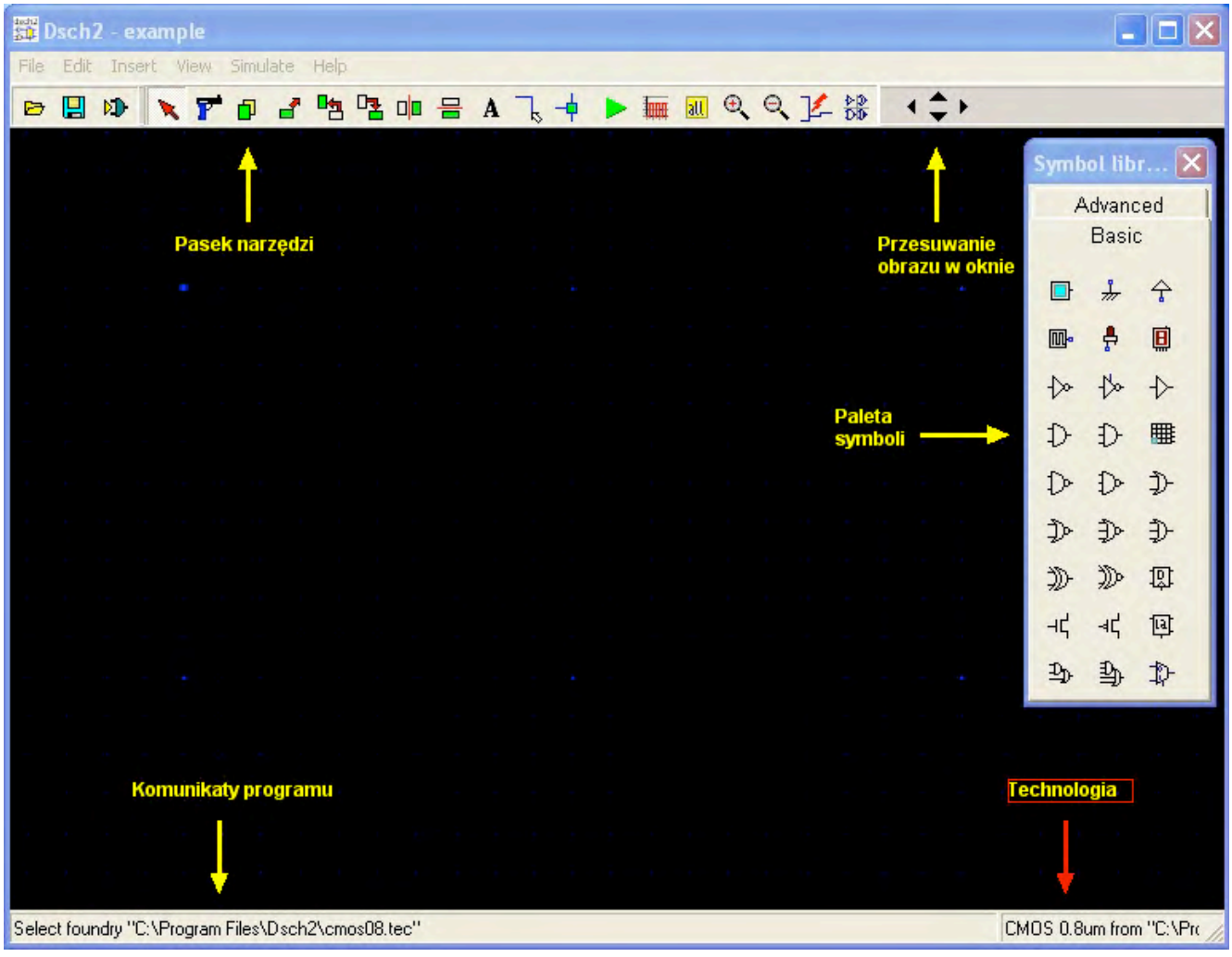
W kolejnych ćwiczeniach otrzymasz szczegółowe wskazówki, jak wykonywać poszczególne czynności. Nie będą jednak omówione wszystkie możliwości i opcje programu.

W katalogu "Dsch2\Html" znajdziesz zwięzłą instrukcję obsługi programu (w języku angielskim) w postaci stron w języku HTML do przeglądania dowolną przeglądarką. Aby je przejrzeć, otwórz plik "index.htm".

Wykonywanie ćwiczeń ułatwi Ci zapoznanie się z prezentacją wideo wprowadzającą do nich. Znajdziesz ją w sekcji "Wykłady VIDEO".

**3. Przebieg ćwiczenia**

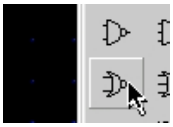
Uruchom program "Dsch2". Zobaczysz okno jak niżej. Na ilustracji kolorem żółtym opisano najważniejsze widoczne obiekty. Zauważysz podobieństwo do okna programu "Microwind2". Wiele narzędzi działa podobnie. Najważniejsza różnica polega na tym, że w programie "Dsch2" będziesz tworzyć i symulować schematy logiczne, a nie topografię układu. Po prawej stronie paleta symboli bramek logicznych i innych obiektów, z których będziesz zestawiać schematy.



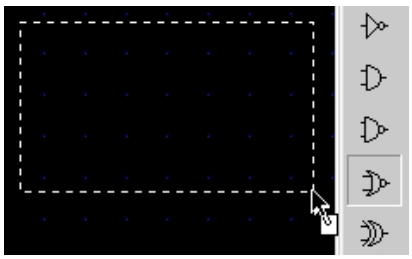
Wykorzystujemy nadal technologię CMOS 0,8 mikrometra. Musimy więc wczytać odpowiedni plik technologiczny. Z menu "File" wybierz "Select Foundry", i wczytaj plik "cmos08.tec". Po wczytaniu możesz zacząć budować schemat logiczny układu.

Będziesz teraz wybierać bramki i zestawiać z nich układ.

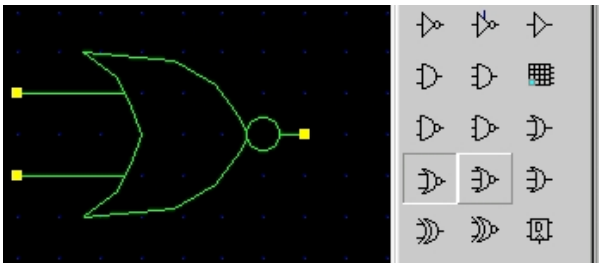
Wybierz z palety dwuwejściową bramkę NOR.



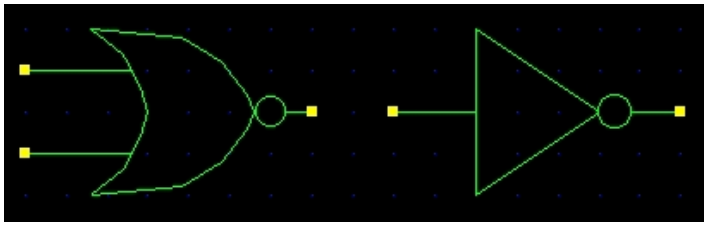
Naciśnij przycisk myszy i ciągnij z palety na pole tworzenia schematu.



Gdy puścisz przycisk, bramka zostanie umieszczona we wskazanym miejscu.

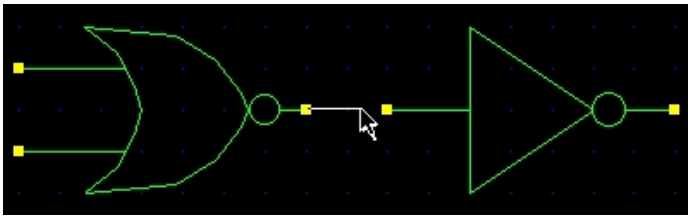


W ten sam sposób wybieraj i umieszczaj dalsze bramki.

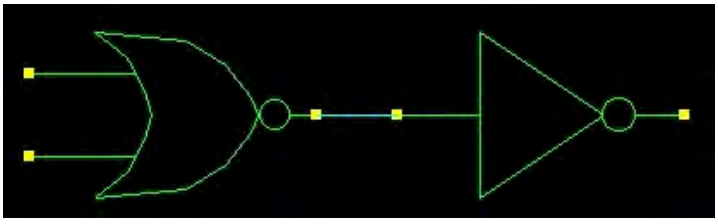


Wybierz z paska narzędzi narzędzie do rysowania połączeń.

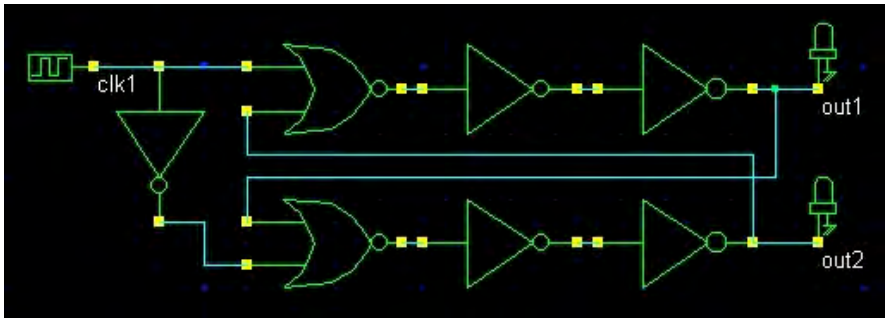
Rysuj połączenie ciągnąc je od wyjścia jednej bramki do wejścia następnej.



Tak wygląda wykonane połączenie.

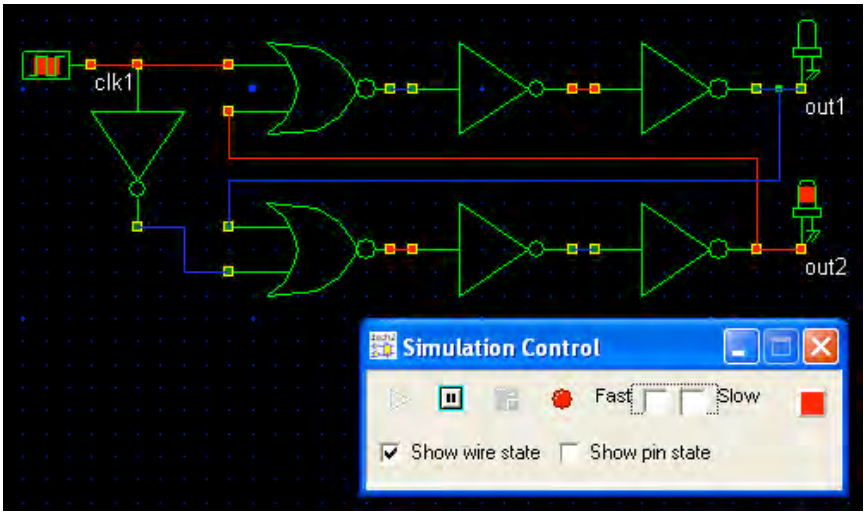


Tak lub podobnie powinien wyglądać gotowy schemat. Na wejściu dodaj generator zegara, na wyjściach dwa symbole diod, będą one pokazywać stan wyjść. Teraz możesz już uruchomić symulator logiczny.



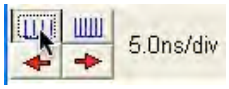

Wybierz z paska narzędzi ikonę powodującą uruchomienie symulacji.

Na ekranie zobaczysz, jak "działa" Twój układ. Możesz przyspieszać lub zwalniać symulację, a także ją zatrzymywać wybierając opcje w oknie "Simulation Control". Sprawdź, jak działają opcje "Show wire state" i "Show pin state".



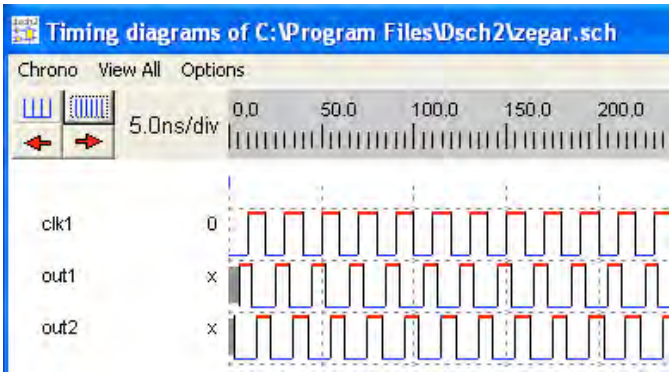

Wybierz z paska narzędzi ikonę otwierającą okno, które pokazuje sygnały w funkcji czasu.


Aby dobrze widzieć przebiegi, ustaw skalę czasu w lewym górnym rogu okna.

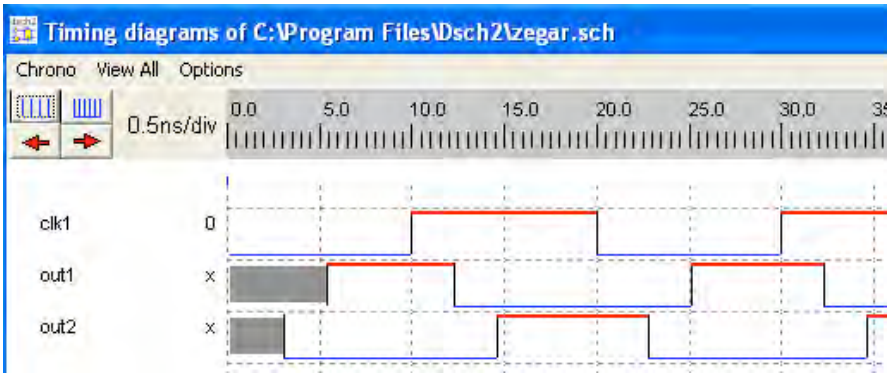



Widać przebiegi na wejściu i obu wyjściach. Symulator uwzględnia czasy propagacji bramek (ich przybliżone wartości zapisane są w pliku technologicznym).

Aby jeszcze dokładniej zbadać zależności czasowe, zmień skalę czasu na 0,5 ns na działkę.




Po zmianie skali dobrze widać, że otrzymujemy dwa przebiegi, których stany "1" nie nakładają się w czasie.




Ćwiczenie zakończone! Teraz już potrafisz zbudować schemat logiczny i sprawdzić jego działanie. Zapisz schemat na dysk, może Ci się jeszcze przydać.

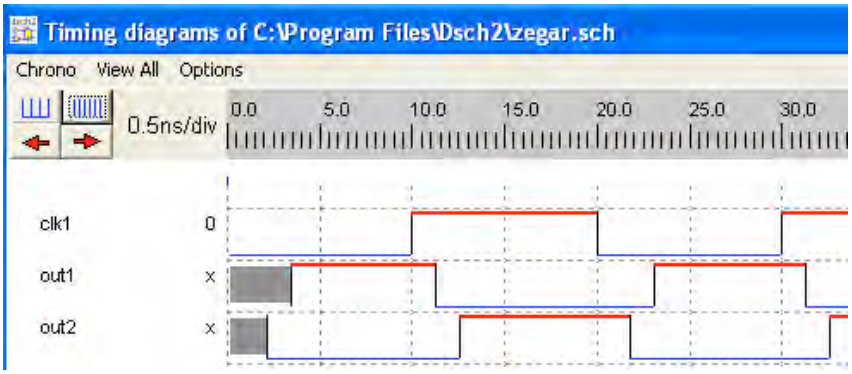
## ĆWICZENIE 2 DO WYKŁADU 8

**Cel ćwiczenia**

Samodzielny trening

**Przebieg ćwiczenia (do samodzielnego wykonania)**

Sprawdź, jaki jest wpływ liczby inwerterów w symulowanym układzie na odstęp czasowy między stanami "1". W tym celu usuń z układu inwertery, zastępując je bezpośrednimi połączeniami. Następnie wykonaj symulację. Jeśli wszystko jest dobrze, wynik powinien wyglądać mniej więcej tak:




Porównując te wyniki z poprzednimi zauważysz, że odstęp między stanami "1" sygnałów wyjściowych zmalał.

## Bibliografia

[1] M. J. Patyra, *"Projektowanie układów MOS w technice VLSI*", Wydawnictwa Naukowo-Technicze, 1993
[2] E. Sicard, *"Microwind & Dsch User's Manual"*, National Institute of Applied Sciences INSA, Toulouse, 2003
(Jest to podręcznik do oprogramowania wykorzystywanego w tym wykładzie, załączony na płycie w postaci pliku PDF)
[3] J. M. Rabaey, *"Digital Integrated Circuits, A Design Perspective*, Prentice Hall, Inc. 1996
[4] J. P. Uyemura, *"CMOS Logic Circuit Design"*, Kluwer Academic Publishers, 2001

# Wykład 9: Pamięci i inne układy o podobnej strukturze

## Wstęp

W wykładzie 9 mowa jest o pamięciach. Trudno znaleźć układ cyfrowy, który nie potrzebowałby pamięci. Potrzebne są zarówno pamięci o stałej zawartości, służące tylko do odczytu, jak i pamięci, których zawartość może być zmieniania. W wykładzie jest mowa o jednych i o drugich. Omawiane są zasady działania komórek pamięci, czyli układów przechowujących pojedyncze bity informacji, oraz organizacja układów pamięci. Omawiane są także inne układy o strukturze regularnej, które wprawdzie nie są zaliczane do pamięci, ale wykazują pewne do nich podobieństwo.

Projektant specjalizowanego układu scalonego na ogół nie projektuje pamięci samodzielnie. Jeżeli w projektowanym układzie potrzebna jest pamięć, jest ona zwykle generowana automatycznie. Jest to możliwe dlatego, że pamięci mają regularną i powtarzalną budowę, co pozwala łatwo zautomatyzować ich projektowanie. To stwierdzenie nie dotyczy pamięci dynamicznych (znanych jako pamięci DRAM) wytwarzanych jako osobne układy scalone. Pamięci takie wymagają specjalnej technologii, a ich projektowanie jest bardzo trudne. Zajmuje się tym tylko kilku wyspecjalizowanych producentów na świecie.

Wykład 9 nie omawia szczegółów projektowania komórek pamięci ani też układów zapisu, odczytu i adresowania. Niemniej ogólne wiadomości o pamięciach potrzebne są każdemu, kto zajmuje się systemami cyfrowymi.

## 9.1. Rodzaje pamięci i ich ogólna budowa

Wykład zaczniemy od przeglądu rodzajów układów pamięci.

Tradycyjnie pamięci półprzewodnikowe są dzielone na:

- pamięci o stałej zawartości, przeznaczone tylko do odczytu (**Read Only Memory** - **ROM**)
- pamięci o swobodnym dostępie, w których w każdej chwili można dokonać zarówno zapisu, jak i odczytu (**Random Access Memory** - **RAM**).

Rozwój technologii wytwarzania pamięci doprowadził jednak do sytuacji, w której granice między pamięciami ROM i RAM zaczęły ulegać zatarciu. Wiele rodzajów pamięci klasyfikowanych jako pamięci ROM umożliwia bowiem wielokrotne programowanie, a więc z punktu widzenia pełnionej funkcji zbliża się do pamięci RAM. Główna różnica polega więc na czym innym. Pamięci klasyfikowane jako RAM są to pamięci *ulotne* - ich zawartość ginie po wyłączeniu zasilania. Pamięci klasyfikowane jako ROM są to pamięci *nieulotne*. Wyłączenie zasilania nie powoduje utraty zawartości pamięci.

Pamięci klasyfikowane jako RAM mają w porównaniu z pozostałymi rodzajami pamięci krótki czas dostępu, tj. czas, jaki upływa od zainicjowania procesu zapisu lub odczytu do zakończenia tego procesu. Pamięci klasyfikowane jako ROM, ale reprogramowalne, tj. takie, których zawartość można wielokrotnie zmieniać, potrzebują do zmiany zawartości znacznie dłuższego czasu niż pamięci RAM. Proces zapisu w tych pamięciach przebiega wolno. Różnica szybkości zapisu w reprogramowalnych pamięciach ROM w porównaniu do pamięci RAM polega na zasadniczo odmiennym charakterze mechanizmów fizycznych, które określają zawartość pamięci. Zmiana zawartości komórki pamięci RAM oznacza *zmianę stanu w układzie elektronicznym* tworzącym tę komórkę, np. zmianę stanu przerzutnika statycznego. W przypadku reprogramowalnych pamięci ROM zmiana zawartości komórki oznacza *zmianę struktury fizycznej* tej komórki lub jednego z jej elementów, np. wytworzenie nowego połączenia elektrycznego lub usunięcie istniejącego. W dalszej części tego wykładu będzie pokazane, na czym taka zmiana struktury fizycznej polega w różnych odmianach pamięci ROM.

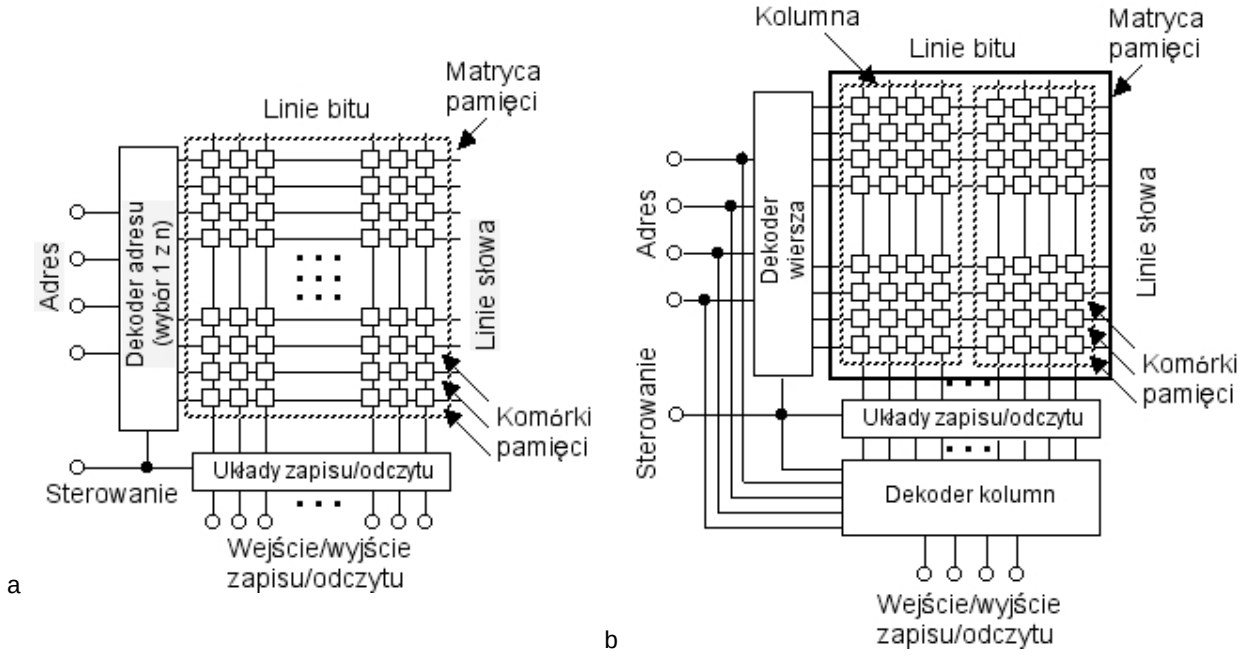
Pamięci RAM dzielą się pod względem sposobu przechowywania informacji na dwa rodzaje: **pamięci statyczne** (ang. static RAM, SRAM) i **dynamiczne** (ang. dynamic RAM, DRAM). W pamięciach statycznych elementarną komórką pamięci przechowującą jeden bit jest przerzutnik statyczny. W pamięciach dynamicznych elementem pamięciowym jest kondensator, a stan jedynki lub zera jest pamiętany jako ładunek w tym kondensatorze lub jego brak.

Pamięci ROM dzielą się na znacznie więcej grup. Są wśród nich pamięci, których zawartość jest określana już w procesie produkcji i nie może być potem zmieniona, są pamięci, które można zaprogramować, ale tylko jeden raz, są pamięci, których zawartość można kasować w całości i programować od początku, i wreszcie są pamięci elektrycznie reprogramowalne, które są najbardziej zbliżone pod względem możliwych zastosowań do pamięci RAM.

Wszyskie te rodzaje pamięci będą omówione w dalszej części wykładu.

Oprócz tego istnieją także inne, podobne do pamięci ROM układy, które umożliwiają realizację funkcji kombinacyjnych w postaci matryc o regularnej budowie. Choć formalnie nie zaliczają się one do pamięci ROM, omówimy je w tym wykładzie ze względu na istotne podobieństwa do pamięci ROM.

Budowę typowych układów pamięci pokazuje rys. 9.1.




Rys. 9.1. Budowa układów pamięci: zasada ogólna (a) i organizacja pamięci o dużej pojemności (b)

Ogólną zasadę budowy układów pamięci ilustruje rys. 9.1a. Trzy główne bloki układu pamięci to *matryca pamięci*, *dekoder adresów* i blok *układów zapisu/odczytu*. Matryca pamięci składa się z pewnej liczby *linii słowa* (poziomych na rys. 9.1) i krzyżujących się z nimi *linii bitu* (pionowych na rys. 9.1). Na każdym skrzyżowaniu linii słowa z linią bitu znajduje się komórka pamięci pamiętająca jeden bit. Komórki pamięci zbudowane są różnie w różnych rodzajach pamięci. Mogą to być pojedyncze tranzystory lub układy bardziej złożone, zawierające kilka elementów.

Dekoder adresów to układ, który dla każdej wartości adresu (tj. kombinacji bitów na wejściu adresowym) uaktywnia do zapisu lub odczytu dokładnie jedną linię słowa (na ogół przez podanie na nią stanu "1"). Wszystkie komórki pamięci położone wzdłuż tej linii będą przedmiotem operacji zapisu lub odczytu. Dla słowa adresowego o długości $n$ bitów matryca pamięci posiada $2^n$ linii słowa.

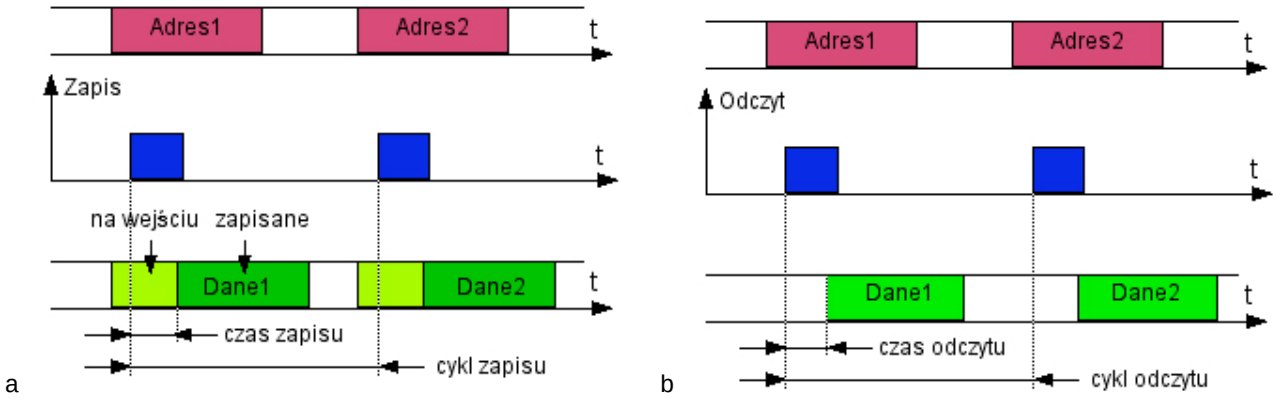
Układy zapisu/odczytu są różne w różnych rodzajach pamięci, ale wszędzie pełnią tę samą rolę. Przy zapisie przetwarzają otrzymane z zewnątrz słowo binarne na wewnętrzne sygnały potrzebne do dokonania zapisu w komórkach pamięci. Przy odczycie odczytane z komórek pamięci stany logiczne (jak zobaczymy, nie zawsze reprezentowane przez napięcia równe 0 i $U_{DD}$) są przetwarzane na słowo binarne podawane na wyjście. Liczba linii bitu decyduje o organizacji pamięci, np. przy 8 liniach zapisywane i odczytywane są słowa ośmiobitowe.

Sterowanie (pokazane symbolicznie jako wejście jednobitowe) określa funkcję wykonywaną w danej chwili przez pamięć, np. zapis lub odczyt.

Najprostsza organizacja pamięci pokazana na rys. 9.1a staje się niepraktyczna, a nawet niemożliwa do realizacji w przypadku pamięci o dużej pojemności. Linie bitu nie mogą być dowolnie długie i nie można do nich dołączać dowolnie dużej liczby komórek pamięci. Jak zobaczymy dalej, długość linii bitu i liczba dołączonych do niej komórek pamięci ma bezpośredni wpływ między innymi na czas zapisu i odczytu. W pamięciach o dużej pojemności matryca jest dzielona na mniejsze części. Przykład takiej organizacji pamięci pokazuje rys.9.1b. Zapisywane i odczytywane są słowa czterobitowe. Matryca jest podzielona na czterobitowe kolumny, i na podstawie adresu wybierana jest nie tylko linia słowa, ale i kolumna. Jeśli kolumn jest $n$, to przy danej pojemności pamięci linie bitu mogą być $n$ razy krótsze.

Dla bardzo dużych pamięci stosuje się podział na mniejsze bloki będące kompletnymi matrycami, które z kolei dzielą się na kolumny. Komplikuje to dekodowanie adresów, ale pozwala zachować rozsądną długość linii bitu.

Zależności czasowe charakterystyczne dla pamięci pokazuje rys. 9.2.




Rys. 9.2. Zależności czasowe przy zapisie (a) i odczycie (b)

Zarówno przy zapisie, jak i przy odczycie trzeba najpierw podać adres na wejście adresowe, a w przypadku zapisu także dane, które mają być zapisane. Potem na wejście sterujące podawany jest sygnał startu zapisu lub odczytu. Przy zapisie upływa pewien czas, zanim dane zostaną zapisane (na rys. 9.2 czas zapisu). Przy odczycie upływa pewien czas, zanim odczytane dane pojawią się na wyjściu (czas odczytu). Cykl zapisu to najkrótszy odcinek czasu, po którym można ponownie dokonać zapisu. Cykl odczytu to najkrótszy odcinek czasu, po którym można ponownie dokonać odczytu. Cechą charakterystyczną pamięci klasyfikowanych jako RAM są krótkie i zbliżone do siebie czasy zapisu i odczytu. W przypadku reprogramowalnych pamięci ROM regułą jest, że czas odczytu może być krótki (porównywalny do czasu odczytu z pamięci RAM), natomiast czas zapisu jest zacznie dłuższy, nawet o kilka rzędów wielkości.

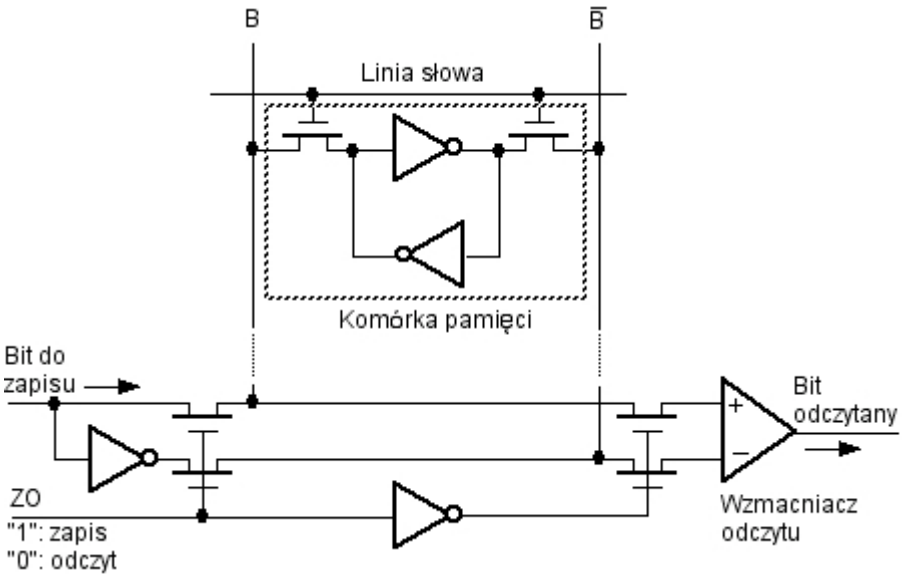
Proste schematy zależności czasowych pokazane na rys. 9.2 służą jedynie do ilustracji zasadniczych pojęć - czasu zapisu, czasu odczytu, cyklu zapisu, cyklu odczytu. Rzeczywiste sekwencje sygnałów przy zapisie i odczycie mogą być inne. Przykładowo, istnieją konstrukcje pamięci statycznych RAM (a także pamięci ROM), w których proces zapisu i odczytu inicjowany jest przez zmianę wartości adresu. Nie jest potrzebny odrębny sygnał startu. Z kolei w pamięciach dynamicznych stosowany jest bardziej złożony zbiór sygnałów sterujących, zależny od wewnętrznej organizacji, na przykład sposób adresowania, w którym na wejście adresowe podawana jest najpierw część adresu określająca linię słowa, a potem część określająca wybór kolumny. W takim przypadku potrzebne są sygnały synchronizujące, które inicjują najpierw wybór wiersza (linii słowa), potem kolumny, a na koniec uruchamiają proces odczytu.

Jeszcze bardziej złożone sposoby sterowania działaniem pamięci występują przy najszybszych dynamicznych pamięciach RAM, zwanych synchronicznymi. Pamięci takie stosuje się na płytach głównych współczesnych komputerów. Działanie tych pamięci jest synchronizowane zegarem. Przy zapisie dane muszą być dostępne na wejściu danych w tym samym cyklu zegara, w którym następuje zapis, natomiast przy odczycie dane są podawane na wyjście z opóźnieniem kilku cykli zegara w stosunku do cyklu, w którym podany został sygnał odczytu. Ta liczba cykli (w oznaczeniu pamięci określana symbolem CL, ang. cycle latency) wynosi od 2 - 3 przy częstotliwościach zegara rzędu 100 MHz aż do 7 - 9 przy częstotliwościach rzędu 1 GHz.Organizacja wewnętrzna pamięci umożliwia odczyt potokowy: podczas oczekiwania na transfer pierwszej porcji odczytanych danych można przesyłać do pamięci kolejne sygnały odczytu, dzięki temu na odczyt następnych porcji danych nie będzie już trzeba czekać równie długo. Ponadto wszystkie współcześnie produkowane pamięci dynamiczne umożliwiają dwa transfery danych na jeden takt zegara, reagując zarówno na rosnące, jak i opadające zbocze sygnału zegarowego. Takie pamięci oznaczane są symbolem DDR (ang. Double Data Rate). Oto przykład zależności czasowych dla bardzo szybkiej pamięci dynamicznej oznaczanej jako DDR3 1600 CL7 (częstotliwość zegara 800 MHz): przy kolejnym odczycie 8 słów pierwsze pojawia się po 8,75 ns, ostatnie już po 13,125 ns. Mimo tych udoskonaleń czas odczytu z pamięci jest jednym czynników poważnie ograniczających wydajność obliczeniową komputerów. Dlatego powszechnie stosowane są podręczne pamięci (ang. cache) o mniejszych pojemnościach, ale znacznie szybsze, które służą do przechowywania danych często i wielokrotnie wykorzystywanych przez procesor. Te pamięci są realizowane jako statyczne pamięci RAM.

Szczegółowe omiawianie wielu istniejących sposobów organizacji pamięci i sterowania ich pracą wykracza poza zakres naszego wykładu. Producenci pamięci podają wszystkie potrzebne informacje w katalogowej dokumentacji technicznej.

## 9.2. Pamięci RAM statyczne i dynamiczne

Komórką pamięci statycznej RAM jest podstawowy przerzutnik statyczny omawiany w wykładzie 8 (rys. 8.5), uzupełniony o dwa tranzystory nMOS pełniące rolę bramek transmisyjnych, które służą do wyboru danej komórki do zapisu i odczytu, w zależności od stanu linii słowa. Schemat takiej sześciotranzystorowej komórki pamięci wraz z (pokazanymi w pewnym uproszczeniu) układami zapisu/odczytu pokazuje rys. 9.3.




Rys. 9.3. Komórka pamięci statycznej wraz z układami zapisu/odczytu

Jak pokazuje rys. 9.3, typowa komórka pamięci statycznej jest połączona poprzez tranzystory nMOS z dwoma liniami bitu, na których pojawia się zapisywany lub odczytywany bit oraz jego negacja. Gdy linia słowa jest w stanie "1", tranzystory nMOS są włączone i komórka komunikuje się z obydwoma liniami bitu. Gdy linia słowa jest w stanie "0", tranzystory nMOS są wyłączone. Przerzutnik statyczny w komórce pamięci trwa w stanie, jaki został ostatnio zapisany. Zapis wymaga wyboru komórki przez podanie "1" na linię słowa oraz podania "1" na wejście sterujące ZO. Otwierają się wówczas bramki transmisyjne zapisu (na schemacie po lewej stronie), i bit do zapisu oraz jego negacja są podawane na linie bitu i jego negacji. Równoczesne podawanie bitu i jego negacji przyspiesza proces zmiany stanu przerzutnika, ponieważ nowy stan podawany jest równocześnie na wejścia obu inwerterów przerzutnika. Przy odczycie (ZO w stanie "0") otwarte są bramki transmisyjne odczytu (z prawej strony na schemacie), a napięcia z linii bitu i jego negacji są podawane na wejście różnicowego wzmacniacza odczytu. Jest to układ analogowy (będzie omawiany bardziej szczegółowo w jednym z dalszych wykładów). Wzmacnia on różnicę napięć między linią bitu i jego negacji. Zastosowanie tego wzmacniacza przyspiesza proces odczytu. Przy odczycie napięcia na liniach bitów zmieniają się stopniowo (potrzebny jest czas na ładowanie lub rozładowywanie pojemności tych linii). Ale już bardzo mała różnica napięć na wejściach wystarcza, aby na wyjściu pojawiło się pełne napięcie $U_{DD}$ lub 0 (w zależności od stanu logicznego odczytywanego z komórki).

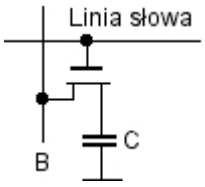
Wzmacniacz różnicowy dokonuje zarazem regeneracji poziomu jedynki (jeśli odczytywana jest jedynka). Jak widać na rys. 9.2, odczyt z komórki pamięci odbywa się poprzez bramkę transmisyjną w postaci pojedynczego tranzystora nMOS. Jak już wiemy, powoduje to degradację poziomu jedynki (wykład 7). Dla oszczędności powierzchni pełnych bramek transmisyjnych CMOS w komórkach pamięci nie używa się.

Pamięci statyczne są najszybciej działającymi pamięciami typu RAM. Są one jednak dość kosztowne i mają ograniczoną pojemność, ponieważ komórki tych pamięci, liczące aż 6 tranzystorów, zajmują dużo miejsca. Zaletą pamięci statycznych jest to, że nie wymagają one do produkcji specjalnej technologii wytwarzania. Mogą być zatem bez trudności stosowane jako na przykład części składowe mikroprocesorów (pamięci podręczne typu "cache"). Jednym z głównych czynników ograniczających szybkość działania pamięci statycznych jest konieczność ładowania lub rozładowywania dużej pasożytniczej pojemności linii bitu. Dlatego, jak już wcześniej mówiliśmy, liczba komórek dołączonych do linii bitu nie może być zbyt duża. Dla pamięci o większej pojemności stosuje się omawiane wcześniej bardziej złożone schematy organizacji pamięci.

W praktyce projektowania układów ASIC bardzo rzadko zdarza się potrzeba zaprojektowania pamięci statycznej od początku, tj. zaprojektowania komórek, całej matrycy dekodera adresów oraz układów wejścia/wyjścia. Profesjonalne systemy projektowania dysponują możliwością automatycznej generacji projektu pamięci o zadanej pojemności i organizacji. Potrzebne do tego dane, schematy i topografie komórek, układów wejścia/wyjścia itp. dostarczają producenci układów.

Pamięci dynamiczne RAM są najbardziej znanym rodzajem pamięci półprzewodnikowych, ponieważ to właśnie te

pamięci są powszechnie stosowane jako pamięci operacyjne komputerów. Zarówno schemat, jak i zasada działania komórki takiej pamięci są bardzo proste.



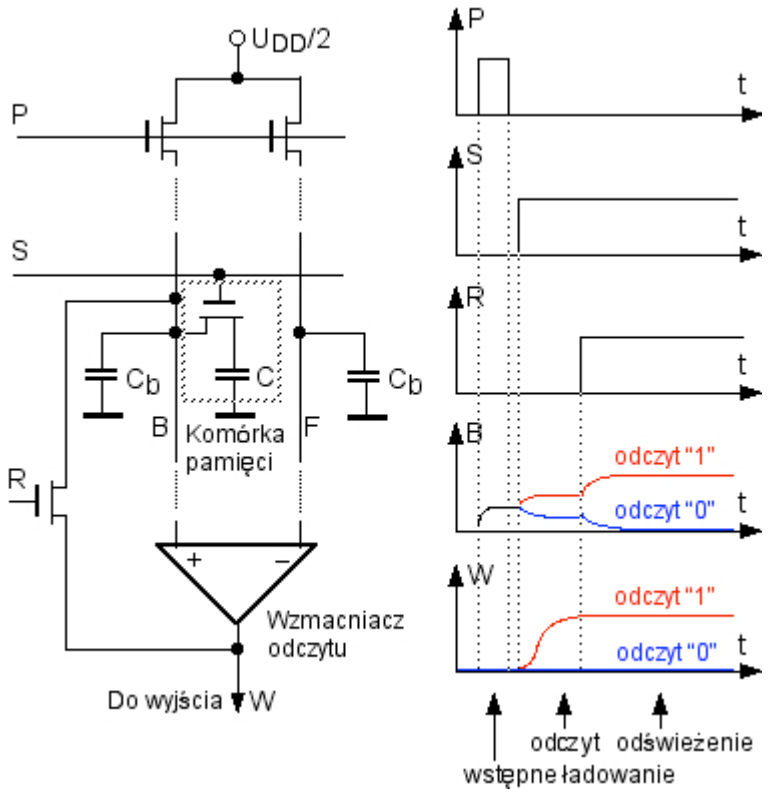

Rys. 9.4. Komórka pamięci dynamicznej RAM

Proces zapisu polega na ładowaniu lub rozładowywaniu pojemności C. Gdy linia słowa jest w stanie "1", a na linię bitu podane jest napięcie $U_{DD}$, pojemność C ładuje się poprzez tranzystor, który jest włączony. Gdy zaś na linii bitu panuje napięcie równe zeru, pojemność C ulega rozładowaniu. W ten sposób zapisywana jest jedynka lub zero. Przy odczycie napięcie z kondensatora podawane jest przez włączony tranzystor na linię bitu.

Ta prosta zasada jest jednak bardzo trudna do praktycznego wykorzystania. Problem stwarza zjawisko podziału ładunku przy odczycie (patrz wykład 6). Pojemność pasożytnicza linii bitu, do której dołączone jest bardzo wiele komórek pamięci, jest wielokrotnie większa od pojemności kondensatora C w pojedynczej komórce. W rezultacie przy odczycie jedynki odczytywane z komórki pamięci napięcie jest wielokrotnie mniejsze od napięcia $U_{DD}$. Typowa wartość odczytywanego napięcia jedynki to kilkadziesiąt mV. Po odczycie napięcie na kondensatorze pozostaje na tym poziomie, toteż następny odczyt jedynki nie jest już możliwy. Wynika z tego, że:

- odczyt wymaga zastosowania wzmacniacza regenerującego właściwy poziom logiczny
- po odczycie konieczne jest odświeżenie zawartości komórki przez ponowny zapis do niej odczytanego stanu logicznego.

Dlatego układy zapisu/odczytu dla pamięci dynamicznych są dużo bardziej skomplikowane, niż w przypadku pamięci statycznych. Rys. 9.5 ilustruje przebieg procesu odczytu i odświeżenia zawartości komórki. Wykorzystuje się tu zasadę wstępnego ładowania, znaną nam już z bramek dynamicznych typu DOMINO, ale tutaj wykorzystaną w nieco inny sposób.




Rys. 9.5. Zasada odczytu z komórki pamięci dynamicznej. Przebieg pokazany linią czerwoną dla przypadku, gdy w komórce zapisana była jedynka, linią niebieską - gdy zapisane było zero.

Każdej linii bitu B towarzyszy fałszywa linia bitu F, do której nie są podłączone żadne komórki pamięci, ale która ma pojemność pasożytniczą o wartości takiej samej (lub zbliżonej), jak linia bitu B. Przed odczytem obie linie są

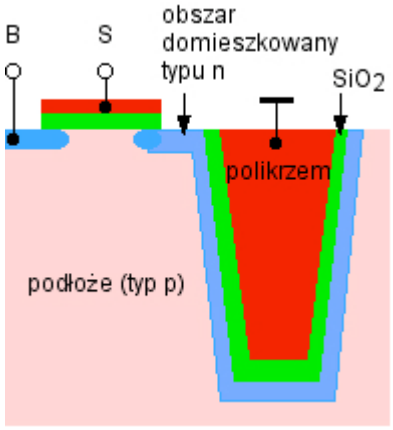
wstępnie ładowane do napięcia równego w przybliżeniu połowie $U_{DD}$. Odbywa się to z chwilą podania jedynki na linię P. Po zakończeniu wstępnego ładowania na liniach B i F panuje jednakowe napięcie podtrzymywane dzięki pojemnościom pasożytniczym tych linii $C_b$. Następnie podawana jest jedynka na linię słowa S. Dzięki temu pojemność C komórki pamięci zostaje dołączona przez tranzystor do linii bitu B. Jeżeli w komórce zapamiętana była jedynka (czyli napięcie na kondensatorze było równe lub bliskie $U_{DD}$), ładunek z kondensatora C doładowuje pojemność pasożytniczą linii bitu $C_b$ i powoduje podskok napięcia o wartości kilkudziesięciu mV. Napięcie na linii F nie zmienia się, powstaje więc różnica napięć na wejściu wzmacniacza odczytu. Jest to wzmacniacz różnicowy działający jako komparator napięcia. Jest tak skonstruowany, że na jego wyjściu pojawia się pełne napięcie jedynki, czyli $U_{DD}$., gdy $U_B > U_F$, a napięcie równe zeru, jeśli $U_B <= U_F$.

Stan z wyjścia wzmacniacza podawany jest na wyjście z pamięci, a także - po podaniu jedynki na bramkę tranzystora oznaczoną R - na linię bitu B. W rezultacie następuje ponowne naładowanie kondensatora C komórki pamięci do napięcia reprezentującego jedynkę. Pokazuje to przebieg napięcia na linii B (rys. 9.5, czerwona linia). Jeśli natomiast w chwili odczytu w komórce zapisane jest zero, czyli kondensator C jest rozładowany, to ładunek odpływa do niego z linii bitu, co powoduje spadek napięcia na tej linii o kilkadziesiąt mV. Różnica napięć na wejściu wzmacniacza różnicowego powoduje pojawienie się na jego wyjściu napięcia równego zeru. To napięcie, czyli "0", jest podawane na wyjście, a także - poprzez tranzystor R - na linię bitu. Napięcie na tej linii spada do zera, co powoduje także rozładowanie do zera pojemności C w komórce pamięci. W ten sposób w komórce pamięci pozostaje zapisane zero. Przebieg napięcia na linii bitu pokazany jest na rys. 9.5 niebieską linią.

Każdy akt odczytu z komórki pamięci powoduje więc zarazem odświeżenie jej zawartości. Komórki wymagają okresowego odświeżania zawartości nawet wtedy, gdy w pamięci nic się nie dzieje. Ładunek w kondensatorze C ulega bowiem powolnemu zanikowi w wyniku istnienia prądów upływu (prąd podprogowy tranzystora komórki, prąd wsteczny złącza p-n drenu). W celu odświeżania wystarczy okresowo odczytywać wszystkie komórki.

Jak widać, działanie pamięci dynamicznej przy odczycie jest dość skomplikowane. Z tego powodu pamięci dynamiczne działają wolniej niż statyczne. Ich wielką zaletą jest bardzo mała powierzchnia zajmowana przez pojedynczą komórkę, zawierającą tylko jeden tranzystor i jeden kondensator. Ta mała powierzchnia umożliwia produkcję pamięci o wielkiej pojemności i niewygórowanej cenie.

W roli kondensatora C w komórce pamięci nie wystarczają pojemności pasożytnicze. Chodzi bowiem o to, by ładunek zgromadzony w tej pojemności był możliwie duży, tak aby zmiana napięcia na linii bitu przy odczycie była dostatecznie duża i umożliwiała bezbłędne zadziałanie wzmacniacza odczytu mimo zjawiska podziału ładunku (przypomnij sobie wzór (8.1) i wzór (8.2)). Tę możliwie dużą pojemność C należy jednak zmieścić na możliwie jak najmniejszej powierzchni. Do tego celu opracowano specjalne technologie wytwarzania układów pamięci, w których kondensatory komórek pamięci są wykonywane w postaci głębokich wnęk wytrawionych w krzemie, których zbocza są pokryte bardzo cienkim dielektrykiem ($SiO_2$), a wnętrze wypełnione polikrzemem (stosowane są też inne sposoby wytwarzania tych kondensatorów). Przekrój przez komórkę z takim kondensatorem pokazuje rys. 9.6. Okładki kondensatora tworzą: polikrzem wypełniający wnękę oraz obszar domieszkowany typu n w podłożu.



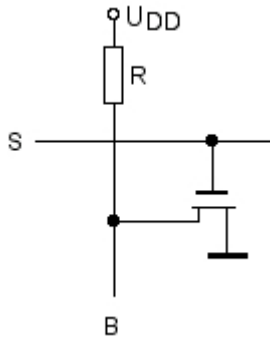

Rys. 9.6. Przekrój przez komórkę pamięci dynamicznej

Nawet stosując taki specjalny kondensator jako element pamięciowy nie unikamy omawianego wyżej problemu podziału ładunku. Sztuka projektowania i produkcji pamięci dynamicznych polega nie tylko na tym, by uzyskać jak największą wartość pojemności komórki C, ale także na tym, by pojemność pasożytnicza linii bitu $C_b$ była możliwie mała. Ogranicza to liczbę komórek pamięci, jaką można dołączyć do jednej linii bitu. Z tego powodu w pamięciach dynamicznych regułą jest stosowanie podziału matrycy pamięci na wiele stosunkowo niedużych bloków. O takiej organizacji pamięci była już mowa wcześniej.

Operacje głębokiego trawienia wnęk, utleniania ich zboczy, wypełniania ich polikrzemem nie są typowe dla zwykłej technologii wytwarzania układów CMOS. Pamięci są produkowane na specjalnie do tego przeznaczonych liniach produkcyjnych. Dlatego w zwykłych układach CMOS pamięci dynamiczne nie są spotykane. Toteż warto wiedzieć, jak działają, ale ich projektowanie wykracza poza zakres tego wykładu.

## 9.3. Pamięci ROM i inne pamięci nieulotne

W technologii CMOS podstawową komórką pamięci ROM jest pojedynczy tranzystor nMOS (rys. 9.7).



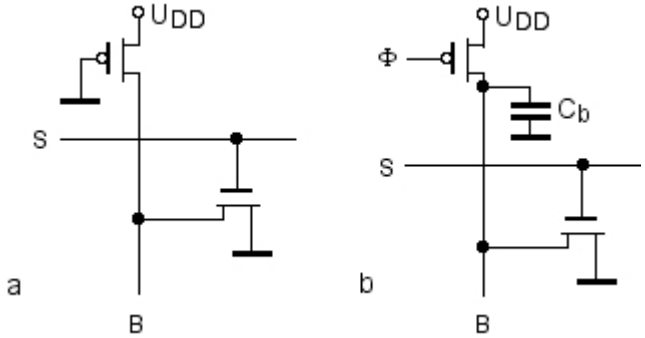

Rys. 9.7. Komórka pamięci ROM - zasadnicza idea

Działanie takiej komórki jest niezwykle proste. Załóżmy, że na linii słowa S panuje stan "1". Jeśli tranzystor jest i przewodzi, to zwiera linię bitu do zera napięcia zasilania, co oczywiście oznacza stan "0". Jeśli tranzystor nie przewodzi lub go w ogóle nie ma, na linii bitu jest niezerowe napięcie reprezentujące stan "1". Widoczny na rys. 9.7. rezystor R jest symbolem elementu, który doprowadza napięcie ze źródła zasilania do linii bitu. W rzeczywistości jednak rezystorów się nie stosuje, ponieważ użycie zwykłej liniowej rezystancji nie dałoby możliwości uzyskania pamięci o dobrych parametrach.

Różne rodzaje pamięci ROM różnią się głównie następującymi cechami:

- w jaki sposób doprowadzane jest niezerowe napięcie do linii bitu,
- w jaki sposób osiąga się stan przewodzenia lub nieprzewodzenia tranzystora stanowiącego komórkę pamięci.

Doprowadzenie niezerowego napięcia do linii bitu można najprościej uzyskać stosując tranzystor pMOS spolaryzowany w taki sposób, by pełnił rolę nieliniowej rezystancji. Połączenie bramki z zerem napięcia zasilania powoduje, że tranzystor pMOS jest zawsze włączony. Ilustruje to rys. 9.8a. Gdy tranzystor nMOS nie przewodzi, na linii bitu panuje napięcie $U_{DD}$. Gdy tranzystor nMOS przewodzi, tworzy się dzielnik napięcia - przewodzą oba tranzystory. Dobierając odpowiednio ich wymiary można osiągnąć na linii bitu napięcie dostatecznie niskie, aby odpowiednio zaprojektowany wzmacniacz odczytu (który nie jest pokazany na rys. 7.8) zinterpretował je jako stan "0". Ten sposób doprowadzenia napięcia do linii bitu jest bardzo prosty, ale ma istotną wadę: pamięć statycznie pobiera prąd, gdy oba tranzystory przewodzą.




Rys. 9.8. Doprowadzenie napięcia do linii bitu: (a) statycznie, (b) dynamicznie

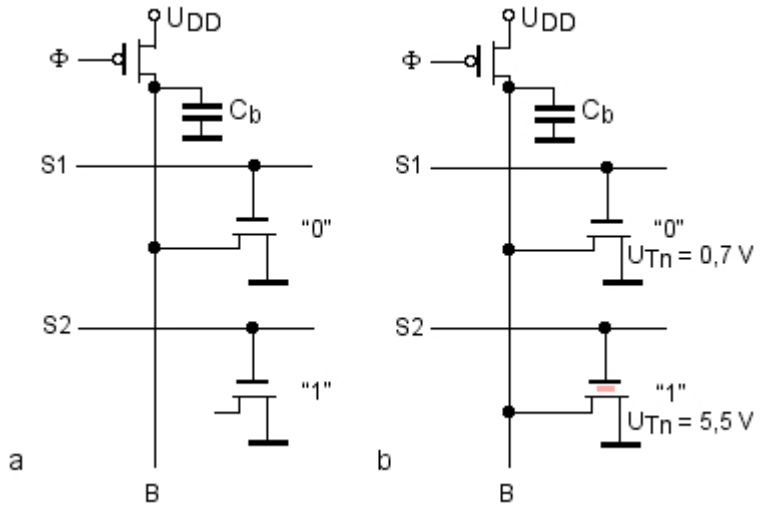
Lepszym rozwiązaniem jest zastosowanie wstępnego ładowania - rys. 9.8b. Tu tranzystor pMOS wstępnie ładuje linię bitu (jej pojemność $C_b$) do napięcia $U_{DD}$, gdy zegar Ø jest w stanie "0". Po przejściu do stanu "1" tranzystor pMOS zostaje wyłączony. Teraz możliwy jest odczyt. Na linię słowa podawana jest "1". Jeśli tranzystor nMOS przewodzi, pojemność $C_b$ rozładowuje się i odczytywane jest "0". W przeciwnym razie napięcie $U_{DD}$ na pojemności $C_b$ pozostaje i odczytywana jest "1". Tę zasadę działania już poznaliśmy, gdy omawiane były bramki typu DOMINO.

Omawianie pamięci ROM zaczniemy od pamięci, których zawartość określona jest już podczas produkcji i nie może być potem zmieniona. Takie pamięci bywają też określane jako pamięci programowane maską, bowiem

rzeczywiście ich zawartość jest określona przez jedną z masek fotolitograficznych, na przykład maskę określającą połączenia w warstwie metalu 1.

Pamięci programowane maską mają zastosowanie tam, gdzie wiadomo na pewno, że nigdy nie będzie potrzeby zmiany zapisanej w pamięci informacji. Przykładem może być program sterujący działaniem mikroprocesora w prostych urządzeniach powszechnego użytku: kalkulatorach, pralkach, sprzęcie audio i TV. Pamięci ROM programowane maską są najtańszym rodzajem pamięci ROM przy produkcji masowej.

Jest kilka sposobów programowania w procesie produkcyjnym. Dwa z nich pokazuje rys. 9.9.



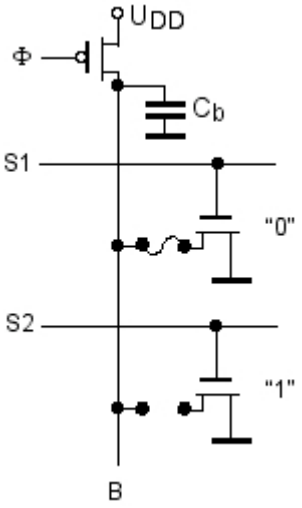

Rys. 9.9. Programowanie maską: (a) przez odłączenie tranzystora (maska kontaktów), (b) przez podwyższenie napięcia progowego (dodatkowa implantacja jonów w obszar kanału)

Pierwszy sposób jest bardzo prosty - polega na braku kontaktu tam, gdzie tranzystor ma się nie włączać. Drugi sposób polega na wprowadzeniu do obszaru kanału tranzystorów, które mają się nie włączać, dodatkowej dawki domieszki podnoszącej napięcie progowe do takiej wartości, że tranzystor nie może być włączony napięciem reprezentującym stan "1".

Sposoby programowania przy użyciu jednej maski są stosowane przez producentów, którym klient zleca wykonanie pamięci o zadanej zawartości nie interesując się, jak to będzie zrobione. Cała struktura pamięci pozostaje bez zmian, nowa zawartość wymaga wykonania tylko jednej nowej maski. Jest to ekonomicznie sensowne rozwiązanie w produkcji masowej.

Jeżeli sami projektujemy pamięć ROM, to sposób zaprogramowania jej zawartości jest niemal dowolny - można na przykład w ogóle nie wykonywać tranzystorów tam, gdzie ma być odczytana jedynka. Wymaga to jednak odpowiedniego zaprojektowania kilku masek, a nie tylko jednej. Pamięci ROM programowane maską są nieprzydatne, jeśli potrzebna jest mała liczba układów pamięci (bo wykonywanie nowej maski jest kosztowne) lub jeśli przewidujemy, że zawartość pamięci może wymagać zmian. W takim przypadku potrzebne są pamięci, które może zaprogramować - jeden raz lub wielokrotnie - użytkownik.

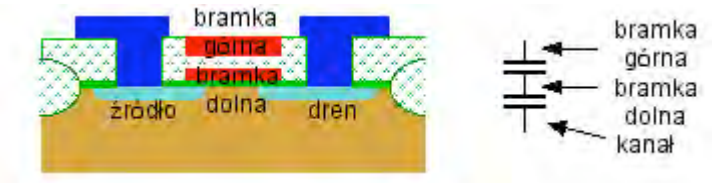
Najprostszy sposób programowania polega na kontrolowanym odłączaniu tranzystorów przez przepalanie połączeń impulsem prądu. Jest to ta sama idea, którą pokazano na rys. 9.9a, z tą różnicą, że odłączanie tranzystorów odbywa się poprzez doprowadzenie odpowiedniego sygnału elektrycznego, a nie w procesie produkcyjnym. Połączenie drenu z linią bitu wykonane jest w specjalny sposób (na przykład jest to bardzo wąska ścieżka polikrzemowa). Bramkę tranzystora oraz linię bitu trzeba spolaryzować napięciem znacznie wyższym od normalnego napięcia $U_{DD}$ . Powoduje to przepływ prądu o na tyle dużym natężeniu, że połączenie drenu z linią bitu ulega przepaleniu. Tę zasadę ilustruje rys. 9.10.




Rys. 9.10. Pamięć jednokrotnie programowalna przez przepalanie połączeń

Ten sposób programowania jest prosty, a układ pamięci tani. Można go oczywiście zaprogramować tylko jeden raz. Pamięci tego typu, niegdyś dość popularne, zostały dziś w dużym stopniu zastąpione przez pamięci wielokrotnie programowalne.

Pamięci wielokrotnie programowalne działają na zasadzie kontrolowanej zmiany napięcia progowego tranzystora. Jest to ta sama idea, co na rys. 9.9b, z tą różnicą, że zmianę napięcia progowego uzyskuje się na drodze elektrycznej, a nie w procesie produkcyjnym. Pamięci działające na tej zasadzie nazywane są ogólnie pamięciami EPROM (ang. Electrically Programmable ROM). Potrzebny jest do tego tranzystor o specjalnej, dwubramkowej konstrukcji - rys. 9.11.




Rys. 9.11. Dwubramkowy tranzystor nMOS stosowany w reprogramowalnych pamięciach ROM

Bramka górna jest podłączona do linii słowa, natomiast bramka dolna nie jest nigdzie podłączona i ze wszystkich stron otoczona dielektrykiem ($SiO_2$). Na kanał oddziaływuje potencjał bramki dolnej. Bramka górna i dolna tworzą pojemnościowy dzielnik napięcia. Jeśli pojemności bramka górna-bramka dolna oraz bramka dolna-kanał są jednakowe, to na nośniki w kanale oddziaływuje połowa napięcia przyłożonego do bramki górnej. Istotą działania omawianego tranzystora jako komórki pamięci jest możliwość trwałego naładowania bramki dolnej ładunkiem elektrycznym. W tym celu trzeba w kanale wytworzyć nośniki o wysokiej energii (t.zw. gorące nośniki). Takie nośniki powstają na przykład w procesie powielania lawinowego w złączu p-n dren-podłoże. Gorące nośniki mogą pokonać barierę dielektryka dzięki zjawisku tunelowania i dotrzeć do bramki dolnej, gdzie zostają uwięzione. Jeśli bramka dolna zostanie naładowana ujemnym ładunkiem elektronów, to do bramki górnej trzeba przyłożyć znacznie wyższe napięcie dodatnie, by włączyć tranzystor. Innymi słowy, poprzez ładowanie bramki dolnej można zmienić napięcie progowe tranzystora "widziane" przez bramkę górną. Napięcie to może się stać tak duże, że tranzystora nie da się włączyć zwykłym napięciem przyłożonym do bramki górnej. W ten sposób programuje się tranzystor służący jako komórka pamięci EPROM. Ładunek zgromadzony w bramce dolnej ma czas rozładowania rzędu 10 lat. Jednak możliwe jest rozładowanie tego ładunku przy pomocy naświetlenia układu pamięci silnym światłem ultrafioletowym (nośniki są uwalniane dzięki zjawisku fotoelektrycznemu). W ten sposób zawartość pamięci może być skasowana, a pamięć zaprogramowana powtórnie.

W nowszych układach pamięci EPROM, zwanych EEPROM (ang. Electrically Erasable Programmable ROM) zmiana zawartości pamięci może być dokonana także na drodze czysto elektrycznej. Do uwalniania nośników z bramki dolnej wykorzystuje się zjawisko znane jako tunelowanie Fowlera-Nordheima. Występuje ono w ultracienkich warstwach dielektrycznych w silnym polu elektrycznym. Aby uzyskać ten efekt, warstwa dielektryczna $SiO_2$ pod dolną bramką musi być niezwykle cienka (rzędu 10 nm lub mniej). Aby wywołać prąd tunelowy Fowlera-Nordheima, trzeba do bramki górnej przyłożyć znaczne napięcie ujemne (kilkanaście V). Wywołuje to ucieczkę ładunku ujemnego z dolnej bramki. Wysokie napięcia potrzebne do programowania wytwarzane są wewnątrz układów pamięci. Z punktu widzenia użytkownika układ pamięci ma jedno typowe napięcie zasilania, np. 5 V lub 3,3 V.

Programowanie pamięci EEPROM było dawniej procesem bardzo powolnym. Dziś istnieją pamięci tego rodzaju

(zwane pamięciami FLASH), które mają czas programowania na tyle krótki, że można je traktować jako specjalny rodzaj pamięci o swobodnym dostępie (RAM). Mają one dłuższy od pamięci statycznych i dynamicznych czas zapisu, ale za to są nieulotne - do podtrzymania ich zawartości nie jest potrzebne żadne źródło zasilania. Są dziś powszechnie stosowane na przykład w popularnych kartach pamięci stosowanych w cyfrowych aparatach fotograficznych, telefonach komórkowych i innych urządzeniach, w których potrzebna jest pamięć nieulotna o możliwie krótkim czasie zapisu. Coraz częściej też zastępują twarde dyski w komputerach przenośnych. Pewną wadą pamięci tego rodzaju jest ich ograniczona trwałość. Procesy zapisu, które są związane z transportem przez cienką warstwę dielektryka nośników o wysokiej energii, powodują powolne pogarszanie się jakości warstwy dielektrycznej. Po określonej liczbie cykli zapisu komórka pamięci przestaje prawidłowo działać. Ta liczba cykli wynosi typowo do kilkudziesięciu tysięcy do kilkuset tysięcy. Nie należą więc do rzadkości przypadki, gdy intensywnie używana pamięć typu FLASH przestaje prawidłowo działać już po niedługim okresie użytkowania. Aby wydłużyć czas użytkowania pamięci, w blokach pamięci stosowanych jako zamiennik hard dysków stosowana jest systematyczna okresowa zmiana przyporządkowania komórek pamięci do adresów, co zapobiega szybszemu zużywaniu się komórek związanych z adresami wykorzystywanymi częściej, niż inne.

Pamięci EEPROM i FLASH wymagają specjalnej technologii wytwarzania. Niewielu producentów oferuje technologie CMOS, w których można w tym samym układzie obok zwykłego układu cyfrowego CMOS wyprodukować moduł pamięci typu EEPROM lub FLASH. Jeśli taka możliwość istnieje, to projekt pamięci o zadanej pojemności i organizacji jest zwykle wykonywany automatycznie przy wykorzystaniu gotowych, opracowanych przez producenta układów zapisu, odczytu, adresowania i samych komórek pamięci.

## 9.4. Układy PLA i inne układy o strukturze matrycowej

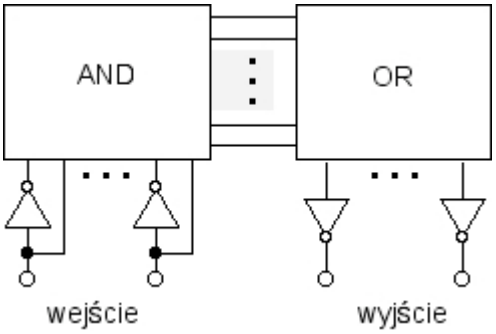
Jeśli spojrzeć na pamięć ROM jako na "czarną skrzynkę", nie interesując się jej wnętrzem, to jej działanie nie różni się od działania układu kombinacyjnego. W obu przypadkach mamy wejście, na które podawane są słowa binarne (w przypadku pamięci jest to wejście adresowe), i wyjście, na którym pojawiają się inne słowa binarne, przy czym każdemu słowu wejściowemu jednoznacznie przyporządkowane jest słowo wyjściowe. Zatem pamięć ROM może być traktowana jako jeden ze sposobów realizacji funkcji kombinacyjnej.

Istnieją takie funkcje kombinacyjne, które wygodniej jest zrealizować w postaci pamięci ROM, niż w tradycyjny sposób zestawiać z bramek NOR, NAND, NOT. Jest tak wtedy, gdy funkcja nie jest określona przez podanie wyrażeń logicznych, lecz w postaci tabeli przypisującej każdemu słowu wejściowemu odpowiednie słowo wyjściowe. Typowym przykładem są translatory kodów. Powszechnie przyjętym sposobem kodowania znaków alfanumerycznych (liter, cyfr itd.) jest kod ośmiobitowy zwany kodem ASCII. Istnieją jednak też inne kody, na przykład kod zwany dalekopisowym, gdzie literom i cyfrom przypisuje się kody pięciobitowe. Kod ASCII i kod dalekopisowy nie są ze sobą związane żadną prostą funkcją logiczną. Najprostszym sposobem tłumaczenia jednego kodu na drugi jest zatem użycie pamięci ROM, w której na przykład kod dalekopisowy danego znaku służy jako adres, a kod ASCII tego samego znaku jest odczytywany spod tego adresu.

Zauważmy równocześnie, że wnętrze pamięci ROM może być uważane za zespół bramek NOR, bowiem do każdej linii bitu dołączona jest równolegle pewna liczba tranzystorów nMOS.

Te spostrzeżenia prowadzą do wniosku, że wykorzystując bramki NOR takie, jak w pamięciach ROM, i tworząc z nich regularne struktury przypominające matryce pamięci, można realizować funkcje kombinacyjne w uporządkowanej, regularnej formie. Takie układy istnieją i zwane są zwykle układami **PLA** (ang. Programmable Logic Arrays). Nie jest to najoszczędniejszy i najbardziej korzytny z punktu widzenia szybkości działania sposób realizacji funkcji kombinacyjnych. Jego wielką zaletą jest jednak regularność budowy układu umożliwiająca łatwą automatyzację projektowania.

Typowy układ PLA składa się z dwóch matryc i pozwala łatwo realizować funkcje kombinacyjne wyrażone w postaci sumy iloczynów. Pierwsza matryca realizuje iloczyny i bywa nazywana matrycą AND, druga realizuje sumy i bywa nazywana matrycą OR. Obie w rzeczywistości składają się z bramek NOR. Funkcja wyrażona jako suma iloczynów zawsze może być przekształcona do postaci, w której występują tylko bramki NOR i inwertery.




Rys. 9.12. Budowa układu PLA

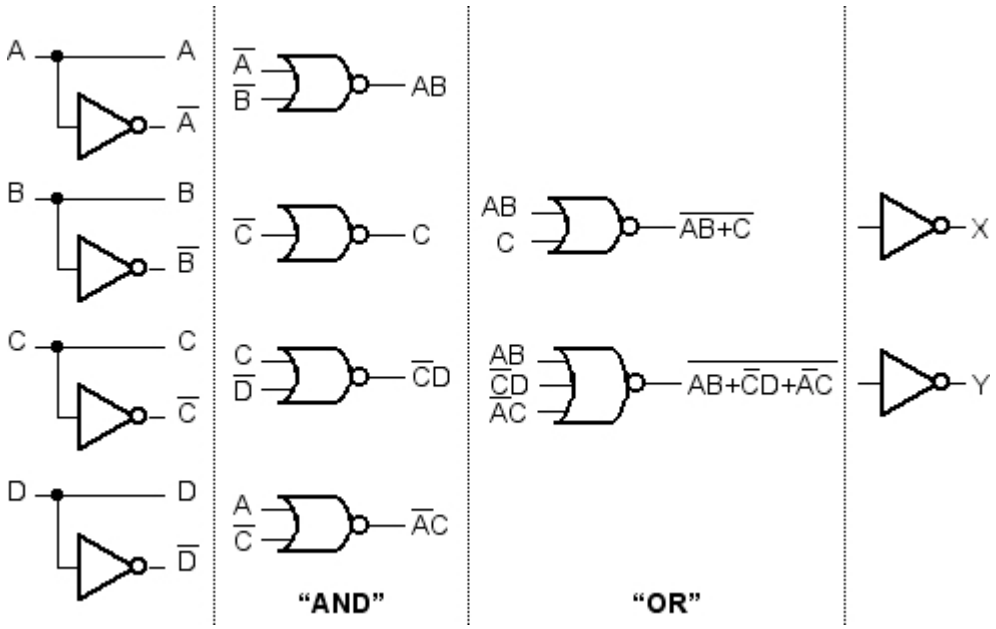
Przekształcenie funkcji logicznej do postaci, która umożliwia realizację tej funkcji w układzie PLA, najłatwiej zilustrować prostym przykładem. Przyjmijmy, że należy zbudować układ PLA realizujący funkcję

$$X = AB + C$$
$$Y = AB + \overline{C}D + \overline{A}C \qquad (9.1)$$

Funkcję tę można zapisać w równoważnej postaci

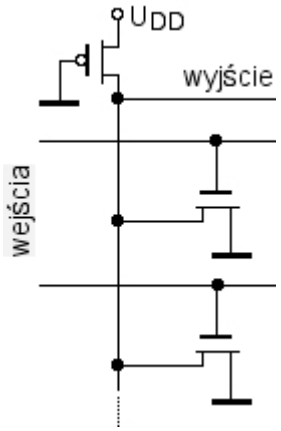
$$X = \overline{\overline{\left(\overline{A} + \overline{B}\right)} + C}$$
$$Y = \overline{\overline{\left(\overline{A} + \overline{B}\right)} + \overline{\left(C + \overline{D}\right)} + \overline{\left(A + \overline{C}\right)}} \qquad (9.2)$$

która prowadzi do następującego schematu



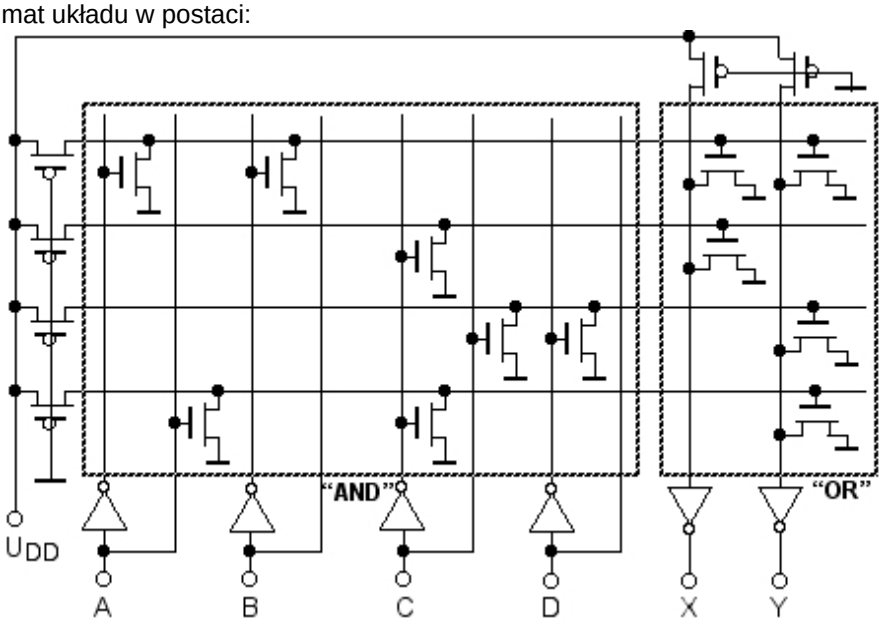

Rys. 9.13. Funkcja (9.1) zrealizowana przy pomocy bramek NOR i inwerterów (symbol bramki NOR z jednym wejściem został użyty dla zachowania regularności, jest to w istocie inwerter)

Jeśli bramki NOR zostaną zbudowane według zasady pokazanej na rys. 9.8a (tj. ze statycznym obciążeniem tranzystorem pMOS)




Rys. 9.14. Bramka NOR taka, jak w pamięciach ROM (patrz rys. 9.8a)

otrzymujemy schemat układu w postaci:




Rys. 9.15. Kompletny układ PLA

Regularność budowy układu PLA umożliwia łatwą automatyzację projektowania. Mając funkcję logiczną w postaci (9.1) można od razu określić, na których skrzyżowaniach wierszy i kolumn w matrycach "AND" i "OR" mają się znaleźć tranzystory. Wadą tego sposobu realizacji funkcji kombinacyjnych jest duża powierzchnia układu PLA - zauważmy, że jest w nim więcej pustych skrzyżowań, niż tranzystorów. Można co prawda powierzchnię trochę zmniejszyć, na przykład usuwając całkowicie niewykorzystane linie, jednak układ zbudowany w tradycyjny sposób z bramek NOR, NAND i inwerterów z reguły zająłby znacznie mniej miejsca.

Układ z bramkami wg rys. 9.14 ma tę wadę, że pobiera prąd, gdy przynajmniej jeden tranzystor nMOS jest włączony. Można tej wady uniknąć stosując bramki dynamiczne typu DOMINO, co jednak komplikuje układ.

Układy PLA były niegdyś bardzo pospolicie stosowane do projektowania złożonych układów kombinacyjnych, np. układów sterowania w mikroprocesorach. Obecnie ich znaczenie zmalało, bowiem układy kombinacyjne można również projektować w sposób zautomatyzowany wykorzystując komórki standardowe. Daje to na ogół porównywalną lub mniejszą powierzchnię i porównywalne lub lepsze parametry elektryczne (szybkość, pobór mocy).

Układy PLA takie, jak omówione wyżej, można porównać do pamięci ROM programowanych maską. Funkcję wykonywaną przez układ PLA określa struktura tego układu ustalona podczas jego produkcji. Nic jednak nie stoi na przeszkodzie, by do budowy układów PLA zastosować te same technologie programowania, które są wykorzystywane w reprogramowalnych pamięciach ROM. W ten sposób powstał niezwykle dziś ważny i użyteczny rodzaj cyfrowych układów scalonych - programowalne matryce bramkowe, znane pod wieloma nazwami zależnymi od wewnętrznej architektury, ale najczęściej określane mianem układów FPGA (skrót od ang. "Field Programmable Gate Array"). O takich układach była już mowa we wcześniejszych wykładach.

Najprostszą programowalną matrycą bramkową może być układ PLA, w którym tranzystory są na każdym skrzyżowaniu linii w matrycy, a te, które nie są potrzebne przy realizacji danej funkcji, są odłączane. Można tu wykorzystać znaną z układów ROM technikę przepalania połączeń (programowanie jest wówczas jednokrotne) lub tranzystory dwubramkowe takie, jak w pamięciach EPROM. Ta ostatnia technologia daje możliwość programowania wielokrotnego.

Układy FPGA osiągnęły jednak dużo wyższy stopień złożoności i uniwersalności, niż ten, który byłby możliwy przy zastosowaniu matryc typu PLA. Typowy układ FPGA zawiera zbiór komórek, których funkcję logiczne są programowalne, oraz zbiór programowalnych połączeń, które umożliwiają skonfigurownie układu o zadanym schemacie logicznym. Układy FPGA zawierają nie tylko bramki kombinacyjne, ale także elementy pamięciowe - przerzutniki, rejestry, a nawet złożone bloki, jak rdzenie procesorów - a więc umożliwiają budowę kompletnych układów cyfrowych o dużej złożoności.

Oprócz technik programowania jednokrotnego i wielokrotnego stosowanych w reprogramowalnych pamięciach ROM w układach FPGA stosowane są jeszcze dwa inne sposoby programowania. Jeden z nich jest odwrotnością przepalanych połączeń. Są to programowalne połączenia (znane pod angielską nazwą "antifuse"). Programowalne połączenie jest to obszar cienkiego dielektryka zlokalizowany między dwoma ścieżkami na dwóch różnych warstwach metalu. Normalnie warstwy te są odizolowane, jednak w obszarze cienkiego dielektryka może zostać wywołane impulsem podwyższonego napięcia przebicie, które prowadzi do trwałego połączenia dwóch warstw metalu. Raz wykonanego połączenia nie można zlikwidować, jest to więc technika programowania jednokrotnego. Są też układy FPGA, w których elementami programującymi połączenia są bramki transmisyjne. Takie układy mają wewnętrzną pamięć typu RAM, w której przechowywana jest informacja konfigurująca - które bramki transmisyjne są włączone, a które nie. Układy te wymagają zaprogramowania po każdym włączeniu zasilania, ponieważ nie dysponują pamięcią nieulotną. Program może zostać wpisany na przykład z zewnętrznej pamięci typu ROM. Zaletą takich układów FPGA jest możliwość zmiany wewnętrznej konfiguracji, a więc i wykonywanej funkcji, w czasie pracy urządzenia. Prowadzi to do nowej, fascynującej i mało dotąd wykorzystywanej w praktyce koncepcji układów samo-rekonfigurowalnych. Można sobie na przykład wyobrazić mikroprocesor, którego architektura wewnętrzna samoczynnie dostosowuje się do aktualnie wykonywanego zadania.

Układy FPGA nie będą jednak dalej omawiane w tym wykładzie, ponieważ ze względu na ich duże praktyczne znaczenie poświęcony jest im odrębny przedmiot.

## Bibliografia

[1] E. Sicard, "*Microwind & Dsch User's Manual*", National Institute of Applied Sciences INSA, Toulouse, 2003
(Jest to podręcznik do oprogramowania wykorzystywanego w tym wykładzie, załączony na płycie w postaci pliku PDF)
[2] J. M. Rabaey, "*Digital Integrated Circuits, a Design Perspective*", Prentice Hall, Inc. 1996

# Wykład 10: Zasady projektowania dużych układów cyfrowych

## Wstęp

Niezbyt długi wykład 10 stanowi zbiór praktycznych rad na temat: jak projektować układ cyfrowy, by zredukować do minimum ryzyko, że zaprojektowany układ nie będzie poprawnie działał.

Wykład dzieli się na dwie części. W pierwszej przedstawiono ogólną zasadę projektowania układów cyfrowych przeznaczonych do realizacji w postaci układów scalonych CMOS. Powinny to być układy synchroniczne. Przedstawiono szereg rozwiązań ryzykownych i niezalecanych, i wskazano jak można je zastąpić lub obejść. W drugiej części jest mowa o problemach, które pojawiają się, gdy układ jest naprawdę duży, tj. liczy setki tysięcy lub miliony bramek, a jego powierzchnia osiąga lub przekracza centymetr kwadratowy.

Ścisłe przestrzeganie reguł omówionych w wykładzie 10 nie daje niestety stuprocentowej pewności, że już pierwszy wyprodukowany egzemplarz układu będzie w pełni poprawnie działał i spełniał wszystkie wymagane warunki techniczne, ale bardzo poważnie zwiększa prawdopodobieństwo, że tak właśnie będzie. Trzeba też powiedzieć, że zaprojektowano wiele zupełnie prawidłowo działających układów, w których łamane są wszelkie reguły przedstawione w wykładzie 10. Jeżeli jednak projektant decyduje się na łamanie reguł, powinien robić to z pełną świadomością możliwych skutków, a projekt powinien być poddany wyjątkowo wnikliwej weryfikacji przy użyciu wszystkich dostępnych metod symulacyjnych.

## 10.1. Zasady projektowania układów synchronicznych

Podstawową zasadą projektowania układów cyfrowych, które mają być zrealizowane jako układy scalone CMOS, jest zasada, że powinny to być układy synchroniczne. Oznacza to, że:

! 
- **wszystkie elementy pamięciowe w układzie (przerzutniki, rejestry, bloki pamięci RAM) mogą zmieniać swój stan tylko w odpowiedzi na sygnał zegara,**
- **we wszystkich elementach pamięciowych zmiana stanu następuje w dokładnie tej samej chwili w odpowiedzi na to samo zbocze sygnału zegara.**

Ta zasada oznacza między innymi, że w układzie występuje tylko jeden główny, globalny sygnał zegara. Ewentualne dodatkowe sygnały (np. zegar dwufazowy o fazach "1" nie nakładających się) są generowane lokalnie z zachowaniem synchronizacji z zegarem głównym.

Tę ogólną zasadę trzeba uzupełnić kilkoma dalszymi regułami. Oto one:

Każdy układ zawierający elementy pamięciowe (przerzutniki, rejestry) musi mieć możliwość ustawienia w nich zadanych "startowych" stanów. Może to być zrealizowane w postaci zewnętrznego sygnału zerującego lub ustawiającego (w terminologii anglojęzycznej sygnał "Reset"). Można również zastosować układ automatycznie generujący sygnał "Reset" przy włączaniu zasilania. Takie układy, zwane "Power-on Reset", są dostępne jako gotowe komórki w bibliotekach dostarczanych przez producentów układów. Jeżeli stosujemy układ "Power-on Reset" do zapewnienia prawidłowego startu, to nawet wtedy nasz układ powinien zawierać też wejście dla zewnętrznego sygnału "Reset", aby była możliwość przywrócenia znanego z góry startowego stanu układu w czasie jego pracy. Taki globalny sygnał "Reset" może działać asynchronicznie (tj. zmieniać stan układu w dowolnej chwili, bez względu na stan sygnału zegara). Jeśli natomiast stosowany jest również lokalny sygnał "Reset" (lokalny oznacza ustawiający stany w elementach pamięciowych tylko w pewnym fragmencie układu), to taki sygnał musi działać synchronicznie.

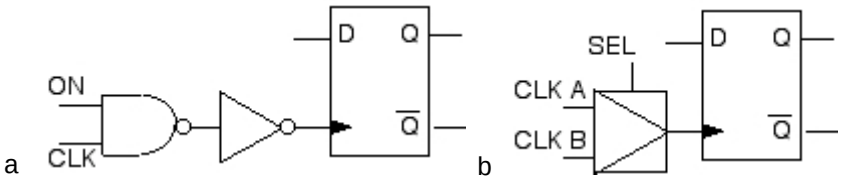
Nie należy używać żadnych układów, których poprawność działania jest uzależniona od wartości opóźnień wnoszonych przez bramki lub połączenia. Opóźnienia mogą się wahać w szerokich granicach, są bowiem uzależnione od rozrzutów produkcyjnych. Nie są w pełni przewidywalne i nie są powtarzalne.

Nie należy samemu zestawiać z bramek kombinacyjnych różnych układów przerzutników takich, jak przerzutniki RS, JK lub T. Używać należy wyłącznie przerzutników typu D. Mogą to być przerzutniki z biblioteki komórek dostarczonej przez producenta układów, a jeśli są samodzielnie projektowane, muszą być starannie zweryfikowane przez symulację elektryczną.

Teraz będą pokazane typowe przykłady rozwiązań układowych, których nie należy stosować, ponieważ są ryzykowne.

Bramkowany zegar

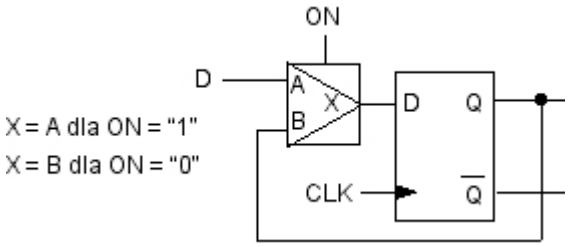
Nie należy sygnału zegara doprowadzać do przerzutników przez bramki NOR, NAND, multipleksery itd. Przykłady takich ryzykownych układów pokazuje rys. 10.1.




Rys. 10.1. Bramkowany zegar

Rys. 10.1a pokazuje układ, który mógłby służyć do kontrolowanego sygnałem ON włączania i wyłączania zegara. Miałoby to taki skutek, że po wyłączeniu zegara przerzutnik nie zapisywałby nowych danych, a na jego wyjściu trwałby ostatni stan sprzed wyłączenia zegara. Rys. 10.1b pokazuje układ, który mógłby służyć do wyboru jednego z dwóch sygnałów zegarowych w zależności od stanu sygnału SEL. Mogłyby to być na przykład dwa sygnały różniące się częstotliwością lub fazą. W obu przypadkach mamy do czynienia z pogwałceniem zasad budowy układów synchronicznych. Żaden z tych układów nie zapewnia pracy przerzutnika synchronicznej z innymi przerzutnikami, ponieważ bramki (rys. 10.1a) lub multiplekser (rys. 10.1b) wprowadzają opóźnienie. Ponadto zmiana stanu sygnału ON lub SEL może spowodować pojawienie się na wejściu przerzutnika w czasie trwania procesu przełączania impulsu, który zostanie potraktowany jako impuls zegarowy i spowoduje wpisanie danej w nieprzewidzianym momencie.

Układ z rys. 10.1a można zastąpić bezpiecznym układem pokazanym na rys. 10.2. W tym układzie, gdy wejście ON jest w stanie "1", każde dodatnie zbocze zegara powoduje wpis nowego stanu wejściowego. Gdy wejście ON jest w stanie "0", każde dodatnie zbocze zegara powoduje ponowne wpisanie stanu panującego na wyjściu, czyli podtrzymanie istniejącego stanu.



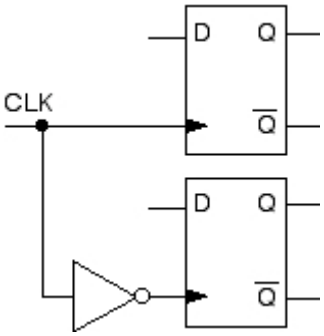

Rys. 10.2. Bezpieczny odpowiednik układu z rys. 8.1a

Układu z rys. 10.1b w ogóle nie należy brać pod uwagę.

Warto jednak dodać, że bramkowanie zegara bywa stosowane w dużych układach jako jeden ze sposobów ograniczania zużycia energii przez układ. Zatrzymanie zegara w bloku funkcjonalnym układu, który w danym momencie nie wykonuje żadnych operacji, sprowadza pobór mocy dynamicznej przez ten blok praktycznie do zera. Jeżeli taki blok zbudowany jest wyłącznie z bramek statycznych, które przy włączonym zasilaniu zachowują swój stan przez dowolnie długi czas, to wyłączenie zegara powoduje "uśpienie" bloku w istniejącym stanie, a ponowne włączenie powoduje kontynuowanie pracy od stanu sprzed uśpienia. Włączanie i wyłączanie zegara musi się jednak odbywać synchronicznie, w ściśle określonych chwilach czasowych.

Taktowanie obu zboczami zegara

Pokazuje to symbolicznie rys. 10.3.



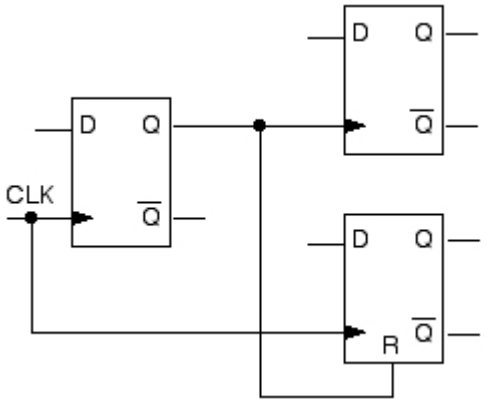

Rys. 10.3. Taktowanie dwoma zboczami zegara

Taki sposób taktowania narusza zasady budowy układów synchronicznych. Może się on niekiedy wydawać atrakcyjny dla przyspieszenia pracy układu - dolny przerzutnik jest gotowy do wpisu o pół taktu zegara wcześniej, niż gdyby oba przerzutniki były taktowane tym samym sygnałem. Jednak zmiany stanu obu przerzutników odbywają się w rzeczywistości w dość źle kontrolowanym odstępie czasu, bowiem wpływa na ten czas współczynnik wypełnienia sygnału zegara oraz opóźnienie inwertera, które mogą podlegać znacznym rozrzutom. Ponadto układ z takim taktowaniem stwarza problemy przy projekcie logicznym i jego weryfikacji oraz przy testowaniu. Należy raczej rozważyć taką konstrukcję układu, w której oba przerzutniki będą taktowane z dwa razy wyższą częstotliwością zegara.

Warto jednak dodać, że oba zbocza zegara są wykorzystywane do taktowania w pamięciach dynamicznych typu DDR (Double-Data Rate) - patrz poprzednia wykład.

Sterowanie wejścia zegarowego lub wejścia "Reset" z wyjścia innego przerzutnika

Takie układy pokazuje symbolicznie rys. 10.4.




Rys. 10.4. Ryzykowne sterowanie wejścia zegarowego i wejścia "Reset"

Te układy te nie są synchroniczne. Żaden z przerzutników sterowanych z wyjścia Q lewego przerzutnika nie zmienia stanu synchronicznie z zegarem CLK.

## 10.2. Problemy projektowania dużych układów

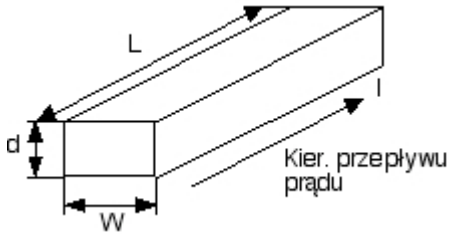
W dużych układach pojawiają się problemy, których dotąd nie omawialiśmy.

Gdy projektowany układ jest duży, niektóre połączenia mogą mieć długość rzędu milimetrów, a nawet centymetrów, a nie mikrometrów. Takie połączenia wprowadzają do układu znaczące pasożytnicze rezystancje, pojemności, a nawet indukcyjności. Czas propagacji sygnałów w długich połączeniach może być prównywalny, a nawet dłuższy od czasu propagacji w bramkach, i może decydować o maksymalnej szybkości działania układu.

Rezystancję ścieżki o długości $L$ i szerokości $W$ można obliczyć ze wzoru

$$R = R_s \frac{L}{W} \tag{10.1}$$

w którym $R_s$ jest rezystancją warstwową (zwaną także "rezystancją na kwadrat") ścieżki rezystora. O rezystancji warstwowej była mowa w wykładzie 4 i tam podana jest jej definicja. Dla jednorodnego prostokątnego obszaru przewodzącego takiego, jak na rys. 10.5, rezystancja warstwowa dana jest wzorem (10.2) (ogólniejsza definicja rezystancji warstwowej - patrz wykład 4).




Rys. 10.5. Prostokątna ścieżka przewodząca, kierunek przepływu prądu pokazuje strzałka

$$R_s = \frac{\rho}{d} \tag{10.2}$$

gdzie $\rho$ jest rezystywnością materiału ścieżki, a $d$ - jej grubością. Miarą rezystancji warstwowej jest om ($\Omega$), ale zwyczajowo używa się miana "om na kwadrat" ($\Omega/\square$). To określenie bierze się stąd, że iloraz $L/W$ można interpretować jako "liczbę kwadratów", jaką można wpisać w ścieżkę o długości $L$ i szerokości $W$. Typowe wartości rezystancji warstwowych wynoszą: dla ścieżek metalu (Al) - 0,03 ... 0,08 $\Omega$ (mniej dla ścieżek wykonanych z miedzi), dla obszarów polikrzemowych 10 ... 30 $\Omega$, dla obszarów źródeł i drenów typu n 10 ... 50 $\Omega$, dla obszarów źródeł i drenów typu p 30 ... 100 $\Omega$. W nowszych technologiach rezystancje warstwowe polikrzemu oraz obszarów źródeł i drenów mają wartości o rząd wielkości mniejsze od podanych wyżej, ponieważ są one pokrywane dobrze przewodzącą warstwą metalo-krzemową (np. $WSi_2$, $TiSi_2$, $PtSi_2$), jednak nadal ich rezystancje warstwowe są o około dwa rzędy wielkości wyższe, niż ścieżek z metalu. Jeśli obliczyć rezystancję ścieżki mającej szerokość 1 µm i długość 1 mm, czyli $L/W$ ("liczbę kwadratów") równą 1000, to ścieżka z metalu będzie miała rezystancję rzędu kilkudziesięciu $\Omega$, zaś ścieżka wykonana z polikrzemu - kilka do kilkudziesięciu k $\Omega$, zależnie od technologii. Widać stąd, że tylko ścieżki metalowe nadają się do wykonywania długich połączeń. Obszary przewodzące polikrzemowe oraz źródeł i drenów tranzystorów mogą służyć jedynie do krótkich połączeń lokalnych.

Znaczącą rezystancję wnoszą także kontakty. Typowe rezystancje kontaktów metalu do obszarów polikrzemu oraz źródeł i drenów tranzystorów wynoszą kilkadziesiąt $\Omega$, kontakty między warstwami metalu mają rezystancję rzędu 1 ... 5 $\Omega$.

Ścieżki są również związane z pojemnością do sąsiadujących obszarów. Przykładowo, pojemność na jednostkę powierzchni ścieżki polikrzemowej do podłoża układu jest rzędu 0,05 ... 0,06 fF/µm$^2$, dla metalu 1 pojemność ta jest rzędu 0,03 fF/µm$^2$. Ścieżka o wymiarach 1 µm x 1 mm ma powierzchnię 1000 µm$^2$, co daje pojemność ścieżki do podłoża na poziomie 60 fF dla polikrzemu i 30 fF dla metalu 1. W rzeczywistości pojemności te są większe, bowiem ścieżka o szerokości porównywalnej z grubością dielektryka, na którym leży, nie jest kondensatorem płaskim. Trzeba doliczyć pojemność obszarów krawędziowych, które mogą mieć pojemność porównywalną, a nawet większą od pojemności obliczonej wyżej.

Rezystancje i pojemności połączeń i kontaktów stanowią problem techniczny w kilku przypadkach:

- w przypadku długich ścież ek sygnałowych problemem jest czas propagacji sygnału związany z pojemnością i rezystancją,
- w przypadku długich ścież ek zasilających problemem jest spadek napięcia zasilania związany z rezystancją (będzie omawiany nieco dalej).

Czas propagacji sygnału w długich połączeniach może być na tyle długi, że układ synchroniczny przestanie działać synchronicznie. Są tu dwa nieco różne zagadnienia: **opóźnienia sygnałów logicznych** i **opóźnienia sygnału zegarowego**. Zajmiemy się najpierw pierwszym zagadnieniem.

Długie połączenie jest elementem RC o stałych rozłożonych. Załóżmy, że na jednym końcu takiego połączenia napięcie skokowo rośnie od 0 do $U_{DD}$. Po jakim czasie ten skok zostanie zaobserwowany na drugim końcu? Można pokazać, że czas propagacji sygnału w połączeniu o całkowitej pojemności C i całkowitej rezystancji R, mierzony jako czas narastania od 0 do $0{,}5U_{DD}$, jest w przybliżeniu równy $0{,}38RC$. Daje to dla ścieżki metalowej o wymiarach 1 µm x 1 mm czas propagacji rzędu $10^{-12}$ s, natomiast w przypadku ścieżki polikrzemowej tak samo obliczony czas propagacji jest rzędu 0,1 ... 1 ns. Jest to czas dłuższy od typowych czasów propagacji bramek CMOS. Trzeba przypomnieć, że w tych obliczeniach założono nieskończenie krótki czas narastania sygnału na początku ścieżki. W rzeczywistości długa ścieżka obciąża sterującą ją bramkę dość znaczną pojemnością, zatem czas narastania na początku ścieżki jest skończony i dość długi (zależy on od wymiarów tranzystorów w bramce sterującej). Wniosek, jaki z tego płynie, jest taki, że projekt bramki, do wyjścia której dołączone jest długie połączenie, musi uwzględniać pojemność tego połączenia. Widać także całkowitą nieprzydatność ścieżek polikrzemowych jako długich połączeń.

Czas propagacji sygnału w ścieżce jest proporcjonalny do stałej czasowej $t = RC$ tej ścieżki. Rezystancja $R$ jest proporcjonalna do ilorazu $L/W$ ścieżki, zaś pojemność do powierzchni, czyli do iloczynu $WLt$ rośnie proporcjonalnie do $L^2$. Ten wynik pokazuje, że jednym z warunków uzyskania dużej szybkości działania układów cyfrowych jest minimalizowanie długości *długich* połączeń. W obecnym stanie mikroelektroniki są one jednym z istotnych czynników ograniczających tę szybkość. Warto dodać, że szacunkowe obliczenia robione były dla ścieżki o długości 1 mm, która we współczesnych dużych układach nie należałaby wcale do najdłuższych. Nie są rzadkością ścież ki o długości rzędu kilkunastu milimetrów. Wówczas nawet czas propagacji w ścieżkach metalowych staje się porównywalny z czasami propagacji sygnałów w bramkach CMOS. Prowadzi to do wniosku, że dalszy wzrost wielkości układów będzie musiał prowadzić do odejścia od koncepcji układu ściśle synchronicznego. Wrócimy do tego zagadnienia w ostatnim wykładzie.

Zajmiemy się teraz zagadnieniem opóźnień sygnału zegara. Rozprowadzenie globalnego sygnału zegara w dużym układzie stwarza problemy z dwóch powodów.

Po pierwsze w dużym układzie całkowita pojemność wszystkich wejść zegarowych jest bardzo duża, może sięgać setek pikofaradów. Oznacza to, że układy generujące sygnał zegara muszą zapewnić krótkie czasy narastania i opadania sygnału przy dużych pojemnościach obciążających. Nie nadają się tu zwykłe bramki z tranzystorami o małych wymiarach. Konieczne są bufory. Zagadnienie to omawialiśmy już w ostatniej części wykładu 8.

Po drugie dla poprawnego działania układu synchronicznego trzeba, aby zbocza sygnału zegara docierały do wszystkich wejść zegarowych w tym samym momencie. Oznacza to, że na drodze od pierwotnego źródła sygnału zegara do każdego wejścia zegarowego powinna znaleźć się taka sama liczba takich samych buforów obciążonych takimi samymi pojemnościami. Również długość połączeń powinna być taka sama. O tym również była mowa w wykładzie 8. Spełnienie tych warunków w dużym układzie jest bardzo trudne. Jest to drugi powód odchodzenia od koncepcji układu ściśle synchronicznego.

W przypadku dużych układów poważnym problemem jest też zaprojektowanie sieci ścież ek zasilania. Rezystancje tych ścieżek i ewentualnie kontaktów mogą powodować znaczne, zmienne w czasie i zakłócające działanie układu spadki napięcia.

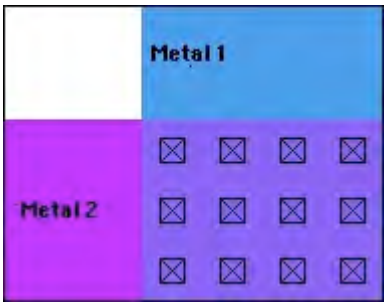
Pobór prądu przez bramki CMOS ma charakter krótkich impulsów o dużej amplitudzie - przypomnij sobie wyniki symulacji inwertera w wykładzie 7. Jeśli dla pojedynczego inwertera szczytowa wartość prądu w impulsie osiąga kilkaset mikroamperów, to dla tysięcy równocześnie przełączających bramek otrzymamy prądy sięgające amperów. Wówczas nawet rezystancje rzędu pojedynczych omów powodują znaczne spadki napięcia. Powoduje to zaburzenia w działaniu układów, w tym między innymi przenikanie zakłóceń między blokami układu (patrz wykład 7 i rys. 7.3). Przy bardzo krótkich czasach narastania impulsów prądu i długich ścieżkach zasilających wpływ na spadki napięć na tych ścieżkach może mieć nie tylko ich rezystancja, ale i indukcyjność.

Aby zminimalizować skutki spadków napięcia na ścieżkach zasilających, trzeba przestrzegać kilku zasad:

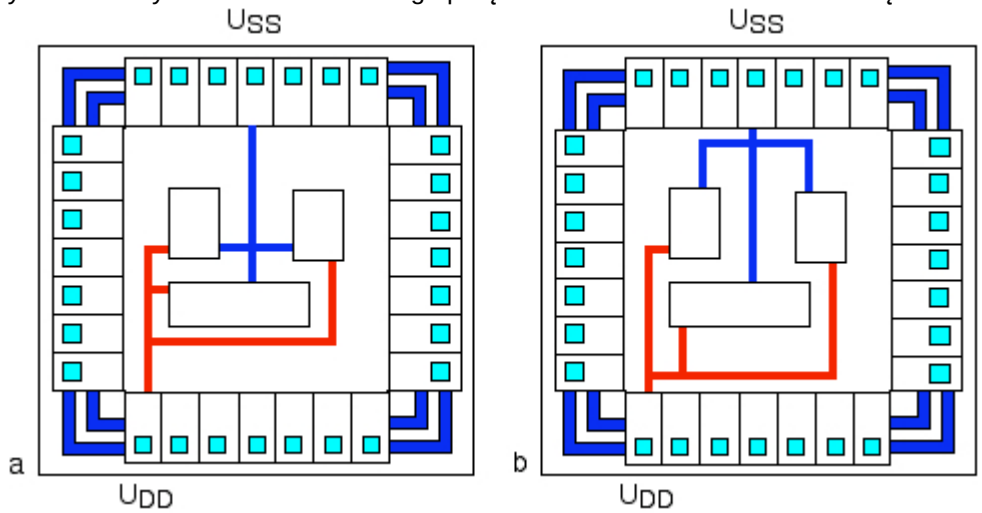
- szerokość ścieżek zasilających powinna być jak największa, a długość jak najmniejsza,
- ścieżki zasilające powinny być, jeśli to możliwe, prowadzone w tej warstwie metalu, która ma najmniejszą

rezystancję warstwową,

- należy unikać przechodzenia ścież ek zasilania z warstwy na warstwę, bo kontakty wprowadzają dodatkową rezystancję; w razie potrzeby stosować wielokrotne kontakty (rys.10.6),
- aby zminimalizować przenikanie zakłóceń, ścieżki zasilania dla różnych bloków powinny biec osobno i spotykać się dopiero przy polach montażowych (rys. 10.7),
- w razie potrzeby stosować więcej niż jedno pole montażowe masy i zasilania, rozdzielając zasilania różnych bloków.



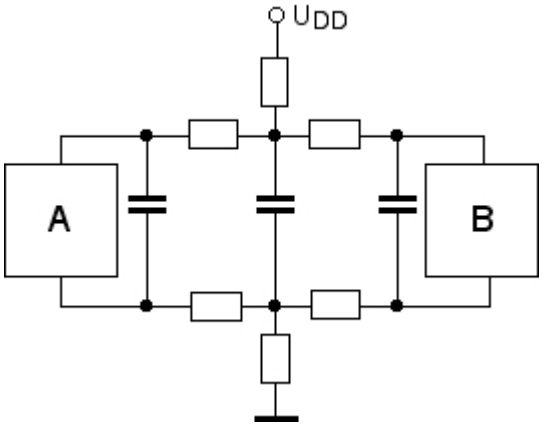

Rys. 10.6. Przykład wielokontaktowego połączenia ścieżki metalu 1 ze ścieżką metalu 2




Rys. 10.7. Rozprowadzenie ścieżek masy ($U_{SS}$) i zasilania ($U_{DD}$): (a) niekorzystne - długie odcinki wspólne wprowadzają wspólną rezystancję, (b) korzystniejsze - ścieżki łączą się blisko pól montażowych

Największe układy liczące miliony elementów pobierają tak duże prądy, że zasilanie i masę trzeba rozdzielać pomiędzy wiele pól montażowych. Poszczególne bloki układu mają odrębne połączenia zasilające, które łączą się dopiero poza obudową układu, na pakiecie drukowanym.

Stsosowane bywają również pojemności odsprzęgające. Ich rola polega na łagodzeniu spadków napięcia - w chwili wzrostu poboru prądu jest on lokalnie dostarczany z naładowanego kondensatora, co zmniejsza prąd płynący przez rezystancję ścieżki ze źródła zasilania. Potem, w okresie małego poboru prądu między momentami przełączania, kondensator doładowuje się ze źródła zasilania ponownie do pełnego napięcia zasilania.




Rys. 10.8. Kondensatory odsprzęgające. Rezystory symbolizują rezystancje ścieżek.

Ta bardzo dawno znana i stosowana w elektronice zasada może być zrealizowana w układzie scalonym przez zastosowanie kondensatorów zbudowanych przy użyciu cienkiego tlenku bramkowego o dużej pojemności jednostkowej. Takie kondensatory powinny znajdować się jak najbliżej bloku, którego napięcie zasilania mają stabilizować. Jednak pojemności, jakie można w ten sposób zrealizować w układzie scalonym, nie są na tyle duże, aby całkowicie wyeliminować niebezpieczeństwo zaburzeń w pracy układu spowodowanych spadkami napięcia zasilania na rezystancji ścieżek.

Przy projektowaniu połączeń, przez które płyną duże prądy, trzeba także przestrzegać maksymalnej obciążalności prądowej ścieżek podawanej przez producenta układów. Przekroczenie dopuszczalnej obciążalności prowadzi do uszkodzeń układów w czasie ich eksploatacji.

# ZADANIA DO WYKŁADU 10

**Zadanie 1**

Dana jest technologia CMOS, w której $U_{Tn}$ = 0,75 V, $U_{Tp}$ = - 0,85 V, $\mu_n/\mu_p$ = 2,9, minimalna długość kanału tranzystora $L$ = 0,7 µm, minimalna szerokość kanału tranzystora $W$ = 1µm (są to te same dane, co w zadaniach do wykładu 7). Napięcie zasilania $U_{DD}$ = 5 V. Rozważ inwerter o minimalnej szerokości kanału tranzystora nMOS i szerokości kanału tranzystora pMOS zapewniającej napięcie przełączania równe 2,5 V (szerokość ta była obliczona w zadaniu 2 do wykładu 7). Załóż, że całkowita pojemność obciążająca inwertera wynosi 50 fF, oraz przyjmij $\mu_n C_{ox}$ = 80 µA/V$^2$, $\mu_p C_{ox}$ = 27 µA/V$^2$. Następnie określ długość ścieżki metalu 1 oraz ścieżki polikrzemu, dla których opóźnienie sygnału ma taką samą wartość, jak dla rozważanego inwertera. Przyjmij, że dla ścieżki metalu rezystancja warstwowa $R_{Smet}$ = 0,028 Ω/□, dla ścieżki polikrzemu rezystancja warstwowa $R_{Spoli}$ = 20 Ω/□, zaś pojemności do podłoża układu wynoszą: dla metalu 1 $C_{met}$ = 0,03 fF/µm$^2$, dla polikrzemu $C_{poli}$ = 0,06 fF/µm$^2$.
Wskazówka: wykorzystaj wyniki z zadań do wykładu 7, użyj zależności podanych w wykładzie 10, część 2.

**Zadanie 2**

Dany jest blok funkcjonalny w układzie scalonym zawierający 1500 bramek. Pojedyncza bramka przy przełączaniu pobiera średnio prąd o wartości szczytowej równej 0,5 mA. Przyjmij, że przeciętnie w bloku w chwili przełączania zmienia stan połowa bramek, a druga połowa nie bierze udziału w przełączaniu. Do bloku napięcie zasilania doprowadzono ścieżką metalu 2 (o rezystancji warstwowej $R_{Smet}$ = 0,02 Ω/□) o długości 200 µm. Oblicz minimalną szerokość ścieżki, dla której spadek napięcia w chwili przełączania nie przekroczy 0,2 V.

## Bibliografia

[1] E. Sicard, *"Microwind & Dsch User's Manual"*, National Institute of Applied Sciences INSA, Toulouse, 2003
(Jest to podręcznik do oprogramowania wykorzystywanego w tym wykładzie, załączony na płycie w postaci pliku PDF)
[2] J. M. Rabaey, *"Digital Integrated Circuits, a Design Perspective*", Prentice Hall, Inc. 1996

# Wykład 11: Testowanie i testowalność układów cyfrowych

## Wstęp

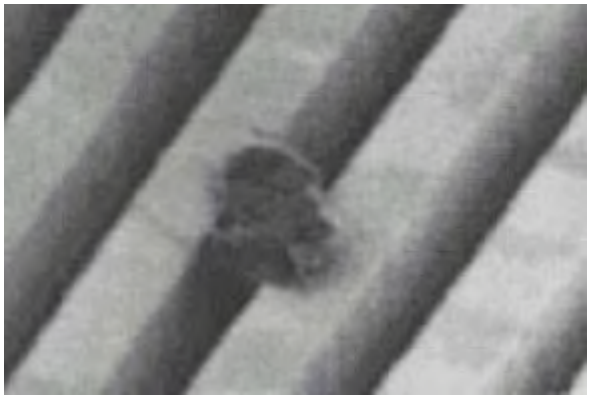
Jak wiemy, w procesach produkcyjnych występują zaburzenia. W wyniku tych zaburzeń część wyprodukowanych układów zawiera defekty, które powodują, że układy nie działają tak, jak powinny. Trzeba je więc testować i odrzucać układy wadliwe. Testowanie dużych układów cyfrowych jest dziś poważnym problemem technicznym. Zbadanie, czy układ zawierający dziesiątki milionów tranzystorów wykonuje poprawnie swe - zwykle bardzo złożone - funkcje dla wszystkich możliwych stanów i sekwencji sygnałów na wejściach jest nie do wykonania w rozsądnie krótkim czasie. Badanie sprawności układu wymaga więc użycia specjalnych metod. Już na etapie projektu architektury i projektu logicznego układu projektant powinien przewidzieć, w jaki sposób układ będzie testowany. Wiele sposobów testowania wymaga bowiem wprowadzenia do układu dodatkowych bloków funkcjonalnych.

Zarys tych zagadnień będzie przedstawiony w wykładzie 11. Jest to ważny wykład. Problemy testowania są często ignorowane przez niedoświadczonych projektantów. W rezultacie otrzymuje się układy, których w ogóle nie da się przetestować. A przecież układ lub system, o którym nie wiadomo, czy działa poprawnie, jest bezużyteczny.

## 11.1. Defekty i uszkodzenia w układach cyfrowych

Typowe bramki cyfrowe CMOS, a zwłaszcza bramki statyczne (takie, jakie były omawiane w wykładzie 7), są mało wrażliwe na parametry tranzystorów. Tranzystory działają jak przełączniki - przewodzą prąd lub nie. Ich charakterystyki mają wpływ na takie parametry, jak szybkość działania i pobór prądu, ale nie na to, czy bramka w ogóle działa, czy też nie. Dlatego głównym problemem w układach cyfrowych nie są uszkodzenia parametryczne, ale uszkodzenia katastroficzne.

W układach cyfrowych CMOS najczęściej obserwujemy uszkodzenia katastroficzne powodowane przez defekty strukturalne (była o nich mowa w wykładzie 2, część 2). Przykład takiego defektu - cząstki przewodzącej, która zwiera ścieżki przewodzące - pokazuje rys. 11.1.



Rys. 11.1. Przykład defektu - zwarcia ścieżek przewodzących (zdjęcie wykonane mikroskopem elektronowym)

Najpospolitsze są defekty powodujące zmiany w połączeniach w układzie: zwarcia ścieżek oraz przerwy w ścieżkach. W układach mających wiele warstw połączeń, i co za tym idzie - dużą liczbę kontaktów między warstwami metalu, dość częste są też uszkodzenia polegające na braku kontaktu między warstwami metalu, co jest zwykle spowodowane niedostatecznym wytrawieniem okna kontaktowego w dielektryku. Zdarzają się też inne defekty, jak na przykład "dziury" w cienkim dielektryku bramkowym tranzystorów MOS powodujące połączenie elektryczne bramki z kanałem i całkowicie zniekształcające charakterystyki uszkodzonego tranzystora.

W dużych cyfrowych układach scalonych spotyka się także uszkodzenia katastroficzne, które mają swe przyczyny w nadmiernym rozrzucie produkcyjnym parametrów elementów, a nie w defektach strukturalnych. Ten nadmierny rozrzut może na przykład powodować, że czasy propagacji sygnałów zegarowych w buforach zegara, które powinny być jednakowe, są w rzeczywistości różne, wobec czego zegar nie dociera jednocześnie do wszystkich bloków układu. Małe różnice zawsze istnieją i nie są szkodliwe. Jeśli jednak różnice w czasie propagacji sygnałów zegara stają się porównywalne z okresem zegara, to układ przestaje być układem synchronicznym. Taki układ może działać prawidłowo pod warunkiem obniżenia częstotliwości zegara, ale bywa i tak, że nie pracuje prawidłowo przy żadnej częstotliwości. Mamy więc wtedy do czynienia z uszkodzeniem katastroficznym. Tutaj jednak nie będziemy zajmować się tego rodzaju uszkodzeniami.

Defekty mogą też powstać w układzie w trakcie jego eksploatacji. Takie defekty w układach prawidłowo zaprojektowanych, prawidłowo wykonanych i prawidłowo eksploatowanych są, jak wiemy, niezwykłe rzadkie. We współcześnie produkowanych układach scalonych są trzy główne mechanizmy uszkodzeń w trakcie eksploatacji:

- elektromigracja powodująca przerwy ścieżek połączeń,
- naprężenia mechaniczne powodujące mechaniczne uszkodzenia płytki układu lub połączeń,
- degradacja właściwości elektrycznych tlenku bramkowego wywołana "gorącymi" (wysokoenergetycznymi) nośnikami ładunku.

Elektromigracja polega na przemieszczaniu się atomów metalu pod wpływem silnego strumienia nośników (czyli prądu o dużej gęstości). Podlegają jej zwłaszcza połączenia aluminiowe (aluminium jest, jak wiemy, bardzo miękkim metalem, czyli jego wiązania międzyatomowe są słabe). Elektromigracja jest procesem, który bardzo nasila się przy podwyższonej temperaturze. Atomy aluminium są powoli przemieszczane z miejsc na ścieżce, gdzie gęstość prądu jest największa (a więc tam, gdzie ścieżki mają zmniejszony przekrój poprzeczny, na przykład są zwężone) do sąsiadujących z takimi miejscami innych fragmentów ścieżki. Po pewnym czasie w miejscu przewężonym ścieżka ulega przerwaniu. Producenci układów podają maksymalną gęstość prądu w ścieżkach, której nie wolno przekraczać. Gęstość ta jest funkcją maksymalnej temperatury, w jakiej pracować ma układ. Projektant nie tylko powinien projektować scieżki o szerokościach dostosowanych do prądów w nich płynących, ale także starać się, by kształty ścieżek były możliwie jak najprostsze, bez zbędnych załamań i przewężeń.

Naprężenia mechaniczne są zwykle wywołane błędami lub wadami produkcyjnymi w montażu układów w obudowach. Przykładowo, użycie niewłaściwego lutu do przylutowania płytki układu do metalowej podstawki może powodować powstawanie naprężeń mechanicznych przy zmianach temperatury na skutek dużych różnic rozszerzalności cieplnej krzemu i metalu. Takie naprężenia mogą powodować mikropęknięcia w płytkach bądź w połączeniach i w konsekwencji uszkodzenie układu.

Degradacja właściwości tlenku bramkowego jest wywołana bombardowaniem tlenku przez nośniki uzyskujące dużą energię w polu elektrycznym. Ten rodzaj uszkodzeń jest typowy zwłaszcza dla tranzystorów stanowiących komórki pamięci w pamięciach typu EEPROM i FLASH. Była o tym mowa w wykładzie 9.

Ponieważ uszkodzenie może wystąpić w układzie w czasie jego eksploatacji, problem testowania to nie tylko problem zbadania układu tuż po jego wyprodukowaniu, ale w niektórych przypadkach także problem sprawdzania poprawności działania układu w czasie, gdy układ działa w urządzeniu. Dotyczy to urządzeń i systemów, których uszkodzenie może wywołać wielkie i nieodwracalne szkody, na przykład katastrofę lub utratę ludzkiego życia.

## 11.2. Modele defektów

Aby móc opracować testy dla układu cyfrowego, musimy wiedzieć, jak wpływają defekty w układzie na wykonywane przez ten układ funkcje. Opis działania układu uszkodzonego, tj. zawierającego defekt, nazywamy **modelem uszkodzenia**. W przypadku układu cyfrowego model uszkodzenia określany jest w odniesieniu do funkcji logicznej wykonywanej przez układ, tj. określa, jak zmienia się funkcja logiczna pod wpływem uszkodzenia.

Najprostszym modelem uszkodzenia jest model znany pod nazwą "stałe zero/stała jedynka" (lub niekiedy "sklejenie do zera/sklejenie do jedynki", w jęz. angielskim "stuck at 0/stuck at 1"). Będziemy go w skrócie oznaczać symbolem "SA0/SA1".

**! Model uszkodzenia "SA0/SA1" polega na założeniu, że każdy defekt w układzie prowadzi do tego, że w jakimś jego węźle stan logiczny nie zmienia się - zawsze jest tam stan "0" lub stan "1". Chodzi tu oczywiście o te węzły w układzie, w których w prawidłowo działającym układzie stan logiczny powinien zmieniać się.**

Model "SA0/SA1" nie ma żadnego uzasadnienia teoretycznego, a symulacje komputerowe i praktyka pokazują, że nie więcej niż kilkanaście procent wszystkich defektów strukturalnych daje w efekcie uszkodzenie typu "SA0/SA1". Znacznie częściej spotyka się uszkodzenia polegające na tym, że zmienia się funkcja logiczna wykonywana przez bramkę czy też blok, w którym wystąpił defekt, ale w żadnym z węzłów nie panuje na stałe stan "0" ani stan "1". Spotyka się defekty, w wyniku których układ kombinacyjny staje się układem z pamięcią (tj. odpowiedź układu zależy nie tylko do aktualnych stanów wejść, ale i od historii zmian stanów). Spotyka się też defekty, w wyniku których w jakimś węźle układu panuje napięcie nie będące ani prawidłowym stanem "0", ani "1". Mimo to model "SA0/SA1" jest powszechnie i niemal wyłącznie stosowany w teorii i praktyce testowania układów cyfrowych. Wieloletnia praktyka pokazała bowiem, że jest to model skuteczny. Otrzymywane przy użyciu tego modelu zestawy testów dla układów cyfrowych pozwalają przetestować te układy z dostateczną dla celów praktycznych wiarygodnością.

Wielką zaletą modelu "SA0/SA1" jest jego prostota, a przede wszystkim niezależność od struktury fizycznej układu. Dla stosowania w praktyce tego modelu wystarczy znać schemat logiczny projektowanego układu. Nie jest konieczna znajomość schematu elektrycznego ani topografii układu. Dalej zobaczymy, w jaki sposób model "SA0/SA1" jest wykorzystywany do otrzymywania zestawów testów dla układów cyfrowych.

Istnieją inne, bliższe rzeczywistości fizycznej modele uszkodzeń. Drogą odpowiednich symulacji komputerowych można na przykład przewidzieć możliwe defekty strukturalne w danej bramce, obliczyć ich prawdopodobieństwo, a następnie określić, jaką funkcję logiczną będzie wykonywać bramka z każdym z tych defektów. W ten sposób uzyskuje się szczegółową informację o rodzajach i prawdopodobieństwie uszkodzeń w danej bramce. Informację tę można potem wykorzystać do określenia najlepszego zestawu testów dla układu zawierającego tak scharakteryzowane bramki. Wadą takiego podejścia jest konieczność brania pod uwagę konkretnej struktury fizycznej układu - topografii bramek i połączeń między nimi. Od tej topografii zależy bowiem prawdopodobieństwo wystąpienia różnych rodzajów defektów strukturalnych. Na przykład prawdopodobieństwo wystąpienia zwarcia między ścieżkami połączeń zależy od ich wzajemnej odległości. Chociaż sposób testowania wychodzący od realnych uszkodzeń i ich prawdopodobieństwa daje możliwość otrzymania lepszych zestawów testów, to jednak konieczność znajomości konkretnej struktury fizycznej układu oraz duża złożoność symulacji, które trzeba wykonać, powodują że ten sposób otrzymywania zestawów testów nie znalazł dotąd zastosowania praktycznego.

W dalszych częściach wykładu będzie używany model "SA0/SA1".

## 11.3. Testowanie układów cyfrowych

Wyobraźmy sobie, że kupujemy w sklepie kalkulator i chcemy sprawdzić, czy działa prawidłowo. W tym celu wykonujemy w sposób mniej lub bardziej przypadkowy kilka obliczeń, których prawidłowy wynik jest nam znany. Wyniki wyświetlone przez kalkulator są poprawne. Czy po takim teście możemy być pewni, że kalkulator działa całkowicie prawidłowo? Nie! Taką pewność uzyskalibyśmy dopiero wtedy, gdybyśmy wykonali wszystkie możliwe działania dla wszystkich możliwych argumentów. Taki test nosi nazwę **wyczerpującego testu funkcjonalnego**. Jest oczywiste, że nie da się go wykonać kupując w sklepie kalkulator.

A gdyby test był wykonywany przez szybki tester, taki, jaki stosowany jest do testowania układów scalonych?

Przypuśćmy, że testujemy układ kombinacyjny mający wejście $n$-bitowe. Jedno słowo $n$-bitowe podane na wejście dla celów testowania będziemy nazywali **wektorem testowym**. Wyczerpujący test funkcjonalny wymaga podania na wejście wszystkich możliwych kombinacji zer i jedynek, czyli $2^n$ wektorów testowych, i zbadania poprawności odpowiedzi układu. Jeżeli testujemy układ sekwencyjny mający wejście $n$-bitowe i $m$ stanów wewnętrznych, wyczerpujący test funkcjonalny wymaga podania na wejście $2^{(n+m)}$ wektorów testowych. Czy to dużo? Wyobraźmy sobie prosty układ, dla którego $n$=25, $m$=50. Potrzebne jest więc $2^{75}$ wektorów testowych. Jest to liczba równa około $3{,}8 \times 10^{22}$. Jeżeli dysponujemy testerem, który potrzebuje 1 μs na jeden wektor testowy, testowanie naszego prostego układu będzie trwało około $10^9$ lat!

Ten prosty przykład (przytoczony za poz. [1] literatury do tego wykładu) pokazuje, że

! **wyczerpujący test funkcjonalny jest niemożliwy do wykonania nawet dla zupełnie prostych układów.**

Potrzebne są więc inne sposoby testowania. Omówimy je nieco dalej.

A ponieważ wyczerpujący test funkcjonalny nie jest możliwy, musimy mieć miarę jakości testów pozwalającą nam oszacować, jaki jest stopień wiarygodności wyników testowania. Innymi słowy, musimy znać odpowiedź na pytanie, jaki procent wszystkich możliwych uszkodzeń w układzie wykryje zestaw wektorów testowych, który nie jest wyczerpującym testem funkcjonalnym. Ten procent będziemy nazywali **poziomem wykrywalności uszkodzeń** dla danego zestawu wektorów testowych. Im niższy ten poziom, tym większe prawdopodobieństwo, że w wyniku testowania zakwalifikowany zostanie jako sprawny układ, który w rzeczywistości jest uszkodzony.

Poziom wykrywalności uszkodzeń jest miarą teoretyczną określaną przy zastosowaniu modelu uszkodzeń "SA0/SA1" i przy dalszych upraszczających założeniach. Jedno z nich to założenie, że uszkodzenie typu "SA0/SA1" może z jednakowym prawdopodobieństwem wystąpić w każdym węźle układu. Tak w rzeczywistości nie jest. W związku z tym poziom wykrywalności uszkodzeń nie wiąże się bezpośrednio i jednoznacznie z faktyczną liczbą układów wadliwych. Oto wyniki badań statystycznych pokazujących typowy związek poziomu wykrywalności uszkodzeń z faktycznym procentem układów wadliwych wśród układów zakwalifikowanych jako sprawne. Dane te *nie są uniwersalne*, zostały uzyskane dla określonych rodzajów układów i określonych sposobów ich testowania, ale dają pewne pojęcie o istniejącej tu zależności:

| Poziom wykrywalności uszkodzeń | Procent układów wadliwych wśród układów zakwalifikowanych jako sprawne |
|---|---|
| 50% | 7% |
| 90% | 3% |
| 95% | 1% |
| 99% | 0,1% |
| 99,9% | 0,01% |

Wprowadzimy jeszcze dwa inne nowe pojęcia: **kontrolowalności** węzłów i **obserwowalności** węzłów w układzie. Metody testowania, o których będzie dalej mowa, wykorzystują model uszkodzenia "SA0/SA1" zakładający, że uszkodzenie polega na ustalonym stanie "0" lub "1" w jakimś węźle układu. Zatem dla sprawdzenia, czy wystąpiło uszkodzenie, trzeba będzie ustawiać określone stany w węzłach i sprawdzać te stany. Mówimy, że

! **badany węzeł jest kontrolowalny, jeśli sekwencja wektorów testowych o skończonej długości pozwala na ustawienie w tym węźle żądanego stanu ("0" lub "1").**

Węzeł jest tym łatwiej kontrolowalny, im krótsza jest ta sekwencja.

Podobnie mówimy, że

**! badany węzeł jest obserwowalny, jeśli sekwencja wektorów testowych o skończonej długości pozwala na pojawienie się na wyjściu układu stanu pozwalającego określić, jaki był stan ustawiony w węźle ("0" lub "1").**

Węzeł jest tym łatwiej obserwowalny, im krótsza jest ta sekwencja.

Dla prawidłowo zaprojektowanych pod względem logicznym układów kombinacyjnych kontrolowalność i obserwowalność węzłów nie stanowi większego problemu. Natomiast w przypadku układów sekwencyjnych zmiana stanu w niektórych wewnętrznych węzłach układu może wymagać niezwykle długich sekwencji wektorów testowych. Jak zobaczymy, istnieją sposoby projektowania układów zasadniczo poprawiające kontrolowalność i obserwowalność węzłów w układach sekwencyjnych.

Posługując się zdefiniowanymi wyżej pojęciami można zadanie określenia sposobu testowania układu oraz wygenerowania odpowiedniego zestawu wektorów testowych sformułować następująco:

**! Celem jest uzyskanie jak najkrótszej sekwencji wektorów testowych zapewniającej wymagany poziom wykrywalności uszkodzeń.**

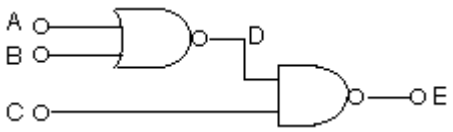
Trzeba się przy tym pogodzić z faktem, że dla dużych i złożonych układów poziom wykrywalności uszkodzeń będzie z reguły niższy niż 100%. Nacisk na to, że sekwencja wektorów testowych powinna być jak najkrótsza, bierze się stąd, że testowanie dużych układów jest kosztowne. Wynika to z bardzo wysokiego kosztu nowoczesnych testerów. Nierzadkie są przypadki, gdy koszt testowania układu jest porównywalny z kosztem jego wyprodukowania. Wprowadziliśmy już pojęcie wyczerpującego testu funkcjonalnego i wiemy, że nie da się go wykonać w rozsądnym czasie. Dlatego powszechnie stosowana jest metoda zwana **testowaniem strukturalnym**. W tej metodzie nie badamy, czy układ prawidłowo wykonuje swą funkcję. Możemy nawet w ogóle nie interesować się, jaka ona jest.

**! Zadaniem testowania strukturalnego jest sprawdzenie, czy w procesie produkcyjnym nie wystąpił defekt naruszający prawidłową strukturę układu.**

Zakłada się przy tym, że układ jest poprawnie zaprojektowany, tzn. brak defektów naruszających prawidłową strukturę układu jest równoznaczny z prawidłowym działaniem układu.

W testowaniu strukturalnym wykorzystuje się model uszkodzenia "SA0/SA1", a sam proces testowania wygląda następująco. Wybiera się konkretny węzeł układu i podaje się na wejścia wektor testowy (lub sekwencję wektorów testowych) ustawiający w tym węźle konkretny stan (na przykład "0"). Następnie obserwując odpowiedź układu na wyjściu określa się, czy ten stan udało się ustawić, czy też nie. Tę samą procedurę powtarza się dla wszystkich węzłow, dla każdego z nich próbując ustawić stan zarówno "0", jak i "1". Jak widać, przy takiej procedurze testowania nie jest potrzebna znajomość funkcji wykonywanej przez układ. Trzeba natomiast znać jego schemat logiczny, od niego bowiem zależy, jakie wektory testowe należy podać na wejścia i jakie powinny być prawidłowe odpowiedzi układu na wyjściach.

Zilustrujmy to prostym przykładem. Niech testowanym układem będzie połączenie dwóch bramek pokazane na rys. 11.2.




Rys. 11.2. Testowany układ kombinacyjny

Przypuśćmy, że chcemy zbadać, czy w węźle D, który jest wewnętrznym węzłem układu, nie występuje uszkodzenie typu "stałe zero". W tym celu staramy się w weźle D ustawić "1" - musimy podać wektor testowy, w którym A="0" i B="0", a stan C jest nieistotny. Aby móc zaobserwować stan węzła D na wyjściu E, trzeba na wejście C podać "1". Innymi słowy, wektor "00x" (x - stan nieistotny) zapewnia kontrolowalność węzła D dla ustawienia w nim "1", zaś wektor "xx1" zapewnia obserwowalność stanu węzła D. Stąd test węzła D ze względu na uszkodzenie "stałe zero" wymaga podania wektora testowego "001". Jeśli w węźle D występuje uszkodzenie "stałe zero", na wyjściu pojawi się "1". W przypadku braku uszkodzenia na wyjściu pojawi się "0". Podobnie, dla testu węzła D ze względu na "stałą jedynkę" trzeba podać jeden z wektorów: "011" lub "101" lub "111".

Testowanie strukturalne ogromnie zmniejsza liczbę wektorów w porównaniu z wyczerpującym testem funkcjonalnym. Dla przetestowania strukturalnego omawianego układu wystarczają dwa wektory. W przypadku wyczerpującego testu funkcjonalnego potrzebne byłoby $2^3$=8 wektorów. W przypadku rzeczywistych układów oszczędność jest daleko większa, bo często jeden wektor testowy testuje skutecznie więcej niż jeden węzeł.

Na omówionej tu zasadzie działają generatory wektorów testowych - programy komputerowe określające optymalne sekwencje wektorów testowych dla układów o danym schemacie logicznym.

Testowanie strukturalne można stosować do układów cyfrowych realizowanych w różny sposób, nawet niekoniecznie w postaci układów scalonych. Istnieje jeszcze jeden sposób testowania układów cyfrowych, nadający się jednak tylko do testowania układów CMOS. Jest to **testowanie prądowe** (znane w literaturze anglojęzycznej jako "IDDQ testing"). Wykorzystuje się tu fakt, że prawidłowo skonstruowane i wykonane bramki CMOS pobierają prąd tylko w czasie przełączania. Gdy na wejściach panuje stan ustalony, bramki CMOS nie pobierają prądu (jeśli pominąć bardzo mały prąd podprogowy tranzystorów i prądy wsteczne złącz źródeł i drenów). Zdecydowana większość defektów, jakie mogą się pojawić w bramkach CMOS, prowadzi do tego, że istnieją takie kombinacje stanów na wejściach bramek, przy których bramki w stanie ustalonym pobierają prąd. Zatem, zamiast obserwować odpowiedzi układu na wyjściach, można badać pobór prądu. Po podaniu każdego wektora testowego należy odczekać, aż ustalą się stany na wyjściu, i dokonać pomiaru prądu pobieranego ze źródła zasilania. Jeśli prąd ten jest o kilka rzędów wielkości większy od prądu w układzie działającym prawidłowo, to możemy mieć pewność, że w układzie występuje co najmniej jeden defekt.

Testowanie prądowe ma wiele zalet. Pozwala ono szybko i łatwo wykryć wiele uszkodzeń, które przy testowaniu strukturalnym wymagałyby podania bardzo długich sekwencji wektorów testowych. Pomiar prądu może być wykonywany wewnątrz układu, przez wprowadzenie do układu **monitorów prądu** - dość prostych układów, które śledzą pobór prądu w czasie pracy układu i sygnalizują przekroczenie jego progowej wartości oznaczającej wystąpienie uszkodzenia. Można więc testować układ w czasie jego normalnej pracy, co nie jest możliwe w przypadku testowania strukturalnego. Możliwe jest nawet budowanie układów samonaprawiających się. W takim układzie bloki funkcjonalne są zdublowane. W razie wykrycia przez monitor prądu, że w którymś bloku wystąpiło uszkodzenie, możliwa jest automatyczna rekonfiguracja układu przez odłączenie bloku uszkodzonego i podłączenie na jego miejsce bloku zapasowego. Możliwość ta jest jednak w praktyce bardzo rzadko wykorzystywana, bowiem układy scalone i bez tego cechują się bardzo wysoką niezawodnością, a zdublowanie wszystkich bloków oznacza podwojenie kosztu układu.

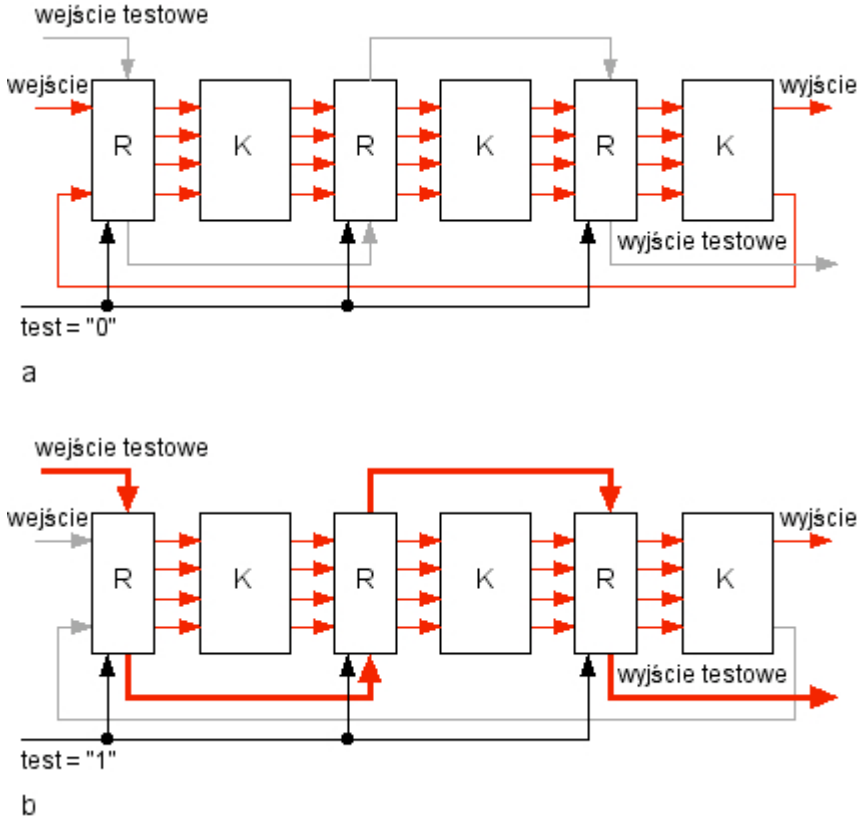
Testowanie prądowe i testowanie strukturalne nie zastępują się wzajemnie, lecz uzupełniają. Są bowiem takie uszkodzenia, których nie wykrywa testowanie prądowe, i takie, których wykrycie w testach strukturalnych wymaga bardzo długich sekwencji wektorów testowych.

Nawet równoczesne zastosowanie obu sposobów testowania nie rozwiązuje do końca problemów testowania układów bardzo dużych i złożonych. W ich przypadku dla uzyskania dostatecznie wysokiego poziomu wykrywalności uszkodzeń potrzebne byłyby sekwencje wektorów testowych o długościach niemożliwych do zaakceptowania. Dlatego rozwinęły się metody projektowania układów łatwo testowalnych i samotestujących się. Jest o nich mowa w następnej części wykładu.

## 11.4. Układy łatwo testowalne i samotestujące się

Zajmiemy się teraz odpowiedzią na pytanie: jak zaprojektować układ cyfrowy, by był on łatwo testowalny (tzn. aby można było uzyskać wysoki poziom wykrywalności uszkodzeń przy umiarkowanej długości sekwencji wektorów testowych)?

Projektowanie układów łatwo testowalnych polega na wyborze rozwiązań poprawiających obserwowalność i kontrolowalność wewnętrznych węzłów. Niekiedy można to osiągnąć przez wprowadzenie zmian do schematu logicznego oraz ewentualnie dodatkowych wejść i wyjść wykorzystywanych tylko przy testowaniu. Dodatkowe wejścia i wyjścia są jednak rozwiązaniem podnoszącym znacznie koszt układu, bowiem pola montażowe zajmują bardzo dużą powierzchnię. Uniwersalnym, skutecznym i powszechnie stosowanym sposobem jest wyposażenie układu w **łańcuch skanujący** (zwany także **ścieżką skanującą**). Łańcuch skanujący jest to utworzony z przerzutników rejestr szeregowy, który umożliwia przekształcenie układu sekwencyjnego w kombinacyjny na czas testowania. Wykorzystuje się tu przerzutniki istniejące w układzie, nie ma potrzeby dodawania nowych. Przerzutniki te mają jednak specjalną konstrukcję umożliwiającą wykorzystanie ich w łańcuchu skanującym.

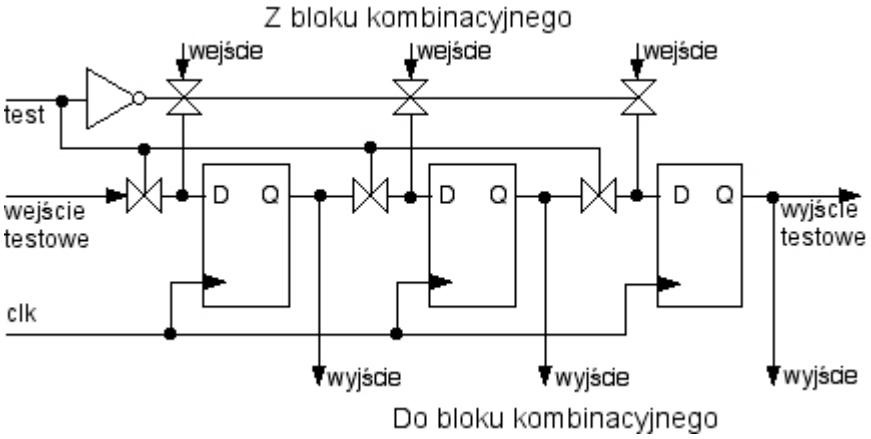



Rys. 11.3. Układ sekwencyjny z łańcuchem skanującym. (a) - praca w trybie normalnym, (b) - praca w trybie testowym.
K: bloki kombinacyjne, R: rejestry złożone z przerzutników.
Kolorem szarym zaznaczono połączenia nieaktywne w danym trybie pracy, pogrubiona czerwona linia: szeregowy łańcuch skanujący

Rys. 11.3 ilustruje ideę łańcucha skanującego. Układ składa się z bloków kombinacyjnych K i z rejestrów R, zaś w każdym rejestrze znajduje się pewna liczba przerzutników typu D. Układ z łańcuchem skanującym ma dwa tryby pracy: zwykły i testowy. W zwykłym trybie pracy każdy z przerzutników w rejestrach działa niezależnie od pozostałych. Układ wykonuje swe zwykłe funkcje. W trybie testowym wszystkie przerzutniki w rejestrach zostają połączone w jeden szeregowy rejestr przesuwający zwany łańcuchem skanującym, zaznaczony na rys. 11.3b pogrubioną czerwoną linią. Teraz możliwe jest szeregowe wpisanie z wejścia testowego do łańcucha skanującego wektora testowego, który z wyjść przerzutników podany będzie bezpośrednio na wewnętrzne węzły układu - wejścia bloków kombinacyjnych. Odpowiedzi są z wyjść wpisywane do przerzutników w łańcuchu skanującym, po czym można je szeregowo wyprowadzić na wyjście testowe i porównać z prawidłowymi. W czasie wyprowadzania odpowiedzi na poprzedni wektor testowy można równocześnie wpisywać następny wektor. Dzięki łańcuchowi skanującemu węzły wewnętrzne połączone z przerzutnikami stają się w trybie testowym łatwo kontrolowalne i obserwowalne.

Budowa przerzutników do łańcucha skanującego nie jest skomplikowana. Idea jest pokazana na rys. 11.4. Przy stanie "1" sygnału "test" następuje wpisywanie szeregowe wektora testowego. Po zakończeniu tego procesu stan

sygnału "test" zmienia się na "0" na czas tak długi, aby w blokach kombinacyjnych ustaliły się odpowiedzi na wyjściach. Odpowiedzi te zostają wpisane do przerzutników, po czym sygnał "test" ponownie otrzymuje wartość "1". Można teraz szeregowo wyprowadzić je na wyjście testowe.

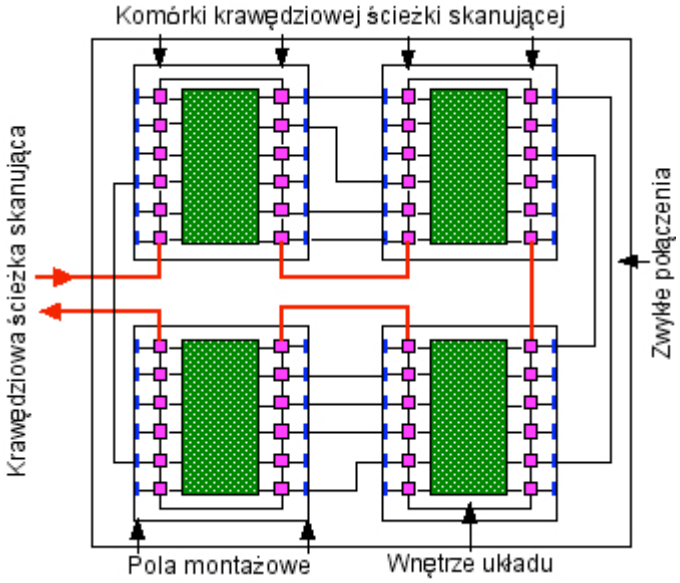



Rys. 11.4. Idea łączenia przerzutników w łańcuch skanujący przy zastosowaniu bramek transmisyjnych sterowanych sygnałem "test"

W bibliotekach komórek standardowych dostarczanych przez producentów układów z reguły znajdziemy dwa rodzaje przerzutnika D - zwykły oraz przystosowany do tworzenia łańcuchów skanujących.

W dużych układach pojedynczy łańcuch skanujący byłby bardzo długi, przez co podawanie kolejnych wektorów testowych w trybie wpisywania szeregowego trwałoby dużo czasu. Dlatego w dużych układach wprowadza się wiele niezależnych łańcuchów skanujących o umiarkowanej długości, które mogą działać w trybie testowania równocześnie.

Łańcuchy skanujące znalazły ważne zastosowanie także w testowaniu całych pakietów drukowanych, urządzeń i systemów. W przypadku wielowarstwowych płytek drukowanych wykonywanych nowoczesnymi technologiami mamy do czynienia z tym samym problemem, który występuje wewnątrz układów scalonych - brakiem dostępu do wewnętrznych węzłów. Co za tym idzie, testowanie pakietów drukowanych zawierających wiele układów scalonych jest równie trudne jak testowanie pojedynczych układów, a czasem nawet trudniejsze. Aby je ułatwić, wbudowuje się do wnętrza scalonych układów cyfrowych specjalne łańcuchy skanujące zwane brzegowymi lub krawędziowymi. Komórki tych łańcuchów znajdują się między wnętrzem układu, a polami montażowymi. W normalnym trybie pracy są one "przezroczyste" dla sygnałów. W trybie testowania są one łączone w łańcuch skanujący obejmujący wszystkie układy na pakiecie. Łańcuch taki pozwala wpisywać szeregowo wektory testowe, które po wpisaniu trafiają bezpośrednio na wejścia układów scalonych, i po otrzymaniu odpowiedzi wyprowadzić je w trybie szeregowym. Łańcuch tego typu bywa nazywany **krawędziową ścieżką skanującą** (z angielskiego "boundary scan path"). Tę ideę ilustruje rys. 11.5.



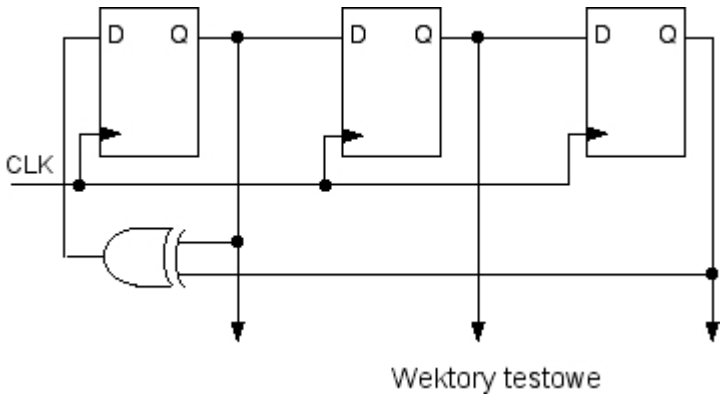

Rys. 11.5. Krawędziowa ścieżka skanująca dla czterech układów scalonych na pakiecie drukowanym

Krawędziowa ścieżka skanująca jest znormalizowana (norma IEEE 1149.1), dzięki czemu można swobodnie

łączyć ze sobą układy różnych producentów. Większość katalogowych cyfrowych układów scalonych jest obecnie wyposażona w taką ścieżkę. Szczegółów technicznych nie będziemy tutaj omawiać.

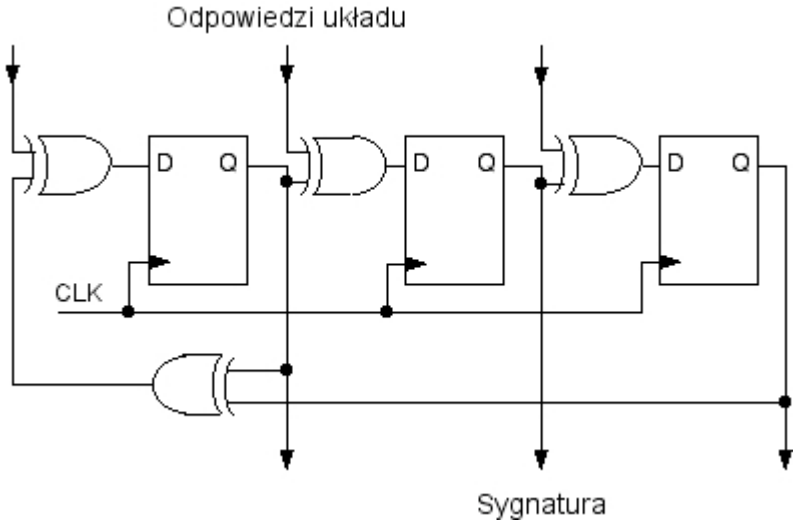
Testowanie staje się jeszcze prostsze, jeśli układ ma wbudowany mechanizm samotestowania. Samotestowanie można zrealizować wbudowując w układ generator wektorów testowych oraz układ analizujący prawidłowość odpowiedzi. Mogłoby się wydawać, że to jest do zrealizowania przez wprowadzenie do układu dwóch pamięci stałych: pamięci wektorów testowych i pamięci poprawnych odpowiedzi, oraz układu porównującego odpowiedzi otrzymane z prawidłowymi. Jest to jednak rozwiązanie niepraktyczne, bo potrzebne do tego pamięci z reguły zajmowałyby w układzie bardzo dużo miejsca. Stosowane zazwyczaj rozwiązanie polega na wprowadzeniu do układu generatora pseudolosowych sekwencji wektorów testowych oraz układu analizy odpowiedzi poddającego kolejne odpowiedzi kompresji prowadzącej do otrzymania t.zw. sygnatury - pojedynczego słowa binarnego, które ma jednoznacznie określoną wartość w przypadku układu sprawnego, natomiast przyjmuje inne wartości, jeśli w testowanym układzie jest uszkodzenie.

Jako generator wektorów testowych stosowany jest zwykle generator wektorów pseudolosowych zbudowany jako rejestr przesuwający ze sprzężeniem zwrotnym poprzez bramki XOR - rys. 11.6.




Rys. 11.6. Trzybitowy generator wektorów pseudolosowych

Taki układ generuje w każdym takcie zegara nowy wektor różny od poprzedniego (pod warunkiem, że stany początkowe przerzutników to nie są same zera). Ciąg tych wektorów ma cechy statystyczne ciągu pseudolosowego. Po wygenerowaniu $2^n$-1 wektorów ($n$ - liczba przerzutników) generowana sekwencja powtarza się. Jednak w testowaniu wystarcza wykorzystywanie tylko początkowej, znacznie krótszej sekwencji. W czasie testowania kolejne wektory podawane są na wejścia testowanego układu, a odpowiedzi na wejście układu tworzącego sygnaturę. Taki układ może być zbudowany podobnie, jak generator wektorów pseudolosowych.




Rys.11.7. Układ tworzący trzybitową sygnaturę

W każdym takcie zegara tworzone jest na wyjściu układu nowe słowo binarne, które zależy od wszystkich poprzednich oraz od ostatniej odpowiedzi układu. Po zakończeniu testowania końcowe słowo jest sygnaturą. Porównywana jest ona z zapamiętaną sygnaturą poprawną (tj. otrzymaną z układu bez uszkodzeń).

Nic nie stoi na przeszkodzie, aby w dużym układzie stosować wszystkie omówione sposoby testowania jednocześnie, dobierając metodę testowania do charakteru i stopnia złożoności testowanego bloku.

Jak widać, testowanie jest na tyle złożonym zagadnieniem, że wymaga brania pod uwagę od samego początku procesu projektowania układu.

## Bibliografia

[1] N. H. E. Weste, K. Eshraghian, "*Principles of CMOS VLSI design, a systems perspective*", 2nd edition, Addison-Wesley Publishing Company, 1993

# Wykład 12: Podstawy układów analogowych

## Wstęp

Wykład 12 wprowadza do problematyki analogowych układów scalonych. Projektowanie tych układów pod wieloma względami różni się od projektowania układów cyfrowych. Wymaga głębszego zrozumienia właściwości elementów i zbudowanych z nich układów, a także praktycznego doświadczenia. Projektowanie układów analogowych z trudem poddaje się formalizacji i jest bardzo trudne do zautomatyzowania. Wszystko to sprawia, że jest w nim więcej sztuki inżynierskiej, a mniej zwykłego rzemiosła.

W wykładzie 12 znajdziesz najpierw porównanie układów cyfrowych z analogowymi z punktu widzenia projektanta. Będzie także mowa o szczególnie ważnych w przypadku układów analogowych zagadnieniach: wpływie temperatury na parametry elementów, wpływie rozrzutów produkcyjnych, wpływie elementów i sprzężeń pasożytniczych. Zaczniesz także poznawać typowe bloki stosowane w układach analogowych. Dalsze przykłady szczegółowych rozwiązań układowych stosowanych w układach analogowych znajdziesz w następnym wykładzie.

## 12.1. Specyfika układów analogowych i problemy ich projektowania

Niezwykle szybki rozwój techniki cyfrowej i układów cyfrowych wywołał tendencję do zastępowania wszędzie gdzie to możliwe analogowej obróbki sygnałów obróbką cyfrową. Jest faktem, że funkcje wypełniane dotąd przez układy analogowe można równie dobrze, a wielu przypadkach lepiej zrealizować cyfrowo, po przekształceniu sygnału analogowego na jego reprezentację cyfrową. Przewaga cyfrowej sygnałów obróbki nad analogową jest w wielu zastosowaniach niewątpliwa. Główne zalety to:

- większe możliwości: w technice cyfrowej można realizować sposoby obróbki sygnałów bardzo trudne lub niemożliwe do realizacji analogowej,
- większa precyzja: technika cyfrowa zapewnia dokładność nieosiągalną w technice analogowej,
- większa elastyczność: programowane układy cyfrowego przetwarzania sygnałów mogą być wykorzystane do bardzo wielu zastosowań, jest też możliwość adaptacji sposobu obróbki sygnału lub parametrów tej obróbki do charakterystycznych cech przetwarzanego sygnału,
- pamięć: sygnał w postaci cyfrowej może być łatwo zapamiętany, nie ma zaś dobrych układów elektronicznych pamięci analogowych.

Mimo to układy analogowe nie znikły i nie znikną. Istnieje wiele zastosowań, w których układy analogowe nie mogą być zastąpione przez cyfrowe. Przykłady takich zastosowań to:

- układy nadawcze i odbiorcze do komunikacji bezprzewodowej,
- układy akwizycji, wzmacniania i kształtowania sygnału z wszelkiego rodzaju czujników i detektorów,
- układy wzmacniające w technice audio i wideo,
- sprzęt pomiarowy,
- układy do niecyfrowego przetwarzania informacji (np. sztuczne sieci neuronowe, układy logiki rozmytej).

Układami stojącymi na pograniczu technik analogowej i cyfrowej są przetworniki analogowo-cyfrowe i cyfrowo-analogowe. Ponadto w układach o czysto cyfrowych zastosowaniach mogą występować problemy projektowania analogowego. Najlepszym przykładem są pamięci dynamiczne RAM, w których wzmacniacze odczytu i niektóre inne bloki funkcjonalne są typowymi układami analogowymi.

Jedną ze specyficznych cech układów analogowych jest to, że chociaż są to zwykle układy małe, wręcz mikroskopijne w porównaniu z mikroprocesorami liczącymi miliony elementów, to ich projektowanie jest dość trudne i pracochłonne, choćby z tego powodu, że topografię układu analogowego w większości przypadków projektuje się w stylu *full custom*. Projektanci dużych układów scalonych typu "System on chip" zawierających zarówno bloki cyfrowe, jak i analogowe twierdzą, że część analogowa takiego systemu zajmuje około 10% - 20% powierzchni układu, ale nakład pracy na zaprojektowanie tej części sięga 80% - 90%.

Skutek uboczny przekonania o tym, że układy analogowe tracą znaczenie i zastosowania, jest taki, że bardzo mało uczelni technicznych kształci projektantów układów analogowych. Jest to dziś na świecie bardzo poszukiwana umiejętność.

Najważniejsze różnice między układami cyfrowymi, a analogowymi z punktu widzenia projektanta ujmuje tablica poniżej:

| Układy cyfrowe | Układy analogowe |
|---|---|
| Sygnał w postaci dyskretnej, przybierający w każdej chwili jedną z dwóch wartości (pomijając stany przejściowe) | Sygnał ciągły, może mieć dowolną wartość z pewnego przedziału |
| Mała wrażliwość na parametry i charakterystyki elementów | Duża wrażliwość na parametry i charakterystyki elementów |
| Mała wrażliwość na rozrzuty produkcyjne | Duża wrażliwość na rozrzuty produkcyjne |
| Zależności temperaturowe parametrów i charakterystyk elementów są mało istotne | Zależności temperaturowe parametrów i charakterystyk elementów są na ogół bardzo istotne |
| Mała wrażliwość na topografię układu | Duża wrażliwość na topografię układu |

| | |
|---|---|
| Elementami układu są tranzystory MOS (dotyczy układów cyfrowych CMOS) | Stosuje się znacznie więcej różnych rodzajów elementów: rezystancje, diody, tranzystory bipolarne |
| Dość niewielka liczba typowych komórek i bloków funkcjonalnych, z których można poskładać dowolny układ. | Wielka rozmaitość układów do różnych zastosowań, i tak najczęściej niewystarczających przy projektowaniu układu do nowego zastosowania |
| Wysoki stopień automatyzacji projektowania, możliwość zaprojektowania układu od poziomu opisu funkcjonalnego do jego topografii | Brak skutecznych narzędzi automatyzacji wielu etapów projektowania, schemat trzeba wymyślić (lub adaptować znane rozwiązania), topografię zaprojektować w stylu *full custom* |
| Dobrze rozwinięte i powszechnie stosowane sformalizowane języki opisu sprzętu (VHDL, Verilog) | Języki opisu układów analogowych są w dość wczesnym stadium rozwoju |
| Testowanie, choć czasochłonne, jest koncepcyjne proste i wymaga jedynie odróżniania poziomu napięcia dla "0" i "1" | Testowanie trudne, wymagające precyzyjnych analogowych pomiarów |

Problemami, które mają stosunkowo niewielkie znaczenie przy projektowaniu układów cyfrowych, natomiast projektantom układów analogowych sprawiają wiele kłopotów, są: zależność parametrów i charakterystyk elementów układu od temperatury oraz rozrzuty produkcyjne.

**Zależności temperaturowe**

W tranzystorach MOS zależne od temperatury są: napięcie progowe $U_T$ oraz ruchliwość nośników w kanale tranzystora μ. Zarówno napięcie progowe, jak i ruchliwość są malejącymi funkcjami temperatury. Napięcie progowe maleje w przybliżeniu liniowo, o 1 ... 3 mV/K. Ruchliwość maleje wg funkcji potęgowej: $\mu \div T^{-a}$. W typowych warunkach pracy tranzystora przeważa wpływ zmian ruchliwości, a to oznacza, że przy napięciach polaryzujących niezależnych od temperatury prąd drenu maleje z temperaturą.

W tranzystorach bipolarnych bardzo silnie (wykładniczo) wzrasta z temperaturą stała $J_{ES0}$ określająca przebieg charakterystyki $I_C$=f($U_{BE}$). W rezultacie przy niezależnych od temperatury napięciach polaryzujących prąd kolektora wykładniczo wzrasta z temperaturą. Jeśli zaś prąd kolektora w układzie jest wymuszony w taki sposób, że nie zależy od temperatury, to napięcie $U_{BE}$ maleje z temperaturą w przybliżeniu liniowo, o około 2 mV/K. Ten fakt jest często wykorzystywany w projektowaniu układów. Od temperatury zależy także współczynnik wzmocnienia prądowego $h_{FE}$. Jego wartość rośnie z temperaturą. Wzrost jest w pierwszym przybliżeniu liniowy, a szybkość wzrostu zależy od koncentracji domieszek w obszarach emitera i bazy tranzystora.

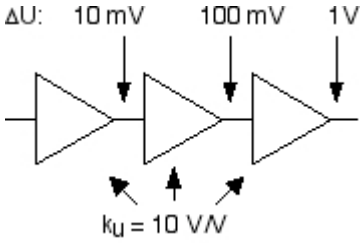
Silna zależność prądu kolektora od temperatury może być przyczyną wewnętrznej niestabilności tranzystora bipolarnego. Jeśli tranzystor polaryzowany jest napięciami niezależnymi od temperatury, to prąd kolektora rośnie z temperaturą, a wraz z prądem rośnie moc wydzielająca się w strukturze tranzystora. To powoduje wzrost temperatury (samopodgrzewanie się tranzystora), a wzrost temperatury powoduje dalszy wzrost prądu. Występuje tu więc zjawisko **niestabilności elektryczno-cieplnej:** dodatnie **elektryczno-cieplne sprzężenie zwrotne**. Może ono w skrajnym przypadku doprowadzić do nieograniczonego wzrostu prądu i temperatury, co oczywiście powoduje zniszczenie tranzystora. Ten efekt jest mało prawdopodobny w tranzystorach pracujących z małymi wartościami prądu i mocy, natomiast jest poważnym problemem w przypadku tranzystorów, w których przy pracy wydziela się duża moc, jak na przykład tranzystory w stopniu wyjściowym akustycznego wzmacniacza mocy.

Tranzystory MOS są stabilne cieplnie, ponieważ wzrost temperatury powoduje w nich spadek wartości prądu drenu, a wiec i wydzielanej mocy. Elektryczno-cieplne sprzężenie zwrotne jest więc w ich przypadku ujemne.

Od temperatury zależy także rezystancja rezystorów półprzewodnikowych. Polikrzemowe rezystory mają dodatni temperaturowy współczynnik rezystancji, jego wartość zależy od poziomu domieszkowania polikrzemu i jest rzędu 0,05%/K ... 0,2%/K (silniej domieszkowane obszary półprzewodnika mają niższe wartości tego współczynnika).

Zależności parametrów elementów od temperatury powodują, że w układach analogowych trzeba stosować specjalne układy stabilizujące punkty pracy (czyli składowe stałe napięć i prądów) tranzystorów. Układy te powinny zapewnić punkty pracy możliwie niezależne od temperatury. Trzeba dodać, że w układach scalonych punkty pracy wszystkich elementów są z reguły powiązane ze sobą. Powoduje to, że nawet bardzo nieznaczne zmiany prądów i napięć w jednym bloku układu mogą powodować katastrofalnie duże zmiany w innych blokach.

Przykładowo rozważmy trzystopiowy wzmacniacz, którego każdy stopień ma wzmocnienie napięciowe $k_u$ = 10. Niech na wyjściu pierwszego stopnia wzmacniacza wystąpi wywołana zmianą temperatury zmiana składowj stałej napięcia równa 10 mV. Ta niewielka zmiana może nie mieć żadnego niekorzystnego wpływu na działanie tego stopnia, ale jest ona wzmacniana i na wyjściu trzeciego stopnia wynosi już 1V!




Rys. 12.1. Ilustracja zjawiska zmian składowych stałych wywołanego zmianami temperatury

Ponadto występuje cieplne sprzężenie pomiędzy elementami układu. Zmiana napięcia o 1V na wyjściu trzeciego stopnia i związana z nią zmiana wielkości mocy wydzielanej w elementach tego stopnia zmienia temperaturę wszystkich stopni układu, w tym także pierwszego, występuje tu więc elektryczno-cieplne sprzężenie zwrotne.

! **Omówione wyżej efekty cieplne zmuszają do stosowania w analogowych układach scalonych takich rozwiązań układowych, które stabilizują napięcia i prądy w układzie czyniąc je mało zależnymi od temperatury.**

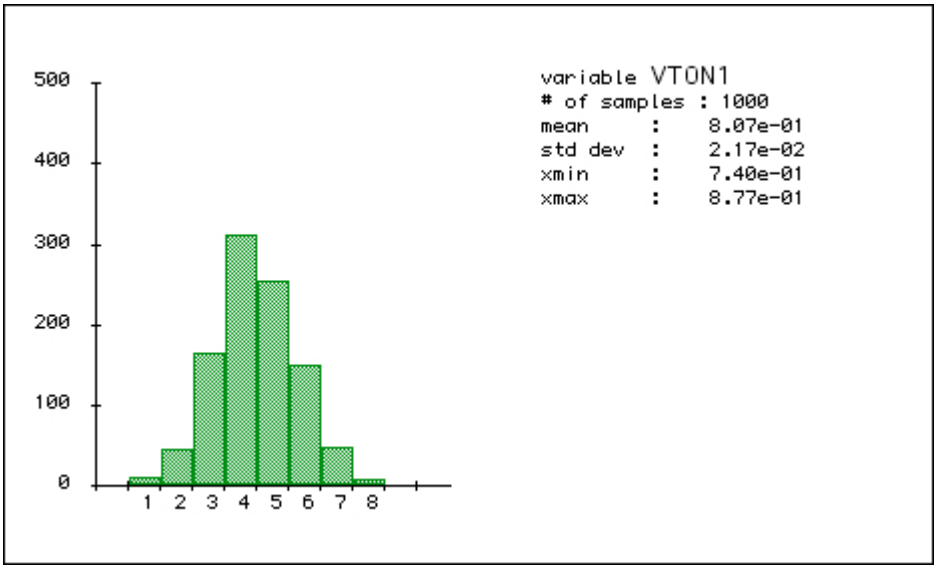
Stosowane są dwie klasy takich rozwiązań. Jedna z nich to stabilne temperaturowo **źródła prądowe** i **źródła napięciowe**. Źródła prądowe są to układy, które wymuszają przepływ w pewnej gałęzi układu prądu o określonym natężeniu. Źródła napięciowe są to układy, które wymuszają określoną różnicę potencjałow między dwoma węzłami układu. Zarówno źródła prądowe, jak i napięciowe mogą służyć m.in. do tego, by zapewnić w układzie stabilne wartości prądów i napięć. Druga klasa rozwiązań to stopnie i bloki robocze (tj. wykonujące operacje na sygnałach elektrycznych - wzmacniacze, filtry itp.), które są wewnętrznie odporne na zmiany charakterystyk i parametrów elementów. Osiąga się to zazwyczaj przez wykorzystanie układów symetrycznych, gdzie zmiany napięć i prądów w jednej części układu są kompensowane takimi samymi zmianami w drugiej, symetrycznej części. Stosowane są także inne sposoby kompensacji zmian napięć i prądów wywołanych zmianami temperatury. Łączne zastosowanie rozwiązań z obu klas daje w rezultacie układy, których odporność na efekty związane ze zmianami temperatury jest bardzo wysoka. Przykłady poznamy dalej.

**Rozrzuty produkcyjne**

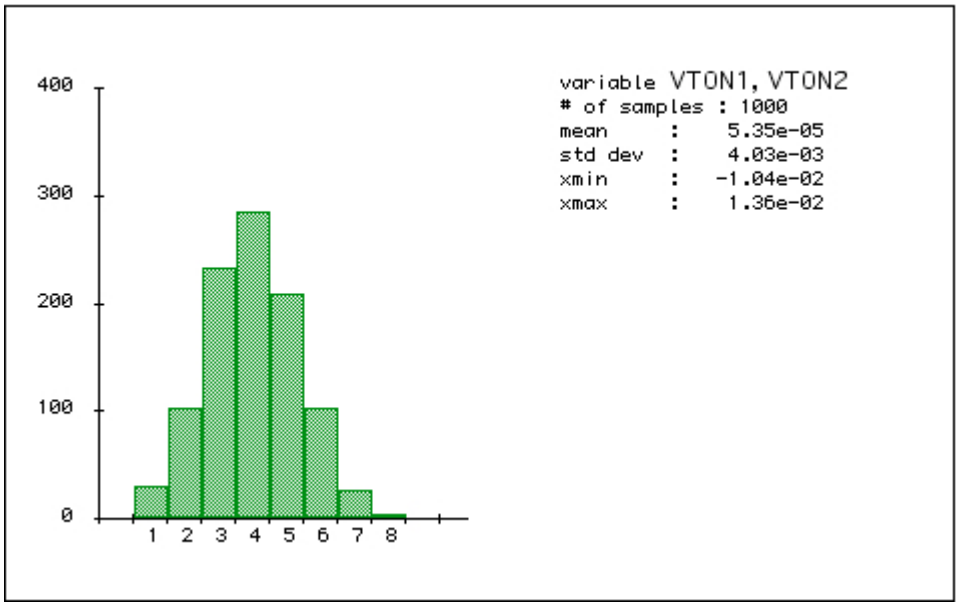
Omówimy teraz rozrzuty produkcyjne dla kilku rodzajów elementów. Ze względu na to, że układy scalone konstruuje się tak, by były niewrażliwe na rozrzuty globalne, skoncentrujemy się na rozrzutach lokalnych.

**Rozrzuty produkcyjne parametrów tranzystorów MOS**

Rozrzut prądu drenu tranzystora MOS wynika z rozrzutów następujących wielkości: napięcia progowego $U_T$, ruchliwości nośników w kanale μ, jednostkowej pojemności tlenku bramkowego $C_{ox}$ oraz wymiarów kanału $W$ i $L$ (patrz wykład 4, wzór 4.4). Wszystkie te rozrzuty mają składową globalną i lokalną. Przykładowo, rys. 12.2 pokazuje rozrzut całkowity (sumę rozrzutu globalnego i lokalnego) napięcia progowego tranzystora nMOS w pewnej technologii CMOS, zaś rys. 12.3 pokazuje histogram rozrzutu lokalnego, czyli różnicy napięć progowych dwóch identycznych tranzystorów znajdujących się obok siebie w tym samym układzie.



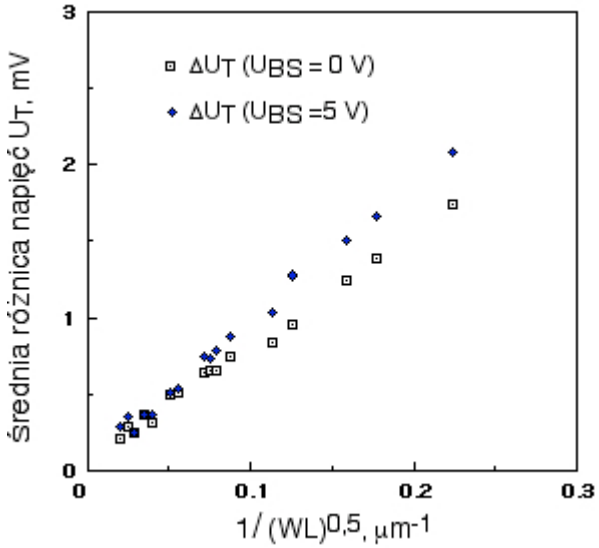

Rys. 12.2. Przykładowy histogram rozrzutu wartości napięcia progowego tranzystorów nMOS dla próbki 1000 tranzystorów z różnych płytek i partii produkcyjnych.




Rys. 12.3. Przykładowy histogram rozrzutu lokalnego wartości napięcia progowego tranzystorów nMOS dla próbki 1000 par tranzystorów. Histogram pokazuje rozrzut różnicy napięć progowych pomiędzy dwoma tranzystorami pary. Wartość średnia tej różnicy jest bardzo bliska zeru (0,0535 mV), miarą wielkości rozrzutu jest odchylenie standardowe, które wynosi około 4 mV.

Wszystkie rozrzuty lokalne w tranzystorze MOS wygodnie jest scharakteryzować odnosząc je do **napięcia niezrównoważenia**. Jest to różnica napięć $U_{GS}$, jakie trzeba przyłożyć do bramek dwóch tranzystorów, aby przy jednakowych wartościach napięcia $U_{DS}$ otrzymać jednakowe wartości prądu drenu $I_D$. Dwa najważniejsze, niezależne od siebie czynniki określające napięcie niezrównoważenia to lokalny rozrzut napięcia progowego (przykładowo pokazany na rys. 12.3) oraz rozrzut wymiarów kanału, a zwłaszcza jego długości $L$.

Rozrzut lokalny napięcia progowego jest odwrotnie proporcjonalny do pierwiastka z powierzchni kanału tranzystora - im większa powierzchnia, tym mniejszy rozrzut. Przykład tej zależności pokazuje rys. 12.4. Zależność ta nosi w mikroelektronice nazwę "prawa Pelgroma".




Rys. 12.4. Przykład zależności lokalnego rozrzutu napięcia progowego od powierzchni bramki tranzystora MOS

Dla oszacowania wpływu rozrzutu wymiarów tranzystora na napięcie niezrównoważenia założymy, że dwa pod każdym innym względem identyczne (w szczególności mające identyczne napięcia progowe) tranzystory MOS różnią się długością kanału o wartość $\Delta L$. Dla uzyskania jednakowych prądów drenu trzeba wtedy spolaryzować bramki tranzystorów napięciami $U_{GS}$ różniącymi się o napięcie niezrównoważenia $\Delta U_{GS}$. Wartość tę można wyznaczyć ze wzoru 4.4.

$$\left|\Delta U_{GS}\right| = \frac{U_{GS} - U_T}{2}\left|\frac{\Delta L}{L}\right| \tag{12.1}$$

Podobną zależność można wyznaczyć dla różnicy szerokości kanałów $\Delta W$:

$$\left|\Delta U_{GS}\right| = \frac{U_{GS} - U_T}{2}\left|\frac{\Delta W}{W}\right| \tag{12.2}$$

Wzory (12.1) i (12.2) pokazują, że dla uzyskania małego napięcia niezrównoważenia (rzędu pojedynczych miliwoltów) fotolitografia musi być niezwykle precyzyjna. Dla uzyskania $\Delta U_{GS}$ < 1 mV, przy $U_{GS}$-$U_T$ = 2 V, otrzymujemy warunek $\Delta W / W$ < 0,001. Warunek ten jest możliwy do spełnienia tylko wtedy, gdy długość kanału $L$ jest wielokrotnie większa od minimalnej długości dopuszczalnej w danej technologii.

"Prawo Pelgroma" (rys.12.4) oraz wzory (12.1) i (12.2) pokazują, że:

! **Dla uzyskania małej wartości napięcia niezrównoważenia pary tranzystorów MOS powierzchnie ich kanałów muszą być dostatecznie duże, jak również każdy z wymiarów $W$, $L$ z osobna musi być dostatecznie duży.**

Z tego powodu tranzystory w układach analogowych mają z reguły zarówno długość, jak i szerokość kanału znacznie większą, niż minimalna dopuszczalna w danej technologii.

Niezerowe napięcie niezrównoważenia powstaje także wtedy, gdy tranzystory są pod każdym względem identyczne, ale różnią się ich temperatury. Ponieważ napięcie progowe tranzystora MOS zmienia się o około 1 ... 3 mV/K, różnica temperatur wynosząca 1K powoduje powstanie napięcia niezrównoważenia o takiej właśnie wartości.

**Rozrzuty produkcyjne parametrów tranzystorów bipolarnych**

Podobne oszacowanie zrobimy dla tranzystora bipolarnego. Napięcie niezrównoważenia zdefiniujemy jako różnicę napięć $U_{BE}$, jakie trzeba przyłożyć do złącz emiter-baza dwóch tranzystorów, aby przy jednakowych wartościach napięcia otrzymać jednakowe wartości prądu kolektora $U_{CE}I_C$. Użyjemy tu wzoru (4.22), w którym dla uproszczenia pominiemy rozrzut parametru $J_{ES0}$. Pozostaje wówczas do rozważenia rozrzut powierzchni złącza emiter-baza $A_E$. Ze wzoru (4.22) wynika, że jeśli dwa pod każdym innym względem identyczne tranzystory różnią się powierzchnią złącza emiter-baza o $\Delta A_E << A_E$, to dla otrzymania jednakowych wartości prądu kolektora potrzebna jest różnica napięć $U_{BE}$ wynosząca

$$\Delta U_{BE} = \frac{kT}{q}\frac{\Delta A_E}{A_E} \qquad (12.3)$$

Jest to warunek znacznie łagodniejszy, niż (12.1) i (12.2). Ponieważ w temperaturze otoczenia wartość $kT/q$ wynosi około 26 mV, to dla uzyskania $\Delta U_{BE}$ < 1 mV otrzymujemy warunek $\Delta A_E / A_E$ < 0,038. Widać, że dla uzyskania małego napięcia niezrównoważenia wymagania dla dokładności fotolitografii są znacznie mniejsze dla tranzystorów bipolarnych, niż dla tranzystorów MOS. Porównywaliśmy tu tylko rozrzuty wynikające z niedokładności fotolitografii, ale słuszne jest stwierdzenie ogólniejsze:

**! Uzyskanie małych rozrzutów lokalnych parametrów tranzystorów jest znacznie łatwiejsze dla tranzystorów bipolarnych, niż dla tranzystorów MOS.**

Podobnie jak dla tranzystorów MOS, dla tranzystorów bipolarnych niezerowe napięcie niezrównoważenia powstaje także wtedy, gdy tranzystory są pod każdym względem identyczne, ale różnią się ich temperatury. Ponieważ napięcie $U_{BE}$ przy stałej wartości prądu kolektora zmienia się o około 2 mV/K, różnica temperatur wynosząca 1K powoduje powstanie napięcia niezrównoważenia o takiej właśnie wartości.

**Rozrzuty produkcyjne rezystancji rezystorów**

Ponieważ w naszym wykładzie koncentrujemy się na technologii CMOS, omówimy tylko rezystory wykonywane jako ścieżki polikrzemowe. Rozrzut globalny rezystancji takich rezystorów jest bardzo duży, może sięgać ±25% ... ±40%. Możliwe jest natomiast osiągnięcie rozrzutu lokalnego, czyli różnicy rezystancji dwóch identycznych, położonych tuż obok siebie rezystorów, na poziomie zaledwie 0,1% ... 0,2% ich wartości nominalnej. Wynika z tego, że:

**! Nie należy projektować układów w taki sposób, by jakiś ich istotny parametr zależał od bezwzględnej wartości rezystancji pojedynczego rezystora. Stosuje się rozwiązania układowe, w których parametry układu zależą od stosunków rezystancji, a nie od ich bezwzględnych wartości.**

Rozrzut lokalny rezystancji rezystorów zależy, podobnie jak dla tranzystorów MOS, od całkowitej powierzchni rezystora (im większa, tym lepiej) i od wymiarów $L$ i $W$ (im większe, tym lepiej). Odpowiednie dane statystyczne podają producenci układów.

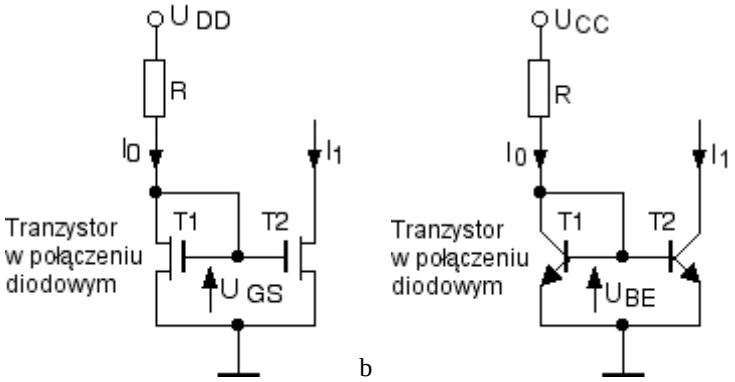
**Rozrzuty produkcyjne pojemności kondensatorów**

W tych rodzajach układów, które będziemy dalej omawiać, w większości przypadków kondensatory nie występują, a jeśli są, to rozrzuty ich pojemności nie są krytyczne.

## 12.2. Źródła prądowe

**Źródłem prądowym** nazywamy układ wymuszający przepływ prądu o zadanym natężeniu w pewnej gałęzi układu. Elementarnym źródłem prądowym jest po prostu pojedynczy tranzystor MOS lub bipolarny. Oba rodzaje tranzystorów mają taki zakres charakterystyk prądowo-napięciowych (były one omawiane w wykładzie 4), w których prąd w obwodzie wyjściowym (drenu $I_D$ lub kolektora $I_C$ - zobacz w dodatku 1) bardzo słabo zależy od napięcia na wyjściu elementu (tj. napięcia dren-źródło $U_{DS}$ lub kolektor-emiter $U_{CE}$). Trzeba więc spolaryzować tranzystor MOS lub bipolarny w taki sposób, by pracował w tym właśnie zakresie charakterystyk. Napięcie w obwodzie wejściowym ($U_{GS}$ lub $U_{BE}$) określa prąd, jaki płynie w obwodzie wyjściowym (patrz wzory: (4.4) dla tranzystora MOS, (4.20) dla tranzystora bipolarnego).

Nie jest jednak obojętne, w jaki sposób polaryzowany jest obwód wejściowy. Gdyby napięcie $U_{GS}$ lub $U_{BE}$ miało stałą, niezależną od temperatury wartość, to w przypadku tranzystora MOS prąd malałby z temperaturą, a w przypadku tranzystora bipolarnego wzrastałby, i to bardzo szybko. To nie jest dopuszczalne w większości zastosowań. Prąd mało zmieniający się z temperaturą można uzyskać, jeśli napięcie polaryzujące $U_{GS}$ lub $U_{BE}$ otrzymuje się jako spadek napięcia na tranzystorze MOS lub bipolarnym *w połączeniu diodowym* - rys. 12.5.




Rys. 12.5. Podstawowe układy źródeł prądowych: (a) MOS, (b) bipolarnego

Zasada działania obu źródeł jest taka sama i opiera się na spostrzeżeniu, że jeśli napięcie $U_{GS}$ lub $U_{BE}$ dla pary identycznych tranzystorów jest takie samo, to takie same muszą być wartości prądów drenów lub kolektorów (oczywiście pod warunkiem, że tranzystory pracują w zakresach napięć, w których słuszne są wzory (4.4) lub (4.20)). Układ z tranzystorami MOS spełnia ten warunek, jeśli napięcie $U_{DS}$ obu tranzystorów jest większe od napięcia nasycenia $U_{DSsat}$. Układ z tranzystorami bipolarnymi działa nawet gdy $U_{CE}$ jest bardzo bliskie zeru.

A zatem w obu układach mamy: $I_1 = I_0$. Chcąc określić wartość $I_1$ musimy określić $I_0$. Prąd ten wynika z odpowiedniego równania

$$U_{GS} + I_0 R = U_{DD} \tag{12.1a}$$

$$U_{BE} + I_0 R = U_{CC} \tag{12.1b}$$

skąd

$$I_0 = \frac{U_{DD} - U_{GS}}{R} \tag{12.2a}$$

$$I_0 = \frac{U_{CC} - U_{BE}}{R} \tag{12.2b}$$

Wartości $U_{GS}$ lub $U_{BE}$ można wyznaczyć z równań opisujących charakterystyki tranzystorów: (4.4) lub (4.20). Z (4.4) mamy

$$U_{GS} = U_T + \sqrt{2I_D \frac{1}{\mu C_{ox}} \frac{L}{W}} \tag{12.3}$$

zaś z (4.20) wynika wzór (4.22), który tu dla wygody powtórzymy

$$U_{BE} = \frac{kT}{q} \ln\left(\frac{I_C}{I_{ES0}}\right) = \frac{kT}{q} \ln\left(\frac{I_C}{J_{ES0} A_E}\right) \tag{12.4}$$

Po podstawieniu (12.3) do (12.2a) otrzymamy równanie kwadratowe ze względu na $I_0$, a po podstawieniu (12.4) do (12.2b) otrzymamy równanie uwikłane. Jednak ścisłe rozwiązania nie są nam tutaj potrzebne. Istotne jest, że w przypadkach obu rodzajów źródeł spełniony jest zwykle warunek: $U_{GS} << U_{DD}$ lub $U_{BE} << U_{CC}$. Wynika to z faktu, że w (12.3) zwykle drugi składnik jest mały wobec $U_T$, a $U_T$ jest poniżej 1 V, zaś z (12.4) można obliczyć, że przy typowych wartościach $I_C$ i $I_{ES0}$ oraz w temperaturze otoczenia $U_{BE}$<1V. Typowa wartość $U_{BE}$ dla krzemowych tranzystorów bipolarnych wynosi ok. 0,7V i słabo (logarytmicznie) zależy od $I_C$. Widzimy więc, że prąd $I_0$ w obu przypadkach dla dostatecznie dużych napięć zasilania $U_{DD}$ lub $U_{CC}$ uzależniony jest głównie od ilorazu $U_{DD}/R$ lub $U_{CC}/R$. Gdyby rezystancja R miała wartość niezależną od temperatury, mielibyśmy prąd także praktycznie niezależny od temperatury. Rezystancje w układach scalonych rosną ze wzrostem temperatury, ale zauważmy, że zarówno $U_T$, jak i $U_{BE}$ maleją ze wzrostem temperatury, czyli we wzorach (12.2a) i (12.2b) mamy do czynienia z ułamkami, w których zarówno liczniki, jak i mianowniki mają wartości rosnące z temperaturą (zakładamy tu oczywiście, że napięcia zasilania od temperatury nie zależą). Występuje więc w mniejszym lub większym stopniu kompensacja zmian temperaturowych. Stabilność temperaturowa prądów wymuszanych przez źródła prądowe wg rys. 12.1 jest całkowicie wystarczająca w większości zastosowań.

W przytoczonych wyżej rozumowaniach obliczaliśmy prąd $I_0$ i zakładaliśmy, że prąd $I_1$ jest mu dokładnie równy. Tak jednak w rzeczywistości nie jest. Prądy $I_0$ i $I_1$ nie są dokładnie równe. Przyczyny różnic:

- napięcia $U_{DS}$ lub $U_{CE}$ obu tranzystorów źródła nie są równe, a prąd, choć słabo, to jednak zależy od tych napięć,
- tranzystory nie są dokładnie takie same (rozrzut produkcyjny),
- tranzystory nie znajdują się w identycznej temperaturze.

W przypadku tranzystora bipolarnego jest jeszcze jedna przyczyna różnicy - prądy baz tranzystorów.

W większości przypadków identyczność prądów $I_0$ i $I_1$ nie ma większego znaczenia. Ważne jest tylko to, że prąd $I_1$ ma określoną, zadaną wartość i mało zależy od temperatury. Są jednak zastosowania, w których prąd $I_1$ powinien powtarzać prąd $I_0$. Mówimy wtedy, że układ pełni rolę **zwierciadła prądowego**. Jeżeli identyczność prądów $I_0$ i $I_1$ jest istotna, to można do niej dążyć przez:

(1) Zastosowanie tranzystorów dużych i z długim kanałem - znacznie dłuższym od minimalnej długości dopuszczalnej w danej technologii.

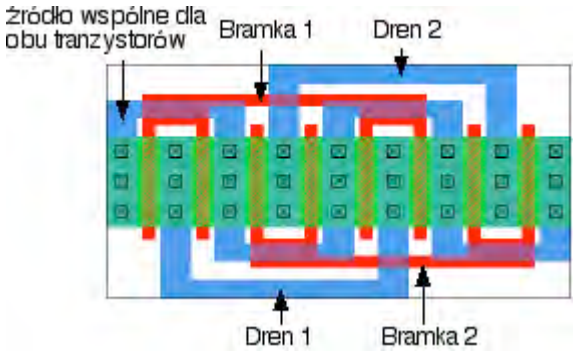
Im dłuższy kanał tranzystora, tym słabszy wpływ napięcia $U_{DS}$ na prąd drenu. Duże wymiary i długi kanał minimalizują także rozrzuty produkcyjne (patrz poprzedni wykład).

(2) Zastosowanie topografii minimalizującej wpływ rozrzutów lokalnych. Reguły są następujące:

- Oba trazystory powinny mieć dokładnie te same wymiary kanałów oraz obszarów źródła i drenu.
- Oba tranzystory powinny mieć tę samą orientację.
- W obu tranzystorach kierunek przepływu prądu powinien być ten sam.
- Tranzystory powinny być umieszczone w możliwie najmniejszej odległości jeden od drugiego.

(3) Umieszczenie tranzystorów w sposób symetryczny względem źródeł ciepła w układzie, aby miały możliwie jednakową temperaturę.

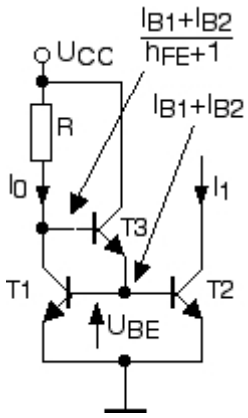
Przykład topografii dwóch tranzystorów MOS spełniającej podane wyżej kryteria pokazuje rys. 12.6. Kanał każdego tranzystora został podzielony na 4 równolegle połączone kanały. Kanały tranzystorów 1 i 2 są wzajemnie przeplecione. W każdym z tranzystorów jest taka sama liczba kanałow, w których prąd płynie z lewej do prawej, i kanałów, w których prąd płynie z prawej do lewej.




Rys. 12.6. Para tranzystorów nMOS - przykład topografii minimalizującej rozrzuty lokalne

Topografie tego rodzaju stosuje się zawsze tam, gdzie niezbędna jest minimalizacja rozrzutów lokalnych, nie tylko w przypadku źródeł prądowych.

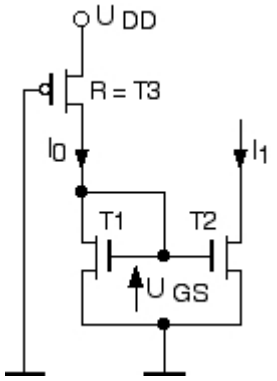
W przypadku tranzystorów bipolarnych wystarcza zachowanie identycznych kształtów i wymiarów tranzystorów. Pojawia się natomiast problem prądów baz. Prąd $I_0$ nie jest równy prądowi kolektora tranzystora, lecz sumie prądu kolektora i dwóch prądów baz. Wprowadza to dodatkową różnicę między prądami $I_0$ i $I_1$. Prąd bazy tranzystora bipolarnego jest $h_{FE}$ razy mniejszy od prądu kolektora. Dla tranzystorów o dużych wartościach $h_{FE}$ (100 ... 200 i więcej) prądy baz można pominąć, ale w układach scalonych można spotkać także tranzystory o wartościach $h_{FE}$ rzędu 10, a nawet mniejszych. Może tak być na przykład w przypadku użycia pasożytniczych tranzystorów bipolarnych w strukturach układów CMOS jako elementów aktywnych. Stosuje się wtedy podstawowe źródło prądowe w wersji wzbogaconej o dodatkowy tranzystor, którego rolą jest dostarczenie prądów baz bezpośrednio ze źródła zasilania - rys. 12.7.




Rys. 12.7. Źródło prądowe ze zredukowanym wpływem prądów baz

Dodatkowy tranzystor redukuje prąd odgałęziający się od prądu $I_0$ $h_{FE}$+1 - krotnie.

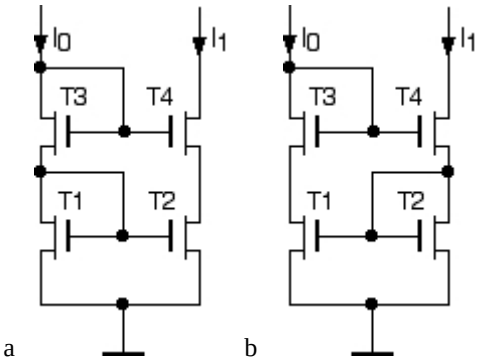
W układach CMOS na ogół rezystor R *nie* jest wykonywany jako zwykły rezystor polikrzemowy. Typowe wartości prądów drenu w analogowych układach CMOS są na poziomie od kilkudziesięciu do kilkuset µA. Przy napięciach zasilania wynoszących kilka V rezystor R musiałby mieć rezystancję rzędu kilkudziesięciu do kilkuset kΩ. Wykonanie takiego rezystora nie ma ekonomicznego sensu ze względu na powierzchnię, jaką musiałby on zająć. Zamiast rezystora stosuje się zwykle odpowiednio spolaryzowany tranzystor MOS o tak dobranych wymiarach, aby płynął przezeń prąd o wymaganym natężeniu. Przykład pokazuje rys. 12.8. Powierzchnia takiego tranzystora jest wielokrotnie mniejsza, niż rezystora wykonanego jako ścieżka polikrzemowa.




Rys. 12.8. Źródło prądowe, w którym rolę rezystancji R pełni tranzystor pMOS

Źródła prądowe są tak powszechnie stosowane w układach analogowych, że warto poznać kilka ich wariantów i odmian mających różne pożyteczne cechy.

W wielu zastosowaniach źródła prądowe powinny wykazywać możliwie jak największą małosygnałową rezystancję wyjściową, tj. zmiany prądu $I_1$ wywołane przez zmiany napięcia na tranzystorze źródła T2 powinny być jak najmniejsze. Dla najprostszego źródła rezystancja wyjściowa $r_{wy}$ jest równa $1/g_{ds}$. Sposobem na powiększenie tej rezystancji jest dodanie w szereg z tranzystorem T2 drugiego tranzystora. Oto dwie wersje źródła o zwiększonej rezystancji wyjściowej: źródło zwane kaskodowym i bardzo podobny układ zwany źródłem Wilsona:




Rys. 12.9. Źródła prądowe o podwyższonej rezystancji wyjściowej: (a) źródło kaskodowe, (b) źródło Wilsona

Zasada działania obu źródeł jest taka sama, jak źródła podstawowego. Dodatkowe tranzystory podnoszą jednak rezystancję wyjściową. Obliczyć ją można zastępując tranzystory ich małosygnałowymi schematami zastępczymi. Dla układu z rys. 12.9a otrzymuje się

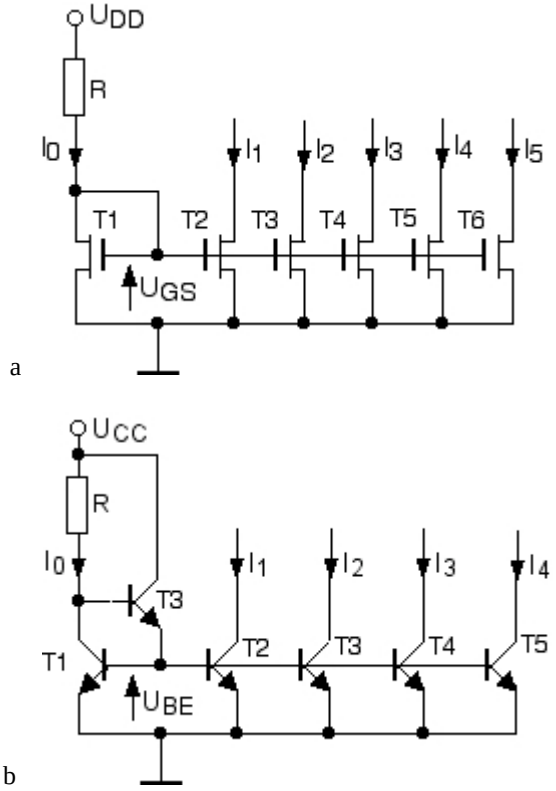
$$r_{wy} = \frac{g_{m4}}{g_{ds2} g_{ds4}} \tag{12.5}$$

zaś dla układu z rys. 12.9b wynik jest nieco bardziej skomplikowany

$$r_{wy} \cong \frac{g_{m1} g_{m4}}{g_{m2} g_{ds4}} r_{in} \tag{12.6}$$

gdzie symbolem $r_{in}$ oznaczono wypadkową rezystancję równoległego połączenia rezystancji $r_{ds1}=1/g_{ds1}$ oraz rezystancji R (nie pokazanej na rys. 12.9), przez którą dostarczany jest prąd $I_0$.

W bardziej złożonych układach występuje wiele źródeł prądowych zasilających różne gałęzie układu prądami o różnym natężeniu. Bardzo pospolicie stosowanym rozwiązaniem jest wówczas użycie jednego tranzystora T1 w połączeniu diodowym, który wytwarza napięcie polaryzujące $U_{GS}$ lub $U_{BE}$ dla wielu źródeł prądowych. Tranzystory tych źródeł (odpowiedniki tranzystora T2 w źródle podstawowym) mają różne szerokości kanałów lub różne powierzchnie złącz emiter-baza i dzięki temu dostarczają prądy o różnym natężeniu. Ilustruje to rys. 12.10.




Rys. 12.10. Zespoły źródeł prądowych: (a) MOS, (b) bipolarnych

W przypadku tranzystorów MOS, zakładając jednakowe długości wszystkich kanałów T1 ... T6, można napisać:

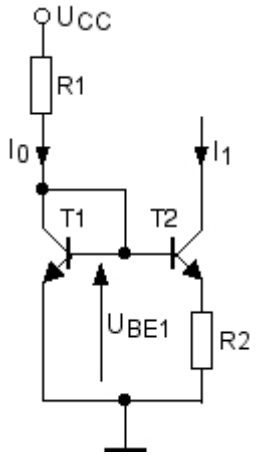
$$\frac{I_0}{W_{T1}} = \frac{I_1}{W_{T2}} = \frac{I_2}{W_{T3}} = \frac{I_3}{W_{T4}} = \frac{I_4}{W_{T5}} = \frac{I_5}{W_{T6}} \tag{12.7}$$

i podobnie w przypadku tranzystorów bipolarnych

$$\frac{I_0}{A_{ET1}} = \frac{I_1}{A_{ET2}} = \frac{I_2}{A_{ET3}} = \frac{I_3}{A_{ET4}} = \frac{I_4}{A_{ET5}} \tag{12.8}$$

W przypadku źródeł z tranzystorami bipolarnymi zastosowany jest dodatkowy tranzystor T3 redukujący wpływ sumy wszystkich prądów baz na prąd $I_0$.

Rys.12.11 pokazuje zmodyfikowany schemat źródła prądowego przydatny wtedy, gdy prąd źródła $I_1$ powinien być bardzo mały. Użycie źródeł podstawowych (jak na rys. 12.5) wymagałoby w takim przypadku zastosowania bardzo dużej rezystancji R, co jest nieekonomiczne lub nawet technicznie niemożliwe.




Rys. 12.11. Źródło prądowe dla bardzo małych prądów

W źródle pokazanym na rys. 12.11 dodatkowy rezystor R2 wprowadza lokalne ujemne sprzężenie zwrotne. Prąd

$I_1$ przepływając przez ten rezystor wywołuje spadek napięcia, który odejmuje się od napięcia $U_{BE1}$. W rezultacie napięcie emiter-baza tranzystora T2 jest mniejsze o wartość $I_1R2$ od $U_{BE1}$, a prąd $I_1$ mniejszy od $I_0$. Pozwala to uzyskać mały prąd $I_1$ przy prądzie $I_0$ na tyle dużym, że rezystor R1 ma rozsądnie małą rezystancję. Wykorzystując zależność (4.22) można wyznaczyć różnicę napięć $U_{BE}$ tranzystorów T1 i T2:

$$U_{BE1} - U_{BE2} = \frac{kT}{q}\ln\left(\frac{I_0}{I_1}\right) = I_1 R_2 \tag{12.9}$$

skąd wynika zależność

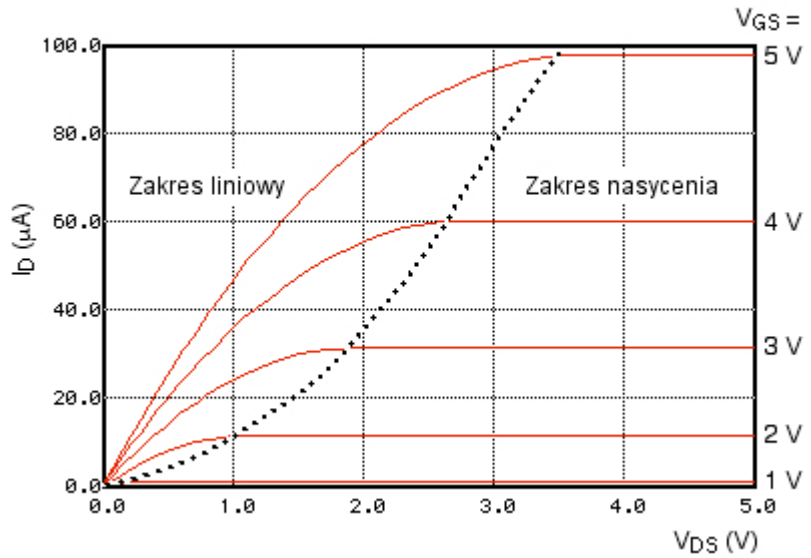
$$I_1 = \frac{kT/q}{R_2}\ln\left(\frac{I_0}{I_1}\right) \tag{12.10}$$

Zależność ta, choć w postaci uwikłanej ze względu na $I_1$, wystarcza do pokazania skutku wprowadzenia rezystora R2. Nic nie stoi na przeszkodzie, by stosunek $I_0/I_1$ wynosił na przykład 100. Otrzymujemy wówczas bardzo mały prąd $I_1$ (na przykład 10 μA) przy dużym prądzie $I_0$ (na przykład 1 mA). Rezystor R1 może więc mieć małą, możliwą do przyjęcia rezystancję. Równocześnie rezystor R2 też nie musi mieć dużej rezystancji, bo napięcie $kT/q$ mnożone przez logarytm stosunku prądów $I_0/I_1$ wynosi kilkadziesiąt do stu kilkudziesięciu mV.

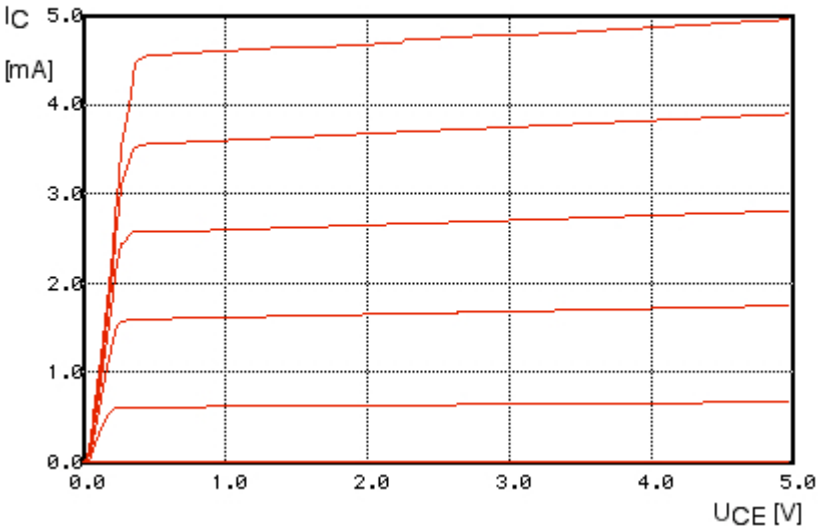
Tę samą ideę budowy źródła dla małych prądów można by zastosować także w przypadku źródeł z tranzystorami MOS. Zależności ilościowe są oczywiście inne. Jednak w przypadku źródeł z tranzystorami MOS nietrudno uzyskać małe wartości prądu korzystając z układu z rys. 12.8 i odpowiednio dobierając wymiary kanału tranzystora T3.

Przy okazji zwróćmy uwagę, że omawiane źródło umożliwia uzyskanie dobrej stabilności temperaturowej prądu $I_1$. Napięcie $kT/q$ w temperaturze otoczenia rośnie o 0,33%/K. Jeśli rezystor R2 ma ten sam temperaturowy współczynnik zmian rezystancji (a jest to wartość łatwa do uzyskania dla rezystorów półprzewodnikowych), to prąd $I_1$ w pierwszym przybliżeniu nie będzie zależał od temperatury (zależność logarytmu prądów od temperatury można zaniedbać).

## 12.2. Dodatek 1: Przykładowe charakterystyki tranzystorów MOS i bipolarnego




Charakterystyki wyjściowe tranzystora MOS




Charakterystyki wyjściowe tranzystora bipolarnego

## 12.3. Źródła napięciowe

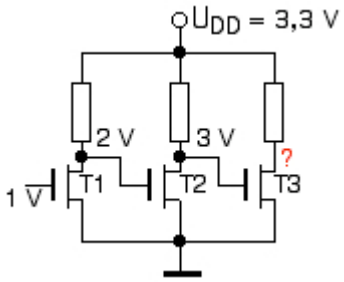
**Źródło napięciowe** wymusza określoną różnicę potencjałów między dwoma węzłami układu. Można wyróżnić trzy typy układów źródeł napięciowych różniące się zastosowaniem i wymaganiami:

- źródła napięć zasilania,
- źródła napięć odniesienia,
- układy przesuwania poziomu składowej stałej.

**Źródła napięcia zasilania** służą do wytworzenia napięcia zasilania o zadanej, stałej wartości napięcia. Mogą to być samodzielne układy scalone, ale istnieją również układy źródeł napięcia zasilania wbudowywane do wnętrza układów scalonych. W tym ostatnim przypadku chodzi zwykle o to, by wewnętrzne bloki układu były zasilane innym napięciem, niż to, które doprowadza z zewnątrz użytkownik układu. W każdym przypadku zadaniem źródła napięcia zasilania jest przede wszystkim zapewnienie, by napięcie dostarczane przez źródło możliwie jak najsłabiej zależało od poboru prądu przez zasilany układ lub blok. Innymi słowy, zasadniczym wymaganiem dla źródła napięcia zasilania jest mała rezystancja wewnętrzna, i pod tym kątem są te układy konstruowane.

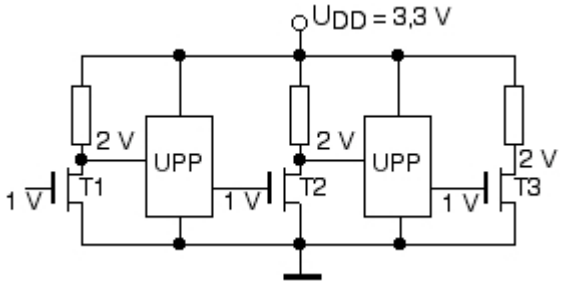
**Źródła napięcia odniesienia** są to układy, których zadaniem jest wytworzenie napięcia służącego jako wzorcowy poziom napięcia, na przykład do porównywania z jakąś inną wartością napięcia wytwarzaną w układzie lub podawaną z zewnątrz na wejście układu. Źródła napięcia odniesienia są na ogół obciążane bardzo małym prądem (w układach MOS jest on najczęściej równy zeru), który nie ulega zmianom. Wobec tego rezystancja wewnętrzna źródła napięcia odniesienia nie ma większego znaczenia. Istotne są natomiast: stałość napięcia w funkcji takich czynników zakłócających, jak wahania napięcia zasilania i wahania temperatury. Stosowane są również do określonych zastosowań źródła napięcia odniesienia mające pewne szczególne cechy, np. proporcjonalność napięcia do temperatury bezwzględnej.

**Układy przesuwania poziomu składowej stałej** umożliwiają połączenie ze sobą stopni lub bloków układu, pomiędzy którymi należy przesyłać sygnały zmienne, a składowe stałe napięć na odpowiednich wejściach i wyjściach różnią się. Typowym przykładem są wzmacniacze kilkustopniowe. Rys. 12.12 pokazuje prosty wzmacniacz trzystopniowy o napięciu zasilania równym 3,3 V (takich układów się w rzeczywistości nie stosuje, ale chodzi tu tylko o ilustrację problemu dopasowania napięć). Załóżmy, że warunki polaryzacji w każdym stopniu są tak dobrane, że napięcie między drenem, a bramką wynosi 1 V. Niech na pierwszej, wejściowej bramce panuje także napięcie 1 V. Wówczas na drenie T1 mamy 2 V, na drenie T2 - 3 V, a na dren T3 nie starcza już napięcia zasilającego!




Rys. 12.12. Ilustracja problemu przesuwania składowej stałej

Wprowadzenie układów przesuwania poziomu składowej stałej rozwiązuje problem - patrz rys. 12.13.



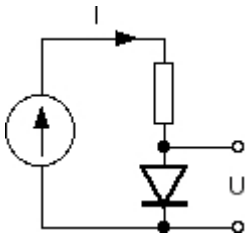

Rys. 12.13. Układy przesuwania poziomu składowej stałej (UPP) wymuszające różnicę napięć równą 1 V rozwiazują problem układu z rys. 12.12

Główne wymaganie dla układów przesuwania poziomu składowej stałej to "przezroczystość" dla składowej zmiennej. Układy te powinny zapewniać transmisję składowej zmiennej z wejścia na wyjście bez tłumienia, zakłóceń i zniekształceń. Wymagania odnoszące się do wymuszanej różnicy napięć mogą być różne, w

zależności od zastosowania.

W naszym wykładzie nie będziemy zajmować się źródłami napięć zasilania. Jest to cała odrębna klasa układów analogowych. Ich bardziej szczegółowe omówienie mogłoby być tematem osobnego wielogodzinnego wykładu. Omówimy natomiast kilka przykładowych rozwiązań układów źródeł napięć odniesienia i układów przesuwania poziomu składowej stałej.

**Pierwotnymi źródłami napięcia odniesienia** będziemy nazywać dwójniki nieliniowe, które cechuje prawie stały, mało zależny od prądu spadek napięcia na pewnym odcinku charakterystyki prądowo-napięciowej. W układach źródeł napięciowych z reguły musi być co najmniej jeden element będący pierwotnym źródłem napięcia odniesienia. Pierwotne źródło napięcia odniesienia jest zwykle wykorzystywane w taki sposób, że wymuszany jest w nim prąd, a towarzyszący mu spadek napięcia jest wykorzystany jako napięcie odniesienia - rys. 12.14.




Rys. 12.14. Typowy sposób wykorzystania pierwotnego źródła napięcia odniesienia. Napięciem jest spadek na elemencie nieliniowym (na rysunku jest to dioda)

**Dioda spolaryzowana w kierunku przewodzenia**

Jest to często spotykane pierwotne źródło napięcia odniesienia. Używany jest zwykle tranzystor bipolarny w połączeniu diodowym. Jedno z zastosowań tego źródła już znamy - w źródłach prądowych. Spadek napięcia na diodzie spolaryzowanej w kierunku przewodzenia, dany zależnością (4.22), jest słabo (logarytmicznie) zależny od prądu płynącego przez diodę i w temperaturze otoczenia wynosi około 0,7 V. Napięcie to maleje o ok. 2 mV/K, dioda nie jest więc źródłem stabilnym temperaturowo.

**Dioda spolaryzowana w zakresie przebicia (zwana potocznie diodą Zenera)**

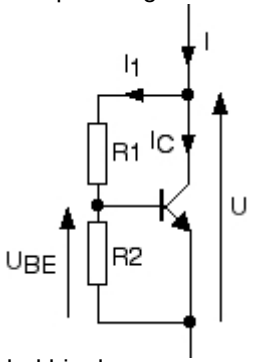
Napięcie przebicia lawinowego złącz p-n w układach scalonych może się wahać w szerokich granicach, od kilku do kilkudziesięciu V. Najniższe napięcia mają zwykle złącza emiter-baza tranzystorów bipolarnych. Napięcie na diodzie spolaryzowanej w zakresie przebicia jest równe napięciu przebicia lawinowego, nieznacznie zależy od prądu płynącego przez diodę i wzrasta z temperaturą. Dla napięć przebicia wynoszących 6 ... 9 V (typowe wartości dla złącz emiter - baza) współczynnik temperaturowy napięcia przebicia wynosi około +3 mV/K. Zatem i to źródło nie jest źródłem stabilnym temperaturowo.

**Tranzystor MOS w połączeniu diodowym**

Źródło używane w układach MOS, znane nam już z układów źródeł prądowych. Spadek napięcia jest określony przez zależność (12.3) , drugi składnik w tej zależności rośnie z pierwiastkiem prądu, nie jest to więc napięcie o dużej stałości. Wpływ drugiego składnika można jednak zminimalizować dobierając odpowiednio małą wartość stosunku wymiarów kanału $L/W$. Zależność od temperatury jest wyraźna, decyduje o niej zmniejszanie się z temperaturą napięcia progowego $U_T$.

**Bipolarny mnożnik $U_{BE}$**

Jest to dwójnik będący połączeniem tranzystora bipolarnego i dwóch rezystorów - rys. 12.15.




Rys. 12.15. Układ bipolarny zwany mnożnikiem $U_{BE}$

Dla zanalizowania działania tego układu pominiemy prąd bazy tranzystora. Mamy wówczas proste zależności:

$$I_1 = \frac{U}{R_1 + R_2} \tag{12.11}$$

$$U_{BE} = R_2 I_1 \tag{12.12}$$

skąd otrzymujemy

$$U = U_{BE} \frac{R_1 + R_2}{R_2} \tag{12.13}$$

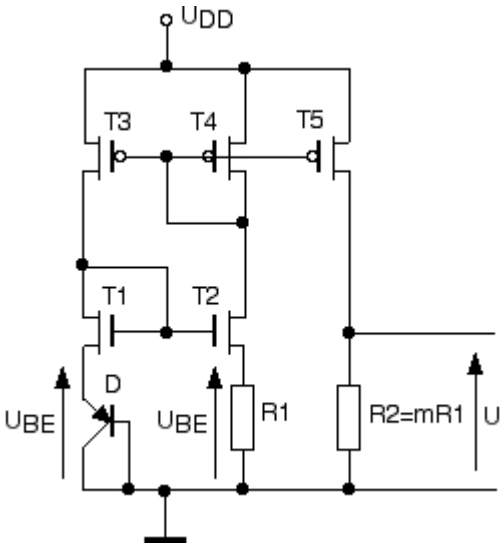
zatem napięcie $U$ jest proporcjonalne do $U_{BE}$. Stąd nazwa układu. Warunkiem działania układu jako pierwotnego źródła napięcia odniesienia jest wymuszenie dostatecznie dużego prądu $I$, tak aby napięcie $U_{BE}$ było dostatecznie duże (około 0.7 V w temperaturze otoczenia) i spełniony był warunek: $I_C >> I_1$. Jeśli warunek ten nie jest spełniony, charakterystyka układu jest liniowa: $I = U/(R_1+R_2)$, ponieważ prąd kolektora tranzystora jest pomijalnie mały. Wzór (12.13) jest wprawdzie nadal słuszny, ale napięcie $U_{BE}$ jest tak małe, że tranzystor praktycznie nie przewodzi. Układ nie spełnia wówczas swej roli.

Zaletą mnożnika $U_{BE}$ jest możliwość uzyskania napięcia o dowolnej wartości przez odpowiedni dobór obu rezystancji. Tak, jak i poprzednie źródła, mnożnik $U_{BE}$ nie jest układem stabilnym temperaturowo. Jeżeli $U = kU_{BE}$ ($k$ wyznaczone przez stosunek rezystancji we wzorze (12.13)), to napięcie $U$ maleje z temperaturą o $k$*2 mV/K.

Jak widać, pierwotne źródła napięć odniesienia nie zapewniają stabilności temperaturowej. Dlatego tam, gdzie potrzebne jest napięcie nie zmieniające się z temperaturą, stosowane są bardziej złożone układy.

Teraz omówimy kilka przykładów **układów źródeł napięć odniesienia.**

Rys. 12.16 przedstawia układ mnożnika $U_{BE}$ wykonanego w technologii CMOS, z wykorzystaniem bipolarnego tranzystora podłożowego p-n-p (wykład 4, część 2 - patrz dodatek 2).



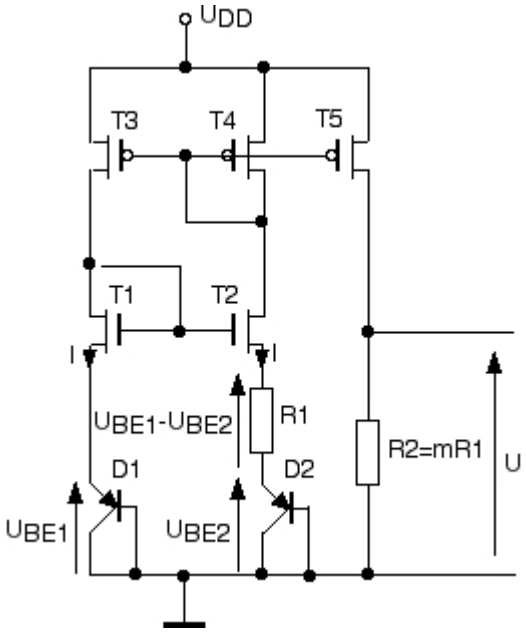

Rys. 12.16. Układ mnożnika $U_{BE}$ w technologii CMOS

Działanie układu jest proste. Dwa skrzyżowane źródła prądowe z tranzystorami T1, T2, T3 i T4 powodują, że jednakowe są prądy: emitera tranzystora bipolarnego w połączeniu diodowym (dlatego oznaczonego D) i rezystora R1. Jednakowe, równe $U_{BE}$, są także napięcia na tranzystorze i na rezystorze. Tranzystory T4 i T5 są identyczne, zatem prądy płynące przez rezystory R1 i R2 są także identyczne. Ponieważ rezystor R2 ma rezystancję $m$-krotnie większą, niż R1, spadek napięcia na nim jest także $m$-krotnie większy. Stąd napięcie $U$ jest równe

$$U = mU_{BE} = \frac{R_2}{R_1} U_{BE} \tag{12.14}$$

Układ mnożnika $U_{BE}$ wg rys. 12.16 daje napięcie $U$ malejące z temperaturą. Po niewielkiej rozbudowie układ

może także dawać napięcie nie malejące, lecz rosnące z temperaturą. Układ taki jest pokazany na rys. 12.17.




Rys. 12.17. Układ mnożnika*kT/q* w technologii CMOS

W tym układzie także prądy $I$ w dwóch równoległych gałęziach są jednakowe, i jednakowe są napięcia źródeł tranzystorów T1 i T2. Tranzystory bipolarne D1 i D2 *nie* są identyczne - różnią się powierzchnią złącza emiter-baza. Niech $A_{E2} = nA_{E1}$. Przy jednakowych prądach $I_C = I$ i większej powierzchni złącza emiterowego tranzystora D2 napięcie $U_{BE2}$ jest mniejsze od $U_{BE1}$, zgodnie z wzorem (4.22). Różnica napięć $U_{BE}$ odkłada się na rezystorze R1 i wynosi

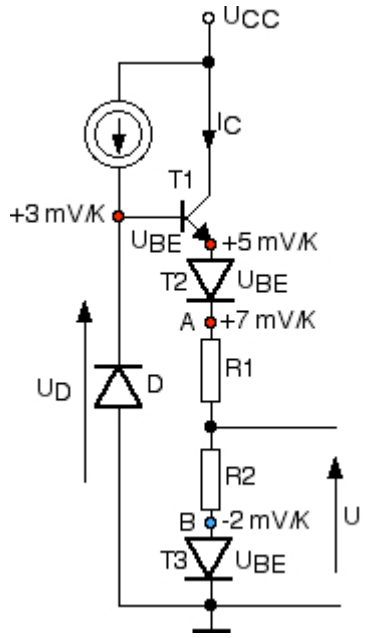
$$U_{R1} = U_{BE1} - U_{BE2} = \frac{kT}{q}\ln\left(\frac{A_{E2}}{A_{E1}}\right) \tag{12.15}$$

To napięcie podlega mnożeniu na rezystancji R2 w taki sam sposób, jak w układzie poprzednim (rys. 12.16). Zatem otrzymujemy

$$U = m\frac{kT}{q}\ln\left(\frac{A_{E2}}{A_{E1}}\right) \tag{12.16}$$

Jak widać, układ ten daje napięcie proporcjonalne do temperatury bezwzględnej T. Taki układ może być wykorzystany na przykład jako czujnik temperatury w elektronicznym termometrze. Można także, łącząc ten układ z układem z rys. 12.16, otrzymać układ źródła napiecia odniesienia dający napięcie niezależne od temperatury. W tym celu można zastosować układ sumowania napięć sumujący oba napięcia z odpowiednimi wagami. Można również postąpić prościej - sumując z wagami bezpośrednio prądy wyjściowe (tj. prądy drenów tranzystorów T5 obu układów) i otrzymany w ten sposób prąd przepuszczając przez rezystor. W takim układzie można uzyskać napięcie praktycznie stałe w szerokim zakresie temperatur.

Na koniec przykład układu z tranzystorami bipolarnymi dającego napięcie niezależne od temperatury - rys. 12.18.




Rys. 12.18. Układ bipolarnego źródła napięcia odniesienia dającego napięcie niezależne od temperatury. Kolorowymi punktami zaznaczono węzły, w których określony jest temperaturowy współczynnik zmian napięcia (względem masy)

Działanie układu polega na wzajemnej kompensacji zależności kilku napięć od temperatury. Diody oznaczone T2 i T3 są to tranzystory takie same, jak T1, w połączeniu diodowym. Dioda D pracuje w zakresie przebicia (prąd w niej wymusza źródło prądowe). Załóżmy, że napięcie na tej diodzie rośnie z temperaturą o 3 mV/K. Na emiterze tranzystora napięcie rośnie o 5 mV/K, w węźle A o 7 mV/K, zaś w węźle B *maleje* o 2 mV/K. Intuicja podpowiada, że odpowiednio dobierając stosunek rezystancji R1 i R2 uda się uzyskać napięcie nie zmieniające się z temperaturą. Tak jest w istocie.

Napięcie wyjściowe $U$ wynosi: $U = I_C R_2 + U_{BE}$. Zaniedbując prądy baz tranzystorów możemy napisać, że prąd $I_C$ jest równy: $I_C = (U_D - 3\,U_{BE})/(R_1 + R_2)$. Stąd otrzymujemy

$$U = \frac{R_2 U_D + (R_1 - 2R_2) U_{BE}}{R_1 + R_2} \qquad (12.17)$$

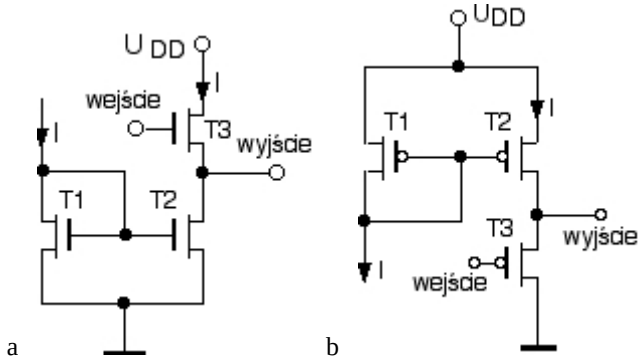
Główne zależności temperaturowe w (12.17) to zależność napięcia $U_D$ (+3 mV/K) i napięcia $U_{BE}$ (-2 mV/K). Zaniedbując zależność rezystancji rezystorów od temperatury łatwo pokazać, że $U$ nie zależy od temperatury, jeśli spełniony jest warunek

$$\frac{R_1 - 2R_2}{R_2} = -\frac{dU_D/dT}{dU_{BE}/dT} = \frac{3}{2} \qquad (12.18)$$

czyli $R_1 = 3{,}5\,R_2$.

Przeprowadzone wyżej analizy i rozumowania są przybliżone, zaprojektowanie rzeczywistych układów wymaga przeprowadzenia symulacji przy użyciu dobrych modeli elementów. Modele te muszą dobrze odzwierciedlać zależności charakterystyk od temperatury.

Na koniec kilka przykładów **układów przesuwania poziomu składowej stałej**. Najprostszym układem przesuwania poziomu składowej stałej jest układ znany jako wtórnik (źródłowy lub emiterowy, w zależności od realizacji w technologii bipolarnej lub MOS). Wtórnik z tranzystorami MOS ma w najprostszym przypadku schemat jak na rys. 12.19. Wersja z tranzystorami nMOS daje możliwość uzyskania na wyjściu napięcia stałego niższego, niż na wejściu, wersja z tranzystorami pMOS - przeciwnie. W obu przypadkach tranzystor T3 działa jako wtórnik (układ ze wspólnym drenem), a tranzystory T1 i T2 tworzą źródło prądowe.



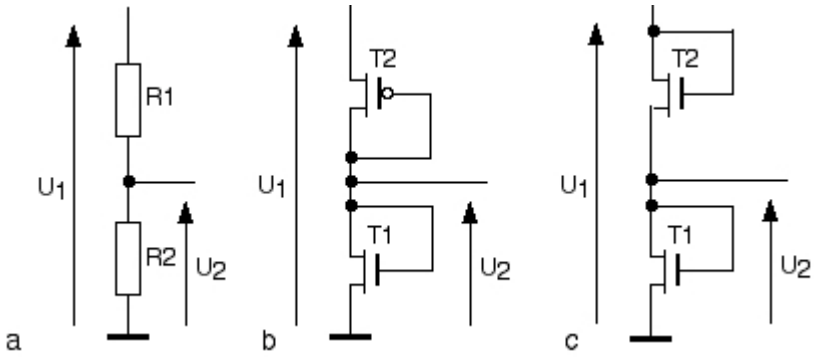

Rys. 12.19. Układy przesuwania poziomu składowej stałej: (a) z tranzystorami nMOS, (b) z tranzystorami pMOS

Różnica między napięciem na wyjściu, a napięciem na wejściu jest dana (co do wartości bezwzględnej) znanym nam już wzorem

$$\Delta U = U_T + \sqrt{\frac{2I}{\mu C_{ox}\left(W_3/L_3\right)}} \tag{12.19}$$

Można ją w pewnym zakresie regulować dobierając prąd $I$ oraz wymiary kanału. Podobne układy buduje się też na tranzystorach bipolarnych.

Jedno źródło napięcia może dostarczyć kilku różnych napięć, jeśli zostanie na wyjściu wyposażone w dzielnik napięcia. Dzielnik napięcia może też służyć jako najprymitywniejsze źródło napięcia odniesienia, jeśli zostanie włączony między $U_{DD}$ i masę. Dzielniki napięcia pokazuje rys. 12.20.




Rys. 12.20. Układy dzielników napięcia: (a) rezystorowy, (b),(c) tranzystorowe

Dzielnik rezystorowy jest najprostszy, ale nie najbardziej ekonomiczny. Rezystory zajmują duże powierzchnie i nie mogą mieć dużych rezystancji, więc dzielnik rezystorowy zwykle pobiera znaczny prąd. Dzielniki tranzystorowe są pod tym względem znacznie korzystniejsze. W przypadku dzielników pokazanych na rys. 12.20 b, c, zakładając prąd wyjściowy równy zeru i przyrównując prądy drenów tranzystorów otrzymuje się

$$U_2 = \frac{m_2}{m_1 + m_2}U_1 + \frac{m_1 U_{T1} - m_2 U_{T2}}{m_1 + m_2} \tag{12.20}$$

gdzie przez $m_1$ i $m_2$ oznaczono

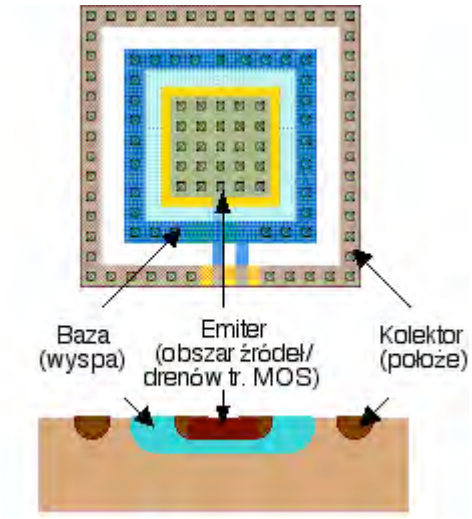
$$m_1 = \sqrt{\mu_1\left(W_1/L_1\right)} \tag{12.21}$$

$$m_2 = \sqrt{\mu_2\left(W_2/L_2\right)} \tag{12.22}$$

W przypadku tranzystora pMOS należy w (12.20) użyć wartości bezwzględnej $U_T$.

Teoretycznie można także zbudować dzielnik napięcia z kondensatorów, jednak jest to rzadko stosowane, bo jakakolwiek większa upływność kondensatora zaburza podział napięcia.

## 12.3. Dodatek 2: Tranzystor bipolarny




Rys. 4.5. Bipolarny tranzystor podłożowy *pnp* w układzie CMOS

# ZADANIA DO WYKŁADU 12

**Zadanie 1**

Oblicz rezystancję rezystora pokazanego na rys. 4.11, dla temperatur T1 = -20$^o$ C i T2 = +120$^o$ C, jeśli wiadomo, że rezystancja warstwowa polikrzemu w temperaturze 27$^o$ C wynosi 25 Ω/□ i rośnie o 0,3%/K, a rezystancja kontaktu jest równa 15 Ω i praktycznie nie zależy od temperatury.

**Zadanie 2**

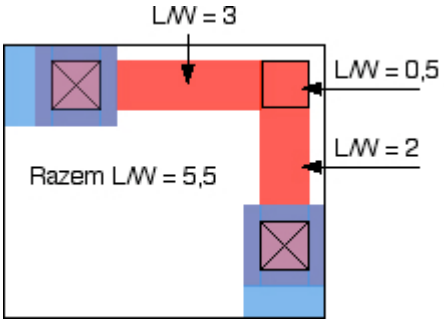
Oblicz, jaka jest minimalna długość bramki tranzystora MOS dla uzyskania napięcia niezrównoważenia mniejszego od 10 mV, dla tranzystora nMOS o $U_{Tn}$ = 0,75 V i dla napięcia $U_{GS}$ = 5V, jeśli wiadomo, że dokładność fotolitografii wynosi ±0.02 µm.

**Zadanie 3**

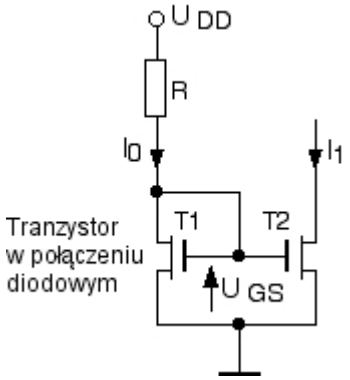
Oblicz, jaka jest minimalna długość boku kwadratowego emitera tranzystora bipolarnego, dla której napięcie niezrównoważenia jest mniejsze od 1 mV, jeśli wiadomo, że dokładność fotolitografii wynosi ±0.2 µm.

**Zadanie 4**

Oblicz, jak dalece może różnić się od jedności stosunek prądów $I_1/I_0$ w podstawowym źródle prądowym MOS (rys. 12.5a), jeśli nominalna wartość napięcia progowego tranzystorów wynosi 0,75 V, rozrzut lokalny tej wartości wynosi 10 mV, nominalna długość kanału tranzystora wynosi 0,25 µm, a rozrzut lokalny tej długości kanału wynosi 0,03 µm? Jak zmieni się wynik, jeśli nominalna długość kanału będzie wynosić 5 µm, a rozrzut pozostanie bez zmiany? Przyjmij, że żadna z pozostałych wielkości określających prąd tranzystora nie wykazuje rozrzutu produkcyjnego, a napięcie $U_{GS}$ = 2,5 V i nie zależy od rozrzutów produkcyjnych.




Rys. 4.11. Rezystor polikrzemowy




Rys. 12.5a. Źródło prądowe MOS

## Bibliografia

[1] F. Maloberti, "*Analog design for CMOS VLSI systems*", Kluwer Academic Publishers, 2001
[2] H. Camenzind, "*Designing Analog Chips*", książka dostępna w internecie:
http://www.designinganalogchips.com/

# Wykład 13: Stopnie wzmacniające i wybrane układy nieliniowe

## Wstęp

Wykład 13 opowiada głównie o tym, jak konstruuje się stopnie wzmacniające. Takie stopnie znajdziemy w prawie każdym układzie analogowym. Wśród nich szczególne miejsce zajmuje wzmacniacz różnicowy. Można śmiało powiedzieć, że bez tego układu nie istniałaby znaczna część współczesnej mikroelektroniki układów analogowych. Dlatego poświęcimy mu nieco więcej uwagi. Pokazane będą nie tylko typowe zastosowania do wzmacniania sygnałów analogowych, ale także przykłady zastosowań do wykonywania operacji nieliniowych.

W przypadku stopni wzmacniających najważniejsze parametry elektryczne to wzmocnienie napięciowe, a także szerokość pasma wzmacnianych częstotliwości i maksymalna amplituda sygnału na wyjściu. W konkretnych zastosowaniach mogą także mieć znaczenie: rezystancja wyjściowa, moc, jaką stopień wzmacniający może oddać do obciążenia, maksymalna szybkość zmiany napięcia na wyjściu. Te zagadnienia będą dyskutowane przy omawianiu konkretnych przykładów stopni wzmacniających.

Łącznie materiał wykładów 12 i 13 daje możliwość zrozumienia działania typowych bloków bardzo wielu różnorodnych układów analogowych i pozwala poznać w zarysie istotę projektowania układów analogowych.

## 13.1. Proste stopnie wzmacniające

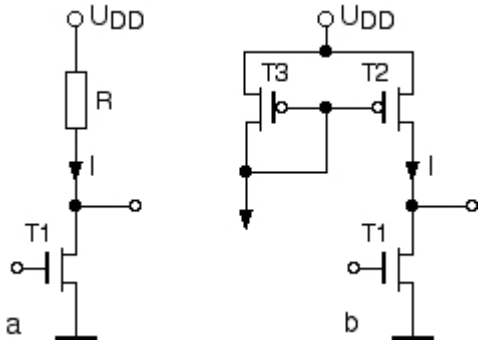
Przedmiotem znacznej części wykładu 13 są wzmacniacze małych sygnałów, tj. sygnałów zmiennych o małej amplitudzie. Sygnały takie z reguły są w układzie elektronicznym nałożone na pewne napięcie stałe, w związku z czym będziemy wyróżniać:

- składową stała napięcia, która jak dotąd będzie oznaczana dużą literą $U$,
- składową zmienną napięcia, której amplitudę będziemy oznaczali małą literą $u$.

Przykładowo: napięcie bramka-źródło tranzystora MOS zawierające obie składowe, w tym sinusoidalną składową zmienną, będzie miało postać:

$$U = U_{GS} + u_{gs}\sin(\omega t + \varphi) \tag{13.1}$$

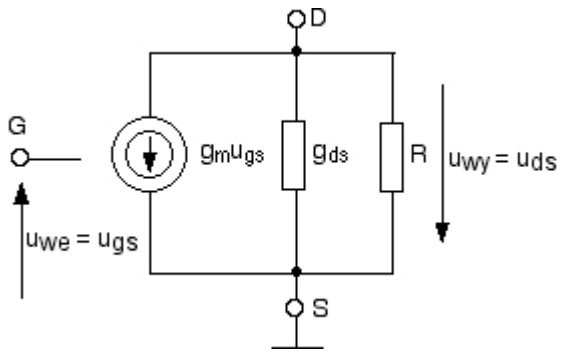
Najprostszy wzmacniacz można zbudować według rys. 13.1a. Składowa zmienna napięcia bramki powoduje powstawanie składowej zmiennej prądu drenu tranzystora. Ta, przepływając przez rezystor obciążający R, powoduje powstawanie na nim składowej zmiennej napięcia wyjściowego, której amplituda jest większa, niż amplituda składowej zmiennej napięcia wejściowego. Wzmacniacz zapewnia więc wzmocnienie napięciowe $k_u = u_{wy}/u_{we}$, gdzie $u_{we}$ jest amplitudą składowej zmiennej napięcia na wejściu, a $u_{wy}$ jest amplitudą składowej zmiennej napięcia na wyjściu.




Rys. 13.1. Zasada budowy najprostszego stopnia wzmacniającego: (a) z obciążeniem w postaci rezystora R, (b) z obciążeniem aktywnym.

Aby układ działał, bramka tranzystora T1 musi być oczywiście spolaryzowana względem źródła napięciem stałym wyższym od progowego, co nie jest pokazane na rysunku.

Prosty układ pokazany na rys. 13.1 jest w rzeczywistości nieprzydatny, ponieważ możliwe do osiągnięcia wzmocnienie napięciowe jest bardzo niewielkie. Posłuży on nam jednak do zilustrowania metody analizy małosygnałowej. Zastępujemy tranzystor T1 przez jego małosygnałowy schemat zastępczy (rys. 4.3 - przypomina go dodatek 1). Będziemy określać wzmocnienie dla bardzo małych częstotliwości, toteż pomijamy wszystkie pojemności. Zakładamy także, że do pominięcia są rezystancje rozproszone źródła i drenu. Dołączamy rezystor R. Jest on włączony między dren i źródło, ponieważ dla składowych zmiennych źródło zasilania jest zwarciem. Ostatecznie otrzymujemy schemat jak na rys. 13.2.




Rys. 13.2. Małosygnałowy schemat zastępczy układu z rys. 13.1a.

Zwróćmy uwagę, że wzmacniacz odwraca fazę sygnału zmiennego. Analizując wartości chwilowe napięć nietrudno przekonać się, że gdy napięcie na bramce rośnie, to na drenie maleje.

Z rysunku 13.2 od razu widać, że napięcie wyjściowe $u_{wy}$ otrzymamy mnożąc prąd źródła prądowego równy $g_m u_{we}$ przez rezystancję złożoną z równolegle połączonych: rezystancji R i konduktancji wyjściowej tranzystora $g_{ds}$. Zatem

$$|k_u| = g_m \frac{1}{g_{ds} + 1/R} \tag{13.2}$$

Konduktancja $g_{ds}$ jest z reguły znacznie mniejsza od konduktancji $1/R$, wobec czego można ją pominąć i zależność (13.2) uprościć do

$$|k_u| = g_m R \tag{13.3}$$

Rezystancja $R$ nie może być dowolna, bowiem jej wartość wraz ze składową stałą $I_D$ prądu drenu oraz napięciem zasilania $U_{DD}$ decydują o punkcie pracy tranzystora, tj. składowej stałej napięcia dren-źródło $U_{DS}$:

$$U_{DS} = U_{DD} - I_D R \tag{13.4}$$

Załóżmy, że tranzystor pracuje w zakresie nasycenia, a napięcie $U_{DS}$ jest równe połowie napięcia zasilania: $U_{DS}=U_{DD}/2$ (przy takiej lub zbliżonej wartości uzyskuje się największą możliwą amplitudę sygnału zmiennego na wyjściu). Wówczas, wykorzystując zależność (4.17), można otrzymać prosty wynik:

$$|k_u| = \frac{U_{DD}}{U_{GS} - U_T} \tag{13.5}$$

Widać, że otrzymane wzmocnienie jest niewiele większe od jedności, bo i licznik, i mianownik we wzorze (13.5) mają wartość rzędu kilku woltów. Nieco lepszy wynik można by otrzymać, gdyby tranzystor pracował w zakresie podprogowym. Posługując się wówczas wzorem (4.18) otrzymamy

$$|k_u| = \frac{U_{DD}}{2n\dfrac{kT}{q}} \tag{13.6}$$

Mianownik (13.6) ma typową wartość około 100 mV, więc można uzyskać wzmocnienie napięciowe rzędu kilkudziesięciu. Jest to nadal niewiele. W dodatku rezystor - jak wiemy - jest elementem o dużej powierzchni, więc nieekonomicznym. Dlatego praktyczne znaczenie ma układ z rys. 13.1b, w którym rezystor zastąpiony jest przez źródło prądowe. Taki układ nosi nazwę wzmacniacza z aktywnym obciążeniem. Jak wiemy, źródło prądowe ma bardzo dużą rezystancję wewnętrzną. W najprostszym układzie (rys.13.1b) jest ona równa $1/g_{dsT2}$ (patrz wykład 12). Zamieniając w schemacie zastępczym z rys. 13.2 rezystor R na rezystancję wyjściową $1/g_{dsT2}$ otrzymujemy

$$|k_u| = \frac{g_{mT1}}{g_{dsT1} + g_{dsT2}} \tag{13.7}$$

Wykorzystując tu wzory (4.14) oraz (4.17) można wzór (13.7) przekształcić do postaci

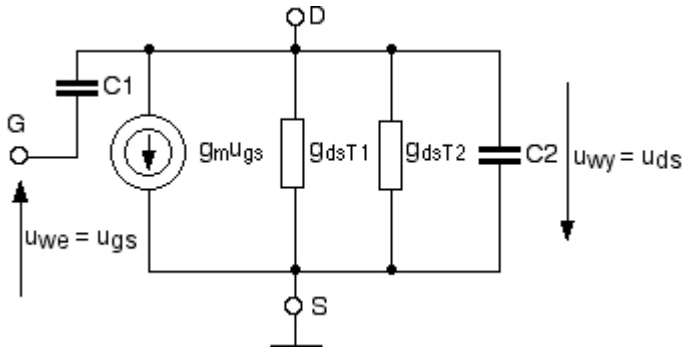
$$|k_u| = \frac{\sqrt{2\mu_{T1} C_{ox} \dfrac{W_{T1}}{L_{T1}}}}{(\lambda_{T1} + \lambda_{T2})\sqrt{I_D}} \tag{13.8}$$

Typowa wartość parametru $\lambda$ wynosi 0,05 $V^{-1}$ ... 0,1 $V^{-1}$. Można się przekonać, że bez trudu uzyskuje się przy typowych wartościach pozostałych parametrów wzmocnienia rzędu 100 ... 200. Wzmocnienie rośnie, gdy *maleje* prąd $I_D$. Maksymalne osiągalne wzmocnienie otrzymamy przy pracy tranzystora w zakresie podprogowym. Wynosi ono

$$|k_u| = \frac{1}{n(\lambda_{T1} + \lambda_{T2})\dfrac{kT}{q}} \tag{13.9}$$

Jest ono w tym zakresie niezależne od prądu $I_D$. Typowa wartość wynosi ok. 250.

Drugim istotnym w wielu zastosowaniach parametrem jest szerokość pasma częstotliwości, w jakim uzyskuje się wzmocnienie. Tę szerokość pasma można scharakteryzować określając częstotliwość graniczną $f_T$. Jest to częstotliwość, przy której wartość bezwzględna wzmocnienia napięciowego spada do jedności: $|k_u| = 1$. Aby ją określić, trzeba schemat zastępczy uzupełnić pojemnościami.




Rys. 13.3. Schemat zastępczy wzmacniacza z aktywnym obciążeniem uzupełniony pojemnościami

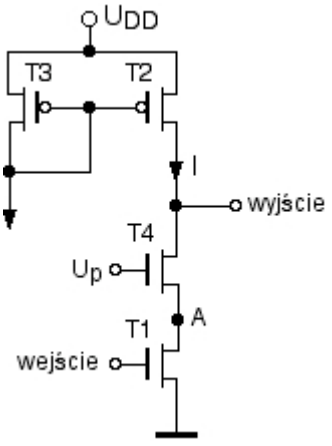
Pojemność C1 w schemacie zastępczym jest to suma wszystkich pojemności włączonych między dren i bramkę. Podobnie, C2 jest sumą wszystkich pojemności obciążających węzeł wyjściowy. Pojemność wejściowa układu nie ma wpływu na częstotliwość $f_T$ (pod warunkiem, że układ jest sterowany z idealnego źródła napięciowego sygnału zmiennego). Analiza wzmocnienia w funkcji częstotliwości przy wykorzystaniu schematu z rys. 13.3 prowadzi do następującego wzoru na częstotliwość $f_T$:

$$f_T = \frac{\sqrt{2 I_D \mu_{T1} C_{ox} \frac{W_{T1}}{L_{T1}}}}{2\pi(C_1 + C_2)} \qquad (13.10)$$

(wzór ten obowiązuje dla tranzystorów pracujących w zakresie nasycenia). Jak widać, szerokość pasma rośnie z prądem $I_D$, przeciwnie niż wzmocnienie dla małych częstotliwości. Oznacza to, że projektant układu ma wybór: albo duże wzmocnienie w wąskim pasmie, albo niewielkie, ale w szerokim pasmie częstotliwości. Taką właściwość, zwaną potocznie wymiennością pasma i wzmocnienia, ma zresztą większość typowych układów wzmacniających, nie tylko układ teraz omawiany.

Istotną zaletą omawianego układu jest duża amplituda sygnału wyjściowego. Jedyne ograniczenie to warunek, aby tranzystory T1 i T2 pozostawały w stanie nasycenia. Oznacza to, że wartość chwilowa napięcia na wyjściu może zmieniać się w granicach od $U_{DSsatT1}$ do $U_{DD}$-$U_{DSsatT2}$. Napięcia nasycenia mają typową wartość rzędu kilkuset mV, a więc napięcie wyjściowe może się zmieniać prawie od $0$ do$U_{DD}$ . Istnieje wiele układów stopni wzmacniających, które dają większe wzmocnienie lub mają inne zalety, ale z reguły ceną za to jest komplikacja schematu prowadząca m.in. do ograniczenia amplitudy sygnału na wyjściu.

Układ wzmacniacza z aktywnym obciążeniem ma pewną słabą stronę: pojemność włączona między wyjście, a wejście ulega pozornemu zwielokrotnieniu (zjawisko zwane efektem Millera). Od strony wejścia pojemność $C_1$ jest widoczna jako $C_1' = C_1(|k_u|+1)$. Dla $|k_u|$ rzędu 100 ... 200 nawet bardzo mała pojemność $C_1$ oznacza znaczne obciążenie pojemnościowe poprzedniego stopnia. Ogranicza to szerokość pasma we wzmacniaczach kilkustopniowych. Efekt ten można wyeliminować we wzmacniaczu w układzie pokazanym na rys. 13.4. Jest on zwany układem kaskodowym.

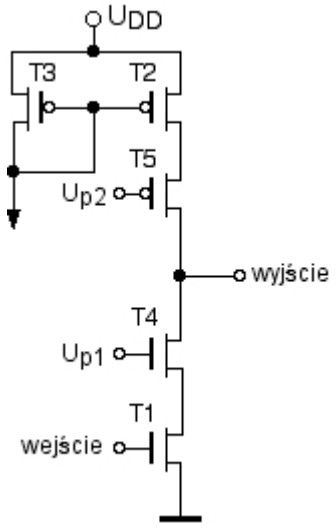



Rys.13.4. Wzmacniacz w układzie kaskody

Dodatkowy tranzystor T4 jest polaryzowany napięciem stałym $U_p$ tak dobranym, by wszystkie tranzystory

pracowały w zakresie nasycenia. Analizując schemat zastępczy tego układu można pokazać, że chociaż wzmocnienie napięciowe układu jako całości jest zbliżone do wzmocnienia układu poprzednio omawianego, to wzmocnienie między węzłem A, a wyjściem jest znacznie mniejsze. W rezultacie wpływ efektu Millera na pojemność wejściową jest znacznie zredukowany.

Można również rozbudować w podobny sposób układ aktywnego obciążenia - rys. 13.5.

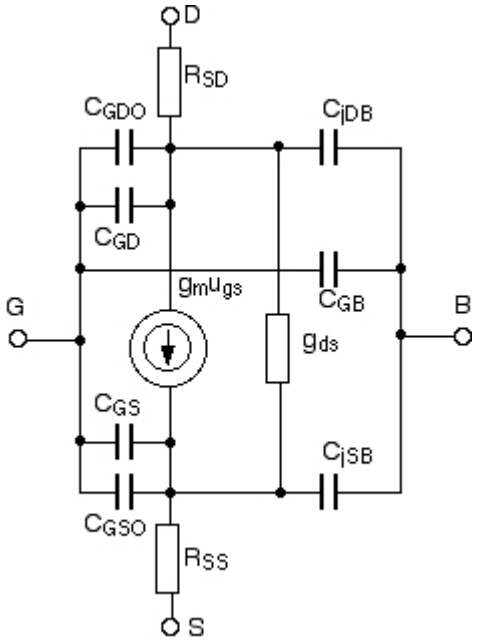



Rys. 13.5. Układ kaskody z kaskodowym obciążeniem aktywnym

Dodatkowy tranzystor T5 znacznie zwiększa rezystancję obciążenia pozwalając osiągnąć wzmocnienie napięciowe nawet o 2 rzędy wielkości wyższe, niż wzmocnienie układu podstawowego (rys. 13.1b). Wadą takiego układu jest konieczność dostarczenia dodatkowych napięć polaryzujących (może tu znaleźć zastosowanie jedno z omawianych wcześniej źródeł i dzielników napięcia). Cechuje go też mniejsza amplituda sygnału na wyjściu, bo musi być zapewniona praca w nasyceniu dwóch tranzystorow nMOS połączonych szeregowo i dwóch tranzystorów pMOS połączonych szeregowo.

Układy omawiane wyżej mogą być także budowane na tranzystorach bipolarnych. Można uzyskać większe wzmocnienia, ponieważ transkonduktancja tranzystora bipolarnego jest znacznie większa, niż tranzystora MOS - patrz wzór (4.30).

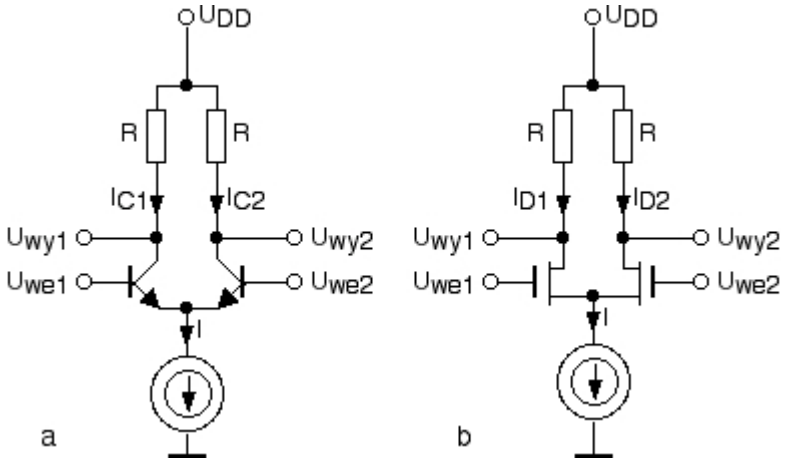
## 13.1. Dodatek 1: Pełny małosygnałowy schemat zastępczy tranzystora MOS




Rys. 4.3. Małosygnałowy schemat zastępczy tranzystora MOS

## 13.2. Wzmacniacz różnicowy

Wzmacniacz różnicowy w różnych odmianach i wariantach jest prawdopodobnie najczęściej stosowanym układem wzmacniającym w analogowych układach scalonych. Jest powszechnie stosowany zarówno w wersji bipolarnej, jak i MOS. Podstawowy układ pokazuje rys. 13.6.




Rys. 13.6. Podstawowy wzmacniacz różnicowy: (a) bipolarny, (b) MOS

Zakładamy, że nie pokazane na rysunku układy polaryzacji zapewniają takie wartości napięć na bazach lub bramkach tranzystorów, że pracują one we właściwych zakresach charakterystyk, tj. tranzystory bipolarne w zakresie polaryzacji normalnej ($U_{CB}<0$, $U_{BE}>0$), a tranzystory MOS w zakresie nasycenia. Założymy, że układ pokazany na rys. 13.6 jest dokładnie symetryczny, tj. tranzystory są identyczne i rezystory mają identyczną rezystancję R. Oba układy są zasilane prądem I wymuszonym przez idealne źródło prądowe. Dla $U_{we1} = U_{we2}$ prąd ten dzieli się po połowie między obie gałęzie układu. Spadek napięcia na obu rezystorach jest jednakowy, a różnicowe napięcie wyjściowe $U_{wy} = U_{wy2} - U_{wy1}$ jest równe zeru. Jeśli istnieje różnica napięć $U_{we} = U_{we2} - U_{we1}$ (zwana różnicowym napięciem wejściowym), to powstaje różnica prądów kolektora lub drenu tranzystorów, spadki napięć na rezystorach są różne i pojawia się różne od zera różnicowe napięcie wyjściowe.

Zależność różnicowego napięcia wyjściowego od różnicowego napięcia wejściowego łatwo otrzymać dla wzmacniacza z tranzystorami bipolarnymi. Prąd kolektora dany jest wzorem (4.20), zatem dla różnicy prądów $I_{C2} - I_{C1}$ mamy

$$I_{C2} - I_{C1} = I_{ES0}\left[\exp\left(\frac{qU_{BE2}}{kT}\right) - \exp\left(\frac{qU_{BE1}}{kT}\right)\right] \tag{13.11}$$

a równocześnie

$$I = I_{C2} + I_{C1} = I_{ES0}\left[\exp\left(\frac{qU_{BE2}}{kT}\right) + \exp\left(\frac{qU_{BE1}}{kT}\right)\right] \tag{13.12}$$

Różnicowe napięcie wyjściowe wynosi

$$U_{wy} = \left(U_{DD} - U_{R2}\right) - \left(U_{DD} - U_{R1}\right) = U_{R1} - U_{R2} = -R\left(I_{C2} - I_{C1}\right) \tag{13.13}$$

Z (13.11), (13.12) i (13.13) można po prostych algebraicznych przekształceniach otrzymać wynik

$$U_{wy} = -IR\tanh\left(\frac{1}{2}\frac{qU_{we}}{kT}\right) \tag{13.14}$$

Funkcja ta ma następujące właściwości:

- dla małych wartości różnicowego napięcia wejściowego ($|U_{we}|<kT/q$) zależność $U_{wy}$ od $U_{we}$ jest praktycznie liniowa,
- dla dostatecznie dużych wartości napięcia wejściowego ($|U_{we}|>2kT/q$) układ działa jako ogranicznik amplitudy,
- małosygnałowe wzmocnienie napięciowe $k_u$ na liniowym odcinku charakterystyki wynosi

$$|k_u| = \frac{qI}{2kT} R = g_m R \tag{13.15}$$

gdzie $g_m$ jest transkonduktancją pojedynczego tranzystora dla prądu $I_C = I/2$.

Dla tranzystorów MOS uzyskanie prostego wzoru opisującego całą charakterystykę przejściową $U_{wy} = f(U_{we})$ nie jest możliwe, bowiem dla dużych wartości różnicowego napięcia wejściowego jeden bądź drugi z tranzystorow przestaje pracować w zakresie nasycenia. Jeżeli ograniczyć się do napięć wejściowych, dla których oba tranzystory znajdują się w stanie nasycenia, to łatwo otrzymać

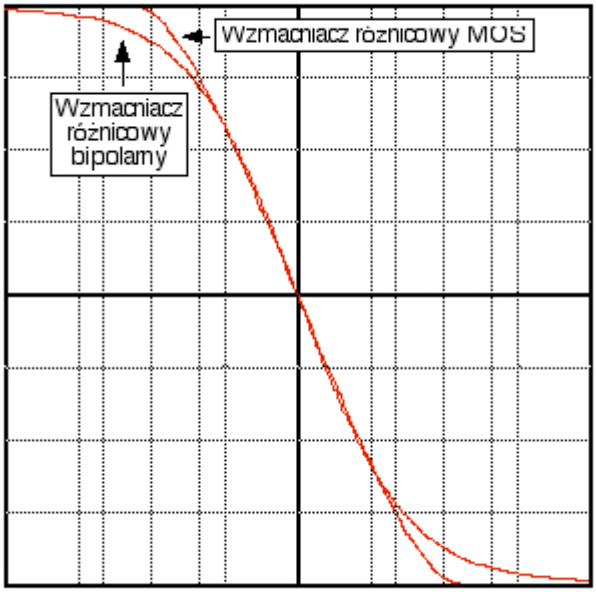
$$U_{wy} = -U_{we} R \sqrt{\mu C_{ox} \frac{W}{L} I} \tag{13.16}$$

skąd dla wzmocnienia napięciowego otrzymujemy

$$|k_u| = R \sqrt{\mu C_{ox} \frac{W}{L} I} = g_m R \tag{13.17}$$

gdzie $g_m$ jest transkonduktancją pojedynczego tranzystora dla prądu $I_D = I/2$.

Jak widać, charakterystyka przejściowa wzmacniacza MOS jest liniowa tak długo, jak długo tranzystory znajdują się w stanie nasycenia. Jednak różnice kształtu obu charakterystyk nie są duże. Rys. 13.7 pokazuje obie charakterystyki (otrzymane przy pomocy symulatora) znormalizowane w taki sposób, by obie miały to samo nachylenie i te same wartości napięcia wyjściowego dla dużych wartości napięcia wejściowego.



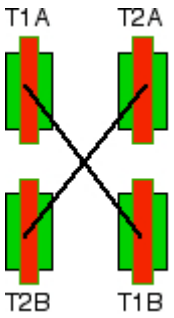

Rys. 13.7. Charakterystyki przejściowe wzmacniaczy różnicowych: bipolarnego i MOS (nierównomierna siatka wynika z nałożenia dwóch wykresów wykonanych w różnych skalach)

Interesującą i wykorzystywaną w praktyce właściwością wzmacniaczy różnicowych jest zależność wzmocnienia napięciowego od prądu źródła prądowego $I$ - liniowa w przypadku wzmacniacza bipolarnego, pierwiastkowa w przypadku wzmacniacza MOS. Jak zobaczymy dalej, umożliwia to wykorzystanie wzmacniacza różnicowego do realizacji pewnych operacji nieliniowych, jak np. operacja mnożenia sygnałów analogowych.

Dużą zaletą wzmacniacza z symetrycznym wejściem i wyjściem jest niewrażliwość na zakłócenia pojawiające się na tle napięcia zasilania. Jeżeli wejściowe napięcie różnicowe jest równe zeru, to i wyjściowe napięcie różnicowe jest równe zeru niezależnie od ewentualnych zakłóceń czy wahań napięcia zasilania. Natomiast napięcie zasilania może mieć wpływ na *amplitudę* sygnału wyjściowego, jeśli nie jest ona równa zeru, poprzez wpływ na prąd zasilający $I$. Zastosowanie źródła prądowego o słabej zależności prądu $I$ od napięcia zasilającego pozwala znacznie zredukować tę zależność.

Symetrię układu zawsze zakłócają w jakimś stopniu lokalne rozrzuty produkcyjne. Decydujące znaczenie mają rozrzuty charakterystyk tranzystorów. Były one omawiane w poprzednich wykładach (wykład 2, część 2 oraz wykład 12, część 1). Zasady minimalizacji wpływu rozrzutów lokalnych omówione dla źródeł prądowych (wykład

12, część 2) obowiązują także dla par tranzystorów we wzmacniaczach różnicowych. Zastosować można topografię pokazaną na rysunku 12.6. Dla uzyskania jeszcze lepszej symetrii stosuje się też topografię zwaną w jęz. angielskim *common centroid* (brak dobrego polskiego tłumaczenia). Topografia taka polega na podziale każdego z tranzystorów pary różnicowej na dwa tranzystory, i połączeniu ich równolegle na krzyż (rys. 13.8). Zachować trzeba przy tym pozostałe reguły: identyczność wszystkich szczegółów topografii każdego z czterech tranzystorów, symetrię wszystkich połączeń itp. Te wymagania powodują, że topografia *common centroid* zajmuje dużo więcej miejsca, niż najprościej zaprojektowana para tranzystorów.




Rys.13.8. Zasada budowy topografii typu common centroid. Tranzystory T1 i T2 pary różnicowej są podzielone na T1A i T1B oraz T2A i T2B, a następnie połączone równolegle na krzyż.

W następnych częściach wykładu zobaczymy, jak szczególne cechy wzmacniacza różnicowego są wykorzystywane w różnych zastosowaniach.

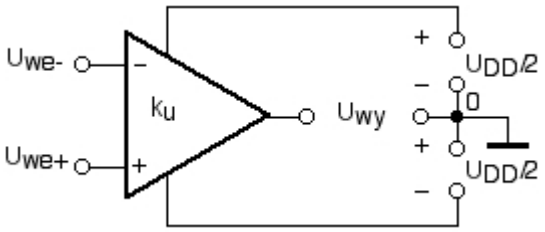
## 13.3. Przykłady zastosowań wzmacniacza różnicowego

**Wzmacniacz operacyjny**

Klasycznym zastosowaniem wzmacniacza różnicowego jest użycie go do budowy wzmacniacza operacyjnego. Jak wiemy, wzmacniacz operacyjny jest to układ elektroniczny mający różnicowe wejście i asymetryczne wyjście (rys. 13.9), a zależność napięcia wyjściowego od wejściowego jest dana wzorem

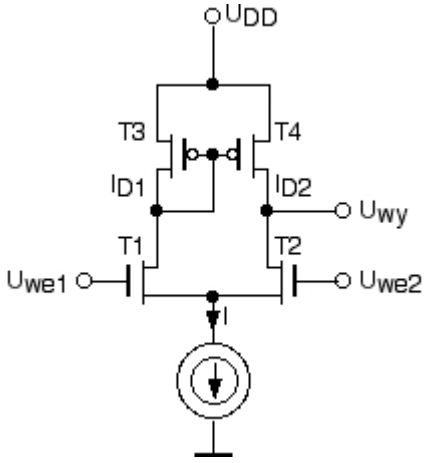
$$U_{wy} = k_u\left(U_{we+} - U_{we-}\right) \tag{13.18}$$

przy czym dąży się do tego, by wzmocnienie napięciowe $k_u$ było jak największe (w t.zw. idealnym wzmacniaczu operacyjnym jest ono nieskończenie wielkie).




Rys. 13.9. Wzmacniacz operacyjny - symbol i sposób zasilania z dwóch źródeł.
Węzeł oznaczony "0" jest węzłem odniesienia dla napięć wejściowych i napięcia wyjściowego. Wejście oznaczone plusem zwane jest wejściem nie odwracającym fazy, wejście oznaczone minusem - wejściem odwracającym fazę.

Ponieważ wzmacniacz operacyjny ma napięcie wyjściowe proporcjonalne do *różnicy* napięć wejściowych, zastosowanie na wejściu wzmacniacza różnicowego jest oczywiste. Jednak taki wzmacniacz będzie różnił się od układu podstawowego (rys. 13.6). Po pierwsze, wyjście wzmacniacza operacyjnego jest asymetryczne. Po drugie, dąży się do uzyskania jak największego wzmocnienia. Wzmacniacz różnicowy o dużym wzmocnieniu i asymetrycznym wyjściu wykorzystuje zasadę aktywnego obciążenia - rys. 13.10.



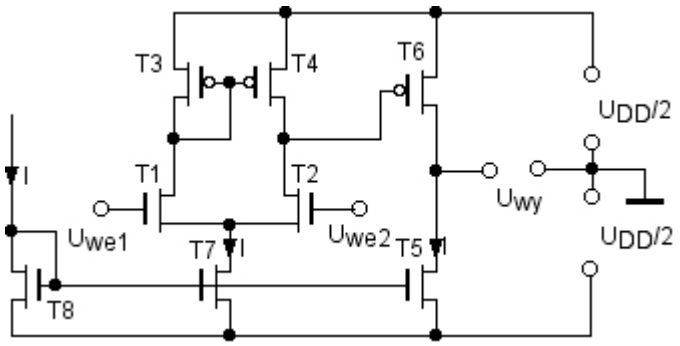

Rys. 13.10. Wzmacniacz różnicowy z aktywnym obciążeniem i asymetrycznym wyjściem

Ze schematu zastępczego tego układu można łatwo otrzymać wartość wzmocnienia napięciowego dla sygnału różnicowego. Wzmocnienie to wynosi

$$\left|k_u\right| = \frac{g_{m1}}{g_{ds2} + g_{ds4}} \tag{13.19}$$

Analogiczny wzmacniacz różnicowy można także zbudować na tranzystorach bipolarnych.

Jeśli dodamy do stopnia wejściowego drugi stopień wzmacniający z aktywnym obciążeniem (zbudowany wg zasady pokazanej na rys. 13.1b, z tym, że tranzystorem wzmiacniającym jest tranzystor pMOS, a obciążającym - nMOS), otrzymamy najprostszy wzmacniacz operacyjny - rys. 13.11.




Rys. 13.11. Prosty transkonduktancyjny wzmacniacz operacyjny (OTA). Pokazano zasilanie z dwóch jednakowych źródeł napięcia.

Wzmacniacz ten ma znaczną rezystancję wyjściową. Można go uważać za swego rodzaju źródło prądowe o prądzie sterowanym napięciem różnicowym na wejściu. Wzmacniacz o takich właściwościach nosi nazwę transkonduktancyjnego, a w literaturze występuje często jako wzmacniacz typu OTA (z ang. *Operational Transconductance Amplifier*). Jego wzmocnienie napięciowe otrzymamy mnożąc wzmocnienie stopnia wejściowego przez wzmocnienie stopnia wyjściowego. Daje to zależność:

$$|k_u| = \frac{g_{mT1} g_{mT5}}{(g_{dsT2} + g_{dsT4})(g_{dsT5} + g_{dsT6})} \tag{13.20}$$

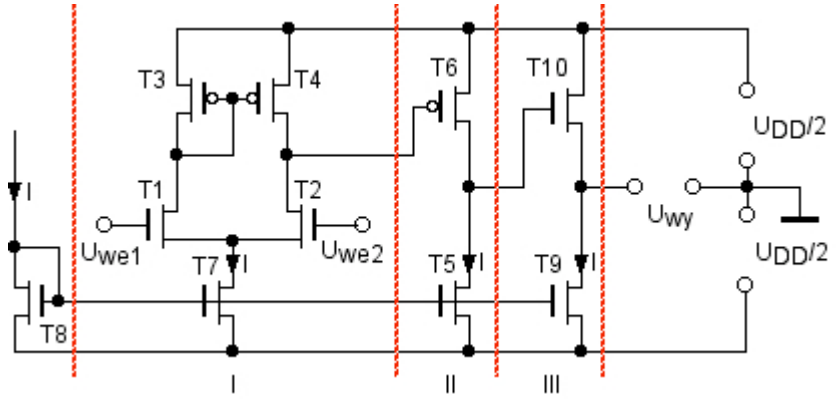
Zauważmy, że w wersji pokazanej na rys. 13.11 prąd drugiego stopnia jest taki sam, jak prąd zasilający stopień wejściowy (zakładając że tranzystory T5, T7 i T8 są identyczne). Ponieważ transkonduktancja $g_m$ jest proporcjonalna do pierwiastka z prądu $I_D$, a konduktancja wyjściowa $g_{ds}$ proporcjonalna do $I_D$, wzmocnienie napięciowe $|k_u|$ jest odwrotnie proporcjonalne do prądu $I$ w układzie.

Układ powinien zapewniać dla wejściowego napięcia różnicowego równego zeru napięcie wyjściowe także równe zeru. Można pokazać, że jeśli wszystkie tranzystory są w stanie nasycenia, to prowadzi to do następującego warunku

$$\frac{W_{T3}}{L_{T3}} \frac{L_{T6}}{W_{T6}} = \frac{W_{T7}}{L_{T7}} \frac{L_{T5}}{2W_{T5}} \tag{13.21}$$

Warunek ten należy w praktyce potraktować jako punkt wyjścia do dokładnego dobrania wymiarów tranzystorów przy zastosowaniu symulacji. Jeśli układ jest zaprojektowany tak, że przy zerowym napięciu różnicowym na wejściu i nominalnych (tj. nie wykazujących rorzutów produkcyjnych) wymiarach kanałów tranzystorów napięcie na wyjściu jest równe zeru, to mówimy że w układzie nie występuje **niezrównoważenie systematyczne**. W praktycznej realizacji pozostaje jednak zawsze **niezrównoważenie losowe** wynikające z lokalnych rozrzutów produkcyjnych.

Większość scalonych wzmacniaczy operacyjnych CMOS to wzmacniacze typu OTA. Jeżeli jednak potrzebny jest wzmacniacz operacyjny o małej rezystancji wyjściowej (tak jest w niektórych zastosowaniach), to trzeba zastosować specjalny stopnień wyjściowy. Cały wzmacniacz staje się bardziej złożony i ma zwykle trzy stopnie: wejściowy wzmacniacz różnicowy, stopień pośredni dostarczający dodatkowego wzmocnienia i zarazem będący układem przesuwania składowej stałej, i stopień wyjściowy o małej rezystancji wyjściowej. Jako stopień wyjściowy może być w najprostszym przypadku użyty układ wtórnika, który poznaliśmy wcześniej w zastosowaniu do przesuwania poziomu składowej stałej (rys. 12.19a). Taki układ pokazuje rys. 13.12.



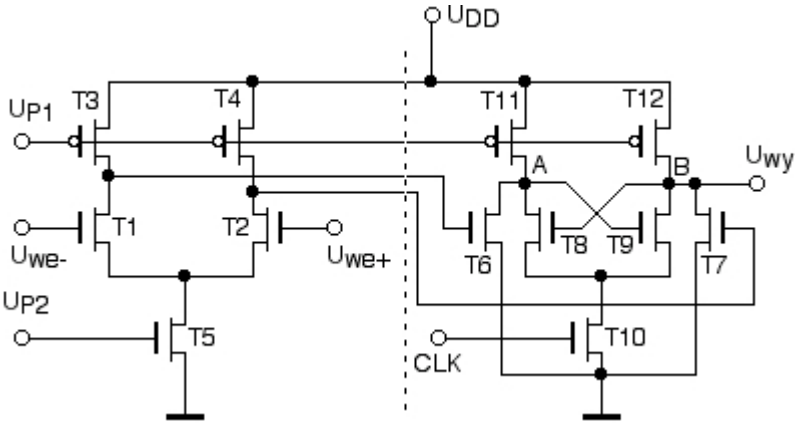

Rys. 13.12. Przykład wzmacniacza operacyjnego ilustrujący podział na 3 stopnie

Dla układu z rys. 13.12 zależność (13.21) nie obowiązuje, ponieważ trzeci stopień wprowadza dodatkowe przesunięcie poziomu napięcia na wyjściu. Układu tego nie będziemy szczegółowo omawiali, posłużył on bowiem tylko do zilustrowania typowej budowy trzystopniowego wzmacniacza operacyjnego. Praktyczne układy wzmacniaczy są bardziej skomplikowane. W szczególności, stosowane są zwykle stopnie wyjściowe w układzie przeciwsobnym, które poznamy nieco dalej.

**Komparator napięcia**

Komparator napięcia jest to podstawowy układ wiążący domeny elektroniki analogowej i cyfrowej. Podobnie jak wzmacniacz operacyjny, ma dwa wejścia i jedno wyjście. Podobnie jak wzmacniacz operacyjny, reaguje na napięcie różnicowe, czyli różnicę napięć na dwóch wejściach. Od wzmacniacza operacyjnego różni się tym, że na jego wyjściu pojawiają się sygnały cyfrowe: napięcia reprezentujące jedynkę logiczną lub zero logiczne w zależności od tego, jaki jest znak napięcia różnicowego na wejściach. Jeśli na wejściu nie odwracającym fazy napięcie jest wyższe, niż na wejściu odwracającym fazę, na wyjściu pojawia się napięcie o wartości jedynki logicznej, w przeciwnym razie pojawia się zero. Takie układy mają bardzo wiele zastosowań. Układy pełniące rolę wzmacniaczy odczytu w pamięciach statycznych i dynamicznych (patrz wykład 9, rys. 9.3 i 9.5) są szczególnym rodzajem komparatorów napięcia. Komparatory napięcia są też niezbędne w układach przetworników analogowo-cyfrowych i wielu innych układach, w których sygnały analogowe są w jakiś sposób mierzone bądź porównywane z innymi sygnałami analogowymi.

Komparatory napięcia dzielą się na swa rodzaje. Pierwszy rodzaj to komparatory działające w sposób ciągły - stan wyjścia (zero lub jedynka) zmienia się po każdej zmianie znaku różnicowego napięcia wejściowego. Tego rodzaju komparator ma budowę bardzo podobną do wzmacniacza operacyjnego. Drugi rodzaj to komparatory z wejściem zegarowym. Dokonują one operacji porównania napięć na wejściach w odpowiedzi na sygnał zegarowy. Taki komparator ma zwykle na wyjściu układ zapamiętujący wynik porównania - zero lub jedynkę - na czas trwania jedynki zegara. Ten rodzaj komparatorów ma szerokie zastosowanie m.in. w przetwornikach analogowo-cyfrowych, gdzie sygnał analogowy jest próbkowany w określonych chwilach czasowych. Przykład prostego układu komparatora z wejściem zegarowym pokazuje rys. 13.13. W tym układzie następuje porównanie napięć na wejściach $U_{we+}$ i $U_{we-}$ w chwili, gdy stan logiczny wejścia zegarowego CLK zmienia się z "0" na "1", a wynik porównania (stan logiczny "0" lub "1" na wyjściu $U_{wy}$) utrzymuje się aż do końca czasu trwania stanu "1" zegara. W czasie, gdy zegar jest w stanie "0", napięcie na wyjściu nie reprezentuje wyniku porównania.




Rys. 13.13. Prosty komparator napięcia z wejściem zegarowym

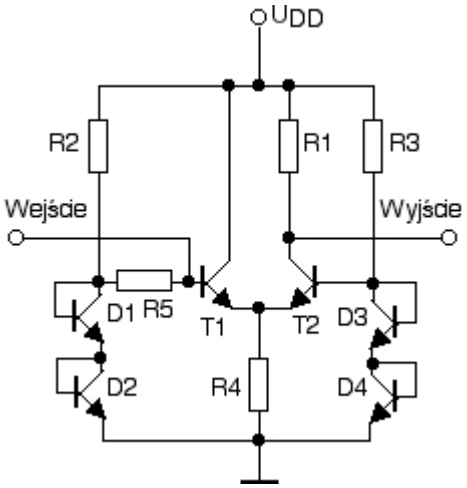
Układ ten składa się z dwóch części. Po lewej tranzystory T1 - T5 tworzą wzmacniacz różnicowy z symetrycznym wyjściem, w którym tranzystor T5 pełni rolę źródła prądowego (dającego prąd o wartości określonej przez napięcie polaryzujące $U_{P2}$), a tranzystory T3 i T4 pełnią rolę aktywnych obciążeń (napięcie polaryzacji $U_{P1}$ musi mieć taką wartość, aby zapewnić pracę tych tranzystorów w zakresie nasycenia). Prawa część układu to przerzutnik zbudowany z tranzystorów T8, T9, T11 i T12, ma on dwa stany stabilne podobnie jak omawiany wcześniej przerzutnik zbudowany z dwóch inwerterów (patrz wykład 8, rys. 8.5). Tranzystor T10 sterowany sygnałem zegara CLK włącza zasilanie tego przerzutnika, gdy zegar jest w stanie "1", i wyłącza, gdy zegar jest w stanie "0". Dla zegara w stanie "0" napięcie wyjściowe $U_{wy}$ zależy od wartości napięć na bramkach tranzystorów T12 i T7 (tranzystor T10 jest wyłączony, więc przez tranzystory T8 i T9 prąd nie płynie). Porównanie napięć $U_{we+}$ i $U_{we-}$ następuje z chwilą przejścia zegara do stanu "1". Tranzystor T10 zostaje włączony, włączając zasilanie przerzutnika. Napięcia na bramkach tranzystorów T6 i T7 decydują o tym, w którym z dwóch możliwych stanów ustawi się w tym momencie przerzutnik. Dla $U_{we+} > U_{we-}$ napięcie na bramce T7 będzie niższe, niż na bramce T6, co spowoduje, że napięcie w węźle B będzie wyższe, niż w węźle A. Dodatnie sprzężenie zwrotne w przerzutniku spowoduje dalszy wzrost napięcia w węźle B i spadek w węźle A, i ostatecznie na wyjściu pojawi się napięcie

bliskie $U_{DD}$, czyli logiczna jedynka. W przeciwnym przypadku na wyjściu pojawi się napięcie bliskie zeru, czyli logiczne zero. Ten stan nie ulegnie już zmianie bez względu na zmiany napięć wejściowych aż do chwili, gdy zegar przejdzie ze stanu "1" do stanu "0". Ponowne przejście zegara do stanu "1" spowoduje kolejny akt porównania napięć wejściowych i zapamiętanie wyniku.

Pokażemy teraz jeszcze inne zastosowanie wzmacniacza różnicowego.

**Wzmacniacz szerokopasmowy**

Wzmacniacz różnicowy można z powodzeniem zastosować jako szerokopasmowy stopień wzmacniający. Szczególnie nadają się do tego wzmacniacze bipolarne ze względu na większe wartości transkonduktancji, co ułatwia uzyskanie dużego wzmocnienia. Układy z tranzystorami bipolarnymi pozwalają także osiągnąć wyższe maksymalne częstotliwości pracy takich układów. Przykład prostego stopnia wzmacniającego z tranzystorami bipolarnymi pokazuje rys. 13.14.




Rys.13.14. Wzmacniacz różnicowy jako wzmacniacz szerokopasmowy

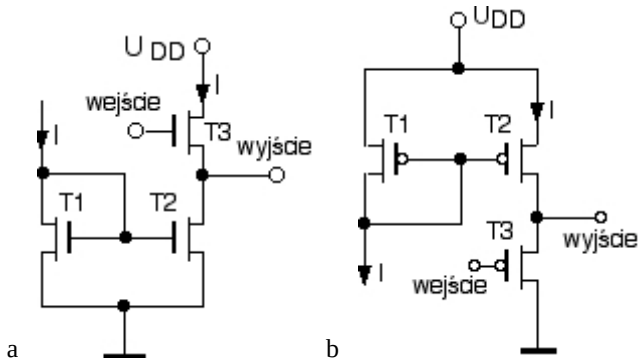
Omawiany układ jest - z punktu widzenia składowych stałych - typowym wzmacniaczem różnicowym, uproszczonym o tyle, że zamiast źródła prądowego zastosowany jest rezystor R4. Tranzystory w połączeniu diodowym D1 - D4 określają napięcia na bazach tranzystorów T1 i T2 i tym samym, wraz z rezystorem R4, określają wartości prądów kolektora. Dla sygnałów zmiennych układ ma asymetryczne wejście i asymetryczne wyjście. Polaryzowane w kierunku przewodzenia diody D3 i D4 stanowią dla sygnałów zmiennych o małej amplitudzie zwarcie, toteż dla takich sygnałów bazę T2 można uważać za uziemioną. Uziemiony dla małych sygnałów jest także (poprzez źródło zasilania) kolektor T1. Zatem dla małych sygnałów układ działa jako dwustopniowy wzmacniacz, w którym pierwszy stopień pracuje w układzie wspólnego kolektora (innymi słowy - jako wtórnik emiterowy), a drugi stopień pracuje w układzie wspólnej bazy. Takie połączenie ma bardzo korzystne właściwości - sprzężenie zwrotne z wyjścia na wejście, które mogłoby być przyczyną niestabilnej pracy układu przy dużych częstotliwościach, w tym układzie praktycznie nie istnieje. Kilka takich stopni połączonych układami przesuwania poziomu składowej stałej może stanowić kompletny wzmacniacz szerokopasmowy. Dodanie na wejściu i wyjściu takiego wzmacniacza obwodów rezonansowych lub filtrów pozwala zastosować całość jako wzmacniacz selektywny. Takie układy są często stosowane np. w sprzęcie radioodbiorczym, telewizyjnym itp. Omawiany układjest szczególnie przydatny przy odbiorze sygnałów z modulacją częstotliwości (FM). Jeśli amplituda sygnału wejściowego jest dostatecznie duża, układ przejawia swoje właściwości wzmacniacza różnicowego - działa jako ogranicznik amplitudy (patrz charakterystyki przejściowe, rys. 13.7). Pozwala to usunąć z sygnału FM modulację amplitudy, która - jeśli występuje - jest sygnałem zakłócającym i pogarsza jakość odbioru.

Podobny układ można też zbudować na tranzystorach MOS. Ze względu na małą transkonduktancję tych tranzystorów trzeba użyć obciążenia aktywnego, co komplikuje układ, ale zasada pozostaje ta sama.

## 13.4. Stopnie wyjściowe

Jeżeli wzmacniacz lub inny układ analogowy ma dostarczać sygnał do obciążenia o małej rezystancji (np. akustyczny wzmacniacz mocy do głośnika o rezystancji wynoszącej kilka omów), to na wyjściu tego układu konieczny jest stopień, który może dostarczyć do obciążenia prąd o odpowiednio dużym natężeniu. Poznamy teraz przykłady takich stopni.

Wtórnik źródłowy już poznaliśmy - jako układ służący do przesuwania poziomu składowej stałej (rys. 12.19, który tutaj dla wygody powtórzymy jako rys. 13.15), a także jako jeden z możliwych stopni wyjściowych wzmacniacza operacyjnego (rys. 13.12).




Rys. 13.15. Układy wtórników źródłowych: (a) z tranzystorami nMOS, (b) z tranzystorami pMOS

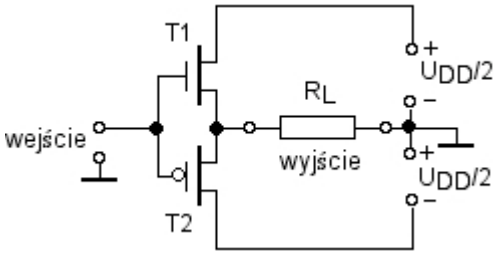
Tutaj dodamy kilka szczegółów interesujących z punktu widzenia działania układów wtórników jako stopni wyjściowych. Jeżeli wtórnik obciążony jest zewnętrzną rezystancją $R_L$, to jego wzmocnienie napięciowe opisuje zależność

$$|k_u| = \frac{g_{m3}}{g_{m3} + g_{ds3} + g_{ds2} + R_L^{-1}} \tag{13.22}$$

Z zależności tej wynika, że wtórnik ma wzmocnienie bliskie 1 tylko wtedy, gdy $g_{m3} >> g_{ds3}+g_{ds2}+R_L^{-1}$. Zatem, znając rezystancję obciążenia $R_L$, należy tak zaprojektować układ wtórnika, aby transkonduktancja $g_{m3}$ była dostatecznie duża.

Układ wtórnika ma tę niezbyt korzystną cechę, że źródło tranzystora T3 jest dołączone do wyjścia, a nie do masy lub zasilania. To oznacza, że w ogólnym przypadku istnieje niezerowe napięcie polaryzacji podłoża względem źródła $U_{BS}$. To napięcie ma wpływ na napięcie progowe tranzystora (wzór 4.5). Napięcie $U_{BS}$, wpływając na $U_T$, pośrednio wpływa więc i na prąd drenu oraz napięcie wyjściowe. Jest to efekt niekorzystny, zmniejszający wzmocnienie napięciowe, a przy dużej amplitudzie sygnału wprowadzający dodatkowo zniekształcenia nieliniowe. W przypadku tranzystora nMOS wykonanego w podłożu układu efekt ten jest nie do uniknięcia. W przypadku tranzystora pMOS można zastosować nietypowe rozwiązanie polegające na wykonaniu dla tranzystora T3 odrębnej wyspy i dołączeniu jej do wyjścia, a nie do napięcia zasilania $U_{DD}$. Wyspa jest w takim przypadku nadal prawidłowo (tj. zaporowo) spolaryzowana względem podłoża, ale jej potencjał jest równy potencjałowi źródła tranzystora T3, co likwiduje efekt zmiany napięcia progowego. Z tego powodu układ z tranzystorami pMOS (rys. 13.15b) jest korzystniejszy od układu z tranzystorami nMOS (rys. 13.15a).

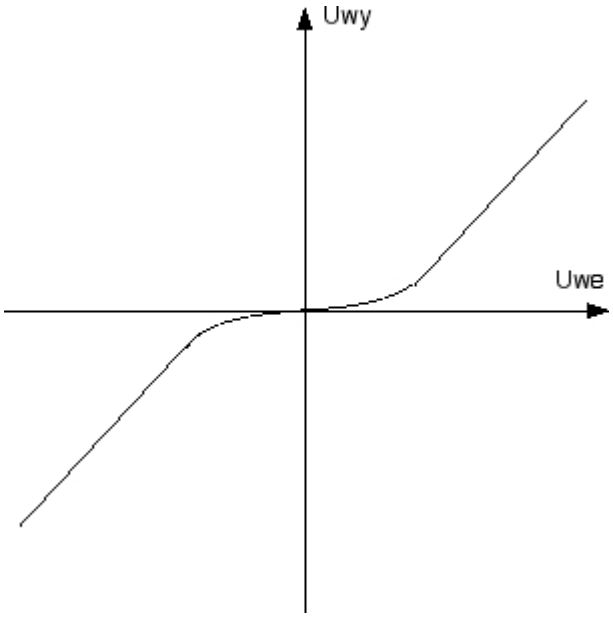
Omówiony wyżej układ wtórnika w praktyce wystarcza, jeśli rezystancja obciążenia jest rzędu kiloomów lub większa. Dla mniejszych rezystancji obciążających (lub gdy układ obciążony jest znaczną pojemnością) potrzebne są inne układy. Stosowane są wtedy stopnie w układzie szeregowo-przeciwsobnym. Zasadę budowy takiego układu, w wersji CMOS, pokazuje rys. 13.16.




Rys. 13.16. Zasada budowy układu szeregowo-przeciwsobnego

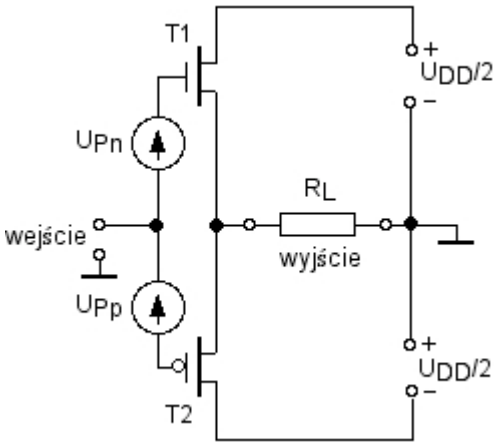
W tym układzie przy dodatnim półokresie sygnału sterującego prąd tranzystora T2 wzrasta, a T1 - maleje. Różnica tych prądów stanowi prąd wyjściowy płynący przez zewnętrzną rezystancję $R_L$. Przy ujemnym półokresie wzrasta prąd T1, a maleje - T2. Oba tranzystory pracują jako wtórniki, których rezystancją obciążającą jest $R_L$. Należy zwrócić uwagę na to, że w tym układzie umownym węzłem odniesienia dla wzmacnianego sygnału jest węzeł łączący dwa źródła zasilania, a nie - jak to zwykle bywa - "minus" zasilania, czyli podłoże układu. Taki sam stopień można zbudować z tranzystorów bipolarnych, zastępując tranzystor nMOS tranzystorem n-p-n, a tranzystor pMOS - tranzystorem p-n-p. Działanie układu będzie analogiczne.

Układ zbudowany tak, jak na rysunku, ma poważną wadę - jego charakterystyka przejściowa wykazuje silną nieliniowość dla napięć wejściowych w okolicy zera. Wynika to z faktu, że aby przez tranzystor T1 płynął prąd, napięcie na jego bramce musi być wyższe od jego napięcia progowego, a aby płynął prąd przez tranzystor T2, napięcie na jego bramce musi spaść poniżej jego napięcia progowego. Istnieje więc taki (dość szeroki!) zakres napięć wejściowych, dla których żaden tranzystor nie przewodzi, a więc prąd przez rezystancję obciążenia nie płynie i napięcie na niej wynosi zero. Charakterystykę przejściową układu ilustrującą to zjawisko pokazuje rys. 13.17.

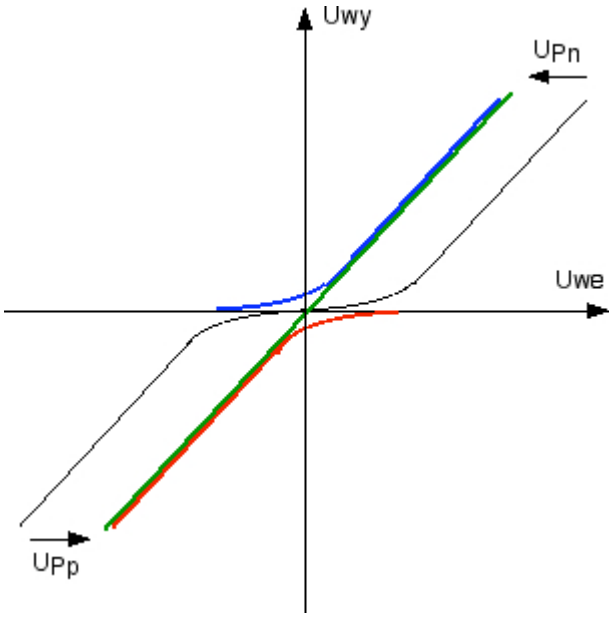



Rys. 13.17. Charakterystyka przejściowa układu pokazanego na rys. 13.16

Tę wadę można usunąć wprowadzając dodatkowe źródła napięcia wstępnie polaryzujące bramki tranzystorów (rys. 13.18). Napięcia te sumują się z napięciem wejściowym. Na charakterystyce odpowiada to przesunięciu górnej połówki charakterystyki w lewo, a dolnej w prawo, i w rezultacie otrzymuje się wypadkową charakterystykę przejściową, która jest liniowa lub bardzo bliska liniowej - jak na rys. 13.19.

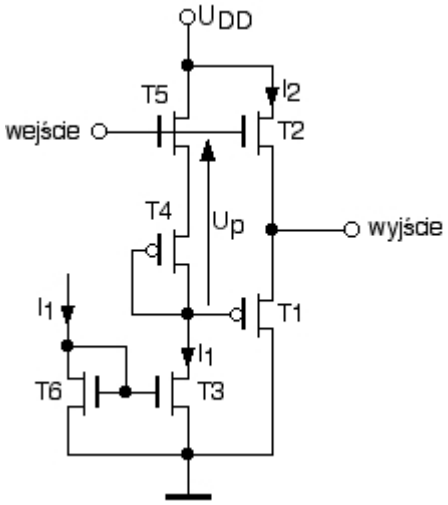



Rys. 13.18. Układ szeregowo-przeciwsobny, zasada wstępnej polaryzacji bramek tranzystorów




Rys. 13.19. Charakterystyka przejściowa układu pokazanego na rys. 13.18

Zwróćmy uwagę, że w obwodzie złożonym z obu źródeł zasilania oraz połączonych szeregowo tranzystorów T1 i T2 płynie pewien stały prąd (zwany prądem spoczynkowym) niezależnie od tego, czy układ jest wysterowany jakimkolwiek sygnałem wejściowym. Gdyby napięcia $U_{Pn}$ i $U_{Pp}$ miały wartości niezależne od temperatury, to prąd spoczynkowy zmieniałby się z temperaturą: malał w przypadku tranzystorów MOS, rósł (i to bardzo szybko) w przypadku tranzystorów bipolarnych. Zatem istotne jest, w jaki sposób wytworzone zostaną napięcia $U_{Pn}$ i $U_{Pp}$. Zależy od tego stabilność pracy układu przy zmianach temperatury. Będzie to pokazane na praktycznych przykładach. Oto praktyczny układ szeregowo-przeciwsobny w wersji CMOS, w którym napięciami $U_{Pn}$ i $U_{Pp}$ są napięcia $U_{GS}$ tranzystorów T4 i T5:




Rys. 13.20. Stopień wyjściowy CMOS w układzie szeregowo-przeciwsobnym - układ praktyczny

W tym układzie sygnał z wejścia steruje bezpośrednio bramką T2, zaś tranzystor T1 jest sterowany z wyjścia wtórnika, jaki tworzy tranzystor T5 wraz ze źródłem prądowym (T3-T6). Dodatkowo tranzystor T5 wraz z tranzystorem T4 (w połączeniu diodowym) stanowią układ przesuwania poziomu składowej stałej zapewniający odpowiednią wartość składowej stałej napięcia na bramce tranzystora T1.

Wartość napięcia stałego $U_p$ decyduje o wartości prądu spoczynkowego $I_2$ tranzystorów wyjściowych T1 i T2. Napięcie to jest sumą $U_{GST5}$ i $U_{GST4}$, a zarazem $U_{GST1}$ i $U_{GST2}$. Zakładając, że wszystkie tranzystory pracują w zakresie nasycenia, i rozpisując odpowiednie wyrażenia dla napięć można pokazać, że jeśli spełnione są warunki

$$\frac{\frac{W_1}{L_1}}{\frac{W_4}{L_4}} = \frac{\frac{W_2}{L_2}}{\frac{W_5}{L_5}} = m \tag{13.23}$$

to prąd $I_2$ ma wartość

$$I_2 = mI_1 \tag{13.24}$$

Można pokazać, że przy właściwie dobranym napięciu $U_p$ charakterystyka przejściowa układu jest liniowa, jeśli dodatkowo spełniony jest warunek jednakowych wydajności prądowych obu tranzystorów, czyli

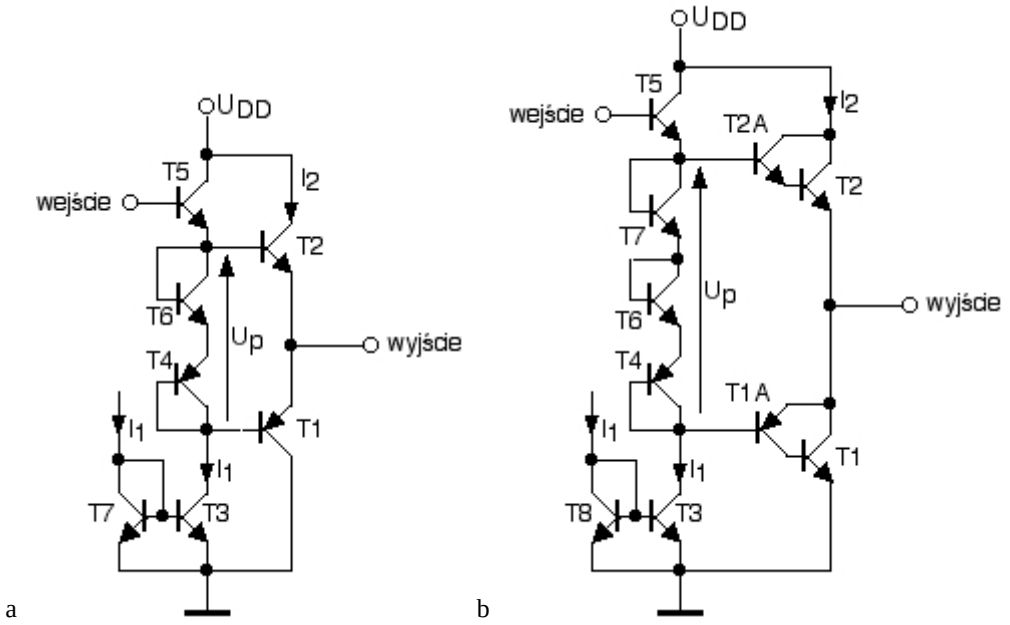
$$\frac{\left(\frac{W}{L}\right)_{T1}}{\left(\frac{W}{L}\right)_{T2}} = \frac{\mu_n}{\mu_p} \tag{13.25}$$

Charakterystyka ta jest jednak liniowa tylko tak długo, jak długo oba tranzystory przewodzą. Gdy jeden z nich zostaje wyłączony, prąd do obciążenia dostarcza tylko drugi, a to oznacza, że charakterystyka staje się nieliniowa (zależność prądu drenu od napięcia bramki jest, jak pamiętamy, w stanie nasycenia funkcją kwadratową).

Konduktancja wyjściowa układu wynosi

$$g_{wy} = g_{m1} + g_{m2} = r_{wy}^{-1} \tag{13.26}$$

W praktyce rezystancja wyjściowa $r_{wy}$ ma typowo wartość rzędu kilkuset omów. To oznacza, że jeśli potrzebny jest stopień wyjściowy sterujący małą rezystancję i dostarczający prąd wyjściowy rzędu setek miliamperów lub kilku amperów, to zwykłe układy CMOS nie są przydatne. Stosuje się wtedy zwykle układy bipolarne. Budowa bipolarnego stopnia wyjściowego jest podobna do budowy układu CMOS. Przykład pokazany jest na rys. 13.21.




Rys. 13.21. Bipolarne wyjściowe stopnie przeciwsobne: (a) zasada budowy, (b) układ praktyczny

Stopień pokazany na rys. 13.21a nie różni się zasadą działania od stopnia CMOS, z tym wyjątkiem, że napięcie polaryzacji $U_p$ otrzymuje się jako spadek napięcia na dwóch tranzystorach w połączeniu diodowym. Jednak w układach bipolarnych tranzystory p-n-p nie są komplementarne do tranzystorów n-p-n i nie nadają się do pracy przy dużych prądach. Dlatego w praktyce stosuje się układ, którego zasadę budowy ilustruje rys. 13.21b. Wyjściowy tranzystor p-n-p jest zastąpiony połączeniem tranzystora p-n-p (T1A) z tranzystorem n-p-n (T1). Ten ostatni jest taki sam, jak tranzystor T2. Połączenie T1A-T1 zachowuje się jak tranzystor p-n-p, ale prąd płynący przez tranzystor T1A jest prądem bazy tranzystora T1, a więc jest $h_{FE}$-krotnie mniejszy od prądu płynącego przez tranzystor T1. Dla lepszej symetrii układu także tranzystor T2 jest zastąpiony przez połączenie T2A-T2. Dla wytworzenia właściwego napięcia polaryzującego $U_p$ trzeba połączyć szeregowo trzy, a nie dwa tranzystory w

połączeniu diodowym. Wartość prądu spoczynkowego $I_2$ określa się przez dobór odpowiedniej proporcji powierzchni złącz emiterowych tranzystorów T1, T2 i tranzystorów T4, T6 i T7.

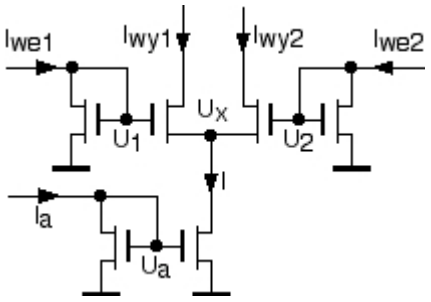
Na zasadzie zilustrowanej rysunkiem 13.21 buduje się stopnie wyjściowe do wielu zastosowań, na przykład do popularnych scalonych wzmacniaczy dużej mocy małej częstotliwości stosowanych w sprzęcie elektroakustycznym. Można osiągnąć moce wyjściowe sięgające dziesiątków W. Powstają jednak wtedy nowe problemy konstrukcyjne. Przy znacznej mocy wydzielanej w układzie jego temperatura może przekraczać temperaturę otoczenia o kilkadziesiąt $^{o}$C. W tej sytuacji niezwykle ważne jest zapewnienie stabilności cieplnej układu. Gdyby napięcie polaryzujące $U_p$ było niezależne od temperatury, to prąd spoczynkowy płynący przez tranzystory wyjściowe wzrastałby wykładniczo (czyli bardzo szybko) ze wzrostem temperatury tych tranzystorów, a jego wzrost powodowałby dalsze samopodgrzewanie się tranzystorów. Występujące tu elektryczno-cieplne dodatnie sprzężenie zwrotne mogłoby łatwo doprowadzić do zniszczenia tranzystorów wyjściowych na skutek nieograniczonego wzrostu prądu $I_2$. Zapobiega temu polaryzacja napięciem, które odkłada się na tranzystorach T4, T6, T7. Jeżeli prąd źródła prądowego $I_1$ mało zależy od temperatury, to napięcie $U_p$ maleje z temperaturą i dodatnie sprzężenie elektryczno-cieplne nie może wystąpić. Jednak spełnione musi być wymaganie, że temperatura wszystkich elementów: T1, T2, T1A, T2A, T4, T6, T7 jest jednakowa.

Do zagadnień cieplnych w układach scalonych wrócimy jeszcze w następnym wykładzie.

## 13.5. Wybrane układy nieliniowe

Układy, które na wyjściu dają napięcie (lub prąd) proporcjonalne do iloczynu dwóch napięć (lub prądów) wejściowych, nazywane są **analogowymi układami mnożącymi**. Znajdują one liczne zastosowania i stanowią jeden z najważniejszych rodzajów układów wykonujących operacje nieliniowe na sygnałach analogowych.

Zasada działania klasycznego układu mnożącego zwanego układem Gilberta polega na wykorzystaniu wykładniczej zależności prądu od napięcia w tranzystorach bipolarnych lub w tranzystorach MOS w zakresie podprogowym. Zasadę działania układu mnożącego Gilberta ilustruje rys. 13.22:




Rys. 13.22. Ilustracja zasady działania układu mnożącego Gilberta

Dla wyjaśnienia zasady działania tego układu przypomnimy zależność prądu drenu od napięcia bramki tranzystora MOS w zakresie podprogowym (wzór 4.8), którą przepiszemy tu w postaci skróconej

$$I_D = I_0 \exp\left(\frac{qU_{GS}}{nkT}\right) \tag{13.27}$$

gdzie $I_0$ jest pewną stałą. Stąd

$$I_{we1} = I_0 \exp\left(\frac{qU_1}{nkT}\right) \tag{13.28a}$$

$$I_{we2} = I_0 \exp\left(\frac{qU_2}{nkT}\right) \tag{13.28b}$$

oraz

$$I_{wy1} = I_0 \exp\left[\frac{q(U_1 - U_x)}{nkT}\right] \tag{13.29a}$$

$$I_{wy2} = I_0 \exp\left[\frac{q(U_2 - U_x)}{nkT}\right] \tag{13.29b}$$

a równocześnie

$$I = I_a = I_{wy1} + I_{wy2} = I_0 \exp\left(\frac{qU_a}{nkT}\right) \tag{13.30}$$

Stąd łatwo otrzymać

$$I = I_0 \exp\left(-\frac{qU_x}{nkT}\right)\left[\exp\left(\frac{qU_1}{nkT}\right) + \exp\left(\frac{qU_2}{nkT}\right)\right] \tag{13.31}$$

oraz

$$I_{wy1} = I_0 \frac{\exp\left(\frac{qU_a}{nkT}\right)\exp\left(\frac{qU_1}{nkT}\right)}{\exp\left(\frac{qU_1}{nkT}\right)+\exp\left(\frac{qU_2}{nkT}\right)} \tag{13.32a}$$

$$I_{wy2} = I_0 \frac{\exp\left(\frac{qU_a}{nkT}\right)\exp\left(\frac{qU_2}{nkT}\right)}{\exp\left(\frac{qU_1}{nkT}\right)+\exp\left(\frac{qU_2}{nkT}\right)} \tag{13.32b}$$

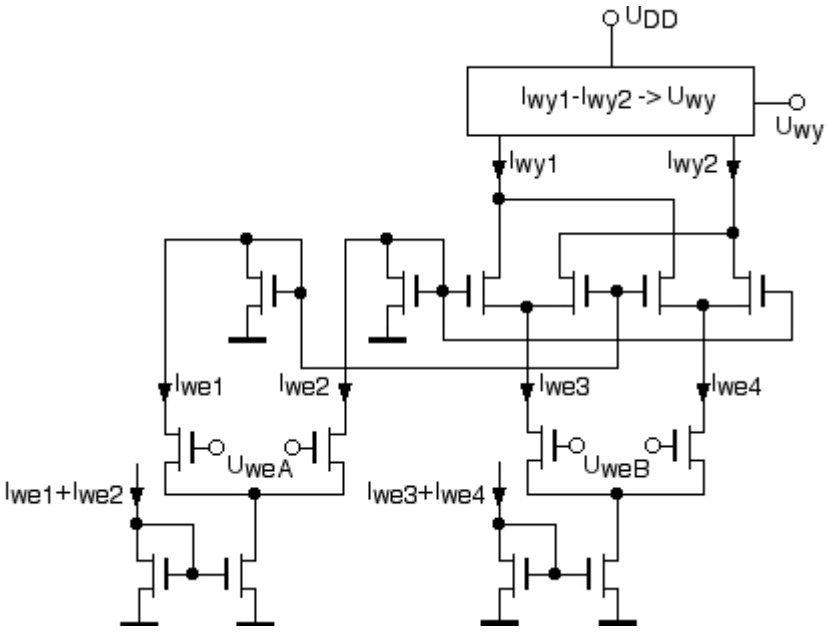
Z zależności (13.32a) i (13.32b) można otrzymać od razu

$$I_{wy1} = \frac{I_a I_{we1}}{I_{we1}+I_{we2}}$$
$$I_{wy2} = \frac{I_a I_{we2}}{I_{we1}+I_{we2}} \tag{13.33}$$

Widać więc, że układ mnoży prądy $I_{we1}$ oraz $I_{we2}$ przez $I_a$, pod dodatkowym warunkiem, że suma $I_{we1}$+$I_{we2}$ pozostaje stała.

Rys. 13.22 ilustruje zasadę działania układu, ale nie jest kompletnym układem. Kompletny układ składa się z dwóch połączonych ze sobą układów takich, jak na rys. 13.22 - patrz rys. 13.23. Rozumując podobnie, jak wyżej, można pokazać, że w tym układzie zachodzi zależność:

$$I_{wy1} - I_{wy2} = \frac{(I_{we1} - I_{we2})(I_{we3} - I_{we4})}{(I_{we3} + I_{we4})} \tag{13.34}$$




Rys. 13.23. Pełny układ mnożący Gilberta

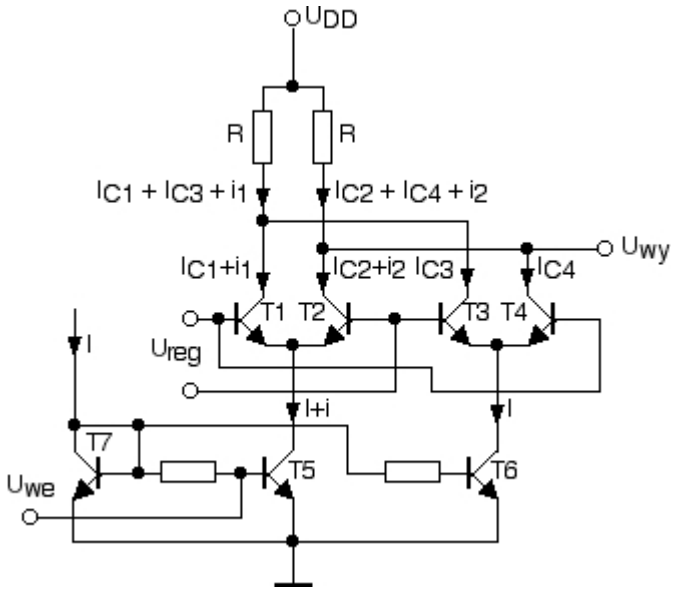
W tym układzie dolne źródła prądowe zapewniają stałość sum prądów (w szczególności $I_{we3}$+$I_{we4}$), zaś wzmacniacze różnicowe, do których wejść doprowadzone są napięcia $U_{weA}$ i $U_{weB}$, służą jako liniowe przetworniki napięcie->prąd, dając na wyjściach różnice prądów $I_{we1}$-$I_{we2}$ oraz $I_{we3}$-$I_{we4}$. Dzięki temu różnica prądów $I_{wy1}$-$I_{wy2}$ jest proporcjonalna do iloczynu $U_{weA}U_{weB}$. Napięcie wyjściowe powstaje w układzie (nie pokazanym w szczegółach) przetwarzającym liniowo różnicę prądów $I_{wy1}$-$I_{wy2}$ na napięcie $U_{wy}$.

Podobny układ można zbudować z tranzystorów bipolarnych (taki był oryginalny układ mnożący Gilberta).

Układy mnożące mają liczne zastosowania. Są używane do realizacji takich funkcji, jak modulacja i demodulacja amplitudy i fazy, przemiana częstotliwości, wytwarzanie sygnału proporcjonalnego do mocy sygnału elektrycznego itp.

Ciekawym zastosowaniem wzmacniczy różnicowych są **układy o regulowanym wzmocnieniu.**

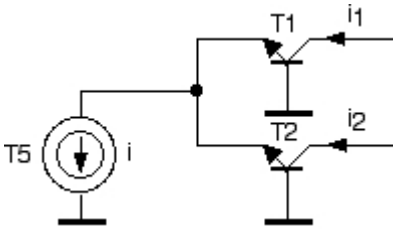
Każdy wzmacniacz różnicowy jest układem o wzmocnieniu napięciowym, którego wielkość można regulować na drodze elektrycznej przez zmianę wartości prądu zasilającego. W przypadku układu bipolarnego wzmocnienie jest wprost proporcjonalne do wartości tego prądu (wzór 13.14), a w przypadku układu MOS - do pierwiastka z tego prądu (wzór 13.16). Możliwy do uzyskania zakres regulacji jest wąski, bowiem prądu zasilającego z różnych względów nie można zmieniać w bardzo szerokim zakresie. Wzmacniacz różnicowy może być jednak użyty w inny sposób - rys. 13.24.




Rys. 13.24. Układ regulacji amplitudy sygnału zmiennego

Układ jest zbudowany z dwóch wzmacniaczy różnicowych zasilanych jednakowymi prądami I przez układ źródeł prądowych (tranzystory T5 - T7). Napięcie $U_{reg}$ doprowadzone jest do wejść obu wzmacniaczy różnicowych, z tym, że wejścia te są skrzyżowane, tj. gdy napięcie na bazie T1 wzrasta, to na bazie T3 maleje, i podobnie z tranzystorami T2 i T4. Wyjścia obu wzmacniaczy są połączone równolegle, toteż gdy prąd $I_{C1}$ maleje, to prąd $I_{C3}$ wzrasta, a ich suma pozostaje stała, i podobnie z prądami $I_{C2}$ i $I_{C4}$. W rezultacie prądy kolektorów mogą się zmieniać w bardzo szerokim zakresie, a mimo to napięcia na kolektorach tranzystorów T1 - T4 pozostają stałe. Sygnał zmienny, którego amplituda podlega regulacji, jest doprowadzony do bazy tranzystora T5, i powoduje powstanie składowej zmiennej $i$ prądu kolektora tranzystora T5. Zasada działania układu polega na tym, że ta składowa zmienna rozdziela się pomiędzy tranzystory T1 i T2 w stopniu zależnym od znaku i wartości napięcia $U_{reg}$.

Podział składowej zmiennej $i$ pomiędzy tranzystory T1 i T2 zależy od ich konduktancji wejściowych $g_{we}$ w układzie wspólnej bazy, bowiem połączenie tranzystorów T5, T1 i T2 można dla składowych zmiennych przedstawić w uproszczeniu następująco:




Rys. 13.25. Rozpływ składowych zmiennych między tranzystory T1 i T2. Źródło prądowe symbolizuje tranzystor T5.
Zakładamy, że bazy tranzystorów są dla składowych zmiennych uziemione przez źródło $U_{reg}$ (patrz rys. 13.24)

Składowe zmienne $i_1$ i $i_2$ są wprost proporcjonalne do konduktancji wejściowych $g_{weT1}$ i $g_{weT2}$, przy czym suma $i_1$ i

$i_2$ jest oczywiście równa $i$. Konduktancje wejściowe *w układzie wspólnej bazy* można wyrazić następująco (wykorzystując zależność (4.24)):

$$g_{we} = \frac{\partial I_E}{\partial U_{BE}} \approx \frac{\partial I_C}{\partial U_{BE}} = g_m = \frac{qI_C}{kT} \tag{13.35}$$

a stąd widać, że *składowe zmienne* $i_1$ i $i_2$ *są wprost proporcjonalne do składowych stałych* $I_{C1}$ i $I_{C2}$. Stąd, wykorzystując podstawową zależność (4.20), łatwo otrzymać:

$$i_2 = \frac{i}{\exp\left(\frac{qU_{reg}}{kT}\right)+1} \tag{13.36}$$

gdzie $U_{reg} = U_{BE1} - U_{BE2}$. Składowa zmienna napięcia wyjściowego $u_{wy}$ jest równa $i_2R$. Dzięki wykładniczej zależności składowej zmiennej $i_2$ od napięcia $U_{reg}$ możliwa jest regulacja amplitudy w bardzo szerokim zakresie (wiele dekad), i to przy niewielkich zmianach napięcia $U_{reg}$. Na tej zasadzie działają m.in. układy regulacji głośności w sprzęcie elektroakustycznym.

Podobny układ można zbudować na tranzystorach MOS *pracujących w zakresie podprogowym*, gdzie obserwuje się wykładniczą zależność prądu drenu od napięcia bramki, a konduktancja wejściowa w układzie wspólnej bramki jest dana zależnością (4.18) podobną do zależności (4.24).

# ZADANIA DO WYKŁADU 13

**Dane do zadań:**

Technologia CMOS: $U_{Tn}$ = 0,75 V, $U_{Tp}$ = - 0,85 V, $\mu_n C_{ox}$ = 80 μA/V$^2$, $\mu_p C_{ox}$ = 27 μA/V$^2$. Wartość stałej $n$ we wzorze (4.8) opisującym charakterystyki w zakresie podprogowym: $n$ = 1,5. Minimalna długość kanału 0,7 μm, minimalna szerokość kanału 1,0 μm. Minimalna szerokość ścież ki polikrzemowej 1,0 μm. Rezystancja warstwowa polikrzemu: 20 Ω/□. Standardowe napięcie zasilania 5 V.

**Zadanie 1**

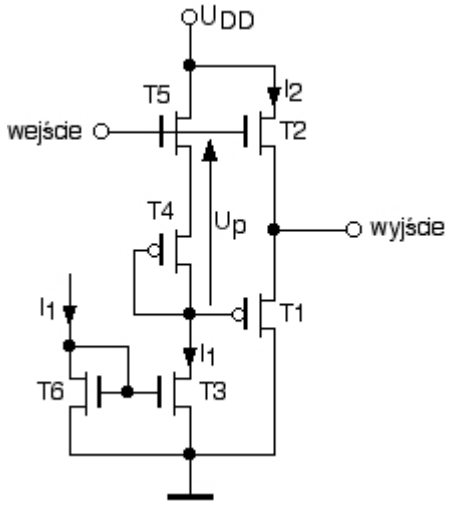
Oceń, jaką maksymalną częstotliwość graniczną $f_T$ można uzyskać w układzie wzmacniacza z aktywnym obciążeniem (wzór (13.10)), jeśli całkowite obciążenie pojemnościowe wzmacniacza (suma $C_1$+$C_2$+pojemność dołączona z zewnątrz) wynosi: (a) 0,01 pF, (b) 1 pF. Wykonaj obliczenia dla tranzystorów nMOS i pMOS, dla maksymalnej szerokości kanału równej 10 μm.

**Zadanie 2**

Oceń maksymalne wzmocnienie napięciowe dla bardzo małych częstotliwości w układzie wzmacniacza z aktywnym obciążeniem dla takich tranzystorów, dla jakich były wykonane obliczenia w poprzednim zadaniu. Przyjmij wartość parametru $\lambda$ równą 0,1 V$^{-1}$ dla obu rodzajów tranzystorów.

**Zadanie 3**

Oszacuj prąd $I_2$ w stopniu szeregowo-przeciwsobnym (rys. 13.23) potrzebny do tego, aby rezystancja wyjściowa stopnia była mniejsza, niż: (a) 10 kΩ, (b) 500 Ω, (c) 5 Ω. Co sądzisz o możliwości uzyskania tych wartości prądu w realnym układzie?




Rys. 13.23. Stopień wyjściowy szeregowo-przeciwsobny

## Bibliografia

[1] F. Maloberti, *"Analog design for CMOS VLSI systems*", Kluwer Academic Publishers, 2001
[2] H. Camenzind, "*Designing Analog Chips*", książka dostępna w internecie:
http://www.designinganalogchips.com/
[3] A. B. Grebene, *"Bipolar and MOS analog integrated circuit design*", John Wiley & Sons, Inc. 1984

# Wykład 14: Pobór mocy układów scalonych

## Wstęp

Tematem wykładu 14 jest pobór mocy w układach scalonych i związane z tym problemy odprowadzania ciepła.

Jest kilka powodów, dla których problemom tym poświęca się bardzo wiele uwagi. Po pierwsze, pobór mocy przez układy i systemy scalone wraz ze wzrostem stopnia ich złożoności i szybkości działania osiągnął dziś poziom, który uniemożliwia dalszy wzrost szybkości i ogranicza wzrost złożoności. Po drugie, coraz więcej urządzeń elektronicznych to urządzenia przenośne, o bateryjnym zasilaniu, naturalne jest więc dążenie do ograniczenia poboru mocy przez te urządzenia. Po trzecie, duży pobór mocy jest zwykle równoznaczny z wysoką temperaturą pracy układu, co negatywnie wpływa na jego niezawodność.

W niektórych przypadkach bardzo mały pobór mocy jest podstawowym warunkiem użyteczności danego urządzenia, na przykład gdy jest to urządzenie elektromedyczne wszczepiane choremu człowiekowi. Takie urządzenie, na przykład stymulator serca, powinno działać przez wiele lat bez konieczności wymiany źródła zasilania, a równocześnie z oczywistych powodów nie może zawierać wielkiej i ciężkiej baterii. We wszczepialnych urządzeniach elektromedycznych wyczerpanie źródł a zasilania oznacza konieczność wszczepienia całego nowego urządzenia, a więc wykonania zabiegu chirurgicznego.

Warto też wspomnieć, że w krajach wysoko uprzemysłowionych systemy elektroniczne (komputery, internet, telefonia przewodowa i komórkowa, domowy sprzęt elektroniczny itp.) konsumują obecnie 10% - 15% produkowanej w tych krajach energii elektrycznej, i procent ten rośnie. Obliczono, że gdyby w krajach Unii Europejskiej udało się obniżyć zużycie energii przez systemy elektroniczne o 30%, to rocznie dałoby to oszczędność około 100 TWh energii. Jest to mniej więcej tyle, ile wynosi całkowite roczne zużycie energii elektrycznej w Holandii, i mniej więcej 2/3 energii elektrycznej produkowanej rocznie przez elektrownie węglowe w Polsce. Obniżanie poboru mocy przez systemy elektroniczne ma więc poważne znaczenie gospodarcze.

W wykładzie 14 omówione są mechanizmy poboru mocy przez układy scalone, sposoby obniżania poboru mocy, specyficzne problemy projektowania układów dużej mocy i wreszcie sposoby odprowadzania wydzielanej mocy, czyli chłodzenia układów.

## 14.1. Moc pobierana przez układy scalone

W pierwszej części wykładu zajmiemy się poborem mocy przez układy CMOS. Omówimy najpierw układy cyfrowe: pobór mocy statycznych bramek kombinacyjnych, na najprostszym przykładzie inwertera, ale zależności i wnioski będą odnosić się do dowolnych statycznych bramek CMOS. Na początku przypomnimy i rozszerzymy wiadomości z wykładu 7.

Przypomnijmy, że prąd, jaki pobiera ze źródła zasilania statyczny inwerter CMOS, ma dwie składowe: dynamiczną i statyczną. **Składowa dynamiczna** poboru prądu pojawia się, gdy zmieniają się stany logiczne. **Składowa statyczna** to prąd, jaki płynie w stanie ustalonym, gdy stany logiczne nie zmieniają się.

Składowa dynamiczna prądu ma dwa składniki. Pierwszy z nich związany jest z ładowaniem i rozładowywaniem pojemności obciążającej inwerter - moc związaną z tym składnikiem nazwiemy w skrócie **mocą przełączania**. Drugi składnik to prąd, który płynie w czasie przełączania z tego powodu, że istnieje taki zakres napięć wejściowych, dla których oba tranzystory inwertera równocześnie przewodzą. Moc związaną z tym składnikiem nazwiemy w skrócie **mocą jednoczesnego przewodzenia**.

Przypomnijmy (wykład 7), że przy każdej zmianie stanu powodującej naładowanie obciążającej inwerter pojemności $C_l$ do napięcia $U_{DD}$ ze źródła zasilania pobierana jest energia o wartości $E_c = C_l U_{DD}^2$. W każdym cyklu zmiany stanów na wyjściu "0"->"1"->"0" następuje jedno naładowanie i jedno rozładowanie. Jeżeli w ciągu sekundy cykli ładowanie-rozładowanie jest $f$, to moc $P_c$ pobierana ze źródła zasilania dana jest zależnością (7.10), którą tu dla wygody powtórzymy jako (14.1):

$$P_c = C_l U_{DD}^2 f \tag{14.1}$$

Drugą składową dynamicznego poboru prądu jest prąd, który płynie przez tranzystory bezpośrednio ze źródła zasilania do masy w okresie, gdy w czasie przełączania oba jednocześnie przewodzą. Gdyby czasy narastania i opadania sygnału na wejściu były równe zeru, pobór mocy związany z tym prądem także byłby równy zeru, bo odcinek czasu, w którym tranzystory równocześnie przewodzą, byłby nieskończenie krótki. Przy różnych od zera czasach $t_r$ i $t_f$ pobór mocy $P_j$ można w przybliżeniu oszacować z zależności (7.11), którą tu dla wygody powtórzymy jako (14.2):

$$P_j = I_{max} U_{DD} \frac{t_r + t_f}{2} f \tag{14.2}$$

gdzie $I_{max}$ jest szczytową wartością prądu płynącego w czasie przełączania przez przewodzące tranzystory.

Suma obu składowych dynamicznego poboru mocy jest, jak widać, wprost proporcjonalna do liczby zmian stanów logicznych w jednostce czasu $f$. W układach taktowanych zegarem oznacza to, że

**! dynamiczny pobór mocy jest proporcjonalny do częstotliwości zegara.**

Moc przełączania (14.1) rośnie proporcjonalnie do kwadratu napięcia zasilania oraz jest proporcjonalna do pojemności obciążającej. Moc jednoczesnego przewodzenia (14.2) jest proporcjonalna do napięcia zasilania w pierwszej potędze, do wartości maksymalnej prądu płynącego bezpośrednio przez tranzystory w czasie, gdy są one równocześnie w stanie przewodzenia, oraz do czasu, w którym podczas przełączania oba tranzystory równocześnie przewodzą. W typowych układach CMOS prądy równoczesnego przewodzenia mają małą wartość w porównaniu z prądami ładowania i rozładowywania pojemności obciążających, a czasy narastania i opadania sygnałów na wejściach bramek są krótkie. Zatem w dynamicznym poborze mocy dominuje moc przełączania. W uproszczonych rozważaniach moc równoczesnego przewodzenia często jest w ogóle pomijana. My jednak w następnej częsci tego wykładu zajmiemy się obiema składowymi.

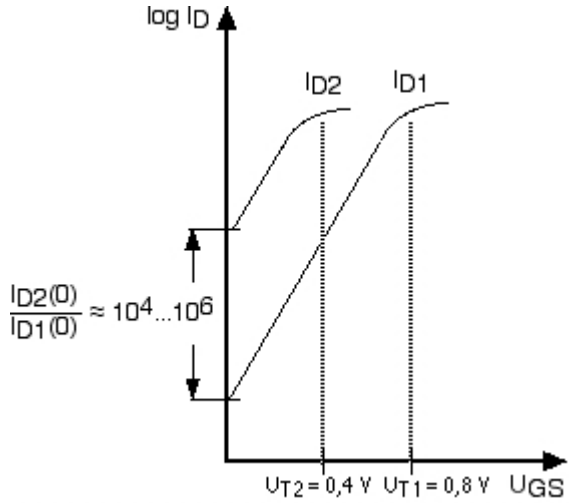
Składowa statyczna prądu pobieranego przez inwerter była w dawniejszych technologiach CMOS uważana za pomijalnie małą, bowiem zarówno w stanie "0" na wejściu, jak i w stanie "1" jeden z połączonych szeregowo tranzystorów - nMOS lub pMOS - jest wyłączony. Nie oznacza to jednak, że tranzystor wyłączony w ogóle nie przewodzi prądu. Jak wiemy (wykład 4), dla napięć bramki mniejszych od napięcia progowego płynie prąd drenu zwany prądem podprogowym. Opisuje go wzór (4.8), który tu dla wygody powtórzymy jako (14.3).

$$I_{Dp} = I_t \frac{W}{L} \exp\left[\frac{q(U_{GS} - U_T)}{nkT}\right]\left[1 - \exp\left(-\frac{qU_{DS}}{kT}\right)\right] \tag{14.3}$$

Dla napięcia bramki $U_{GS}$ = 0, czyli dla tranzystora wyłączonego, oraz dla $U_{DS} >> kT/q$ wzór (14.3) można uprościć do postaci

$$I_{Dp} = I_t \frac{W}{L} \exp\left(-\frac{qU_T}{nkT}\right) \tag{14.4}$$

W starszych technologiach prąd podprogowy wyłączonego tranzystora był tak mały, że można go było zaniedbać. Jednak w technologiach, w których długość bramki jest rzędu 100 nm i mniej, napięcia progowe tranzystorów są znacznie mniejsze, niż w starszych technologiach. Skracanie kanału tranzystora zmusza do zmniejszania napięcia zasilania układu, a przy niższym napięciu zasilania niższe musi być również napięcie progowe (wrócimy do tego zagadnienia w następnym wykładzie). Posługując się wzorem (14.4) można pokazać, że prąd podprogowy rośnie o rząd wielkości, gdy napięcie progowe maleje o wartość równą $2{,}3nkT/q$. W temperaturze otoczenia wartość ta wynosi od 60 mV (dla $n$=1) do 90 mV (dla $n$=1,5). Stąd łatwo policzyć, że zmniejszenie napięcia progowego z wartości 0,7 ... 0,8 V (typowej dla układów o napięciu zasilania 5V) do 0,3 ... 0,4 V (typowej dla układów o napięciu zasilania rzędu 1 V) daje wzrost prądu podprogowego o 4 ... 6 rzędów wielkości, a nawet więcej. Ilustruje to rys. 14.1. Prąd podprogowy jest wówczas na tyle duży, że związany z nim pobór mocy może stanowić znaczącą część całkowitego poboru mocy układu.




Rys. 14.1. Wzrost wartości prądu podprogowego w wyłączonym tranzystorze przy zmniejszaniu napięcia progowego

Ponadto w tranzystorach o nanometrowych wymiarach kanału pojawiają się jeszcze inne prądy powiększające całkowity statyczny pobór prądu. Najważniejsze z nich to prąd tunelowy w złączach *pn* źródeł i drenów oraz prąd tunelowy przez dielektryk bramki. Prądy tunelowe w złączach powodują odpływ prądu do podłoża (w tranzystorach nMOS) lub wyspy (w tranzystorach pMOS). Prąd tunelowy płynący przez dielektryk bramki osiąga znaczące wartości, gdy grubość warstwy tego dielektryka zmniejsza się do pojedynczych nanometrów. Jego zależność od grubości dielektryka opisana jest funkcją wykładniczą, wzrost wartości tego prądu przy zmniejszaniu grubości dielektryka staje się bardzo gwałtowny, gdy grubość ta maleje poniżej 2 nm. Tranzystor MOS, w którym występuje prąd tunelowy bramki, nie może już być uważany za tranzystor z izolowaną bramką. Wejścia bramek cyfrowych z takimi tranzystorami pobierają prąd, którego kierunek i wartość zależy od napięcia na wejściu, czyli od stanu logicznego. Zjawisko to nie tylko powoduje wzrost całkowitego statycznego poboru prądu, ale może zmieniać działanie układu cyfrowego, bowiem prądy wejściowe bramek obciążają wyjścia bramek poprzednich (sterujących). W skrajnych przypadkach może to powodować zmiany poziomów napięć zera i jedynki i prowadzić nawet do błędów w działaniu układu.

Na szczęście zarówno prądy tunelowe w złączach, jak i prądy tunelowe bramek tranzystorów mogą być wyeliminowane lub zredukowane do nieistotnego poziomu środkami technologicznymi. Prądy tunelowe w złączach nie wystąpią, jeśli koncentracje domieszek po obu stronach warstw zaporowych złącz nie będą zbyt wysokie. Sposobem na wyeliminowanie prądów tunelowych bramek jest użycie dielektryków innych niż czysty dwutlenek krzemu ($SiO_2$). Zastosowanie warstwy dielektrycznej o przenikalności dielektrycznej większej, niż przenikalność $SiO_2$, umożliwia zwiększenie grubości dielektryka bez pogarszania parametrów tranzystora - wartość $C_{ox}$, która decyduje o wartości prądu drenu tranzystora (wzory 4.2, 4.4), jest proporcjonalna do ilorazu

współczynnika przenikalności dielektrycznej i grubości dielektryka. Takie dielektryki, np. dwutlenek krzemu z domieszką hafnu, są stosowane w układach CMOS z długościami bramki 65 nm i poniżej. Zarówno ukształtowanie rozkładów domieszek tak, by nie występowały prądy tunelowe w złączach, jak i zastąpienie czystego $SiO_2$ przez inną warstwę dielektryczną, poważnie komplikuje procesy produkcyjne, jest jednak możliwe. Dlatego

! **statyczny pobór mocy jest uzależniony przede wszystkim od wartości prądu podprogowego w tranzystorach znajdujących się w stanie wyłączenia, czyli gdy $U_{GS}$=0.**

Omówione wyżej zależności odnoszą się do wszystkich bramek statycznych CMOS, nie tylko do inwerterów. Natomiast dla bramek dynamicznych, np. dla bramek typu DOMINO omawianych w pierwszej części wykładu 8, i wszystkich innych bramek i bloków wymagających taktowania, do mocy pobieranej przez same bramki trzeba doliczyć moc związaną z taktowaniem zegarem. Sygnał zegarowy będący periodycznym ciągiem zer i jedynek powoduje przeładowywanie pojemności w układzie nawet wtedy, gdy układ nie pracuje, tj. stany logiczne w nim nie zmieniają się.

! **W dużych układach, np. w układach mikroprocesorów, moc pobierana przez układy generowania sygnałów zegarowych oraz bufory zegara może sięgać, a nawet przekraczać 40% całej mocy dynamicznej pobieranej przez układ.**

---

W układach analogowych pobór prądu bezpośrednio wynika z wybranego rozwiązania układowego i wymaganych parametrów układu. Przykładowo, dla stopni wzmacniających prąd je zasilający określa wartości wzmocnienia napięciowego oraz częstotliwości granicznej $f_T$ - przykłady w wykładzie 13 (wzory 13.8, 13.10). Warto zauważyć pewną prawidłowość: układy wzmacniające są tym "szybsze", tj. mają wyższą częstotliwość graniczną $f_T$, im większy jest ich pobór prądu. Jest tu jakościowa (choć nie ilościowa) analogia do układów cyfrowych, w przypadku których także większa "szybkość" (częstotliwość zegara) związana jest z większym poborem prądu.

Szczególnym przypadkiem są analogowe wzmacniacze dużej mocy. W nich problemem nie jest całkowity pobór prądu, bo ten uzależniony jest od tego, jaka jest rezystancja obciążenia i amplituda sygnału wyjściowego, lecz sprawność energetyczna. Sprawnością energetyczną nazywamy stosunek mocy oddawanej przez wzmacniacz do obciążenia do całkowitej mocy dostarczonej ze źródła zasilania. Im większa sprawność energetyczna, tym mniejszy procent całkowitej mocy pobieranej ze źródła zasilania rozpraszany jest w elementach układu, a zatem tym mniej wydziela się w nich ciepła, co jest z każdego punktu widzenia korzystne.

## 14.2. Sposoby zmniejszania poboru mocy w układach cyfrowych

Teraz zapoznamy się ze sposobami redukcji poboru mocy w układach cyfrowych CMOS. Omówimy najpierw metody obniżania mocy przełączania. Przypomnijmy, ż e moc przełączania bramek statycznych jest proporcjonalna do (wzór 14.1):

- pojemności obciążającej bramki $C_l$,
- kwadratu napięcia zasilania $U_{DD}^{\ 2}$,
- liczby zmian stanu logicznego w jednostce czasu *f*.

Z tej zależności wynikają wszystkie sposoby obniżania poboru mocy przełączania.

**Redukcja pojemności w układzie** nie tylko zmniejsza pobór mocy, ale skraca czasy propagacji sygnału w bramkach, umożliwiając zwiększenie szybkości działania układu. Projektant układu ma wpływ na te pojemności projektując topografię układu. Reguły są proste:

- minimalna długość bramek tranzystorów, jaką dopuszcza dana technologia (szerokość wynika z założeń dotyczących charakterystyk i parametrów bramek, patrz wykład 7),
- jak najmniejsze powierzchnie obszarów źródeł i drenów tranzystorów,
- jak najkrótsze połączenia.

Projektant może oczywiście starać się postępować według tych reguł tylko wtedy, gdy projektuje układ lub jego fragment w stylu *full custom*. Taki sposób projektowania w dziedzinie układów cyfrowych jest jednak rzadkością. Przy automatycznej syntezie topografii w stylu komórek standardowych profesjonalne systemy projektowania pozwalają projektantowi w pewnym stopniu wpływać tylko na rozmieszczenie komórek i długość połączeń. Ingerencja projektanta w proces projektowania ma w praktyce sens tylko wtedy, gdy istnieje możliwość znaczącego skrócenia połączen szczególnie długich. Ma to jednak w większym stopniu wpływ na szybkość działania układu, niż na pobór mocy. Szybkość działania układu może być bowiem ograniczona przez czas propagacji sygnału nawet w jednym tylko bardzo długim połączeniu, podczas gdy całkowity pobór mocy zależy od sumy wszystkich pojemności w układzie.

W nowszych technologiach CMOS pojemności połączeń są również redukowane na drodze technologicznej, przez stosowanie warstw dielektrycznych o przenikalności dielektrycznej niższej, niż przenikalność dielektryczna $SiO_2$. Chodzi tu oczywiście o warstwy dielektryczne pomiędzy kolejnymi poziomami połączeń, a nie o dielektryk bramkowy tranzystorów. Ten ostatni bowiem powienien mieć, jak już wiemy, przenikalność dielektryczną możliwie jak największą.

**Redukcja napięcia zasilania** bardzo skutecznie zmniejsza pobór mocy. Równocześnie jednak maleje szybkość działania układu, bowiem rosną czasy propagacji sygnałów w bramkach (wzory 7.7 - 7.8). Nawet jeśli szybkość działania układu nie ma w konkretnym zastosowaniu większego znaczenia, obniżanie napięcia zasilania ma swoje granice. Po pierwsze, układy o niższym napięciu zasilania mają niższy poziom jedynki logicznej, a więc i mniejszą odporność na zakłócenia (patrz wykład 7). Po drugie, obniżanie napięcia zasilania zmusza do obniżania napięć progowych tranzystorów. Aby bramki statyczne działały prawidłowo i dostatecznie szybko, suma wartości bezwzględnych napięć progowych tranzystorów nMOS i pMOS powinna być mniejsza od wartości napięcia zasilania. Niskie napięcia progowe oznaczają jednak, jak już wiemy, znacznie zwiększony pobór mocy statycznej (wzór 14.4). W praktyce stosowane są rozwiązania polegające na tym, że w dużych układach są bloki zasilane napięciami o różnych wartościach - niższej dla bloków nie mających krytycznego wpływu na szybkość działania układu, wyższej dla pozostałych. Niektórzy producenci dostarczają kilka wersji bibliotek komórek standardowych dostosowanych do różnych napięć zasilania. Jest to możliwe dlatego, że w wielu współczesnych technologiach istnieje możliwość produkowania w jednym układzie tranzystorów o kilku różnych grubościach tlenku bramkowego i wartościach napięć progowych. Stosowanie w tym samym układzie więcej niż jednego napięcia zasilania komplikuje jednak projekt, bowiem różne wartości napięcia zasilania oznaczają też różne poziomy jedynki logicznej. Przesyłanie danych między blokami zasilanymi różnymi napięciami wymaga więc dodatkowych, pomocniczych układów zmieniających odpowiednio poziom jedynki logicznej.

**Redukcja liczby zmian stanów logicznych w jednostce czasu** jest możliwa do osiągnięcia na kilka sposobów. Najprostszym jest oczywiście obniżenie częstotliwości zegara, co jednak propocjonalnie obniża szybkość działania układu. Jednym z rozwiązań kompromisowych jest sterowanie częstotliwością zegara uzależnioną od złożoności i priorytetu zadania wykonywanego przez układ. Na takie rozwiązania pozwala konstrukcja niektórych mikroprocesorów, przy czym sterowanie częstotliwością zegara wykonywane jest zwykle przez oprogramowanie. Innym oczywistym sposobem jest "usypianie" bloków w układzie, które w danej chwili nie wykonują żadnej funkcji. Można to osiągnąć przez wyłączenie zegara (co powoduje zmniejszenie do zera mocy przełączania, w tym również mocy związanej z sygnałem zegara), a także przez wyłączenie zasilania bloku (co powoduje

zmniejszenie do zera także mocy statycznej). Nie w każdym przypadku jest to jednak możliwe. Gdy blok zbudowany jest wyłącznie z bramek statycznych, zatrzymanie zegara w bloku w danej chwili nieaktywnym (czyli gdy na wejściach bloku stany logiczne nie zmieniają się) powoduje, że wszystkie stany logiczne pozostają bez zmian aż do chwili "obudzenia" bloku. Natomiast w przypadku bloku zbudowanego z bramek dynamicznych po krótkim czasie braku aktywności i braku sygnału zegarowego wszystkie stany logiczne zostają "zapomniane". Podobnie jest w przypadku bloku z bramek statycznych, jeśli wyłączone zostało jego zasilanie. Dlatego bloków statycznych nie można "usypiać" przez wyłączenie zasilania, a bloków dynamicznych także przez wyłączenie zegara, jeśli istotne jest, aby w czasie "uśpienia" blok zachowywał swoje stany logiczne. Niemniej, zasada "usypiania" bloków nieaktywnych jest obecnie szeroko stosowana w dużych układach. Istnieją konstrukcje mikroprocesorów zbudowanych wyłącznie z bramek statycznych. W takim mikroprocesorze częstotliwość zegara może być dowolnie niska, a nawet okresowo równa zeru. Mikroprocesor jest wówczas "uśpiony" w całości, nie pracuje, ale podejmuje działanie od stanu, w którym został "uśpiony", gdy zegar zostaje ponownie włączony. Najbardziej znane konstrukcje tego rodzaju to mikroprocesory z rodziny ARM, powszechnie stosowane w telefonach komórkowych i innych urządzeniach przenośnych wymagających maksymalnej oszczędności energii.

Znaczne oszczędności mocy przełączania można by osiągnąć eliminując całkowicie zegar. Zegara nie wymagają **układy asynchroniczne**. Idea układu asynchronicznego polega na tym, że każdy blok logiczny zaczyna wykonywać operację na danych wejściowych, gdy otrzymuje sygnał startu, a po zakończeniu operacji, gdy wyniki pojawiają się na wyjściach, układ wysyła sygnał startu do innych bloków, dla których te wyniki są danymi wejściowymi. Budowa takich układów jest teoretycznie możliwa; zauważmy, że bloki kombinacyjne zbudowane z bramek statycznych nie wymagają zegara do prawidłowego działania. Budowa dużych układow asynchronicznych zawierających obok bloków kombinacyjnych także elementy pamięciowe (przerzutniki, rejestry, bloki pamięci) związana jest jednak z koniecznością pokonania szeregu trudnych problemów. Zasadniczym problemem jest sposób generacji sygnałów startu - w jaki sposób ustalić, że wyniki na wyjściu konkretnego bloku są gotowe, nie będą już ulegać zmianom i mogą być użyte przez inne bloki? Zaproponowano różne rozwiązania tego problemu, ale wprowadzają one do układów wiele komplikacji. Narzędzia komputerowego wspomagania projektowania układów asynchronicznych dopiero powstają. Toteż układy asynchroniczne nie weszły, jak dotąd, do praktyki przemysłowej.

---

Zajmiemy się teraz zagadnieniem **redukcji mocy statycznej**. Polega ona na redukcji sumy prądów podprogowych, które płyną w stanie ustalonym przez tranzystory znajdujące się w stanie wyłączenia. Najbardziej oczywistym sposobem jest "usypianie" przez wyłączanie zasilania bloków, które w danym momencie są nieaktywne. Nie jest to jednak dopuszczalne, jeśli blok w okresie "uśpienia" ma zachować swój stan logiczny. Innym wykorzystywanym w praktyce sposobem jest zmniejszanie prądów podprogowych przez zwiększanie wartości napięć progowych tranzystorów (patrz wzór 14.4). Jak wspomniano wyżej, w wielu współczesnych technologiach CMOS możliwe jest wytworzenie w tym samym układzie tranzystorów o różnych wartościach napięć progowych, i w związku z tym wielu producentów dostarcza kilka wersji bibliotek komórek standardowych. Komórki zbudowane z tranzystorów o wyższych napięciach progowych są stosowane w tych częściach układu, które nie wymagają dużej szybkości działania. Komórki zbudowane z tranzystorów o niższych wartościach napięć progowych, które zapewniają krótsze czasy propagacji sygnału (patrz wzory 7.7 - 7.8), używane są wyłącznie w blokach, których szybkość jest krytyczna z punktu widzenia szybkości działania całego układu.

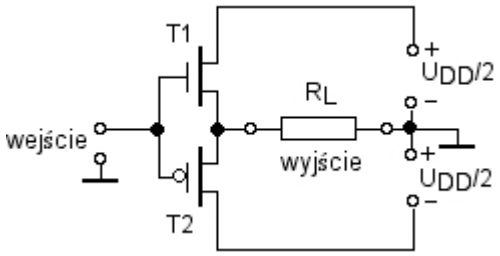
Istnieje też inny sposób obniżania wartości prądów podprogowych, polegający na wykorzystaniu zależności napięcia progowego tranzystora od napięcia polaryzacji podłoża (napięcie $U_{BS}$, wzór 4.5). Do wysp (lub podłoża), na których znajdują się tranzystory, doprowadza się napięcie polaryzujące zaporowo źródła tranzystorów względem wyspy (lub podłoża). Napięcie takie powoduje wzrost napięcia progowego, a więc zmniejszenie wartości prądów podprogowych (kosztem zmniejszenia szybkości działania układu). Ten sposób umożliwia sterowanie statycznym poborem mocy i szybkością działania układu w pracującym układzie (lub jakimś jego bloku). Wadą jest komplikacja układu - potrzebne jest źródło napięciowe generujące dodatkowe napięcie polaryzujące oraz układ sterujący wartością tego napięcia.

We współczesnych systemach projektowania istnieją specjalne programy ułatwiające automatyzację projektowania z zastosowaniem wymienionych wyżej sposobów redukcji poboru mocy przełączania i mocy statycznej.

## 14.3. Analogowe układy dużej mocy

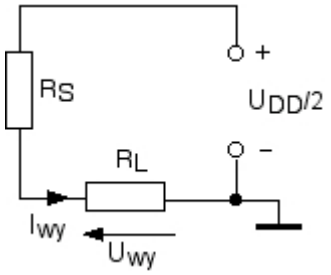
Zajmiemy się teraz stopniami wyjściowymi, szukając odpowiedzi na pytania: jaka jest i czym jest ograniczona maksymalna moc oddawana do obciążenia. Przedstawimy tu bardzo uproszczoną analizę, pozwoli ona jednak na wyciągnięcie uogólniających wniosków.

Typowe stopnie dużej mocy budowane są jako stopnie szeregowo-przeciwsobne. Ich idea, omawiana w wykładzie 13, zilustrowana jest rys. 13.16, który dla wygody tu powtórzymy jako rys. 14.2. Na rysunku pokazany jest układ z tranzystorami MOS, ale nasze rozważania będą się równie dobrze odnosić do układu z tranzystorami bipolarnymi.




Rys. 14.2. Zasada budowy układu szeregowo-przeciwsobnego

Dla obliczenia maksymalnego prądu, jaki może płynąć przez rezystancję obciążenia $R_L$, założymy, że tranzystory można zastąpić pewnymi zastępczymi rezystancjami $R_S$, których wartości można określić z charakterystyk prądowo-napięciowych tranzystora w warunkach maksymalnego wysterowania na wejściu (np. w przypadku tranzystora nMOS przyłożenia napięcia wejściowego $U_{GS}$ równego napięciu $U_{DD}/2$). Rozważmy dodatni półokres napięcia, gdy przewodzi wyłącznie górny tranzystor. Mamy wówczas prosty obwód, w którym połączone są szeregowo tranzystor T1 reprezentowany przez rezystancję $R_S$, rezystancja obciążenia $R_L$ i źródło napięcia o napięciu równym połowie całkowitej wartości napięcia zasilania.




Rys. 14.3. Schemat zastępczy do obliczenia maksymalnego prądu wyjściowego

Maksymalny prąd, jaki może płynąć przez rezystancję obciążenia, wynosi

$$I_{wymax} = \frac{U_{DD}}{2\left(R_L + R_S\right)} \tag{14.5}$$

zatem maksymalne napięcie wyjściowe wynosi

$$U_{wymax} = \frac{U_{DD} R_L}{2\left(R_L + R_S\right)} \tag{14.6}$$

moc maksymalna wydzielająca się w obciążeniu jest równa

$$P_{wymax} = \frac{U_{DD}^2 R_L}{4\left(R_L + R_S\right)^2} \tag{14.7}$$

zaś moc tracona w tranzystorze wynosi

$$P_s = \frac{U_{DD}^2 R_S}{4\left(R_L + R_S\right)^2} \tag{14.8}$$

Są to wartości chwilowe. Na ich podstawie można obliczyć maksymalną moc dla sygnału sinusoidalnego i ewentualnie sygnałów o innych kształtach. Ale nie będzie to nam potrzebne, bo zależność (14.7) dostatecznie ilustruje ograniczenia maksymalnej mocy wyjściowej i pozwala je przedyskutować.

Zależność (14.7) pokazuje, że do zwiększania mocy wyjściowej można dążyć na kilka sposobów: przez podnoszenie napięcia zasilania $U_{DD}$ oraz przez obniżanie rezystancji obciążenia $R_L$ i rezystancji zastępczej $R_S$. Przedyskutujemy te sposoby z punktu widzenia wymagań, jakie narzucają one na tranzystory stopnia dużej mocy.

Podnoszenie napięcia zasilania wymaga podwyższania napięć dopuszczalnych dla tranzystorów. W przypadku tranzystorów MOS jest to maksymalne dopuszczalne napięcie dren-źródło $U_{DS}$ oraz maksymalne napięcie dopuszczalne napięcie bramki $U_{GS}$. W przypadku tranzystorów bipolarnych jest to napięcie dopuszczalne kolektor-emiter $U_{CE}$. Ponadto w obu przypadkach dostatecznie wysokie musi być napięcie przebicia złącz wyspa-podłoże. Nie wdając się w szczegółowe rozważania na temat konstrukcji tranzystorów (wykracza to poza zakres przedmiotu "Układy scalone") wystarczy w tym miejscu stwierdzić, że podwyższanie wymienionych wyżej napięć oznacza:

- dla tranzystorów MOS: zwiększanie długości kanału oraz obniżanie koncentracji domieszek w kanale, oraz zwiększanie grubości tlenku bramkowego,
- dla tranzystorów bipolarnych: zwiększanie grubości bazy oraz odległości między złączem baza-emiter, a granicą warstwy zagrzebanej oraz obniżanie koncentracji domieszek w obszarze kolektora.

Obniżanie koncentracji domieszek jest potrzebne dlatego, że - jak wiadomo z teorii przyrządów półprzewodnikowych - napięcie przebicia każdego złącza *pn* jest tym wyższe, im niższe są koncentracje domieszek w obszarach złącza. Powiększanie długości kanału, grubości tlenku bramkowego, grubości bazy, odległości między złączem baza-emiter, a granicą warstwy zagrzebanej wynika z konieczności utrzymania natężenia pola elektrycznego w bezpiecznych granicach, a także uniknięcia przebić skrośnych (np. zwarcia warstw zaporowych złącz źródła i drenu w tranzystorze MOS). Jednak wszystkie te zmiany powodują wzrost rezystancji zastępczej $R_S$. W tranzystorze MOS zwiększanie długości kanału oraz grubości tlenku bramkowego powoduje spadek wartości prądu drenu, co jest równoznaczne ze wzrostem zastępczej rezystancji włączonego tranzystora. W przypadku tranzystora bipolarnego następuje wzrost rezystancji rozproszonych w strukturze tranzystora, przede wszystkim rezystancji obszaru kolektora, która w przypadku tranzystora bipolarnego jest głównym składnikiem rezystancji zastępczej $R_S$. Widać więc, że zmiany w konstrukcji tranzystora umożliwiające zastosowanie wyższych napięć zasilania prowadzą do wzrostu rezystancji zastępczej $R_S$. Wzrost tej rezystancji jest bardzo niekorzystny, bowiem gdy wzrasta rezystancja $R_S$, to moc oddawana do obciążenia maleje (patrz wzór 14.7). Zatem zmianom w konstrukcji tranzystora umożliwiającym podniesienie napięcia zasilania powiny towarzyszyć zmiany zapobiegające wzrostowi rezystancji $R_S$. Powrócimy do tego zagadnienia nieco niżej.

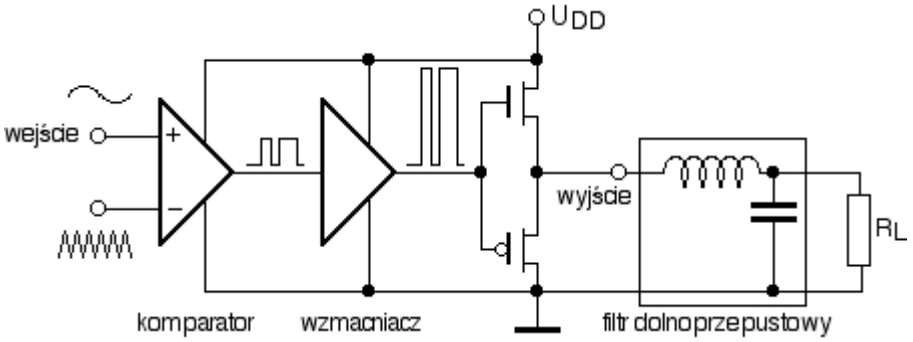
Obniżanie rezystancji obciążenia $R_L$ powoduje wzrost mocy oddawanej do obciążenia tylko tak długo, jak długo rezystancja $R_L$ jest znacznie większa od zastępczej rezystancji $R_S$. Gdyby rezystancja $R_L$ stała się mniejsza od $R_S$, dalsze jej obniżanie nie prowadziłoby już do wzrostu mocy oddawanej do obciążenia, lecz przeciwnie - powodowałoby jej spadek. Równocześnie rosłaby moc tracona w tranzystorze (wzory 14.7 i 14.8). Widzimy więc, że zarówno zwiększanie napięcia zasilania, jak i zmniejszanie rezystancji obciążenia ostatecznie zmusza do zmniejszania zastępczej rezystancji tranzystora $R_S$.

Jeżeli założyć, że takie cechy struktury tranzystora, jak rozkłady domieszek, długość kanału i grubość tlenku bramkowego (w tranzystorze MOS) czy grubość bazy (w tranzystorze bipolarnym) są optymalne z punktu widzenia parametrów tranzystora oraz wymaganej wartości napięcia zasilania, to jedynym sposobem zmniejszania zastępczej rezystancji $R_S$ jest zwiększanie powierzchni przekroju poprzecznego obszaru, przez który płynie prąd drenu bądź kolektora. Dla tranzystora MOS oznacza to zwiększanie szerokości kanału, dla tranzystora bipolarnego - zwiększanie powierzchni złącza emiter-baza (oraz ewentualnie powierzchni kontaktu do obszaru kolektora). W obu przypadkach rosnąć będzie powierzchnia zajmowana przez tranzystor. A ponieważ koszt układu jest - jak wiemy - proporcjonalny do jego powierzchni, ograniczenie maksymalnej mocy oddawanej do obciążenia przez układ scalony może mieć charakter ekonomiczny - układ scalony dużej mocy może się w konkretnym zastosowaniu okazać droższy od innych możliwych rozwiązań.

Z tego punktu widzenia tranzystory bipolarne są elementami korzystniejszymi od tranzystorów MOS, bowiem przy danej powierzchni zajmowanej przez tranzystor i danych napięciach polaryzujących tranzystory bipolarne przewodzą znacznie większy prąd od tranzystorów MOS, co oznacza, że ich zastępcze rezystancje są znacznie mniejsze, a więc i znacznie mniejsze mogą być ich powierzchnie. Z tego powodu do produkcji układów dużej mocy, np. układów stosowanych w elektronice samochodowej, stosowane są specjalne technologie BiCMOS umożliwiające wykonanie w jednym układzie zarówno układów CMOS, jak i tranzystorów bipolarnych obu typów przewodnictwa o dobrych parametrach elektrycznych. Technologie te jednak są dużo bardziej skomplikowane i bardziej kosztowne zarówno od prostej technologii CMOS, jak i od prostej technologii bipolarnej, jakie były przedstawione w wykładzie 3. Ponadto w przypadku tranzystorów bipolarnych dużej mocy istnieje ryzyko pojawienia się niestabilności elektryczno-cieplnej prowadzącej do zniszczenia tranzystora (o zjawisku tym była mowa w wykładzie 13). Z tego powodu układy, w których rolę elementów dużej mocy pełnią tranzystory MOS, są bardziej niezawodne. Dlatego opracowano też technologie pozwalające wytwarzać tranzystory MOS mające duże wartości stosunku $W/L$ kanału przy umiarkowanej powierzchni. Kanały takich tranzystorów wytwarzane są w objętości płytki półprzewodnika, a nie na powierzchni, jak w klasycznej technologii CMOS. Te technologie także są dużo bardziej złożone i bardziej kosztowne od zwykłej technologii CMOS. Ogólnie technologie mikroelektroniczne przystosowane specjalnie do wytwarzania analogowych i analogowo-cyfrowych układów dużej mocy określane są angielskim terminem "smart power". Omawianie ich wykracza poza zakres tego wykładu.

Maksymalną moc, jaką może oddać układ scalony do obciążenia, ogranicza oczywiście także konieczność odprowadzania ciepła wydzielającego się w tranzystorach wyjściowego stopnia mocy. Przy danej mocy oddawanej do obciążenia ilość ciepła wydzielającego się w tranzystorach stopnia mocy zależy od sprawności energetycznej tego stopnia. Sprawnością energetyczną nazywamy stosunek mocy oddawanej przez wzmacniacz do obciążenia do całkowitej mocy dostarczonej ze źródła zasilania. Im niższa sprawność energetyczna, tym większy - przy danej mocy oddawanej do obciążenia - pobór mocy ze źródła zasilania, a także tym więcej wydziela się ciepła, które trzeba odprowadzić od tranzystorów stopnia mocy.

Sprawność energetyczna zależy nie tylko od budowy stopnia mocy i parametrów jego elementów, ale i od rodzaju sygnałów wzmacnianych przez ten stopień. Można pokazać, że maksymalna osiągalna teoretycznie sprawność szeregowo-przeciwsobnego stopnia mocy dla czystego sygnału sinusoidalnego wynosi nieco ponad 78%. Sprawność energetyczna osiąga wartość bliską 100% dla sygnałów o charakterze impulsów prostokątnych zero-jedynkowych, tj. takich, dla których tranzystory stopnia mocy są albo całkowicie wyłączone i nie przewodzą, albo są maksymalnie wysterowane. W pierwszym przypadku prąd nie płynie, więc i moc pobierana oraz oddawana do obciążenia są równe zeru. W drugim przypadku moc oddawaną do obciążenia i traconą w tranzystorze określają wzory (14.7) i (14.8). Jeśli spełniony jest warunek $R_L >> R_S$, moc $P_S$ jest pomijalnie mała wobec $P_{wy\ max}$, a sprawność zbliża się do 100%. To spostrzeżenie doprowadziło do powstania nowej klasy układów wzmacniaczy mocy, zwanych wzmacniaczami klasy D. Zasada działania takiego wzmacniacza jest następująca. Na wejściu sygnał zmienny jest przekształcany na ciąg impulsów prostokątnych o stałej amplitudzie i długości (czasie trwania) proporcjonalnej do chwilowej wartości napięcia wejściowego. Służy do tego komparator napięcia, do którego wejść doprowadzony jest sygnał wejściowy oraz napięcie piłokształtne o częstotliwości wielokrotnie większej od częstotliwości sygnału wejściowego. Tak otrzymane impulsy są wzmacniane i doprowadzone do wejścia szeregowo-przeciwsobnego stopnia mocy. Prąd dostarczany przez stopień mocy ma więc charakter ciągu impulsów, a wartość średnia tego prądu jest proporcjonalna do ich długości. Aby lepiej odtworzyć sygnał wejściowy, pomiędzy wyjściem stopnia mocy, a obciążeniem umieszcza się filtr indukcyjno-pojemnościowy, który uśrednia prąd w czasie. Rys. 14.4 przedstawia schemat blokowy wzmacniacza w klasie D.




Rys. 14.4. Schemat blokowy wzmacniacza mocy w klasie D

Stopień mocy w tak działającym wzmacniaczu ma sprawność bliską 100%. Dlatego układy scalonych wzmacniaczy mocy do takich zastosowań, jak np. telefony komórkowe, aparaty słuchowe dla słabo słyszących czy też komputery przenośne są dziś najczęściej budowane właśnie jako wzmacniacze klasy D. Pierwsze konstrukcje takich wzmacniaczy nie zapewniały wysokiej wierności odtwarzanego dźwięku, ale szereg udoskonaleń (wyższa częstotliwość próbkowania sygnału wejściowego, wprowadzenie ujemnego sprzężenia zwrotnego i in.) doprowadziło do tego, że wzmacniacze klasy D zaczynają się pojawiać także w sprzęcie

elektroakustycznym wysokiej jakości. Wzmacniacze klasy D znajdują też zastosowania innego rodzaju, na przykład jako układy sterujące pracą silników elektrycznych prądu stałego.

## 14.4. Chłodzenie układów scalonych

Ze wszystkich znanych rodzajów układów scalonych układy CMOS są najbardziej oszczędne, jeśli chodzi o pobór mocy. Jest to zresztą jedna z przyczyn, dla których wyparły one niemal całkowicie inne rodzaje układów. Przed 40 laty, gdy cyfrowe układy CMOS były nowością, ich częstotliwości pracy nie przekraczały pojedynczych MHz, liczba elementów w układzie nie przekraczała kilku tysięcy, a analogowych układów CMOS nie było w ogóle, ilość ciepła wydzielająca się w pracujących układach była znikoma. W miarę rozwoju technologii wzrastała częstotliwość pracy układów oraz gęstość upakowania elementów, i problem odprowadzania wydzielającego się ciepła stawał się coraz poważniejszy (zobacz dodatek 1). Dziś jest to bariera uniemożliwiająca dalszy wzrost częstotliwości pracy układów cyfrowych.

Pracujący układ scalony musi być chłodzony w taki sposób, aby nie została przekroczona jego maksymalna dopuszczalna temperatura pracy. Temperatura ta jest określona na takim poziomie, aby zapewnić układowi wymaganą trwałość i niezawodność. Krótkotrwałe i niewielkie przekroczenie dopuszczalnej temperatury zwykle nie powoduje natychmiastowego uszkodzenia, natomiast jeśli przekroczenie jest duże i długotrwałe lub często się powtarza, czas bezawaryjnej pracy układu ulega znacznemu skróceniu.

Jak wspomniano w wykładzie 11, zjawiskami, które najczęściej wywołują uszkodzenia pracujących układów, są:

- elektromigracja powodująca uszkodzenia połączeń,
- defekty w płytce półprzewodnikowej i uszkodzenia płytki powstające w wyniku naprężeń mechanicznych,
- degradacja jakości tlenku bramkowego wywołana nośnikami ładunku o wysokiej energii (zwanymi "gorącymi nośnikami").

Uszkodzenia (upływności, korozja) mogą też być powodowane przedostawaniem się zanieczyszczeń i wilgoci przez nieszczelne obudowy układów.

Wszystkie te mechanizmy powstawania uszkodzeń nasilają się przy wzroście temperatury, przy czym zależności od temperatury są bardzo silne. Zależność częstości występowania uszkodzeń od temperatury ma charakter wykładniczy.

Maksymalne dopuszczalne temperatury pracy wynoszą od $75^{o}$C do $180^{o}$C. Najniższe są dopuszczalne temperatury pracy układów montowanych w obudowach z tworzywa sztucznego. Obudowy te nie są zbyt szczelne, a współczynniki rozszerzalności cieplnej tworzyw różnią się znacznie od rozszerzalności metali i półprzewodnika, co powoduje naprężenia mechaniczne przy zmianach temperatury i przyczynia się do wzrostu częstości uszkodzeń. Wyższe dopuszczalne temperatury pracy mają układy w obudowach ceramicznych. Wysoka maksymalna temperatura pracy musi być brana pod uwagę już w czasie projektowania układu. Przykładowo, im wyższa temperatura pracy, tym mniejsza dopuszczalna gęstość prądu w ścieżkach połączeń, a więc ścieżki muszą mieć odpowiednio większe szerokości.

W przypadku cyfrowych układów CMOS maksymalna dopuszczalna temperatura pracy może także być ograniczona z powodu spadku szybkości działania układu przy wzroście temperatury. Wzrost temperatury bramek CMOS powoduje wydłużenie ich czasów propagacji sygnału (ponieważ ze wzrostem temperatury maleją wartości prądów drenu tranzystorów, patrz wykład 12). A więc w wyższej temperaturze układ działa wolniej. Zatem po przekroczeniu pewnej temperatury układ może przestać spełniać wymagania na parametry elektryczne (np. maksymalną częstotliwość zegara), i będzie działał wadliwie mimo tego, iż nie będzie fizycznie uszkodzony.

Zajmiemy się teraz sposobami chłodzenia układów scalonych. Ciało, którego temperatura jest wyższa od temperatury otoczenia, oddaje energię cieplną do otoczenia przez przewodnictwo cieplne oraz przez promieniowanie (podczerwień). W przypadku układów scalonych przewodnictwo cieplne jest głównym mechanizmem oddawania ciepła.

Jeśli strumień ciepła płynie od źródła, w którym wydziela się moc $P$, do odbiornika o temperaturze $T_0$, to temperatura źródła $T_1$ wynika z prostego równania

$$P = \frac{T_1 - T_0}{R_T} \tag{14.9}$$

w którym $R_T$ jest rezystancją termiczną między źródłem, a odbiornikiem. Rezystancja termiczna dana jest zależnością analogiczną do rezystancji elektrycznej - jeśli przepływ ciepła następuje przez prostopadłościan o długości między źródłem ciepła, a odbiornikiem wynoszącej $L$, oraz polu przekroju $S$, to rezystancja termiczna

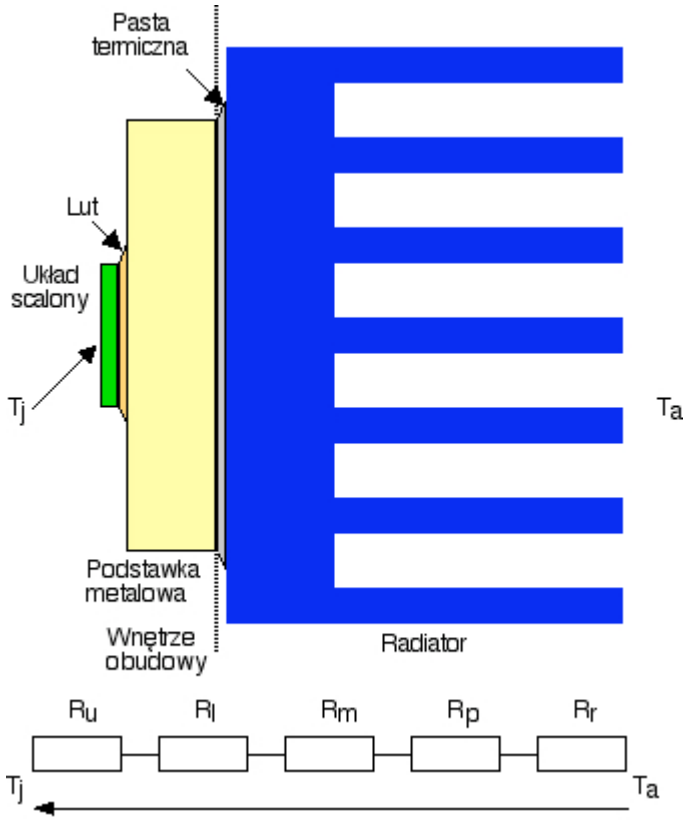
tego prostopadłościanu wynosi

$$R_T = \frac{1}{\lambda_T}\frac{L}{S} \tag{14.10}$$

gdzie $\lambda_T$ jest przewodnością cieplną materiału, z którego wykonany jest prostopadłościan. Przewodność cieplna krzemu wynosi około 150 W/(m*K). Przewodność cieplna metali stosowanych w mikroelektronice jest większa, np. dla miedzi, która jest bardzo dobrym przewodnikiem ciepła, przewodność cieplna wynosi około 400 W/(m*K), dla aluminium - około 240 W/(m*K).

Rezystancja termiczna jest wielkością podlegającą podobnym regułom, jak rezystancja elektryczna. Jeśli na przykład strumień ciepła przepływa kolejno przez kilka ośrodków charakteryzujących się określonymi rezystancjami termicznymi (czyli z punktu widzenia transportu ciepła "połączonymi szeregowo"), to wypadkowa rezystancja termiczna jest sumą rezystancji termicznych tych ośrodków. Jeśli przepływ ciepła odbywa się dwiema równoległymi, niezależnymi drogami (np. płytka z układem scalonym jest chłodzona z obu stron), to wypadkową rezystancję termiczną oblicza się tak samo, jak dla rezystancji elektrycznej dwóch rezystorów połączonych równolegle. Wygodnie jest więc posługiwać się schematami zastępczymi transportu ciepła analogicznymi do schematów elektrycznych.

Rys. 14.5 pokazuje układ scalony wraz metalową podstawką, do której jest umocowany wewnątrz obudowy i zewnętrznym radiatorem, oraz zastępczy schemat cieplny.




Rys. 14.5. Przykładowy schemat chłodzenia układu scalonego. Pokazano tylko szczegóły istotne z punktu widzenia odprowadzania ciepła.

Płytka z układem ma na powierzchni temperaturę $T_j$. Płytka ta jest przylutowana lutem złotym do metalowej podstawki. Rezystancja termiczna między powierzchnią płytki, a warstwą lutu wynosi $R_u$, rezystancja termiczna warstwy lutu - $R_l$, rezystancja termiczna między warstwą lutu, a powierzchnią podstawki dostępną na zewnątrz obudowy - $R_m$. Pomiędzy tą powierzchnią, a powierzchnią radiatora znajduje się warstwa pasty termicznej. Jej zadaniem jest poprawa przewodzenia ciepła między metalową podstawką, a radiatorem. Rezystancja termiczna warstwy pasty wynosi $R_p$. Wreszcie $R_r$ jest rezystancją termiczną charakteryzującą wymianę ciepła między radiatorem, a otoczeniem, którym jest zazwyczaj powietrze. Temperatura otoczenia wynosi $T_a$. Jeżeli w układzie wydziela się moc $P$, to temperatura powierzchni układu $T_j$ wynosi

$$T_j = T_a + P\left(R_u + R_l + R_m + R_p + R_r\right) \qquad (14.11)$$

Suma rezystancji $R_u+R_l+R_m$ zależy od szczegółów konstrukcji układu i obudowy: wymiarów i grubości płytki półprzewodnikowej, grubości warstwy lutu, materiału i grubości podstawki, na której umocowany jest układ. Na te szczegóły ma wpływ konstruktor i producent układu. Suma tych rezystancji jest podawana przez producenta układu jako rezystancja termiczna układ - obudowa (oznaczana często symbolem $R_{thjc}$). Użytkownik układu decyduje o rezystancjach $R_p$ i $R_r$, ale trzeba pamiętać, że nawet gdyby rezystancje te były równe zeru, to całkowita rezystancja nie będzie nigdy mniejsza, niż $R_{thjc}=R_u+R_l+R_m$. Zatem

! **jeśli maksymalna dopuszczalna temperatura układu wynosi $T_{jmax}$, a temperatura otoczenia $T_a$, to maksymalna moc, jaka może się bezpiecznie wydzielać w układzie, przy jakimkolwiek sposobie chłodzenia układu, bez względu na rodzaj i wielkość radiatora, nie będzie mogła być większa, niż $P_{max}=(T_{jmax}-T_a)/R_{thjc}$.**

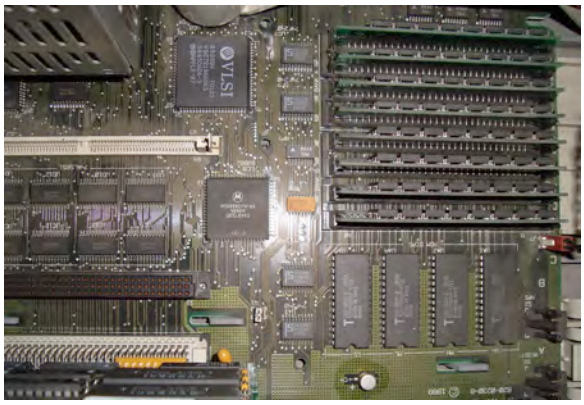
Typowa wartość rezystancji $R_{thjc}$ może wahać się w granicach od ułamka do kilku $^{o}$C/W. Rezystancje zewnętrzne $R_p$ i $R_r$ mogą się zmieniać w bardzo szerokich granicach. Jeśli stosujemy (jak to zwykle ma miejsce) gotowe radiatory, to w danych technicznych producenta znajdziemy wartości rezystancji termicznej $R_r$ dla różnych warunków chłodzenia. Wartości te zależą od sposobu wymuszania obiegu powietrza (naturalny ruch powietrza lub chłodzenie wymuszone dmuchawą), orientacji radiatora (żeberka w pionie ułatwiają naturalny ruch ku górze ogrzanego powietrza) itp. Radiatory barwione na kolor czarny mają nieco mniejszą rezystancję termiczną, ponieważ pewna część energii cieplnej jest wypromieniowywana w postaci podczerwieni.

Producenci układów scalonych oprócz rezystancji $R_{thjc}$ podają jeszcze drugą wartość rezystancji oznaczaną często $R_{thja}$. Jest to rezystancja termiczna między powierzchnią układu, a otoczeniem, gdy obudowa układu *nie jest umieszczona na żadnym radiatorze*. Rezystancja $R_{thja}$ jest wielokrotnie większa od $R_{thjc}$. Odprowadzanie ciepła w takim przypadku odbywa się przez bezpośrednią wymianę ciepła między obudową układu, a otaczającym powietrzem. Część ciepła odpływa też poprzez wyprowadzenia układu do płytki drukowanej.

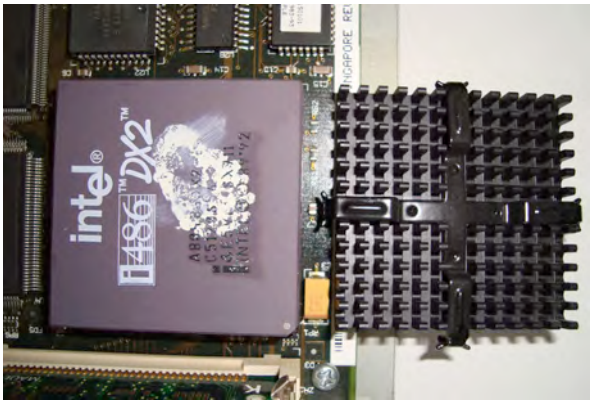
Dla układów wydzielających niewielką moc, dla których chłodzenie przy użyciu radiatora nie jest przewidziane, podawana jest tylko rezystancja $R_{thja}$. Ma ona wysoką wartość, rzędu kilkuset $^{o}$C/W, co oznacza, że maksymalna moc, jaka może się wydzielać w układzie w typowych warunkach, jest poniżej 1W.

Podane wyżej zależności są wystarczające do obliczeń w przypadkach, gdy moc wydzielana w układzie nie zmienia się w czasie. Jeśli natomiast moc ulega dużym zmianom, na przykład gdy pobór prądu i wydzielanie ciepła mają charakter impulsowy, trzeba uwzględniać dynamikę przepływu ciepła. Elementy przewodzące ciepło charakteryzują się nie tylko rezystancją, ale i pojemnością cieplną, a zastępczy schemat transportu ciepła jest analogiczny do schematów elektrycznych z elementami R i C. Szczegółowe omawianie tego zagadnienia wykracza poza zakres wykładu "Układy scalone", obszerniejsze omówienia można znaleźć w literaturze, np. w poz. [4] literatury do tego wykładu.

## 14.4. Dodatek 1: Historia chłodzenia mikroprocesorów

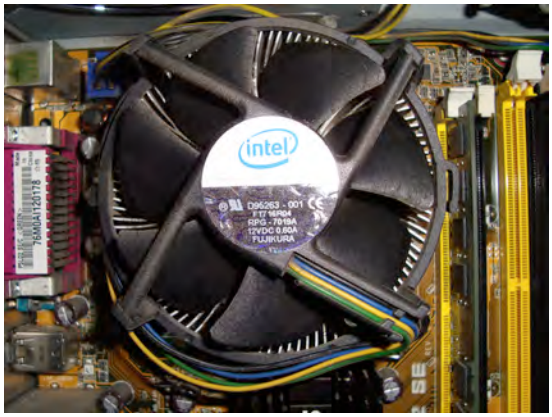


Fragment wnętrza komputera, druga połowa lat 80 XX w. - żaden układ scalony nie wymaga specjalnego chłodzenia. Częstotliwość zegara: 16 MHz.




Fragment wnętrza komputera, początek lat 90 XX w. - niewielki radiator (zdjęty, widoczny obok procesora). Częstotliwość zegara: 33 MHz.

Fragment wnętrza komputera, koniec lat 90 XX w. - solidny radiator na płycie z procesorem, za radiatorem dmuchawa. Częstotliwość zegara: 800 MHz.




Fragment wnętrza komputera, około roku 2005 - bezpośredni nadmuch powietrza na radiator umocowany na procesorze. Częstotliwość zegara: 2 GHz.

Fragment wnętrza komputera, około roku 2008 - moduł chłodzenia cieczą zespołu procesorów, obok (wyjęty do zdjęcia) zespół dwóch dmuchaw. Częstotliwość zegara: 2,7 GHz.

# ZADANIA DO WYKŁADU 14

**Zadanie 1**

Dany jest układ scalony w obudowie z tworzywa sztucznego, dla którego rezystancja termiczna układ-otoczenie $R_{thja}$ wynosi 300 $^{o}$C/W, a maksymalna temperatura pracy układu wynosi 85 $^{o}$C. Jaka może być maksymalna wartość mocy wydzielanej w układzie dla temperatur otoczenia: (a) 25 $^{o}$C, (b) 45 $^{o}$C?

**Zadanie 2**

Mikroprocesor wydziela w czasie pracy moc 100 W. Rezystancja termiczna układ-obudowa $R_{thjc}$ wynosi 0,4 $^{o}$C/W. Maksymalna temperatura pracy układu wynosi 150 $^{o}$C. W jakiej maksymalnie temperaturze otoczenia może pracować ten mikroprocesor, jeśli rezystancja termiczna radiatora (wraz z rezystancją warstwy pasty termicznej) wynosi 0,8 $^{o}$C/W?

**Zadanie 3**

W czasie pracy mikroprocesora wydziela się moc statyczna wynosząca $P_{st}$ = 20 W oraz moc dynamiczna, która zależy od częstotliwości zegara następująco: $P_{dyn} = k*f$, gdzie $k$ = 30 W/GHz. Z jaką maksymalną częstotliwością zegara może być eksploatowany ten mikroprocesor, jeśli jego maksymalna temperatura pracy wynosi 150 $^{o}$C, maksymalna temperatura otoczenia, w jakiej wymagana jest poprawna i niezawodna praca, wynosi 50 $^{o}$C, rezystancja termiczna $R_{thjc}$ wynosi 0,4 $^{o}$C/W, a rezystancja termiczna radiatora (wraz z rezystancją warstwy pasty termicznej) wynosi 0,6 $^{o}$C/W? O ile wzrosłaby ta częstotliwość zegara, gdy moc statyczna została zredukowana do zera?

**Zadanie 4**

Akustyczny wzmacniacz mocy ma dla przeciętnego sygnału muzycznego sprawność energetyczną 60%. Wzmacniacz jest eksploatowany z radiatorem w postaci płaskiej płyty o pewnej powierzchni $A$. Oszacuj, ile razy mniejsza będzie mogła być ta powierzchnia, jeśli wzmacniacz zostanie zastąpiony wzmacniaczem w klasie D o sprawności 95%? Należy przyjąć następujące założenia upraszczające: (1) rezystancja termiczna układ-obudowa $R_{thjc}$ jest pomijalnie mała wobec rezystancji termicznej radiatora, (2) rezystancja termiczna radiatora jest odwrotnie proporcjonalna do jego powierzchni.

## Bibliografia

[1] P. Horowitz, W.Hill, "Sztuka elektroniki", WKiŁ Warszawa, wyd. 9, 2009

Dużo praktycznych informacji o chłodzeniu elementów półprzewodnikowych można znaleźć w internecie, np:

[2] P. Górecki, "*Tranzystory dla początkujących - Parametry termiczne*", Elektronika dla wszystkich, nr 7, 1998, str. 22 - 25, http://www.edw.com.pl/pdf/k01/31_13.pdf
[3] P. Górecki, "*Radiatory w sprzęcie elektronicznym*", Elektronika dla wszystkich, nr 12, 1999, str. 34 - 39, http://www.edw.com.pl/pdf/k01/48_07.pdf
[4] Podręcznik firmy NXP (dawniej Philips Semiconductors): http://www.nxp.com/documents/thermal_design/SC18_CHAPTER_5.pdf

# Wykład 15: Przyszłość mikroelektroniki

## Wstęp

Ostatni wykład to spojrzenie w przyszłość. Czy mikroelektronika będzie nadal rozwijać się w tak zawrotnym tempie, jak dotąd? Gdzie są granice i skąd się biorą? Określa je oczywiście w pierwszym rzędzie technologia, a ta - choćby najdoskonalsza - nie może pokonać praw natury. Istnieją granice rozwoju wyznaczone przez prawa fizyki i dane nam przez przyrodę właściwości materiałów. Oprócz technologicznych problemów i granic rozwoju omawiane są też problemy projektowania coraz bardziej złożonych układów i systemów, oraz przedstawione są niektóre nowe idee w dziedzinie przetwarzania informacji metodami odmiennymi od powszechnie przyjętych.

Wykład kończy się podsumowaniem. Jest w nim mowa o tym, czego nie było w programie przedmiotu "Układy scalone", i co warto dalej studiować, by stać się w pełni profesjonalnym konstruktorem układów i systemów elektronicznych.

## 15.1. Skalowanie i reguły Moore'a

W pierwszej części pierwszego wykładu przytoczona została obserwacja zwana powszechnie "prawem Moore'a", głosząca, że liczba tranzystorów w układach scalonych podwaja się mniej więcej co dwa lata. Z tą prognozą związana jest inna, głosząca że minimalny wymiar tranzystora (długość kanału) maleje dwukrotnie mniej więcej co pięć - sześć lat. Rzeczywiście, od początkowej długości 10 µm (rok 1971) doszliśmy dziś (rok 2010) do 32 nm. Wraz z długością kanału zmniejszane były też inne wymiary w układach CMOS, choć nie zawsze w takiej samej proporcji.

Redukowanie wymiarów tranzystorów, a zwłaszcza zmniejszanie długości bramki, ma kluczowe znaczenie dla wzrostu złożoności i szybkości działania układów CMOS. Rozważymy to zagadnienie nieco bardziej szczegółowo. Założymy dla uproszczenia, że jedynymi pojemnościami w układzie są pojemności bramek tranzystorów MOS proporcjonalne do ich powierzchni, czyli do iloczynu $WL$.

Załóżmy, że zmniejszamy proporcjonalnie wszystkie wymiary tranzystora MOS, ale nie zmieniamy napięć panujących w układzie, także napięcie progowe tranzystora i ruchliwość nośników pozostają bez zmian. Załóżmy, że szerokość i długość kanału oraz grubość dielektryku bramkowego zostają podzielone przez ten sam stały współczynnik $S$ zwany **współczynnikiem skalowania**. Konsekwencje tego są następujące:

- powierzchnia bramek tranzystorów maleje jak $1/S^2$,
- pojemność bramek tranzystorów maleje jak $1/S$ (bo powierzchnia maleje jak $1/S^2$, ale grubość dielektryku maleje $S$-krotnie),
- współczynniki $K_p$, $K_n$ (wzór 4.13) rosną $S$-krotnie,
- prądy w układzie *rosną* $S$-krotnie,
- **czasy przełączania bramek maleją jak $1/S^2$** (co wynika ze wzrostu prądów w stosunku $S$ i malenia pojemności w stosunku $1/S$),
- moc pobierana przez układ rośnie $S$-krotnie,
- **natężenie pól elektrycznych w kanale i w dielektryku bramkowym rośnie $S$-krotnie**,
- **gęstość mocy wydzielanej w układzie (moc na jednostkę powierzchni) rośnie $S^3$-krotnie** (zakładamy tu, że całkowita powierzchnia układu maleje w takim samym stopniu, jak powierzchnia bramek tranzystorów, czyli jak $1/S^2$).

Dwa ostatnie stwierdzenia pokazują, że zmniejszanie wymiarów bez zmiany napięć, zwane **skalowaniem przy stałym napięciu**, ma swoje granice. Stosowano je mniej więcej do roku 1992, gdy długość bramki osiągnęła 0,5 µm (owo stałe napięcie zasilania miało wartość 5 V). Przy długości bramki 0,35 µm uznano, że zarówno natężenia pól elektrycznych, jak i gęstość wydzielającej się w układzie mocy osiągnęły już maksimum, którego nie można przekroczyć. Dlatego kolejne generacje technologii cechowało coraz niższe dopuszczalne napięcia zasilania: 3,3 V, następnie 2,5 V, 1,8 V, 1,2 V. Napięcie zasilania malało mniej więcej proporcjonalnie do zmniejszania długości bramki. Oznacza to **skalowania przy stałym natężeniu pól elektrycznych**: dzielimy przez $S$ nie tylko wymiary, ale także napięcie zasilania oraz napięcie progowe. Prowadzi to do następujących skutków:

- powierzchnia bramek tranzystorów maleje jak $1/S^2$,
- pojemność bramek tranzystorów maleje jak $1/S$ (bo powierzchnia maleje jak $1/S^2$, ale grubość dielektryku maleje $S$-krotnie),
- współczynniki $K_p$, $K_n$ (wzór 4.13) rosną $S$-krotnie,
- prądy w układzie *maleją* jak $1/S$,
- **czasy przełączania bramek maleją jak $1/S$** (wprawdzie prądy maleją jak $1/S$, ale pojemności także maleją jak $1/S$, a ponieważ maleją również napięcia, to ładunki maleją jak $1/S^2$,
- moc pobierana przez układ *maleje* jak $1/S^2$ (ponieważ maleją zarówno prądy, jak i napięcia),
- **natężenie pól elektrycznych w kanale i w dielektryku bramkowym nie zmienia się**,
- **gęstość mocy wydzielanej w układzie (moc na jednostkę powierzchni) nie zmienia się** (zakładamy tu, że całkowita powierzchnia układu maleje w takim samym stopniu, jak powierzchnia bramek tranzystorów, czyli jak $1/S^2$).

Jak widać, skalowanie przy stałym natężeniu pól elektrycznych prowadzi do zwiększania szybkości działania układów mimo obniżania napięcia zasilającego, przy czym można uniknąć katastrofalnego wzrostu gęstości wydzielanej mocy.

Jednak dalsze zmniejszanie napięcia zasilania poniżej 0,9 V ... 1 V nie jest możliwe. Przyczynę już znamy (wykład 14) - jest nią statyczny pobór mocy. Zmniejszaniu napięcia zasilania musi towarzyszyć proporcjonalne zmniejszanie napięć progowych tranzystorów, a to powoduje wykładniczy wzrost prądu podprogowego - wzór 14.4. Toteż dla technologii z bramką o długości 90 nm i mniejszej napięcie zasilania układu nie jest już obniżane. Wróciliśmy więc do skalowania przy stałym napięciu - ale niezupełnie. Gdy długość kanału maleje, a napięcie dren-źródło nie zmienia się, natężenie pola elektrycznego wzrasta. W silnych polach elektrycznych ruchliwość nośników maleje. Grubości tlenku bramkowego nie można zmniejszać w takiej samej skali, jak długości kanału, ze względu na ryzyko przebicia oraz tunelowy prąd upływu bramki. Nie rośnie zatem pojemność jednostkowa bramki $C_{ox}$. Nie rośnie rucliwość ani $C_{ox}$, nie rosną więc współczynniki $K_p$, $K_n$. W rezultacie proste skalowanie poniżej 90 nm nie poprawia już parametrów tranzystorów, a nawet może je pogarszać. Niewielką poprawę parametrów tranzystorów osiąga się przez wykorzystanie takich zabiegów, jak wprowadzenie naprężeń mechanicznych o kontrolowanej wielkości do obszarów kanałów (to zwiększa ruchliwość nośników), zastąpienie $SiO_2$ przez dielektryki o wyższej przenikalności (to zwiększa pojemność jednostkową bramki $C_{ox}$), czy też wprowadzanie do obszarów kanałów skomplikowanego niejednorodnego rozkładu koncentracji domieszek zwiększającego odporność tranzystorów na wysoką wartość natężenia pól elektrycznych oraz na efekt "krótkiego kanału" (zależności napięcia progowego tranzystorów od długości kanału). Ponieważ skalowanie nie prowadzi już, jak dawniej, do istotnego zwiększenia szybkości działania układów, głównym celem zmniejszania wymiarów staje się możliwość zwiększania liczby tranzystorów w układzie o danej powierzchni, co pozwala na budowanie układów i systemów cyfrowych o coraz bardziej złożonych i wydajnych architekturach.

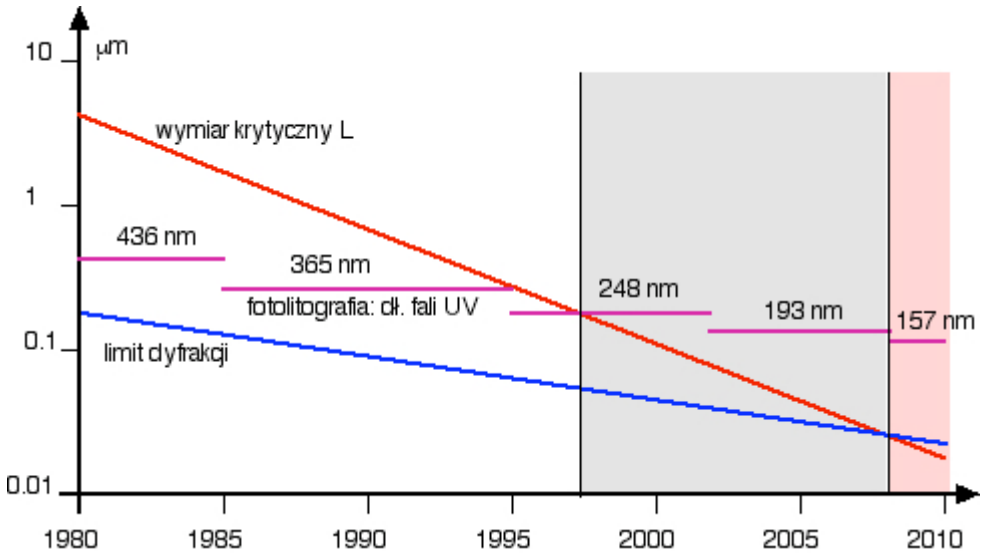
Przewiduje się, że kres skalowania nastąpi przy długości bramki rzędu 6 - 8 nm. Jedną z przyczyn będzie prąd tunelowy nośników przechodzących bezpośrednio ze źródła do drenu. Wprawdzie już dziś trwają prace nad elementem zwanym tunelowym tranzystorem MOS, w którym prąd tunelowy może być kontrolowany przez pole elektryczne wytwarzane przez bramkę. Jednak praktyczna przydatność takiego elementu jest niepewna. Pewne natomiast jest to, że granica na drodze dotychczasowego rozwoju mikroelektroniki jest już niedaleko. Granicę tę wyznaczają prawa fizyki. Oprócz niej istnieją też bariery technologiczne i ekonomiczne. O nich będzie mowa w następnej części wykładu.

## 15.2. Bariery i granice rozwoju

Przyjrzymy się teraz problemom technologicznym narastającym w miarę rozwoju mikroelektroniki, niektórym rozwiązaniom tych problemów, a także problemom projektowania, i ich ekonomicznym konsekwencjom.

**Fotolitografia**

Jak wiemy, rozwój mikroelektroniki był i jest w pierwszym rzędzie uwarunkowany zmniejszaniem wymiarów elementów w układach scalonych. Zmniejszanie wymiarów jest możliwe dzięki opanowywaniu technologii fotolitograficznych o coraz lepszej zdolności rozdzielczej i dokładności. Przez długi czas uważano, że minimalny wymiar, jaki można dobrze odwzorować techniką fotolitografii, nie może być mniejszy od długości fali świetlnej użytej do naświetlania fotolitograficznej emulsji światłoczuł ej (zwanej gwarowo fotorezystem). Zdolność rozdzielczą ograniczają bowiem zjawiska dyfrakcji oraz interferencji, ujawniające się na odległościach rzędu długości fali świetlnej. Dlatego do naświetlania wykorzystuje się ultrafiolet. Z biegiem lat wykorzystywane staje się promieniowanie ultrafioletowe o coraz mniejszej długości fali (rys.15.1).

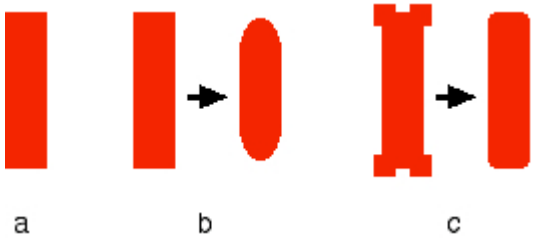



Rys. 15.1. Historia fotolitografii

Obecnie stosowane jest do najbardziej zaawansowanych procesów promieniowanie o długości fali 248 nm, 193 nm i 157 nm, ale w produkcji są układy scalone, w których minimalny wymiar jest na poziomie 32 nm! W drugiej połowie lat dziewięćdziesiątych XX w. zaczęto produkować układy o wymiarach mniejszych od długości fali świetlnej wykorzystywanej w fotolitografii, a teraz mamy już do czynienia także z wymiarami mniejszymi od limitu wynikającego z dyfrakcji. Jak to jest możliwe? Osiąga się to przez zastosowanie

- wstępnego zniekształcania obrazu na maskach,
- masek z kontrastem fazowym,
- podwójnego naświetlania,
- fotolitografii immersyjnej.

Wstępne zniekształcanie obrazu polega na celowej deformacji kształtów na maskach, która kompensuje zniekształcenia powodowane dyfrakcją, interferencją, a także innymi czynnikami (na przykład zależnością szybkości procesów trawienia fotorezystu od kształtu i wymiarów trawionych obszarów i od ich otoczenia). Idea jest pokazana na rys. 15.2. Wstępne deformacje obrazów na maskach są wprowadzane przez odpowiednie programy komputerowe, konstruktor nie musi o tych deformacjach wiedzieć.




Rys. 15.2. Zasada korekcji (wstępnego zniekształcania) obrazu na maskach:
(a) pożądany kształt paska polikrzemu, (b) kształt na masce i kształt otrzymany w układzie bez korekcji, (c) kształt na masce z korekcją i kształt otrzymany w układzie.

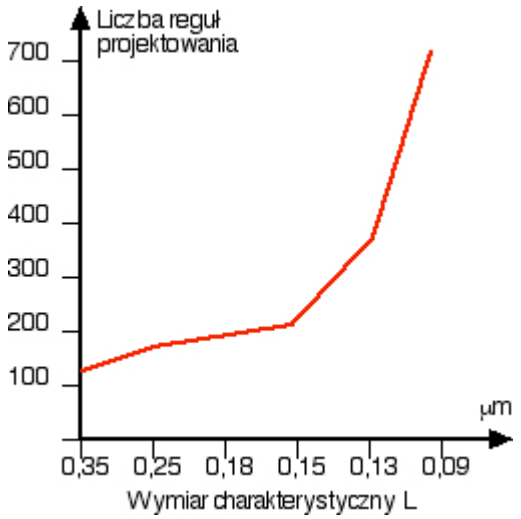
Maski z kontrastem fazowym wykorzystują zjawisko interferencji do podniesienia kontrastu obrazu na granicy obszarów. Takie maski nie są płaskie. W pobliżu krawędzi, które mają być ostro odwzorowane, występują wypukłości i wgłębienia, dzięki którym promienie ultrafioletowe docierające w pobliże krawędzi różnymi drogami mają różne fazy. Tam, gdzie fotorezyst ma być naświetlony, w wyniku interferencji następuje wzrost natężenia fali świetlnej. W obszarach, które mają pozostać nie naświetlone, interferencja powoduje wygaszenie fali świetlnej.

Podwójne naświetlanie polega w uproszczeniu na tym, że proces naświetlania wykonywany jest dwukrotnie, przy zastosowaniu dwóch różnych masek. Zbiór wszystkich obiektów geometrycznych definiowanych w danym procesie fotolitografii dzielony jest na dwa podzbiory, obiekty z jednego z nich trafiają na jedną z masek, z drugiego - na drugą. Dzięki temu na każdej z masek odległości między obiektami są większe, wymagania dla rozdzielczości fotolitografii są nieco mniejsze, a po wykonaniu obu procesów naświetlania i wywołaniu fotorezystu otrzymujemy komplet obiektów.

Fotolitografia immersyjna polega na zanurzeniu naświetlanej płytki z fotorezystem w cieczy o wysokim współczynniku załamania światła, co daje efekt optyczny zwiększenia rozdzielczości. Cieczą taką może być po prostu woda o odpowiednio wysokiej czystości.

Panuje przekonanie, że łączne zastosowanie tych metod umożliwi zmniejszanie wymiarów elementów do poziomu 20 nm bez konieczności wprowadzenia radykalnych zmian do sposobu wykonywania fotolitografii.

Opisane wyżej udoskonalenia procesu fotolitografii wymuszają wprowadzenie licznych nowych geometrycznych reguł projektowania. Ich liczba gwałtownie rośnie w najnowszych generacjach technologii układów scalonych (rys. 15.3).




Rys. 15.3. Wzrost liczby reguł projektowania towarzyszący zmniejszaniu wymiarów w układach scalonych

**Technologia bramki tranzystora**

W najbardziej zaawansowanych technologiach CMOS dielektryk bramkowy nie jest już czystym dwutlenkiem krzemu. Jak już wiemy (wykład 14), dielektryk bramki nie może być zbyt cienki, aby uniknąć tunelowego prądu bramki oraz przebicia, a równocześnie powinien zapewniać możliwie dużą jednostkową pojemność $C_{ox}$. Aby równocześnie spełnić te dwa warunki, stosowane są dielektryki o przenikalności dielektrycznej wyższej, niż przenikalność $SiO_2$. Materiałów takich poszukiwano od bardzo dawna. Muszą one spełniać szereg dodatkowych warunków fizycznych, chemicznych i technologicznych, takich jak łatwość otrzymania materiału o odpowiednim składzie, łatwość nałożenia go na krzem, brak szkodliwego oddziaływania na powierzchnię krzemu (zanieczyszczanie, szkodliwe naprężenia mechaniczne, powstawanie defektów), dobre i stabilne właściwości mechaniczne, odporność na zmiany temperatury itp. Po wielu latach eksperymentów opanowano stosowanie jako dielektryku bramkowego warstw $SiO_2$ z domieszką atomów azotu, a ostatnio hafnu. Technologia wytwarzania takich warstw jest daleko bardziej skomplikowana, niż utlenianie termiczne powierzchni krzemu, jakie wystarcza dla wytworzenia jako tlenku bramkowego czystej warstwy $SiO_2$.

Do niedawna materiałem bramki był zawsze domieszkowany krzem polikrystaliczny typu *n*. Wytworzenie domieszkowanej warstwy takiego polikrzemu było prostym procesem. Dziś w zaawansowanych technologiach konieczne jest stosowanie dwóch różnych typów krzemu polikrystalicznego: typu *n* dla tranzystorów nMOS, typu *p* dla tranzystorów pMOS. Zwiększa to liczbę procesów domieszkowania. Wadą polikrzemu jest też to, że będąc półprzewodnikiem o stosunkowo ograniczonej koncentracji nośników ładunku podlega zjawisku zmiany rozkładu

koncentracji tych nośników przy zmianach polaryzacji bramki. Przykładowo, w bramce polikrzemowej typu *n*, po przyłożeniu dodatniego napięcia do bramki, następuje odpływ elektronów od powierzchni granicznej bramka-dielektryk bramkowy. Powstaje obszar zubożony w nośniki ładunku, co daje taki rezultat, jak gdyby warstwa dielektryczna była grubsza, niż jest w rzeczywistości. Im mniejsza jest rzeczywista grubość warstwy dielektryku bramkowego, tym bardziej efekt ten wpływa na parametry tranzystorów. Zastępuje się więc krzem polikrystaliczny warstwami metali trudnotopliwych. Jednak wytworzenie takich bramek jest technologicznie trudniejsze, niż bramek polikrzemowych.

**Technologia kanału tranzystora**

W najdawniejszych technologiach CMOS kanał był obszarem jednorodnie domieszkowanym w kierunku równoległym do powierzchni (między źródłem i drenem), a w kierunku prostopadłym rozkład domieszek był określony przez pojedynczy proces implantacji jonów, którego celem było zapewnienie właściwej wartości napięcia progowego. Dziś procesów domieszkowania obszaru kanału jest znacznie więcej. Dwa lub więcej procesów implantacji jonów kształtują dość skomplikowany rozkład domieszek w kierunku prostopadłym do powierzchni. Celem jest zarówno zapewnienie właściwej wartości napięcia progowego, jak i uzyskanie korzystnej zależności tego napięcia od napięcia polaryzacji podłoża tranzystora $U_{BS}$. W kierunku równoległym do powierzchni rozkład domieszek także nie jest równomierny. W pobliżu źródeł i drenów wytwarzane są obszary tego samego typu przewodnictwa, co źródła i dreny, ale o niższej koncentracji domieszek. Służą one takiemu ukształtowaniu natężenia pola elektrycznego w warstwach zaporowych źródeł i drenów, aby uniknąć ostrego maksimum tego natężenia. Nieco dalej od granicy obszarów źródeł i drenów wprowadzana jest domieszka tego samego typu, co podłoże tranzystora, w celu lokalnego zwiększenia jej koncentracji. Takie obszary zmniejszają zależność napięcia progowego tranzystora od długości kanału. W sumie kanał może być poddawany czterem, pięciu, a nawet większej liczbie procesów domieszkowania wykonywanych na różnych etapach procesu wytwarzania układu.

W ceku zwiększenia ruchliwości nośników w kanale poddaje się obszar kanału i jego okolice dodatkowym operacjom, których celem jest wywołanie w kanale niewielkich naprężeń mechanicznych o ściśle kontrolowanym kierunku i wielkości. Jedną z takich operacji może być np. wprowadzenie do kanału pewnej ilości atomów germanu - materiału, który jest półprzewodnikiem z tej samej grupy układu okresowego, co krzem, jego atomy nie są więc donorami ani akceptorami, ale mają większy promień atomowy od atomów krzemu, więc ich obecność w sieci monokrzyształu wywołuje naprężenia.

**Kontakty, źródła, dreny, połączenia**

Szczególne problemy przy zmniejszaniu wymiarów elementów stwarzają kontakty. Rezystancja kontaktu jest odwrotnie proporcjonalna do jego powierzchni. Gdy wymiary liniowe kontaktu maleją dwukrotnie, to jego rezystancja czterokrotnie wzrasta. Dlatego wymiary kontaktów maleją wolniej, niż wymiary kanałów tranzystorów. Sposoby wytwarzania dobrych, niezawodnych kontaktów o małej powierzchni i równocześnie małej rezystancji są jednym z istotnych problemów technologicznych współczesnej mikroelektroniki.

Obszary domieszkowane źródeł i drenów zachodzą pod obszar bramki, a odległość, na jaką zachodzą, jest proporcjonalna do głębokości tych obszarów. Toteż skracaniu kanałów tranzystorów musi towarzyszyć zmniejszanie głębokości źródeł i drenów. Problemy, jakie temu towarzyszą, są związane zarowno z technologią, jak i z elektrycznymi właściwościami tych obszarów. Bardzo płytkie obszary domieszkowane są trudne do wykonania, ponieważ w procesach produkcyjnych mikroelektroniki nie do uniknięcia są operacje wykonywane w wysokiej temperaturze, a każda taka operacja wywołuje dyfuzję domieszek, które z płytkiej i silnie domieszkowanej warstwy przypowierzchniowej dyfundują w głąb podłoża układu scalonego. Ponadto do bardzo płytkich obszarów domieszkowanych trudno wykonać dobre kontakty. Bardzo płytkie obszary źródeł i drenów cechuje znaczna rezystancja warstwowa, co pogarsza charakterystyki tranzystorów. Ten problem znajduje częściowe rozwiązanie w postaci warstw krzemowo-metalicznych (TiSi, PtSi i in.) wytwarzanych na powierzchni obszarów domieszkowanych. Zmniejszają one znacznie rezystancje tych obszarów.

Podstawowym metalem stosowanym do wykonywania połączeń w układach scalonych jest aluminium. Aluminium nie może być jednak stosowane jako metal bezpośrednio kontaktujący się z bardzo płytkimi obszarami domieszkowanymi, ponieważ atomy aluminium (które jest w krzemie domieszką akceptorową) mają tendencję do głębokiej penetracji i w przypadku kontaktów do obszarów bardzo płytkich są powodem zwarć. Dlatego kontakty są strukturami wielowarstwowymi, gdzie bezpośrednio z krzemem kontaktuje się inny metal, a dopiero z nim aluminium. Wpływ długich połączeń na opóźnienia w układach cyfrowych (patrz wykład 10) doprowadził do wprowadzenia miedzi jako metalu, z którego wykonuje się połączenia (miedź ma mniejszą od aluminium rezystywność). Technologia połączeń miedzianych jest jednak skomplikowana i kosztowna (patrz wykład 3).

Do niedawna w układach scalonych wykonywano co najwyżej dwie warstwy połączeń. We współczesnych technologiach CMOS tych warstw jest znacznie więcej - od pięciu do dziesięciu i więcej - co bardzo ułatwia uzyskanie dużej gęstości upakowania elementów. Przy dwóch warstwach połączeń właśnie połączenia, a nie tranzystory, zajmowały większą część powierzchni układu. Duża gęstość upakowania elementów sprzyja też zmniejszaniu długości połączeń, co ma wpływ na szybkość działania układu.

**Technologia CMOS SOI (SOI: ang. "Silicon on insulator" - "krzem na dielektryku")**

Jest to sposób produkcji układów CMOS różniący się klasycznego tym, że układy wytwarzane są w cienkiej warstwie krzemu, która oddzielona jest od krzemowej płytki podłożowej warstwą dielektryczną $SiO_2$. Technologia CMOS SOI w pewnym stopniu rozwiązuje niektóre z problemów technologii standardowych: tranzystory mają nieco lepsze parametry (w tym mniejszy prąd podprogowy), zredukowane są pojemnościach pasożytnicze, w znacznym stopniu wyeliminowane pasożytnicze sprzężenia przez podłoże (wykład 4). Wadą jest gorsze odprowadzanie ciepła i wyższy koszt.

**Rozrzuty produkcyjne**

Im bardziej skomplikowane i subtelne są procesy produkcyjne mikroelektroniki, tym łatwiej o zaburzenia, a więc tym większe są rozrzuty produkcyjne (patrz wykłady 2 i 12). Przy dzisiejszym stanie technologii mikroelektronicznych nie ma już żadnych możliwości zmniejszania niektórych rodzajów rozrzutów, ponieważ ich istnienie wynika z fundamentalnych praw przyrody. Przykładem jest domieszkowanie obszaru kanału tranzystora MOS. Można łatwo obliczyć, że w tranzystorach o wymiarach rzędu kilkudziesięciu nanometrów całkowita liczba atomów domieszek w elektrycznie aktywnym obszarze pod bramką tranzystora wynosi od zaledwie kilku do kilkudziesięciu atomów. A ponieważ tory atomów wprowadzanych do półprzewodnika metodą implantacji jonów mają charakter losowy, przy tak małej całkowitej liczbie atomów nieuchronne są znaczne losowe różnice w ich liczbie w poszczególnych tranzystorach. Prowadzi to do lokalnego rozrzutu napięcia progowego tranzystorów, którego nie da się usunąć ani nawet zmniejszyć żadnymi środkami technologicznymi. Rozrzut ten wynika bowiem wprost z ziarnistego charakteru budowy materii. Nie jest znany żaden proces technologiczny, który umożliwiałby wprowadzanie do półprzewodnika pojedynczych atomów w z góry określone, jednakowe w każdym tranzystorze położenia.

**Tranzystory bipolarne**

Tranzystory bipolarne mają pod wieloma względami przewagę nad tranzystorami MOS, zwłaszcza w układach analogowych (patrz wykłady 4, 12, 13, 14). Jednak szybki rozwój technologii CMOS, z jakim mamy do czynienia, spowodował znaczne zmniejszenie zainteresowania tymi tranzystorami poczynając od końca lat osiemdziesiątych XX wieku. Obecnie jednak tranzystory bipolarne znajdują nowe zastosowania - jako elementy czynne w układach do systemów komunikacji bezprzewodowej takich, jak telefonia GSM, Bluetooth, komputerowe sieci bezprzewodowe WiFi itp. Wymienione tu systemy działają w pasmach częstotliwości od 900 MHz do kilku GHz. Ten zakres częstotliwości jest już osiągalny dla najnowocześniejszych technologii CMOS, jednak tranzystory bipolarne pozwalają budować układy o wyższych parametrach, a ich maksymalne użyteczne częstotliwości pracy sięgają dziesiątków GHz. *Nie są* to jednak pasożytnicze struktury bipolarne w układach CMOS, o których była mowa w wykładzie 4. Do zastosowań w tym zakresie częstotliwości opracowano bardzo wyrafinowane technologie BiCMOS, w których oprócz standardowych tranzystorów nMOS i pMOS wytwarza się tranzystory bipolarne o częstotliwościach granicznych $f_T$ sięgających 200 GHz.

Warto wiedzieć, że jedna z wytwórni takich układów, dysponująca oryginalnymi opracowanymi w Europie technologiami, znajduje się we Frankfurcie nad Odrą, tuż obok polskiej granicy - patrz http://www.ihp-microelectronics.com/).

**Problemy projektowania**

Wielkim problemem mikroelektroniki jest nienadążanie metod projektowania za możliwościami stwarzanymi przez rozwój technologii. Szacuje się, że wydajność pracy projektanta, mierzona maksymalną liczbą tranzystorów w układzie zaprojektowanym w określonej jednostce czasu, od wielu lat rośnie o około 30% wolniej od maksymalnej liczby tranzystorów, jaką mogą posiadać układy w kolejnych generacjach technologii. To oznacza, że coraz trudniejsze, bardziej pracochłonne i kosztowne jest pełne wykorzystanie w nowych systemach możliwości stwarzanych przez postęp technologiczny. Jest to mało widoczna, ale w praktyce bardzo poważna bariera rozwojowa.

Nie jest to jedyny problem związany z projektowaniem. Problemem jest także jakość projektów. Im bardziej subtelne są technologie i wytwarzane przy ich użyciu elementy, im bardziej złożone są wykonywane z tych elementów układy, i im wyższe stawia się tym układom wymagania użytkowe, tym więcej różnorodnych zjawisk w układzie należy wziąć pod uwagę, aby mieć pewność, że zaprojektowany układ będzie działał zgodnie z oczekiwaniami. Trudne do uwzględnienia w projektowaniu układów są między innymi:

- zjawiska związane z transmisją sygnałów w długich połączeniach: wpływ pojemności, rezystancji i indukcyjności, sprzężeń między połączeniami, zakłóceń i przesłuchów (wykład 10),
- zjawiska związane z przenikaniem zakłóceń przez podłoże układu (wykład 4),
- wpływ rozrzutów produkcyjnych (wykłady 2 i 12),
- problemy testowania i testowalności (wykład 11).

Rozwiązań problemu pracochłonności projektowania i wydajności pracy projektanta poszukuje się obecnie przede wszystkim poprzez:

- doskonalenie metod i oprogramowania do automatycznej syntezy systemów cyfrowych na podstawie opisu funkcji systemu,
- przygotowywanie projektów w postaci przydatnej do powtórnego wykorzystania.

Ta ostatnia tendencja doprowadziła do wyodrębnienia się specyficznego podejścia do projektowania polegającego na tworzeniu i udostępnianiu tak zwanych bloków IP (nazwa pochodzi od ang. "Intellectual Property block" - co oznacza dosłownie "blok stanowiący własność intelektualną", czyli będący przedmiotem prawa autorskiego). Blok IP jest to najczęściej nadający się do automatycznej syntezy kod w języku opisu sprzętu (Verilog, VHDL - patrz wykład 5). Kod taki może być użyty do syntezy struktury logicznej i topografii układu dla różnych technologii wytwarzania. Istnieją i są dostępne (odpłatnie lub nawet bezpłatnie) tego typu bloki IP będące opisami popularnych mikrokontrolerów, układów peryferyjnych i różnych układów pomocniczych. Typowy sposób użycia takich bloków IP polega na połączeniu ich w zestaw funkcjonalny potrzebny w przewidywanym zastosowaniu, ewentualnym uzupełnieniu o projekty bloków, których brakuje, i automatycznym wygenerowaniu projektu specjalizowanego układu o architekturze ściśle dostosowanej do konkretnego zastosowania. Istnieją także bloki IP w postaci gotowych projektów topografii. Te mogą być oczywiście użyte tylko do zaprojektowania układów dla konkretnej technologii wytwarzania.

Przewiduje się, że te dwa kierunki rozwoju metod projektowania będą nadal intensywnie rozwijane, bo trudno sobie wyobrazić zaprojektowanie układu mającego $10^{10}$ elementów bez użycia automatycznej syntezy i powtarzalnych bloków.

**Konsekwencje ekonomiczne**

Ekonomicznym skutkiem wciąż rosnącej złożoności i subtelności procesów technologicznych z jednej strony, a trudności i pracochłonności projektowania coraz większych i szybszych układów z drugiej strony, jest gwałtownie wzrastający koszt projektów i prototypów układów. Przeciętny koszt projektu i prototypu układu przewidzianego do produkcji w technologii CMOS z bramką o długości 65 nm jest o blisko dwa rzędy wielkości większy od kosztu dla starszej technologii z bramką o długości 0,35 µm. To zaś oznacza, że technologie najbardziej zaawansowane mogą stać się niedostępne dla mniejszych firm elektronicznych, które będą zmuszone ograniczyć się w projektach układów ASIC do technologii starszych i mniej kosztownych. Poszukuje się nowych metodologii projektowania i prototypowania, które pozwoliłyby przezwyciężyć tę ekonomiczną barierę. Zauważmy jednak, że istnieje ogromnie dużo zastosowań, do których technologie najbardziej zaawansowane i kosztowne wcale nie są potrzebne.

Na zakończenie warto dodać, że w historii mikroelektroniki już wielokrotnie przewidywano bliski kres rozwoju zarówno ze względu na bariery techniczne, jak i ekonomiczne. Takim kresem miała być niegdyś długość kanału tranzystora równa 1 µm, a następnie 0,1 µm, i obie te granice zostały przekroczone. Prawdziwym kresem będzie osiągnięcie wymiarów rzędu kilku odległości międzyatomowych w sieci krystalicznej półprzewodnika. Zaczną wtedy dominować efekty kwantowe całkowicie zmieniające działanie elementów układu. W tym kierunku idą badania nad nanotechnologiami i nanoelektroniką - to znaczy nad opracowaniem zupełnie nowych, odmiennych od znanych dziś, metod przetwarzania informacji bezpośrednio wykorzystujących zjawiska kwantowe.

## 15.3. Nowe architektury i metody przetwarzania informacji

### Nowe sposoby realizacji systemów cyfrowych

Systemy cyfrowe zbudowane są ze sprzętu i działającego na tym sprzęcie oprogramowania. Tradycyjnie każdy z tych składników tworzy się osobno. Takie podejście ma szereg zalet, przede wszystkim elastyczność i uniwersalność. To samo oprogramowanie może być (do pewnych granic) użytkowane na różnych wersjach danej platformy sprzętowej. Ten sam sprzęt może być wykorzystany w różny sposób, a nawet do zupełnie różnych celów przez wymianę oprogramowania. Ceną elastyczności i uniwersalności jest jednak niska efektywność wykorzystania zasobów sprzętu i możliwości oprogramowania. Współczesne komputery dysponujące ogromnymi i wciąż rosnącymi możliwościami obliczeniowymi, graficznymi, multimedialnymi są przez dużą część użytkowników wykorzystywane głównie do prostych prac takich jak edycja tekstów czy też obsługa poczty elektronicznej...

W zastosowaniach, w których elastyczność i uniwersalność nie jest potrzebna, bo nie przewiduje się zmian funkcji użytkowych systemu, można postawić pytanie: czy te funkcje mają być realizowane przez oprogramowanie, czy też bezpośrednio przez sprzęt? Z jednej strony, możliwe jest użycie sprzętu standardowego: mikroprocesorów, mikrokontrolerów, pamięci i układów peryferyjnych i zrealizowanie wszystkich funkcji użytkowych poprzez oprogramowanie. Z drugiej strony, możliwa jest też realizacja funkcji użytkowych wyłącznie przez sprzęt, tj. przez układy elektroniczne zaprojektowane specjalnie do realizacji tych funkcji. Pierwsze podejście jest łatwiejsze do realizacji. Drugie ma jednak istotne zalety: czysto sprzętowa realizacja zapewnia dokładne dostosowanie sprzętu do wykonywanych funkcji, co daje produkt technicznie doskonalszy i w produkcji wielkoseryjnej zwykle mniej kosztowny. Jeśli jednak funkcje użytkowe są bardzo skomplikowane, czysto sprzętowa realizacja może być zbyt złożona, a zaprojektowanie odpowiednich układów niezwykle pracochłonne i kosztowne. W każdym konkretnym przypadku można więc postawić pytanie: jaka część funkcji użytkowych systemu ma być realizowana sprzętowo, a jaka przez oprogramowanie, i jakie powinny być elementy sprzętu realizujące owo oprogramowanie?

Do niedawna odpowiedź na takie pytanie oparta była głównie na intuicji i doświadczeniu projektantów systemów. Dziś dzięki postępom metod projektowania w mikroelektronice, w szczególności metod automatycznej syntezy układów cyfrowych, a także dzięki postępom w dziedzinie inżynierii oprogramowania, rozwijana jest metodologia projektowania systemów realizowanych w sposób mieszany, sprzętowo-programowy, zwana po angielsku *hardware-software codesign*. Istnieje oprogramowanie pozwalające opisywać przy definiowaniu funkcji i struktury systemu obie warstwy - sprzętową i programową - łącznie, i kontynuować proces projektowania aż do otrzymania zarówno projektu układów realizujących funkcje sprzętowe, jak i niezbędnego oprogramowania. Prowadzi to do układów będących kompletnymi funkcjonalnie systemami, zwanych "*system on chip*". Takie układy mogą zawierać bloki będące specjalizowanymi mikroprocesorami przeznaczonymi do wykonywania określonych operacji, potrzebne do wykonywanych funkcji układy peryferyjne i komunikacyjne (w tym - być może - także bloki analogowe), pamięci ROM i RAM oraz wbudowane oprogramowanie (ang. "*embedded software*"), czyli oprogramowanie zapisane na stałe w wewnętrznej pamięci ROM. Na świecie (a także w Polsce) projektuje się coraz więcej takich układów.

Praktyczna realizacja układów typu "system on chip" napotyka jednak na coraz większe trudności w miarę wzrostu złożoności tych układów. Jednym z wielkich i coraz trudniejszych problemów jest zapewnienie synchronicznej pracy układów o bardzo dużej powierzchni, ze względu na opóźnienia wynikające ze skończonych czasów propagacji sygnałów, w tym sygnałów zegarowych (patrz wykład 10). Odpowiedzią na to są układy lokalnie synchroniczne, ale w całości działające asynchronicznie. W takich układach każdy blok cyfrowy jest zaprojektowanym w tradycyjny sposób układem synchronicznym. Komunikacja między blokami odbywa się jednak asynchronicznie. Wymaga to potraktowania całego układu na podobieństwo sieci lokalnej, w której są działający niezależnie nadawcy i odbiorcy sygnałów, oraz zdefiniowania i zrealizowania protokołów komunikacji między nimi. Układy takie są nazywane "*network on chip*". Jest to koncepcja nowa, rozwijana od niedawna. Rozwijane są także metody projektowania układów całkowicie asynchronicznych (patrz wykład 14). Układy te nie weszły jeszcze do praktyki przemysłowej.

### Nietradycyjne metody przetwarzania informacji

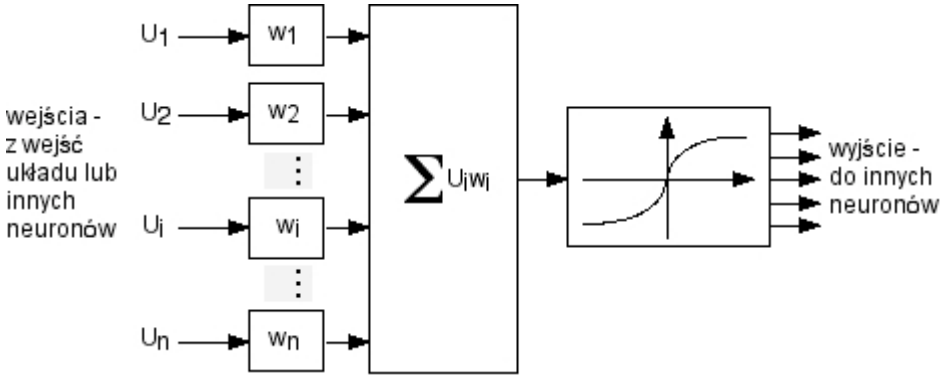
Zupełnie odmienną odpowiedzią na problemy złożoności, szybkości, poboru mocy itd. może być także zastosowanie niekonwencjonalnych metod przetwarzania informacji. Dwie z nich omówione są niżej.

### Sztuczne sieci neuronowe

Komputery były kiedyś nazywane w Polsce mózgami elektronowymi, jednak ich zasada działania ma niewiele wspólnego z działaniem sieci nerwowych, czyli biologicznych systemów przetwarzania informacji. Współczesne komputery, niezależnie od szczegółów ich architektury, wywodzą się w swej koncepcji od maszyny Turinga i komputera von Neumanna. Są to proste i bardzo eleganckie modele teoretyczne. Jednak ich podstawą jest zasada przetwarzania szeregowego. Każde zadanie jest podzielone na elementarne kroki wykonywane kolejno w czasie. Nie zmienia tej zasadniczej koncepcji istnienie mikroprocesorów wykonujących rownocześnie więcej niż jedną instrukcję, komputerów wieloprocesorowych i superkomputerów złożonych z tysięcy procesorów

pracujących równocześnie. Systemy biologiczne działają inaczej. Pojedyncze neurony, będące biologicznymi "bramkami logicznymi", działają bardzo wolno (ich "częstotliwość graniczna" jest na poziomie 10 Hz, dzięki czemu nie dostrzegamy nieciągłości ruchu obserwując film wyświetlany z częstotliwością 24 klatek na sekundę). Siłą systemów biologicznych jest masowe przetwarzanie równoległe. Ich "schemat logiczny" to olbrzymia liczba neuronów połączonych w trójwymiarową sieć, w której neurony wzajemnie komunikują się ze sobą. Chociaż wciąż bardzo daleko do pełnego zrozumienia, w jaki sposób taka sieć wykonuje błyskawicznie i bezbłędnie takie funkcje, jak rozpoznawanie obrazów czy rozumienie mowy, nie mówiąc już o bardziej złożonych funkcjach intelektualnych, to jednak działanie pojedynczych neuronów i ich prostych połączeń jest dość dobrze poznane. Można więc pokusić się o zbudowanie elektronicznych odpowiedników neuronów w nadziei, że połączenie w sieć dostatecznie wielkiej ich liczby da nam prawdziwy "mózg elektronowy", tj. układ do przetwarzania informacji działający tak, jak systemy biologiczne. Taką sztuczną sieć neuronową można by nauczyć sprawnego wykonywania różnych funkcji, z którymi doskonale sobie radzą biologiczne sieci neuronowe, a które wciąż stanowią problem dla istniejących obecnie komputerów i ich oprogramowania. Inną zaletą takich sieci jest ich odporność na uszkodzenia. Brak lub nieprawidłowe działanie pojedynczych neuronów nie ma większego wpływu na działanie sieci jako całości.

Elektronicznym modelem pojedynczego neuronu jest układ pokazany na rys. 15.4.




Rys. 15.4. Elektroniczny model neuronu

Sygnały z wejść są sumowane z wagami, a następnie suma jest podawana na wyjście przez układ nieliniowy o charakterystyce progowej. Wyjście łączy się z wejściami wielu innych neuronów. Funkcja realizowana przez taką sieć zależy od wartości współczynników wagowych, z jakimi podawane są na układ sumujący sygnały z poszczególnych wejść. Sieć można "nauczyć" wykonywania określonej funkcji zmieniając odpowiednio wartości tych współczynników. Dlatego sztuczna sieć neuronowa musi mieć, oprócz samych neuronów, także układy umożliwiające zmienianie i zapamiętywanie współczynników wagowych.

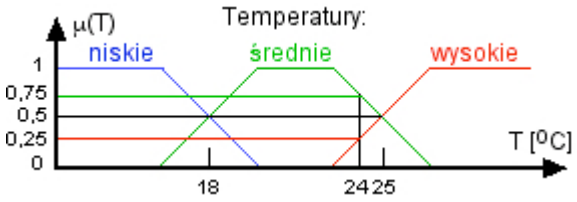
Układ taki, jak na rys. 15.4, można zbudować jako układ analogowy (wtedy sygnałami są wartości napięć lub prądów na wejściach i wyjściach). Można również zbudować układ cyfrowy realizujący opisaną funkcję. Na wejściach są wówczas liczby reprezentowane binarnie. Układ analogowy jest znacznie prostszy, ale działa mniej dokładnie (co zresztą na ogół nie ma w sztucznych sieciach neuronowych większego znaczenia). Problemem w układach analogowych jest realizacja pamięci wag (ponieważ nie jest znany prosty sposób realizacji pamięci sygnałów analogowych). Istnieje wiele układów realizujących ideę sztucznych sieci neuronowych zarówno w wersji analogowej, jak i cyfrowej. Jednak są to układy zbyt małe, aby ich działanie mogło być porównywalne pod względem skuteczności do działania sieci biologicznych. Realizowane są układy mające od kilkuset do kilku tysięcy sztucznych neuronów. Ponadto połączenia w sieciach biologicznych są bardziej złożone, ich trójwymiarowej struktury nie daje się odwzorować w sztucznych sieciach neuronowych. Mimo to podstawowe właściwości sieci biologicznych, takie jak efekt "uczenia się" sieci, dają się zademonstrować w sieciach sztucznych. Również komputerowe symulacje działania sztucznych sieci neuronowych pokazują, że sieci takie są w stanie naśladować działanie sieci biologicznych - pod warunkiem osiągnięcia odpowiedniej złożoności (liczby neuronów).

Sztuczne sieci neuronowe są dobrym przykładem poszukiwań nowych sposobów przetwarzania informacji, inspirowanych obserwacjami działania żywych organizmów. Jednak zakres praktycznych zastosowań sztucznych sieci neuronowych jest, jak na razie, dość ograniczony.

**Układy logiki rozmytej**

W przeciwieństwie do sztucznych sieci neuronowych układy logiki rozmytej (ang. *fuzzy logic*) okazały się bardzo przydatne w wielu praktycznych zastosowaniach. Służą one do budowy układów sterowania złożonymi obiektami, a także do rozwiązywania problemów diagnostyki, klasyfikacji itp. Pozwalają skutecznie poradzić sobie z problemami, dla których nie jest znany dobry opis matematyczny lub też opis ten jest tak skomplikowany, że nie daje się w praktyce wykorzystać.

Zasada działania układów logiki rozmytej wywodzi się z obserwacji, że człowiek jest w stanie praktycznie sterować obiektami lub procesami o dużej złożoności nie rozwiązając żadnych równań matematycznych, lecz kierując się jedynie zbiorem prostych reguł opartych na doświadczeniu i praktycznym treningu. Ta obserwacja legła u podstaw formalizmu matematycznego zwanego teorią *zbiorów rozmytych* (ang. *fuzzy sets*). Zbiór rozmyty jest uogólnieniem pojęcia zbioru w rozumieniu klasycznej teorii mnogości. W klasycznej teorii mnogości element E należy do zbioru Z lub nie, nie ma innej możliwości. W teorii zbiorów rozmytych element E charakteryzuje się *stopniem przynależności do zbioru*, który jest wyrażony liczbą z przedziału [0,1]. Takie uogólnienie pojęcia zbioru pozwoliło wprowadzić reprezentację danych przybliżonych i ich nieostrą klasyfikację, z jaką najczęściej mamy do czynienia w życiu. Można to zilustrować przykładem podziału temperatury w pomieszczeniu na trzy zbiory: temperatur niskich, średnich i wysokich. Elementami tych zbiorów są wartości temperatury wyrażone na przykład w stopniach Celsjusza. Gdyby te trzy zbiory temperatur były zbiorami w tradycyjnym rozumieniu, trzeba by zdefiniować ostre granice, na przykład zakładając, że temperatury niskie to wszystkie temperatury niższe od +18 $^{o}$C, wysokie to wszystkie wyższe od +25 $^{o}$C, a pozostałe należą do zbioru temperatur średnich. Wtedy temperatura +18 $^{o}$C jest średnia, podobnie jak +25 $^{o}$C, temperatura +17,99 $^{o}$C jest niska, a +25,01$^{o}$C - wysoka. W przypadku zbiorów rozmytych możemy zdefiniować stopień przynależności temperatur do zbiorów przy użyciu *funkcji przynależności*, na przykład takich, jak na rys. 15.5. Funkcja przynależności przypisuje każdemu elementowi zbioru wartość stopnia przynależności do tego zbioru.




Rys. 15.5. Ilustracja pojęcia zbioru rozmytego na przykładzie podziały temperatur na trzy zbiory rozmyte: temperatur niskich, średnich i wysokich.

Na rysunku pokazano przykładowe funkcje przynależności do tych zbiorów. Temperatura 24 $^{o}$C należy do zbioru temperatur średnich w stopniu 0,75 i do zbioru temperatur wysokich w stopniu 0,25.

Przy funkcjach przynależności zdefiniowanych tak, jak na rys. 15.5, można powiedzieć, że temperatura +18 $^{o}$C należy ze stopniem przynależności 0,5 do zbioru temperatur średnich, i ze stopniem przynależności 0,5 do zbioru temperatur niskich. Im niższa temperatura, tym niższy jej stopień przynależności do zbioru temperatur średnich, a wyższy - do zbioru temperatur niskich. Taka rozmyta klasyfikacja bardziej odpowiada naszej intuicji mówiącej, że nie ma jakiejś naturalnej ostrej granicy między temperaturami niskimi, a średnimi. Podobnie dla temperatur średnich i wysokich.

Dla zbiorów rozmytych definiuje się operacje logiczne NOT, AND i OR poprzez zdefiniowanie odpowiednich działań na funkcjach przynależności:

B = NOT A jeśli $\mu_B = 1 - \mu_A$

C = A AND B jeśli $\mu_C = \min(\mu_A, \mu_B)$

C = A OR B jeśli $\mu_C = \max(\mu_A, \mu_B)$

gdzie A, B, C są to zbiory rozmyte, zaś $\mu_A$, $\mu_B$ i $\mu_C$ są to funkcje przynależności do tych zbiorów.

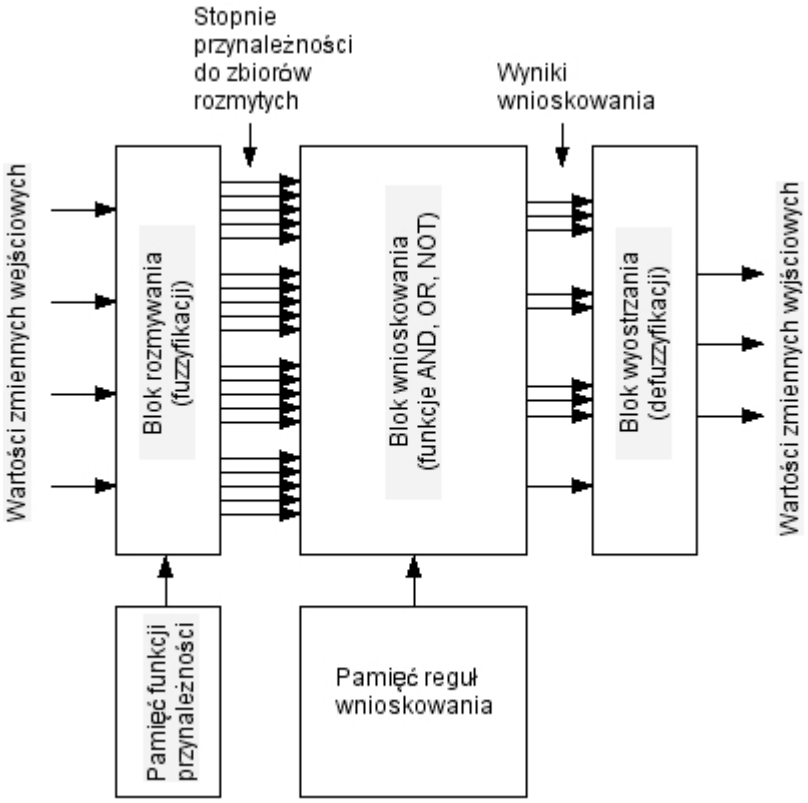
Przypuśćmy, że chcemy opisać pewien algorytm klasyfikacji lub sterowania w terminologii zbiorów rozmytych. Niech będzie to na przykład algorytm sterowania grzejnikiem w pomieszczeniu. Należy zdefiniować zbiory rozmyte dla temperatury wewnątrz i na zewnątrz pomieszczenia (są to zmienne wejściowe) oraz dla ustawienia pokrętła regulatora (jest to zmienna wyjściowa). Następnie należy określić *reguły wnioskowania*. Reguły takie mają postać, która może być zilustrowana przykładem następujących dwóch reguł:

JEŻELI temperatura wewnętrzna NALEŻY DO ZBIORU temperatur "niskich" I temperatura zewnętrzna NALEŻY DO ZBIORU temperatur "upał" TO ustawienie pokrętła NALEŻY DO ZBIORU "brak grzania".

JEŻELI temperatura wewnętrzna NALEŻY DO ZBIORU temperatur "niskich" I temperatura zewnętrzna NALEŻY DO ZBIORU temperatur "mróz" TO ustawienie pokrętła NALEŻY DO ZBIORU "maksymalne grzanie".

W tego rodzaju regułach słowa NALEŻY DO ZBIORU oznaczają określenie stopnia przynależności, a słowo I oznacza funkcję AND. W regułach wnioskowania może wystąpić także słowo LUB, czyli funkcja OR. Dla konkretnych wartości zmiennych wejściowych przy danych regułach wnioskowania określony jest kształt funkcji przynależności do zbiorów rozmytych dla zmiennej wyjściowej. Pozwala to na określenie konkretnej wartości tej zmiennej - w naszym przykładzie właściwego ustawienia pokrętła regulatora grzejnika. Ten przykład jest trywialny, lecz w ten sam sposób definiuje się algorytmy sterowania dla bardzo złożonych obiektów lub procesów. Reguły te

definiuje w języku naturalnym człowiek - ekspert, który potrafi sterować danym obiektem. Zdefiniowanie zbiorów rozmytych dla zmiennych wejściowych i wyjściowych oraz reguł wnioskowania wystarcza, by zbudować układ realizujący algorytm sterowania. Taki układ ma następującą ogólną strukturę:




Rys. 15.6. Struktura układu sterowania rozmytego. W bloku rozmywania aktualnym wartościom zmiennych wejściowych przypisywane są stopnie przynależności do zbiorów rozmytych. Blok wnioskowania wykonuje na wartościach stopni przynależności operacje wynikające z reguł wnioskowania. Blok wyostrzania konstruuje kształt wyjściowych funkcji przynależności i na tej podstawie określa wartości zmiennych wyjściowych.

Zauważmy, że operacje logiczne na zbiorach rozmytych są równoważne prostym operacjom arytmetycznym na wartościach stopni przynależności, które mogą przybierać dowolne wartości z przedziału [0,1]. Operacje wykonywane w blokach układu sterowania rozmytego są operacjami o charakterze analogowym i mogą być realizowane przez odpowiednio zaprojektowane układy analogowe. Zmienne wejściowe, wyjściowe oraz wartości stopni przynależności są wówczas reprezentowane przez zmieniające się w sposób ciągły wartości napięć (lub prądów). Można również budować układy cyfrowe. Zmienne wejściowe, wyjściowe oraz wartości stopni przynależności są wówczas reprezentowane przez liczby wyrażone w postaci binarnej. Pominiemy tu szczegóły realizacji zarówno układów w wersji analogowej, jak i cyfrowej.

Metody logiki rozmytej wykazały mimo swej prostoty zdumiewającą skuteczność w wielu trudnych zadanich technicznych. Układy sterowania oparte na tych metodach są obecnie produkowane jako uniwersalne, programowalne układy analogowe lub (częściej) cyfrowe, oraz jako układy specjalizowane realizujące jeden konkretny algorytm. Mimo iż logika rozmyta jest w pewnym sensie uogólnieniem tradycyjnej logiki dwuwartościowej, a blok wnioskowania jest odpowiednikiem tradycyjnego kombinacyjnego układu cyfrowego, układy realizujące algorytmy logiki rozmytej nie zastąpią tradycyjnych układów cyfrowych. Stanowią jednak bardzo cenne ich uzupełnienie.

## 15.4. Podsumowanie

Skromna objętość wykładu nie pozwoliła omówić wielu interesujących i ważnych tematów, a wiele innych z konieczności potraktowanych było powierzchownie. Warto mieć świadomość, co z zagadnień istotnych i interesujących zostało pominięte.

**Układy RF (skrót od ang. "Radio Frequency") i mikrofalowe**

Monolityczne krzemowe układy CMOS i BiCMOS mogą być dziś wykorzystywane przy częstotliwościach znacznie przekraczających 1 GHz, a więc należących już do zakresu mikrofal. Nie omawialiśmy specyfiki tego rodzaju układów ani ich projektowania. Jest ona częściowo reprezentowana w innym przedmiocie. Jest to dziedzina mikroelektroniki obecnie bardzo szybko rozwijająca się ze względu na liczne zastosowania.

Obok elementow czynnych oraz elementów biernych R i C w układach RF i mikrofalowych stosowane są też scalone indukcyjności o wartościach rzędu pojedynczych nH. Nie omawialiśmy ich w tym wykładzie.

**Układy SC (skrót od ang. "Switched Capacitors")**

Jest to specyficzna grupa układów analogowych zwana układami z przełączanymi pojemnościami. W tych układach wykorzystuje się przepływy ładunków pomiędzy pojemnościami sterowane przez odpowiednio włączane i wyłączane tranzystory MOS. Owe przepływy ładunków pozwalają symulować w układach rezystory o dużych rezystancjach. W połączeniu z kondensatorami oraz układami wzmacniającymi można tworzyć różnorodne filtry aktywne typu RC. Układy z przełączanymi pojemnościami są stosowane do budowy filtrów działających przy częstotliwościach niskich i średnich. Filtry takie można przestrajać zmieniając częstotliwość zegara sterującego przełączaniem pojemności. Do wytwarzania takich układów stosowane są specjalne odmiany technologii CMOS, w których wytwarza się dwie warstwy polikrzemu. Przełączane pojemności realizowane są jako kondensatory zbudowane z dwóch warstw polikrzemu rozdzielonych bardzo cienką warstwą dielektryczną ($SiO_2$).

**Układy małej mocy**

W niektórych zastosowaniach moc pobierana z zasilania musi być ekstremalnie mała (na przykład we wszczepialnych urządzeniach elektromedycznych, jak stymulatory serca itp., pobór prądu powinien być na poziomie pojedynczych µA). Istnieją specjalne metody projektowania takich układów, zarówno analogowych, jak i cyfrowych.

**Układy dużej mocy i wysokonapięciowe**

Wprawdzie w wykładzie 14 wspomniane były stopnie wzmacniające dużej mocy, jednak wiele problemów technicznych występujących przy projektowaniu takich stopni nie zostało nawet zasygnalizowanych. W niektórych zastosowaniach (np. w elektronice motoryzacyjnej) potrzebne są układy o wyższych napięciach zasilania - przystosowane do pracy przy napięciach rzędu kilkunastu - kilkudziesięciu V, a także układy wytrzymujące przepięcia rzędu kilkudziesięciu V zarówno występujące w zasilaniu, jak i na wejściach i wyjściach układu. O specyfice takich układów nie było mowy.

**Układy wykonujące różne bardziej złożone funkcje analogowe**

W lekcjach poświęconych układom analogowym przedstawione zostały jedynie niektóre podstawowe bloki pomocnicze i funkcjonalne pozwalające zorientować się ogólnie w problematyce układów analogowych. Istnieje wielka liczba układow wykonujących różne funkcje bardziej złożone, ale z konieczności trzeba je było pominąć.

**Automatyczna synteza układów cyfrowych**

Była ona kilkakrotnie wspominana, ale bez omawiania szczegółów, a przede wszystkim bez przedstawienia języków opisu sprzętu takich, jak Verilog i VHDL. Języki opisu sprzętu są przedstawione (co prawda w nieco innym zastosowaniu) w innych przedmiotach.

---

**Wykład "Układy scalone" - i co dalej?**

Jeśli interesuje Cię specjalizacja w dziedzinie projektowania sprzętu elektronicznego, warto poznać znacznie głębiej problemy projektowania specjalizowanych układów scalonych. W dziedzinie układów cyfrowych należy przede wszystkim zapoznać się z metodami automatycznej syntezy (a w tym z metodami projektowania i realizacji układów cyfrowych w technice układów programowalnych FPGA). W dziedzinie układów analogowych szczegółowych specjalizacji jest wiele. Jedną z najbardziej aktualnych jest obecnie specjalizacja w dziedzinie projektowania układów wielkiej częstotliwość (RF) i mikrofalowych.

Wyrażając nadzieję, że przedmiot "Układy scalone" stanowi dobry wstęp do opanowywania umiejętności projektowania w mikroelektronice, autor tego podręcznika życzy Ci sukcesów w dalszych studiach i powodzenia w zawodzie elektronika, który jest jednym z najbardziej fascynujących zawodów naszych czasów.

## Bibliografia

[1] International Technology Roadmap for Semiconductors

Jest to dostępny w Internecie ( http://public.itrs.net/) obszerny zbiór dokumentów opracowany przez duży zespół specjalistów z całego świata i aktualizowany co dwa - trzy lata, omawiający stan i perspektywy rozwojowe mikroelektroniki we wszystkich jej aspektach. Stanowi niezastąpione źródło informacji, definiuje główne kierunki badań i cele do osiągnięcia.