2019

Układy scalone

CZĘŚĆ IV: O UKŁADACH ANALOGOWYCH, A TAKŻE O PRZYSZŁOŚCI MIKROELEKTRONIKI

WIESŁAW KUŹMICZ

UKŁAD SCALONY, UKŁAD ANALOGOWY, ŹRÓDŁO PRĄDOWE, ŹRÓDŁO NAPIĘCIOWE, WZMACNIACZ RÓŻNICOWY, WZMACNIACZ OPERACYJNY, KOMPARATOR NAPIĘCIA, UKŁAD MNOŻĄCY, UKŁAD DUŻEJ MOCY, POBÓR MOCY, CHŁODZENIE UKŁADÓW SCALONYCH, FINFET, FDSOI, TRAZNYSTOR TUNELOWY, TRANZYSTOR BEZZŁĄCZOWY

CZĘŚĆ CZWARTA MATERIAŁÓW OMAWIAJĄCYCH UKŁADY SCALONE POŚWIĘCONA JEST GŁÓWNIE UKŁADOM ANALOGOWYM. OMÓWIONE SĄ SPECYFICZNE PROBLEMY PROJEKTOWANIA TAKICH UKŁADÓW I TYPOWE ROZWIĄZANIA TYCH PROBLEMÓW. PRZEDSTAWIONE SĄ PODSTAWOWE BLOKI ANALOGOWE, W TYM TAKŻE UKŁADY WYJŚCIOWE DUŻEJ MOCY. OSTATNIA CZĘŚĆ JEST POŚWIĘCONA PRZEDSTAWIENIU KIERUNKÓW DALSZEGO RZOWOJU MIKROELEKTRONIKI.

Spis treści

1 Wstęp: o czym tu będzie mowa

Droga Czytelniczko, Drogi Czytelniku: w ostatniej części materiałów, w których omawiane są układy scalone, będzie mowa głównie o układach analogowych realizowanych w postaci scalonych układów CMOS. 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. Każdy rodzaj układów analogowych ma swoją własną teorię i metodologię projektowania. Nawet w obrębie tej samej klasy układów analogowych wymagania mogą być bardzo różne. Weźmy dla przykładu układy wzmacniaczy. W jednym zastosowaniu najważniejsza może być wartość wzmocnienia napięciowego, w innym – wzmocnienia mocy, w jeszcze innym – maksymalnej częstotliwości wzmacnianego sygnału, w jeszcze innym – poziom szumów własnych, a może być i tak, że jednakowo ważne mogą być różne parametry, co zmusza do kompromisów projektowych. W każdym z tych przypadków inny może być schemat układu, odmienne wzory projektowe, inne i inaczej projektowane elementy układu. Dlatego w naszym przeglądzie problematyki układów analogowych poruszone będą tylko pewne ogólne zasady oraz przykłady najprostszych rozwiązań kilku rodzajów najczęściej spotykanych rodzajów układów analogowych. Będą się tu pojawiać także układy z tranzystorami bipolarnymi, które – jak zobaczymy – w układach analogowych mogą spisywać się lepiej od tranzystorów MOS.

Na zakończenie omówimy także główne problemy, z jakimi mamy do czynienia na drodze do dalszego rozwoju mikroelektroniki. Jednym z nich, bardzo poważnym, jest problem wydzielanej w układach scalonych mocy, która już sporo lat temu osiągnęła poziom niemożliwy do przekroczenia, toteż zajmiemy się tym problemem. Omówimy też inne bariery rozwojowe. Spróbujemy przewidzieć, jaka przyszłość czeka mikroelektronikę.

2 Specyfika i problemy projektowania układów analogowych

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 i przewodowej (WiFi, Bluetooth, Ethernet, USB itp.)
- 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. analogowe sztuczne sieci neuronowe, analogowe 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.

2.1 Układy analogowe a układy cyfrowe

Najważniejsze różnice między układami cyfrowymi, a analogowymi z punktu widzenia projektanta ujmuje tabela poniżej:

Tabela 2-1. Różnice między układami cyfrowymi, a analogowymi

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
W układach nie występują sygnały zmienne o małej amplitudzie.	W wielu układach występują sygnały zmienne o małej amplitudzie, co wymaga wprowadzenia pojęcia parametrów małosygnałowych elementów oraz rozróżniania składowej stałej i składowej zmiennej napięć i prądów
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: rezystory, 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 mniej rozwinięte
Testowanie, choć czasochłonne, jest koncepcyjne proste i wymaga jedynie odróżniania poziomu napięcia dla „0" i „1"	Testowanie trudne, wymagające precyzyjnych analogowych pomiarów

2.2 Problemy projektowania specyficzne dla układów analogowych

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.

2.2.1 Zależności temperaturowe parametrów elementów i ich skutki

Dobra rada: zanim zaczniesz czytać dalej, przypomnij sobie (część I, punkty 3.1.5 i 3.1.6) podstawowe wiadomości o charakterystykach tranzystorów MOS i bipolarnych.

W tranzystorach MOS zależne od temperatury są: napięcie progowe V_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/^{0}C. 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 tranzystora maleje z temperaturą.

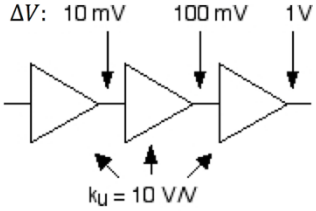
W tranzystorach bipolarnych bardzo silnie (wykładniczo) wzrasta z temperaturą stała J_{ES0} określająca przebieg charakterystyki $I_C = f(V_{BE})$. W rezultacie przy niezależnych od temperatury napięciach polaryzujących prąd kolektora wykładniczo wzrasta z temperaturą. Jeśli zaś prąd kolektora w układzie jest wymuszony w taki sposób, że nie zależy od temperatury, to napięcie V_{BE} maleje z temperaturą w przybliżeniu liniowo, o około 2 mV/^{0}C. 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 (na co jednak projektant układu nie ma żadnego wpływu).

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 więc 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%/^{0}C ... 0,3%/^{0}C (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 trzystopniowy wzmacniacz (rysunek 2-1; trójkąty są symbolami wzmacniacza stosowanymi w układach analogowych), którego każdy stopień ma wzmocnienie napięciowe $k_u = 10$. Niech na wyjściu pierwszego stopnia wzmacniacza wystąpi wywołana zmianą temperatury zmiana składowej 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!



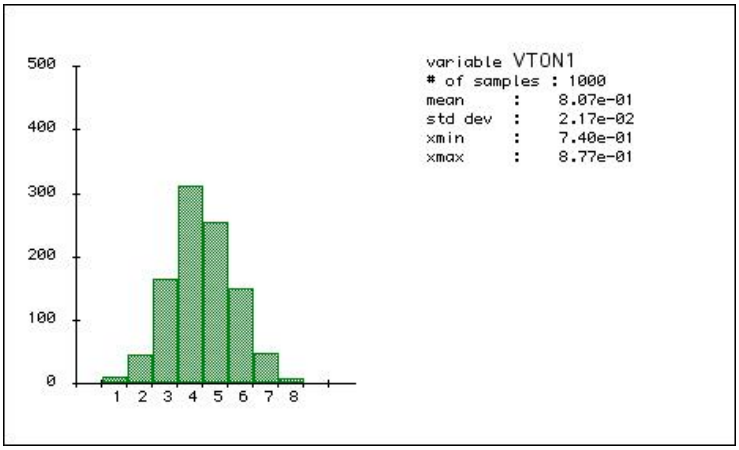
Rysunek 2-1. Ilustracja zjawiska zmian składowych stałych napięcia wywołanych zmianami temperatury

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 potencjałów 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 dostatecznie wysoka. Przykłady poznamy dalej.

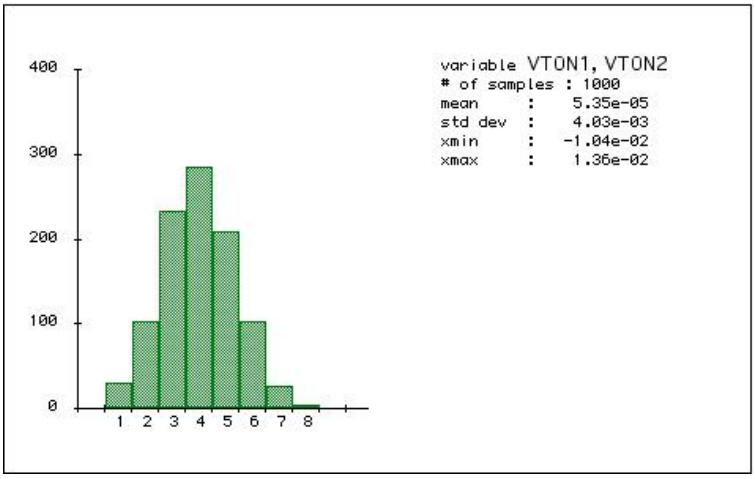
2.2.2 Rozrzuty parametrów tranzystorów MOS

Omawiając rozrzuty produkcyjne skupimy się na rozrzutach lokalnych, tj. takich, które powodują, że elementy zaprojektowane jako identyczne w rzeczywistości nie mają identycznych charakterystyk i parametrów – a to dlatego, że układy scalone konstruuje się tak, by były niewrażliwe na rozrzuty globalne (co to rozrzuty globalne i lokalne? Przypomnij sobie część I, punkt 4.2.6).

Rozrzut prądu drenu tranzystora MOS wynika z rozrzutów następujących wielkości: napięcia progowego V_T, ruchliwości nośników w kanale μ, jednostkowej pojemności tlenku bramkowego C_{ox} oraz wymiarów kanału W i L (część I, zależności 3-4 i 3-5). Wszystkie te rozrzuty mają składową globalną i lokalną. Przykładowo, rysunek 2-2 pokazuje rozrzut całkowity (sumę rozrzutu globalnego i lokalnego) napięcia progowego tranzystora nMOS w pewnej technologii CMOS, zaś rysunek 2-3 pokazuje histogram rozrzutu lokalnego, czyli różnicy napięć progowych dwóch identycznych tranzystorów znajdujących się obok siebie w tym samym układzie. Wartość średnia tej różnicy jest bardzo bliska zeru (0,0535 mV), miarą wielkości rozrzutu jest odchylenie standardowe, które wynosi około 4 mV.



Rysunek 2-2. Przykładowy histogram rozrzutu wartości napięcia progowego tranzystorów nMOS dla próbki 1000 tranzystorów z różnych płytek i partii produkcyjnych



Rysunek 2-3. Przykładowy histogram rozrzutu lokalnego wartości napięcia progowego tranzystorów nMOS dla próbki 1000 par tranzystorów. Histogram pokazuje rozrzut różnicy napięć progowych pomiędzy dwoma tranzystorami pary

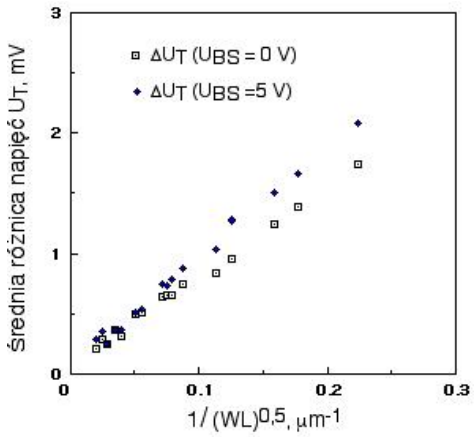
Rozrzuty lokalne w tranzystorze MOS wygodnie jest scharakteryzować odnosząc je do napięcia niezrównoważenia.

DEFINICJA

Napięciem niezrównoważenia pary tranzystorów MOS nazywamy różnicę napięć bramki względem źródła V_{GS}, jakie trzeba przyłożyć do bramek dwóch tranzystorów, aby przy jednakowych wartościach napięcia drenu względem źródła V_{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 rysunku 2-3) 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 rysunek 2-4. Zależność ta nosi w mikroelektronice nazwę „prawa Pelgroma".



Rysunek 2-4. Przykład zależności lokalnego rozrzutu napięcia progowego od powierzchni bramki tranzystora MOS

Teraz oszacujemy wpływ rozrzutu wymiarów kanału tranzystora na napięcie niezrównoważenia. Dla oszacowania tego wpływu założymy, że dwa pod każdym innym względem identyczne (w szczególności mające identyczne napięcia progowe) tranzystory MOS różnią się długością kanału o wartość ΔL. Dla uzyskania jednakowych prądów drenu trzeba wtedy spolaryzować bramki tranzystorów napięciami V_{GD} różniącymi się o napięcie niezrównoważeniaΔV_{GS}. Wartość tę (dla tranzystora pracującego w zakresie nasycenia) można wyznaczyć różniczkując zależność 3-5 z części I i zamieniając pochodne na małe przyrosty. Otrzymujemy

Równanie 2-1

$$|\Delta V_{GS}| = \frac{V_{GS}-V_T}{2}\left|\frac{\Delta L}{L}\right|$$

Podobną zależność można wyznaczyć dla różnicy szerokości kanałów ΔW:

Równanie 2-2

$$|\Delta V_{GS}| = \frac{V_{GS}-V_T}{2}\left|\frac{\Delta W}{W}\right|$$

Wzory 2-1 i 2-2 pokazują, że dla uzyskania małego napięcia niezrównoważenia (rzędu pojedynczych miliwoltów) fotolitografia musi być niezwykle precyzyjna. Dla uzyskania $\Delta V_{GS} < 1$ mV, przy $V_{GS} - V_T = 2$ V, otrzymujemy warunek $\Delta L/L < 0{,}001$. Warunek ten jest możliwy do spełnienia tylko wtedy, gdy długość kanału L jest wielokrotnie większa od minimalnej długości dopuszczalnej w danej technologii.

„Prawo Pelgroma" (rysunek 2-4) oraz wzory 2-1 i 2-2 pokazują, że dla uzyskania małej wartości napięcia niezrównoważenia pary identycznych tranzystorów MOS powierzchnie kanałów tych tranzystorów 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/^{0}C, różnica temperatur wynosząca 1^{0}C powoduje powstanie napięcia niezrównoważenia o takiej właśnie wartości.

2.2.3 Rozrzuty parametrów tranzystorów bipolarnych

Podobne oszacowanie zrobimy dla tranzystora bipolarnego.

DEFINICJA

Napięciem niezrównoważenia pary tranzystorów bipolarnych nazywamy różnicę napięć bazy względem emitera V_{BE}, jakie trzeba przyłożyć do baz dwóch tranzystorów, aby przy jednakowych wartościach napięcia kolektora względem emitera V_{CE} otrzymać jednakowe wartości prądu kolektora I_C.

Użyjemy tu wzorów 3-12 i 3-13 z części I, w których dla uproszczenia pominiemy rozrzut parametru J_{ES0}. Pozostaje wówczas do rozważenia rozrzut powierzchni złącza emiter-baza A_E. Z tych wzorów wynika, że jeśli dwa pod każdym innym względem identyczne tranzystory różnią się powierzchnią złącza emiter-baza o ΔA_E, to dla otrzymania jednakowych wartości prądu kolektora potrzebna jest różnica napięć V_{BE} wynosząca

Równanie 2-3

$$|V_{BE}| = \frac{kT}{q}\left|\frac{\Delta A_E}{A_E}\right|$$

W temperaturze otoczenia wartość kT/q wynosi około 26 mV, więc dla uzyskania $\Delta V_{BE} < 1$ 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 V_{BE} przy stałej wartości prądu kolektora zmienia się o około 2 mV/^{0}C, różnica temperatur wynosząca 1^{0}C powoduje powstanie napięcia niezrównoważenia o takiej właśnie wartości.

2.2.4 Rozrzuty produkcyjne rezystancji rezystorów

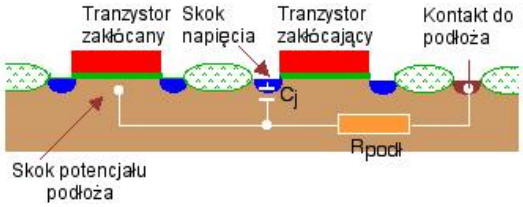
Rezystory występują często w układach analogowych. Omówimy rezystory wykonywane jako ścieżki polikrzemowe. Rozrzut globalny rezystancji takich rezystorów jest bardzo duży, może sięgać ±25% ... ±40%. Możliwe jest natomiast osiągnięcie rozrzutu lokalnego, czyli różnicy rezystancji dwóch identycznych, położonych 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.

2.2.5 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.

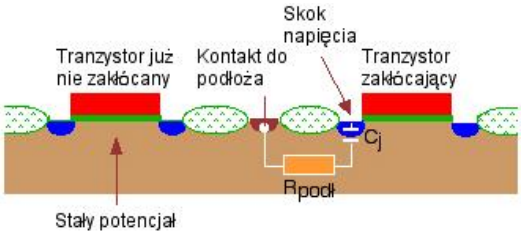
2.2.6 Elementy i sprzężenia pasożytnicze

Elementy i sprzężenia pasożytnicze były wspomniane w części pierwszej, punkt 3.1.8. 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 takie oddziaływania pasożytnicze, których uwzględnienie jest znacznie trudniejsze. Należą do tej grupy sprzężenia przez podłoże. Teraz je omówimy. Mechanizm sprzężenia przez podłoże ilustruje rysunek 2-5.



Rysunek 2-6. Mechanizm sprzężenia przez podłoże. Rezystor symbolizuje rezystancję rozproszoną między drenem tranzystora zakłócającego, a kontaktem uziemiającym podłoże

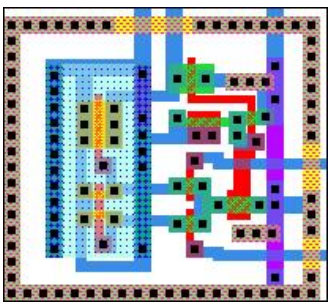
Mechanizm ten można opisać następująco. Jeśli na obszarze drenu tranzystora zakłócającego pojawi się skok napięcia, to zmieni się ładunek zgromadzony w pojemności złącza *p-n* dren-podłoże. Spowoduje to przepływ impulsu prądu przez podłoże, a ponieważ podłoże jest obszarem o dość znacznej rezystywności, wystąpić musi spadek napięcia. Potencjał podłoża w okolicy tranzystora zakłócanego zmieni się, a tym samym zmieni się napięcie V_{BS} tego tranzystora. To, jak wiemy, powoduje zmianę napięcia progowego tranzystora (patrz wzór 3-6 w części I). Zmiana napięcia progowego wywoła zmianę wartości prądu drenu. W ten sposób zakłócenia przenoszą się przez podłoże (podłoże w tym przypadku może oznaczać 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 rysunek 2-6.



Rysunek 2-5. 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" zakłóceń, 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 gęsto 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 (rysunek 2-7). W przypadku podłoża są to kontakty uziemiające, a w przypadku obszaru wyspy – kontakty połączone z plusem zasilania V_{DD}. Jeżeli w tym samym układzie występują zarówno bloki cyfrowe, jak i analogowe, otoczenie bloków cyfrowych i analogowych pierścieniami ochronnymi jest z reguły niezbędne. W układach cyfrowych CMOS występują bowiem skoki napięcia od zera do napięcia zasilania V_{DD}, toteż układy te są źródłem zakłóceń o dużej amplitudzie.

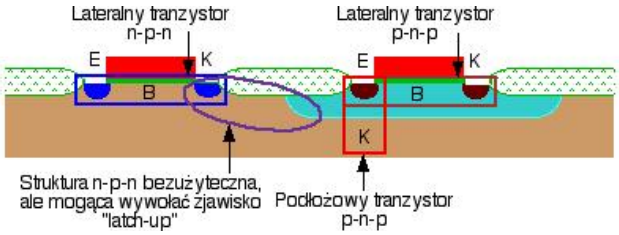


Rysunek 2-7. Mały blok analogowy otoczony pierścieniem kontaktów uziemiających podłoże

2.2.7 Struktury bipolarne szkodliwe i pożyteczne

We współczesnej mikroelektronice królują układy CMOS, więc zajmiemy się tranzystorami bipolarnymi w tych układach. W układzie scalonym CMOS tranzystory bipolarne występują zwykle tylko jako elementy pasożytnicze, ale niektóre z nich mogą być wykorzystane jako aktywne elementy w układzie.

W strukturze układu CMOS można wyróżnić struktury bipolarne *n-p-n* i *p-n-p* pokazane na rysunku 2-8.

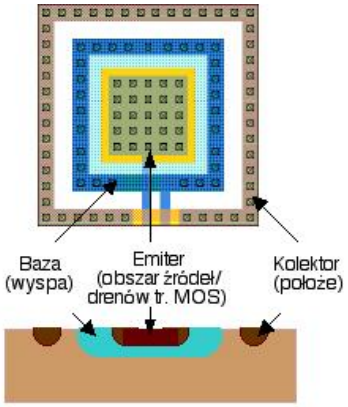


Rysunek 2-8. Struktury bipolarne występujące w układach CMOS

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 rysunku 2-8 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 źródła/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 (ang. „latch-up"). 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 (patrz poprzedni punkt – sprzężenia przez podłoże). Dlatego ważne jest dołączanie do wyspy, a także do podłoża, kontaktów ustalających potencjały tych obszarów (była o tym mowa w poprzednim punkcie).

Z punktu widzenia zastosowań w układach analogowych interesująca – bo może być użyteczna – jest struktura zwana podłożowym tranzystorem *p-n-p*, jaka istnieje pomiędzy obszarami źródła/drenu tranzystora pMOS oraz obszarami wyspy i podłoża. Ten układ obszarów *p-n-p* bywa wykorzystywany jako aktywny tranzystor bipolarny. 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 rysunek 2-9.



Rysunek 2-9. Bipolarny tranzystor podłożowy p-n-p: widok z góry i przekrój. Przekrój w uproszczeniu: nie pokazano obszarów grubego tlenku oraz kontaktów i metalizacji

Użyteczność tej struktury jest ograniczona przez fakt, że kolektor tego tranzystora jest zarazem podłożem układu scalonego, a podłoże – jak już wiemy (część I, punkt 4.2.1) – musi być połączone z „minusem" zasilania. A więc omawianego tranzystora nie można użyć w dowolnym miejscu w schemacie układu, lecz tylko tam, gdzie potrzebny jest tranzystor bipolarny *p-n-p* mający kolektor połączony z „minusem" zasilania. Zobaczymy dalej, że takie sytuacje w układach analogowych zdarzają się.

3 Wybrane układy analogowe

Chociaż przeważająca część produkowanych dziś układów scalonych to układy CMOS, jednak – jak zobaczymy dalej – układy analogowe z tranzystorami bipolarnymi mają szereg zalet i są takie obszary zastosowań, gdzie użyte mogą być tylko tranzystory bipolarne. Dlatego obok układów z tranzystorami MOS będą też pokazywane przykłady układów z tranzystorami bipolarnymi.

3.1 Bloki pomocnicze

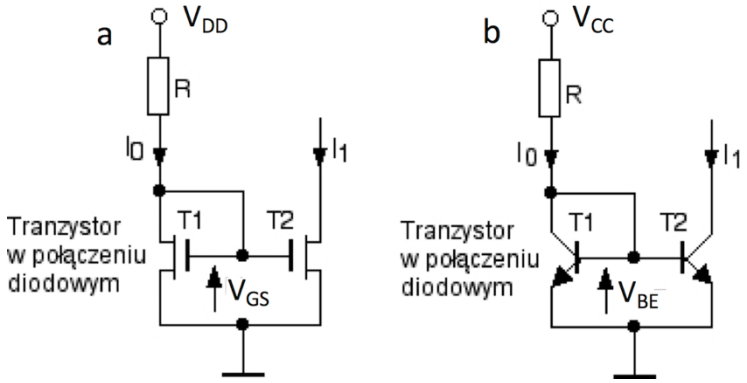
Omówimy tu wspomniane wcześniej układy, które zapewniają stabilną pracę analogowego układu scalonego przy zmianach temperatury – źródła prądowe i napięciowe. Źródła te mogą też pełnić w układach inne funkcje.

3.1.1 Źródła prądowe

DEFINICJA

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 (patrz część I, rysunki 3-16 i 3-18), w których prąd w obwodzie wyjściowym (drenu I_D lub kolektora I_C) bardzo słabo zależy od napięcia na wyjściu elementu (tj. napięcia dren-źródło V_{DS} lub kolektor-emiter V_{CE}). Trzeba więc spolaryzować tranzystor MOS lub bipolarny w taki sposób, by pracował w tym właśnie zakresie charakterystyk. Napięcie w obwodzie wejściowym (V_{GS} lub V_{BE}) określa prąd, jaki płynie w obwodzie wyjściowym (patrz wzory w części I: 3-5 dla tranzystora MOS, 3-12 dla tranzystora bipolarnego). Nie jest jednak obojętne, w jaki sposób polaryzowany jest obwód wejściowy. Gdyby napięcie V_{GS} lub V_{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 V_{GS} lub V_{BE} otrzymuje się jako spadek napięcia na tranzystorze MOS lub bipolarnym w połączeniu diodowym – rysunek 3-1.



Rysunek 3-1. Podstawowe układy źródeł prądowych: (a) MOS, (b) bipolarnego

Zasada działania obu źródeł jest taka sama i opiera się na spostrzeżeniu, że jeśli napięcie V_{GS} lub V_{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 wspomniane wyżej wzory w części I: 3-5 dla tranzystora MOS, 3-12 dla tranzystora bipolarnego). Układ z tranzystorami MOS spełnia ten warunek, jeśli napięcie V_{DS} obu tranzystorów jest większe od napięcia nasycenia V_{DSsat}. Układ z tranzystorami bipolarnymi działa nawet gdy napięcie V_{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

Równanie 3-1

$$I_0 = \frac{V_{DD} - V_{GS}}{R}$$

dla źródła z tranzystorem MOS, i

Równanie 3-2

$$I_0 = \frac{V_{CC} - V_{BE}}{R}$$

dla źródła z tranzystorem bipolarnym.

Wartości V_{GS} lub V_{BE} można wyznaczyć ze wzorów opisujących charakterystyki tranzystorów:

Równanie 3-3

$$V_{GS} = V_T + \sqrt{2I_D \frac{1}{\mu C_{ox}} \frac{L}{W}}$$

dla źródła z tranzystorem MOS, i

Równanie 3-4

$$V_{BE} = \frac{kT}{q} \ln\left(\frac{I_C}{J_{ES0} A_E}\right)$$

dla źródła z tranzystorem bipolarnym.

Po podstawieniu zależności 3-3 do 3-1 otrzymamy równanie kwadratowe ze względu na I_0, a po podstawieniu zależności 3-4 do 3-2 otrzymamy równanie uwikłane. Jednak ścisłe rozwiązania nie są nam tutaj potrzebne. Istotne jest, że w przypadkach obu rodzajów źródeł spełniony jest zwykle warunek: $V_{GS} \ll V_{DD}$ lub $V_{BE} \ll V_{CC}$. Wynika to z faktu, że w zależności 3-4 zwykle drugi składnik jest mały wobec V_T, a V_T jest poniżej 1 V, zaś z zależności 3-4 można obliczyć, że w temperaturze otoczenia i przy typowych wartościach I_C i J_{ES0} $V_{BE} < 1$V. Typowa wartość V_{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 V_{DD} lub V_{CC} uzależniony jest głównie od ilorazu V_{DD}/R lub V_{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 V_T, jak i V_{BE} maleją ze wzrostem temperatury, czyli we wzorach 3-1 i 3-2 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 rysunku 3-1 jest całkowicie wystarczająca w większości zastosowań.

W przytoczonych wyżej rozumowaniach obliczaliśmy prąd I_0 i zakładaliśmy, że prąd I_1 jest mu dokładnie równy. Układy źródeł prądowych, w których prąd w jednej z gałęzi układu powtarza prąd w innej gałęzi, nazywamy zwierciadłami prądowymi. Ale nie każde źródło prądowe jest zarazem zwierciadłem prądowym.

W praktyce prądy I_0 i I_1 w układach obecnie omawianych nie są dokładnie równe. Jest kilka przyczyn różnic:

- napięcia V_{DS} lub V_{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, czyli układ pełni rolę zwierciadła prądowego. Jeżeli identyczność prądów I_0 i I_1 jest istotna, to można do niej dążyć przez:

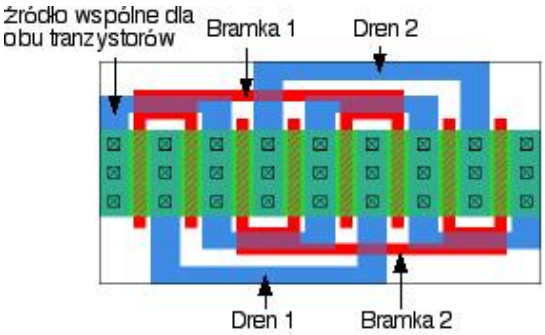
(1) Zastosowanie tranzystorów MOS dużych i z długim kanałem – znacznie dłuższym od minimalnej długości dopuszczalnej w danej technologii. Im dłuższy kanał tranzystora, tym słabszy wpływ napięcia V_{DS} na prąd drenu. Duże wymiary i długi kanał minimalizują także rozrzuty produkcyjne.

(2) Zastosowanie topografii minimalizującej wpływ rozrzutów lokalnych. Reguły są następujące:

- Oba tranzystory 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.

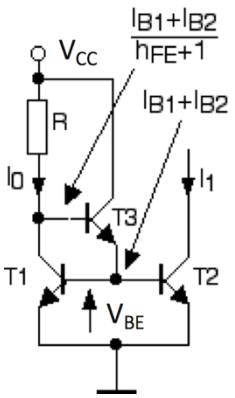
(3) Umieszczenie tranzystorów w sposób symetryczny względem źródeł ciepła w układzie, aby miały możliwie jednakową temperaturę.

Przykład topografii dwóch tranzystorów MOS spełniającej podane wyżej kryteria pokazuje rysunek 3-2. 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 kanałów, w których prąd płynie z lewej do prawej, i kanałów, w których prąd płynie z prawej do lewej.



Rysunek 3-2. Para tranzystorów nMOS - przykład topografii minimalizującej rozrzuty lokalne

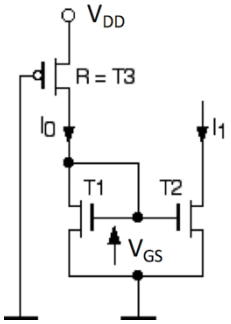
W przypadku tranzystorów bipolarnych wystarcza zachowanie identycznych kształtów i wymiarów tranzystorów. Pojawia się natomiast problem prądów baz. Prąd I_0 nie jest równy prądowi kolektora tranzystora, lecz sumie prądu kolektora i dwóch prądów baz. Wprowadza to dodatkową różnicę między prądami I_0 i 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. Stosuje się wtedy podstawowe źródło prądowe w wersji wzbogaconej o dodatkowy tranzystor, którego rolą jest dostarczenie prądów baz bezpośrednio ze źródła zasilania – rysunek 3-3.
Dodatkowy tranzystor redukuje prąd odgałęziający się od prądu I_0 h_{FE}+1 - krotnie.



Rysunek 3-3. Bipolarne źródło prądowe ze zredukowanym wpływem prądów baz

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

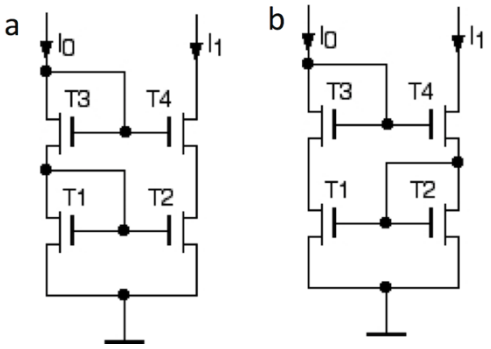
zasilania wynoszących kilka V rezystor R musiałby mieć rezystancję rzędu kilkudziesięciu do kilkuset kΩ. Wykonanie takiego rezystora nie ma ekonomicznego sensu ze względu na powierzchnię, jaką musiałby on zająć. Zamiast rezystora stosuje się zwykle odpowiednio spolaryzowany tranzystor MOS o tak dobranych wymiarach, aby płynął przezeń prąd o wymaganym natężeniu. Przykład pokazuje rysunek 3-4. Powierzchnia takiego tranzystora jest wielokrotnie mniejsza niż rezystora wykonanego jako ścieżka polikrzemowa.



Rysunek 3-4. Źródło prądowe, w którym rolę rezystancji R pełni tranzystor pMOS

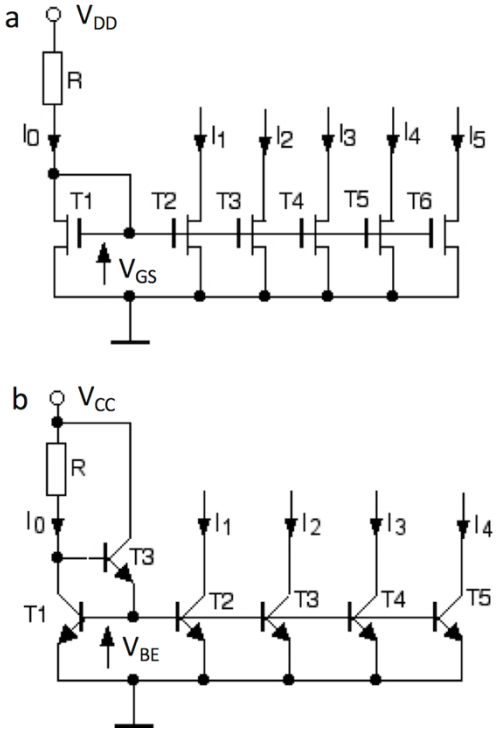
Źródła prądowe są tak powszechnie stosowane w układach analogowych, że warto poznać kilka ich wariantów i odmian mających różne pożyteczne cechy.

W wielu zastosowaniach źródła prądowe powinny wykazywać możliwie jak największą małosygnałową rezystancję wyjściową, tj. zmiany prądu I_1 wywołane przez zmiany napięcia na drenie tranzystora T2 powinny być jak najmniejsze (więcej o parametrach małosygnałowych przeczytasz dalej). Sposobem na powiększenie tej rezystancji jest dodanie w szereg z tranzystorem T2 drugiego tranzystora. Rysunek 3-5 pokazuje dwie wersje źródła o zwiększonej rezystancji wyjściowej: źródło zwane kaskodowym i bardzo podobny układ zwany źródłem Wilsona.



Rysunek 3-5. Źródła prądowe o podwyższonej rezystancji wyjściowej: (a) źródło kaskodowe, (b) źródło Wilsona

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 V_{GS} lub V_{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 rysunek 3-6.



Rysunek 3-6. Zespoły źródeł prądowych: (a) MOS, (b) bipolarnych

W przypadku tranzystorów MOS, zakładając jednakowe długości *L* wszystkich kanałów T1 ... T6, można napisać:

Równanie 3-5

$$\frac{I_0}{W_{T1}} = \frac{I_1}{W_{T2}} = \frac{I_2}{W_{T3}} = \frac{I_3}{W_{T4}} = \frac{I_4}{W_{T5}} = \frac{I_5}{W_{T6}}$$

W przypadku tranzystorów bipolarnych prądy są proporcjonalne do powierzchni złącz emiterowych:

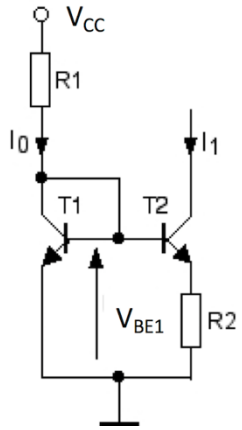
Równanie 3-6

$$\frac{I_0}{A_{ET1}} = \frac{I_1}{A_{ET2}} = \frac{I_2}{A_{ET3}} = \frac{I_3}{A_{ET4}} = \frac{I_4}{A_{ET5}}$$

W przypadku źródeł z tranzystorami bipolarnymi zastosowany jest dodatkowy tranzystor T3 redukujący wpływ sumy wszystkich prądów baz na prąd I_0.

Rysunek 3-7 pokazuje zmodyfikowany schemat bipolarnego źródła prądowego przydatny wtedy, gdy prąd źródła I_1 powinien być bardzo mały. Użycie źródeł podstawowych (jak na rysunku 3-1) wymagałoby w takim przypadku zastosowania bardzo dużej rezystancji R, co jest nieekonomiczne lub nawet technicznie niemożliwe. W źródle pokazanym na rysunku 3-7 dodatkowy rezystor R2 wprowadza lokalne ujemne sprzężenie zwrotne. Prąd I_1 przepływając przez ten rezystor wywołuje spadek napięcia, który odejmuje się od napięcia V_{BE1}. W

rezultacie napięcie emiter-baza tranzystora T2 jest mniejsze o wartość I_1R_2 od V_{BE1}, a prąd I_1 jest 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 3-12 z części I można wyznaczyć różnicę napięć V_{BE} tranzystorów T1 i T2:



Rysunek 3-7. Źródło prądowe dla bardzo małych prądów

Równanie 3-7

$$V_{BE1} - V_{BE2} = \frac{kT}{q}\ln\left(\frac{I_0}{I_1}\right) = I_1R_2$$

skąd wynika zależność

Równanie 3-8

$$I_1 = \frac{kT/q}{R_2}\ln\left(\frac{I_0}{I_1}\right)$$

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 rysunku 3-4 i odpowiednio dobierając wymiary kanału tranzystora T3.

Przy okazji zwróćmy uwagę, że źródło pokazane na rysunku 3-7 umożliwia uzyskanie bardzo dobrej stabilności temperaturowej prądu I_1. Napięcie kT/q w temperaturze otoczenia rośnie o 0,33%/^{0}C. 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).

3.1.2 Źródła napięciowe

DEFINICJA

Źródłem napięciowym nazywamy układ wymuszający zadaną różnicę potencjałów między dwoma węzłami układu.

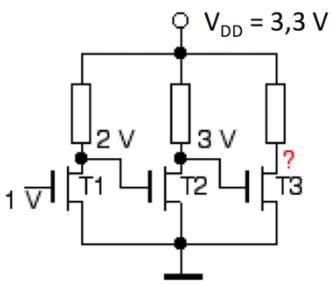
Można wyróżnić trzy typy układów źródeł napięciowych różniące się zastosowaniem i wymaganiami:

- 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.

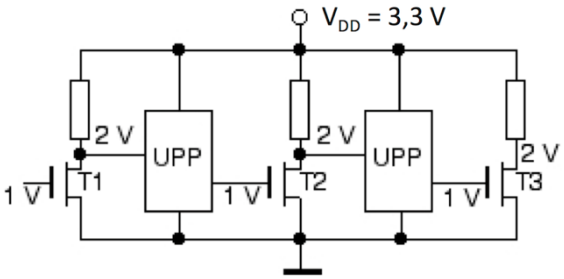
Źródła napięcia odniesienia są to układy, których zadaniem jest wytworzenie napięcia służącego jako wzorcowy poziom napięcia, na przykład do porównywania z jakąś inną wartością napięcia wytwarzaną w układzie lub podawaną z zewnątrz na wejście układu. Źródła napięcia odniesienia są na ogół obciążane bardzo 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. W niektórych zastosowaniach stosowane są również 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. Rysunek 3-8 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!



Rysunek 3-8. Ilustracja problemu przesuwania poziomu składowej stałej

Wprowadzenie układów przesuwania poziomu składowej stałej rozwiązuje ten problem – patrz rysunek 3-9.



Rysunek 3-9. Układy przesuwania poziomu składowej stałej (UPP) wymuszające różnicę napięć równą 1 V rozwiązują problem układu z rys. 3-8

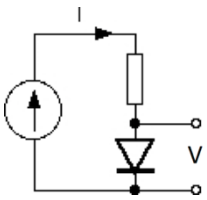
Główne wymaganie dla układów przesuwania poziomu składowej stałej to „przezroczystość" dla składowej zmiennej. Układy te powinny zapewniać transmisję składowej zmiennej z wejścia na wyjście bez tłumienia, zakłóceń i zniekształceń. Wymagania odnoszące się do wymuszanej różnicy napięć mogą być różne, w 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 wielogodzinnego wykładu. Omówimy natomiast kilka przykładowych rozwiązań układów źródeł napięć odniesienia i układów przesuwania poziomu składowej stałej. Zaczniemy od omówienia elementów, które będą nazywane pierwotnymi źródłami napięcia odniesienia.

DEFINICJA

Pierwotnym źródłem napięcia odniesienia będziemy nazywać dwójnik nieliniowy, który cechuje prawie stały, bardzo 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 – przykład na rysunku 3-10.



Rysunek 3-10. Typowy sposób wykorzystania pierwotnego źródła napięcia odniesienia. Napięciem jest spadek na elemencie nieliniowym (na rysunku jest to dioda)

Jako pierwotne źródła napięcia odniesienia bywają stosowane:

- dioda spolaryzowana w kierunku przewodzenia,
- dioda spolaryzowana w zakresie przebicia (zwana potocznie diodą Zenera),
- tranzystor MOS w połączeniu diodowym,
- bipolarny mnożnik V_{BE}.

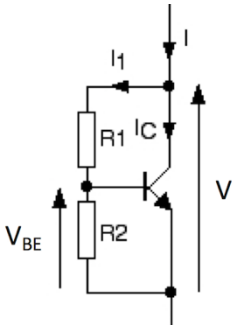
Dioda spolaryzowana w kierunku przewodzenia – jest to często spotykane pierwotne źródło napięcia odniesienia. Używany jest zwykle tranzystor bipolarny w połączeniu diodowym. Jedno z zastosowań tego źródła już znamy – w źródłach prądowych. Spadek napięcia na diodzie spolaryzowanej w kierunku przewodzenia, dany

zależnością 3-14 z części I, 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/^{0}C, dioda nie jest więc źródłem stabilnym temperaturowo.

W przypadku diody spolaryzowanej w zakresie przebicia 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/^{0}C. Zatem i to źródło nie jest źródłem stabilnym temperaturowo.

Tranzystor MOS w połączeniu diodowym jest to źródło używane w układach MOS, znane nam już z układów źródeł prądowych. Spadek napięcia jest określony przez zależność 3-3 , 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 V_T.

Bipolarny mnożnik V_{BE} jest to dwójnik będący połączeniem tranzystora bipolarnego i dwóch rezystorów – rysunek 3-11.



Rysunek 3-11. Układ zwany bipolarnym mnożnikiem V_{BE}

Dla zanalizowania działania tego układu pominiemy prąd bazy tranzystora. Mamy wówczas proste zależności:

Równanie 3-9

$$I_1 = \frac{V}{R_1+R_2}$$

oraz

Równanie 3-10

$$V_{BE} = R_2 I_1$$

Skąd otrzymujemy

Równanie 3-11

$$V = V_{BE}\frac{R_1+R_2}{R_2} = kV_{BE}$$

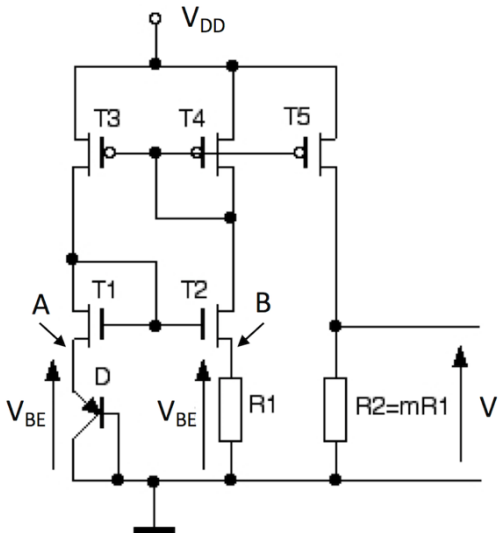
zatem napięcie V jest proporcjonalne do V_{BE}. Stąd nazwa układu. Warunkiem działania układu jako pierwotnego źródła napięcia odniesienia jest wymuszenie dostatecznie dużego prądu *I*, tak aby napięcie V_{BE} było dostatecznie duże (około 0.7 V w temperaturze otoczenia) i spełniony był warunek: $I_C \gg I_1$. Jeśli warunek ten nie jest spełniony, charakterystyka układu jest liniowa: $I = V/(R_1 + R_2)$, ponieważ prąd kolektora

tranzystora jest pomijalnie mały. Wzór 3-11 jest wprawdzie nadal słuszny, ale napięcie V_{BE} jest tak małe, że tranzystor praktycznie nie przewodzi. Układ nie spełnia wówczas swej roli.

Zaletą mnożnika V_{BE} jest możliwość uzyskania napięcia o dowolnej (ale tylko większej od V_{BE}) wartości przez odpowiedni dobór obu rezystancji. Tak, jak i poprzednie źródła, mnożnik V_{BE} nie jest układem stabilnym temperaturowo. Jeżeli $V = kV_{BE}$ (*k* wyznaczone przez stosunek rezystancji we wzorze 3-11), to napięcie V maleje z temperaturą o *k**2 mV/^{0}C.

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. Omówimy kilka takich układów.

Rys. 12.16 przedstawia układ mnożnika V_{BE} wykonanego w technologii CMOS, z wykorzystaniem bipolarnego tranzystora podłożowego *p-n-p* (patrz punkt 2.2.7). Nie jest to jednak dwójnik, jak mnożnik bipolarny, lecz układ bardziej złożony.



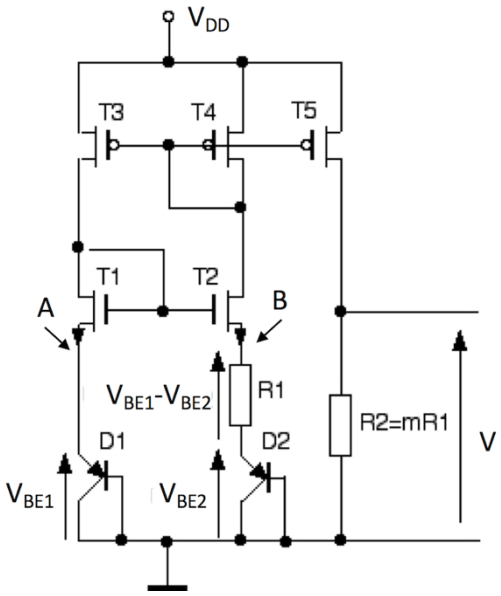
Rysunek 3-12. Układ mnożnika VBE w technologii CMOS

Zakładamy, że tranzystory T1 i T2 są identyczne, oraz identyczne są także tranzystory T3, T4 i T5. Górne zwierciadło prądowe (tranzystory pMOS T3 i T4) wymusza przepływ dwóch prądów o identycznej wartości w gałęziach, w których znajdują się: tranzystor nMOS T1 i tranzystor bipolarny *p-n-p* w połączeniu diodowym oznaczony D (lewa gałąź) oraz tranzystor nMOS T2 i rezystor R1 (prawa gałąź). A ponieważ jednakowe są prądy płynące przez tranzystory T1 i T2, to jednakowe muszą być ich napięcia bramka-źródło, skąd wnioskujemy, że jednakowe są napięcia w węzłach A i B. Zatem jednakowe, równe V_{BE}, są także napięcia na tranzystorze bipolarnym i na rezystorze R1. Tranzystory T4 i T5 są identyczne, zatem prądy płynące przez rezystory R1 i R2 są także identyczne. Ponieważ rezystor R2 ma rezystancję równą mR_1, spadek napięcia na nim jest równy mV_{BE}. Innymi słowy, napięcie wyjściowe układu jest równe

Równanie 3-12

$$V = mV_{BE} = \frac{R_2}{R_1}V_{BE}$$

Układ mnożnika V_{BE} według rysunku 3-12 daje napięcie *V* malejące z temperaturą, podobnie jak mnożnik bipolarny. Po niewielkiej rozbudowie układ może także dawać napięcie nie malejące, lecz rosnące z temperaturą. Układ taki jest pokazany na rysunku 3-13. Zastosowano w nim dwa podłożowe tranzystory bipolarne D1 i D2, różniące się powierzchnią złącza emiterowego: $A_{E2} = nA_{E1}$, gdzie $n > 1$.



Rysunek 3-13. Układ mnożnika kT/q w technologii CMOS

Podobnie jak w poprzednim układzie prądy w gałęziach z tranzystorami T3 i T4 są równe, jak również równe są napięcia w węzłach a i B. Wynika z tego, że na rezystorze R1 odkłada się teraz różnica napięć V_{BE} na tranzystorach bipolarnych D1 i D2. Ponieważ powierzchnia złącza emiterowego D2 jest większa, niż D1, napięcie V_{BE2} jest mniejsze, niż V_{BE1}(wzór 3-14 w części I). Różnica napięć $V_{BE1} - V_{BE2}$ odkładająca się na rezystorze R1 wynosi

Równanie 3-13

$$V_{BE1} - V_{BE2} = \frac{kT}{q} \ln\left(\frac{A_{E2}}{A_{E1}}\right)$$

To napięcie ulega mnożeniu na rezystancji R2 podobnie, jak w poprzednim układzie, a więc napięcie wyjściowe układu wynosi

Równanie 3-14

$$V = m\frac{kT}{q} \ln\left(\frac{A_{E2}}{A_{E1}}\right)$$

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 rysunku 3-12, otrzymać układ źródła napięcia 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.

Zauważmy przy okazji, że w układzie z rysunku 3-13 napięcie wyjściowe zależy tylko od stosunku dwóch rezystancji oraz od stosunku powierzchni złącz emiterowych dwóch tranzystorów bipolarnych. Stosunki te nie zależą od temperatury i wykazują małą wrażliwość na rozrzuty produkcyjne.

Do kategorii źródeł napięciowych zaliczane są też układy przesuwania poziomu napięcia stałego. Omówione będą przy okazji omawiania układów wzmacniaczy, bo tam są potrzebne. Wspomnimy tam też o dzielnikach napięcia, które bywają stosowane wtedy, gdy w układzie potrzebne jest wiele różnych napięć.

3.2 Proste stopnie wzmacniające

Będziemy teraz zajmować się układami wzmacniaczy, tj. układów, które służą do zwiększania amplitudy sygnałów zmiennych. Wzmacniacze służą często do wzmacniania sygnałów o amplitudach bardzo małych. Przykładowo, amplituda sygnału na wejściu antenowym radioodbiornika może być rzędu mikrowoltów, satelitarny sygnał odbierany przez odbiornik GPS jest jeszcze znacznie słabszy, sygnały pochodzenia biologicznego, takie jak np. sygnał EKG, mają amplitudy od mikrowoltów do pojedynczych miliwoltów. Będziemy więc mieli w układach do czynienia równocześnie z prądami i napięciami stałymi oraz z prądami i napięciami zmiennymi o małej amplitudzie. Przykładowo, przebieg napięcia w funkcji czasu może wyglądać tak: $V(t) = V_0 + v_m \sin(\omega t)$, gdzie V_0 nazywać będziemy składową stałą napięcia, a v_m – amplitudą składowej zmiennej. Podobnie dla prądu: $I(t) = I_0 + i_m \sin(\omega t)$, gdzie I_0 nazwiemy składową stałą prądu, a i_m – amplitudą składowej zmiennej. Składowe stałe napięć i prądów będziemy oznaczali dużymi literami: V, I, zaś amplitudy składowych zmiennych małymi literami: v, i. W przypadku elementów czynnych i nieliniowych (tranzystory, diody) przyjęło się określać wartości składowych stałych napięć polaryzujących i prądów mianem punktu pracy danego elementu.

W rozważaniu układów, w których występują sygnały o małych amplitudach, wygodnie jest posługiwać się parametrami elementów, które noszą nazwę parametrów małosygnałowych. Zajmiemy się nimi w następnym punkcie.

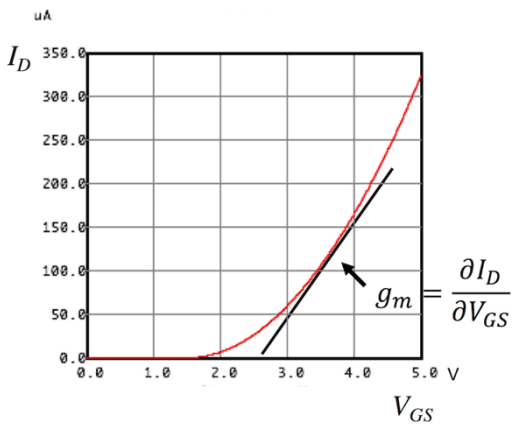
Jeśli pojęcie parametrów małosygnałowych, a także definicje tych parametrów dla tranzystorów MOS i bipolarnych, są Ci znane, możesz punkt 3.2.1 pominąć.

3.2.1 Parametry małosygnałowe tranzystorów i ich schematy zastępcze

Charakterystyki prądowo-napięciowe tranzystorów są, jak wiemy (część I, punkty 3.1.3 i 3.1.4), funkcjami nieliniowymi, ale przy rozważaniu sygnałów o bardzo małej amplitudzie można lokalnie te charakterystyki przybliżyć funkcjami liniowymi – pochodnymi funkcji opisujących te charakterystyki. Dla tranzystora MOS definiujemy dwie takie pochodne.

DEFINICJA

Transkonduktancją g_m tranzystora MOS nazywamy pochodną zależności prądu drenu od napięcia bramka-źródło: $g_m = \frac{\partial I_D}{\partial V_{GS}}$.

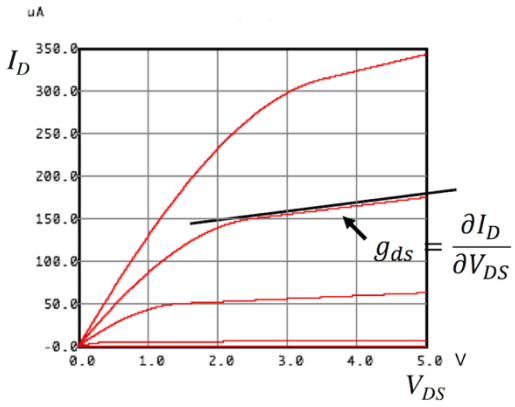


Rysunek 3-14. Ilustracja definicji transkonduktancji

Definicję tę ilustruje rysunek 3-14, na którym pokazana jest zależność prądu drenu od napięcia bramka-źródło. Im większą wartość ma transkonduktancja, tym silniej zależą zmiany prądu drenu od zmian napięcia bramka-źródło.

DEFINICJA

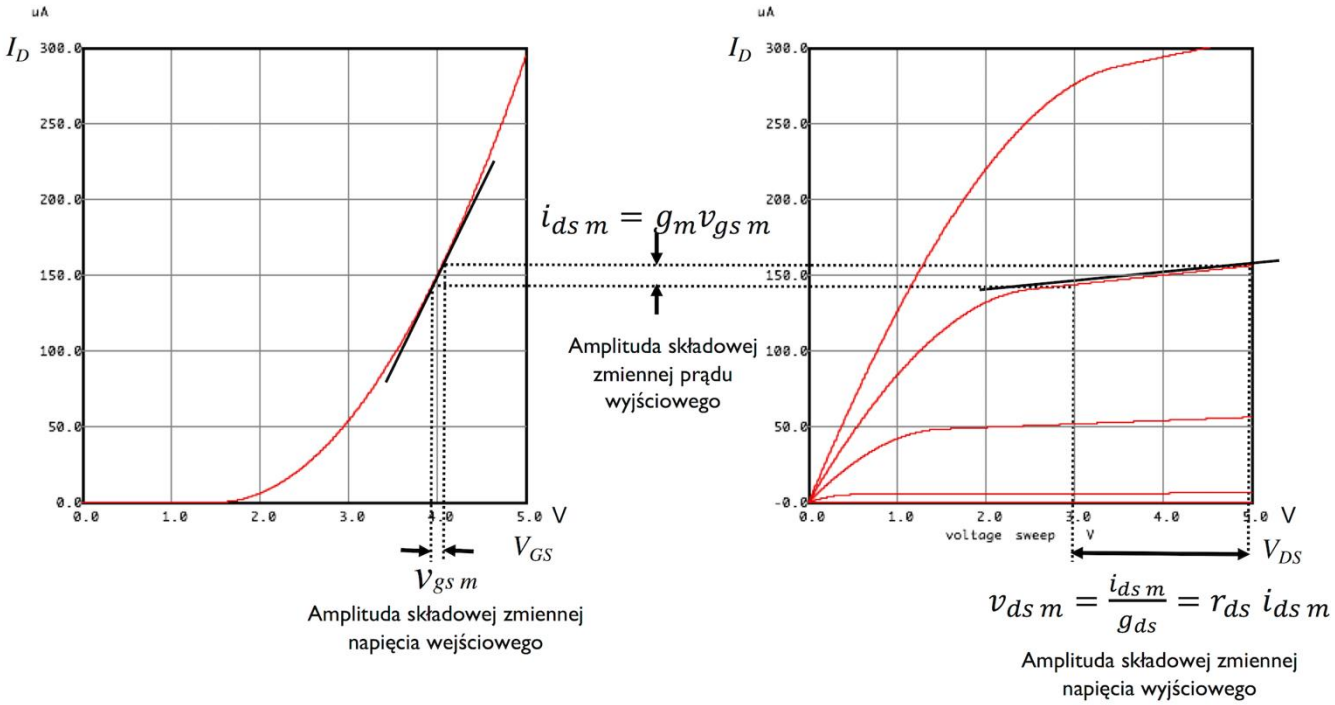
Konduktancją wyjściową g_{ds} tranzystora MOS nazywamy pochodną zależności prądu drenu od napięcia dren-źródło: $g_{ds} = \frac{\partial I_D}{\partial V_{DS}}$.



Rysunek 3-15. Ilustracja definicji konduktancji wyjściowej

Definicję tę ilustruje rysunek 3-15, na którym pokazana jest zależność prądu drenu od napięcia dren-źródło. Im większą wartość ma konduktancja wyjściowa, tym silniej zależą zmiany prądu drenu od zmian napięcia dren-źródło. Często używane jest też pojęcie rezystancji wyjściowej, która jest odwrotnością konduktancji: $r_{ds} = 1/g_{ds}$.

Do czego służy transkonduktancja i konduktancja wyjściowa? Pokazuje to rysunek 3-16.



Rysunek 3-16. Określanie amplitud napięć i prądów zmiennych przy użyciu parametrów małosygnałowych

Przyjmijmy, że między bramkę i źródło tranzystora MOS przyłożone jest napięcie zmienne v_{gs}, którego (bardzo mała) amplituda wynosi $v_{gs\,m}$. Wykorzystując pojęcie transkonduktancji możemy obliczyć wartość amplitudy prądu zmiennego w obwodzie drenu jako

Równanie 3-15

$$i_{ds\,m} = \frac{\partial I_D}{\partial V_{GS}} v_{gs\,m} = g_m v_{gs\,m}.$$

Ilustruje to lewa część rysunku 3-16.

Jeśli prąd zmienny o amplitudzie $i_{ds\,m}$ występuje w obwodzie drenu tranzystora, towarzyszyć temu musi amplituda napięcia zmiennego wynosząca

Równanie 3-16

$$v_{ds\,m} = \frac{1}{\frac{\partial I_D}{\partial V_{DS}}} i_{ds\,m} = \frac{i_{ds\,m}}{g_{ds}} = i_{ds\,m} r_{ds}.$$

Ilustruje to prawa strona rysunku 3-16.

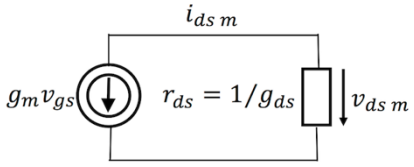
Łącząc zależności 3-15 i 3-16 otrzymujemy

Równanie 3-17

$$v_{ds\,m} = \frac{g_m v_{gs\,m}}{g_{ds}}$$

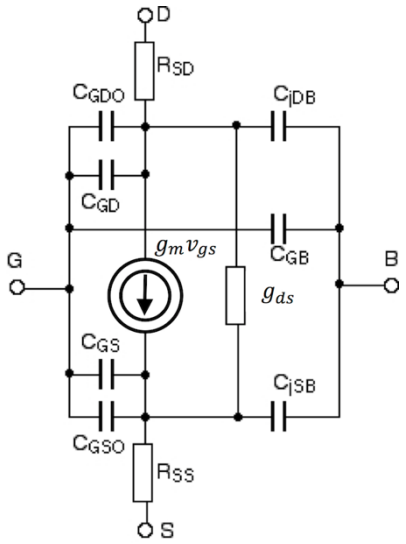
Powyższe rozumowanie pozwala określić wzmocnienie napięciowe $k_u = v_{ds\,m}/v_{gs\,m}$. W omawianym przypadku jest ono równe ilorazowi g_m/g_{ds}. Ten iloraz bywa nazywany wzmocnieniem wewnętrznym tranzystora (ang. „intrinsic gain").

Zauważmy, że zależności 3-15 do 3-17 można interpretować jako wynikające ze schematu pokazanego na rysunku 3-17, w którym występuje idealne źródło prądowe wymuszające prąd $i_{ds\,m}$ zgodnie z zależnością 3-15, a prąd ten przepływając przez rezystancję r_{ds} wymusza na niej spadek napięcia równy $v_{ds\,m}$ zgodnie z zależnością 3-16. Jest to najprostszy schemat zastępczy tranzystora dla sygnałów zmiennych o małej amplitudzie.



Rysunek 3-17. Schemat układu odpowiadającego wzorom 3-15 do 3-17

Pełny małosygnałowy schemat zastępczy tranzystora MOS zawiera także inne elementy: pojemności i rezystancje obszarów źródła i drenu. Ma on cztery węzły zewnętrzne: bramki (G), źródła (S), drenu (D) i podłoża (B) – rysunek 3-18.



Rysunek 3-18. Pełny schemat zastępczy tranzystora MOS

Małosygnałowe schematy zastępcze tranzystorów i innych elementów zawierają wyłącznie liniowe rezystancje i pojemności (w zakresie bardzo wielkich częstotliwości mogą zawierać także indukcyjności). Przydatność schematów zastępczych polega na tym, że jeśli w schemacie układu elektronicznego zastąpimy tranzystory i inne elementy ich schematami zastępczymi, to powstały w ten sposób schemat zastępczy całego układu będzie zawierał wyłącznie elementy liniowe, i będzie można analizować jego działanie oraz wyprowadzać potrzebne wzory metodami teorii obwodów liniowych.

Wartości parametrów małosygnałowych zależą od punktu pracy tranzystora (czyli wartości napięć polaryzujących i prądów), co jest oczywiste, gdy popatrzymy na rysunki 3-14 i 3-15. Projektant układu ma więc wpływ na wartości parametrów małosygnałowych określając wartości składowych stałych napięć polaryzujących i prądów.

Posługując się modelem matematycznym tranzystora MOS (część I, punkt 3.1.5, wzory 3-3 do 3-8) można wyprowadzić wzory określające zależności parametrów małosygnałowych od punktu pracy tranzystora. Dla zakresu liniowego otrzymujemy

Równanie 3-18

$$g_m = KV_{DS}$$

Równanie 3-19

$$g_{ds} = K(V_{GS} - V_T - V_{DS})$$

a dla zakresu nasycenia

Równanie 3-20

$$g_m = K(V_{GS} - V_T) = \sqrt{2KI_D} = \frac{2I_D}{V_{GS} - V_T}$$

(wszystkie trzy postacie wzoru 3-20 są równoważne)

Równanie 3-21

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

Przydatna bywa też wartość transkonduktancji w zakresie podprogowym

Równanie 3-22

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

W powyższych wzorach oznaczono dla skrócenia zapisu

Równanie 3-23

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

Parametrów małosygnałowych i schematu zastępczego tranzystora bipolarnego nie będziemy szczegółowo rozważać, warto jedynie dla porównania z tranzystorem MOS określić wartość transkonduktancji:

Równanie 3-24

$$g_m = \frac{\partial I_C}{\partial V_{BE}} = \frac{q}{kT} I_C$$

3.2.2 Porównanie tranzystorów MOS i bipolarnych jako elementów wzmacniających

Możemy teraz porównać tranzystory MOS i bipolarne z punktu widzenia zastosowań w układach wzmacniaczy. Porównamy wartość transkonduktancji - parametru, od którego w stopniach wzmacniających zależy wartość wzmocnienia napięciowego. Stosunek tych transkonduktancji wynosi

Równanie 3-25

$$\frac{g_{m(BIP)}}{g_{m(MOS)}} = \frac{\frac{qI_C}{kT}}{\frac{2I_D}{V_{GS}-V_T}} = \frac{I_C}{I_D}\frac{V_{GS}-V_T}{2\frac{kT}{q}}$$

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 $V_{GS} - V_T$ w typowych warunkach pracy tranzystora MOS jest wielokrotnie większa. Dużo wyższa transkonduktancja czyni tranzystor bipolarny elementem korzystniejszym w układach analogowych, zwłaszcza w układach wzmacniaczy. Są też dalsze cechy tego elementu dające mu przewagę. W tablicy poniżej pokazane są główne różnice między tranzystorami MOS i bipolarnymi.

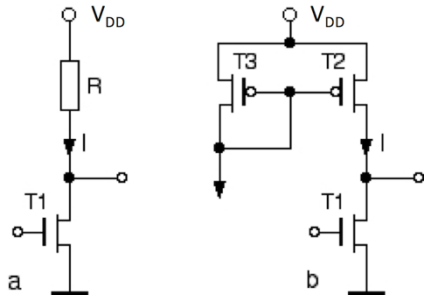
Tabela 2

Właściwość	Tranzystor MOS	Tranzystor bipolarny
Transkonduktancja		Większa niż w tranzystorze MOS
Max. częstotliwość pracy	Kilkanaście GHz	Kilkaset GHz
Rezystancja wejściowa	Praktycznie nieskończenie wielka (bramka izolowana)	Mała (niepomijalny prąd bazy)
Poziom szumów		Mniejszy niż w tranzystorze MOS
Rozrzuty produkcyjne		Mniejsze niż w tranzystorze MOS

Tranzystory bipolarne mają w układach analogowych wyraźną przewagę. Przez wiele lat układy analogowe były wykonywane wyłącznie przy użyciu tranzystorów bipolarnych. Dziś popularność technologii CMOS oraz względy ekonomiczne spowodowały, że układy analogowe CMOS są z powodzeniem stosowane, zwłaszcza jako bloki analogowe w układach typu „system on chip”.

3.2.3 Układy prostych stopni wzmacniających

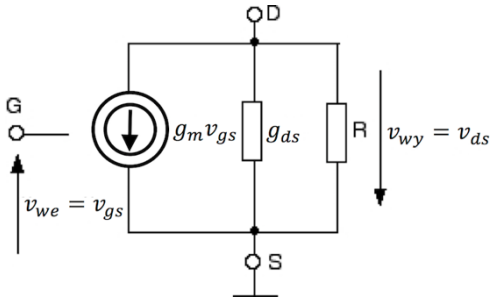
Najprostszy wzmacniacz można zbudować według rysunku 3-19.



Rysunek 3-19. Zasada budowy najprostszego stopnia wzmacniającego: (a) z obciążeniem w postaci rezystora R, (b) z obciążeniem aktywnym

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 $|k_u| = |v_{wy}/v_{we}|$, gdzie v_{we} jest amplitudą składowej zmiennej napięcia na wejściu, a v_{wy} jest amplitudą składowej zmiennej napięcia na wyjściu. 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 rysunku 3-19a 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 (rysunek 3-18). 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 rysunku 3-20.



Rysunek 3-20. Małosygnałowy schemat zastępczy układu z rysunku 3-19a

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 3-20 widać, że napięcie wyjściowe v_{wy} otrzymamy mnożąc prąd źródła prądowego równy $g_m v_{we}$ przez rezystancję złożoną z równolegle połączonych: rezystancji R i konduktancji wyjściowej tranzystora g_{ds} . Ponieważ równoległe konduktancje się sumują, otrzymujemy

Równanie 3-26

$$|k_u| = g_m \frac{1}{g_{ds}+\frac{1}{R}}$$

Konduktancja g_{ds} jest z reguły znacznie mniejsza od konduktancji 1/R, wobec czego można ją pominąć i zależność 3-26 uprościć do

Równanie 3-27

$$|k_u| = g_m R$$

Rezystancja R nie może być dowolna, bowiem jej wartość wraz ze składową stałą I_D prądu drenu oraz napięciem zasilania V_{DD} decydują o punkcie pracy tranzystora, tj. składowej stałej napięcia dren-źródło V_{DS}:

Równanie 3-28

$$V_{DS} = V_{DD} - I_D R$$

Załóżmy, że tranzystor pracuje w zakresie nasycenia, a napięcie V_{DS} jest równe połowie napięcia zasilania: $V_{DS} = V_{DD}/2$ (przy takiej lub zbliżonej wartości uzyskuje się największą możliwą amplitudę sygnału zmiennego na wyjściu). Wówczas, wykorzystując zależność 3-20, można otrzymać prosty wynik:

Równanie 3-29

$$|k_u| = \frac{V_{DD}}{V_{DS}-V_T}$$

Widać, że otrzymane wzmocnienie jest niewiele większe od jedności, bo i licznik, i mianownik we wzorze 3-29 mają wartość rzędu kilku woltów. Nieco lepszy wynik można by otrzymać, gdyby tranzystor pracował w zakresie podprogowym. Posługując się wówczas wzorem 3-22 otrzymamy

Równanie 3-30

$$|k_u| = \frac{V_{DD}}{2n\frac{kT}{q}}$$

Mianownik w zależności 3-30 ma typową wartość około 100 mV, więc można uzyskać wzmocnienie napięciowe rzędu kilkudziesięciu. Jest to ilustracja reguły ogólniejszej: największe wzmocnienie napięciowe wzmacniacza z tranzystorem MOS można zazwyczaj uzyskać, gdy pracuje on w zakresie podprogowym. Dlatego jest to zakres pracy interesujący dla projektantów układów analogowych.

Wzmocnienie, jakie można uzyskać, gdy tranzystor pracuje w zakresie podprogowym (wzór 3-30), jest nadal niewielkie. 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 układzie z rysunku 3-19b jest ona równa $1/g_{dsT2}$. Zamieniając w schemacie zastępczym z rysunku 3-19b rezystor R na rezystancję wyjściową $1/g_{dsT2}$ otrzymujemy

Równanie 3-31

$$|k_u| = \frac{g_{mT1}}{g_{dsT1}+g_{dsT2}}$$

Wykorzystując tu wzory 3-20 oraz 3-21 można wzór 3-31 przekształcić do postaci

Równanie 3-32

$$|k_u| = \frac{\sqrt{2\mu_{T1}C_{ox}\frac{W_{T1}}{L_{T1}}}}{(\lambda_{T1}+\lambda_{T2})\sqrt{I_D}}$$

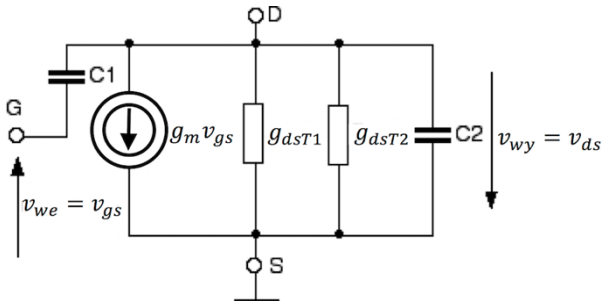
Typowa wartość parametru λ wynosi 0,05 V^{-1} ... 0,1 V^{-1}. Można się przekonać, że bez trudu uzyskuje się przy typowych wartościach pozostałych parametrów wzmocnienia rzędu 100 ... 200. Wzmocnienie rośnie, gdy maleje prąd I_D. Maksymalne osiągalne wzmocnienie otrzymamy przy pracy tranzystora w zakresie podprogowym. Wynosi ono

Równanie 3-33

$$|k_u| = \frac{1}{n(\lambda_{T1}+\lambda_{T2})\frac{kT}{q}}$$

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.



Rysunek 3-21. 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 (czyli pojemność między bramką i źródłem) 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 rysunku 3-21 prowadzi do następującego wzoru na częstotliwość f_T:

Równanie 3-34

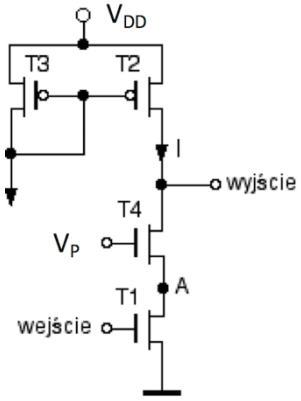
$$f_T = \frac{\sqrt{2I_D\mu_{T1}C_{ox}\frac{W_{T!}}{L_{T!}}}}{2\pi(C_1+C_2)}$$

(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 $V_{DSsatT1}$ do $V_{DD} - V_{DSsatT2}$. Napięcia nasycenia mają typową wartość rzędu

kilkudziesięciu do kilkuset mV, a więc napięcie wyjściowe może się zmieniać prawie od 0 do V_{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.

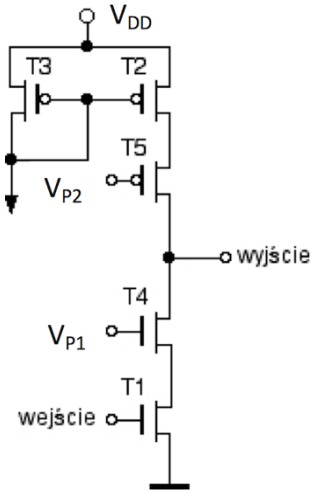
Układ wzmacniacza z aktywnym obciążeniem ma pewną słabą stronę: pojemność włączona między wyjście, a wejście ulega pozornemu zwielokrotnieniu (zjawisko zwane efektem Millera). Od strony wejścia pojemność C1 jest widoczna jako $C1' = C1(|k_u| + 1)$. Ogranicza to szerokość pasma we wzmacniaczach kilkustopniowych. Dla $|k_u|$ rzędu 100 ... 200 nawet bardzo mała pojemność C1 oznacza znaczne obciążenie pojemnościowe poprzedniego stopnia. Efekt ten można wyeliminować we wzmacniaczu w układzie pokazanym na rysunku 3-22. Jest on zwany układem kaskodowym.



Rysunek 3-22. Wzmacniacz w układzie kaskody

Dodatkowy tranzystor T4 jest polaryzowany napięciem stałym V_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 – rysunek 3-23.



Rysunek 3-23. 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 (rysunek 3-19b). Wadą takiego układu jest konieczność dostarczenia dodatkowych napięć polaryzujących (może tu znaleźć zastosowanie jedno z omawianych wcześniej źródeł). Cechuje go też mniejsza amplituda sygnału na wyjściu, bo musi być zapewniona praca w nasyceniu dwóch tranzystorów nMOS połączonych szeregowo i dwóch tranzystorów pMOS połączonych szeregowo. Jest to szczególnie poważny problem w przypadku układów CMOS

wytwarzanych w najbardziej zaawansowanych technologiach, w których maksymalne napięcie zasilania jest rzędu 1…1,5 V.

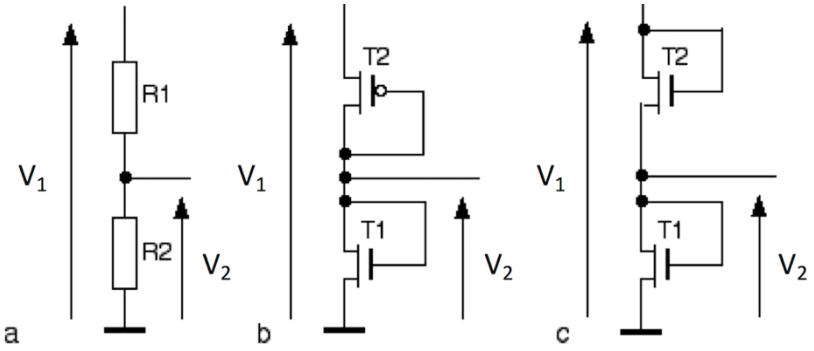
UWAGA

We wszystkich wzorach na wzmocnienie napięciowe występowała wartość bezwzględna tego napięcia. Nie interesowała nas bowiem faza tego napięcia. Wszystkie omawiane jednostopniowe układy wzmacniające odwracają fazę sygnału zmiennego.

3.2.4 Układy polaryzacji we wzmacniaczach

W układach wzmacniaczy omawianych w poprzednim punkcie pominięte były elementy potrzebne do doprowadzenia do bramek tranzystorów stałych napięć polaryzujących określających ich punkty pracy, jak na przykład napięcia V_{P1} i V_{P2} w układzie z rysunku 3-23. Do tego mogą służyć omawiane w punkcie 3.1.2 źródła napięciowe, ale stosowane bywają po prostu dzielniki napięcia zasilającego, a gdy dwa stopnie wzmacniacza są ze sobą bezpośrednio sprzężone, właściwą polaryzację drugiego stopnia najczęściej zapewnia układ przesuwania poziomu składowej stałej (wspomniany w punkcie 3.1.2).

Dzielniki napięcia pokazuje rysunek 3-24.



Rysunek 3-24. 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 rysunku 3-24b i 3-24c, zakładając prąd wyjściowy równy zeru i przyrównując prądy drenów tranzystorów otrzymuje się

Równanie 3-35

$$V_2 = \frac{m_2}{m_1+m_2} V_1 + \frac{m_1 V_{T1} - m_2 V_{T2}}{m_1+m_2}$$

gdzie

Równanie 3-36

$$m_1 = \sqrt{\mu_1 \frac{W_1}{L_1}}$$

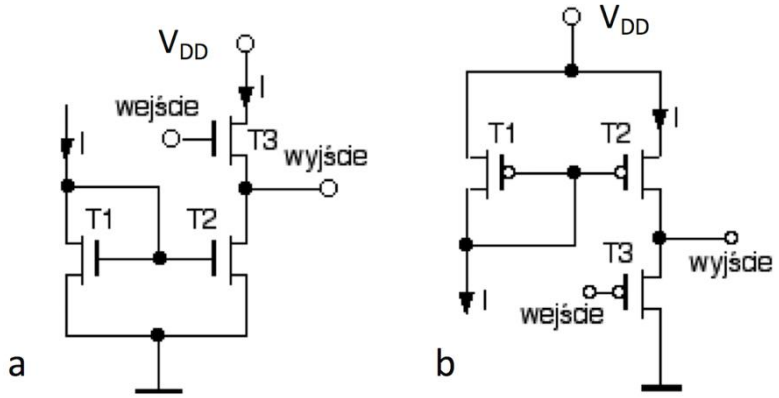
Równanie 3-37

$$m_2 = \sqrt{\mu_2 \frac{W_2}{L_2}}$$

W przypadku tranzystora pMOS należy we wzorze 3-35 użyć wartości bezwzględnej napięcia progowego.

Teoretycznie można także zbudować dzielnik napięcia z kondensatorów, jednak jest to rzadko stosowane, bo jakakolwiek upływność kondensatora zaburza podział napięcia.

Przykład układów przesuwania poziomu składowej stałej pokazuje rysunek 3-25.



Rysunek 3-25. Układy przesuwania poziomu składowej stałej: (a) z tranzystorami nMOS, (b) z tranzystorami pMOS

Najprostszym układem przesuwania poziomu składowej stałej jest układ znany jako wtórnik. Wtórnik z tranzystorami MOS ma w najprostszym przypadku schemat jak na rysunku 3-25. 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. Układy te przenoszą składową zmienną z wejścia na wyjście praktycznie bez zmiany amplitudy. Różnica między napięciem na wyjściu, a napięciem na wejściu jest dana (co do wartości bezwzględnej) wzorem

Równanie 3-38

$$\Delta V = V_T + \sqrt{\frac{2I}{\mu C_{ox}\frac{W_{T3}}{L_{T3}}}}$$

Różnicę napięć ΔV można w pewnym zakresie regulować dobierając prąd *I* oraz wymiary kanału. Podobne układy buduje się też na tranzystorach bipolarnych.

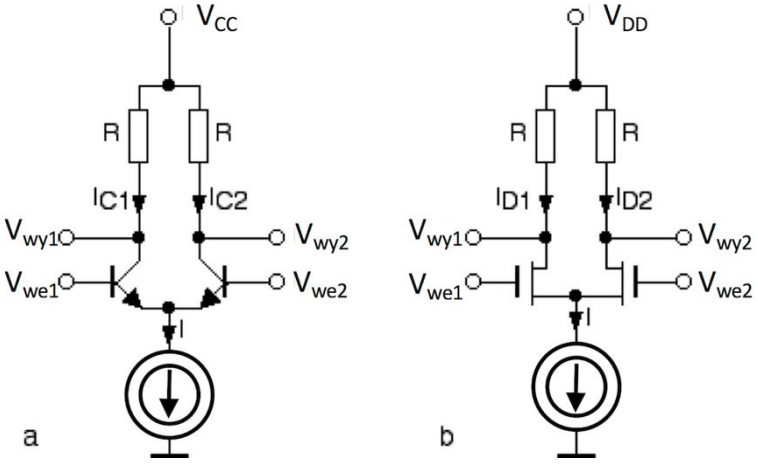
3.3 Wzmacniacz różnicowy i przykłady jego zastosowań

Wzmacniacz różnicowy zajmuje wśród układów wzmacniających szczególne miejsce. 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 innych zastosowań.

3.3.1 Podstawowy układ wzmacniacza różnicowego

Podstawowy układ pokazuje rysunek 3-26. 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 ($V_{CB} < 0, V_{BE} > 0$), a tranzystory MOS w zakresie nasycenia. Założymy, że układ pokazany na rysunku 3-26 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 $V_{we1} = V_{we2}$ prąd ten dzieli się po połowie między obie gałęzie układu. Spadek napięcia na obu rezystorach jest jednakowy, a różnicowe napięcie wyjściowe $V_{wy} =$

$V_{wy2} - V_{wy1}$ jest równe zeru. Jeśli istnieje różnica napięć $V_{we} = V_{we2} - V_{we1}$ (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.



Rysunek 3-26. Podstawowy wzmacniacz różnicowy: (a) bipolarny, (b) MOS

Zależność różnicowego napięcia wyjściowego od różnicowego napięcia wejściowego łatwo otrzymać dla wzmacniacza z tranzystorami bipolarnymi. Prąd kolektora dany jest wzorem 3-12 z części I, zatem dla różnicy prądów $I_{C2} - I_{C1}$ mamy

Równanie 3-39

$$I_{C2} - I_{C1} = I_{ES0}\left(e^{\frac{qV_{BE2}}{kT}} - e^{\frac{qV_{BE1}}{kT}}\right)$$

a równocześnie

Równanie 3-40

$$I = I_{C2} + I_{C1} = I_{ES0}\left(e^{\frac{qV_{BE2}}{kT}} + e^{\frac{qV_{BE1}}{kT}}\right)$$

Różnicowe napięcie wyjściowe wynosi

Równanie 3-41

$$V_{wy} = (V_{DD} - V_{R2}) - (V_{DD} - V_{R1}) = V_{R1} - V_{R2} = -R(I_{C2} - I_{C1})$$

Z tych trzech zależności można po prostych przekształceniach otrzymać

Równanie 3-42

$$V_{wy} = -R\tanh\left(\frac{1}{2}\frac{qV_{we}}{kT}\right)$$

Funkcja ta ma następujące właściwości:

- dla małych wartości różnicowego napięcia wejściowego ($|V_{we}| < \frac{kT}{q}$) zależność V_{wy} od V_{we} jest praktycznie liniowa,
- dla dostatecznie dużych wartości różnicowego napięcia wejściowego ($|V_{we}| > 2\frac{kT}{q}$) układ działa jako ogranicznik amplitudy,
- małosygnałowe wzmocnienie napięciowe k_u na liniowym odcinku charakterystyki wynosi

Równanie 3-43

$$|k_u| = \frac{qI}{2kT}R = g_m R$$

gdzie g_m jest transkonduktancją pojedynczego tranzystora dla prądu kolektora $I_C = I/2$.

Dla tranzystorów MOS uzyskanie prostego wzoru opisującego całą charakterystykę przejściową $V_{wy} = f(V_{we})$ nie jest możliwe, bowiem dla dużych wartości różnicowego napięcia wejściowego jeden bądź drugi z tranzystorów przestaje pracować w zakresie nasycenia. Jeżeli ograniczyć się do napięć wejściowych, dla których oba tranzystory znajdują się w stanie nasycenia, to łatwo otrzymać

Równanie 3-44

$$V_{wy} = -V_{we} R \sqrt{\mu C_{ox} \frac{W}{L} I}$$

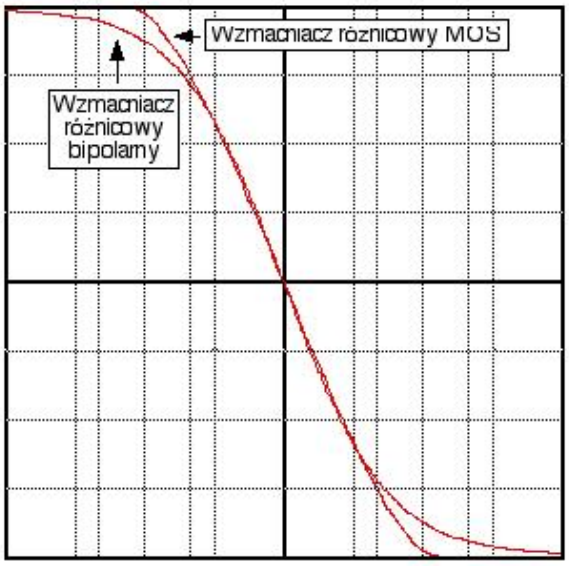
skąd dla wzmocnienia napięciowego otrzymujemy

Równanie 3-45

$$|k_u| = R \sqrt{\mu C_{ox} \frac{W}{L} I} = g_m R$$

gdzie g_m jest transkonduktancją pojedynczego tranzystora dla prądu $I_D = I/2$.

Gdy oba tranzystory znajdują się w stanie nasycenia, charakterystyka przejściowa wzmacniacza MOS jest liniowa. Jednak różnice kształtu charakterystyk wzmacniacza bipolarnego i MOS 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.



Rysunek 3-27. 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 wcześniej (punkty 2.2.2 i 2.2.3). Zasady minimalizacji wpływu rozrzutów lokalnych omówione dla źródeł prądowych obowiązują także dla par tranzystorów we wzmacniaczach różnicowych. Zastosować można topografię pokazaną na rysunku 3-2. 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 (o dwukrotnie mniejszej szerokości kanału w przypadku tranzystorów MOS lub dwukrotnie mniejszej powierzchni złącza emiterowego w przypadku tranzystorów bipolarnych) i połączeniu ich równolegle na krzyż (rysunek 3-28). 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.

Rysunek 3-28. 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

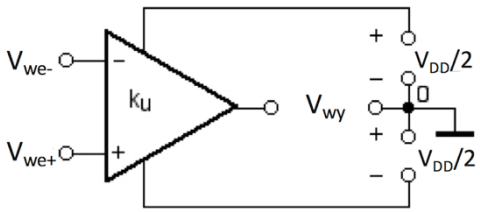
3.3.2 Wzmacniacz operacyjny

Klasycznym zastosowaniem wzmacniacza różnicowego jest użycie go do budowy wzmacniacza operacyjnego. Wzmacniacz operacyjny jest to układ elektroniczny mający różnicowe wejście i asymetryczne wyjście (rysunek 3-29), a zależność napięcia wyjściowego od różnicowego napięcia wejściowego jest dana wzorem

Równanie 3-46

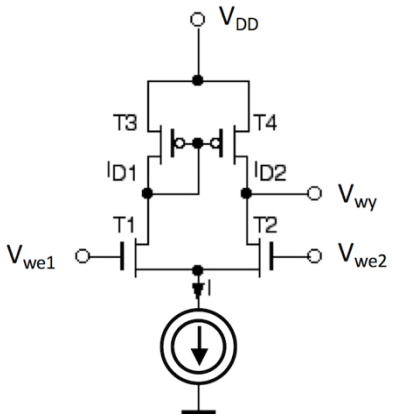
$$V_{wy} = k_u(V_{we+} - V_{we-})$$

przy czym dąży się do tego, by wzmocnienie napięciowe k_u było jak największe (w t.zw. idealnym wzmacniaczu operacyjnym jest ono nieskończenie wielkie).



Rysunek 3-29. 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 (rysunek 3-26). 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 – rysunek 3-30.



Rysunek 3-30. Wzmacniacz różnicowy z aktywnym obciążeniem i asymetrycznym wyjściem

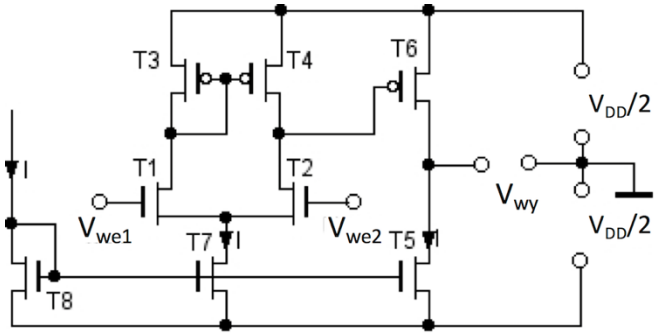
Ze schematu zastępczego tego układu można łatwo otrzymać wartość wzmocnienia napięciowego dla sygnału różnicowego. Wzmocnienie to wynosi

Równanie 3-47

$$|k_u| = \frac{g_{mT1}}{g_{dsT2}+g_{dsT4}}$$

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 rysunku 3-19b, z tym, że tranzystorem wzmacniającym jest tranzystor pMOS, a obciążającym - nMOS), otrzymamy najprostszy kompletny wzmacniacz operacyjny – rysunek 3-31.



Rysunek 3-31. 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 (ang. „Operational Transconductance Amplifier"). Jego wzmocnienie napięciowe otrzymamy mnożąc wzmocnienie stopnia wejściowego przez wzmocnienie stopnia wyjściowego. Daje to zależność:

Równanie 3-48

$$|k_u| = \frac{g_{mT1}g_{mT6}}{(g_{dsT2}+g_{dsT4})(g_{dsT5}+g_{dsT6})}$$

Zauważmy, że w wersji pokazanej na rysunku 3-31 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 drenu I_D (wzór 3-20), a konduktancja wyjściowa g_{ds} proporcjonalna do prądu drenu I_D (wzór 3-21), wzmocnienie napięciowe $|k_u|$ jest odwrotnie proporcjonalne do prądu *I* w układzie.

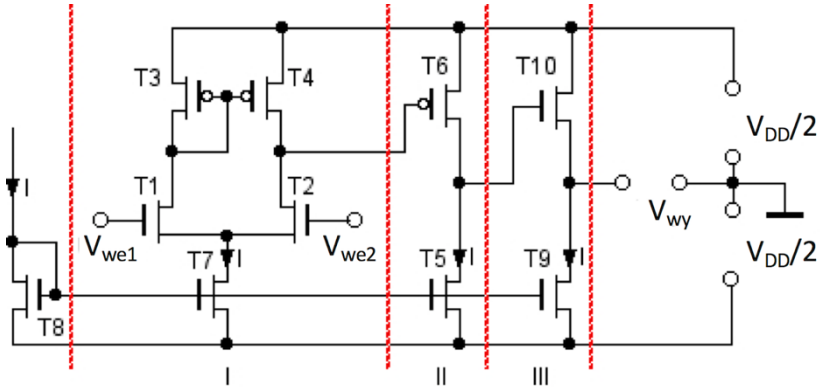
Układ powinien zapewniać dla wejściowego napięcia różnicowego równego zeru napięcie wyjściowe także równe zeru. Można pokazać, że jeśli wszystkie tranzystory są w stanie nasycenia, to prowadzi to do następującego warunku

Równanie 3-49

$$\left(\frac{W}{L}\right)_{T3}\left(\frac{L}{W}\right)_{T6} = \left(\frac{W}{L}\right)_{T7}\left(\frac{L}{2W}\right)_{T5}$$

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 rozrzutów 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 (rysunek 3-25a). Przykład takiego układu pokazuje rysunek 3-32.



Rysunek 3-32. Przykład wzmacniacza operacyjnego ilustrujący podział na 3 stopnie

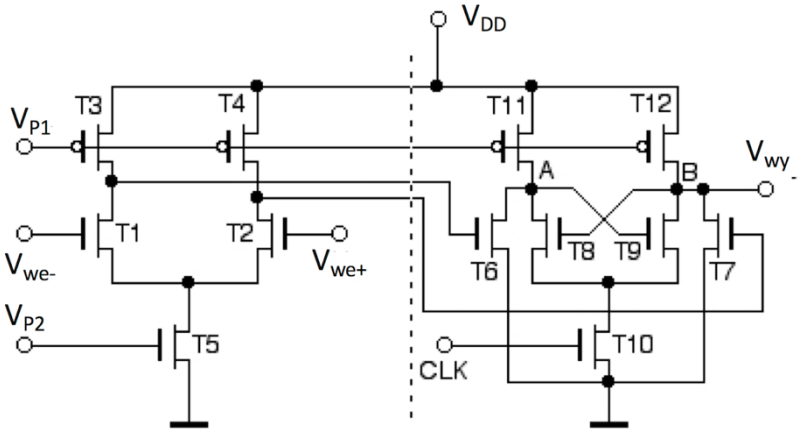
Dla układu z rysunku 3-32 zależność 3-49 nie obowiązuje, ponieważ trzeci stopień wprowadza dodatkowe przesunięcie poziomu napięcia na wyjściu. Układu tego nie będziemy szczegółowo omawiali, posłużył on bowiem tylko do zilustrowania przykładowej budowy trzystopniowego wzmacniacza operacyjnego. Praktyczne układy wzmacniaczy są bardziej skomplikowane. W szczególności, stosowane są zwykle stopnie wyjściowe w układzie przeciwsobnym, które poznamy nieco dalej.

3.3.3 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 (część III, rysunki 4-3 i 4-5) są szczególnym rodzajem komparatorów napięcia. Komparatory napięcia są też niezbędne w układach przetworników analogowo-cyfrowych i w wielu innych układach, w których sygnały analogowe są w jakiś sposób mierzone bądź porównywane z napięciami stałymi lub innymi sygnałami analogowymi.

Komparatory napięcia dzielą się na dwa 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 rysunek 3-33. W tym układzie następuje porównanie napięć na wejściach V_{we+} i V_{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 V_{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.



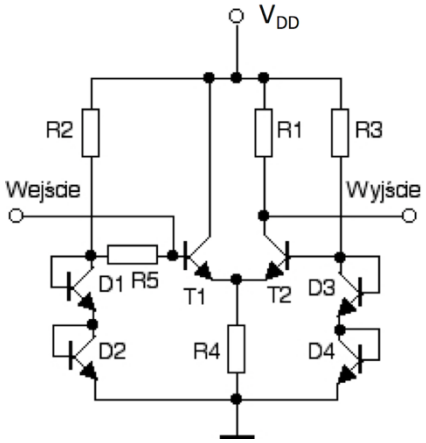
Rysunek 3-33. Prosty komparator napięcia z wejściem zegarowym

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 V_{P2}), a tranzystory T3 i T4 pełnią rolę aktywnych obciążeń (napięcie polaryzacji V_{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 część III, rysunek 3-4). 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 V_{wy} zależy od wartości napięć na bramkach tranzystorów T6 i T7 (tranzystor T10 jest wyłączony, więc przez tranzystory T8 i T9 prąd nie płynie). Porównanie napięć V_{we+} i V_{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 $V_{we+} > V_{we-}$ 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 V_{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.

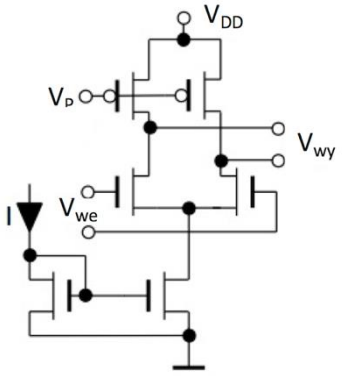
3.3.4 Wzmacniacz szerokopasmowy

Wzmacniacz różnicowy może być użyty do budowy wzmacniaczy szerokopasmowych, tj. przeznaczonych do wzmacniania sygnałów w szerokim zakresie częstotliwości. Pokażemy najpierw ciekawe rozwiązanie takiego stopnia wzmacniającego na tranzystorach bipolarnych:



Rysunek 3-34. Bipolarny 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 (tranzystory w połączeniu diodowym) 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 układ jest 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, rysunek 3-27). Pozwala to usunąć z sygnału FM modulację amplitudy, która – jeśli występuje – jest sygnałem zakłócającym i pogarsza jakość odbioru.



Rysunek 3-35. Wzmacniacz różnicowy MOS jako wzmacniacz szerokopasmowy

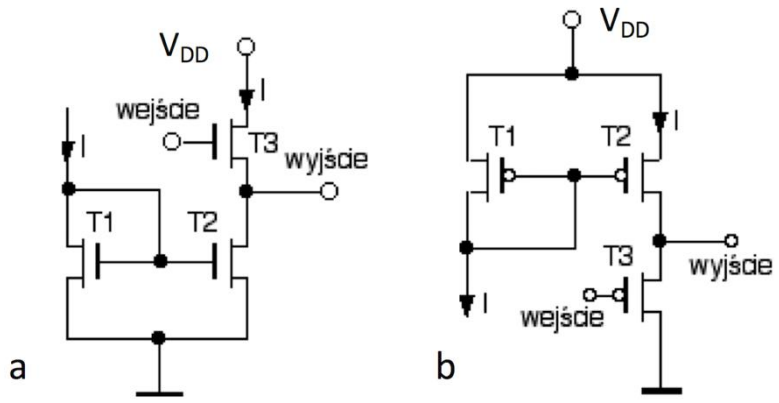
Podobny układ można też zbudować na tranzystorach MOS. Ze względu na małą transkonduktancję tych tranzystorów trzeba użyć obciążenia aktywnego, nie stosuje się również diod i rezystorów do polaryzacji tranzystorów, ale zasada pozostaje ta sama. Przykład pokazuje rysunek 3-35.

3.4 Układy 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.

3.4.1 Wtórnik

Jako stopień wyjściowy może być w najprostszym przypadku użyty układ wtórnika, który poznaliśmy wcześniej w zastosowaniu do przesuwania poziomu składowej stałej (rysunek 3-25a) i jako stopień wyjściowy wzmacniacza operacyjnego (rysunek 3-32). Powtórzymy tu dla wygody schematy z rysunku 3-25.



Kopia rysunku 3-25: Układy przesuwania poziomu składowej stałej: (a) z tranzystorami nMOS, (b) z tranzystorami pMOS

Dodamy też 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

Równanie 3-50

$$|k_u| = \frac{g_{mT3}}{g_{mT3}+g_{dsT3}+g_{dsT2}+\frac{1}{R_L}}$$

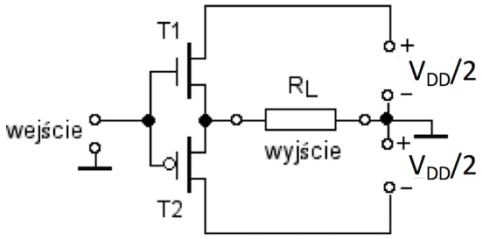
Z zależności tej wynika, że wtórnik ma wzmocnienie bliskie 1 tylko wtedy, gdy $g_{mT3} \gg g_{dsT3} + g_{dsT2} + \frac{1}{R_L}$. Zatem, znając rezystancję obciążenia R_L należy tak zaprojektować układ wtórnika, aby transkonduktancja g_{mT3} była dostatecznie duża.

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 V_{BS}. To napięcie ma wpływ na napięcie progowe tranzystora (część I, wzór 3-6). Napięcie polaryzacji podłoża V_{BS}, wpływając na napięcie progowe V_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 V_{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 (rysunek 3-25b) jest korzystniejszy od układu z tranzystorami nMOS (rysunek 3-25a).

W najnowszych, zaawansowanych technologiach jest możliwość wykonania, obok znanych nam wysp typu *n*, także odrębnych, odizolowanych od podłoża i niezależnie od podłoża polaryzowanych wysp typu *p* dla tranzystorów nMOS (takie technologie są dość skomplikowane, nie omawialiśmy ich). Wówczas tranzystor nMOS T3 (rysunek 3-25a) może być umieszczony na takiej wyspie, podłączonej do źródła tranzystora, i negatywny efekt omówiony wyżej znika.

3.4.2 Układ szeregowy przeciwsobny

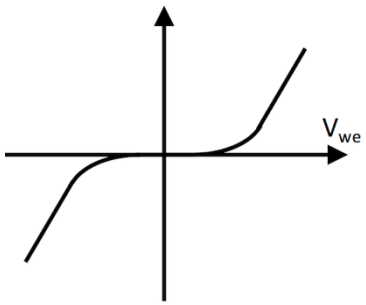
Omówiony wyżej układ wtórnika w praktyce wystarcza, jeśli rezystancja obciążenia jest rzędu kiloomów lub większa. Dla mniejszych rezystancji obciążających (lub gdy układ obciążony jest znaczną pojemnością) potrzebne są inne układy. Stosowane są wtedy stopnie w układzie szeregowo-przeciwsobnym. Zasadę budowy takiego układu, w wersji CMOS, pokazuje rysunek 3-36:



Rysunek 3-36. Zasada budowy układu szeregowo-przeciwsobnego

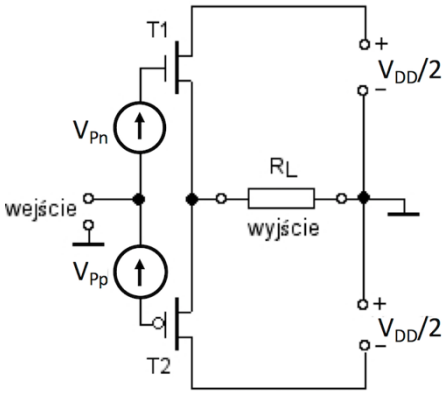
W tym układzie przy dodatnim półokresie sygnału sterującego prąd tranzystora T2 wzrasta, a T1 - maleje. Różnica tych prądów stanowi prąd wyjściowy płynący przez zewnętrzną rezystancję R_L. Przy ujemnym półokresie wzrasta prąd T1, a maleje - T2. Oba tranzystory pracują jako wtórniki, których rezystancją 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 3-36, 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. Jakościowo charakterystykę przejściową układu ilustrującą to zjawisko pokazuje rysunek 3-37.

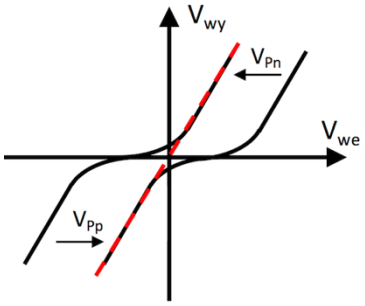


Rysunek 3-37. Charakterystyka przejściowa układu pokazanego na rysunku 3-36

Tę wadę można usunąć wprowadzając dodatkowe źródła napięcia wstępnie polaryzujące bramki tranzystorów (rysunek 3-38). 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 rysunku 3-39.



Rysunek 3-38. Układ szeregowo-przeciwsobny, zasada wstępnej polaryzacji bramek tranzystorów

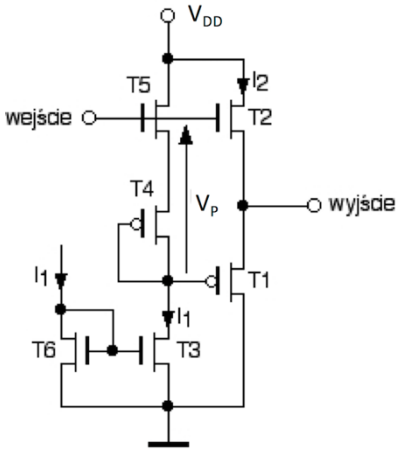


Rysunek 3-39. Charakterystyka przejściowa (czerwona linia) układu pokazanego na rysunku 3-38

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 V_{Pn} i V_{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 V_{Pn} i V_{Pp}. Zależy od tego stabilność pracy układu przy zmianach temperatury. Pokażemy to na praktycznych przykładach. Oto praktyczny układ szeregowo-przeciwsobny w wersji CMOS, w którym napięciami V_{Pn} i V_{Pp} są napięcia V_{GS} tranzystorów T4 i T5:

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.



Rysunek 3-38. Stopień wyjściowy CMOS w układzie szeregowo-przeciwsobnym - układ praktyczny

Wartość napięcia stałego V_P decyduje o wartości prądu spoczynkowego I_2 tranzystorów wyjściowych T1 i T2. Napięcie to jest sumą V_{GST5} i V_{GST4}, a tym samym V_{GST1} i V_{GST2}. Zakładając, że wszystkie tranzystory pracują w zakresie nasycenia, i rozpisując odpowiednie wyrażenia dla napięć można pokazać, że jeśli spełnione są warunki

Równanie 3-51

$$\frac{\frac{W_{T1}}{L_{T1}}}{\frac{W_{T4}}{L_{T4}}} = \frac{\frac{W_{T2}}{L_{T2}}}{\frac{W_{T5}}{L_{T5}}} = m$$

to prąd I_2 ma wartość

Równanie 3-52

$$I_2 = mI_1$$

Można pokazać, że przy właściwie dobranym napięciu V_P charakterystyka przejściowa układu jest liniowa, jeśli dodatkowo spełniony jest warunek jednakowych wydajności prądowych obu tranzystorów, czyli

Równanie 3-53

$$\frac{\frac{W_{T1}}{L_{T1}}}{\frac{W_{T2}}{L_{T2}}} = \frac{\mu_n}{\mu_p}$$

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

Równanie 3-54

$$g_{wy} = \frac{1}{r_{wy}} = g_{mT1} + g_{mT2}$$

W praktyce rezystancja wyjściowa r_{wy} ma zwykle 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. Dzieje się tak dlatego, że dla osiągnięcia dużej wartości

transkonduktancji g_m i dostarczania dużego prądu do obciążenia dużą wartość musi mieć prąd drenu (wzór 3-20). Ale tranzystory MOS, aby dostarczać prądy rzędu amperów, musiałyby mieć olbrzymie wartości stosunku *W/L.* Takie tranzystory MOS, jakie można wytworzyć w standardowych technologiach CMOS (również tych najbardziej zaawansowanych), musiałyby zajmować olbrzymie powierzchnie krzemu. Nie miałoby to technicznego i ekonomicznego sensu.

Istnieją jednak konstrukcje tranzystorów MOS dużej mocy, które pozwalają osiągnąć duże wartości stosunku *W/L* przy umiarkowanej powierzchni zajmowanej przez tranzystor. Kanały tych tranzystorów znajdują się w głębi płytki krzemowej, ukośnie lub pionowo w stosunku do powierzchni. Produkcja układów z takimi tranzystorami jest technologicznie dużo bardziej skomplikowana niż w przypadku standardowych technologii CMOS, i nie będziemy jej omawiać. Układy z tranzystorami dużej mocy MOS znajdują zastosowanie m.in. w elektronice motoryzacyjnej, ze względu na to, że tranzystory MOS są mniej narażone na uszkodzenia przy pracy w wysokiej temperaturze. Dlaczego? O tym będzie mowa dalej.

Prostszym i od dawna stosowanym sposobem budowy układów scalonych dostarczających duży prąd do obciążenia jest zastosowanie układów na tranzystorach bipolarnych. O tym także będzie mowa dalej.

3.5 Układy nieliniowe

Układy omawiane do tej pory w punktach 3.2 i 3.3 zaliczają się do kategorii układów liniowych (wyjątek stanowi komparator napięcia). Układy liniowe to takie, które wykonują na sygnale operacje opisane funkcją liniową, na przykład taką, jak wzór 3-46 – w określonym zakresie napięć sygnał wyjściowy jest wprost proporcjonalny do sygnału wejściowego. Istnieją także układy analogowe wykonujące na sygnale operacje nieliniowe. Zapoznamy się z dwoma przykładami.

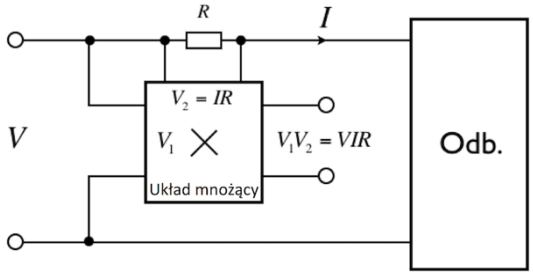
3.5.1 Układy mnożące

Analogowy układ mnożący jest to układ, który ma dwa wejścia i jedno wyjście, a jego charakterystyka przejściowa jest opisana zależnością:

Równanie 3-55

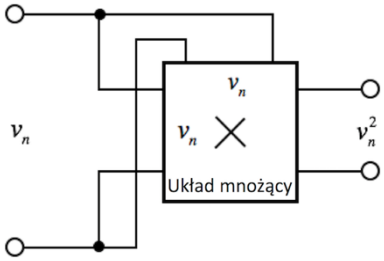
$$V_{wy} = kV_{we1}V_{we2}$$

gdzie V_{we1}i V_{we2} są napięciami na wejściach, V_{wy} jest napięciem na wyjściu, a k jest stałym współczynnikiem. Analogowe układy mnożące mają wiele zastosowań. Można przy ich pomocy budować inne analogowe układy nieliniowe: układy dzielące, podnoszące do kwadratu itp. Układy mnożące służą też do wykonywania operacji modulacji amplitudowej i częstotliwościowej, jak również demodulacji sygnałów z modulacją amplitudy lub częstotliwości i jako mieszacze, czyli układy, które z sygnałów o dwóch różnych częstotliwościach tworzą sygnał mający składowe o częstotliwościach będących sumą i różnicą częstotliwości. Oto przykłady takich zastosowań. Układ do pomiaru poboru mocy:



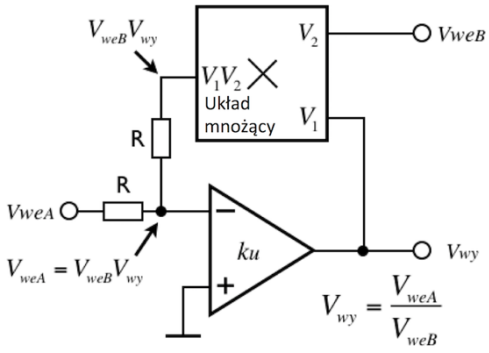
Rysunek 3-39. Układ do pomiaru poboru mocy

Układ do pomiaru kwadratu amplitudy sygnału (taki układ potrzebny jest na przykład do pomiaru mocy szumów):



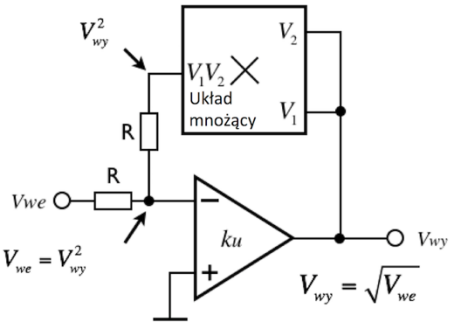
Rysunek 3-40. Układ do pomiaru kwadratu amplitudy sygnału zmiennego

Układ dzielący, wykorzystuje układ mnożący i wzmacniacz operacyjny:



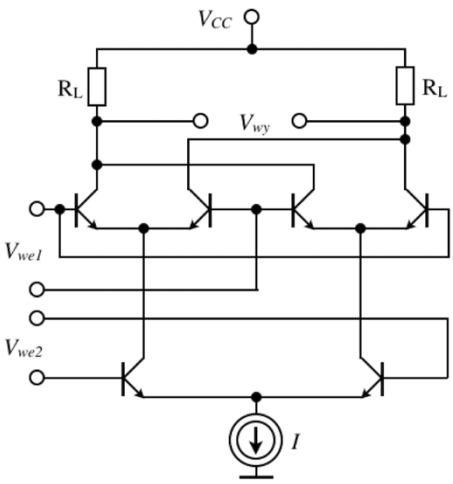
Rysunek 3-41. Układ wykonujący operację dzielenia

Układ pierwiastkujący, wykorzystuje układ mnożący i wzmacniacz operacyjny:



Rysunek 3-42. Układ wykonujący operację pierwiastkowania

Teraz zobaczymy, jak zbudować scalony układ mnożący. Łatwiej to zrobić w technologii bipolarnej. Służy do tego układ zwany układem Gilberta.



Rysunek 3-43. Podstawowy układ Gilberta

Efekt mnożenia bierze się stąd, że dla dwóch górnych wzmacniaczy różnicowych, sterowanych napięciem V_{we1}, źródłami prądowymi są dwie gałęzie dolnego wzmacniacza różnicowego sterowanego napięciem V_{we2}. Wykorzystując zależności 3-39 do 3-42 możemy dla omawianego układu otrzymać

Równanie 3-56

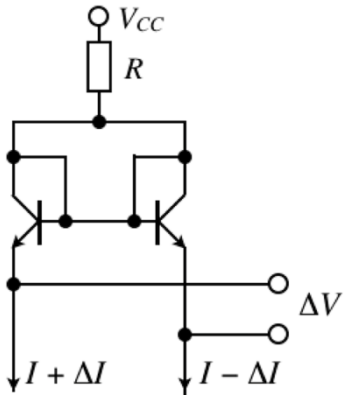
$$V_{wy} = IR_L \tanh\left(\frac{qV_{we1}}{2kT}\right) \tanh\left(\frac{qV_{we2}}{2kT}\right)$$

Dla małych amplitud napięć wejściowych (mniejszych od kT/q) można wyrażenie 3-56 uprościć do postaci

Równanie 3-57

$$V_{wy} \approx \frac{IR_L}{4}\left(\frac{kT}{q}\right)^2 V_{we1}V_{we2}$$

Układ z rysunku 3-45 można jednak rozbudować tak, że będzie prawidłowo mnożył także sygnały o dużej amplitudzie. Potrzebny jest do tego układ, którego charakterystyka jest odwrotna do funkcji tangens hiperboliczny. Oto taki układ:



Rysunek 3-44. Układ o charakterystyce odwrotnej do tangensa hiperbolicznego

Funkcja odwrotna do tangensa hiperbolicznego może być wyrażona tak:

Równanie 3-58

$$\tanh^{-1}(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right)$$

skąd dla układu z rysunku 3-46 można, obliczając różnicę spadków napięć na dwóch tranzystorach w połączeniach diodowych, otrzymać

Równanie 3-59

$$\Delta V = \frac{2kT}{q}\tanh^{-1}\left(\frac{\Delta I}{I}\right)$$

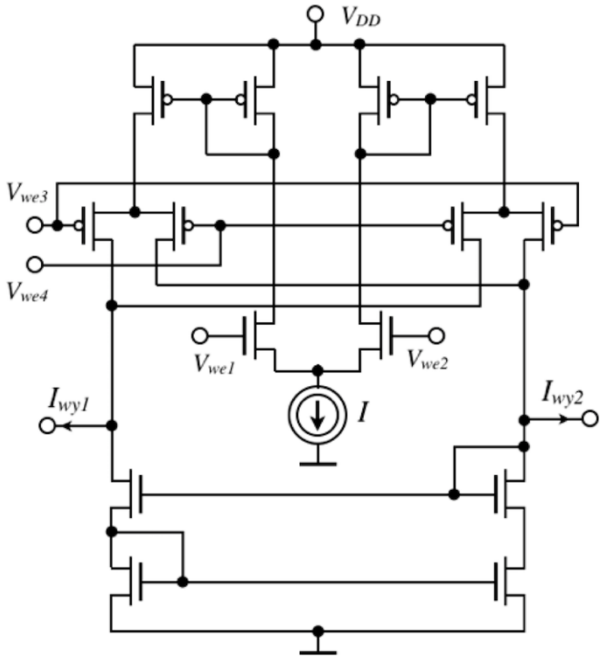
Łącząc odpowiednio schematy 3-45 i 3-46 otrzymamy pełny układ mnożący.

Nieco trudniej jest zbudować dobry układ mnożący CMOS, bowiem charakterystyki tranzystorów MOS nie podsuwają żadnego prostego rozwiązania. Istnieją układy wykorzystujące charakterystyki podprogowe (czyli funkcje wykładnicze, przez analogię do układu bipolarnego Gilberta). Jest sporo innych rozwiązań. Poniżej układ, w którym wykorzystana jest liniowa część charakterystyki wzmacniacza różnicowego MOS (porównaj zależność 3-44). Oto układ zbudowany podobnie do bipolarnego układu Gilberta, w którym różnica prądów wyjściowych jest wynikiem operacji mnożenia

Równanie 3-60

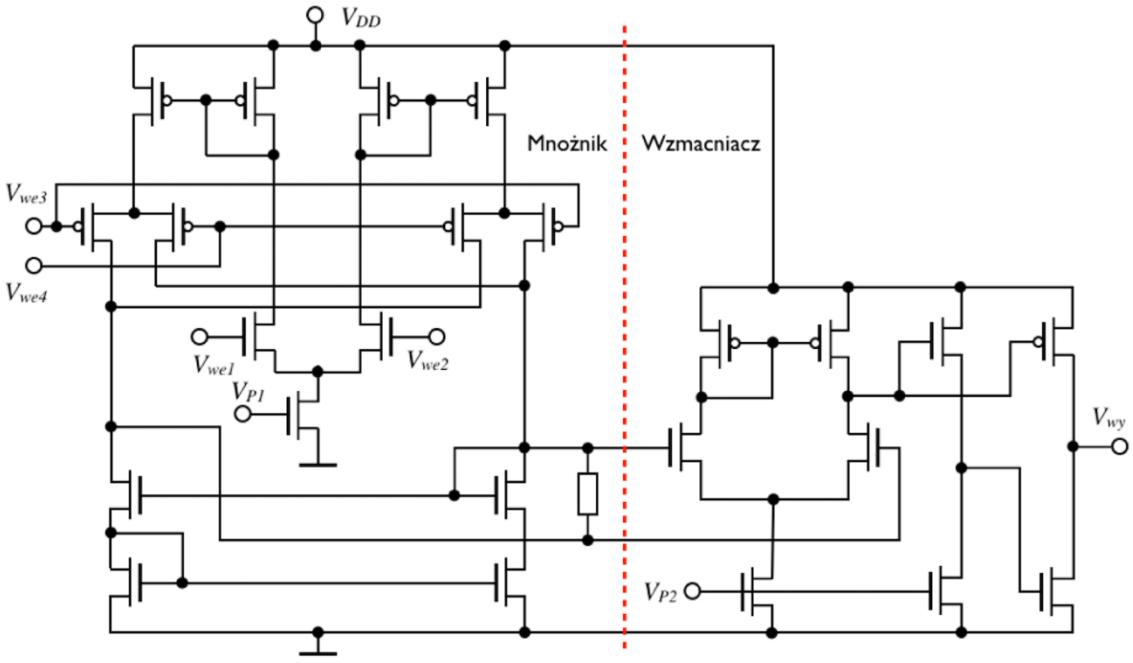
$$I_{wy1} - I_{wy2} = a(V_{we1} - V_{we2})(V_{we3} - V_{we4})$$

gdzie a jest stałym współczynnikiem zależnym od wymiarów kanałów tranzystorów i ich punktów pracy.



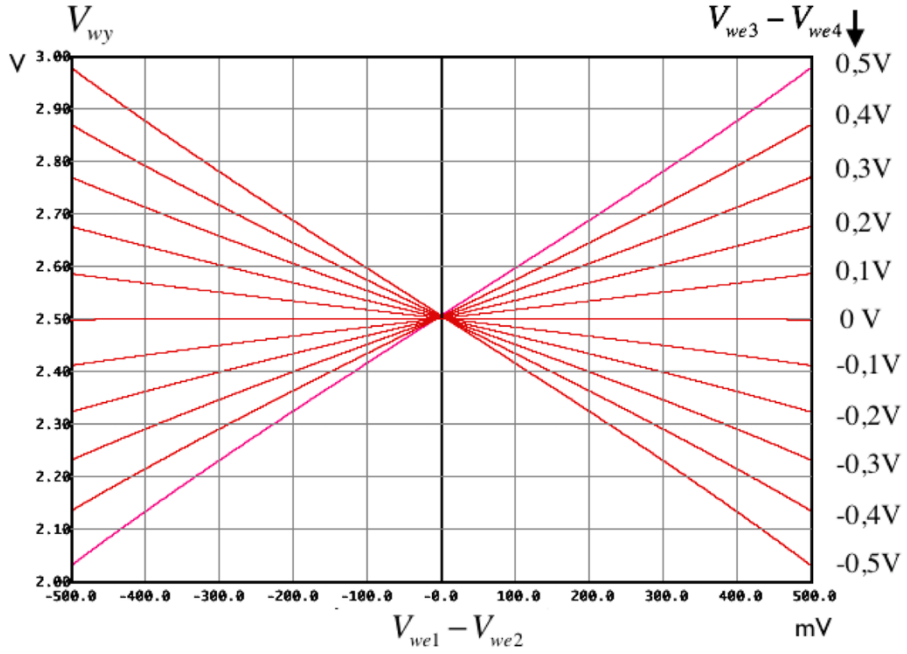
Rysunek 3-45. Podstawowy układ mnożący MOS

Jeśli prąd będący różnicą $I_{wy1} - I_{wy2}$ przepuścić przez rezystor i spadek napięcia na tym rezystorze wzmocnić przy użyciu wzmacniacza operacyjnego, uzyskamy kompletny układ mnożący MOS:



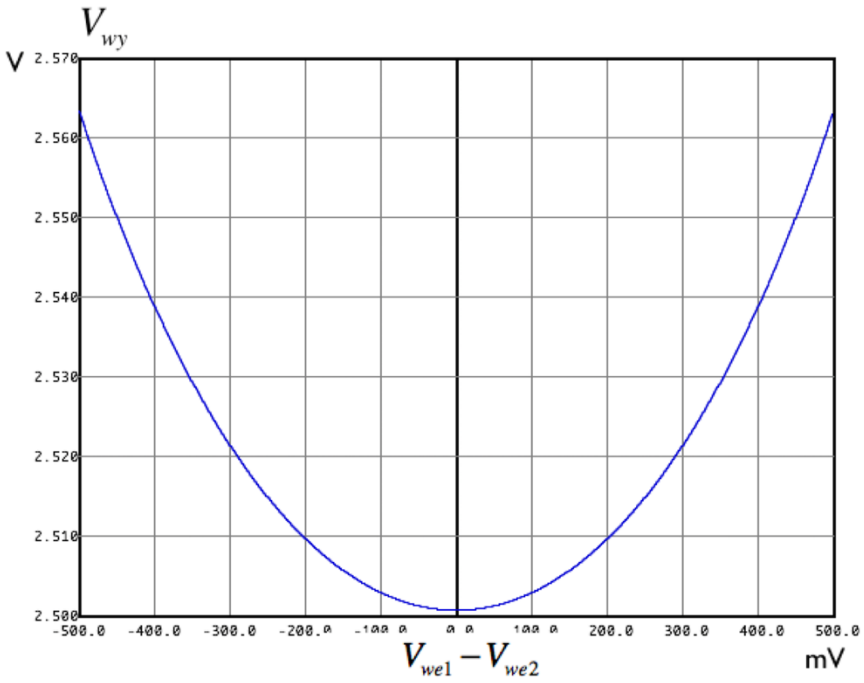
Rysunek 3-46. Kompletny układ mnożący MOS

W tym układzie bez trudu rozpoznamy omawiany poprzednio układ mnożący (rysunek 3-47) i wzmacniacz operacyjny. Dla przejrzystości na rysunku 3-48 nie pokazano źródeł napięć polaryzujących V_{P1} i V_{P2}. Następne dwa rysunki pokazują charakterystyki tego układu otrzymane przy użyciu symulatora układów elektronicznych:



Rysunek 3-49. Charakterystyki przejściowe układu mnożącego MOS

Rysunek 3-50 pokazuje charakterystykę układu, gdy to samo napięcie doprowadzone jest do obu wejść – układ generuje parabolę, czyli „podnosi napięcie do kwadratu”.



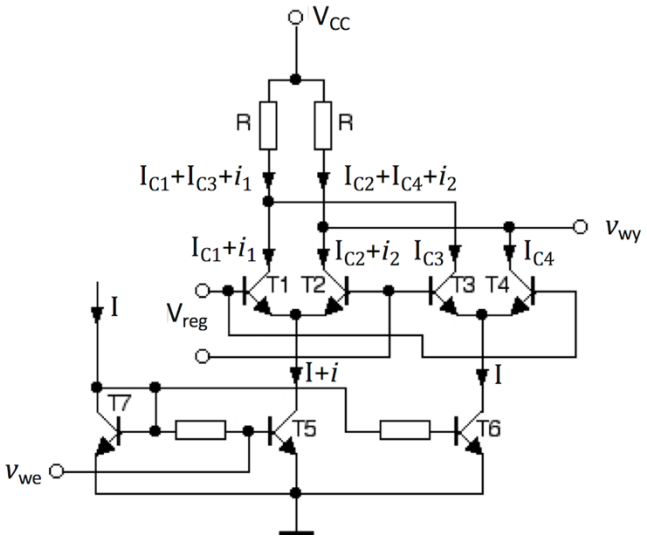
Rysunek 3-50. Charakterystyka przejściowa układu mnożącego MOS ze zwartymi wejściami

3.5.2 Układy regulacji amplitudy napięcia zmiennego

Układy regulacji amplitudy napięcia zmiennego to specyficzny rodzaj wzmacniaczy o zmiennym współczynniku wzmocnienia napięciowego k_u. Wartość tego współczynnika zależy od napięcia regulującego V_{reg}, które jest napięciem stałym lub wolnozmiennym. Układy takie mają kilka ważnych zastosowań. Jedno z nich to automatyczna regulacja wzmocnienia – potrzebna praktycznie w każdym urządzeniu radioodbiorczym, od zwykłego radioodbiornika do odbiorników w profesjonalnym sprzęcie radiokomunikacyjnym czy w telefonach komórkowych. Wartość napięcia na wejściu odbiornika może wahać się w zakresie wielu rzędów wielkości, na przykład, gdy odbiornik przemieszcza się od bliskiej okolicy nadajnika aż do granicy zasięgu. Zwykłe układy wzmacniaczy o stałym wzmocnieniu działają prawidłowo tylko w pewnym zakresie amplitud sygnału wejściowego. Gdy sygnał jest słaby, wzmocnienie może być zbyt małe. Gdy jest zbyt silny, wzmacniacz jest

przesterowany, co powoduje jego wadliwe działanie. Innym przykładem zastosowań układów regulacji amplitudy napięcia zmiennego jest elektroniczna regulacja poziomu dźwięku w urządzeniach elektroakustycznych. Kiedyś taka regulacja odbywała się przy użyciu potencjometru – mechanicznie regulowanego dzielnika napięcia. Dziś w powszechnym użyciu jest regulacja przy pomocy pilota, który wysyła do urządzenia sygnał w podczerwieni, ten jest odbierany i przekształcany na napięcie, które steruje elektronicznym układem regulacji amplitudy sygnału akustycznego. W obu tych zastosowaniach istotne jest, że regulacja musi obejmować bardzo szeroki zakres, bowiem w obu przypadkach amplituda sygnału regulowanego może się zmieniać nawet o kilkanaście rzędów wielkości. Aby układ regulacji był skuteczny w tak szerokim zakresie, jego charakterystyka regulacji (czyli zależność amplitudy wyjściowej regulowanego sygnału od napięcia regulującego) powinna być wykładnicza lub zbliżona do funkcji wykładniczej. W przypadku zastosowań w elektroakustyce dodatkowym argumentem jest to, że ucho ludzkie reaguje logarytmicznie na poziom dźwięku, więc regulacja o charakterze funkcji wykładniczej jest subiektywnie odbierana jako liniowa zmiana poziomu dźwięku.

Wymaganie wykładniczej charakterystyki regulacji w naturalny sposób sugeruje użycie tranzystorów bipolarnych z ich wykładniczą zależnością prądu kolektora od napięcia emiter-baza (część I, wzór 3-12). Oto ciekawy przykład takiego układu, w którym zastosowano wzmacniacze różnicowe w nietypowy sposób:

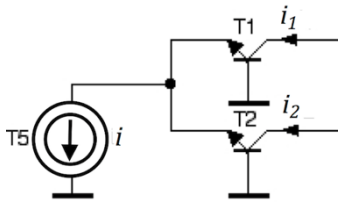


Rysunek 3-51. Układ regulacji amplitudy napięcia zmiennego

Układ jest zbudowany z dwóch wzmacniaczy różnicowych zasilanych jednakowymi prądami *I* przez układ źródeł prądowych (tranzystory T5, T6, T7). Napięcie V_{reg} doprowadzone jest do wejść obu wzmacniaczy różnicowych, z tym, że wejścia te są skrzyżowane: 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 poszczególnych tranzystorów mogą się zmieniać w bardzo szerokim zakresie, a mimo to napięcia na kolektorach tranzystorów T1 – T4 pozostają stałe. Sygnał zmienny v_{we}, 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 prądu rozdziela się pomiędzy tranzystory T1 i T2 w stopniu zależnym od znaku i wartości napięcia V_{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 następującym schematem zastępczym dla składowych zmiennych (gdzie tranzystor T5 jest reprezentowany jako źródło prądu i):

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:



Rysunek 3-52. Rozpływ składowych zmiennych między tranzystory T1 i T2

Równanie 3-61

$$g_{we} = \frac{\partial I_E}{\partial V_{BE}} \approx \frac{\partial I_C}{\partial V_{BE}} = g_m = \frac{q}{kT} I_C$$

gdzie wykorzystano fakt, że w tranzystorze bipolarnym spolaryzowanym normalnie prąd kolektora bardzo mało różni się od prądu emitera, oraz zastosowano wzór 3-24. Skoro składowe zmienne i_1 i i_2 są wprost proporcjonalne do konduktancji wejściowych g_{weT1} i g_{weT2}, a te konduktancje są wprost proporcjonalne do składowych stałych prądów kolektora I_{C1}i I_{C2}, to z tego wynika, że składowe zmienne i_1 i i_2 podzielą się między tranzystory T1 i T2 w takim samym stosunku, jak prądy kolektorów tych tranzystorów I_{C1}i I_{C2}. A o podziale prądu I na prądy I_{C1}i I_{C2} decyduje napięcie regulujące V_{reg}. Stąd po prostych przekształceniach z wykorzystaniem wzoru 3-12 z części I można otrzymać dla składowej zmiennej i_2:

Równanie 3-62

$$i_2 = \frac{i}{e^{\frac{qV_{reg}}{kT}} + 1}$$

gdzie $V_{reg} = V_{BE1} - V_{BE2}$. Składowa zmienna napięcia wyjściowego v_{wy} jest równa $i_2 R$. Dzięki wykładniczej zależności składowej zmiennej i_2 od napięcia V_{reg} możliwa jest regulacja amplitudy w bardzo szerokim zakresie (wiele dekad), i to przy niewielkich zmianach napięcia V_{reg}.

Układu z rysunku 3-51 nie da się powtórzyć w technologii CMOS, ponieważ tranzystory MOS nie mają wykładniczej zależności prądu drenu od napięcia dren-źródło (z wyjątkiem zakresu podprogowego, który jednak obejmuje znacznie węższy zakres napięć i prądów, niż zakres, w jakim obowiązuje zależność 3-12 z części I dla tranzystora bipolarnego). Efekt regulacji o charakterystyce zbliżonej do wykładniczej uzyskuje się budując układy, w których zależność amplitudy napięcia zmiennego od napięcia regulującego daje się w pewnym zakresie lepiej czy gorzej przybliżyć funkcją wykładniczą. Analiza tych układów jest dość skomplikowana, tu je pominiemy.

4 Układy dużej mocy i problemy cieplne w układach scalonych

Ograniczanie poboru mocy jest obecnie jednym z najważniejszych zagadnień w przypadku dużych układów cyfrowych. Omówimy to zagadnienie, będzie też mowa o układach analogowych dużej mocy i w końcu o praktycznym problemie chłodzenia układów scalonych.

4.1 Problem poboru mocy przez układy cyfrowe

4.1.1 Pobór mocy we współczesnych układach CMOS

Najpierw przypomnienie. Wróćmy do części III, punkt 2.2.3, była tam mowa o dwóch składowych mocy pobieranej przez układy cyfrowe: mocy statycznej P_{stat} (część III, wzór 2-9) i mocy dynamicznej mającej dwie składowe: moc P_C związaną z ładowaniem i rozładowywaniem pojemności w układzie (zwana też mocą

przełączania) oraz moc P_j wynikającą z równoczesnego przewodzenia tranzystorów pMOS i nMOS w stanie przejściowym podczas przełączania. Suma obu składowych dynamicznego poboru mocy jest wprost proporcjonalna do liczby zmian stanów logicznych w jednostce czasu f (część III, wzory 2-10 i 2-11). W układach taktowanych zegarem oznacza to, że dynamiczny pobór mocy jest proporcjonalny do częstotliwości zegara.

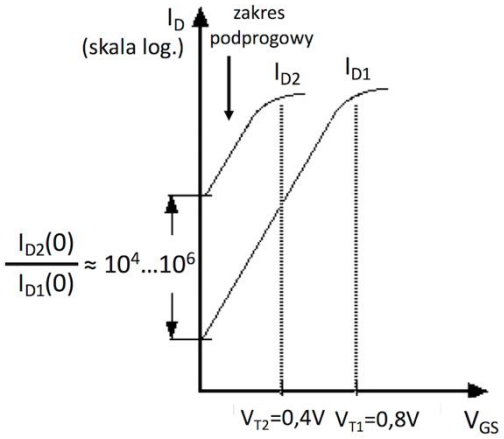
Moc przełączania P_C rośnie proporcjonalnie do kwadratu napięcia zasilania oraz jest proporcjonalna do pojemności obciążającej. Moc jednoczesnego przewodzenia P_j jest proporcjonalna do napięcia zasilania w pierwszej potędze, do wartości maksymalnej prądu płynącego bezpośrednio przez tranzystory w czasie, 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.

Składowa statyczna prądu pobieranego przez bramki była w dawniejszych technologiach CMOS uważana za pomijalnie małą, bowiem w każdej prawidłowo zbudowanej bramce statycznej dla każdej kombinacji stanów na wejściach co najmniej jeden z połączonych szeregowo tranzystorów – nMOS lub pMOS – jest wyłączony i teoretycznie nie ma bezpośredniej drogi dla przepływu prądu ze źródła zasilania. Nie jest jednak tak, że tranzystor wyłączony w ogóle nie przewodzi prądu. Jak wiemy (część I, wzór 3-8), dla napięć bramki mniejszych od napięcia progowego płynie prąd drenu zwany prądem podprogowym. Dla napięcia bramki $V_{GS} = 0$, czyli dla tranzystora wyłączonego oraz dla $V_{DS} \gg \frac{kT}{q}$ wzór 3-8 z części I można uprościć do postaci

Równanie 4-1

$$I_D = I_t \frac{W}{L} e^{-\frac{qV_T}{nkT}}$$

W starszych technologiach, w których napięcia progowe V_T były na poziomie 1V, 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 dalej). Posługując się wzorem 4-1 można pokazać, że prąd podprogowy rośnie o rząd wielkości, gdy napięcie progowe maleje o wartość równą $2{,}3n\frac{kT}{q}$. 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 ... 1 V (typowej dla układów o napięciu zasilania 5V) do 0,3 ... 0,4 V (typowej dla układów o napięciu zasilania rzędu 1 V i poniżej) daje wzrost prądu podprogowego o 4 ... 6 rzędów wielkości. Ilustruje to rys. 14.1. Prąd podprogowy jest wówczas na tyle duży, że związany z nim



Rysunek 4-1. Wzrost wartości prądu podprogowego w wyłączonym tranzystorze MOS

pobór mocy może stanowić znaczącą część całkowitego poboru mocy układu – w układach z bramką o długości poniżej 65 nm może to być nawet do 50% całkowitej mocy pobieranej przez układ w czasie pracy.

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 *p-n* 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 3-4 i 3-5 w części I), jest proporcjonalna do ilorazu współczynnika przenikalności dielektrycznej i grubości dielektryka. Dielektryki o zwiększonej przenikalności dielektrycznej, 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 $V_{GS} = 0$.

Omówione wyżej zależności odnoszą się do wszystkich bramek statycznych CMOS, nie tylko do inwerterów. Natomiast dla bramek dynamicznych, np. dla bramek typu DOMINO (część III, punkt 3.1.2), 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.

Wraz ze zmniejszaniem się wymiarów tranzystorów, wzrostem ich liczby w dużych układach oraz wzrostem częstotliwości taktowania coraz większa moc wydziela się na coraz mniejszej powierzchni układu. Już w okolicach roku 2005 gęstość wydzielanej mocy osiągnęła poziom, którego nie da się przekroczyć, bowiem nie ma sposobu na odprowadzenie od układu dowolnie dużej ilości wydzielającego się ciepła. To spowodowało, że zahamowany został wzrost częstotliwości zegara, a wzrost wydajności obliczeniowej dużych układów cyfrowych

odbywa się w dużej mierze przez doskonalenie ich architektur. W ciągu ostatnich kilku lat pojawiły się jednak też nowe konstrukcje tranzystorów MOS umożliwiające zmniejszenie statycznego poboru mocy. Były one wspomniane w części I, punkt 4.2.3, a zajmiemy się nimi, gdy będziemy omawiać przyszłość mikroelektroniki. Teraz natomiast wskazane będą sposoby zmniejszania poboru mocy, jakie może zastosować projektant układu cyfrowego niezależnie od tego, w jakim wariancie technologii CMOS jego układ będzie produkowany.

4.1.2 Sposoby ograniczania poboru mocy

Omówimy najpierw metody obniżania mocy przełączania. Przypomnijmy, że moc przełączania bramek statycznych jest proporcjonalna do (część III, wzór 2-10):

- pojemności obciążającej bramki C_l,
- kwadratu napięcia zasilania V_{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 część III, punkt 2.2.4),
- 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 połączeń 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 powinien 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 (część III, punkt 2.2.2). 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 punkt 2.2.1 w części III). 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. 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 proporcjonalnie 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 często 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 energooszczędne mikroprocesory z rodziny ARM, powszechnie stosowane w smartfonach, tabletach i innych urządzeniach przenośnych wymagających minimalnego poboru mocy.

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 układów 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. Istnieją natomiast układy globalnie asynchroniczne, lokalnie synchroniczne. Są to układy typu „system on chip", w których jest kilka odrębnych bloków cyfrowych, np. rdzeni mikroprocesorów, które

wewnątrz są układami synchronicznymi, ale wymieniają dane między sobą asynchronicznie, przesyłając te dane przez wewnętrzną sieć działającą na podobnych zasadach, jak sieci zewnętrzne, na przykład sieć Ethernet. Mają swoje protokoły transmisji, sygnały startu i końca transmisji itp. Takie układy bywają nazywane „network on chip". Można tu osiągnąć oszczędności energii, jeśli nie wszystkie bloki muszą pracować równocześnie, a poza tym w takich układach unika się problemu synchronizacji bardzo wielkiego układu przy pomocy pojedynczego zegara (patrz punkt 3.2.6 w części III).

Istnieją też układowe i systemowe sposoby redukcji mocy statycznej. Polegają one na redukcji sumy prądów podprogowych, które płyną w stanie ustalonym przez tranzystory znajdujące się w stanie wyłączenia. Najbardziej oczywistym sposobem jest wspomniane wcześniej „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 (wzór 4-1). 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, używane są wyłącznie w blokach, których szybkość jest krytyczna z punktu widzenia szybkości działania całego układu.

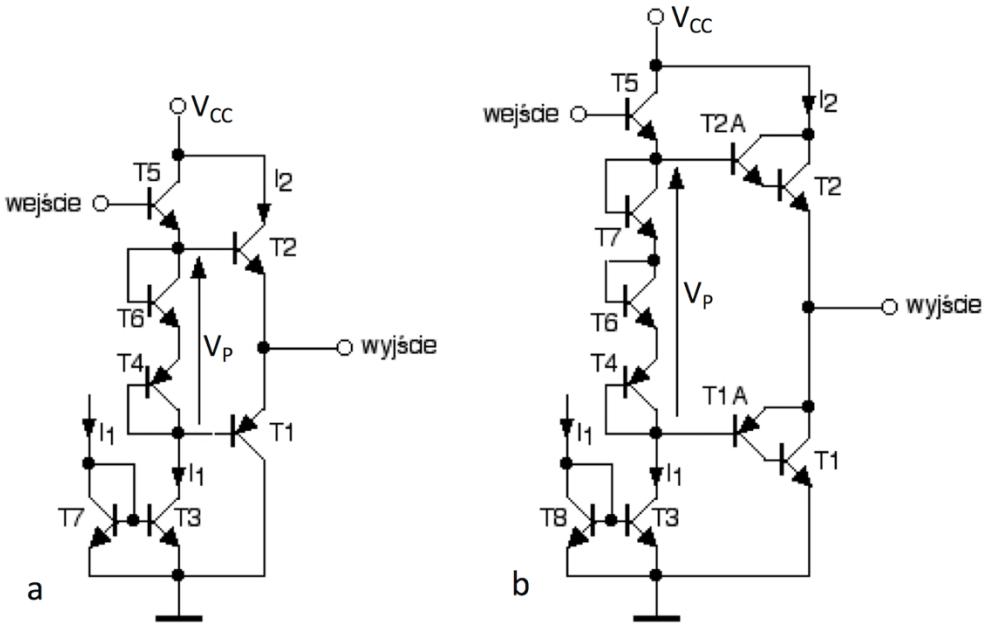
Istnieje też inny sposób obniżania wartości prądów podprogowych, polegający na wykorzystaniu zależności napięcia progowego tranzystora od napięcia polaryzacji podłoża (napięcie V_{BS}, wzór 3-6 w części I). Do wysp (lub podłoża), na których znajdują się tranzystory, doprowadza się napięcie polaryzujące 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. Ten sposób jest dość mało skuteczny w przypadku tradycyjnych technologii CMOS, natomiast może być skutecznie wykorzystany w układach wytwarzanych w nowoczesnej, jednej z przyszłościowych, technologii CMOS FDSOI (patrz część I, punkt 4.2.3). Wrócimy dalej do omówienia tej technologii.

We współczesnych systemach projektowania istnieją specjalne programy ułatwiające automatyzację projektowania z zastosowaniem wymienionych wyżej sposobów redukcji poboru mocy przełączania i mocy statycznej.

4.2 Układy dużej mocy i ich specyficzne problemy cieplne

Jak zbudować układ scalony dużej mocy? I jak duża może być moc oddawana przez taki układ do obciążenia? Co ją ogranicza? To są zagadnienia, które teraz omówimy. W punkcie 3.4.2 pokazany był układ szeregowy przeciwsobny MOS. W układach CMOS wytwarzanych typowymi, standardowymi technologiami taki układ nie może być układem dużej mocy ze względu na to, że w standardowych technologiach CMOS tranzystory MOS oddające do obciążenia prądy o dużym natężeniu musiałyby zajmować ogromne powierzchnie. Istnieją, jak wspomnieliśmy w punkcie 3.4.2, technologie CMOS, w których tranzystory MOS nadające się do układów dużej mocy można wytworzyć. Są to jednak technologie skomplikowane i kosztowne.

Tu przedstawimy problemy układów dużej mocy na przykładzie układów szeregowych przeciwsobnych zbudowanych z tranzystorów bipolarnych. Tak od bardzo dawna wytwarzano i nadal się wytwarza układy scalone dużej mocy. Rysunek 4-2 przedstawia układ szeregowy przeciwsobny w wersji bipolarnej. Zasada działania układu jest taka sama, jak układu MOS, ale szczegóły się różnią.

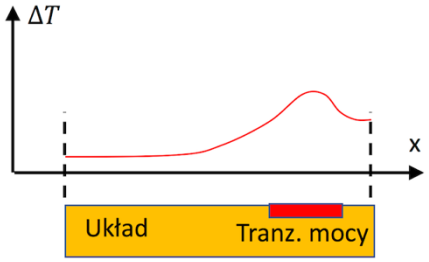


Rysunek 4-2. Bipolarne wyjściowe stopnie przeciwsobne: (a) zasada budowy, (b) układ praktyczny

Krzemowe tranzystory bipolarne w temperaturze otoczenia osiągają znaczącą wartość prądu kolektora, gdy napięcie baza-emiter osiąga wartość około 0,7 V (przykładowa charakterystyka diody krzemowej – rysunek 3-8 w części I, dla prądu kolektora w funkcji napięcia V_{BE} charakterystyka wygląda podobnie). Można powiedzieć, że ta wartość napięcia jest w pewnym sensie odpowiednikiem napięcia progowego tranzystora MOS, choć w przypadku tranzystora bipolarnego pojęcia napięcia progowego się nie używa. W związku z tym bipolarny układ szeregowy przeciwsobny wymaga, podobnie jak układ CMOS, doprowadzenia wstępnego napięcia polaryzującego między bazy tranzystorów bipolarnych, dla usunięcia nieliniowości takiej, jak pokazana na rysunku 3-37, i otrzymania liniowej charakterystyki przejściowej podobnej do tej z rysunku 3-39. Owo napięcie polaryzacji V_P otrzymuje się jako spadek napięcia na tranzystorach T4 i T6 (rysunek 4-2a) w połączeniu diodowym. Jeśli tranzystory T1, T2, T4 i T6 znajdują się w tej samej temperaturze, to prądy I_1 i I_2 są wprost proporcjonalne. Jeśli prąd I1 wymuszany przez źródło prądowe T3 jest temperaturowo stabilny, to i prąd I_1 jest stabilny. Układowi nie grozi zniszczenie na skutek niekontrolowanego wzrostu tego prądu z temperaturą i dodatniego sprzężenia elektryczno-termicznego (patrz punkt 2.2.1).

Układ z rysunku 4-2a nie jest jednak układem często spotykanym w praktyce. W układach bipolarnych tranzystory *p-n-p* nie są pod względem parametrów 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 rysunek 4-2b. 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 V_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.

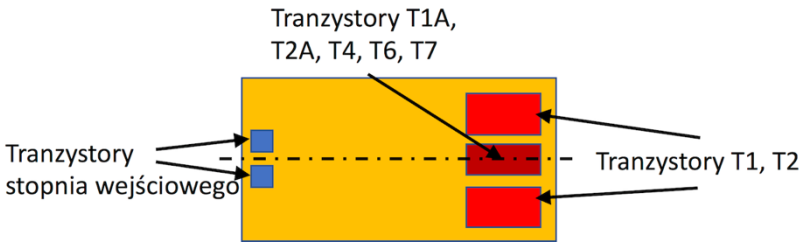
Na zasadzie zilustrowanej rysunkiem 4-2 buduje się stopnie wyjściowe do wielu zastosowań, na przykład do popularnych scalonych wzmacniaczy dużej mocy małej częstotliwości stosowanych w sprzęcie elektroakustycznym. Można osiągnąć moce wyjściowe sięgające dziesiątków W. Powstają jednak wtedy nowe problemy konstrukcyjne. Przy znacznej mocy wydzielanej w układzie jego temperatura może przekraczać temperaturę otoczenia o kilkadziesiąt ^{o}C. Jednym z problemów konstrukcyjnych jest wówczas zapewnienie



Rysunek 4-3. Rozkład temperatury w układzie, w którym wydziela się duża moc. ΔT- przyrost temperatury pod wpływem wydzielanej mocy

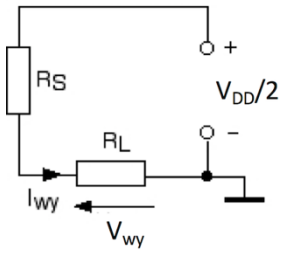
izotermiczności tranzystorów stopnia wyjściowego (T1, T1A, T2, T2A, T4, T6, T7), a od niej zależy, czy układ będzie stabilny. Na powierzchni płytek krzemowych z układami dużej mocy obserwuje się bowiem rozkład temperatur daleki od równomiernego. Różnice temperatur w obrębie układu mogą sięgać kilkudziesięciu stopni. Ilustruje to poglądowo rysunek 4-3.

Aby zapewnić możliwie dobre warunki stabilności układu, pożądane jest ulokowanie tranzystorów T1, T1A, T2, T2A, T4, T6, T7 jak najbliżej siebie, tranzystorów mocy T1, T2 symetrycznie względem osi symetrii układu, a pozostałych elementów, w tym zwłaszcza stopnia wejściowego układu (jest to zwykle wzmacniacz różnicowy) jak najdalej od stopnia wyjściowego (rysunek poglądowy 4-4).



Rysunek 4-4. Tak może wyglądać rozmieszczenie elementów w układzie dużej mocy

Jaka jest i czym jest ograniczona maksymalna moc oddawana do obciążenia przez omawiany układ? Przedstawimy tu bardzo uproszczoną analizę, pozwoli ona jednak na wyciągnięcie uogólniających wniosków. Rozważymy schemat zastępczy układu szeregowego przeciwsobnego dla dodatniego półokresu napięcia sterującego tranzystory mocy. Wróćmy do rysunku 3-36. Przy dodatnim półokresie napięcia sterującego przewodzi górny tranzystor T1, dolny jest wyłączony i można go pominąć. 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 V_{GS} równego napięciu $V_{DD}/2$). Gdy przewodzi wyłącznie górny tranzystor, mamy prosty obwód, w którym połączone są szeregowo tranzystor T1 (MOS lub bipolarny) reprezentowany przez rezystancję R_S, rezystancja obciążenia R_L i źródło napięcia o napięciu równym połowie całkowitej wartości napięcia zasilania (rysunek 4-5).



Rysunek 4-5. Schemat zastępczy do obliczenia maksymalnego prądu wyjściowego

Maksymalny prąd, jaki może płynąć przez rezystancję obciążenia, wynosi

Równanie 4-2

$$I_{wy\ max} = \frac{V_{DD}}{2(R_L+R_S)}$$

zatem maksymalne napięcie wyjściowe wynosi

Równanie 4-3

$$V_{wy\ max} = \frac{V_{DD}R_L}{2(R_L+R_S)}$$

a stąd moc maksymalna wydzielająca się w obciążeniu jest równa

Równanie 4-4

$$P_{wy\ max} = \frac{V_{DD}^2 R_L}{4(R_L+R_S)^2}$$

zaś moc tracona w tranzystorze wynosi

Równanie 4-5

$$P_{tr} = \frac{V_{DD}^2 R_S}{4(R_L+R_S)^2}$$

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 4-4 dostatecznie ilustruje ograniczenia maksymalnej mocy wyjściowej i pozwala je przedyskutować.

Zależność 4-4 pokazuje, że do zwiększania mocy wyjściowej można dążyć na kilka sposobów: przez podnoszenie napięcia zasilania V_{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 V_{DS} oraz maksymalne napięcie dopuszczalne napięcie bramki V_{GS}. W przypadku tranzystorów bipolarnych jest to napięcie dopuszczalne kolektor-emiter V_{CE}. Ponadto w obu przypadkach dostatecznie wysokie musi być napięcie przebicia odpowiednich złącz *p-n*. Nie wdając się w szczegółowe rozważania na temat konstrukcji tranzystorów (co wykracza to poza zakres naszych rozważań) 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 obszaru kolektora oraz obniżanie koncentracji domieszek w obszarze kolektora.

Obniżanie koncentracji domieszek jest potrzebne dlatego, że – jak wiadomo z teorii przyrządów półprzewodnikowych - napięcie przebicia każdego złącza *p-n* 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 obszaru kolektora 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 niekorzystny, bowiem gdy wzrasta rezystancja R_S, to moc oddawana do obciążenia maleje (wzór 4-4). Zatem zmianom w konstrukcji tranzystora umożliwiającym podniesienie napięcia zasilania powinny 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 4-4 i 4-5). 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. Ponadto w przypadku tranzystorów

bipolarnych dużej mocy istnieje ryzyko pojawienia się wewnętrznej niestabilności elektryczno-cieplnej tranzystora prowadzącej do jego zniszczenia zwanej wtórnym przebiciem elektryczno-cieplnym. Zjawisko to może wystąpić w tranzystorach o dużej powierzchni, w których przy dużej wydzielanej mocy na powierzchni tranzystora rozkład temperatury nie jest równomierny (patrz rysunek 4-3). Ponieważ prąd przewodzony przez tranzystor bipolarny rośnie wykładniczo z temperaturą, obszary gorętsze będą obciążone większym prądem, co będzie prowadzić do jeszcze większej koncentracji prądu w tych obszarach kosztem obszarów o niższej temperaturze i w konsekwencji może doprowadzić do lokalnego wzrostu temperatury prowadzącego do nieodwracalnego uszkodzenia tranzystora. Zjawisku temu nie zapobiega pokazany na rysunku 4-2 sposób polaryzacji tranzystorów stopnia mocy, bowiem zjawisko wtórnego przebicia elektryczno-cieplnego polega na zmianie wewnętrznego rozkładu gęstości prądu i temperatury w tranzystorze bez zmiany całkowitego prądu płynącego przez tranzystor.

Z powodu ryzyka wtórnego przebicia elektryczno-cieplnego układy, w których rolę elementów dużej mocy pełnią tranzystory MOS, są bardziej niezawodne. W nich do tego zjawiska nie dochodzi, ponieważ ze wzrostem temperatury prąd w tranzystorze MOS maleje, a nie rośnie. Dlatego opracowano też wspomniane już wcześniej technologie pozwalające wytwarzać tranzystory MOS mające duże wartości stosunku *W/L* kanału przy umiarkowanej powierzchni. Te technologie 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 naszych rozważań.

4.3 Sprawność energetyczna układów dużej mocy

4.3.1 Pojęcie sprawności energetycznej

Maksymalna moc, jaką może oddać układ scalony do obciążenia, jest oczywiście ograniczona także przez 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 więc tym więcej wydziela się ciepła, które trzeba odprowadzić od tranzystorów stopnia mocy.

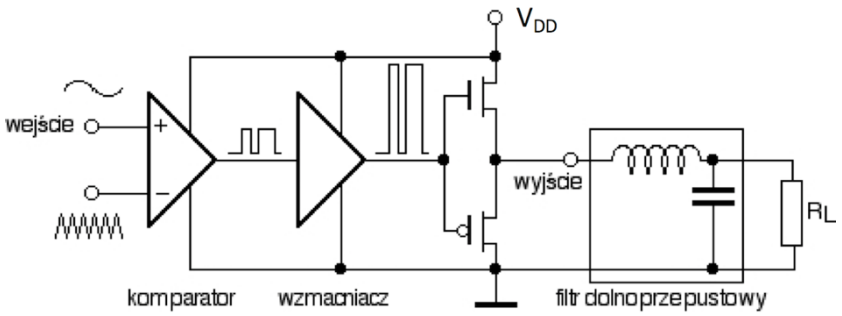
Sprawność energetyczna zależy nie tylko od budowy stopnia mocy i parametrów jego elementów, ale i od rodzaju sygnałów wzmacnianych przez ten stopień. Można pokazać, że maksymalna osiągalna teoretycznie sprawność szeregowego przeciwsobnego stopnia mocy dla czystego sygnału sinusoidalnego wynosi nieco 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 4-4 i 4-5. Jeśli spełniony jest warunek $R_L \gg R_S$, moc P_S wydzielana w tranzystorach jest pomijalnie mała wobec $P_{wy\,max}$, a sprawność zbliża się do 100%.

4.3.2 Układ o najwyższej sprawności – wzmacniacz w klasie D

Spostrzeżenie, że przy sygnałach impulsowych sprawność energetyczna zbliża się do 100%, 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, a 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 szeregowego 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. Rysunek 4-6 przedstawia schemat blokowy wzmacniacza 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 wyższej jakości. Wzmacniacze klasy D znajdują też zastosowania innego rodzaju, na przykład jako układy sterujące pracą silników elektrycznych prądu stałego.



Rysunek 4-6. Schemat blokowy wzmacniacza mocy w klasie D

4.4 Odprowadzanie ciepła od układów scalonych

4.4.1 Dopuszczalna temperatura pracy układów

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. Jak już wiemy, 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.

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.

Elektromigracja polega na tym, że przy dużej gęstości prądu w ścieżkach połączeń ukierunkowany ruch elektronów powoduje wybijanie niektórych atomów metalu z ich położeń i transport w inne miejsce. W rezultacie w miejscach, gdzie występuję zwiększona gęstość prądu (np. przewężenia ścieżki), ścieżka powoli staje się coraz cieńsza i w końcu następuje przerwanie połączenia.

Naprężenia mechaniczne występują w wyniku różnicy rozszerzalności cieplnej krzemu i materiałów z nim połączonych w obudowie układu. Powodują powstawanie defektów sieci krystalicznej półprzewodnika, a w skrajnym przypadku mogą doprowadzić do pęknięcia płytki krzemowej.

Degradacja jakości tlenku bramkowego była już wspominana przy okazji omawiania pamięci nieulotnych typu *flash* (część III, punkt 4.2.6 . Zjawisko to dotyczy jednak, choć w mniejszym stopniu, także wszystkich innych tranzystorów w układach CMOS. Prowadzi do zmian wartości napięcia progowego, co zmienia parametry układu, a może też spowodować jego całkowitą niesprawność.

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.

4.4.2 Chłodzenie układów scalonych

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 ciepła, w którym wydziela się moc P, do odbiornika o temperaturze T_0, to temperatura źródła ciepła T_1 wynika z prostego równania

Równanie 4-6

$$P = \frac{T_1 - T_0}{R_T}$$

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

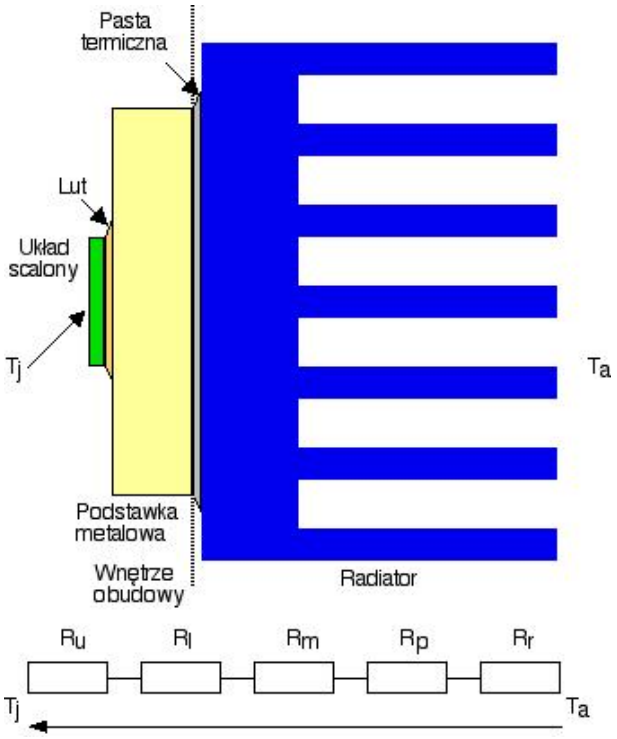
Równanie 4-7

$$R_T = \frac{1}{\lambda_T}\frac{L}{S}$$

gdzie λ_T jest przewodnością cieplną materiału, z którego wykonany jest prostopadłościan. Przewodność cieplna krzemu wynosi około 150 W/(m*°C). Przewodność cieplna metali stosowanych w mikroelektronice jest większa, np. dla miedzi, która jest bardzo dobrym przewodnikiem ciepła, przewodność cieplna wynosi około 400 W/(m*°C), dla aluminium - około 240 W/(m*°C).

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.

Rysunek 5-1 pokazuje układ scalony wraz metalową podstawką, do której jest umocowany wewnątrz obudowy i zewnętrznym radiatorem, oraz zastępczy schemat cieplny.



Rysunek 4-7. 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

Równanie 4-8

$$T_j = T_a + P\left(R_u + R_l + R_m + R_p + R_r\right)$$

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 $P_{max} = \left(T_{jmax} - T_a\right)/R_{thjc}$.
Typowa wartość rezystancji R_{thjc} może wahać się w granicach od ułamka do kilku °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 °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 naszych rozważań.

5 Przyszłość mikroelektroniki

Kończąc poznawanie świata mikroelektroniki warto zastanowić się nad tym, co będzie dalej z tą dziedziną techniki. Przez dziesiątki lat rozwój mikroelektroniki był jednoznacznie kojarzony ze zmniejszaniem się

wymiarów tranzystora, wzrostem szybkości działania układów i stopnia ich komplikacji. Dziś mamy do czynienia z trudnymi do przekroczenia barierami: poboru mocy, którego nie można dalej zwiększać, wymiarów tranzystora, których nie da się zmniejszać bez końca, wreszcie stopnia komplikacji technologii i jej kosztu. Czy są w zasięgu wzroku sposoby pokonania lub ominięcia tych barier? A co, jeśli ich się pokonać nie da?

5.1 Co można osiągnąć w ramach klasycznej technologii CMOS

5.1.1 Reguły skalowania i „prawo Moore'a"

Przez dziesiątki lat postępy mikroelektroniki wytyczała wspomniana na samym początku części I (punkt 2.2.3, rysunek 2-4) reguła zwana prawem Moore'a. Wymiary tranzystorów były zmniejszane w każdej kolejnej generacji technologii. Minimalny wymiar tranzystora – długość kanału – zmniejszano dwukrotnie co mniej więcej 5 – 6 lat. „Prawo Moore'a" było rodzajem samospełniającej się przepowiedni. Czy będzie nadal się spełniać?

Redukowanie wymiarów tranzystorów, a zwłaszcza zmniejszanie długości kanału, ma kluczowe znaczenie dla wzrostu złożoności i szybkości działania układów CMOS, ale także dla ich poboru mocy. 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 (patrz wzory 2-1 i 2-2 w części III) 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 kanału tranzystora. Oznacza to *skalowanie 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 (patrz wzory 2-1 i 2-2 w części III) 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 $1/S^2$).

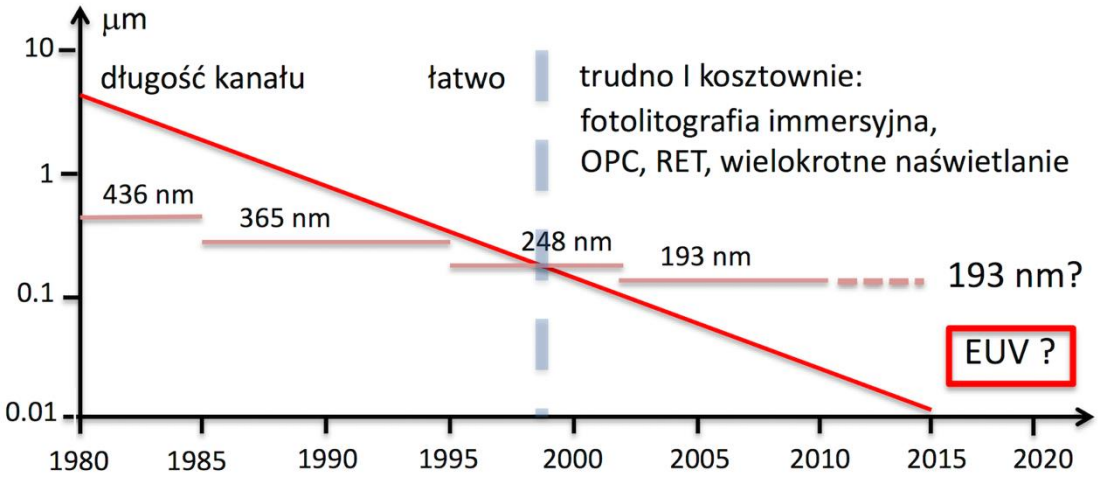
Jak widać, skalowanie przy stałym natężeniu pól elektrycznych prowadzi do zwiększania szybkości działania układów mimo obniżania napięcia zasilającego, przy czym można uniknąć katastrofalnego wzrostu gęstości wydzielanej mocy. Reguły skalowania nie uwzględniają jednak problemu statycznego poboru mocy, który ujawnił się przy długościach kanału poniżej 100 nanometrów, gdy napięcie zasilania układów trzeba było zmniejszyć do 1V lub nieco mniej. Zmniejszaniu napięcia zasilania musi towarzyszyć proporcjonalne zmniejszanie napięć progowych tranzystorów, a to powoduje wykładniczy wzrost prądu podprogowego – wzór 4-1 i rysunek 4-1. 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 w kanale 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}. Ponieważ nie rośnie ruchliwość ani C_{ox}, nie rosną 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 poniżej 100 nm nie prowadzi już, jak dawniej, do istotnego zwiększenia szybkości działania bramek, głównym celem zmniejszania wymiarów stała się możliwość zwiększania liczby tranzystorów w układzie o danej powierzchni, co pozwala budować układy i systemy cyfrowe o coraz bardziej złożonych i wydajnych architekturach. Gdy jednak na powierzchni układu umieszcza się coraz więcej coraz mniejszych tranzystorów, gęstość mocy wydzielanej w układzie rośnie do poziomu, którego przy żadnym sensownym technicznie sposobie chłodzenia nie da się przekroczyć. Trochę pomagają sposoby ograniczania poboru mocy omówione w punkcie 4.1.2. Układy CMOS z tradycyjną budową tranzystora MOS, mimo ulepszeń wspomnianych wyżej i omówionych w punkcie 4.1.2 systemowych sposobów ograniczania poboru mocy, osiągają kres możliwości przy długości kanału na poziomie 28 nanometrów. Dalsze skalowanie układów z takimi tranzystorami nie ma już sensu, tym bardziej, że komplikacje technologii powodują, że przestaje też działać reguła mówiąca, że układy z mniejszymi tranzystorami są tańsze. Istnieją oszacowania, z których wynika, że w przypadku technologii CMOS z tranzystorami o tradycyjnej budowie najniższy koszt przeciętnej bramki logicznej osiągany jest dla technologii z długością kanału 28 nm. Dla krótszych kanałów koszt już nie maleje, lecz rośnie.

A przecież dziś są w produkcji układy z tranzystorami o kanałach znacznie krótszych niż 28 nm. Jak to możliwe? Czy to ma sens techniczny i ekonomiczny? O tym w następnych punktach.

5.1.2 Fotolitografia – historia i przyszłość

Skalowanie – łatwiej powiedzieć (i policzyć teoretycznie), trudniej zrobić. O możliwościach technologicznych redukcji wymiarów tranzystorów decyduje fotolitografia. W części I, w punkcie 4.1.5, omówiona była na najprostszym przykładzie naświetlania płytki półprzewodnika pokrytej fotorezystem przez leżącą na tej płytce maskę. Taki sposób naświetlania stosowano tylko w początkowym okresie rozwoju mikroelektroniki. Jak wspomniano w części I, od dziesiątków lat stosuje się fotolitografię projekcyjną. 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". Aby osiągnąć



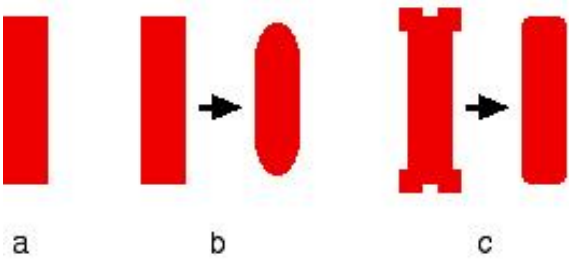
Rysunek 5-1. Zmiany długości kanału w układach CMOS i zmiany długości fali światła stosowanego w fotolitografii

maksymalną możliwą zdolność rozdzielczą, stosuje się światło monochromatyczne o małej długości fali – ultrafiolet. Kierując się zasadami klasycznej optyki uważano, że nie da się precyzyjnie odwzorować kształtów o wymiarach porównywalnych lub mniejszych od długości fali światła służącego do naświetlania. Granicę tę jednak udało się przekroczyć, co pokazuje rysunek 5-1.

Jak widać, tranzystory o długościach kanału rzędu kilkunastu nanometrów są wytwarzane przy zastosowaniu procesu fotolitograficznego o długości fali 193 nanometry. Jak to jest możliwe? Osiąga się to przez zastosowanie

- wstępnego zniekształcania obrazu na maskach,
- masek z kontrastem fazowym,
- fotolitografii immersyjnej,
- podwójnego naświetlania.

Wstępne zniekształcanie obrazu (ang. optical proximity correction, OPC) 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



Rysunek 5-2. Zasada korekcji (wstępnego zniekształcania) obrazu na maskach: (a) pożądany kształt paska polikrzemu, (b) kształt na masce i kształt otrzymany bez korekcji, (c) kształt na masce z korekcją i kształt otrzymany

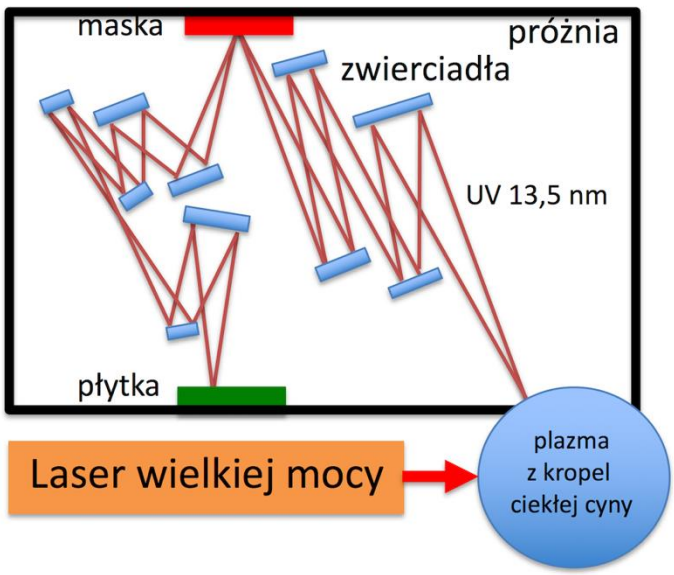
i od ich otoczenia). Idea jest pokazana na rysunku 5-2. Wstępne deformacje obrazów na maskach są wprowadzane przez odpowiednie programy komputerowe, konstruktor nie musi o tych deformacjach wiedzieć. Maski z kontrastem fazowym wykorzystują zjawisko interferencji do podniesienia kontrastu obrazu na granicy 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.

Fotolitografia immersyjna polega na zanurzeniu naświetlanej płytki z fotorezystem w cieczy o wysokim współczynniku załamania światła. W takiej cieczy długość fali świetlnej jest mniejsza. Cieczą taką może być po prostu woda o odpowiednio wysokiej czystości.

Podwójne naświetlanie (ang. double patterning) 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 mniejsze, a po wykonaniu obu procesów naświetlania i wywołaniu fotorezystu otrzymujemy komplet obiektów.

Wymienione wyżej sposoby odwzorowywania kształtów o wymiarach znacznie mniejszych od długości fali światła (ang. resolution enhancement techniques, RET) bardzo utrudniają projektowanie topografii układów. Pojawiają się liczne nowe skomplikowane geometryczne reguły projektowania, ich liczba w zaawansowanych technologiach sięga tysiąca i więcej. Bardzo poważnie rośnie także koszt przygotowania kompletu masek produkcyjnych.

Dlaczego zatem nie zastosować jeszcze krótszych fal światła? Od wielu lat trwają prace nad fotolitografią przy zastosowaniu promieniowania o długości fali 13,5 nm (ang. extreme UV, EUV). Urządzenie do naświetlania EUV działa na zupełnie innej zasadzie niż wcześniej stosowane urządzenia do fotolitografii. Po pierwsze nie nadaje się żadne ze stosowanych wcześniej źródeł światła. Po drugie naświetlanie musi odbywać się w próżni, bo powietrze bardzo silnie tłumi promieniowanie o tak krótkiej fali. Po trzecie nie jest znany żaden materiał optyczny przezroczysty dla tak krótkofalowego promieniowania, więc nie wchodzi w grę żaden obiektyw, układ



Rysunek 5-3. Schemat urządzenia EUV do fotolitografii

optyczny urządzenia do fotolitografii składa się wyłącznie ze zwierciadeł o niezwykle wysokiej precyzji wykonania. Poglądowy schemat urządzenia do fotolitografii EUV przedstawia rysunek 5-3. Źródłem promieniowania jest plazma powstająca pod wpływem naświetlania światłem lasera o wielkiej mocy mikroskopijnych kropelek ciekłej cyny, które są w sposób ciągły dostarczane do specjalnej komory. Wzbudzone w ten sposób promieniowanie jest ogniskowane na masce przez skomplikowany system zwierciadeł, maska jest również zwierciadlana, a nie przezroczysta, po odbiciu od maski promieniowanie zawierające obraz maski jest rzutowane przez drugi system zwierciadeł na naświetlaną płytkę. O stopniu komplikacji tej technologii i koszcie urządzenia EUV niech świadczy fakt, że ma ono wielkość dwupiętrowego domu, a do transportu wymaga rozmontowania na części składowe.

Urządzenia takie są od kilku lat produkowane (w europejskiej firmie ASML), ale nie osiągnęły dotąd wydajności dostatecznej do zastosowań w seryjnej produkcji. Problem w tym, że w układzie optycznym urządzenia EUV do naświetlanej płytki dociera tylko kilka procent energii promieniowanej ze źródła. Na każdym zwierciadle tracone jest bowiem około 30% tej energii. Naświetlanie każdej płytki trwa więc długo, zbyt długo z punktu widzenia opłacalnej ekonomicznie przepustowości linii produkcyjnej. Z fotolitografią EUV związane są też dalsze problemy. Potrzebny jest inny od obecnie używanych fotorezyst, o wyższej czułości. Potrzebne są środki ochrony masek przed zanieczyszczeniami, bowiem pyłek o średnicy jednego nanometra na masce będzie odwzorowany na naświetlanej płytce powodując defekt w układzie scalonym. Produkowane obecnie (początek roku 2019) urządzenia do fotolitografii EUV są na razie wykorzystywane tylko do eksperymentów i przygotowywania przyszłych procesów produkcyjnych. Jednak producenci opracowujący najbardziej zaawansowane technologie mają nadzieję, że wprowadzenie do produkcji fotolitografii EUV poważnie obniży koszt produkcji, bowiem zmniejszy się liczba masek i operacji fotolitografii (nie będzie potrzeby podwójnego naświetlania), a same maski staną się tańsze, uproszczone będzie bowiem ich przygotowanie. Czy i kiedy tak się stanie? Na razie trudno przewidzieć.

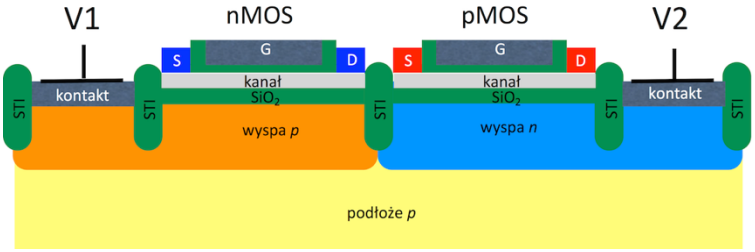
5.1.3 Nadal CMOS, ale z tranzystorami o innej budowie

W części I wspomniane były dwie nowe struktury tranzystorów MOS: FinFET i FDSOI (punkt 4.2.3, rysunki 4-17 i 4-18). Z takimi tranzystorami można nadal wytwarzać układy CMOS. Ich główną zaletą jest redukcja prądu podprogowego o kilkadziesiąt procent, a tym samym ograniczenie mocy statycznej pobieranej przez układy cyfrowe. Można to wykorzystać na dwa sposoby, w zależności od zastosowania układu: albo pozostawić bez zmian częstotliwość taktowania układu, uzyskując obniżenie poboru mocy (na przykład w urządzeniach mobilnych – dłuższa praca po naładowaniu baterii), albo zwiększyć wydajność obliczeniową układu przez podniesienie częstotliwości taktowania, bo obniżenie mocy statycznej umożliwia zwiększenie mocy dynamicznej bez przekroczenia dopuszczalnej mocy całkowitej.

Tranzystory FinFET są trudne w produkcji. Wytrawienie milionów kanałów tranzystorów w postaci cienkich pionowych pasków krzemu w taki sposób, aby żaden z tych pasków nie uległ złamaniu (a krzem monokrystaliczny jest materiałem kruchym) to wyzwanie dla technologów. W porównaniu z tą technologią technologia tranzystorów FDSOI jest bez porównania prostsza, prostsza nawet od tradycyjnej technologii wytwarzania tranzystorów w układach CMOS, bo struktura tranzystora jest taka sama, jak w technologiach tradycyjnych, a w dodatku nie są potrzebne procesy wprowadzania domieszek w obszar kanału – tranzystor ma bardzo dobre parametry bez nich. Trudniej natomiast przygotować podłoże. Wytworzenie ultra-cienkiej monokrystalicznej warstwy krzemu odizolowanej dielektrykiem od krzemowego podłoża nie jest łatwe, ale proces ten jest już dobrze opanowany na skalę przemysłową.

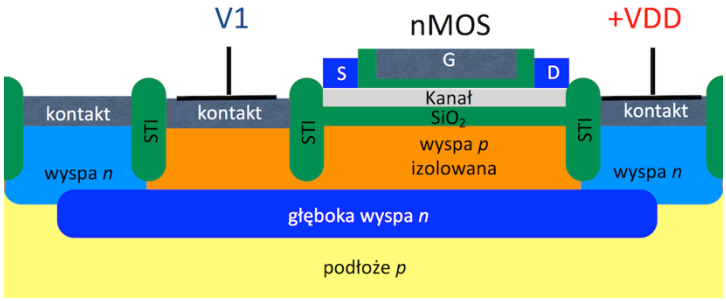
Tranzystory FinFET przydatne są przede wszystkim do układów cyfrowych, natomiast do układów analogowych nie bardzo się nadają. Ich specyficzna geometria powoduje, że nie ma w ich przypadku możliwości swobodnego

kształtowania stosunku szerokości do długości kanału *W/L*. Aby zwielokrotnić szerokość *W*, trzeba połączyć równolegle odpowiednią liczbę tranzystorów. Gdyby zaszła potrzeba zwiększenia długości kanału *L*, trzeba by łączyć tranzystory szeregowo. Natomiast tranzystory FDSOI są uniwersalne, nadają się równie dobrze do układów cyfrowych, jak i analogowych. Co więcej, w ich przypadku istnieje możliwość wykorzystania wyspy, nad którą znajduje się tranzystor, a która jest od niego odizolowana cienką warstwą SiO_2, jako drugiej bramki. Rysunek 5-4 ilustruje tę możliwość.



Rysunek 5-4. Przekrój przez tranzystory w technologii FDSOI pokazujący możliwość wykorzystania wysp jako drugich bramek

Napięcie V1 przyłożone jest do podłoża poprzez wyspę typu *p* i polaryzuje wszystkie wyspy tego typu, natomiast napięcie V2 polaryzuje tylko wyspę typu *n*, do której jest przyłożone. Musi być oczywiście zapewniona zaporowa polaryzacja tej wyspy względem podłoża. Można odizolować od podłoża także wyspę typu *p*, wprowadzając pod nią obszar typu *n* (obszar zwany trzecią wyspą, ang. „triple well"). Musi on być oczywiście spolaryzowany zaporowo wobec obszarów typu *p* poprzez doprowadzenie doń dodatniego napięcia zasilania – rysunek 5-5.

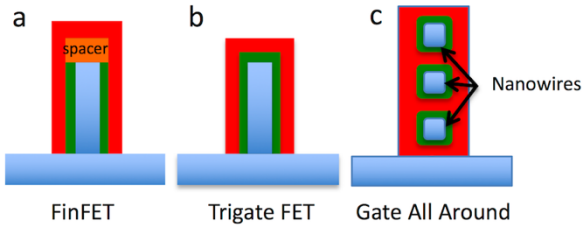


Rysunek 5-5. Tranzystor nMOS w technologii FDSOI nad wyspą izolowaną

Stosując izolowane wyspy typu *p* można indywidualnie doprowadzać napięcia polaryzujące zarówno do wysp typu *p*, jak i do wysp typu *n*. Daje to zupełnie nowe możliwości w układach zarówno cyfrowych, jak i analogowych. Polaryzacja wysp pod tranzystorami może być wykorzystana do sterowania napięciem progowym tranzystorów, co daje w układach cyfrowych możliwość automatycznej zmiany warunków pracy układu: przyłożenie do wysp napięć obniżających napięcia progowe – układ działa szybciej przy większym poborze mocy; przyłożenie napięć podwyższających napięcia progowe – układ działa wolniej, ale oszczędza energię. W układach analogowych wyspy mogą być polaryzowane sygnałami zmiennymi, co stwarza zupełnie nowe możliwości rozwiązań układowych.

Zarówno technologie FinFET, jak i FDSOI są już stosowane na skalę przemysłową. Są to niewątpliwie technologie, które będą stosowane równolegle przez następne lata. W technologii FinFET produkowane są układy z tranzystorami o długości bramki poniżej 10 nm, w FDSOI – obecnie najmniejsze tranzystory mają długość kanału 14 nm. Można się spodziewać, że wprowadzenie fotolitografii EUV pozwoli prędzej czy później produkować układy z tranzystorami o długości kanału 5 nm. Poniżej tej długości raczej nie da się zejść, bowiem na przeszkodzie stoją prawa fizyki. Jeszcze mniejsze tranzystory dałoby się zrobić, ale nie będą one prawidłowo działać, bo pojawi się prąd tunelowy bezpośrednio płynący między źródłem i drenem.

W laboratoriach są natomiast eksperymentalne tranzystory będące wersjami rozwojowymi tranzystorów typu FinFET. Pokazuje je rysunek 5-6.

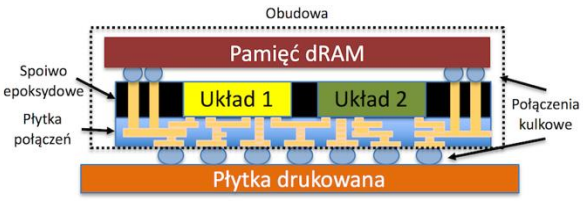


Rysunek 5-6. Rozwój koncepcji tranzystorów FinFET (na rysunku angielskie nazwy struktur). Rysunek pokazuje przekroje pionowe przez tranzystory.

W tych kolejnych wersjach celem jest jak najdokładniejsze otoczenie obszaru kanału bramką, co sprzyja dalszej redukcji prądu podprogowego.

5.1.4 CMOS w pionie, czyli układy trójwymiarowe

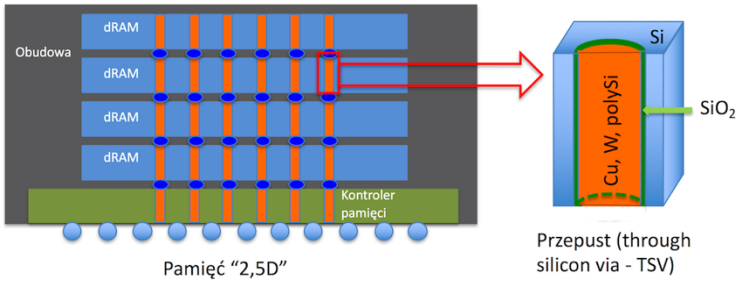
Idea trójwymiarowych układów scalonych jest znana od wielu lat, lecz dopiero ostatnio zaczęła wchodzić do praktyki przemysłowej. Układ trójwymiarowy to nic innego jak stos płytek z układami ułożonych jedna nad drugą i połączonych mechanicznie i elektrycznie w taki czy inny sposób (niektórzy twierdzą, że nie są to „prawdziwe" układy trójwymiarowe i nazywają je układami „2,5D"). Główną zaletą takich układów są bardzo krótkie połączenia. Przykładowo, układy pamięci operacyjnej komputera umieszczone bezpośrednio nad mikroprocesorem w odległości rzędu milimetra to w porównaniu z tradycyjnym sposobem montażu na płycie drukowanej kilkudziesięciokrotne skrócenie połączeń. To oznacza znaczne zmniejszenie pojemności tych połączeń i duże skrócenie czasu propagacji sygnałów między procesorem i pamięcią, a także pewne zmniejszenie poboru mocy, bowiem pojemności pasożytnicze ścieżek połączeń na zewnątrz układu ładują się i rozładowują przez tranzystory w układzie tak samo, jak pojemności wewnątrz układu, powiększając pobór mocy dynamicznej (punkt 4.1.1). W urządzeniach mobilnych ma też znaczenie zmniejszenie liczby odrębnych układów we własnych obudowach, co pozwala zmniejszyć wymiary tych urządzeń.



Rysunek 5-7. Sposób montażu kilku układów we wspólnej obudowie

Rysunek 5-7 pokazuje jeden ze stosowanych w praktyce sposobów budowy układów „2,5D", zwany „system in package" (SiP). Dwa układy cyfrowe umieszczone są obok siebie na pomocniczej płytce, w której wykonane są wielowarstwowe połączenia (płytka ta nosi angielską nazwę „interposer"). Układy są zalane warstwą spoiwa epoksydowego. Nad tymi układami znajduje się układ pamięci dynamicznej. Pamięć połączona jest elektrycznie z płytką pomocniczą przepustami w spoiwie epoksydowym. Widoczne na rysunku połączenia kulkowe są jednym ze sposobów wykonania wyprowadzeń z układu scalonego. Obudowa z takimi wyprowadzeniami nosi angielską nazwę „ball grid array" (BGA).

Układy pamięci dynamicznych także bywają łączone w pionowe stosy – rysunek 5-8. Płytki z układami pamięci dRAM umieszczone są jedna nad drugą i połączone elektrycznie poprzez przepusty w krzemie zwane „through silicon via" (TSV). Są to wytrawione w krzemie na wylot cylindryczne otwory, których ścianki są pokryte dielektrykiem, a wnętrze wypełnione metalem lub polikrzemem. Pod układami pamięci znajduje się układ kontrolera łączący pamięci z zewnętrznym światem. Wszystko razem umieszczone jest w obudowie z tworzywa sztucznego.



Rysunek 5-8. Układ pamięci dynamicznej "2,5D"

Układy trójwymiarowe przy swoich zaletach mają też poważną wadę: odprowadzanie wydzielającego się w nich ciepła jest znacznie utrudnione w porównaniu z układami montowanymi w sposób tradycyjny w zwykłych obudowach.

5.2 A co będzie po klasycznych układach CMOS?

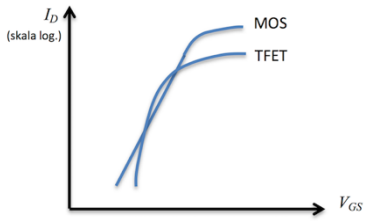
Czy technologię CMOS będzie można zastąpić czymś innym? Innym rodzajem tranzystorów i układów? Materiałami innymi niż krzem? A może w ogóle innym sposobem przetwarzania informacji? Zobaczymy, jakie są pomysły i próby zastąpienia tradycyjnych krzemowych układów CMOS czymś zasadniczo innym.

5.2.1 Tunelowy tranzystor MOS

Tranzystor tunelowy MOS (ang. "tunnel field effect transistor" – TFET) ma strukturę podobną do zwykłego tranzystora MOS, ale źródło i dren są przeciwnego typu przewodnictwa (rysunek 5-9), a między nimi znajduje się półprzewodnik bardzo słabo domieszkowany.

Rysunek 5-9. Przekrój przez tranzystor tunelowy TFET

Tranzystor zostaje włączony, gdy do drenu i bramki doprowadzone są dodatnie napięcia polaryzujące względem źródła. Przepływ prądu następuje w wyniku efektu tunelowego, co jest możliwe, gdy odległość między obszarami źródła i drenu jest bardzo mała. Zaletą tranzystora tunelowego jest bardzo mały prąd w tranzystorze wyłączonym, co w układach oznacza bardzo mały statyczny pobór mocy. Niestety prąd w stanie włączenia jest także bardzo mały, co powoduje, że tranzystor taki jest mało użyteczny. Poglądowe porównanie charakterystyk obu rodzajów tranzystorów pokazuje rysunek 5-10.

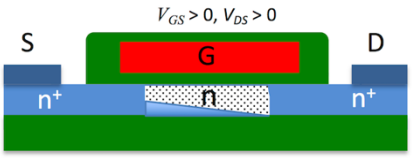


Rysunek 5-10. Poglądowe porównanie charakterystyk $I_D=f(V_{GS})$ tranzystorów MOS i TFET

Tranzystory TFET dotąd nie wyszły poza laboratoria badawcze.

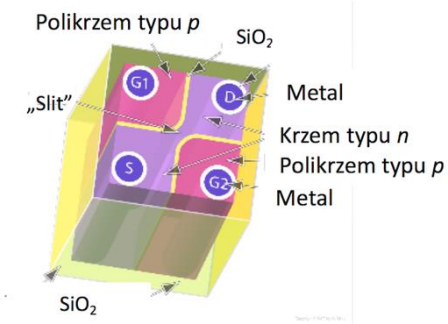
5.2.2 Bezzłączowe tranzystory MOS

Bezzłączowe tranzystory MOS są czymś w rodzaju rezystorów o sterowanej bramką rezystancji. Kanał o bardzo małej grubości jest tego samego typu, co źródło i dren. Napięcie bramki zależnie od wielkości i znaku powoduje przyciąganie nośników ładunku do kanału lub ich wypychanie, co powoduje zmiany rezystancji kanału. Idea tranzystora bezzłączowego pokazana jest na rysunku 5-11.



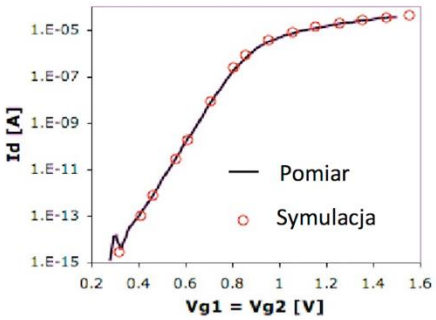
Rysunek 5-13. Idea tranzystora bezzłączowego MOS

Charakterystyki prądowo-napięciowe tranzystorów bezzłączowych są podobne do charakterystyk tradycyjnych tranzystorów MOS, mogą być zarówno z kanałem typu *n*, jak i typu *p*. Można więc z nich budować układy podobne do już nam znanych. Tranzystory bezzłączowe cechuje wyjątkowo korzystny stosunek prądu tranzystora włączonego do prądu tranzystora w stanie wyłączenia – może on sięgać nawet 9 rzędów wielkości, czego nie osiągają inne rodzaje tranzystorów, które omawialiśmy. Oznacza to, że statyczny pobór mocy w układach z takimi tranzystorami będzie wyjątkowo mały. Rysunek 5-12 pokazuje jedną z praktycznie przebadanych struktur tranzystorowych tego typu – ang. „vertical slit field effect transistor” (VeSFET).



Rysunek 5-11. Tranzystor bezzłączowy typu VeSFET (rysunek © W. Mały)

Tranzystor taki ma dwie symetryczne bramki. Jedną z nich można wykorzystać do regulacji napięcia progowego tranzystora (podobnie jak w tranzystorach FDSOI – punkt 5.1.3), możliwe są też inne, jeszcze mało zbadane, ciekawe rozwiązania układowe. Rysunek 5-13 pokazuje zmierzoną charakterystykę podprogową takiego tranzystora.



Rysunek 5-12. Przykładowa charakterystyka podprogowa tranzystora VeSFET (obie bramki zwarte)

Warto dodać, że koncepcja tranzystora VeSFET jest autorstwa Wojciecha Małego, profesora Carnegie Mellon University w Pittsburgu (USA) oraz Politechniki Warszawskiej. Czy tranzystory bezzłączowe wejdą do praktyki przemysłowej? Nie wiadomo. Są pod wieloma względami lepsze od omawianych w punkcie 5.1.3 tranzystorów FinFET i FDSOI, ale nie ma jeszcze przemysłowej technologii tranzystorów bezzłączowych, podczas gdy układy z tranzystorami FinFET i FDSOI są produkowane przez wielu wytwórców.

5.2.3 A może inne materiały?

Wciąż jeszcze w niewielkich ilościach produkowane są układy na arsenku galu (GaAs), ale ich znaczenie maleje. Nie jest to materiał przyszłościowy, a poza tym nie jest przydatny do układów cyfrowych. Bardzo „modny” jest grafen, materiał o niezwykłych właściwościach, między innymi cechuje go bardzo duża ruchliwość elektronów, ale nie jest to półprzewodnik. Proponowano kilka sposobów otrzymania półprzewodzącego grafenu, ale w takim

grafenie ruchliwość nośników ładunku nie jest już wielka. Czy będą układy scalone z grafenu? W obecnej chwili wydaje się to wątpliwe. Podobnie z innymi materiałami, na przykład z nanorurkami węglowymi, czy też z półprzewodnikami wieloskładnikowymi. Gdyby nawet znaleziono materiał konkurencyjny dla krzemu, jego praktyczne zastosowanie będzie wymagało wieloletnich prac technologicznych.

5.2.4 Czy będą komputery kwantowe?

Wciąż wiele się o nich mówi, ale opublikowana ostatnio praca teoretyczna wyklucza możliwość zbudowania takiego komputera o praktycznej przydatności (co nie znaczy, że w mikroelektronice nie wykorzystuje się efektów kwantowych – na przykład w pamięci nieulotnej MRAM wykorzystującej zjawiska związane ze spinem elektronu, część III, punkt 4.2.7).

5.3 Nowe architektury układów i systemów, nowe metody przetwarzania informacji

5.3.1 Uniwersalność kontra specjalizacja

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, jak już wiemy, „system on chip”. Takie układy mogą zawierać bloki będące specjalizowanymi

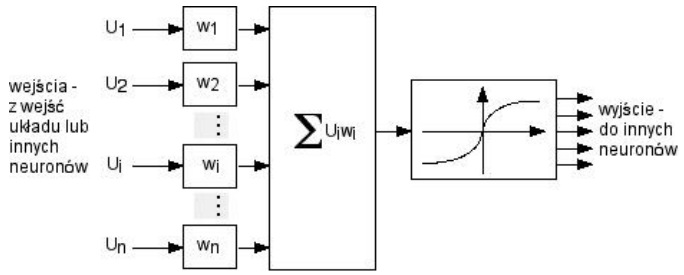
mikroprocesorami przeznaczonymi do wykonywania określonych operacji, potrzebne do wykonywanych funkcji układy peryferyjne i komunikacyjne (w tym – być może – także bloki analogowe), pamięci ROM i RAM oraz wbudowane oprogramowanie (ang. „embedded software"), czyli oprogramowanie zapisane na stałe w wewnętrznej pamięci ROM. Na świecie, a także w Polsce, projektuje się coraz więcej takich układów.

Jednym z kierunków rozwoju układów i systemów mikroelektronicznych jest więc specjalizacja. Przykładem są układy przeznaczone do wykonywania obliczeń równoległych. Już jakiś czas temu zauważono, że procesory graficzne, obsługujące wyświetlanie obrazu na monitorach, doskonale się nadają do takich obliczeń, ponieważ składają się z setek, a nawet tysięcy prostych układów arytmetycznych wykonujących równocześnie niezbyt skomplikowane obliczenia dla poszczególnych pikseli obrazu. Istnieje wiele algorytmów umożliwiających zrównoleglenie obliczeń, na przykład w przypadku operacji na macierzach, gdzie operacje arytmetyczne można wykonywać równocześnie na wszystkich lub znacznej części elementów macierzy. Procesory graficzne nie są układami o uniwersalnym zastosowaniu, bo nie każdy algorytm umożliwia wykonywanie obliczeń równoległych, ale tam, gdzie można je zastosować, ogromnie przyspieszają obliczenia. Inny przykład układów o architekturze wyspecjalizowanej do konkretnych zastosowań to bardzo obecnie rozwijane układy realizujące sprzętowo algorytmy w dziedzinie zwanej sztuczną inteligencją. Z pewnością będzie w przyszłości coraz więcej podobnych przykładów.

5.3.2 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.

Pierwsza z ich to 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 równocześnie 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 rysunku 5-14.



Rysunek 5-14. Elektroniczny model neuronu

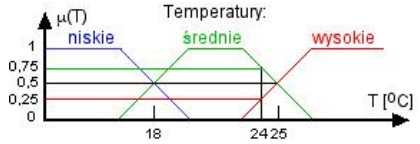
Sygnały z wejść są sumowane z wagami, a następnie suma jest podawana na wyjście przez układ nieliniowy o charakterystyce progowej. Wyjście łączy się z wejściami wielu innych neuronów. Funkcja realizowana przez taką sieć zależy od wartości współczynników wagowych, z jakimi podawane są na układ sumujący 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 rysunku 5-14, 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 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. 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. Rozwój metod sztucznej inteligencji prowadzi między innymi do budowy wspomnianych wyżej specjalizowanych procesorów realizujących wielopoziomowe sieci umożliwiające realizację algorytmów znanych jako „deep learning".

Innym przykładem nietradycyjnych metod przetwarzania informacji są układy realizujące operacje logiki rozmytej. 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 rozwiązując 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 °C, wysokie to wszystkie wyższe od +25 °C, a pozostałe należą do zbioru temperatur średnich. Wtedy temperatura +18 °C jest średnia, podobnie jak +25 °C, temperatura +17,99 °C jest



Rysunek 5-15. Ilustracja pojęcia zbioru rozmytego na przykładzie podziału temperatur na trzy zbiory rozmyte: temperatur niskich, średnich i wysokich. Temperatura 24 °C należy do zbioru temperatur średnich w stopniu 0,75 i do zbioru temperatur wysokich w stopniu 0,25

niska, a +25,01°C – wysoka. W przypadku zbiorów rozmytych możemy zdefiniować *stopień przynależności* temperatur do zbiorów przy użyciu *funkcji przynależności*, na przykład takich, jak na rysunku 5-15. Funkcja przynależności przypisuje każdemu elementowi zbioru wartość stopnia przynależności do tego zbioru.
Przy funkcjach przynależności zdefiniowanych tak, jak na rysunku 5-15, można powiedzieć, że temperatura +18 °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 μ_A, μ_B i μ_C są to funkcje przynależności do tych zbiorów.

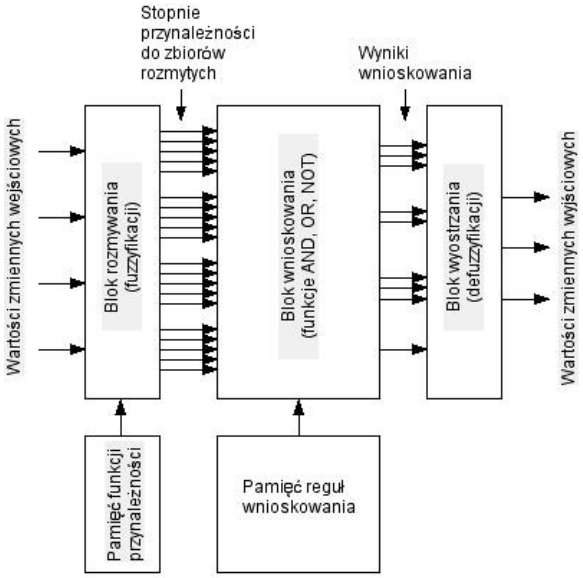
Przypuśćmy, że chcemy opisać pewien algorytm klasyfikacji lub sterowania w terminologii zbiorów rozmytych. Niech będzie to na przykład algorytm sterowania grzejnikiem w pomieszczeniu. Należy zdefiniować zbiory rozmyte dla temperatury wewnątrz i na zewnątrz pomieszczenia (są to zmienne wejściowe) oraz 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 strukturę pokazaną na rysunku 5-16.



Rysunek 5-16. 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 zadaniach 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.

5.4 Dla starych technologii nadal jest miejsce

Rozwój technologii mikroelektronicznych, o którym była mowa wyżej, był i jest nadal dyktowany przez ambicje i rynkową walkę kilku największych producentów, którzy skupiają się na ogromnym rynku układów do komputerów, tabletów, smartfonów. Ale to nie znaczy, że innych zastosowań i innych rynków nie ma lub nie mają one znaczenia. Jest przecież elektronika motoryzacyjna, jest elektronika medyczna, rozwija się „Internet rzeczy" (ang. „Internet of Things", IoT), mikroelektronika jest obecna w sprzęcie AGD, we wszelkiego rodzaju kartach identyfikacyjnych i płatniczych, w zabawkach dla dzieci i wielu innych zastosowaniach. We wszystkich tych zastosowaniach super-wydajne systemy cyfrowe nie są wcale potrzebne (wyjątkiem w elektronice motoryzacyjnej są systemy sterowania w samochodach autonomicznych). W zależności od zastosowań

najważniejsze mogą być takie cechy, jak niezawodność, mały pobór mocy, niska cena, cyberbezpieczeństwo. Dlatego starsze technologie wcale nie wymierają. Na świecie jest wielu producentów, którzy z powodzeniem i ekonomicznymi sukcesami produkują układy w technologiach CMOS z kanałem o długościach powyżej 100 nanometrów: 0,35 μm, 0,5 μm, 0,7 μm i więcej. Technologie te zostały dawno opracowane i wdrożone, linie produkcyjne się zamortyzowały, procesy produkcyjne i komplety masek do fotolitografii nie kosztują wiele, toteż układy produkowane w tych technologiach są tanie. Z tych właśnie technologii często korzystają projektanci układów specjalizowanych (ASIC), bo w większości przypadków te układy nie potrzebują bardziej zaawansowanych technologii. Co więcej, z technicznego punktu widzenia dla niektórych klas układów te „stare" technologie są lepsze od najnowocześniejszych. Przykładowo, w technice układów analogowych jest wiele bardzo dobrych układowych rozwiązań, które jednak wymagają napięć zasilania rzędu 5V, niedostępnych w technologiach najnowszych. Dlatego technologie starsze nie tylko nie zanikają, ale przeciwnie – nadal się w nie inwestuje. W ostatnich latach powstało na świecie kilkanaście nowych linii produkcyjnych, a nawet całych fabryk przeznaczonych dla starszych technologii. Jednym z kierunków ich rozwoju jest specjalizacja. Powstają linie technologiczne, na których można wytwarzać układy scalone wraz z elementami optoelektronicznymi, układy do nietypowych zastosowań (na przykład wysokonapięciowe), układy zintegrowane z czujnikami i mikromechanizmami itp.

6 Podsumowanie

W podsumowaniu warto najpierw zwrócić uwagę na to, o czym w tych materiałach nie było mowy. A potem zachęcić do dalszych studiów. Mikroelektronika jest fascynującą dziedziną techniki – powie to każdy, kto się nią zajmuje – a przy tym dziedziną, która ukształtowała i nadal przekształca całą naszą współczesną cywilizację.

6.1 O czym nie było mowy

Skromna objętość tych materiałów 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.

6.1.1 Układy RF (ang. „Radio frequency") I mikofalowe

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 to dziedzina mikroelektroniki obecnie bardzo szybko rozwijająca się ze względu na liczne zastosowania.

Obok elementów czynnych oraz elementów biernych R i C w układach RF i mikrofalowych stosowane są też scalone indukcyjności o wartościach rzędu pojedynczych nanohenrów. Nie omawialiśmy ich w tym wykładzie.

6.1.2 Układy SC (ang. „Switched capacitors")

Jest to specyficzna grupa układów analogowych zwana w polskiej nomenklaturze 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 zwykle 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.

6.1.3 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.

6.1.4 Układy dużej mocy I wysokonapięciowe

Wprawdzie w punkcie 4.2 omawiane 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.

6.1.5 Układy wykonujące różne bardziej złożone funkcje analogowe

W punktach poświęconych układom analogowym przedstawione zostały jedynie niektóre podstawowe bloki pomocnicze i funkcjonalne pozwalające zorientować się ogólnie w problematyce układów analogowych. Istnieje wielka liczba układów wykonujących różne funkcje bardziej złożone, ale z konieczności trzeba je było pominąć.

6.1.6 Automatyczna synteza układów i systemów 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.

6.2 Zakończenie

Warto na zakończenie jeszcze raz podkreślić, o czym była już mowa w części I, że wszystkie technologie mikroelektroniczne, i te starsze, i te najnowocześniejsze, są dostępne dla polskiego inżyniera. A więc droga Czytelniczko, drogi Czytelniku, jeśli chcesz projektować nowoczesne i ambitne układy specjalizowane – nic nie stoi na przeszkodzie. Wystarczy kontynuować studia w tym kierunku, do czego wstępem są niniejsze materiały. 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. A potem – genialne pomysły, świetne projekty i wielkie sukcesy, czego Wam życzy autor tych materiałów.