2018

Układy scalone

CZĘŚĆ III: UKŁADY CYFROWE

WIESŁAW KUŹMICZ

UKŁAD SCALONY, MIKROPROCESOR, UKŁAD CYFROWY, TRANZYSTOR MOS, BRAMKA CYFROWA, BRAMKA STATYCZNA, BRAMKA DYNAMICZNA, PRZERZUTNIK, REJSTR, PAMIĘĆ PÓŁPRZEWODNIKOWA, TESTOWANIE

CZĘŚĆ TRZECIA MATERIAŁÓW OMAWIAJĄCYCH UKŁADY SCALONE DOTYCZY UKŁADÓW I SYSTEMÓW CYFROWYCH. PRZEDSTAWIONE SĄ RÓŻNE RODZAJE BRAMEK LOGICZNYCH, OMÓWIONE WYMAGANIA DLA TYCH BRAMEK I ICH PARAMETRY, PRZEDSTAWIONE SĄ W ZARYSIE ZASADY DZIAŁANIA PAMIĘCI PÓŁPRZEWODNIKOWYCH I ICH ORGANIZACJA, OMAWIANE SĄ PROBLEMY TESTOWANIA UKŁADÓW CYFROWYCH I SPOSOBY POKONYWANIA TYCH PROBLEMÓW.

Spis treści

1 Wstęp: o czym tu będzie mowa

Droga Czytelniczko, Drogi Czytelniku: teraz będzie mowa o cyfrowych układach scalonych. W części pierwszej przedstawione były sposoby budowy z tranzystorów MOS prostych bramek kombinacyjnych CMOS. Tu omówimy je bardziej szczegółowo, pokazując jak i od czego zależą ich parametry takie, jak czasy propagacji sygnału czy też pobór mocy. Wspomnimy też inne rodzaje bramek: bramki zwane transmisyjnymi, bramki dynamiczne, bramki trójstanowe. Kolejnym tematem będą przerzutniki, dzięki którym można budować układy sekwencyjne (nie pamiętasz co to? – sprawdź w części I) oraz komórki pamięci i budowa całych pamięci. Omówimy też problemy i metody testowania układów cyfrowych. Zobaczymy, że testowanie dużych układów cyfrowych jest trudne – a przecież układ, którego poprawności działania nie można sprawdzić, jest bezużyteczny. W ćwiczeniu praktycznym zobaczymy, jak można zbadać działanie prostego układu cyfrowego przy użyciu symulatora układów logicznych.

2 Statyczne bramki kombinacyjne CMOS

2.1 Podstawy

Omówimy na początku podstawowe wymagania, jakie powinny spełniać bramki logiczne, a także główne parametry charakteryzujące te bramki. Będziemy mówić o bramkach kombinacyjnych. Bramki realizujące funkcje pamięciowe (przerzutniki, rejestry, komórki pamięci) będą omawiane dalej.

Będziemy się zajmować na razie bramkami statycznymi. Takie właśnie bramki są stosowane w zdecydowanej większości układów cyfrowych CMOS do budowy bloków logiki kombinacyjnej.

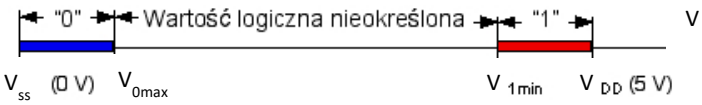
> **DEFINICJA**
>
> *Bramka statyczna jest to bramka mająca tę własność, że jak długo włączone jest napięcie zasilania, a stany logiczne na wejściach nie zmieniają się, to i stany logiczne na wyjściach nie zmieniają się.*

Najważniejszymi cechami i właściwościami bramek cyfrowych są:

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

2.1.1 Poziomy logiczne i ich regeneracja

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



Rysunek 2-1. Definicja poziomów logicznych (przykład dla napięcia zasilania 5V)

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

DEFINICJA

Mówimy, że bramka logiczna regeneruje poziom logiczny „0", jeżeli dla dowolnej kombinacji stanów logicznych na wejściach reprezentowanych przez napięcia na granicach odpowiednich zakresów (tj. V_{0max} dla stanów „0" i V_{1min} dla stanów „1"), dla których na wyjściu bramki mamy stan „0", stan ten jest reprezentowany przez napięcie wewnątrz zakresu zera, tj. napięcie V spełniające warunek $V<V_{0max}$.

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

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

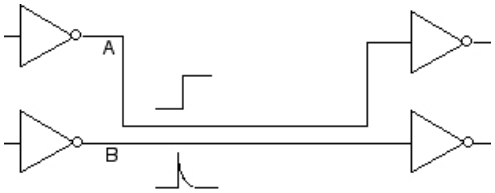
Warto dodać, że istnieją bramki nie mające zdolności regeneracji poziomów logicznych, a nawet takie, które powodują degradację tych poziomów, tj. dają na wyjściu napięcia spoza dopuszczalnego zakresu. Jednak w każdym złożonym układzie przynajmniej część bramek musi mieć zdolność regenerowania poziomów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

2.1.2 Zakłócenia i ich tłumienie

Drugim ważnym wymaganiem jest zdolność bramek do tłumienia zakłóceń. Skąd się biorą zakłócenia? Zakłócenia w układach cyfrowych mają zwykle charakter impulsów nakładających się na prawidłowy poziom napięcia reprezentujący „0" lub „1". Impulsy takie mogą pochodzić z kilku źródeł:

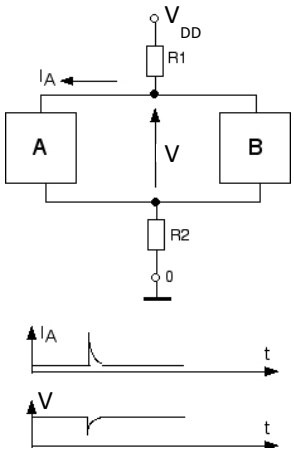
- pasożytnicze sprzężenia pomiędzy połączeniami (pojemnościowe, indukcyjne),
- sprzężenia poprzez wspólne zasilanie,
- zewnętrzne pola elektromagnetyczne,
- promieniowanie jonizujące.

Wpływ sprzężenia między połączeniami ilustruje symbolicznie rysunek 2-2. Jeśli na jednej z dwóch położonych blisko siebie ścieżek pojawia się skokowa zmiana stanu logicznego, to na drugiej pojawia się krótki impuls wynikający z istnienia pojemności pomiędzy tymi ścieżkami (indukcyjność wzajemna też istnieje, ale ma zwykle drugorzędne znaczenie).



Rysunek 2-3. Zakłócenie wywołane sprzężeniem pojemnościowym: zmiana napięcia na ścieżce A indukuje impuls w ścieżce B

Sprzężenie poprzez wspólne zasilanie jest wywołane skończoną, niezerową rezystancją połączeń między blokami logicznymi, a ich zasilaniem. Jak zobaczymy dalej, bramki CMOS pobierają prąd w postaci krótkich, stromych impulsów. Rysunek 2-3 pokazuje, jak pobór prądu o takim charakterze wywołuje powstawanie i przenikanie zakłóceń impulsowych. Impuls prądu I_A pobieranego przez blok logiczny A związany jest ze spadkiem napięcia na rezystancjach R1 i R2, i co za tym idzie wywołuje chwilowe zmniejszenie napięcia zasilającego V. Ten spadek napięcia o charakterze krótkiego impulsu jest widoczny nie tylko dla bloku A, ale i dla bloku B, ponieważ ich napięcie zasilania V jest wspólne.



Rysunek 2-2. Zakłócenie wywołane chwilowym spadkiem napięcia zasilania spowodowanym impulsowym poborem prądu

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

Promieniowanie jonizujące wywołuje w półprzewodniku generację par elektron-dziura, co powoduje przepływ impulsów zwiększonego prądu wstecznego w spolaryzowanych zaporowo złączach *p*-*n*, np. złączach źródeł i drenów tranzystorów. Może to zakłócać przede wszystkim działanie układów zwanych dynamicznymi, w których informacja jest reprezentowana przez ładunek zgromadzony w pojemności (o takich układach będzie mowa dalej).

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

DEFINICJA

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

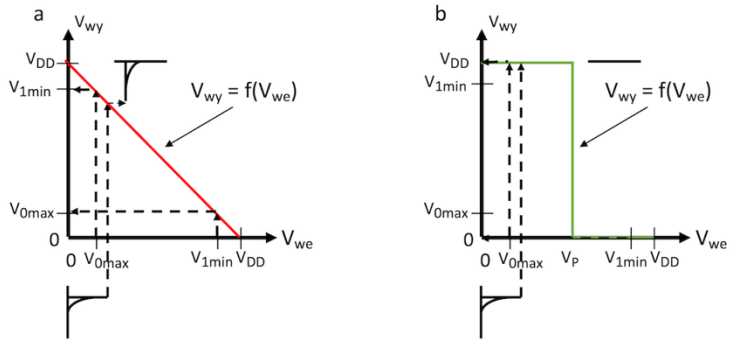
2.1.3 Charakterystyka przejściowa bramki

DEFINICJA

Charakterystyką przejściową bramki nazywamy zależność jej napięcia wyjściowego od wejściowego.

Zastanówmy się teraz, jak pogodzić regenerację poziomów logicznych z tłumieniem zakłóceń. Nietrudno pokazać, że te wymagania są do pewnego stopnia sprzeczne. Pokażemy to na hipotetycznym przykładzie „liniowego inwertera". Przypomnijmy (część I): inwerter jest to najprostsza bramka logiczna wykonująca funkcję negacji NOT: „0" na wejściu daje „1" na wyjściu, i odwrotnie. Taką funkcję mógłby na przykład wykonywać układ elektroniczny o liniowej charakterystyce przejściowej $V_{wy} = V_{DD} - V_{we}$ pokazanej na rysunku 2-4a. Jednak taki układ ani nie regeneruje poziomów logicznych (V_{1min} na wejściu daje V_{0max} na wyjściu, i odwrotnie), ani nie tłumi zakłóceń (impuls zakłócający na tle zera logicznego na wejściu pojawia się na tle jedynki logicznej na wyjściu z taką samą amplitudą). Wniosek z tego jest taki, że bramka logiczna mająca równocześnie zdolność regeneracji poziomów logicznych i tłumienia zakłóceń nie może mieć liniowej charakterystyki przejściowej.

Charakterystyka przejściowa „inwertera idealnego", łączącego zdolność do regeneracji poziomów logicznych i do tłumienia zakłóceń, pokazana jest na rysunku 2-4b. Widać, że V_{1min} na wejściu daje 0 na wyjściu, a V_{0max} na wejściu daje V_{DD} na wyjściu. Widać także, że impuls zakłócający w ogóle nie pojawia się na wyjściu, ale pod warunkiem, że jego amplituda nie przekracza napięcia przełączania inwertera V_P.

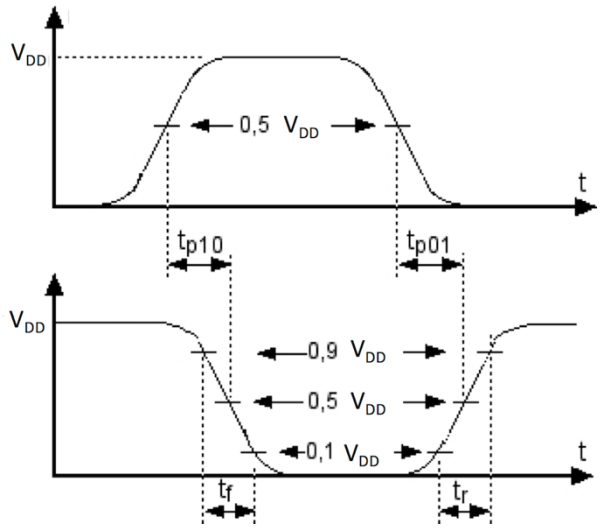


Rysunek 2-4. Inwerter o liniowej charakterystyce przejściowej (a) i o idealnej charakterystyce przejściowej (b)

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

2.1.4 Czas propagacji sygnału

Kolejnym wymaganiem dla bramek jest możliwie krótki czas propagacji sygnału. Od niego w dużym stopniu zależy, jak szybko będą działały układy cyfrowe z tymi bramkami. Po zmianie stanów na wejściu zmiana stanów na wyjściu bramki cyfrowej nie następuje natychmiast. Miarami szybkości działania bramki są czasy propagacji sygnału t_{p10} i t_{p01}. Istotne są też czasy narastania t_r i opadania t_f sygnału na wyjściu. Są one zdefiniowane na rysunku 2-5 dla najprostszego przypadku inwertera. Czasy propagacji są zdefiniowane w odniesieniu do punktów na osi czasu, w których napięcie ma wartość połowy poziomu logicznego jedynki. Czasy narastania i opadania są zdefiniowane w odniesieniu do punktów na osi czasu, w których napięcie osiąga wartość 0,1 i 0,9 poziomu logicznego jedynki.



Rysunek 2-5. Czasy propagacji sygnału, narastania i opadania na przykładzie przebiegów na wejściu i wyjściu inwertera

Często definiuje się dla bramki jeden czas propagacji sygnału t_p, który jest średnią arytmetyczną czasów t_{p10} i t_{p01}.

2.1.5 Pobór mocy

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

2.1.6 Kierunkowość

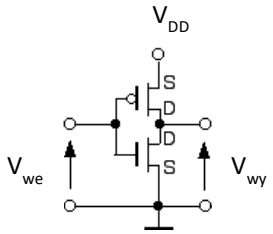
Jeszcze jedną cechą bramek cyfrowych, która ma w niektórych przypadkach znaczenie, jest kierunkowość. Kierunkowość jest właściwością bramek polegającą na tym, że sygnały propagują się tylko w jedną stronę - od wejść do wyjść. Mówiąc precyzyjniej, stany wyjść określane są przez stany wejść, ale stany wejść nie są uzależnione od stanów wyjść. Jak zobaczymy, nie wszystkie bramki mają tę właściwość.

Dodajmy jako ciekawostkę, że istnieją też bramki zwane odwracalnymi. Są to bramki, w których stany na wejściu decydują o stanach na wyjściu, ale i na odwrót – stany logiczne na wyjściach jednoznacznie określają stany wejść. Postulowano na gruncie termodynamiki, że teoretycznie takie bramki mogłyby działać bez strat mocy. W praktyce tak się nie dzieje, bo tranzystory nie są idealnymi wyłącznikami – mają różną od zera rezystancję w stanie włączenia i różną od nieskończoności rezystancję w stanie wyłączenia. Bramki odwracalne nie znalazły więc praktycznego zastosowania. Są jednak nadal przedmiotem badań, bo ich teoria ma między innymi związek z tak zwanymi komputerami kwantowymi. Tego ciekawego tematu nie będziemy tu jednak rozwijać.

2.2 Zależności ilościowe w bramkach NOT, NOR i NAND

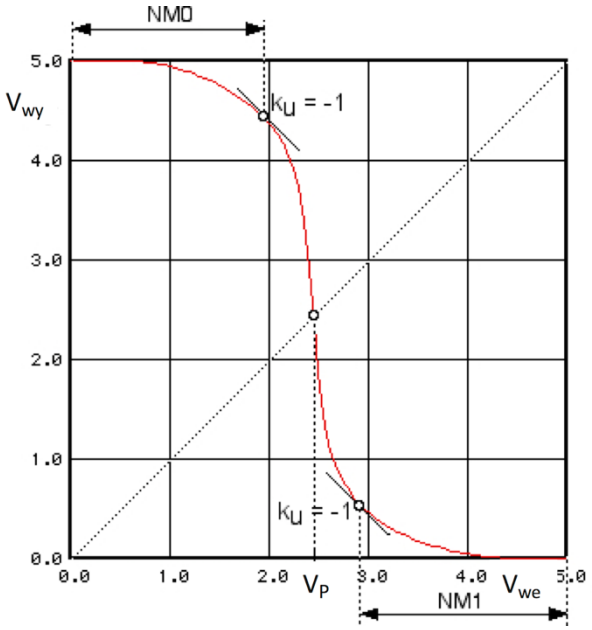
2.2.1 Charakterystyka przejściowa inwertera

Schematy i zasady działania bramek NOT, NOR i NAND były przedstawione w części I, punkt 3.2.2. Przypomnimy dla wygody ich schematy. Zacznijmy od omawiania bramki NOT zwanej, jak wiemy, inwerterem – rysunek 2-6.



Rysunek 2-6. Inwerter CMOS

Charakterystyka przejściowa inwertera, czyli zależność napięcia na wyjściu od napięcia na wejściu, nie da się określić w postaci jednego prostego wzoru. Wynika to z tego, że gdy napięcie na wejściu zmienia się w zakresie między 0 a V_{DD}, to każdy z tranzystorów przechodzi przez wszystkie trzy stany: podprogowy, liniowy i nasycenia (porównaj część I, punkt 3.1.5), a w każdym z tych stanów prąd drenu tranzystora jest opisany inną zależnością od napięć. Charakterystykę przejściową można łatwo otrzymać przy użyciu symulatora układów elektronicznych. Taka przykładowa charakterystyka jest pokazana na rysunku 2-7.



Rysunek 2-7. Charakterystyka przejściowa inwertera CMOS, z zaznaczonym napięciem przełączania Vp i marginesami zakłóceń

Na charakterystyce przejściowej inwertera zaznaczono trzy charakterystyczne punkty. Dwa z nich są to punkty, w których wzmocnienie napięciowe k_u (tj. pochodna dV_{wy}/dV_{we}; więcej o pojęciu wzmocnienia napięciowego znajdziesz w części IV) ma wartość bezwzględną równą 1. Te punkty wyznaczają maksymalne wartości amplitudy zakłóceń, dla których zakłócenia są tłumione. Odpowiednie wartości amplitud są oznaczone symbolami NM0 – od strony „0”, i NM1 – od strony „1” (symbol NM pochodzi od angielskiego terminu „noise margin”). Wartości NM0 i NM1 nazywamy marginesami zakłóceń. Trzeci punkt to umowne napięcie przełączania inwertera V_P. Jest to punkt przecięcia charakterystyki z prostą o równaniu $V_{wy} = V_{we}$.

Ze wszystkich znanych sposobów realizacji inwertera inwerter CMOS ma charakterystykę przejściową najbardziej zbliżoną do idealnej. Przy odpowiednim doborze wymiarów kanałów tranzystorów można uzyskać

charakterystykę symetryczną o napięciu przełączania równym połowie napięcia zasilania. Taka charakterystyka oznacza także maksymalną odporność na zakłócenia - oba marginesy zakłóceń są wówczas jednakowe. W obszarze przejściowym charakterystyka jest bardzo stroma, co oznacza, że punkty na charakterystyce wyznaczające marginesy zakłóceń leżą niezbyt daleko od punktu wyznaczającego napięcie przełączania.

W obszarze przejściowym oba tranzystory pracują w zakresie nasycenia. Wartość napięcia przełączania można więc oszacować przyrównując oba prądy drenu wyznaczane z zależności dla zakresu nasycenia (część I, wzór 3-5):

Równanie 2-1

$$K_n(V_P - V_{Tn})^2 = K_p\left(V_{DD} - V_P - \left|V_{Tp}\right|\right)^2$$

gdzie oznaczono

Równanie 2-2

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

W tych wzorach oznaczenia jak w części I: V_T jest napięciem progowym tranzystora, W jest szerokością kanału, L – jego długością, μ jest ruchliwością nośników ładunku (elektronów w tranzystorze nMOS, dziur w tranzystorze pMOS) w kanale, C_{ox} – pojemnością dielektryku bramkowego na jednostkę powierzchni. K będzie nazywany współczynnikiem przewodności tranzystora, indeks „n" oznacza wielkość odnoszącą się do tranzystora nMOS, zaś indeks „p" oznacza wielkość odnoszącą się do tranzystora pMOS. Rozwiązując równanie 2-1 otrzymujemy:

Równanie 2-3

$$V_p = \frac{V_{Tn} + \sqrt{r}(V_{DD} - |V_{Tp}|)}{1 + \sqrt{r}}$$

gdzie

Równanie 2-4

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

Z równania 2-3 można wyznaczyć wartość r dla wymaganej wartości napięcia przełączania V_p:

Równanie 2-5

$$r = \left[\frac{V_p - V_{Tn}}{(V_{DD} - |V_{Tp}|) - V_p}\right]^2$$

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

Równanie 2-6

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

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

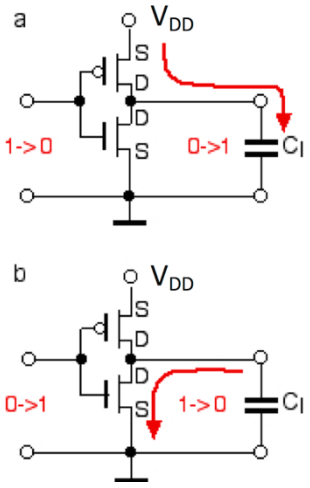
Naturalna ruchliwość elektronów μ_n jest 2 ... 2,5 raza większa od ruchliwości dziur μ_p, więc dla uzyskania symetrii charakterystyki przejściowej kanał tranzystora pMOS powinien być 2 ... 2,5 raza szerszy od kanału tranzystora nMOS. Reguła ta obowiązuje jednak tylko dla starszych wersji technologii CMOS. W bardziej zaawansowanych technologiach stosowane są dodatkowe zabiegi technologiczne. Ich celem jest zwiększenie ruchliwości nośników obu rodzajów. Wówczas stosunek ruchliwości, o którym mowa wyżej, nie obowiązuje. Często jest tak, że obie ruchliwości są sobie bliskie. Toteż jeżeli zależy nam na dokładnej symetrii charakterystyki przejściowej, to dobranie stosunku szerokości kanałow tranzystorów wymaga wykonania kilku symulacji elektrycznych z zastosowaniem dokładnych modeli tranzystorów, bowiem zależności, z których korzystamy, są przybliżone, a ponadto warunek równości napięć progowych $V_{Tn} = |V_{Tp}|$ nie zawsze jest spełniony.

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

2.2.2 Czasy przełączania inwertera

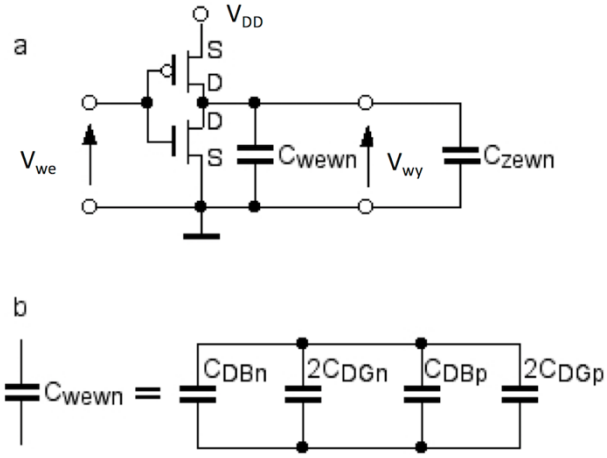
Zajmiemy się teraz czasami przełączania. Terminem „czasy przełączania” będziemy ogólnie określać czasy narastania i opadania sygnału na wyjściu oraz czasy propagacji sygnału. Czasy propagacji sygnału decydują o szybkości działania bramki w układzie. Czasy narastania i opadania sygnału sterującego bramkę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

Zarówno czasy propagacji, jak i czasy narastania i opadania są uwarunkowane szybkością ładowania lub rozładowywania pojemności, jaką obciążony jest inwerter - patrz rysunek 2-8. Gdy stan na wyjściu zmienia się z "0" na "1", pojemność obciążająca C_l ładuje się w wyniku przepływu prądu ze źródła zasilania przez tranzystor pMOS. Gdy stan na wyjściu zmienia się z "1" na "0", pojemność obciążająca C_l rozładowuje się w wyniku przepływu prądu przez tranzystor nMOS.



Rysunek 2-8. Ładowanie (a) i rozładowywanie (b) pojemności obciążającej C_l przy przełączaniu inwertera

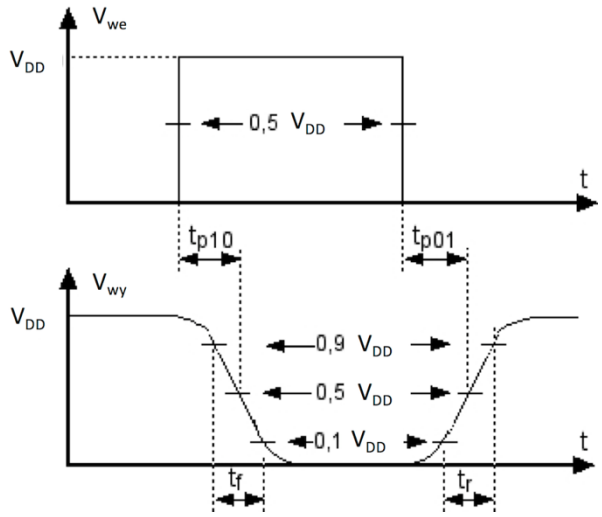
Na pojemność obciążającą C_l składają się wewnętrzne pojemności samego inwertera oraz pojemności zewnętrzne w stosunku do niego, takie jak suma pojemności wejściowych innych bramek obciążających inwerter i połączeń prowadzących do tych bramek – rysunek 2-9.



Rysunek 2-9. Inwerter obciążony pojemnością wewnętrzną i zewnętrzną (a) oraz składniki pojemności wewnętrznej (b)

Pojemności wewnętrzne obciążające inwerter to suma pojemności złączowych drenów tranzystorów nMOS i pMOS oraz suma podwojonych pojemności dren-bramka tych tranzystorów (przypomnij sobie punkt 3.1.5 w części I). Dlaczego podwojonych? Otóż pojemności te są w rzeczywistości włączone między węzeł wyjściowy, a wejściowy. W procesie zmiany stanów logicznych napięcie na wejściu zmienia się od zera do V_{DD} (lub odwrotnie), a napięcie na wyjściu zmienia się w przeciwnym kierunku: od V_{DD} do zera (lub odwrotnie). W rezultacie napięcie na pojemnościach włączonych między wyjściem, a wejściem zmienia się o $2V_{DD}$. Pojemności te w schemacie z rysunku 2-9 są przeniesione na wyjście, gdzie przy zmianie stanów logicznych napięcie zmienia się tylko o wartość równą V_{DD}. Zatem aby pojemności po przeniesieniu na wyjście gromadziły taki sam ładunek, jak w miejscu, w którym rzeczywiście występują, ich wartości trzeba podwoić. Zabieg polegający na przeniesieniu tych pojemności na wyjście bardzo ułatwia oszacowanie czasów przełączania.

Załóżymy teraz dla uproszczenia, że inwerter CMOS jest sterowany sygnałem o kształcie idealnego impulsu prostokątnego, tj. o czasach narastania i opadania równych zeru. Wówczas czas propagacji liczyć należy od chwili, w której nastąpiła skokowa zmiana napięcia na wejściu do chwili, w której napięcie wyjściowe osiągnęło (malejąc lub rosnąc, zależnie od kierunku zmiany) wartość równą $0{,}5V_{DD}$ (patrz rysunek 2-10, który jest zmodyfikowaną wersją rysunku 2-5).



Rysunek 2-10. Czasy propagacji przy wyidealizowanym sygnale wejściowym

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

Równanie 2-7

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

Równanie 2-8

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

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

Z punktu widzenia szybkości działania układu logicznego korzystne jest na ogół, by czasy propagacji t_{p10} i t_{p01} miały zbliżone wartości. Ze wzorów 2-7 i 2-8 widać, że jednakowe czasy propagacji t_{p10} i t_{p01} osiąga się przy jednakowych wartościach napięć progowych i jednakowych wartościach współczynników K_n i K_p. Są to te same warunki, które zapewniają symetryczną charakterystykę przejściową inwertera.

2.2.3 Pobór mocy

Na koniec zajmiemy się oszacowaniem poboru mocy. Jest to bardzo ważny problem, w dzisiejszym stanie technologii CMOS nie można już dalej zwiększać szybkości układów CMOS, bo na przeszkodzie stoi wzrost poboru mocy. Dlatego temu problemowi poświęcona jest osobny punkt w części IV. Na razie tylko podstawowe informacje i szacunkowe wzory.

Prąd, jaki pobiera ze źródła zasilania statyczny inwerter CMOS, ma dwie składowe: statyczną i dynamiczną. Składowa statyczna to prąd, jaki płynie w stanie ustalonym, gdy stany logiczne nie zmieniają się. Prąd ten ma małą wartość, bowiem zarówno w stanie „0” na wejściu, jak i w stanie „1” jeden z połączonych szeregowo tranzystorów - nMOS lub pMOS - jest wyłączony, nie przewodzi. Statyczny prąd ma kilka składników, z których najistotniejszy jest zwykle prąd podprogowy tego z tranzystorów MOS, który jest w danej chwili wyłączony (o prądzie podprogowym była mowa w części I, punkt 3.1.5). Jeżeli sumę wszystkich prądów składających się na prąd statyczny nazwiemy prądem statycznego upływu I_{stat}, to moc statyczna P_{stat} pobierana przez inwerter wynosi

Równanie 2-9

$$P_{stat} = I_{stat} V_{DD}$$

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

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

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

Równanie 2-10

$$P_C = C_l V_{DD}^2 f$$

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

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

Równanie 2-11

$$P_j = I_{max} V_{DD} \frac{t_r + t_f}{2} f$$

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

Łączny pobór mocy jest sumą mocy określonych wzorami 2-9, 2-10 i 2-11, przy czym zazwyczaj dominuje moc związana z ładowaniem-rozładowywaniem pojemności P_C.

2.2.4 Kryteria wyboru wymiarów tranzystorów

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

- maksymalna szybkość działania,
- minimalny pobór mocy,
- minimalna powierzchnia.

Dla uzyskania maksymalnej szybkości działania, czyli możliwie krótkich czasów propagacji, trzeba przede wszystkim ustalić, jaki charakter ma pojemność obciążająca. Dwa skrajne przypadki to: (a) dominująca pojemność zewnętrzna $C_{zewn} \gg C_{wewn}$, i (b) dominująca pojemność wewnętrzna $C_{zewn} \ll C_{wewn}$. Jeśli

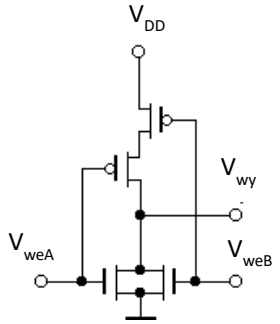
dominuje pojemność zewnętrzna, której wartość nie zależy od wymiarów tranzystorów w inwerterze, to im większa szerokość kanałów tranzystorów, tym krótsze czasy propagacji (patrz wzory 2-7 i 2-8). Jeśli natomiast dominuje pojemność wewnętrzna, to poszerzanie kanałów nie jest celowe. Ze wzrostem szerokości kanałów tranzystorów proporcjonalnie rosną powierzchnie złącz drenów, czyli pojemności C_{DB}, a także powierzchnie bramek i zakładek bramek nad drenami, czyli w sumie pojemności C_{DG}. Ze wzrostem szerokości kanałów rośnie więc proporcjonalnie pojemność C_l, a czasy propagacji pozostają bez zmiany. Większe szerokości kanałów są w tej sytuacji wręcz szkodliwe, bo wzrasta niepotrzebnie prąd ładowania i rozładowywania pojemności, a więc pobór mocy. Zarówno w przypadku (a), jak i w przypadku (b), zachowywany jest zwykle stosunek W_p/W_n zapewniający symetryczną charakterystykę przejściową dla zapewnienia maksymalnej odporności na zakłócenia i jednakowych czasów propagacji t_{p10} i t_{p01}.

Minimalny pobór mocy układu uzyskuje się przy minimalnej sumie wszystkich pojemności obciążających bramki, zgodnie ze wzorem 2-10. Reguła jest więc prosta: wszystkie wymiary tranzystorów (bramek, ale także obszarów źródeł i drenów) powinny być jak najmniejsze. Możliwe są tu dwa przypadki. W pierwszym przypadku minimalną długość i szerokość kanału mają tranzystory nMOS, natomiast tranzystory pMOS mają kanały poszerzone dla zapewnienia maksymalnej odporności na zakłócenia. W drugim przypadku również tranzystory pMOS mają oba wymiary minimalne. Charakterystyki przejściowe nie są wówczas symetryczne. Napięcie przełączania jest mniejsze od $0{,}5V_{DD}$, a margines zakłóceń od strony zera logicznego ulega zmniejszeniu. W wielu przypadkach margines ten pozostaje jednak wystarczająco duży, zwłaszcza w przypadku układów zasilanych napięciem 5V.

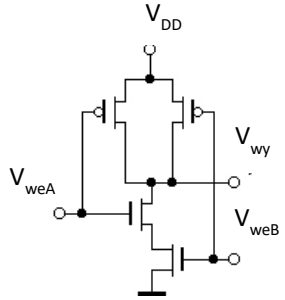
Minimalną powierzchnię układu również uzyskamy zmniejszając do wartości minimalnych dopuszczalnych wymiary kanałów tranzystorów (a także obszarów źródeł i drenów). Jest to więc przypadek już omówiony. Warto jednak w tym miejscu dodać, że we współczesnych technologiach CMOS o powierzchni układu cyfrowego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

2.2.5 Charakterystyki przejściowe bramek NOR i NAND

Nasza znajomość parametrów i charakterystyk inwertera da się łatwo uogólnić na bramki wielowejściowe NAND i NOR. Dla wygody przypomnimy ich schematy, które były omówione w części I, punkt 3.2.2.



Rysunek 2-12. Dwuwejściowa bramka NOR



Rysunek 2-11. Dwuwejściowa bramka NAND

Rozważmy na początek charakterystyki przejściowe i napięcie przełączania bramki NOR. Rozróżnić trzeba dwa przypadki: przełączania równoczesnego obu wejść i przełączania tylko jednego wejścia.

Rozważymy najpierw przełączanie równoczesne obu wejść. Gdy na oba wejścia doprowadzone są sygnały identycznie zmieniające się w czasie, wejścia te można potraktować jako zwarte. Bramka NOR staje się wówczas równoważna inwerterowi, w którym rolę tranzystora nMOS pełnią dwa tranzystory połączone równolegle, a rolę tranzystora pMOS - dwa tranzystory połączone szeregowo. Do obliczenia napięcia przełączania można użyć zależności 2-3, w której wartość r daną wzorem 2-4 zastąpimy zmodyfikowaną wartością r':

Równanie 2-12

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

Uogólniając ten wynik dla bramki NOR o N wejściach otrzymujemy

Równanie 2-13

$$V_p = \frac{V_{Tn}+\frac{1}{N}\sqrt{r}(V_{DD}-|V_{Tp}|)}{1+\frac{1}{N}\sqrt{r}}$$

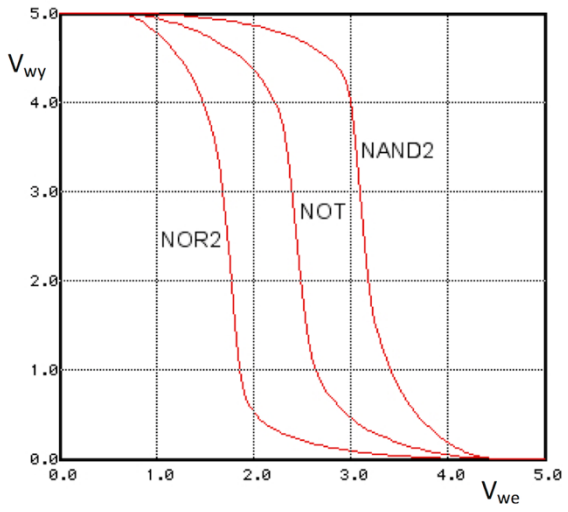
Analogicznie rozumując dla bramki NAND o N wejściach otrzymujemy

Równanie 2-14

$$V_p = \frac{V_{Tn}+N\sqrt{r}(V_{DD}-|V_{Tp}|)}{1+N\sqrt{r}}$$

W obu wzorach, 2-13 i 2-14, r dane jest wzorem 2-4.

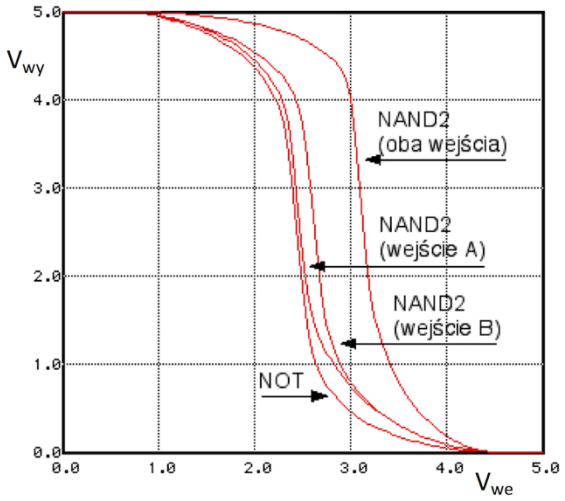
Z zależności 2-13 i 2-14 widać, że jeśli do budowy bramek NOR i NAND użyjemy takich samych tranzystorów, jak dla inwertera, to przy jednoczesnym przełączaniu wejść otrzymamy wartość napięcia przełączania różną od napięcia przełączania inwertera. Różnica jest tym większa, im większa jest liczba wejść bramki.



Rysunek 2-13.Charakterystyki przejściowe jednoczesnego przełączania obu wejść dla dwuwejściowych bramek NOR i NAND zbudowanych z takich samych tranzystorów, jak inwerter, w porównaniu z charakterystyką tego inwertera

Na rysunku 2-13 pokazano charakterystyki przejściowe inwertera oraz bramek dwuwejściowych NOR i NAND zbudowanych z takich samych tranzystorów jak inwerter. Zmiana napięcia przełączania jest niekorzystna, bowiem zmniejsza odporność bramki na zakłócenia od strony zera (NOR) lub od strony jedynki (NAND).

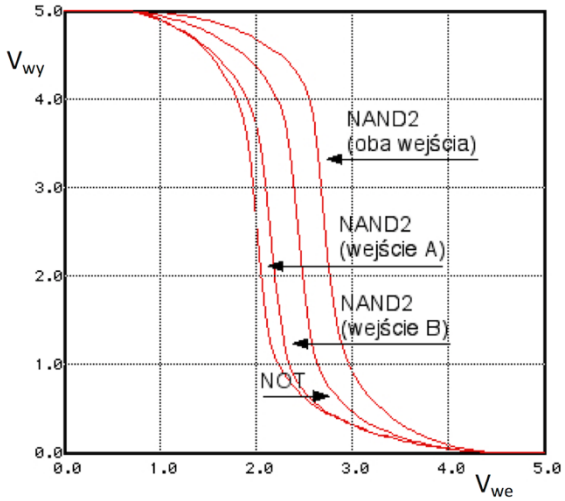
Przy przełączaniu tylko na jednym wejściu przesunięcie charakterystyki także występuje, ale jest niewielkie. Przykładowo, dla bramki NAND odpowiednie charakterystyki wyglądają następująco:



Rysunek 2-14. Charakterystyki przejściowe jednoczesnego przełączania obu wejść oraz każdego wejścia z osobna dla dwuwejściowej bramki NAND zbudowanej z takich samych tranzystorów, jak inwerter, w porównaniu z charakterystyką tego inwertera

Przy okazji zauważmy, że charakterystyki przejściowe z wejść A i B nie są identyczne. Oba wejścia są ekwiwalentne z punktu widzenia funkcji logicznej bramki, ale pod względem elektrycznym się różnią. Różnica polega na tym, że tranzystor nMOS połączony z wejściem B ma źródło zwarte z „minusem" zasilania, czyli z podłożem, natomiast tranzystor nMOS połączony z wejściem A ma źródło połączone z drenem tranzystora B, a to jest węzeł elektryczny, na którym napięcie jest ogólnie biorąc wyższe w stosunku do „minusa" zasilania, czyli podłoża. Wobec tego w tym tranzystorze występuję zależność napięcia progowego od polaryzacji podłoża względem źródła. Była o tym mowa w części I, punkt 3.1.5, równanie 3-6. Różnica między tymi charakterystykami jest jednak na tyle mała, że nie ma praktycznego znaczenia.

Bardziej korzystne charakterystyki można uzyskać poszerzając kanały tranzystorów połączonych szeregowo.



Rysunek 2-15. Charakterystyki przejściowe jednoczesnego przełączania obu wejść oraz każdego wejścia z osobna dla dwuwejściowej bramki NAND, w której kanały tranzystorów nMOS poszerzono dwukrotnie w stosunku do tranzystora nMOS inwertera, w porównaniu z charakterystyką tego inwertera

Stosowana jest tu prosta reguła: kanały te poszerza się tylokrotnie, ile tranzystorów jest połączonych szeregowo. Wówczas stosunek W/L dla całego łańcucha połączonych szeregowo tranzystorów jest taki sam, jak dla pojedynczego tranzystora przed poszerzeniem. W rezultacie charakterystyki przełączania dla poszczególnych wejść nieco się pogarszają, za to poprawia się charakterystyka jednoczesnego przełączania. Rysunek 2-15 pokazuje takie charakterystyki dla dwuwejściowej bramki NAND, w której dwukrotnie zwiększono szerokość kanałów tranzystorów nMOS.

Jak widać, zwymiarowanie tranzystorów w taki sposób, że tranzystory w połączeniu równoległym pozostają bez zmiany, a w połączeniu szeregowym są poszerzone tylokrotnie, ile jest ich w łańcuchu, daje w rezultacie charakterystyki przełączania bliskie symetrycznej charakterystyce inwertera, zapewniające dostateczną odporność bramki na zakłócenia przy wszystkich kombinacjach stanów wejść. Dotyczy to zarówno bramek NOR, jak i NAND. Jednak przesunięcia charakterystyk w stosunku do symetrycznej charakterystyki inwertera, widoczne na rysunku 2-15, rosną przy wzroście liczby wejść. Dlatego statyczne bramki kombinacyjne CMOS nie mogą mieć dowolnie dużej liczby wejść.

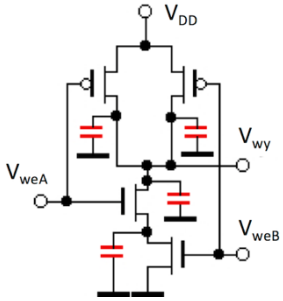
UWAGA

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

2.2.6 Czasy przełączania bramek NOR i NAND

Czasy propagacji sygnału w bramkach NOR i NAND są określone przez ten sam mechanizm, co w inwerterze - ładowanie pojemności obciążającej poprzez tranzystory pMOS, rozładowywanie poprzez tranzystory nMOS. Do oszacowania czasów propagacji można użyć zależności 2-7 i 2-8, w których trzeba podstawić wartości L/W dla N połączonych równolegle (i równocześnie włączonych) tranzystorów, czyli szerokość pojedynczego tranzystora należy pomnożyć przez N. Czas propagacji wówczas maleje N-krotnie, co jest z reguły korzystne. Gdy N tranzystorów połączonych jest szeregowo, wówczas przez N należy pomnożyć długość kanału pojedynczego tranzystora L. W tym, bardzo niekorzystnym, przypadku czas propagacji rośnie N-krotnie. Nadmiernemu wydłużeniu czasu propagacji przeciwdziała reguła poszerzania kanałów tranzystorów połączonych szeregowo, o której była mowa wyżej. Jeżeli w szeregowym połączeniu N tranzystorów kanały są poszerzone N-krotnie, to *w pierwszym przybliżeniu* niekorzystny efekt N tranzystorów połączonych szeregowo jest skompensowany N-krotnym poszerzeniem ich kanałów.

W rzeczywistości jednak większa liczba tranzystorów w bramce wydłuża czasy propagacji także dlatego, że suma pojemności ładowanych i rozładowywanych jest większa. W połączeniu równoległym sumują się pojemności złączowe drenów wszystkich tranzystorów. W połączeniu szeregowym dochodzą dodatkowe pojemności związane z węzłami wewnętrznymi w łańcuchu połączonych szeregowo tranzystorów. Pojemności te przedstawia rysunek 2-16 na przykładzie bramki NAND.



Rysunek 2-16. Pojemności w bramce NAND, które ulegają ładowaniu i rozładowywaniu przy zmianach stanów logicznych

Większa suma pojemności w bramkach wielowejściowych wydłuża czasy propagacji. Jest to drugi powód, dla którego nie używa się bramek o dowolnej liczbie wejść. Powtórzmy więc jeszcze raz: w praktyce nie stosuje się bramek statycznych NOR i NAND o liczbie wejść większej niż 4.

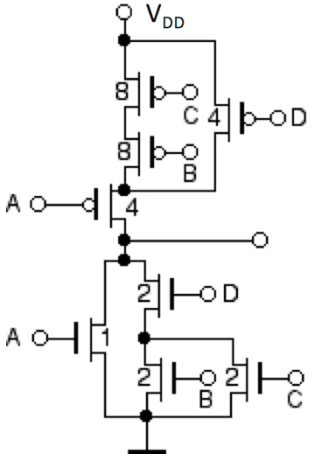
Podsumujmy: projektowanie bramek NOR i NAND w najprostszym przypadku odbywa się następująco. Dla tranzystorów połączonych równolegle zachowuje się te same wymiary, które określone zostały dla inwertera. Dla tranzystorów połączonych szeregowo zwiększa się szerokość kanału tylokrotnie, ile jest tranzystorów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

2.2.7 Bramki złożone: AND-OR-INVERT, OR-AND-INVERT

W części I, punkt 3.2.2, zobaczyliśmy, że funkcje statycznych bramek kombinacyjnych CMOS nie ograniczają się do NOT, NOR i NAND. Przez połączenia równoległe i szeregowe tranzystorów można utworzyć bramki realizujące bardziej złożone funkcje, zwane w skrócie bramkami AOI lub OAI. Bramki AOI oraz OAI mogą mieć różne liczby wejść, także nieparzyste, i mogą mieć więcej wejść niż 4. Nie należy jedynie budować bramek, w których byłyby łańcuchy szeregowo łączonych tranzystorów o długości większej niż 4.

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

Wymiarowanie tranzystorów w bramkach AOI i OAI polega, tak jak i w przypadku bramek NOR i NAND, na poszerzaniu tranzystorów w połączeniach szeregowych. Wykonuje się to przez znajdowanie w schemacie bramki łańcuchów tranzystorów i poszerzanie ich kanałów odpowiednio do ich liczby w łańcuchu. Rysunek 2-17 pokazuje przykład.



Rysunek 2-17. Przykład wymiarowania tranzystorów. Dla przejrzystości schematu połączenia bramek tranzystorów z wejściami zaznaczono tylko literami. Liczby oznaczają szerokość kanałów tranzystorów względem pewnej szerokości jednostkowej

W przykładzie z rysunku 2-17 przyjęto założenie, że stosunek ruchliwości nośników μ_n/μ_p wynosi 2, czyli w przypadku inwertera tranzystor pMOS powinien mieć kanał 2 razy szerszy od tranzystora nMOS. Znajdujemy najpierw łańcuchy tranzystorów nMOS. Są dwa takie łańcuchy: BD i CD. W obu kanały poszerzamy dwukrotnie. Dla określenia szerokości kanałów tranzystorów pMOS zaczynamy od najkrótszego łańcucha szeregowego: AD.

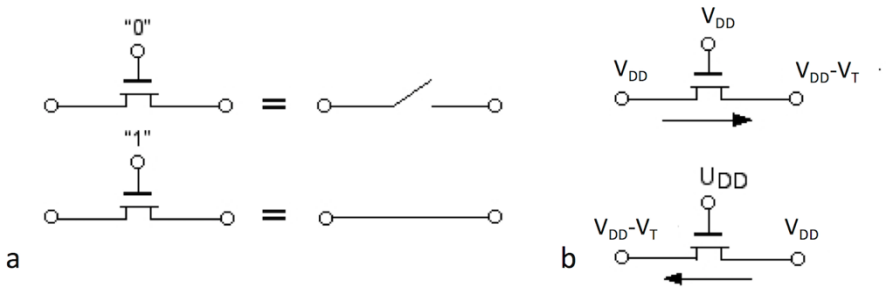
Kanały poszerzamy dwukrotnie, a ponieważ są one i tak 2 razy szersze od kanałów tranzystorów nMOS, otrzymujemy wymiary podane na rysunku 2-17. Dla określenia wymiarów tranzystorów B i C zauważmy, że stanowią one połączenie szeregowe z tranzystorem A, który ma już nadany wymiar. Kanały tranzystorów B i C musimy poszerzyć 4 razy, a wtedy ich szeregowe połączenie będzie równoważne tranzystorowi D. Ponieważ tranzystory pMOS są i tak 2 razy szersze od kanałów tranzystorów nMOS, otrzymujemy ostatecznie wymiary podane na rysunku 2-17.

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

2.3 Bramki transmisyjne i trójstanowe

2.3.1 Bramki transmisyjne

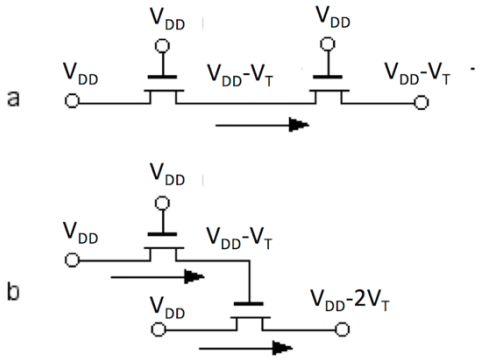
Bramki transmisyjne działają inaczej niż bramki dotąd omawiane. Są one odpowiednikami sterowanego wyłącznika. W zależności od stanu logicznego sygnału sterującego bramka transmisyjna przepuszcza sygnał z wejścia na wyjście lub nie. Najprostszą bramką transmisyjną jest pojedynczy tranzystor nMOS – rysunek 2-18.



Rysunek 2-18. Tranzystor nMOS jako najprostsza bramka transmisyjna: (a) zasada działania, (b) zjawisko degradacji jedynki logicznej

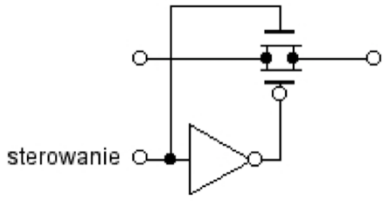
Bramka transmisyjna nie wykazuje kierunkowości – jeśli jest otwarta, może transmitować sygnał w obu kierunkach. Może to być wygodne w niektórych zastosowaniach, ale najczęściej jest to właściwość kłopotliwa.

Bramka transmisyjna w postaci pojedynczego tranzystora nMOS ma ponadto istotną wadę: wprowadza degradację poziomu jedynki logicznej (patrz rysunek 2-18b). Degradacja polega na zmniejszeniu napięcia jedynki o wartość równą w przybliżeniu napięciu progowemu tranzystora V_T, a bierze się stąd, że aby tranzystor przewodził, musi istnieć między bramką i źródłem różnica napięć równa co najmniej V_T. Dlatego nie można sterować bramki transmisyjnej sygnałem już zdegradowanym, np. pochodzącym z wyjścia innej bramki transmisyjnej. Można natomiast łączyć bramki transmisyjne szeregowo - patrz rysunek 2-19.



Rysunek 2-19. Dozwolone (a) i niedozwolone (b) łączenie bramek transmisyjnych degradujących jedynkę logiczną

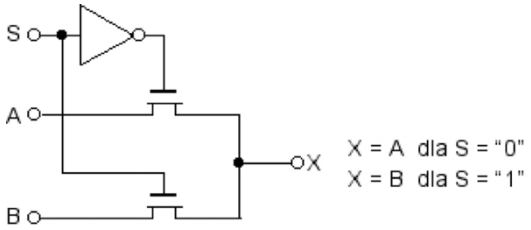
Degradacji jedynki można uniknąć stosując pełną bramkę transmisyjną CMOS złożoną z pary tranzystorów – nMOS i pMOS – połączonych równolegle i sterowanych przeciwnymi stanami logicznymi. Tranzystor pMOS zapewnia transmisję jedynki logicznej bez degradacji, a tranzystor nMOS podobnie zapewnia transmisję zera. Jednak cała bramka łącznie z niezbędnym do jej sterowania inwerterem ma aż 4 tranzystory (rysunek 2-20).



Rysunek 2-20. Pełna czterotranzystorowa bramka transmisyjna CMOS

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

Bramki transmisyjne pozwalają prosto realizować niektóre funkcje kombinacyjne, np. układy multiplekserów i demultiplekserów. Przykład pokazuje rysunek 2-21.



Rysunek 2-21. Multiplekser na bramkach transmisyjnych

Układ z rysunku 2-21 pokazuje prostotę realizacji multipleksera przy użyciu bramek transmisyjnych, ale zarazem pozwala też zilustrować pewne niebezpieczeństwa takiej realizacji. Układ działa prawidłowo pod warunkiem, że nie dochodzi do sytuacji równoczesnego włączenia obu bramek transmisyjnych. Gdyby obie bramki były włączone równocześnie, a stany wejść A i B byłyby różne, na wyjściu powstałby niedopuszczalny konflikt. W dodatku ze względu na dwukierunkową transmisję sygnału przez bramkę transmisyjną sygnał z wejścia A oddziaływałby bezpośrednio na stan wejścia B, i odwrotnie. Skutki tej niedozwolonej sytuacji nie są możliwe do przewidzenia bez znajomości szczegółów budowy bramek dostarczających sygnały na wejścia A i B. W stanie ustalonym do równoczesnego włączenia obu bramek oczywiście dojść nie może, natomiast może się to zdarzyć podczas zmiany stanu logicznego wejścia sterującego. Wyobraźmy sobie, że początkowy stan wejścia sterującego S to "0". Włączona jest wówczas górna bramka transmisyjna, zaś dolna - wyłączona. Gdy stan wejścia S zmienia się na "1", dolna bramka zostaje włączona, a górna wyłączona, ale górna bramka wyłączona jest z opóźnieniem wynikającym z niezerowego czasu propagacji sygnału w inwerterze. W rezultacie może pojawić się taki odcinek czasu, w którym obie bramki transmisyjne przewodzą równocześnie. Nie musi, ale może spowodować to błędne działanie układu z multiplekserem - zależy to od tego, jak zbudowany jest cały układ.

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

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

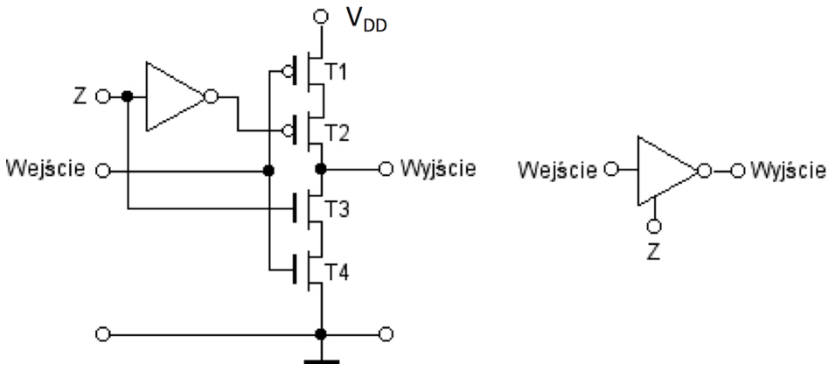
Równanie 2-15

$$R_t = \frac{1}{K_n(V_{DD}-V_{Tn})+K_p(V_{DD}-|V_{Tp}|)}$$

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

2.3.2 Bramki trójstanowe

Specjalnym rodzajem bramek statycznych są bramki trójstanowe. Są to bramki, których wyjście może być w stanie zera, jedynki lub wysokiej impedancji. W tym ostatnim przypadku wyjście bramki może być uważane za odłączone od układu. Pozwala to na przykład dołączyć do tego samego węzła elektrycznego wyjścia kilku bramek, z których w każdym momencie wszystkie z wyjątkiem jednej są w stanie wysokiej impedancji. Najprostszą bramką trójstanową jest inwerter trójstanowy. Jego schemat pokazany jest na rysunku 2-22. Można go traktować jako zwykły inwerter skojarzony z bramką transmisyjną.



Rysunek 2-22. Inwerter trójstanowy: schemat i symbol

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

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

3 Bramki dynamiczne i przerzutniki

Bramki statyczne, omawiane w poprzednim punkcie, nie są jedynym sposobem realizacji układów kombinacyjnych CMOS. W pewnych przypadkach stosuje się układy zwane dynamicznymi. Terminem „układy dynamiczne" określana jest szeroka klasa układów, w których wartości logiczne - zera i jedynki - są reprezentowane przez ładunki gromadzone w pojemnościach. Jednak głównym obszarem zastosowań układów dynamicznych nie jest logika kombinacyjna, lecz układy sekwencyjne. Praktycznie każdy układ cyfrowy jest układem sekwencyjnym, tj. zawiera elementy pamięciowe: przerzutniki, rejestry, a nawet całe bloki pamięci. Będziemy je teraz omawiać.

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

3.1 Istota układów dynamicznych i przykład ich zastosowań

3.1.1 Istota i cechy układów dynamicznych

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

Układy dynamiczne, niezależnie od budowy i przeznaczenia, mają pewne wspólne cechy:
1. Pojemności, w których gromadzony jest ładunek, są zawsze związane z pewnymi upływnościami, takimi jak prądy wsteczne złącz *p-n* i prądy podprogowe tranzystorów MOS. W związku z tym czas przechowywania ładunku w pojemnościach jest ograniczony. Naładowana do napięcia V_{DD} pojemność ulega stopniowemu rozładowaniu. Czas tego rozładowania może być mierzony milisekundami w temperaturze otoczenia, ale maleje do mikrosekund dla temperatur o kilkadziesiąt stopni wyższych, bo ze wzrostem temperatury bardzo szybko rosną prądy upływu. Wynika z tego, że wartość „1" zapisana jako ładunek w pojemności powinna być w krótkim czasie odczytana i wykorzystana, a jeżeli ma być przechowywana przez długi czas, to wymagać będzie okresowego odświeżania (czyli odczytu i ponownego zapisu tej samej wartości).
2. Układy dynamiczne wymagają taktowania, konieczny jest więc sygnał zegara. Większość układów dynamicznych wymaga przy tym dość precyzyjnego taktowania, co stwarza problemy z generacją sygnałów zegarowych i ich rozprowadzeniem w dużych układach.
3. W związku z tym, że czas przechowywania ładunku w pojemności jest ograniczony, układy dynamiczne nie mogą działać z dowolnie małą częstotliwością zegara. Zbyt mała częstotliwość powoduje błędy w działaniu układu.
4. W układach dynamicznych nie mamy do czynienia ze statycznymi charakterystykami przejściowymi, a odporność na zakłócenia jest definiowana inaczej niż dla bramek statycznych.
5. Układy dynamiczne mogą wprowadzać degradację jedynki logicznej, zarówno ze względu na rozładowywanie pojemności w czasie, jak i ze względu na zjawisko podziału ładunku. Polega ono na tym, że przy odczycie pojemność, w której zgromadzony został ładunek reprezentujący stan „1", zostaje połączona równolegle z inną, często znacznie większą pojemnością. Zgodnie z prawami elektrostatyki napięcie V na pojemności jest proporcjonalne do ładunku Q i odwrotnie proporcjonalne do pojemności C

Równanie 3-1

$$V = \frac{Q}{C}$$

a więc jeśli pewna pojemność C_1 zostanie naładowana do napięcia V_{DD} i zostanie w niej zgromadzony ładunek $Q = C_1 V_{DD}$, a potem zostanie ona połączona równolegle z drugą pojemnościąC_2, to napięcie spadnie do wartości

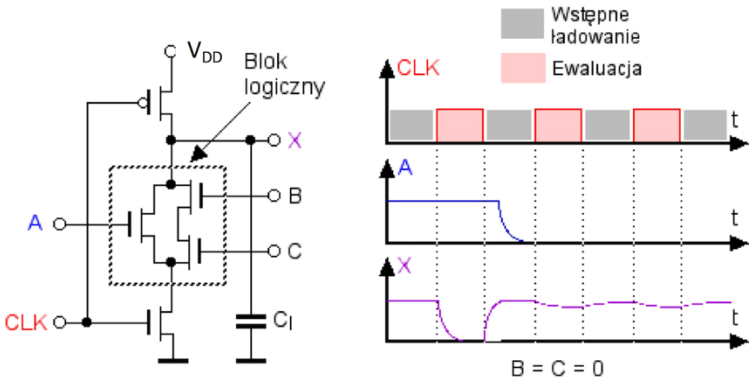
Równanie 3-2

$$V = V_{DD} \frac{C_1}{C_1 + C_2}$$

Jeżeli $C_2 > C_1$, degradacja jest znaczna i wymaga zastosowania specjalnych układów regenerujących poziom jedynki. Jest to bardzo poważny problem na przykład w pamięciach dynamicznych RAM. Wymienione wyżej cechy układów dynamicznych powodują, że projektowanie tych układów jest trudniejsze, niż bramek statycznych. Niemniej, układy dynamiczne mają szereg zalet i dlatego warto się z nimi zapoznać.

3.1.2 Przykład: kombinacyjne bramki dynamiczne typu DOMINO

Dobrym przykładem układów dynamicznych są dość często stosowane w praktyce dynamiczne bramki kombinacyjne zwane bramkami typu DOMINO. Przykładowa bramka typu DOMINO pokazana jest na rysunku 3-1.



Rysunek 3-1. Schemat przykładowej bramki typu DOMINO i przebiegi napięć w funkcji czasu (przy założeniu, że na wejściach B i C są cały czas zera)

Bramka składa się z dwóch części: bloku logicznego, w którym odpowiednie połączenia szeregowe i równoległe tranzystorów nMOS określają wykonywaną funkcję kombinacyjną, oraz dwóch dodatkowych tranzystorów: dolnego nMOS i górnego pMOS, które są okresowo na zmianę włączane i wyłączane sygnałem zegara CLK. Działanie bramki odbywa się w dwóch fazach: wstępnego ładowania (gdy zegar CLK jest w stanie „0”) i ewaluacji (gdy zegar CLK jest w stanie „1”). W fazie wstępnego ładowania włączony jest górny tranzystor pMOS, a dolny tranzystor nMOS jest wyłączony. Niezależnie od stanów wejść A, B i C pojemność C_l obciążająca węzeł wyjściowy X ładuje się ze źródła zasilania do napięcia V_{DD} reprezentującego „1”. Gdy stan zegara zmienia się z „0” na „1”, zaczyna się faza ewaluacji. Górny tranzystor pMOS zostaje wyłączony, a włącza się dolny tranzystor nMOS. Teraz stan na wyjściu zależy od stanów wejść A, B i C. Jeśli tranzystor A jest włączony, pojemność C_l rozładowuje się. To samo dzieje się, gdy włączone są równocześnie tranzystory B i C. W tych przypadkach na wyjściu pojawia się (po upływie czasu potrzebnego na rozładowanie pojemnościC_l) zero. Jeśli stany wejść są takie, że pojemność nie może się rozładować, na wyjściu pozostaje stan „1”. Rysunek 3-1 pokazuje działanie przykładowej bramki. Po pierwszej fazie wstępnego ładowania tranzystor A jest włączony, więc w fazie ewaluacji na wyjściu pojawia się „0”. Po dwóch następnych fazach wstępnego ładowania wszystkie trzy tranzystory A, B i C są wyłączone, więc na wyjściu w fazach ewaluacji utrzymuje się „1”.

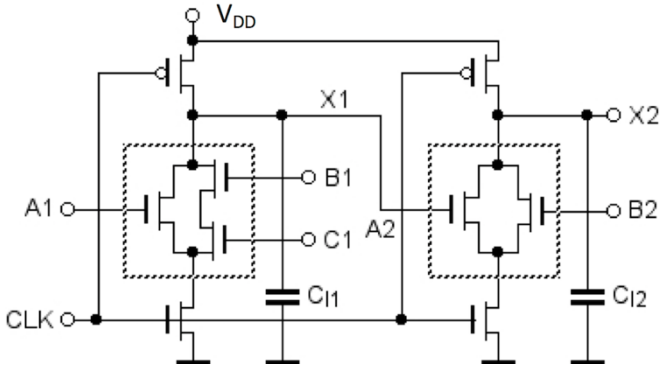
Bramka omawianego typu ma kilka istotnych zalet:

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

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

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

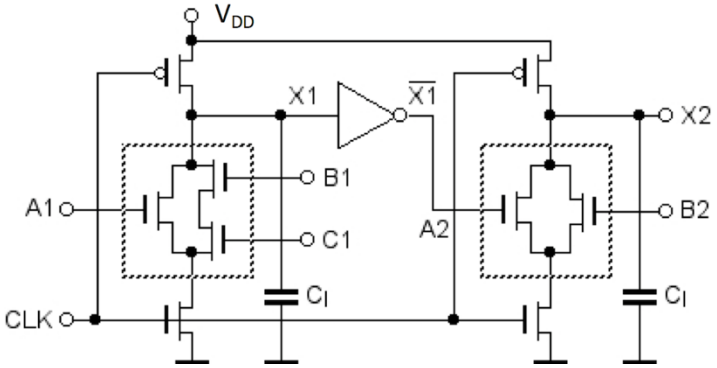
Dodatkowym utrudnieniem jest to, że wyjść bramek dynamicznych takich, jak pokazano na rysunku 3-1, nie wolno łączyć bezpośrednio z wejściami bramek tego samego rodzaju. Skąd to ograniczenie? Bierze się ono stąd, że w układzie kombinacyjnym typu DOMINO wszystkie bramki taktowane są tym samym sygnałem zegara. Wobec tego fazy ewaluacji zaczynają się we wszystkich bramkach w tym samym momencie. Problem powstaje stąd, że w chwili rozpoczęcia fazy ewaluacji wyjścia wszystkich bramek są w stanie „1”. Spójrzmy na rysunek 3-2, pokazujący dwie bramki typu DOMINO, przy czym druga z nich otrzymuje wejściowy sygnał logiczny wprost z wyjścia pierwszej.



Rysunek 3-2. Bramka typu DOMINO sterująca drugą podobną bramkę

Gdy zaczyna się faza ewaluacji, na wyjściu X1 mamy zawsze stan „1”. Zatem w tym momencie tranzystor drugiej bramki połączony z jej wejściem A2 jest włączony i pojemność C_{l2} zaczyna się rozładowywać. Dopiero po upływie pewnego czasu, potrzebnego na rozładowanie pojemności C_{l1}, stan X1 zmieni się, być może, na „0” (zależy to, oczywiście, od stanów logicznych na wejściach A1, B1 i C1). Ale w tym momencie pojemność C_{l2} może być już rozładowana do tego stopnia, że na wyjściu X2 będziemy mieli „0” lub stan nieokreślony pomiędzy „0”, a „1”, a nie jedynkę. Zatem układ da nam wynik błędny.

Można tego uniknąć w prosty sposób rozdzielając bramki typu DOMINO statycznymi inwerterami - patrz rysunek 3-3.



Rysunek 3-3. Bramka typu DOMINO sterująca drugą podobną bramkę poprzez inwerter

Inwerter powoduje, że na wejście A2 drugiej bramki na początku fazy ewaluacji podawane jest zawsze „0", a nie „1". Dopiero po ustaleniu się właściwego stanu wyjścia X1 stan wejścia A2 zmieni się, być może, na „1". Nie ma niebezpieczeństwa błędnych zadziałań, ponieważ w kolejnych stopniach układu bramki „czekają", aż w poprzednich stopniach ustalą się docelowe, prawidłowe stany. Stąd zresztą pochodzi nazwa „domino". Każdy układ kombinacyjny zbudowany z bramek dynamicznych rozdzielonych inwerterami działa w taki sposób, że w czasie trwania fazy ewaluacji właściwe poziomy logiczne ustalają się stopniowo w kolejnych, coraz dalszych od wejściach bramkach. Twórcom tego sposobu realizacji układów kombinacyjnych skojarzyło się to z dziecięcą zabawą polegającą na pionowym ustawieniu jeden za drugim klocków domina, a następnie pchnięciu pierwszego klocka, który przewracając się popycha i przewraca drugi, ten przewraca następny i w końcu cały wąż klocków się kładzie.

Dodatkowe inwertery oczywiście wprowadzają pewne opóźnienia, ale nawet z nimi układ typu DOMINO może działać szybciej od układu z bramek statycznych. Pojawia się jednak pewne dodatkowe, często kłopotliwe ograniczenie - nie ma inwertera typu DOMINO, a bramki (z uwzględnieniem statycznych inwerterów) realizują tylko funkcje nie zawierające nigdzie negacji. Mówiąc ściśle, negacje (w postaci dodatkowych statycznych inwerterów) są możliwe na wejściu i na wyjściu bloku kombinacyjnego zbudowanego z bramek DOMINO, ale nie mogą występować wewnątrz tego bloku. Komplikuje to projekt logiczny. Istnieją inne wersje układów dynamicznych, w których stosowanie inwerterów statycznych rozdzielających poszczególne stopnie nie jest potrzebne. Nie będziemy ich tutaj omawiać.

Dla uniknięcia nieporozumień trzeba zwrócić uwagę, że pod względem realizowanej funkcji logicznej układy z rysunków 3-2 i 3-3 nie są równoważne. Projekt logiczny układu przeznaczonego do realizacji w technice bramek dynamicznych musi być dostosowany do rodzaju użytych bramek. Wadą bramek dynamicznych jest mniejsza od bramek statycznych odporność na zakłócenia. Impuls zakłócający, którego amplituda jest większa od napięcia progowego tranzystorów w bloku logicznym, a czas trwania dostatecznie długi, może spowodować fałszywe rozładowanie pojemności C_l, a tym samym błąd w działaniu bramki. Inną słabą stroną bramek dynamicznych jest dodatkowy pobór mocy wynikający z obecności sygnału zegarowego. Bramki dynamiczne są taktowane nawet wtedy, gdy stany na ich wejściach nie zmieniają się. Oznacza to dodatkowy pobór mocy przez generator zegara.

Biorąc to wszystko pod uwagę widzimy, że stosowanie bramek dynamicznych ma sens tylko wtedy, gdy układ z bramkami statycznymi nie pozwala osiągnąć niezbędnej szybkości działania. Projektowanie układów z bramkami dynamicznymi jest znacznie trudniejsze niż układów z bramkami statycznymi, i z reguły wymaga

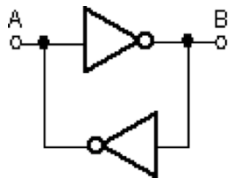
szczegółowych symulacji elektrycznych. Dodajmy, że w bibliotekach komórek standardowych nie ma bramek dynamicznych. Trzeba je więc projektować w stylu „*full custom*”.

3.2 Przerzutniki

Zajmiemy się teraz elementarnymi układami pamięciowymi, jakimi są przerzutniki. Najprostszy przerzutnik dwustabilny jest układem statycznym. Takie przerzutniki używane są jako komórki pamięci w statycznych pamięciach o swobodnym dostępie (Static Random Access Memory – sRAM). Będzie o nich mowa dalej. Natomiast większość przerzutników używanych poza pamięciami należy do klasy układów dynamicznych.

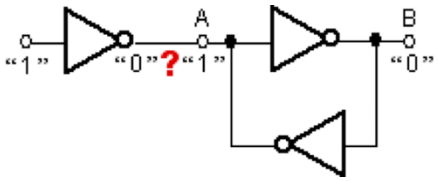
3.2.1 Podstawowy przerzutnik statyczny

Podstawowy przerzutnik statyczny powstaje przez połączenie dwóch inwerterów w taki sposób, że sygnał z wyjścia pierwszego inwertera jest podawany na wejście drugiego, a sygnał z wyjścia drugiego inwertera jest podawany na wejście pierwszego (rysunek 3-4). Taki układ, jak łatwo się przekonać, ma dwa samopodtrzymujące się stany stabilne: gdy w węźle A jest stan „1”, to w węźle B „0” i odwrotnie. Taki układ można więc użyć jako elementarną komórkę pamiętającą jeden bit informacji.



Rysunek 3-4. Podstawowy przerzutnik statyczny

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

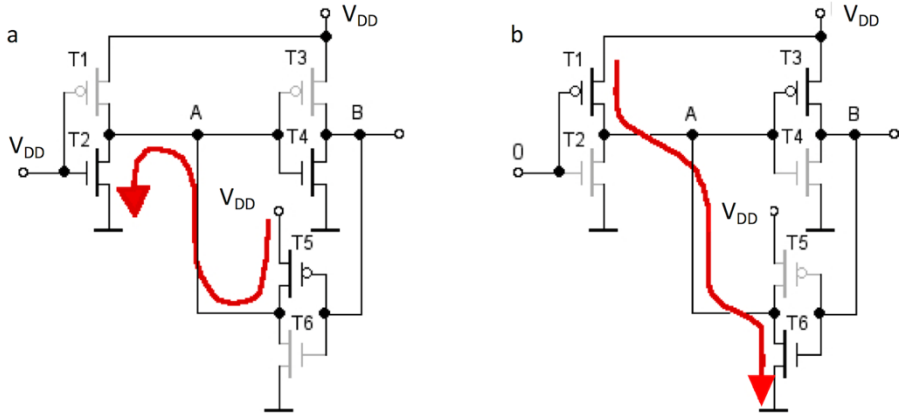


Rysunek 3-5. Przełączanie przerzutnika przez inwerter

Analiza układu z rysunku 3-5 jako układu logicznego nie pozwala stwierdzić, czy w węźle A ustali się „0”, czy też „1”. Trzeba zbadać wartość napięcia w węźle A. Schemat zastępczy układu (rysunek 3-6a), w którym pominięto (zaznaczono kolorem szarym) tranzystory znajdujące się w stanie odcięcia, pokazuje, że napięcie w węźle A określone jest przez dzielnik napięcia złożony z tranzystorów T5 i T2. Oba znajdują się w stanie przewodzenia. Jeśli tranzystor T2 będzie miał bardzo małą rezystancję w stosunku do tranzystora T5, to napięcie w węźle A spadnie do wartości bliskiej zeru. Wówczas ulegnie przełączeniu inwerter złożony z tranzystorów T3 i T4, w węźle B pojawi się „1”, czyli napięcie równe V_{DD}, które przełączy inwerter złożony z tranzystorów T5 i T6. W ten sposób nastąpi zmiana stanu przerzutnika. Małą rezystancję tranzystora T2 osiąga się przez dobór odpowiednio dużej szerokości kanału. Przy przełączaniu w drugą stronę (tj. zmianie stanu w węźle A z „0” na „1”) dzielnik napięcia tworzą tranzystory T1 i T6 (rysunek 3-6b). Tranzystor T1 musi mieć bardzo dużą szerokość kanału, aby wymusić w węźle A napięcie bliskie V_{DD}.

Zatem zmiana stanu przerzutnika wymaga sterowania go z inwertera mającego tranzystory (T1 i T2) o szerokościach kanału znacznie większych, niż szerokości kanałów tranzystorów T5 i T6 w przerzutniku. Nie ma

prostego wzoru pozwalającego obliczyć wymagane szerokości kanałów tranzystorów w inwerterze sterującym. Należy dobrać te szerokości korzystając z symulacji elektrycznej. Jako wartość startową można wybrać szerokości kanałów tranzystorów w inwerterze sterującym (T1, T2) 3 razy większe od szerokości kanałów tranzystorów w przerzutniku (T5, T6).

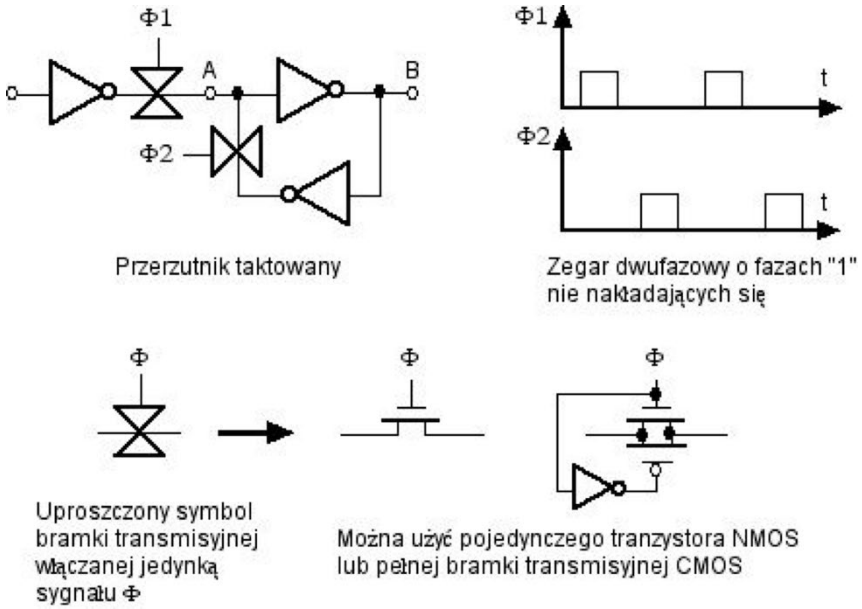


Rysunek 3-6. Przełączanie przerzutnika: (a) dzielnik napięcia T2 - T5, (b) dzielnik napięcia T6 - T1

Omawiany przerzutnik znajduje zastosowanie głównie jako podstawowa komórka pamięci statycznych sRAM, natomiast w innych zastosowaniach stosuje się różne wersje przerzutników dynamicznych.

3.2.2 Podstawowy przerzutnik dynamiczny

Dodając do podstawowego przerzutnika dwie bramki transmisyjne można zbudować przerzutnik, który będzie się przełączał przy sterowaniu z dowolnego inwertera (lub innej bramki), bez względu na wymiary tranzystorów. Idea polega na tym, że na czas przełączania przerywa się pętlę dodatniego sprzężenia zwrotnego występującą w przerzutniku. Układ zbudowany według tej idei pokazany jest na rysunku 3-7.



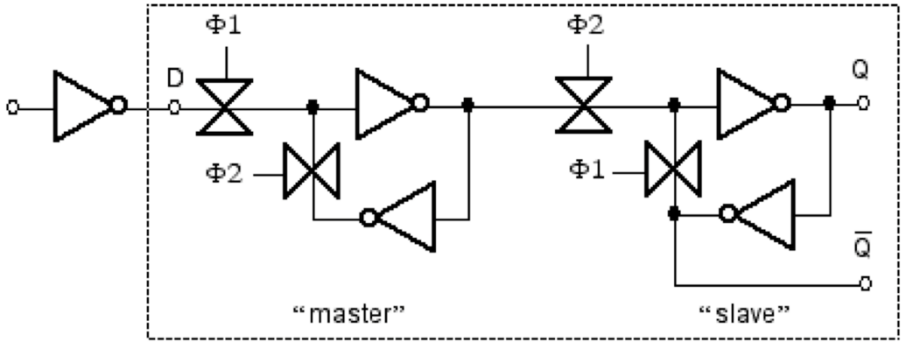
Rysunek 3-7. Przerzutnik dynamiczny wykorzystujący bramki transmisyjne

Gdy zegar Φ1 jest w stanie „1”, na wejście A podawany jest sygnał z inwertera sterującego. W tym czasie Φ2 jest w stanie „0”, a więc sygnał z dolnego inwertera w przerzutniku nie dociera do węzła A. Wobec tego poziomy logiczne w węzłach A i B ustalają się bez trudu. Po zmianie stanu Φ2 na „1”, a Φ1 na „0” następuje utrwalenie stanów w węzłach A i B. Układ działa prawidłowo, gdy fazy „1” obu zegarów nie nakładają się w czasie. W tych odcinkach czasowych, w których oba sygnały zegarowe są w stanie „0”, stany logiczne w węzłach A i B są podtrzymywane dzięki istniejącym w tych węzłach pojemnościom pasożytniczym. Układ ten należy więc do klasy układów dynamicznych. Zwany jest jednak także często pseudo-statycznym, bowiem gdy Φ2 jest w stanie „1”,

oba inwertery przerzutnika wzajemnie podtrzymują swoje stany tak samo, jak w omawianym poprzednio przerzutniku statycznym.

3.2.3 Przerzutnik dynamiczny typu D

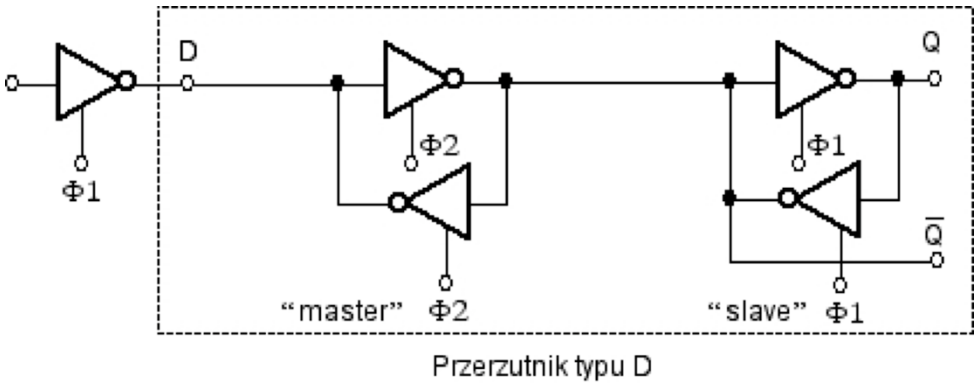
Z dwóch przerzutników taktowanych można łatwo zbudować przerzutnik typu D (patrz punkt 3.2.1 w części pierwszej). Jest to najczęściej używany rodzaj przerzutnika w układach cyfrowych CMOS – w większości układów nie używa się w ogóle innych przerzutników. Przerzutnik typu D zapamiętuje sygnał wejściowy i opóźnia go o jeden takt zegara. Schemat takiego przerzutnika zwanego przerzutnikiem „master-slave" (dosłownie po polsku „pan-niewolnik") jest pokazany na rysunku 3-8.



Rysunek 3-8. Przerzutnik typu D

Gdy zegar Φ1 jest w stanie „1", następuje wpisanie stanu wejścia do pierwszego stopnia („master"). Podczas jedynki zegara Φ2 następuje przepisanie do drugiego stopnia („slave"). Tu również ważne jest, aby jedynki Φ1 i Φ2 nie nakładały się w czasie, bowiem równoczesne otwarcie wszystkich bramek transmisyjnych uniemożliwia prawidłowe działanie układu.

Dość często spotyka się przerzutniki typu D, w których jednak zastosowano zamiast bramek transmisyjnych inwertery trójstanowe. Takie przerzutniki działają dokładnie tak samo. Sygnały zegarowe Φ1 i Φ2 włączają lub wyłączają stan wysokiej impedancji. Schemat takiego przerzutnika pokazuje rysunek 3-9.

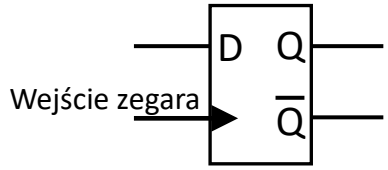


Rysunek 3-9. Przerzutnik typu D z inwerterami trójstanowymi

Przerzutnik z inwerterami trójstanowymi wymaga taktowania zegarem dwufazowym o fazach nie nakładających się, podobnie jak przerzutnik z bramkami transmisyjnymi. Jak zobaczymy dalej, zapewnienie właściwego taktowania zegarem dwufazowym o fazach nie nakładających się może być poważnym problemem technicznym w dużych układach.

Rysunek 3-10 pokazuje symbol przerzutnika typu D używany w schematach układów logicznych. Zauważmy, że w tym symbolu jest tylko jedno wejście zegarowe, a nie dwa wejścia dla zegara dwufazowego. Z powodu problemów z taktowaniem zegarem dwufazowym (o czym była mowa wyżej) w praktycznych rozwiązaniach przerzutników jest tylko wejście dla sygnału zegarowego odpowiadającego zegarowi Φ1 (patrz rysunki 3-8, 3-

9), a drugi sygnał zegarowy $\Phi 2$ jest generowany lokalnie w układzie przerzutnika (czego nie ma na rysunkach 3-8, 3-9).

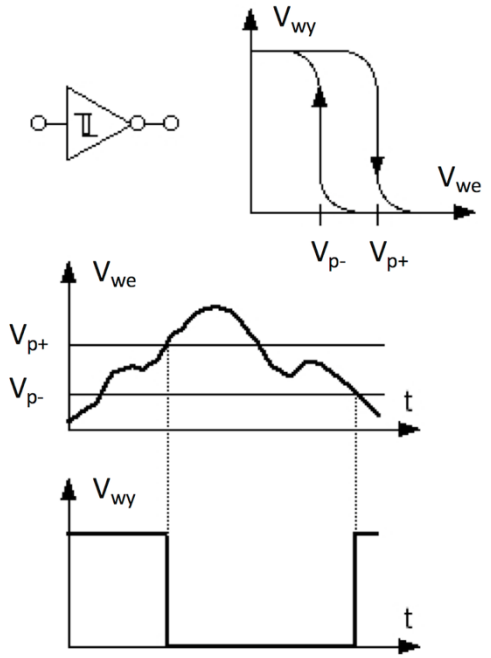


Rysunek 3-10. Symbol przerzutnika typu D

Łącząc w łańcuch przerzutniki typu D można zbudować szeregowy rejestr przesuwający. O rejestrach będzie mowa nieco dalej.

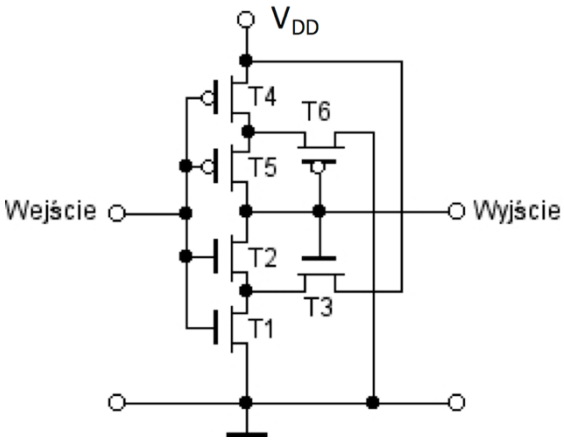
3.2.4 Przerzutnik Schmitta (inwerter z histerezą)

Szczególnym rodzajem przerzutnika jest układ zwany przerzutnikiem Schmitta. Jest to inwerter z histerezą - charakterystyka przejściowa jest inna przy przełączaniu „0”->„1” niż przy przełączaniu „1”->„0”. Taki układ bywa stosowany tam, gdzie trzeba przekształcić sygnał o niezbyt regularnym kształcie w ciąg prawidłowych zer i jedynek. Symbol, charakterystyki i typowe zastosowanie przerzutnika Schmitta ilustruje rysunek 3-11.



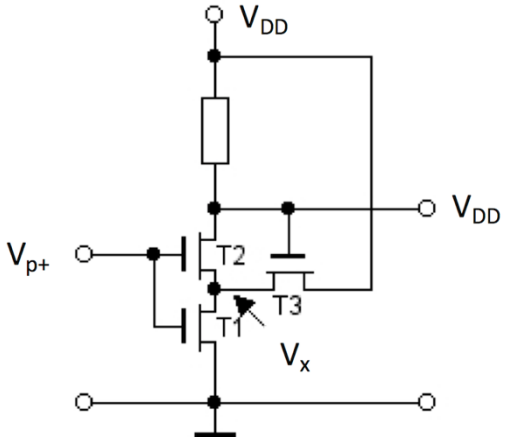
Rysunek 3-11. Przerzutnik Schmitta: symbol, typowe charakterystyki oraz zastosowanie do kształtowania sygnału

Schemat przerzutnika Schmitta zrealizowanego jako bramka CMOS pokazuje rysunek 3-12.



Rysunek 3-12. Schemat przerzutnika Schmitta

Zanalizujemy najpierw proces przełączania od stanu „0” do stanu „1” na wejściu. Pozwoli nam to także określić napięcie przełączania V_{p+}. Przyjmiemy dla uproszczenia, że napięcie V_{p+} jest to graniczne napięcie, po przekroczeniu którego zaczyna się proces przełączania. Zatem gdy na wejściu panuje napięcie V_{p+}, to na wyjściu jest jeszcze stan „1”, czyli napięcie V_{DD}. W analizie pomocny będzie uproszczony schemat przerzutnika (rysunek 3-12), w którym szeregowe połączenie przewodzących tranzystorów pMOS (T4 i T5) zastąpiono pewną nieistotną z punktu widzenia naszych rozważań rezystancją.



Rysunek 3-13. Uproszczony schemat zastępczy przerzutnika Schmitta dla przełączania "0"->"1" na wejściu

Przełączenie, czyli spadek napięcia na wyjściu do wartości 0, zaczyna się w momencie, gdy zaczyna przewodzić tranzystor T2, czyli gdy napięcie między jego źródłem, a drenem przekroczy jego napięcie progowe V_{Tn}. Zatem proces przełączania zaczyna się, gdy spełniony jest warunek

Równanie 3-3

$$V_{p+} - V_x = V_{Tn}$$

W momencie początku przełączania tranzystor T1 jest już na pewno włączony, bowiem potencjał bramki obu tranzystorów, T1 i T2, jest taki sam, a potencjał źródła tranzystora T1 jest na pewno niższy, niż potencjał źródła tranzystora T2, a więc spełniony jest warunek $V_{GS1} > V_{Tn}$. Napięcie V_x jest w momencie początku przełączania określone przez dzielnik napięcia, jaki tworzą dwa przewodzące tranzystory: T1 i T3 (T3 jest włączony, bo w chwili rozpoczęcia procesu przełączania na jego bramce jest jeszcze stan „1”, czyli napięcie V_{DD}). Przyrównując prądy drenu tranzystorów T1 i T3 (prąd drenu tranzystora T2 na początku procesu przełączania jest do pominięcia) otrzymujemy

Równanie 3-4

$$K_{n1}\left(V_{p+} - V_{Tn}\right)^2 = K_{n3}[(V_{DD} - V_x) - V_{Tn}]^2$$

Łącząc to równanie z warunkiem 3-3 po przekształceniach otrzymujemy wyrażenie pozwalające oszacować napięcie V_{p+}:

Równanie 3-5

$$V_{p+} = \frac{V_{DD} + V_{Tn}\sqrt{\frac{K_{n1}}{K_{n3}}}}{1 + \sqrt{\frac{K_{n1}}{K_{n3}}}}$$

gdzie K_{n1} i K_{n3} są to współczynniki przewodności tranzystorów T1 i T3 (definicja współczynnika przewodności: wzór 2-2). Zależność 3-5 daje dość mało dokładne oszacowanie między innymi dlatego, że napięcia progowe tranzystorów T1, T2 i T3 nie są identyczne. Źródła tranzystorów T2 i T3 nie są połączone z podłożem układu (o

potencjale równym 0), lecz z węzłem, w którym panuje wyższe napięcie V_x. To oznacza niezerowe napięcie polaryzacji podłoża V_{BS} tych tranzystorów, co powoduje wzrost napięcia progowego zgodnie z zależnością 3-6 z części I. Zatem po wstępnym określeniu wymiarów tranzystorów należy wartość V_{p+} uściślić przy pomocy symulacji i ewentualnie wprowadzić korektę wymiarów. Zauważmy, że tylko wymiary tranzystorów T1 i T3 mają wpływ na napięcie V_{p+}. Wymiary tranzystora T2 można przyjąć takie same, jak T1.

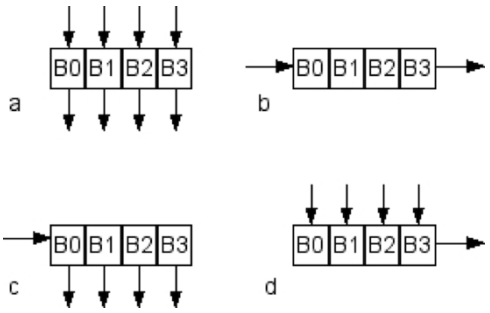
Rozumowanie analogiczne do przytoczonego wyżej pozwala określić napięcie przełączania V_{p-}. Wynosi ono

Równanie 3-6

$$V_{p-} = \frac{(V_{DD} - |V_{Tp}|)\sqrt{\frac{K_{p4}}{K_{p6}}}}{1+\sqrt{\frac{K_{p4}}{K_{p6}}}}$$

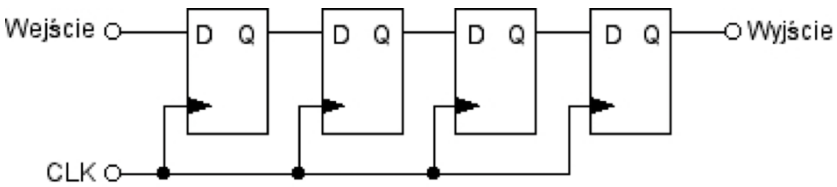
3.2.5 Rejestry

Rejestry służą do zapisu, przechowywania i odczytu grup bitów - typowo od kilku do kilkuset bitów. Istnieją rejestry równoległe i szeregowe. Do rejestrów równoległych zapisuje się i odczytuje wszystkie bity równocześnie. Rejestry szeregowe umożliwiają szeregowe wpisywanie kolejnych bitów i podobnie odczyt. Istnieją też rejestry szeregowo-równoległe, umożliwiające np. szeregowe wpisywanie i równoległy odczyt, lub odwrotnie.



Rysunek 3-14. Rodzaje rejestrów: (a) równoległy (wpis i odczyt równoległy), (b) szeregowy przesuwający (wpis i odczyt szeregowy), (c) szeregowo-równoległy (wpis szeregowy, odczyt równoległy) (d) równoległo-szeregowy (wpis równoległy, odczyt szeregowy)

Rejestry równoległe to zespoły przerzutników jednobitowych. Bardziej interesujące są rejestry szeregowe. Zwane są one przesuwającymi, ponieważ przy szeregowym wpisywaniu i odczycie kolejne bity są "przesuwane" przy każdym takcie zegara w kierunku od wejścia do wyjścia. Typowym rejestrem przesuwającym jest rejestr zbudowany z przerzutników typu D.



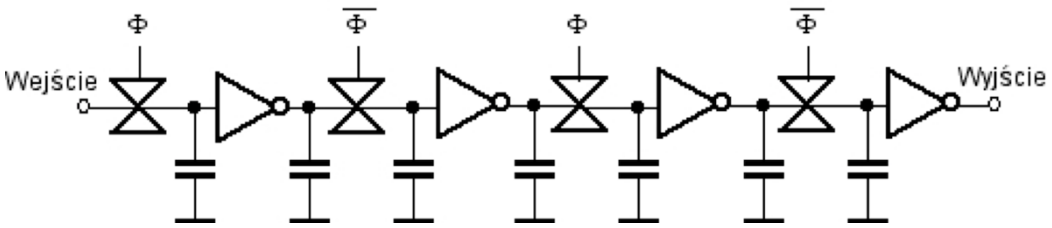
Rysunek 3-15. Rejestr szeregowy przesuwający z przerzutników typu D

Każdy takt zegara CLK powoduje przesunięcie zapisanych w przerzutnikach bitów o jedną pozycję w prawo i wczytanie nowego bitu z wejścia.

Rejestr pokazany na rysunku 3-15 jest taktowany pojedynczym sygnałem zegara. Tymczasem w przerzutnikach typu D takich, jak omawiane poprzednio (rysunki 3-8 i 3-9) wymagany jest zegar dwufazowy o fazach „1” nie nakładających się. Teoretycznie możliwe są dwa rozwiązania: albo druga faza zegara jest generowana lokalnie

w każdym przerzutniku (takie rozwiązanie, wspomniane wcześniej, jest zwykle stosowane w przerzutnikach wchodzących w skład bibliotek komórek standardowych), albo oba zegary są generowane centralnie i doprowadzane do każdego przerzutnika. To drugie rozwiązanie jest technicznie kłopotliwe. Do tego zagadnienia będziemy jeszcze wracać.

Oprócz rejestrów złożonych z przerzutników pseudo-statycznych (takich, jak te z rysunków 3-8 i 3-9) bywają też stosowane rejestry w pełni dynamiczne. Takie rejestry mają znacznie prostszą budowę. Pojedynczy stopień rejestru składa się z inwertera, bramki transmisyjnej oraz pojemności, która służy jako element pamięciowy:



Rysunek 3-16. Dynamiczny rejestr szeregowy przesuwający

W rejestrze dynamicznym kolejne bramki transmisyjne (mogą to być pojedyncze tranzystory nMOS) są otwierane i zamykane na przemian sygnałem zegara i jego negacją. Otwarta bramka powoduje podanie sygnału z wyjścia poprzedniego inwertera na wejście następnego. Łatwo się przekonać, że w tej sytuacji jeden pełny takt zegara powoduje przesunięcie o dwa stopnie. Rejestr dynamiczny wymaga ciągłego taktowania, a częstotliwość zegara nie może być dowolnie mała, jest ograniczona od dołu zjawiskiem upływności rozładowującej kondensatory.

3.2.6 Zegary i taktowanie

Zegar jest niezbędny w każdym układzie sekwencyjnym, a także w układach kombinacyjnych z bramkami dynamicznymi. W praktyce zegar jest potrzebny w każdym układzie cyfrowym, ponieważ układy cyfrowe przeznaczone do realizacji jako scalone układy CMOS są układami synchronicznymi. Stosowane są dwa sposoby generacji sygnałów zegarowych:

- zegar generowany w systemie poza układem,
- zegar generowany w układzie.

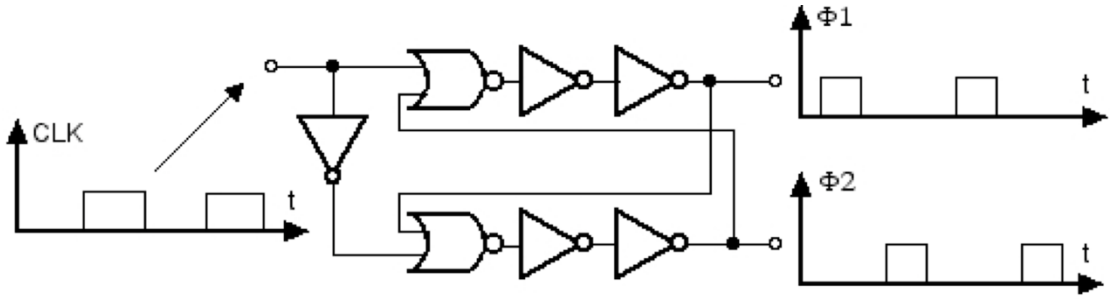
W pierwszym przypadku układ ma wejście dla zewnętrznego sygnału zegara. Sygnału tego zwykle nie dostarcza się bezpośrednio do bramek w układzie. Steruje on wewnętrznymi układami formującymi sygnały zegarowe i dostarczającymi je w wymagane miejsca.

Drugi przypadek zachodzi wtedy, gdy nie ma potrzeby synchronizowania działania układu z innymi układami, na przykład wtedy, gdy układ stanowi funkcjonalną całość. Przykłady takich układów: układ do zegarka elektronicznego, układ prostego kalkulatora itp. W praktyce nie ma potrzeby samodzielnego projektowania oscylatorów, które generują sygnał zegara. Producenci układów ASIC dostarczają gotowe projekty takich oscylatorów w bibliotekach komórek standardowych. Na ogół trzeba do takiego oscylatora dołączyć z zewnątrz rezonator kwarcowy, który ustala częstotliwość zegara.

W dużym i złożonym układzie może być potrzebne wiele różnych sygnałów zegarowych różniących się fazą, a nawet częstotliwością. Są one wytwarzane z głównego sygnału zegarowego, dostarczanego z zewnątrz lub generowanego w układzie. Sygnał o częstotliwości *n*-krotnie mniejszej od głównego, gdzie *n* jest całkowitą wielokrotnością 2, uzyskuje się bez trudu przez użycie dzielnika częstotliwości. Sygnał o częstotliwości *n*-krotnie wyższej zsynchronizowany z sygnałem głównym jest trudniej uzyskać. Potrzeba takiego sygnału występuje z reguły tylko wtedy, gdy sygnał główny jest dostarczany z zewnątrz. Do synchronizacji wykorzystuje się układy

pętli fazowej („phase-locked loop”, w skrócie PLL) dla synchronizacji wewnętrznego generatora sygnału o częstotliwości *nf* z zewnętrznym sygnałem o częstotliwości *f*. Układy PLL są skomplikowane, a ich projektowanie trudne. Temat ten wykracza poza ramy naszej opowieści o układach scalonych.

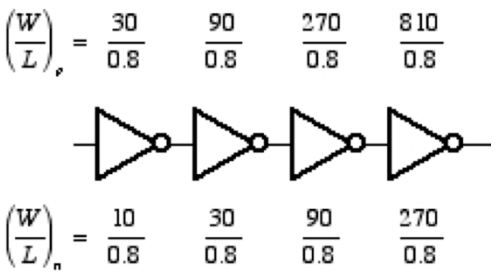
Częstym, a znanym nam już przypadkiem jest przypadek zegara dwufazowego o fazach „1” nie nakładających się. Takie dwa sygnały zegarowe można łatwo wygenerować przy użyciu układu pokazanego na rysunku 3-17.



Rysunek 3-17. Układ generujący dwa sygnały zegarowe o fazach "1" nie nakładających się

Odstęp czasu między stanami „1” obu sygnałów jest regulowany opóźnieniami wprowadzanymi przez inwertery wprowadzone w pętlę sprzężenia zwrotnego. W razie potrzeby można wprowadzić zamiast dwóch - cztery inwertery lub większą (zawsze parzystą) ich liczbę.

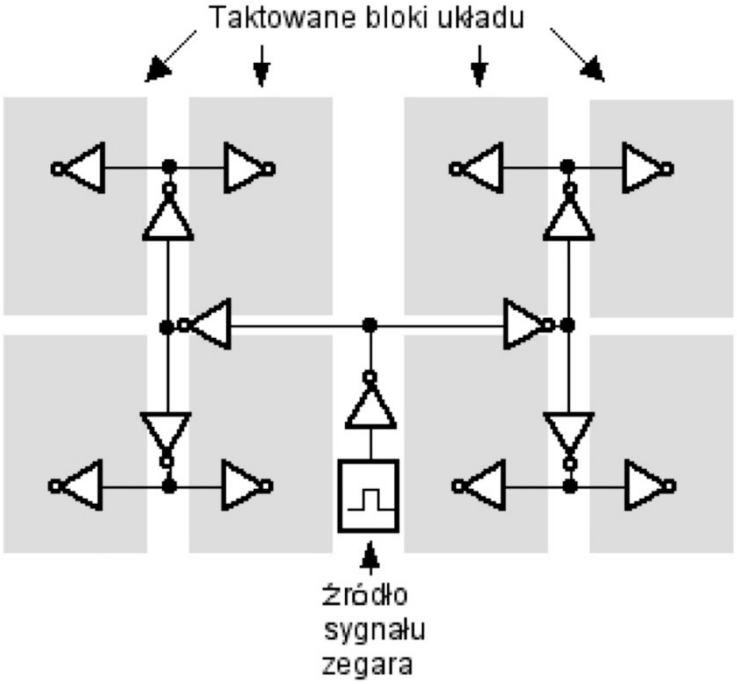
Rozprowadzanie sygnału zegara w dużych układach jest poważnym problemem technicznym. W idealnym przypadku dla każdego taktowanego danym sygnałem zegarowym elementu układu (jak przerzutnik, bramka transmisyjna, rejestr itp.) stany jedynki i zera zegara powinny zaczynać się i kończyć w dokładnie tych samych momentach. Tylko wtedy układ jest rzeczywiście synchroniczny. W praktyce spełnienie tego warunku jest trudne. W dużych układach sygnały zegara są doprowadzone równocześnie do bardzo wielu wejść zegarowych w przerzutnikach, bramkach dynamicznych itp., a to oznacza, że układ, który jest źródłem sygnału zegara, jest obciążony bardzo dużą pojemnością. Dlatego stosuje się bufory, których zadaniem jest zapewnić jak najkrótsze czasy narastania i opadania impulsów sygnału zegarowego mimo dużej pojemności obciążającej. Takim buforem jest zwykle inwerter o stosunku *W/L* tranzystorów wynoszącym kilkaset, a nawet kilka tysięcy. Tranzystory takie mają wiele równolegle połączonych kanałów. Przykład konstrukcji takich tranzystorów był pokazany w części II na rysunku 4-4 - obejrzyj duże tranzystory nMOS i pMOS znajdujące się przy górnej krawędzi bloku pokazanego na tym rysunku. Inwerter z takimi tranzystorami nie może być sterowany bezpośrednio z inwertera lub bramki o „zwykłych” wymiarach tranzystorów, tj, o stosunku *W/L* rzędu 1 ... 10, ponieważ bardzo duże tranzystory same stwarzają duże obciążenie pojemnościowe, zbyt duże dla zwykłego inwertera lub bramki. Bufory zwykle buduje się jako łańcuchy inwerterów o stopniowo rosnących stosunkach *W/L* tranzystorów. Liczba buforów w takim łańcuchu może być nieparzysta lub parzysta, w zależności o tego, czy chcemy uzyskać na wyjściu sygnał zanegowany, czy też nie. Przykład tak zbudowanego bufora pokazuje rysunek 3-18. Prostą praktyczną regułą jest trzykrotne powiększanie szerokości tranzystorów w kolejnych inwerterach bufora.



Rysunek 3-18. Czterostopniowy bufor zegara, pokazane przykładowe wymiary tranzystorów w kolejnych stopniach

Na zasadzie pokazanej na rysunku 3-18 buduje się nie tylko bufory dla sygnału zegara, ale także bufory dla wszystkich innych sygnałów, które trzeba dostarczyć do obciążenia o dużej pojemności wejściowej (a także małej rezystancji).

W dużych układach liczących miliony bramek całkowita pojemność wszystkich wejść zegarowych jest tak duża, że pojedynczy bufor musiałby mieć absurdalnie duży stosunek *W/L* tranzystorów. W takim przypadku układ dzieli się na fragmenty mające mniej więcej jednakowe pojemności wejść zegara i do każdego z tych fragmentów doprowadza się sygnał zegara przez system buforów tworzący strukturę drzewiastą. Pokazuje to rysunek 3-19.



Rysunek 3-19. Rozprowadzenie i buforowanie sygnału zegara przy zastosowaniu drzewa typu H

W strukturze pokazanej na rysunku 3-19 (zwanej drzewem typu H) sygnał zegara na drodze do każdego bloku przechodzi przez tę samą liczbę identycznych i jednakowo obciążonych buforów. Jeżeli na dodatek uda się w topografii układu zachować geometryczną identyczność wszystkich symetrycznych segmentów drzewa, to także długość drogi dla zegara jest w sensie odległości geometrycznej taka sama dla każdego bloku. Jeżeli wszystkie taktowane bloki układu mają jednakowe pojemności obciążające bufory zegara, to uzyskuje się dokładnie jednakowy czas propagacji sygnałów zegarowych z ich źródła do wszystkich bloków układu (mogą jednak istnieć różnice wynikające z lokalnych zaburzeń procesu produkcyjnego).

Szczególne trudności mogą powstać, jeśli w dużym układzie trzeba rozprowadzić dwa lub więcej sygnałów zegarowych o ściśle określonych zależnościach czasowych. Dotyczy to na przykład omawianego wcześniej zegara dwufazowego o fazach „1" nie nakładających się w czasie. Nawet niewielkie różnice w czasach propagacji tych dwóch sygnałów mogą spowodować, że w jakimś bloku układu jedynki należące do dwóch różnych sygnałów zegara nałożą się w czasie, co spowoduje błędne działanie bloku. Dlatego unika się rozprowadzania w dużych układach więcej niż jednego sygnału zegara. Znacznie bezpieczniej jest dodatkowe sygnały zegarowe generować lokalnie w blokach, w których są potrzebne.

Przy projektowaniu buforów zegara i sposobu rozprowadzenia zegara należy pamiętać, że celem zazwyczaj nie jest uzyskanie jak najmniejszego opóźnienia między źródłem tego sygnału, a taktowanymi blokami. Celem jest uzyskanie jak najkrótszych czasów narastanie i opadania impulsów zegarowych oraz jednakowego czasu propagacji tych impulsów do wszystkich taktowanych bloków. Osiągnięcie tego celu zapewnia prawidłową, synchroniczną pracę układu jako całości.

4 Pamięci i inne układy o strukturze matrycowej

Trudno znaleźć układ cyfrowy, który nie potrzebowałby pamięci. Potrzebne są zarówno pamięci o stałej zawartości, służące tylko do odczytu, jak i pamięci, których zawartość może być zmieniania. Będziemy mówić o jednych i o drugich. Omawiane będą zasady działania komórek pamięci, czyli układów przechowujących pojedyncze bity informacji oraz organizacja układów pamięci. Omawiane będą także inne układy o strukturze regularnej, które wprawdzie nie są zaliczane do pamięci, ale wykazują pewne do nich podobieństwo.

Projektant specjalizowanego układu scalonego na ogół nie projektuje pamięci samodzielnie. Jeżeli w projektowanym układzie potrzebna jest pamięć, jest ona zwykle generowana automatycznie. Jest to możliwe dlatego, że pamięci mają regularną i powtarzalną budowę, co pozwala łatwo zautomatyzować ich projektowanie. To stwierdzenie nie dotyczy pamięci dynamicznych (znanych jako pamięci dRAM) wytwarzanych jako osobne układy scalone. Pamięci takie wymagają specjalnej technologii, a ich projektowanie jest bardzo trudne. Zajmuje się tym tylko kilku wyspecjalizowanych producentów na świecie.

Nie będziemy rozważać szczegółów projektowania komórek pamięci ani też układów zapisu, odczytu i adresowania. Niemniej ogólne wiadomości o pamięciach potrzebne są każdemu, kto zajmuje się systemami cyfrowymi.

4.1 Rodzaje pamięci

4.1.1 Pamięci ROM i RAM

Tradycyjnie pamięci półprzewodnikowe są dzielone na:

- pamięci o stałej zawartości, przeznaczone tylko do odczytu („Read Only Memory” - ROM)
- pamięci o swobodnym dostępie, w których w każdej chwili można dokonać zarówno zapisu, jak i odczytu („Random Access Memory” - RAM).

Pamięci RAM dzielą się pod względem sposobu przechowywania informacji na dwa rodzaje: pamięci statyczne (ang. „static RAM”, sRAM) i dynamiczne (ang. „dynamic RAM”, dRAM). W pamięciach statycznych elementarną komórką pamięci przechowującą jeden bit jest przerzutnik statyczny. W pamięciach dynamicznych elementem pamięciowym jest kondensator, a stan jedynki lub zera jest pamiętany jako ładunek w tym kondensatorze lub jego brak.

Pamięci ROM dzielą się na znacznie więcej grup. Są wśród nich pamięci, których zawartość jest określana już w procesie produkcji i nie może być potem zmieniona, są pamięci, które można zaprogramować, ale tylko jeden raz, są pamięci, których zawartość można kasować w całości i programować od początku, i wreszcie są pamięci elektrycznie reprogramowalne, które są najbardziej zbliżone pod względem możliwych zastosowań do pamięci RAM.

Pamięci klasyfikowane jako RAM mają w porównaniu z pozostałymi rodzajami pamięci krótki czas dostępu, tj. czas, jaki upływa od zainicjowania procesu zapisu lub odczytu do zakończenia tego procesu. Pamięci klasyfikowane jako ROM, ale reprogramowalne, tj. takie, których zawartość można wielokrotnie zmieniać, potrzebują do zmiany zawartości znacznie dłuższego czasu niż pamięci RAM. Proces zapisu w tych pamięciach przebiega wolniej. Różnica szybkości zapisu w reprogramowalnych pamięciach ROM w porównaniu do pamięci RAM polega na zasadniczo odmiennym charakterze mechanizmów fizycznych, które określają zawartość pamięci. Zmiana zawartości komórki pamięci RAM oznacza zmianę stanu w układzie elektronicznym tworzącym tę komórkę, np. zmianę stanu przerzutnika statycznego. W strukturze komórki pamięci, w jej schemacie i w parametrach jej elementów nie zachodzą żadne zmiany. W przypadku reprogramowalnych pamięci ROM

zmiana zawartości komórki oznacza zmianę struktury fizycznej tej komórki lub właściwości jednego z jej elementów, np. wytworzenie nowego połączenia elektrycznego, usunięcie istniejącego lub zmiana napięcia progowego tranzystora. Dalej zobaczymy, na czym taka zmiana struktury fizycznej polega w różnych odmianach pamięci ROM.

4.1.2 Pamięci ulotne i nieulotne

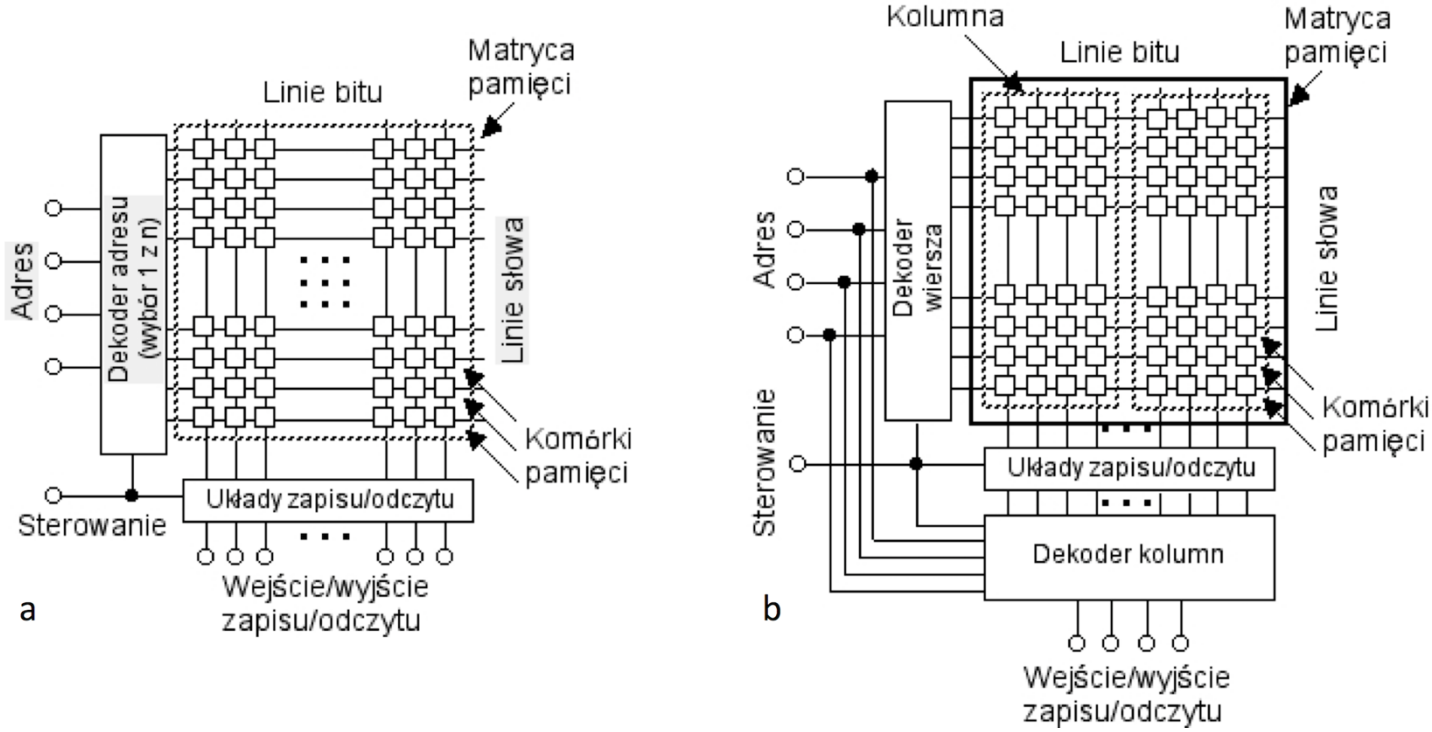
Istotny jest również inny podział: na pamięci ulotne i nieulotne. Pamięci klasyfikowane jako pamięci ulotne to takie, których zawartość ginie po wyłączeniu zasilania. Pamięci klasyfikowane jako pamięci nieulotne to takie, w których wyłączenie zasilania nie powoduje utraty zawartości.

Rozwój technologii wytwarzania pamięci doprowadził do sytuacji, w której granice między pamięciami ROM i RAM zaczęły ulegać zatarciu. Wiele rodzajów pamięci klasyfikowanych jako pamięci ROM, które są pamięciami nieulotnymi, umożliwia bowiem wielokrotne programowanie, a więc z punktu widzenia pełnionej funkcji zbliża się do pamięci RAM. Stopniowo zmniejsza się również czas zapisu w reprogramowalnych pamięciach ROM. Idealną pamięcią byłaby pamięć nieulotna o czasie zapisu i odczytu porównywalnym z pamięciami ROM. Prace nad takimi pamięciami trwają i zaczynają dawać obiecujące rezultaty, o czym będzie mowa dalej.

4.2 Budowa i zasady działania układów pamięci

4.2.1 Ogólna struktura układów pamięci

Ogólną zasadę budowy układów pamięci ilustruje rysunek 4-1a. Trzy główne bloki układu pamięci to matryca pamięci, dekoder adresów i blok układów zapisu/odczytu. Matryca pamięci składa się z pewnej liczby linii słowa (poziomych na rysunku 4-1) i krzyżujących się z nimi linii bitu (pionowych na rysunku 4-1). Na każdym skrzyżowaniu linii słowa z linią bitu znajduje się komórka pamięci pamiętająca jeden bit. Komórki pamięci zbudowane są różnie w różnych rodzajach pamięci. Mogą to być pojedyncze tranzystory lub układy bardziej złożone, zawierające kilka elementów.



Rysunek 4-1. Budowa układów pamięci: zasada ogólna (a) i organizacja pamięci o dużej pojemności (b)

Dekoder adresów to układ, który dla każdej wartości adresu (tj. kombinacji bitów na wejściu adresowym) uaktywnia do zapisu lub odczytu dokładnie jedną linię słowa (na ogół przez podanie na nią stanu „1”). Wszystkie komórki pamięci położone wzdłuż tej linii będą przedmiotem operacji zapisu lub odczytu. Dla słowa adresowego o długości *n* bitów matryca pamięci posiada 2*n* linii słowa.

Układy zapisu/odczytu są różne w różnych rodzajach pamięci, ale wszędzie pełnią tę samą rolę. Przy zapisie przetwarzają otrzymane z zewnątrz słowo binarne na wewnętrzne sygnały potrzebne do dokonania zapisu w komórkach pamięci. Przy odczycie odczytane z komórek pamięci stany logiczne (jak zobaczymy, nie zawsze reprezentowane przez napięcia równe 0 i V_{DD}) są przetwarzane na słowo binarne podawane na wyjście. Liczba linii bitu decyduje o organizacji pamięci, np. przy 8 liniach zapisywane i odczytywane są słowa ośmiobitowe.

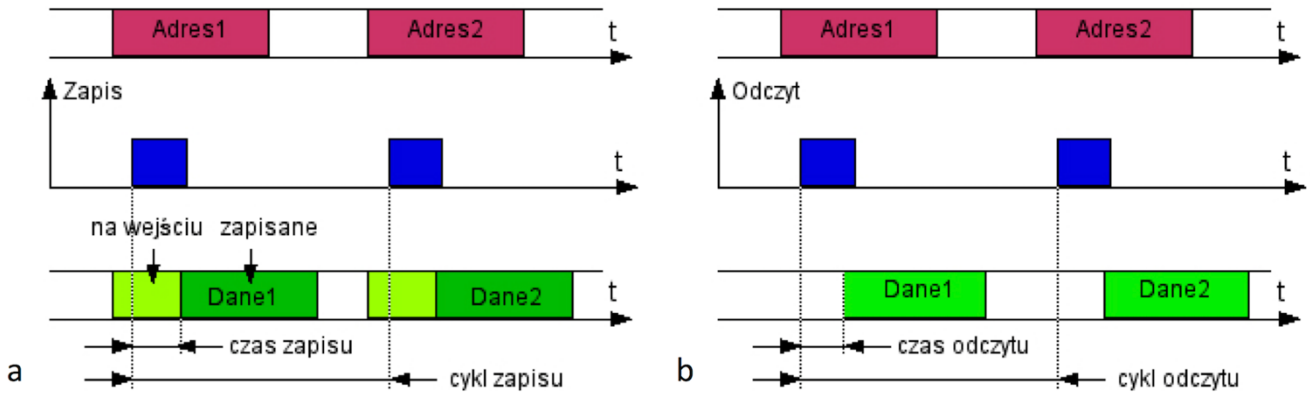
Sterowanie (pokazane symbolicznie jako wejście jednobitowe) określa funkcję wykonywaną w danej chwili przez pamięć, np. zapis lub odczyt.

Najprostsza organizacja pamięci pokazana na rysunku 4-1a staje się niepraktyczna, a nawet niemożliwa do realizacji w przypadku pamięci o dużej pojemności. Linie bitu nie mogą być dowolnie długie i nie można do nich dołączać dowolnie dużej liczby komórek pamięci. Jak zobaczymy dalej, długość linii bitu i liczba dołączonych do niej komórek pamięci ma bezpośredni wpływ między innymi na czas zapisu i odczytu. W pamięciach o dużej pojemności matryca jest dzielona na mniejsze części. Przykład takiej organizacji pamięci pokazuje rysunek 4-1b. Zapisywane i odczytywane są słowa czterobitowe. Matryca jest podzielona na czterobitowe kolumny i na podstawie adresu wybierana jest nie tylko linia słowa, ale i kolumna. Jeśli kolumn jest *n*, to przy danej pojemności pamięci linie bitu mogą być *n* razy krótsze.

Dla bardzo dużych pamięci stosuje się podział na mniejsze bloki będące kompletnymi matrycami, które z kolei dzielą się na kolumny. Komplikuje to dekodowanie adresów, ale pozwala zachować rozsądną długość linii bitu.

4.2.2 Zależności czasowe w pamięciach

Zanim przejdziemy do omawiania budowy i zasad działania poszczególnych rodzajów pamięci, poznamy w zarysie zależności czasowe przy zapisie i odczycie.



Rysunek 4-2. Zależności czasowe przy zapisie (a) i odczycie (b)

Zarówno przy zapisie, jak i przy odczycie trzeba najpierw podać adres na wejście adresowe, a w przypadku zapisu także dane, które mają być zapisane. Potem na wejście sterujące podawany jest sygnał startu zapisu lub odczytu. Przy zapisie upływa pewien czas, zanim dane zostaną zapisane (na rysunku 4-2 czas zapisu). Przy odczycie upływa pewien czas, zanim odczytane dane pojawią się na wyjściu (czas odczytu). Cykl zapisu to najkrótszy odcinek czasu, po którym można ponownie dokonać zapisu. Cykl odczytu to najkrótszy odcinek czasu, po którym można ponownie dokonać odczytu. Cechą charakterystyczną pamięci klasyfikowanych jako RAM są krótkie i zbliżone do siebie czasy zapisu i odczytu. W przypadku reprogramowalnych pamięci ROM regułą jest, że czas odczytu może być krótki (porównywalny do czasu odczytu z pamięci RAM), natomiast czas zapisu jest znacznie dłuższy.

Proste schematy zależności czasowych pokazane na rysunku 4-2 służą jedynie do ilustracji zasadniczych pojęć - czasu zapisu, czasu odczytu, cyklu zapisu, cyklu odczytu. Rzeczywiste sekwencje sygnałów przy zapisie i odczycie mogą być bardziej skomplikowane. Przykładowo, istnieją konstrukcje pamięci statycznych RAM (a także pamięci ROM), w których proces zapisu i odczytu inicjowany jest przez zmianę wartości adresu. Nie jest potrzebny odrębny sygnał startu. Z kolei w pamięciach dynamicznych stosowany jest bardziej złożony zbiór sygnałów sterujących, zależny od wewnętrznej organizacji, na przykład sposób adresowania, w którym na wejście adresowe podawana jest najpierw część adresu określająca linię słowa, a potem część określająca wybór kolumny. W takim przypadku potrzebne są sygnały synchronizujące, które inicjują najpierw wybór wiersza (linii słowa), potem kolumny, a na koniec uruchamiają proces odczytu.

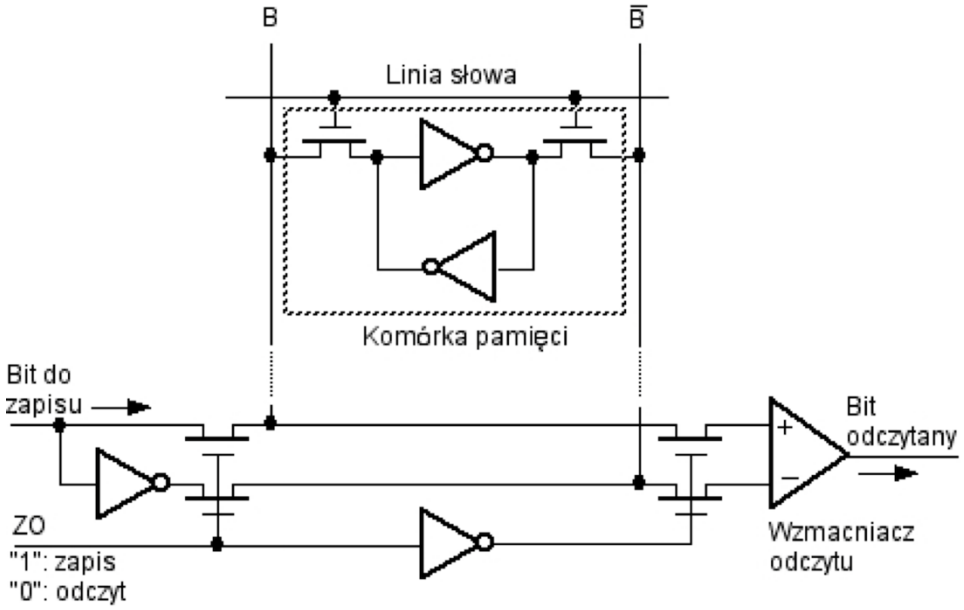
Jeszcze bardziej złożone sposoby sterowania działaniem pamięci występują przy najszybszych dynamicznych pamięciach RAM, zwanych synchronicznymi. Pamięci takie stosuje się na płytach głównych współczesnych komputerów. Działanie tych pamięci jest synchronizowane zegarem. Przy zapisie dane muszą być dostępne na wejściu danych w tym samym cyklu zegara, w którym następuje zapis, natomiast przy odczycie dane są podawane na wyjście z opóźnieniem kilku cykli zegara w stosunku do cyklu, w którym podany został sygnał odczytu. Ta liczba cykli (w oznaczeniu pamięci określana symbolem CL, ang. „cycle latency") wynosi od 2 - 3 przy częstotliwościach zegara rzędu 100 MHz aż do 7 - 9 przy częstotliwościach rzędu 1 GHz. Organizacja wewnętrzna pamięci umożliwia odczyt potokowy: podczas oczekiwania na transfer pierwszej porcji odczytanych danych można przesyłać do pamięci kolejne sygnały odczytu, dzięki temu na odczyt następnych porcji danych nie będzie już trzeba czekać równie długo. Ponadto wszystkie współcześnie produkowane pamięci dynamiczne umożliwiają dwa transfery danych na jeden takt zegara, reagując zarówno na rosnące, jak i opadające zbocze sygnału zegarowego. Takie pamięci oznaczane są symbolem DDR (ang. „Double Data Rate"). Oto przykład zależności czasowych dla szybkiej pamięci dynamicznej oznaczanej jako DDR3 1600 CL7 (częstotliwość zegara 800 MHz): przy kolejnym odczycie 8 słów pierwsze pojawia się po 8,75 nanosekundy, ostatnie już po 13,125 nanosekundy. Mimo tych udoskonaleń czas odczytu z pamięci jest jednym czynników poważnie ograniczających wydajność obliczeniową komputerów. Dlatego powszechnie stosowane są podręczne pamięci (ang. cache) o mniejszych pojemnościach, ale znacznie szybsze, które służą do przechowywania danych często i wielokrotnie wykorzystywanych przez procesor. Te pamięci są realizowane jako statyczne pamięci RAM.

Szczegółowe omawianie wielu istniejących sposobów organizacji pamięci i sterowania ich pracą wykracza poza zakres tych materiałów. Producenci pamięci podają wszystkie potrzebne informacje w katalogowej dokumentacji technicznej.

4.2.3 Statyczna pamięć RAM

Komórką pamięci statycznej RAM jest podstawowy przerzutnik statyczny omawiany w punkcie 3.2.1, uzupełniony o dwa tranzystory nMOS pełniące rolę bramek transmisyjnych, które służą do wyboru danej komórki do zapisu i odczytu, w zależności od stanu linii słowa. Schemat takiej sześciotranzystorowej komórki pamięci wraz z (pokazanymi w pewnym uproszczeniu) układami zapisu/odczytu pokazuje rysunek 4-3. Jak pokazuje ten rysunek, typowa komórka pamięci statycznej jest połączona poprzez tranzystory nMOS z dwoma liniami bitu, na których pojawia się zapisywany lub odczytywany bit oraz jego negacja. Gdy linia słowa jest w stanie „1", tranzystory nMOS są włączone i komórka komunikuje się z obydwoma liniami bitu. Gdy linia słowa jest w stanie „0", tranzystory nMOS są wyłączone. Przerzutnik statyczny w komórce pamięci trwa w stanie, jaki został ostatnio zapisany. Zapis wymaga wyboru komórki przez podanie „1" na linię słowa oraz podania „1" na wejście sterujące ZO. Otwierają się wówczas bramki transmisyjne zapisu (na schemacie po lewej stronie), i bit do zapisu oraz jego negacja są podawane na linie bitu i jego negacji. Równoczesne podawanie bitu i jego negacji przyspiesza proces zmiany stanu przerzutnika, ponieważ nowy stan podawany jest równocześnie na wejścia obu inwerterów przerzutnika. Przy odczycie (ZO w stanie „0") otwarte są bramki transmisyjne odczytu (z prawej

strony na schemacie), a napięcia z linii bitu i jego negacji są podawane na wejście różnicowego wzmacniacza odczytu. Ten wzmacniacz jest to komparator napięcia porównujący napięcia na linii bitu i negacji bitu. Jest to układ analogowy (będzie omawiany bardziej szczegółowo dalej). Wzmacnia on różnicę napięć między linią bitu i jego negacji. Zastosowanie tego wzmacniacza przyspiesza proces odczytu. Przy odczycie napięcia na liniach bitów zmieniają się stopniowo (potrzebny jest czas na ładowanie lub rozładowywanie pojemności tych linii). Ale już bardzo mała różnica napięć na wejściach wystarcza, aby na wyjściu pojawiło się pełne napięcie V_{DD} lub 0 (w zależności od stanu logicznego odczytywanego z komórki).



Rysunek 4-3. Komórka statycznej pamięci RAM wraz z układami zapisu/odczytu

Komparator napięcia jako wzmacniacz odczytu dokonuje zarazem regeneracji poziomu jedynki (jeśli odczytywana jest jedynka). Jak widać na rysunku 4-3, odczyt z komórki pamięci odbywa się poprzez bramkę transmisyjną w postaci pojedynczego tranzystora nMOS. Jak już wiemy, powoduje to degradację poziomu jedynki (punkt 2.3.1). Dla oszczędności powierzchni pełnych bramek transmisyjnych CMOS w komórkach pamięci nie używa się.

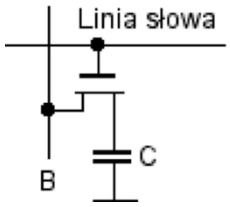
Pamięci statyczne są najszybciej działającymi pamięciami typu RAM. Są one jednak dość kosztowne i mają ograniczoną pojemność, ponieważ komórki tych pamięci, liczące aż 6 tranzystorów, zajmują dużo miejsca. Zaletą pamięci statycznych jest to, że nie wymagają one do produkcji specjalnej technologii wytwarzania. Mogą być zatem bez trudności stosowane jako na przykład części składowe mikroprocesorów (pamięci podręczne typu "cache"). Jednym z głównych czynników ograniczających szybkość działania pamięci statycznych jest konieczność ładowania lub rozładowywania dużej pasożytniczej pojemności linii bitu. Dlatego, jak już wcześniej mówiliśmy, liczba komórek dołączonych do linii bitu nie może być zbyt duża. Dla pamięci o większej pojemności stosuje się omawiane wcześniej bardziej złożone schematy organizacji pamięci.

W praktyce projektowania układów ASIC bardzo rzadko zdarza się potrzeba zaprojektowania pamięci statycznej od początku, tj. zaprojektowania komórek, całej matrycy dekodera adresów oraz układów wejścia/wyjścia. Profesjonalne systemy projektowania dysponują możliwością automatycznej generacji projektu pamięci o zadanej pojemności i organizacji. Potrzebne do tego dane, schematy i topografie komórek, układów wejścia/wyjścia itp. dostarczają producenci układów.

4.2.4 Dynamiczna pamięć RAM

Pamięci dynamiczne RAM są najbardziej znanym rodzajem pamięci półprzewodnikowych, ponieważ to właśnie te pamięci są powszechnie stosowane jako pamięci operacyjne komputerów. Zarówno schemat (rysunek 4-4),

jak i zasada działania komórki takiej pamięci są bardzo proste. Proces zapisu polega na ładowaniu lub rozładowywaniu pojemności C. Gdy linia słowa jest w stanie „1”, a na linię bitu podane jest napięcie V_{DD}, pojemność C ładuje się poprzez tranzystor, który jest włączony. Gdy zaś na linii bitu panuje napięcie równe zeru, pojemność C ulega rozładowaniu. W ten sposób zapisywana jest jedynka lub zero. Przy odczycie napięcie z kondensatora podawane jest przez włączony tranzystor na linię bitu.



Rysunek 4-4. Komórka pamięci dynamicznej RAM

Ta prosta zasada jest jednak bardzo trudna do praktycznego wykorzystania. Problem stwarza zjawisko podziału ładunku przy odczycie (patrz punkt 3.1.1). Pojemność pasożytnicza linii bitu, do której dołączone jest bardzo wiele komórek pamięci, jest wielokrotnie większa od pojemności kondensatora C w pojedynczej komórce. W rezultacie przy odczycie jedynki odczytywane z komórki pamięci napięcie jest wielokrotnie mniejsze od napięcia V_{DD}. Typowa wartość odczytywanego napięcia jedynki to kilkadziesiąt miliwoltów. Po odczycie napięcie na kondensatorze pozostaje na tym poziomie, toteż następny odczyt jedynki nie jest już możliwy. Wynika z tego, że:

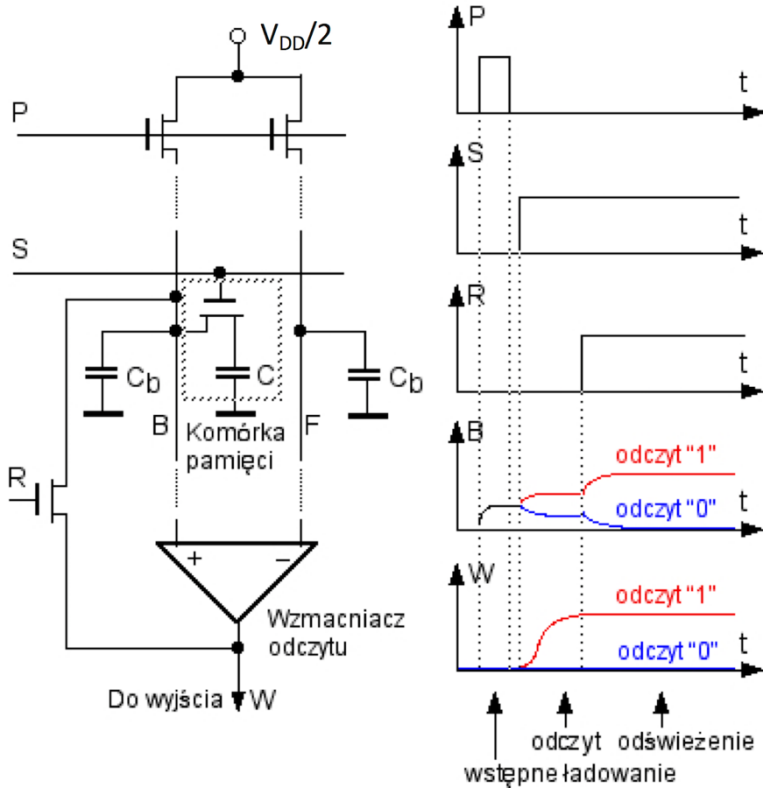
- odczyt wymaga zastosowania wzmacniacza regenerującego właściwy poziom logiczny,
- po odczycie konieczne jest odświeżenie zawartości komórki przez ponowny zapis do niej odczytanego stanu logicznego.

Dlatego układy zapisu/odczytu dla pamięci dynamicznych są dużo bardziej skomplikowane niż w przypadku pamięci statycznych. Rysunek 4-5 ilustruje w uproszczeniu układ odczytu oraz przebieg procesu odczytu i odświeżenia zawartości komórki. Wykorzystuje się tu zasadę wstępnego ładowania, znaną nam już z bramek dynamicznych typu DOMINO (punkt 3.1.2), ale tutaj wykorzystaną w nieco inny sposób.

Każdej linii bitu B towarzyszy fałszywa linia bitu F, do której nie są podłączone żadne komórki pamięci, ale która ma pojemność pasożytniczą o wartości takiej samej (lub zbliżonej), jak linia bitu B. Przed odczytem obie linie są wstępnie ładowane do napięcia równego w przybliżeniu połowie V_{DD}. Odbywa się to z chwilą podania jedynki na linię P. Po zakończeniu wstępnego ładowania na liniach B i F panuje jednakowe napięcie podtrzymywane dzięki pojemnościom pasożytniczym tych linii C_b. Następnie podawana jest jedynka na linię słowa S. Dzięki temu pojemność C komórki pamięci zostaje dołączona przez tranzystor do linii bitu B. Jeżeli w komórce zapamiętana była jedynka (czyli napięcie na kondensatorze było równe lub bliskie V_{DD}), ładunek z kondensatora C doładowuje pojemność pasożytniczą linii bitu C_b i powoduje podskok napięcia o wartości kilkudziesięciu miliwoltów. Napięcie na linii F nie zmienia się, powstaje więc różnica napięć na wejściu wzmacniacza odczytu. Jest to, podobnie jak w układach pamięci statycznej, komparator napięcia. Jest tak skonstruowany, że na jego wyjściu pojawia się pełne napięcie jedynki, czyli V_{DD}, gdy $V_B > V_F$, a napięcie równe zeru, jeśli $V_B \leq V_F$.

Stan z wyjścia wzmacniacza odczytu podawany jest na wyjście z pamięci, a także – po podaniu jedynki na bramkę tranzystora oznaczoną R – na linię bitu B. W rezultacie następuje ponowne naładowanie kondensatora C komórki pamięci do napięcia reprezentującego jedynkę. Pokazuje to przebieg napięcia na linii B (rysunek 4-5, czerwona linia). Jeśli natomiast w chwili odczytu w komórce zapisane jest zero, czyli kondensator C jest rozładowany, to ładunek odpływa do niego z linii bitu, co powoduje spadek napięcia na tej linii o kilkadziesiąt miliwoltów. Różnica napięć na wejściu wzmacniacza różnicowego powoduje pojawienie się na jego wyjściu napięcia równego zeru. To napięcie, czyli „0”, jest podawane na wyjście, a także - poprzez tranzystor R - na linię

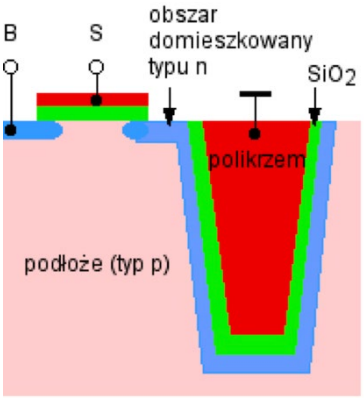
bitu. Napięcie na tej linii spada do zera, co powoduje także rozładowanie do zera pojemności C w komórce pamięci. W ten sposób w komórce pamięci pozostaje zapisane zero. Przebieg napięcia na linii bitu pokazany jest na rysunku 4-5 niebieską linią.



Rysunek 4-5. Zasada odczytu z komórki pamięci dynamicznej. Przebieg pokazany linią czerwoną dla przypadku, gdy w komórce zapisana była jedynka, linią niebieską - gdy zapisane było zero

Każdy akt odczytu z komórki pamięci powoduje więc zarazem odświeżenie jej zawartości. Komórki wymagają okresowego odświeżania zawartości nawet wtedy, gdy w pamięci nic się nie dzieje. Ładunek w kondensatorze C ulega bowiem powolnemu zanikowi w wyniku istnienia prądów upływu (prąd podprogowy tranzystora komórki, prąd wsteczny złącza *p-n* drenu). W celu odświeżania wystarczy okresowo odczytywać wszystkie komórki.

Jak widać, działanie pamięci dynamicznej przy odczycie jest dość skomplikowane. Z tego powodu pamięci dynamiczne działają wolniej niż statyczne. Ich wielką zaletą jest bardzo mała powierzchnia zajmowana przez pojedynczą komórkę, zawierającą tylko jeden tranzystor i jeden kondensator. Ta mała powierzchnia umożliwia produkcję pamięci o wielkiej pojemności i niewygórowanej cenie. W roli kondensatora C w komórce pamięci nie wystarczają pojemności pasożytnicze. Chodzi bowiem o to, by ładunek zgromadzony w tej pojemności był możliwie duży, tak aby zmiana napięcia na linii bitu przy odczycie była dostatecznie duża i umożliwiała bezbłędne zadziałanie wzmacniacza odczytu mimo zjawiska podziału ładunku (patrz punkt 3.1.1). Tę możliwie dużą pojemność C należy jednak zmieścić na możliwie jak najmniejszej powierzchni. Do tego celu opracowano specjalne technologie wytwarzania układów pamięci, w których kondensatory komórek pamięci są wykonywane w postaci głębokich wnęk wytrawionych w krzemie, których zbocza są pokryte bardzo cienkim dielektrykiem (SiO_2), a wnętrze wypełnione polikrzemem (stosowane są też inne sposoby wytwarzania tych kondensatorów). Przekrój przez komórkę z takim kondensatorem pokazuje rysunek 4-6. Okładki kondensatora tworzą: polikrzem wypełniający wnękę oraz obszar domieszkowany typu *n* w podłożu.



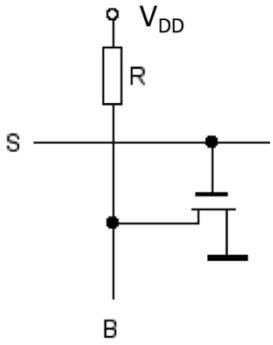
Rysunek 4-6. Przekrój przez komórkę pamięci dynamicznej

Nawet stosując taki specjalny kondensator jako element pamięciowy nie unikamy omawianego wyżej problemu podziału ładunku. Sztuka projektowania i produkcji pamięci dynamicznych polega nie tylko na tym, by uzyskać jak największą wartość pojemności komórki C, ale także na tym, by pojemność pasożytnicza linii bitu C_b była możliwie mała. Ogranicza to liczbę komórek pamięci, jaką można dołączyć do jednej linii bitu. Z tego powodu w pamięciach dynamicznych regułą jest stosowanie podziału matrycy pamięci na wiele stosunkowo niedużych bloków. O takiej organizacji pamięci była już mowa wcześniej.

Operacje głębokiego trawienia wnęk, utleniania ich zboczy, wypełniania ich polikrzemem nie są typowe dla zwykłej technologii wytwarzania układów CMOS. Pamięci są produkowane na specjalnie do tego przeznaczonych liniach produkcyjnych. Dlatego w zwykłych układach CMOS pamięci dynamiczne nie są spotykane. Produkuje się je jako samodzielne układy scalone. Toteż warto wiedzieć, jak działają, ale ich projektowania nie będziemy omawiać.

4.2.5 Pamięci nieulotne CMOS programowane jednokrotnie

W układach CMOS podstawową komórką pamięci nieulotnych był dotąd pojedynczy tranzystor nMOS (rysunek 4-7). O różnych rodzajach pamięci z takimi komórkami będzie teraz mowa. W ostatnim czasie zaczęły się jednak pojawiać pamięci nieulotne, które wykorzystują jako nośnik informacji cyfrowej spin elektronu (czyli w praktyce stan namagnesowania pewnej warstwy magnetycznej). Będzie o nich mowa w punkcie 4.2.7.



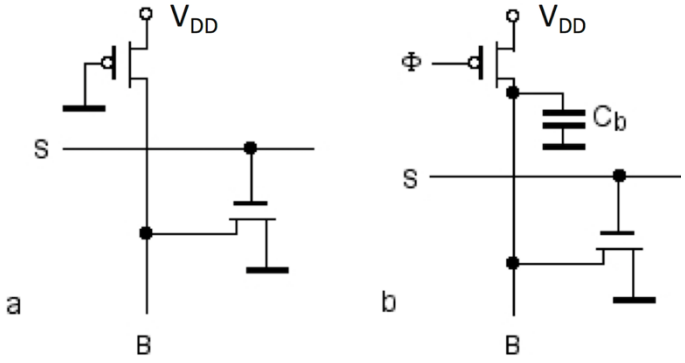
Rysunek 4-7. Komórka pamięci nieulotnej CMOS - zasadnicza idea

Działanie takiej komórki jest niezwykle proste. Załóżmy, że na linii słowa S panuje stan „1”. Jeśli tranzystor przewodzi, to zwiera linię bitu do zera napięcia zasilania, co oczywiście oznacza stan „0”. Jeśli tranzystor nie przewodzi lub go w ogóle nie ma, na linii bitu jest niezerowe napięcie reprezentujące stan „1”. Widoczny na rysunku 4-7 rezystor R jest symbolem elementu, który doprowadza napięcie ze źródła zasilania do linii bitu. W rzeczywistości jednak rezystorów się nie stosuje, ponieważ użycie zwykłej liniowej rezystancji nie dałoby możliwości uzyskania pamięci o dobrych parametrach.

Różne rodzaje pamięci nieulotnych różnią się głównie następującymi cechami:

- w jaki sposób doprowadzane jest niezerowe napięcie do linii bitu,
- w jaki sposób osiąga się stan przewodzenia lub nieprzewodzenia tranzystora stanowiącego komórkę pamięci.

Doprowadzenie niezerowego napięcia do linii bitu można najprościej uzyskać stosując tranzystor pMOS spolaryzowany w taki sposób, by pełnił rolę nieliniowej rezystancji. Połączenie bramki z zerem napięcia zasilania powoduje, że tranzystor pMOS jest zawsze włączony. Ilustruje to rysunek 4-8a. Gdy tranzystor nMOS nie przewodzi, na linii bitu panuje napięcie V_{DD}. Gdy tranzystor nMOS przewodzi, tworzy się dzielnik napięcia - przewodzą oba tranzystory. Dobierając odpowiednio ich wymiary można osiągnąć na linii bitu napięcie dostatecznie niskie, aby odpowiednio zaprojektowany wzmacniacz odczytu (który nie jest pokazany na rysunku 4-8) zinterpretował je jako stan „0". Ten sposób doprowadzenia napięcia do linii bitu jest bardzo prosty, ale ma istotną wadę: pamięć statycznie pobiera prąd, gdy oba tranzystory przewodzą.



Rysunek 4-8. Doprowadzenie napięcia do linii bitu: (a) statycznie, (b) dynamicznie

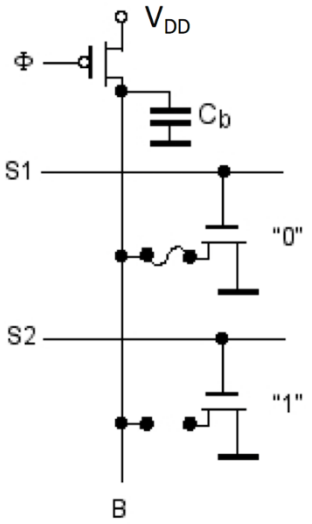
Lepszym rozwiązaniem jest zastosowanie wstępnego ładowania – rysunek 4-8b. Tu tranzystor pMOS wstępnie ładuje linię bitu (jej pojemność C_b) do napięcia V_{DD}, gdy zegar Φ jest w stanie „0". Po przejściu do stanu „1" tranzystor pMOS zostaje wyłączony. Teraz możliwy jest odczyt. Na linię słowa podawana jest „1". Jeśli tranzystor nMOS przewodzi, pojemność C_b rozładowuje się i odczytywane jest „0". W przeciwnym razie napięcie V_{DD} na pojemności C_b pozostaje i odczytywana jest „1". Tę zasadę działania już poznaliśmy, gdy omawiane były bramki typu DOMINO.

Omawianie pamięci nieulotnych zaczniemy od pamięci, których zawartość określona jest już podczas produkcji i nie może być potem zmieniona. Takie pamięci bywają też określane jako pamięci programowane maską, bowiem rzeczywiście ich zawartość jest określona przez jedną lub kilka masek fotolitograficznych, na przykład maskę określającą połączenia w warstwie metalu 1. W najprostszym przypadku tam, gdzie ma być zapisana jedynka, może w ogóle nie być tranzystora, lub też może on być odłączony od linii bitu.

Pamięci programowane maską mają zastosowanie tam, gdzie wiadomo na pewno, że nigdy nie będzie potrzeby zmiany zapisanej w pamięci informacji. Przykładem może być program sterujący działaniem mikroprocesora w prostych urządzeniach powszechnego użytku: kalkulatorach, pralkach, sprzęcie audio i TV. Pamięci programowane maską są najtańszym rodzajem pamięci nieulotnych przy produkcji masowej. Pamięci programowane maską są nieprzydatne, jeśli potrzebna jest mała liczba układów pamięci (bo wykonywanie nowej maski jest kosztowne) lub jeśli przewidujemy, że zawartość pamięci może wymagać zmian. W takim przypadku potrzebne są pamięci, które można zaprogramować – jeden raz lub wielokrotnie.

Jednym ze sposobów programowania pamięci nieulotnych jest przepalanie połączeń. Odłączanie tranzystorów odbywa się poprzez doprowadzenie odpowiedniego sygnału elektrycznego, a nie w procesie produkcyjnym.

Połączenie drenu z linią bitu wykonane jest w specjalny sposób (na przykład jest to bardzo wąska ścieżka polikrzemowa). Bramkę tranzystora oraz linię bitu trzeba spolaryzować napięciem znacznie wyższym od normalnego napięcia V_{DD}. Powoduje to przepływ prądu o na tyle dużym natężeniu, że połączenie drenu z linią bitu ulega przepaleniu. Tę zasadę ilustruje rysunek 4-9.

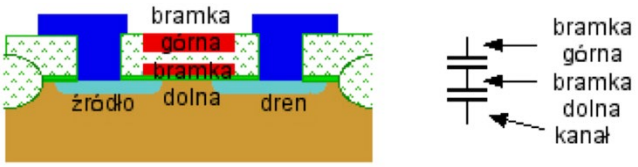


Rysunek 4-9. Pamięć jednokrotnie programowalna przez przepalanie połączeń

Takie pamięci mogą być oczywiście zaprogramowane tylko jeden raz. Były one swego czasu popularne, ale obecnie są już praktycznie wyparte przez pamięci wielokrotnie programowalne, znane pod ogólną nazwą EPROM (ang. Electrically Programmable ROM).

4.2.6 Reprogramowalne pamięci nieulotne CMOS

Pamięci wielokrotnie programowalne działają na zasadzie kontrolowanej zmiany napięcia progowego tranzystora. Zmianę napięcia progowego uzyskuje się na drodze elektrycznej, a nie w procesie produkcyjnym. Potrzebny jest do tego tranzystor o specjalnej, dwubramkowej konstrukcji. Idea takiego tranzystora pokazana jest na rysunku 4-10.



Rysunek 4-10. Dwubramkowy tranzystor nMOS stosowany w reprogramowalnych pamięciach EPROM

Bramka górna jest podłączona do linii słowa, natomiast bramka dolna nie jest nigdzie podłączona i ze wszystkich stron otoczona dielektrykiem (SiO_2). Na kanał oddziałuje potencjał bramki dolnej. Bramka górna i dolna tworzą pojemnościowy dzielnik napięcia. Jeśli pojemności bramka górna-bramka dolna oraz bramka dolna-kanał są jednakowe, to na nośniki w kanale oddziałuje połowa napięcia przyłożonego do bramki górnej. Istotą działania omawianego tranzystora jako komórki pamięci jest możliwość trwałego naładowania bramki dolnej ładunkiem elektrycznym. W tym celu trzeba w kanale wytworzyć nośniki o wysokiej energii (t.zw. gorące nośniki). Takie nośniki powstają na przykład w procesie powielania lawinowego (patrz część I, punkt 3.1.2) w złączu *p-n* dren-podłoże. Gorące nośniki mogą pokonać barierę dielektryka dzięki zjawisku tunelowania i dotrzeć do bramki dolnej, gdzie zostają uwięzione. Tunelowanie jest zjawiskiem kwantowym, polegającym na tym, że – w dużym uproszczeniu – nośniki ładunku w pewnych warunkach mogą przenikać przez bardzo cienkie warstwy dielektryczne.

Jeśli bramka dolna zostanie naładowana ujemnym ładunkiem elektronów, to do bramki górnej trzeba przyłożyć znacznie wyższe napięcie dodatnie, by włączyć tranzystor. Innymi słowy, poprzez ładowanie bramki dolnej

można zmienić napięcie progowe tranzystora „widziane” przez bramkę górną. Napięcie to może się stać tak duże, że tranzystora nie da się włączyć zwykłym napięciem przyłożonym do bramki górnej. W ten sposób programuje się tranzystor służący jako komórka pamięci EPROM. Ładunek zgromadzony w bramce dolnej ma czas rozładowania rzędu 10 lat. Jednak możliwe jest rozładowanie tego ładunku przy pomocy naświetlenia układu pamięci silnym światłem ultrafioletowym (nośniki są uwalniane dzięki zjawisku fotoelektrycznemu). W ten sposób zawartość pamięci może być skasowana, a pamięć zaprogramowana powtórnie.

W nowszych układach pamięci EPROM, zwanych EEPROM (ang. Electrically Erasable Programmable ROM) zmiana zawartości pamięci może być dokonana także na drodze czysto elektrycznej. Do uwalniania nośników z bramki dolnej wykorzystuje się zjawisko znane jako tunelowanie Fowlera-Nordheima. Występuje ono w ultracienkich warstwach dielektrycznych w silnym polu elektrycznym. Aby uzyskać ten efekt, warstwa dielektryczna SiO_2 pod dolną bramką musi być niezwykle cienka (rzędu 10 nanometrów lub mniej). Aby wywołać prąd tunelowy Fowlera-Nordheima, trzeba do bramki górnej przyłożyć znaczne napięcie ujemne (kilkanaście V). Wywołuje to ucieczkę ładunku ujemnego z dolnej bramki. Programowanie, a w przypadku pamięci o organizacji NAND także odczyt, wymaga napięć dodatnich i ujemnych o wartościach wynoszących kilkanaście V. Napięcia te są generowane wewnątrz układu pamięci przy zastosowaniu układów pomocniczych zwanych pompami ładunkowymi. Zasada wytwarzania podwyższonego napięcia w pompie ładunkowej polega na ładowaniu kilku połączonych szeregowo kondensatorów, każdego z nich napięciem bliskim napięciu zasilania. Dzięki połączeniu szeregowemu napięcia się sumują i w ten sposób otrzymuje się napięcie wyższe od napięcia zasilania. Działanie pomp ładunkowych jest dość skomplikowane, i nie będzie tu omawiane. Z punktu widzenia użytkownika układ pamięci ma jedno typowe napięcie zasilania, np. 5 V lub 3,3 V.

Pamięci EEPROM wymagają specjalnej technologii wytwarzania dwubramkowych tranzystorów. Niewielu producentów oferuje technologie CMOS, w których można w tym samym układzie obok zwykłego układu cyfrowego CMOS wyprodukować moduł pamięci typu EEPROM. Jeśli taka możliwość istnieje, to projekt pamięci o zadanej pojemności i organizacji jest zwykle wykonywany automatycznie przy wykorzystaniu gotowych, opracowanych przez producenta bloków zapisu, odczytu, adresowania i samych komórek pamięci.

Programowanie pamięci EEPROM było dawniej procesem bardzo powolnym. Dziś istnieją pamięci tego rodzaju (zwane pamięciami „flash”), które mają czas programowania na tyle krótki, że można je traktować jako specjalny rodzaj pamięci o swobodnym dostępie (RAM). Mają one dłuższy od pamięci statycznych i dynamicznych czas zapisu, ale za to są nieulotne - do podtrzymania ich zawartości nie jest potrzebne żadne źródło zasilania. Są dziś powszechnie stosowane na przykład w popularnych kartach pamięci stosowanych w cyfrowych aparatach fotograficznych, telefonach komórkowych i innych urządzeniach, w których potrzebna jest pamięć nieulotna o możliwie krótkim czasie zapisu. Coraz częściej też zastępują twarde dyski w komputerach przenośnych (tzw. dyski SSD – ang. Solid-State Drive). Pamięci typu *flash* mają krótszy czas zapisu i odczytu, niż tradycyjne dyski twarde, a poza tym są odporne mechanicznie, bo nie zawierają precyzyjnych części ruchomych.

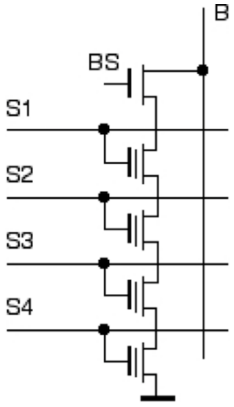
Pewną wadą pamięci tego rodzaju jest ich ograniczona trwałość. Procesy zapisu, które są związane z transportem przez cienką warstwę dielektryka nośników o wysokiej energii, powodują powolne pogarszanie się jakości warstwy dielektrycznej. Po określonej liczbie cykli zapisu komórka pamięci przestaje prawidłowo działać. Ta liczba cykli wynosi typowo do kilku tysięcy do kilkuset tysięcy. Nie należą więc do rzadkości przypadki, gdy intensywnie używana pamięć typu *flash* przestaje prawidłowo działać już po niedługim okresie użytkowania. Aby wydłużyć czas użytkowania pamięci, w blokach pamięci stosowanych jako zamiennik twardych dysków (dyskach SSD) stosowana jest systematyczna okresowa zmiana przyporządkowania komórek pamięci do adresów, co zapobiega szybszemu zużywaniu się komórek związanych z adresami wykorzystywanymi częściej,

niż inne. Dyski SSD mają też bloki komórek zapasowych (zwykle o łącznej pojemności 7% - 10% pojemności nominalnej dysku), zastępujących automatycznie bloki z komórkami, które uległy zużyciu.

Niektóre rodzaje pamięci typu "flash" mają organizację odmienną od omawianych dotąd rodzajów pamięci. Wyróżnia się dwa sposoby organizacji tych pamięci, noszą one nazwy „pamięci typu NOR” i „pamięci typu NAND”.

Pamięci typu NOR mają matryce zorganizowane tak samo, jak pamięci innych rodzajów, tj. na każdym skrzyżowaniu linii słowa z linią bitu znajduje się pojedynczy tranzystor. Dzięki temu przy odczycie jest natychmiastowy indywidualny dostęp do każdej komórki pamięci. Te pamięci są wykorzystywane tam, gdzie zawartość pamięci zmieniana jest stosunkowo rzadko, natomiast ważny jest jak najkrótszy czas odczytu. Tak może być na przykład w pamięciach zawierających mikroprogramy sterujące systemami cyfrowymi („firmware”, BIOS).

Pamięci typu NAND mają organizację pokazaną niżej: do linii bitu dołączony jest nie pojedynczy tranzystor, lecz łańcuch tranzystorów połączonych szeregowo (stąd nazwa – analogia do łańcuchów tranzystorów nMOS w statycznych bramkach NAND).



Rysunek 4-11. Zasada organizacji pamięci flash typu NAND

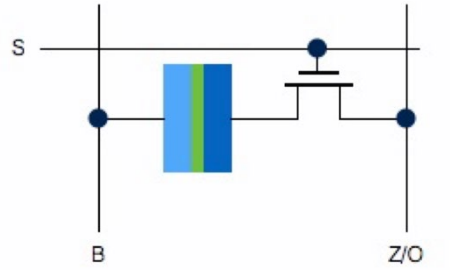
Każdy z dwubramkowych tranzystorów, których bramki dołączone są do linii słowa S1 ... S4, stanowi jedną komórkę pamięci. Łańcuch tych tranzystorów (może ich być więcej, niż cztery) dołączony jest do linii bitu poprzez zwykły tranzystor nMOS. Sygnał BS na bramce tego tranzystora umożliwia wybór danego łańcucha tranzystorów do zapisu lub odczytu. Przy odczycie linia słowa, dla której ma być dokonany odczyt, polaryzowana jest normalnym napięciem odczytu, a wszystkie pozostałe znacznie wyższymi napięciami - takimi, przy których tranzystory ulegają włączeniu – przewodzą – bez względu na to, czy zapisane w nich jest „0”, czy „1”. Tranzystor, z którego dokonywany jest odczyt, przewodzi lub nie, i dzięki temu cały łańcuch przewodzi lub nie. Podobnie adresowane są tranzystory przy zapisie. Organizacja pamięci typu NAND daje przy umiejętnym zaprojektowaniu topografii oszczędność powierzchni, a ponadto ułatwia kasowanie zawartości całych bloków pamięci.

4.2.7 Pamięci nieulotne MRAM w układach CMOS

Jest to szczególny rodzaj komórek pamięci, w których wykorzystuje się tunelowy efekt magnetorezystancyjny. Są one na pograniczu pamięci ROM i RAM, ponieważ są nieulotne, a równocześnie mają krótkie czasy zapisu i odczytu. Zasadniczą częścią komórki pamięci MRAM jest złącze składające się z trzech części: przewodzącego materiału magnetycznego namagnesowanego trwale, przewodzącego materiału magnetycznego podlegającego przemagnesowywaniu (magnetycznie „miękkiego”) oraz bardzo cienkiej warstwy dielektrycznej między nimi. Przez takie złącze może przepływać prąd dzięki zjawisku tunelowemu, przy czym prawdopodobieństwo przejść tunelowych zależy od tego, jakie są kierunki pola magnetycznego w obu magnetycznych obszarach. Gdy są zgodne, prąd tunelowy jest znaczny, w przeciwnym razie jest dużo mniejszy (rysunek 4-12).

Rysunek 4-12. Idea tunelowego złącza magnetorezystancyjnego

Komórka pamięci magnetorezystancyjnej składa się ze złącza tunelowego i jednego tranzystora:



Rysunek 4-13. Komórka pamięci STT-MRAM

Zapis do komórki MRAM polega na zmianie kierunku namagnesowania obszaru magnetycznie "miękkiego". Dokonuje się to przez przepuszczanie prądu między liniami bitu B i zapisu/odczytu Z/O w jednym lub drugim kierunku. Zapis w komórce można w pewnym uproszczeniu przedstawić następująco: elektrony płynące od obszaru namagnesowanego do „miękkiego" zachowują spin zgodny z kierunkiem namagnesowania materiału trwale namagnesowanego i przenoszą ten spin do obszaru „miękkiego" namagnesowując go zgodnie z kierunkiem namagnesowania obszaru trwale namagnesowanego. Przepływ w przeciwnym kierunku powoduje, że elektrony o spinie zgodnym z kierunkiem namagnesowania obszaru „miękkiego" uchodzą z tego obszaru pozostawiając elektrony o spinie przeciwnym. W rezultacie obszar „miękki" zmienia kierunek pola magnetycznego. Ten proces „przekazywania spinu" nosi w jęz. angielskim nazwę „spin-transfer torque", stąd pamięć działająca według tej zasady określana jest często skrótem „STT-MRAM".

Technologicznie złącza tunelowe wytwarza się ponad tranzystorami, pomiędzy ścieżkami metalizacji, co umożliwia lokowanie komórek pamięci MRAM w najbliższym sąsiedztwie tranzystorów tworzących układy logiczne. Stwarza to nowe możliwości powiązania pamięci z układami logicznymi. Możliwości te są jeszcze przedmiotem badań, ponieważ układy CMOS z komórkami MRAM dopiero (w roku 2019) wchodzą do produkcji u kilku producentów. Produkowane są natomiast pamięci MRAM jako odrębne układy.

4.2.8 Pamięci ROM jako układy kombinacyjne

Jeśli spojrzeć na pamięć ROM jako na „czarną skrzynkę", nie interesując się jej wnętrzem, to jej działanie nie różni się od działania układu kombinacyjnego. W obu przypadkach mamy wejście, na które podawane są słowa binarne (w przypadku pamięci jest to wejście adresowe), i wyjście, na którym pojawiają się inne słowa binarne, przy czym każdemu słowu wejściowemu jednoznacznie przyporządkowane jest słowo wyjściowe. Zatem pamięć ROM może być traktowana jako jeden ze sposobów realizacji funkcji kombinacyjnej.

Istnieją takie funkcje kombinacyjne, które wygodniej jest zrealizować w postaci pamięci ROM, niż w tradycyjny sposób zestawiać z bramek NOR, NAND, NOT. Jest tak wtedy, gdy funkcja nie jest określona przez podanie wyrażeń logicznych, lecz w postaci tabeli przypisującej każdemu słowu wejściowemu odpowiednie słowo wyjściowe. Typowym przykładem są translatory kodów. Powszechnie przyjętym sposobem kodowania znaków

alfanumerycznych (liter, cyfr itd.) jest kod ośmiobitowy zwany kodem ASCII. Istnieją jednak też inne kody, na przykład kod zwany dalekopisowym, gdzie literom i cyfrom przypisuje się kody pięciobitowe. Kod ASCII i kod dalekopisowy nie są ze sobą związane żadną prostą funkcją logiczną. Najprostszym sposobem tłumaczenia jednego kodu na drugi jest zatem użycie pamięci ROM, w której na przykład kod dalekopisowy danego znaku służy jako adres, a kod ASCII tego samego znaku jest odczytywany spod tego adresu.

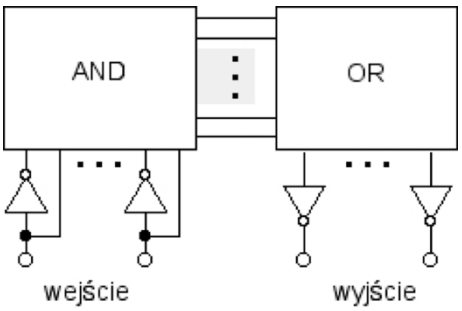
4.3 Układy o strukturze matrycowej

4.3.1 Układy PLA

Zauważmy, że wnętrze pamięci ROM może być uważane za zespół bramek NOR, bowiem do każdej linii bitu dołączona jest równolegle pewna liczba tranzystorów nMOS, tak samo jak równoległe połączenia tranzystorów nMOS w statycznych bramkach NOR (patrz punkt 3.2.2 w części pierwszej).

To spostrzeżenie prowadzi do wniosku, że wykorzystując bramki NOR takie, jak w pamięciach ROM, i tworząc z nich regularne struktury przypominające matryce pamięci, można realizować funkcje kombinacyjne w uporządkowanej, regularnej formie. Takie układy istnieją i zwane są zwykle układami PLA (ang. Programmable Logic Arrays). Nie jest to najoszczędniejszy i najbardziej korzystny z punktu widzenia szybkości działania sposób realizacji funkcji kombinacyjnych. Jego wielką zaletą jest jednak regularność budowy układu umożliwiająca łatwą automatyzację projektowania.

Typowy układ PLA składa się z dwóch matryc i pozwala łatwo realizować funkcje kombinacyjne wyrażone w postaci sumy iloczynów. Pierwsza matryca realizuje iloczyny i bywa nazywana matrycą AND, druga realizuje sumy i bywa nazywana matrycą OR. Obie w rzeczywistości składają się z bramek NOR. Funkcja wyrażona jako suma iloczynów zawsze może być przekształcona do postaci, w której występują tylko bramki NOR i inwertery.



Rysunek 4-14. Ogólny schemat układu PLA

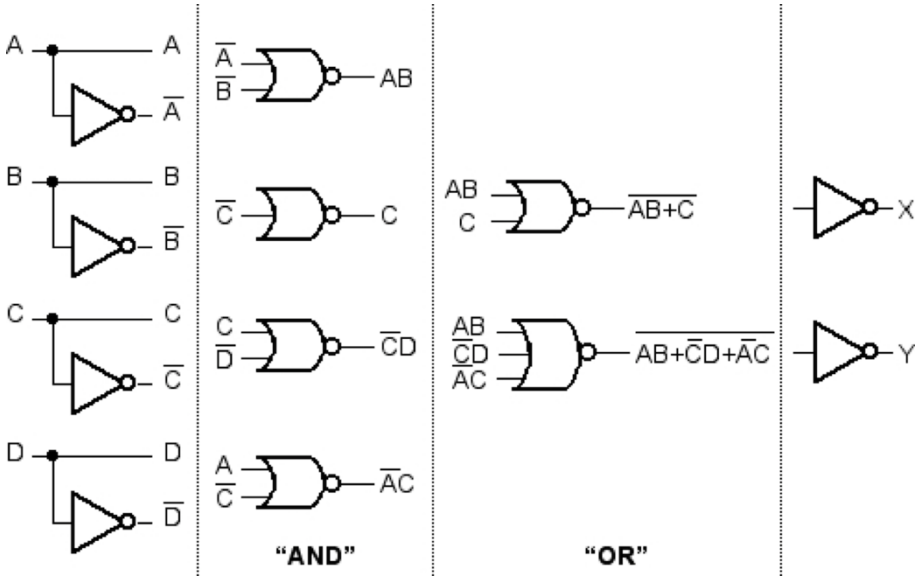
Przekształcenie funkcji logicznej do postaci, która umożliwia realizację tej funkcji w układzie PLA, najłatwiej zilustrować prostym przykładem. Przyjmijmy, że należy zbudować układ PLA realizujący funkcję

$$X = AB + C$$
$$Y = AB + \bar{C}D + \bar{A}C$$

Funkcję tę można zapisać w równoważnej logicznie postaci

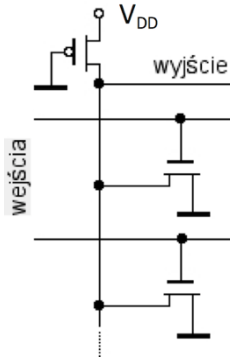
$$X = \overline{\overline{(\bar{A} + \bar{B})} + C}$$
$$Y = \overline{\overline{(\bar{A} + \bar{B})} + \overline{(C + \bar{D})} + \overline{(A + \bar{C})}}$$

która prowadzi do następującego schematu:



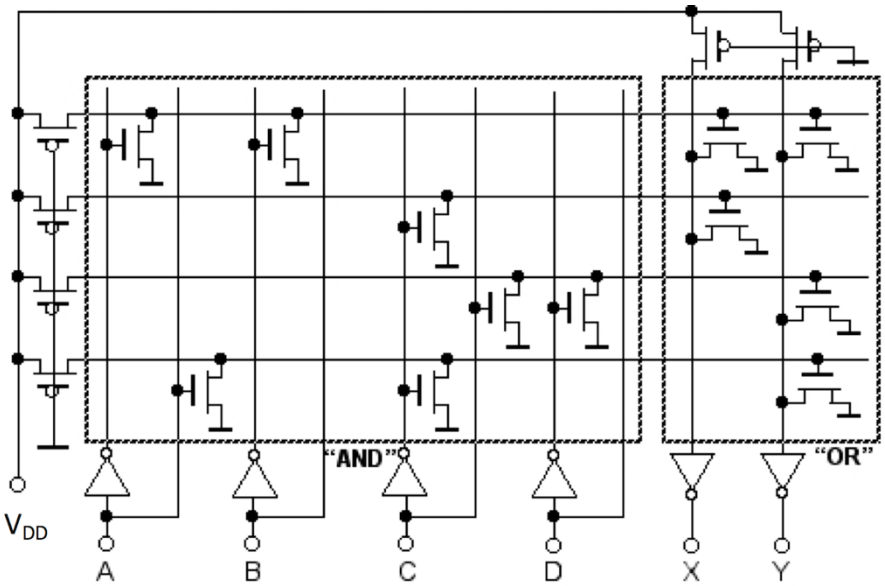
Rysunek 4-15. Przykładowa struktura układu PLA

Jeśli bramki NOR zostaną zbudowane tak, jak w wersji statycznej pamięci ROM (tj. ze statycznym obciążeniem tranzystorem pMOS)



Rysunek 4-16. Realizacja funkcji NOR dla matrycy PLA

otrzymujemy schemat układu w postaci:



Rysunek 4-17. Przykładowy kompletny układ PLA

Regularność budowy układu PLA umożliwia łatwą automatyzację projektowania. Mając funkcję logiczną w postaci sum iloczynów można od razu określić, na których skrzyżowaniach wierszy i kolumn w matrycach "AND" i "OR" mają się znaleźć tranzystory. Wadą tego sposobu realizacji funkcji kombinacyjnych jest duża powierzchnia

układu PLA - zauważmy, że jest w nim więcej pustych skrzyżowań, niż tranzystorów. Można co prawda powierzchnię trochę zmniejszyć, na przykład usuwając całkowicie niewykorzystane linie, jednak układ zbudowany w tradycyjny sposób z bramek NOR, NAND i inwerterów zapewne zająłby znacznie mniej miejsca.

Układ z bramkami obciążanymi statycznie ma tę wadę, że pobiera prąd, gdy przynajmniej jeden tranzystor nMOS jest włączony. Można tej wady uniknąć stosując bramki dynamiczne typu DOMINO, co jednak znacznie komplikuje układ.

Układy PLA były niegdyś bardzo pospolicie stosowane do projektowania złożonych układów kombinacyjnych, np. układów sterowania w mikroprocesorach. Obecnie ich znaczenie bardzo zmalało, bowiem układy kombinacyjne można również projektować w sposób zautomatyzowany wykorzystując komórki standardowe. Daje to na ogół mniejszą powierzchnię i porównywalne lub lepsze parametry elektryczne (szybkość, pobór mocy).

4.3.2 Programowalne układy cyfrowe

Układy PLA takie, jak omówione wyżej, można porównać do pamięci ROM programowanych maską. Funkcję wykonywaną przez układ PLA określa struktura tego układu ustalona podczas jego produkcji. Nic jednak nie stoi na przeszkodzie, by do budowy układów PLA zastosować te same technologie programowania, które są wykorzystywane w reprogramowalnych pamięciach ROM. W ten sposób powstał niezwykle dziś ważny i użyteczny rodzaj cyfrowych układów scalonych - programowalne matryce bramkowe, znane pod wieloma nazwami zależnymi od wewnętrznej architektury, ale najczęściej określane mianem układów FPGA (skrót od ang. "Field Programmable Gate Array"). O takich układach była już mowa w części 2 punkt 4.3.2.

Najprostszą programowalną matrycą bramkową może być układ PLA, w którym tranzystory są na każdym skrzyżowaniu linii w matrycy, a te, które nie są potrzebne przy realizacji danej funkcji, są odłączane. Można tu wykorzystać znaną z układów ROM technikę przepalania połączeń (programowanie jest wówczas jednokrotne) lub tranzystory dwubramkowe takie, jak w pamięciach EPROM. Ta ostatnia technologia daje możliwość programowania wielokrotnego.

Układy FPGA osiągnęły jednak dużo wyższy stopień złożoności i uniwersalności, niż ten, który byłby możliwy przy zastosowaniu matryc typu PLA. Typowy układ FPGA zawiera zbiór komórek, których funkcję logiczne są programowalne, oraz zbiór programowalnych połączeń, które umożliwiają skonfigurowanie układu o zadanym schemacie logicznym. Układy FPGA zawierają nie tylko bramki kombinacyjne, ale także elementy pamięciowe – przerzutniki, rejestry, a nawet rdzenie mikroprocesorów – a więc umożliwiają budowę kompletnych układów cyfrowych o dużej złożoności.

Oprócz technik programowania jednokrotnego i wielokrotnego stosowanych w reprogramowalnych pamięciach nieulotnych w układach FPGA stosowane są jeszcze dwa inne sposoby programowania. Jeden z nich jest odwrotnością przepalanych połączeń. Są to programowalne połączenia (znane pod angielską nazwą „antifuse”). Programowalne połączenie jest to obszar cienkiego dielektryka zlokalizowany między dwoma ścieżkami na dwóch różnych warstwach metalu. Normalnie warstwy te są odizolowane, jednak w obszarze cienkiego dielektryka można przy pomocy impulsu podwyższonego napięcia wywołać przebicie, które prowadzi do trwałego połączenia dwóch warstw metalu. Raz wykonanego połączenia nie można zlikwidować, jest to więc technika programowania jednokrotnego. Są też układy FPGA, w których elementami programującymi połączenia są bramki transmisyjne. Takie układy mają wewnętrzną pamięć typu RAM, w której przechowywana jest informacja konfigurująca - które bramki transmisyjne są włączone, a które nie. Układy te wymagają zaprogramowania po każdym włączeniu zasilania, ponieważ nie dysponują pamięcią nieulotną. Program może

zostać wpisany na przykład z zewnętrznej pamięci typu ROM. Zaletą takich układów FPGA jest możliwość zmiany wewnętrznej konfiguracji, a więc i wykonywanej funkcji, w czasie pracy urządzenia. Prowadzi to do nowej, fascynującej i mało dotąd wykorzystywanej w praktyce koncepcji układów samo-rekonfigurowalnych. Można sobie na przykład wyobrazić mikroprocesor, którego architektura wewnętrzna samoczynnie dostosowuje się do aktualnie wykonywanego zadania.

5 Problemy projektowania dużych układów cyfrowych

5.1 Ogólne zasady projektowania układów cyfrowych CMOS

Podstawową zasadą projektowania układów cyfrowych, które mają być zrealizowane jako układy scalone CMOS, jest zasada, że powinny to być układy synchroniczne. Oznacza to, że:

- wszystkie elementy pamięciowe w układzie (przerzutniki, rejestry, bloki pamięci RAM) mogą zmieniać swój stan tylko w odpowiedzi na sygnał zegara,
- we wszystkich elementach pamięciowych zmiana stanu następuje w dokładnie tej samej chwili w odpowiedzi na to samo zbocze sygnału zegara.

Ta zasada oznacza między innymi, że w układzie występuje tylko jeden główny, globalny sygnał zegara. Ewentualne dodatkowe sygnały (np. zegar dwufazowy o fazach „1" nie nakładających się) są generowane lokalnie z zachowaniem synchronizacji z zegarem głównym.

Tę ogólną zasadę trzeba uzupełnić kilkoma dalszymi regułami. Oto one:

Każdy układ zawierający elementy pamięciowe (przerzutniki, rejestry) musi mieć możliwość ustawienia w nich zadanych „startowych" stanów. Może to być zrealizowane w postaci zewnętrznego sygnału zerującego lub ustawiającego (w terminologii anglojęzycznej sygnał „Reset"). Można również zastosować układ automatycznie generujący sygnał „Reset" przy włączaniu zasilania. Takie układy, zwane „Power-on Reset", są dostępne jako gotowe komórki w bibliotekach dostarczanych przez producentów układów. Jeżeli stosujemy układ „Power-on Reset" do zapewnienia prawidłowego startu, to nawet wtedy nasz układ powinien zawierać też wejście dla zewnętrznego sygnału „Reset", aby była możliwość przywrócenia znanego z góry startowego stanu układu w czasie jego pracy. Taki globalny sygnał „Reset" może działać asynchronicznie (tj. zmieniać stan układu w dowolnej chwili, bez względu na stan sygnału zegara). Jeśli natomiast stosowany jest również lokalny sygnał „Reset" (lokalny oznacza ustawiający stany w elementach pamięciowych tylko w pewnym fragmencie układu), to taki sygnał musi działać synchronicznie.

Nie należy używać żadnych układów, których poprawność działania jest uzależniona od wartości opóźnień wnoszonych przez bramki lub połączenia. Opóźnienia mogą się wahać w szerokich granicach, są bowiem uzależnione od rozrzutów produkcyjnych. Nie są w pełni przewidywalne i nie są powtarzalne.

Nie należy samemu zestawiać z bramek kombinacyjnych różnych układów przerzutników takich, jak przerzutniki RS, JK lub T. Używać należy wyłącznie przerzutników typu D. Mogą to być przerzutniki z biblioteki komórek dostarczonej przez producenta układów, a jeśli są samodzielnie projektowane, muszą być starannie zweryfikowane przez symulację elektryczną.

5.2 Praktyczne problemy projektowania dużych układów

W dużych układach pojawiają się problemy, których dotąd nie omawialiśmy.

Gdy projektowany układ jest duży, niektóre połączenia mogą mieć długość rzędu milimetrów, a nawet centymetrów, a nie mikrometrów. Takie połączenia wprowadzają do układu znaczące pasożytnicze rezystancje, pojemności, a nawet indukcyjności. Czas propagacji sygnałów w długich połączeniach może być porównywalny, a nawet dłuższy od czasu propagacji w bramkach i może decydować o maksymalnej szybkości działania układu.

Rezystancję ścieżki o długości *L* i szerokości *W* można obliczyć ze wzoru 3-18 z części I, punkt 3.1.7, który tu dla wygody powtórzymy

Równanie 5-1

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

gdzie R_S jest rezystancją warstwową (zwaną także „rezystancją na kwadrat") ścieżki rezystora. O rezystancji warstwowej była mowa w części I, punkt 3.1.7, i tam podana jest jej definicja. Dla jednorodnego prostopadłościennego obszaru przewodzącego rezystancja warstwowa dana jest wzorem

Równanie 5-2

$$R_s = \frac{\rho}{d}$$

gdzie ρ jest rezystywnością materiału ścieżki, a *d* - jej grubością. Miarą rezystancji warstwowej jest om (Ω), ale zwyczajowo używa się miana „om na kwadrat" (Ω/□). To określenie bierze się stąd, że iloraz *L/W* można interpretować jako „liczbę kwadratów", jaką można wpisać w ścieżkę o długości *L* i szerokości *W*. Typowe wartości rezystancji warstwowych wynoszą: dla ścieżek metalu (Al) - 0,03 ... 0,08 Ω (mniej dla ścieżek wykonanych z miedzi), dla obszarów polikrzemowych 10 ... 30 Ω, dla obszarów źródeł i drenów typu *n* 10 ... 50 Ω, dla obszarów źródeł i drenów typu *p* 30 ... 100 Ω. W nowszych technologiach rezystancje warstwowe polikrzemu oraz obszarów źródeł i drenów mają wartości o rząd wielkości mniejsze od podanych wyżej, ponieważ są one pokrywane dobrze przewodzącą warstwą metalo-krzemową (np. WSi_2, $TiSi_2$, $PtSi_2$), jednak nadal ich rezystancje warstwowe są o około dwa rzędy wielkości wyższe, niż ścieżek z metalu. Jeśli obliczyć rezystancję ścieżki mającej szerokość 1 µm i długość 1 mm, czyli *L/W* („liczbę kwadratów") równą 1000, to ścieżka z metalu będzie miała rezystancję rzędu kilkudziesięciu omów, zaś ścieżka wykonana z polikrzemu - kilka do kilkudziesięciu kiloomów, zależnie od technologii. Widać stąd, że tylko ścieżki metalowe nadają się do wykonywania długich połączeń. Obszary przewodzące polikrzemowe oraz źródeł i drenów tranzystorów mogą służyć jedynie do krótkich połączeń lokalnych.

Znaczącą rezystancję wnoszą także kontakty. Typowe rezystancje kontaktów metalu do obszarów polikrzemu oraz źródeł i drenów tranzystorów wynoszą kilkadziesiąt Ω, kontakty między warstwami metalu mają rezystancję rzędu 1 ... 5 Ω.

Ścieżki są również związane z pojemnością do sąsiadujących obszarów. Przykładowo, pojemność na jednostkę powierzchni ścieżki polikrzemowej do podłoża układu jest rzędu 0,05 ... 0,06 fF/µm^2, dla metalu 1 pojemność ta jest rzędu 0,03 fF/µm^2. Ścieżka o wymiarach 1 µm x 1 mm ma powierzchnię 1000 µm^2, co daje pojemność ścieżki do podłoża na poziomie 60 fF dla polikrzemu i 30 fF dla metalu 1. W rzeczywistości pojemności te są większe, bowiem ścieżka o szerokości porównywalnej z grubością dielektryka, na którym leży, nie jest kondensatorem płaskim. Trzeba doliczyć pojemność obszarów krawędziowych, które mogą mieć pojemność porównywalną, a nawet większą od pojemności obliczonej wyżej.

Rezystancje i pojemności połączeń i kontaktów stanowią problem techniczny w kilku przypadkach:

- w przypadku długich ścieżek sygnałowych problemem jest czas propagacji sygnału związany z pojemnością i rezystancją,
- w przypadku długich ścieżek zasilających problemem jest spadek napięcia zasilania związany z rezystancją (będzie omawiany nieco dalej).

Czas propagacji sygnału w długich połączeniach może być na tyle długi, że układ synchroniczny przestanie działać synchronicznie. Są tu dwa nieco różne zagadnienia: opóźnienia sygnałów logicznych i opóźnienia sygnału zegarowego. Zajmiemy się najpierw pierwszym zagadnieniem.

Długie połączenie jest elementem RC o stałych rozłożonych. Załóżmy, że na jednym końcu takiego połączenia napięcie skokowo rośnie od 0 do V_{DD}. Po jakim czasie ten skok zostanie zaobserwowany na drugim końcu? Można pokazać, że czas propagacji sygnału w połączeniu o całkowitej pojemności C i całkowitej rezystancji R, mierzony jako czas narastania od 0 do 0,5V_{DD}, jest w przybliżeniu równy 0,38*RC*. Daje to dla ścieżki metalowej o wymiarach 1 µm x 1 mm czas propagacji rzędu 10^{-12} s, natomiast w przypadku ścieżki polikrzemowej tak samo obliczony czas propagacji jest rzędu 0,1 ... 1 ns. Jest to czas dłuższy od czasów propagacji bramek CMOS, w przypadku najbardziej zaawansowanych technologii nawet znacznie dłuższy. Trzeba przypomnieć, że w tych obliczeniach założono nieskończenie krótki czas narastania sygnału na początku ścieżki. W rzeczywistości długa ścieżka obciąża sterującą ją bramkę dość znaczną pojemnością, zatem czas narastania na początku ścieżki jest skończony i dość długi (zależy on od wymiarów tranzystorów w bramce sterującej). Wniosek, jaki z tego płynie, jest taki, że projekt bramki, do wyjścia której dołączone jest długie połączenie, musi uwzględniać pojemność tego połączenia. Widać także całkowitą nieprzydatność ścieżek polikrzemowych jako długich połączeń.

Czas propagacji sygnału w ścieżce jest proporcjonalny do stałej czasowej $t = RC$ tej ścieżki. Rezystancja *R* jest proporcjonalna do ilorazu *L/W* ścieżki, zaś pojemność do powierzchni, czyli do iloczynu *WL* (nie uwzględniamy tu pojemności obszarów krawędziowych). Z tego wynika, że stała czasowa *t* rośnie proporcjonalnie do L^2. Ten wynik pokazuje, że jednym z warunków uzyskania dużej szybkości działania układów cyfrowych jest minimalizowanie długości długich połączeń. W obecnym stanie mikroelektroniki są one jednym z istotnych czynników ograniczających tę szybkość. Warto dodać, że szacunkowe obliczenia robione były dla ścieżki o długości 1 mm, która we współczesnych dużych układach nie należałaby wcale do najdłuższych. Nie są rzadkością ścieżki o długości rzędu kilkunastu milimetrów. Wówczas nawet czas propagacji w ścieżkach metalowych staje się porównywalny z czasami propagacji sygnałów w bramkach CMOS. Prowadzi to do wniosku, że dalszy wzrost wielkości układów będzie musiał prowadzić do odejścia od koncepcji układu ściśle synchronicznego. Wrócimy jeszcze do tego zagadnienia.

Zajmiemy się teraz zagadnieniem opóźnień sygnału zegara. Rozprowadzenie globalnego sygnału zegara w dużym układzie stwarza problemy z dwóch powodów.

Po pierwsze w dużym układzie całkowita pojemność wszystkich wejść zegarowych jest bardzo duża, może sięgać setek pikofaradów. Oznacza to, że układy generujące sygnał zegara muszą zapewnić krótkie czasy narastania i opadania sygnału przy dużych pojemnościach obciążających. Nie nadają się tu zwykłe bramki z tranzystorami o małych wymiarach. Konieczne są bufory. Zagadnienie to omawialiśmy już wcześniej (punkt 3.2.6).

Po drugie dla poprawnego działania układu synchronicznego trzeba, aby zbocza sygnału zegara docierały do wszystkich wejść zegarowych w tym samym momencie. Oznacza to, że na drodze od pierwotnego źródła sygnału zegara do każdego wejścia zegarowego powinna znaleźć się taka sama liczba takich samych buforów obciążonych takimi samymi pojemnościami. Również długość połączeń powinna być taka sama. O tym również

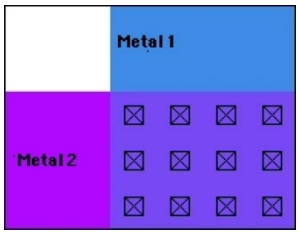
była mowa w punkcie 3.2.6. Spełnienie tych warunków w dużym układzie jest bardzo trudne. Jest to drugi powód odchodzenia od koncepcji układu ściśle synchronicznego.

W przypadku dużych układów poważnym problemem jest też zaprojektowanie sieci ścieżek zasilania. Rezystancje tych ścieżek i ewentualnie kontaktów mogą powodować znaczne, zmienne w czasie i zakłócające działanie układu spadki napięcia.

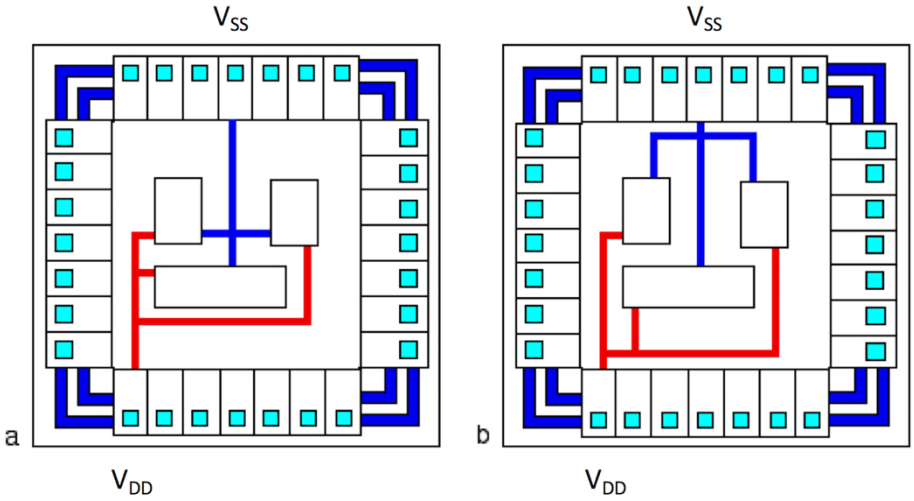
Pobór prądu przez bramki CMOS ma charakter krótkich impulsów o dużej amplitudzie. Jeśli dla pojedynczego inwertera szczytowa wartość prądu w impulsie osiąga kilkaset mikroamperów, to dla tysięcy równocześnie przełączających bramek otrzymamy prądy sięgające amperów. Wówczas nawet rezystancje rzędu pojedynczych omów powodują znaczne spadki napięcia. Powoduje to zaburzenia w działaniu układów, w tym między innymi przenikanie zakłóceń między blokami układu (patrz punkt 2.1.2). Przy bardzo krótkich czasach narastania impulsów prądu i długich ścieżkach zasilających wpływ na spadki napięć na tych ścieżkach może mieć nie tylko ich rezystancja, ale i indukcyjność.

Aby zminimalizować skutki spadków napięcia na ścieżkach zasilających, trzeba przestrzegać kilku zasad:

- szerokość ścieżek zasilających powinna być jak największa, a długość jak najmniejsza,
- ścieżki zasilające powinny być, jeśli to możliwe, prowadzone w tej warstwie metalu, która ma najmniejszą rezystancję warstwową,
- należy unikać przechodzenia ścieżek zasilania z warstwy na warstwę, bo kontakty wprowadzają dodatkową rezystancję; w razie potrzeby stosować wielokrotne kontakty (rysunek 5-1),
- aby zminimalizować przenikanie zakłóceń, ścieżki zasilania dla różnych bloków powinny biec osobno i spotykać się dopiero przy polach montażowych (rysunek 5-2),
- w razie potrzeby stosować więcej niż jedno pole montażowe masy i zasilania, rozdzielając zasilania różnych bloków.



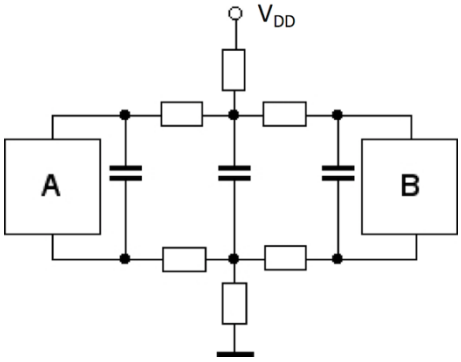
Rysunek 5-1. Przykład wielokontaktowego połączenia ścieżki metalu 1 ze ścieżką metalu 2



Rysunek 5-2. Rozprowadzenie ścieżek masy (V_{SS}) i zasilania (V_{DD}): (a) niekorzystne - długie odcinki wspólne wprowadzają wspólną rezystancję, (b) korzystniejsze - ścieżki łączą się blisko pól montażowych

Największe układy liczące miliony elementów pobierają tak duże prądy, że zasilanie i masę trzeba rozdzielać pomiędzy wiele pól montażowych. Poszczególne bloki układu mają odrębne połączenia zasilające, które łączą się dopiero poza obudową układu, na pakiecie drukowanym.

Stosowane bywają również pojemności odsprzęgające. Ich rola polega na łagodzeniu spadków napięcia – w chwili wzrostu poboru prądu jest on lokalnie dostarczany z naładowanego kondensatora, co zmniejsza prąd płynący przez rezystancję ścieżki ze źródła zasilania. Potem, w okresie małego poboru prądu między momentami przełączania, kondensator doładowuje się ze źródła zasilania ponownie do pełnego napięcia zasilania. Ta bardzo dawno znana i stosowana w elektronice zasada może być zrealizowana w układzie scalonym przez zastosowanie kondensatorów (rysunek 5-3). Takie kondensatory powinny znajdować się jak najbliżej bloku, którego napięcie zasilania mają stabilizować. Jednak pojemności, jakie można w ten sposób zrealizować w układzie scalonym, nie są na tyle duże (patrz punkt 3.1.7 w części I), aby całkowicie wyeliminować niebezpieczeństwo zaburzeń w pracy układu spowodowanych spadkami napięcia zasilania na rezystancji ścieżek.



Rysunek 5-3. Kondensatory odsprzęgające. Rezystory symbolizują rezystancje ścieżek

Przy projektowaniu połączeń, przez które płyną duże prądy, trzeba także przestrzegać maksymalnej obciążalności prądowej ścieżek podawanej przez producenta układów. Przekroczenie dopuszczalnej obciążalności prowadzi do uszkodzeń układów w czasie ich eksploatacji.

6 Testowanie i testowalność układów scalonych

Jak wiemy, w procesach produkcyjnych występują zaburzenia. W wyniku tych zaburzeń część wyprodukowanych układów zawiera defekty, które powodują, że układy nie działają tak, jak powinny. Trzeba je więc testować i odrzucać układy wadliwe. Testowanie dużych układów cyfrowych jest dziś poważnym problemem technicznym. Zbadanie czy układ zawierający dziesiątki milionów tranzystorów wykonuje poprawnie swe – zwykle bardzo złożone – funkcje dla wszystkich możliwych stanów i sekwencji sygnałów na wejściach jest nie do wykonania w rozsądnie krótkim czasie. Badanie sprawności układu wymaga więc użycia specjalnych metod. Już na etapie projektu architektury i projektu logicznego układu projektant powinien przewidzieć, w jaki sposób układ będzie testowany. Wiele sposobów testowania wymaga bowiem wprowadzenia do układu dodatkowych bloków funkcjonalnych. Problemy testowania są często ignorowane przez niedoświadczonych projektantów. W rezultacie otrzymuje się układy, których w ogóle nie da się przetestować. A przecież układ lub system, o którym nie wiadomo, czy działa poprawnie, jest bezużyteczny.

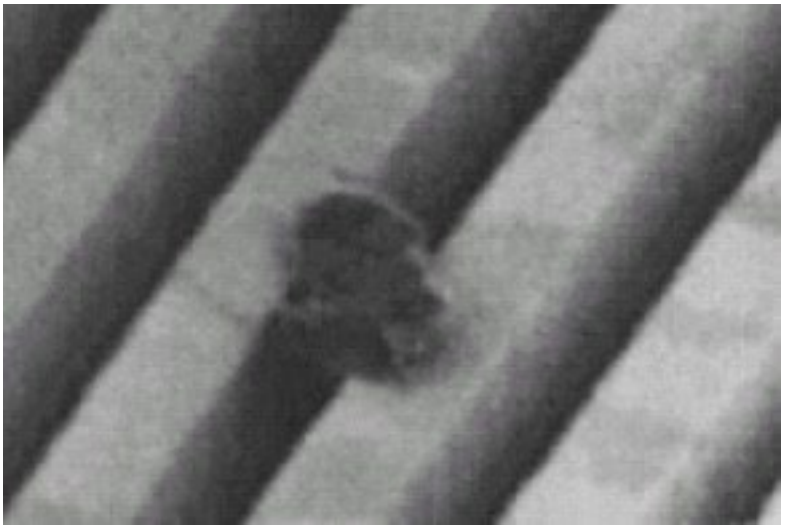
6.1 O defektach i uszkodzeniach w układach cyfrowych

6.1.1 Defekty i uszkodzenia produkcyjne

Typowe bramki cyfrowe CMOS, a zwłaszcza bramki statyczne, są mało wrażliwe na parametry tranzystorów. Tranzystory działają jak przełączniki - przewodzą prąd lub nie. Ich charakterystyki mają wpływ na takie parametry, jak szybkość działania i pobór prądu, ale nie na to, czy bramka w ogóle działa, czy też nie. Dlatego

głównym problemem w układach cyfrowych nie są uszkodzenia parametryczne, ale uszkodzenia katastroficzne (o defektach i uszkodzeniach była mowa w części I, punkt 4.2.6).

W układach cyfrowych CMOS najczęściej obserwujemy uszkodzenia katastroficzne powodowane przez defekty strukturalne. Przykład takiego defektu – cząstki przewodzącej, która zwiera ścieżki przewodzące – pokazuje poniżej zdjęcie mikroskopowe.



Rysunek 6-1. Przykład defektu - zwarcia ścieżek przewodzących (zdjęcie wykonane mikroskopem elektronowym)

Najpospolitsze są defekty powodujące zmiany w połączeniach w układzie: zwarcia ścieżek oraz przerwy w ścieżkach. W układach mających wiele warstw połączeń, i co za tym idzie - dużą liczbę kontaktów między warstwami metalu, dość częste są też uszkodzenia polegające na braku kontaktu między warstwami metalu, co jest zwykle spowodowane niedostatecznym wytrawieniem okna kontaktowego w dielektryku. Zdarzają się też inne defekty, jak na przykład „dziury" w cienkim dielektryku bramkowym tranzystorów MOS powodujące połączenie elektryczne bramki z kanałem i całkowicie zniekształcające charakterystyki uszkodzonego tranzystora.

W dużych cyfrowych układach scalonych spotyka się także uszkodzenia katastroficzne, które mają swe przyczyny w nadmiernym rozrzucie produkcyjnym parametrów elementów, a nie w defektach strukturalnych. Ten nadmierny rozrzut może na przykład powodować, że czasy propagacji sygnałów zegarowych w buforach zegara, które powinny być jednakowe, są w rzeczywistości różne, wobec czego zegar nie dociera jednocześnie do wszystkich bloków układu. Małe różnice zawsze istnieją i nie są szkodliwe. Jeśli jednak różnice w czasie propagacji sygnałów zegara stają się porównywalne z okresem zegara, to układ przestaje być układem synchronicznym. Taki układ może działać prawidłowo pod warunkiem obniżenia częstotliwości zegara, ale bywa i tak, że nie pracuje prawidłowo przy żadnej częstotliwości. Mamy więc wtedy do czynienia z uszkodzeniem katastroficznym. Tutaj jednak nie będziemy zajmować się tego rodzaju uszkodzeniami.

6.1.2 Defekty i uszkodzenia eksploatacyjne

Defekty mogą też powstać w układzie w trakcie jego eksploatacji. Takie defekty w układach prawidłowo zaprojektowanych, prawidłowo wykonanych i prawidłowo eksploatowanych są, jak wiemy, niezwykle rzadkie. We współcześnie produkowanych układach scalonych są trzy główne mechanizmy uszkodzeń w trakcie eksploatacji:

- elektromigracja powodująca przerwy ścieżek połączeń,
- naprężenia mechaniczne powodujące mechaniczne uszkodzenia płytki układu lub połączeń,

- degradacja właściwości elektrycznych tlenku bramkowego wywołana "gorącymi" (wysokoenergetycznymi) nośnikami ładunku.

Elektromigracja polega na przemieszczaniu się atomów metalu pod wpływem silnego strumienia nośników (czyli prądu o dużej gęstości). Podlegają jej zwłaszcza połączenia aluminiowe (aluminium jest, jak wiemy, bardzo miękkim metalem, czyli jego wiązania międzyatomowe są słabe). Elektromigracja jest procesem, który bardzo nasila się przy podwyższonej temperaturze. Atomy aluminium są powoli przemieszczane z miejsc na ścieżce, gdzie gęstość prądu jest największa (a więc tam, gdzie ścieżki mają zmniejszony przekrój poprzeczny, na przykład są zwężone) do sąsiadujących z takimi miejscami innych fragmentów ścieżki. Po pewnym czasie w miejscu przewężonym ścieżka ulega przerwaniu. Producenci układów podają maksymalną gęstość prądu w ścieżkach, której nie wolno przekraczać. Gęstość ta jest funkcją maksymalnej temperatury, w jakiej pracować ma układ. Projektant nie tylko powinien projektować ścieżki o szerokościach dostosowanych do prądów w nich płynących, ale także starać się, by kształty ścieżek były możliwie jak najprostsze, bez zbędnych załamań i przewężeń.

Naprężenia mechaniczne są zwykle wywołane błędami lub wadami produkcyjnymi w montażu układów w obudowach. Przykładowo, użycie niewłaściwego lutu do przylutowania płytki układu do metalowej podstawki może powodować powstawanie naprężeń mechanicznych przy zmianach temperatury na skutek dużych różnic rozszerzalności cieplnej krzemu i metalu. Takie naprężenia mogą powodować mikropęknięcia w płytkach bądź w połączeniach i w konsekwencji uszkodzenie układu.

Degradacja właściwości tlenku bramkowego jest wywołana bombardowaniem tlenku przez nośniki uzyskujące dużą energię w polu elektrycznym. Ten rodzaj uszkodzeń jest typowy zwłaszcza dla tranzystorów stanowiących komórki pamięci w pamięciach typu EEPROM i *flash*. Była o tym mowa wcześniej (punkt 4.2.6).

Ponieważ uszkodzenie może wystąpić w układzie w czasie jego eksploatacji, problem testowania to nie tylko problem zbadania układu tuż po jego wyprodukowaniu, ale w niektórych przypadkach także problem sprawdzania poprawności działania układu w czasie, gdy układ działa w urządzeniu. Dotyczy to urządzeń i systemów, których uszkodzenie może wywołać wielkie i nieodwracalne szkody, na przykład katastrofę lub utratę ludzkiego życia.

6.1.3 Modele uszkodzeń

Aby móc opracować testy dla układu cyfrowego, musimy wiedzieć, jak wpływają defekty w układzie na wykonywane przez ten układ funkcje.

> **DEFINICJA**
>
> *Opis działania układu uszkodzonego, tj. zawierającego defekt, nazywamy modelem uszkodzenia.*

W przypadku układu cyfrowego model uszkodzenia określany jest w odniesieniu do funkcji logicznej wykonywanej przez układ, tj. określa, jak zmienia się funkcja logiczna pod wpływem uszkodzenia.

Najprostszym modelem uszkodzenia jest model znany pod nazwą „stałe zero/stała jedynka” (lub niekiedy „sklejenie do zera/sklejenie do jedynki”, w jęz. angielskim „stuck at 0/stuck at 1”). Będziemy go w skrócie oznaczać symbolem „SA0/SA1”.

DEFINICJA

Model uszkodzenia „SA0/SA1" polega na założeniu, że każdy defekt w układzie prowadzi do tego, że w jakimś jego węźle stan logiczny nie zmienia się - zawsze jest tam stan „0" lub stan „1".

Chodzi tu oczywiście o te węzły w układzie, w których w prawidłowo działającym układzie stan logiczny powinien zmieniać się.

Model „SA0/SA1" nie ma żadnego uzasadnienia teoretycznego, a symulacje komputerowe i praktyka pokazują, że nie więcej niż kilkanaście procent wszystkich defektów strukturalnych daje w efekcie uszkodzenie typu „SA0/SA1". Znacznie częściej spotyka się uszkodzenia polegające na tym, że zmienia się funkcja logiczna wykonywana przez bramkę czy też blok, w którym wystąpił defekt, ale w żadnym z węzłów nie panuje na stałe stan „0" ani stan „1". Spotyka się defekty, w wyniku których układ kombinacyjny staje się układem z pamięcią (tj. odpowiedź układu zależy nie tylko do aktualnych stanów wejść, ale i od historii zmian stanów). Spotyka się też defekty, w wyniku których w jakimś węźle układu panuje napięcie nie będące ani prawidłowym stanem „0", ani „1". Mimo to model „SA0/SA1" jest powszechnie i niemal wyłącznie stosowany w teorii i praktyce testowania układów cyfrowych. Wieloletnia praktyka pokazała bowiem, że jest to model skuteczny. Otrzymywane przy użyciu tego modelu zestawy testów dla układów cyfrowych pozwalają przetestować te układy z dostateczną dla celów praktycznych wiarygodnością.

Wielką zaletą modelu „SA0/SA1" jest jego prostota, a przede wszystkim niezależność od struktury fizycznej układu. Dla stosowania w praktyce tego modelu wystarczy znać schemat logiczny projektowanego układu. Nie jest konieczna znajomość schematu elektrycznego ani topografii układu. Dalej zobaczymy, w jaki sposób model „SA0/SA1" jest wykorzystywany do otrzymywania zestawów testów dla układów cyfrowych.

Istnieją inne, bliższe rzeczywistości fizycznej modele uszkodzeń. Drogą odpowiednich symulacji komputerowych można na przykład przewidzieć możliwe defekty strukturalne w danej bramce, obliczyć ich prawdopodobieństwo, a następnie określić, jaką funkcję logiczną będzie wykonywać bramka z każdym z tych defektów. W ten sposób uzyskuje się szczegółową informację o rodzajach i prawdopodobieństwie uszkodzeń w danej bramce. Informację tę można potem wykorzystać do określenia najlepszego zestawu testów dla układu zawierającego tak scharakteryzowane bramki. Wadą takiego podejścia jest konieczność brania pod uwagę konkretnej struktury fizycznej układu - topografii bramek i połączeń między nimi. Od tej topografii zależy bowiem prawdopodobieństwo wystąpienia różnych rodzajów defektów strukturalnych. Na przykład prawdopodobieństwo wystąpienia zwarcia między ścieżkami połączeń zależy od ich wzajemnej odległości. Chociaż sposób testowania wychodzący od realnych uszkodzeń i ich prawdopodobieństwa daje możliwość otrzymania lepszych zestawów testów, to jednak konieczność znajomości konkretnej struktury fizycznej układu oraz duża złożoność symulacji, które trzeba wykonać, powodują, że ten sposób otrzymywania zestawów testów nie znalazł dotąd zastosowania praktycznego.

Dalej będzie używany model „SA0/SA1".

6.2 Testowanie układów cyfrowych

6.2.1 Metody testowania i generacji testów

Wyobraźmy sobie, że kupujemy w sklepie kalkulator i chcemy sprawdzić, czy działa prawidłowo. W tym celu wykonujemy w sposób mniej lub bardziej przypadkowy kilka obliczeń, których prawidłowy wynik jest nam znany. Wyniki wyświetlone przez kalkulator są poprawne. Czy po takim teście możemy być pewni, że kalkulator

działa całkowicie prawidłowo? Nie! Taką pewność uzyskalibyśmy dopiero wtedy, gdybyśmy wykonali wszystkie możliwe działania dla wszystkich możliwych argumentów. Taki test nosi nazwę wyczerpującego testu funkcjonalnego. Jest oczywiste, że nie da się go wykonać kupując w sklepie kalkulator.

A gdyby test był wykonywany przez szybki tester, taki, jaki stosowany jest do testowania układów scalonych?

Przypuśćmy, że testujemy układ kombinacyjny mający wejście *n*-bitowe. Jedno słowo *n*-bitowe podane na wejście dla celów testowania będziemy nazywali wektorem testowym. Wyczerpujący test funkcjonalny wymaga podania na wejście wszystkich możliwych kombinacji zer i jedynek, czyli 2^n wektorów testowych, i zbadania poprawności odpowiedzi układu. Jeżeli testujemy układ sekwencyjny mający wejście *n*-bitowe i *m* stanów wewnętrznych, wyczerpujący test funkcjonalny wymaga podania na wejście $2^{(n+m)}$ wektorów testowych. Czy to dużo? Wyobraźmy sobie prosty układ, dla którego *n*=25, *m*=50. Potrzebne jest więc 2^{75} wektorów testowych. Jest to liczba równa około $3{,}8 \times 10^{22}$. Jeżeli dysponujemy testerem, który potrzebuje 1 µs na jeden wektor testowy, testowanie naszego prostego układu będzie trwało około 10^9 lat!

Ten prosty przykład (zaczerpnięty z literatury) pokazuje, że wyczerpujący test funkcjonalny jest niemożliwy do wykonania nawet dla zupełnie prostych układów. Potrzebne są więc inne sposoby testowania. Omówimy je nieco dalej.

A ponieważ wyczerpujący test funkcjonalny nie jest możliwy, musimy mieć miarę jakości testów pozwalającą nam oszacować, jaki jest stopień wiarygodności wyników testowania. Innymi słowy, musimy znać odpowiedź na pytanie, jaki procent wszystkich możliwych uszkodzeń w układzie wykryje zestaw wektorów testowych, który nie jest wyczerpującym testem funkcjonalnym. Ten procent będziemy nazywali poziomem wykrywalności uszkodzeń dla danego zestawu wektorów testowych. Im niższy ten poziom, tym większe prawdopodobieństwo, że w wyniku testowania zakwalifikowany zostanie jako sprawny układ, który w rzeczywistości jest uszkodzony.

Poziom wykrywalności uszkodzeń jest miarą teoretyczną określaną przy zastosowaniu modelu uszkodzeń „SA0/SA1" i przy dalszych upraszczających założeniach. Jedno z nich to założenie, że uszkodzenie typu „SA0/SA1" może z jednakowym prawdopodobieństwem wystąpić w każdym węźle układu. Tak w rzeczywistości nie jest. W związku z tym poziom wykrywalności uszkodzeń nie wiąże się bezpośrednio i jednoznacznie z faktyczną liczbą układów wadliwych. Poniżej (tabela 6-1) przykładowe wyniki badań statystycznych pokazujących typowy związek poziomu wykrywalności uszkodzeń z faktycznym procentem układów wadliwych wśród układów zakwalifikowanych jako sprawne. Dane te *nie są uniwersalne*, zostały uzyskane dla określonych rodzajów układów i określonych sposobów ich testowania, ale dają pewne pojęcie o istniejącej tu zależności.

Tabela 6-1. Poziom wykrywalności uszkodzeń a liczba faktycznie wadliwych układów - przykład

Poziom wykrywalności uszkodzeń	Procent układów wadliwych wśród układów zakwalifikowanych jako sprawne
50%	7%
90%	3%
95%	1%
99%	0,1%
99,9%	0,01%

Wprowadzimy jeszcze dwa inne nowe pojęcia: kontrolowalności węzłów i obserwowalności węzłów w układzie. Metody testowania, o których będzie dalej mowa, wykorzystują model uszkodzenia „SA0/SA1" zakładający, że

uszkodzenie polega na ustalonym stanie „0” lub „1” w jakimś węźle układu. Zatem dla sprawdzenia, czy wystąpiło uszkodzenie, trzeba będzie ustawiać określone stany w węzłach układu i sprawdzać te stany.

DEFINICJA

Mówimy, że dany węzeł jest kontrolowalny, jeśli sekwencja wektorów testowych o skończonej długości pozwala na ustawienie w tym węźle żądanego stanu („0” lub „1”).

Węzeł jest tym łatwiej kontrolowalny, im krótsza jest ta sekwencja.

DEFINICJA

Mówimy, że dany węzeł jest obserwowalny, jeśli sekwencja wektorów testowych o skończonej długości pozwala na pojawienie się na wyjściu układu stanu pozwalającego określić, jaki był stan ustawiony w węźle („0” lub „1”).

Węzeł jest tym łatwiej obserwowalny, im krótsza jest ta sekwencja.

Dla prawidłowo zaprojektowanych pod względem logicznym układów kombinacyjnych kontrolowalność i obserwowalność węzłów nie stanowi większego problemu. Natomiast w przypadku układów sekwencyjnych zmiana stanu w niektórych wewnętrznych węzłach układu może wymagać niezwykle długich sekwencji wektorów testowych. Jak zobaczymy, istnieją sposoby projektowania układów zasadniczo poprawiające kontrolowalność i obserwowalność węzłów w układach sekwencyjnych.

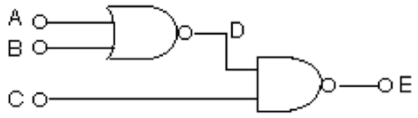
Posługując się zdefiniowanymi wyżej pojęciami można zadanie określenia sposobu testowania układu oraz wygenerowania odpowiedniego zestawu wektorów testowych sformułować następująco: celem jest uzyskanie jak najkrótszej sekwencji wektorów testowych zapewniającej wymagany poziom wykrywalności uszkodzeń.

Trzeba się przy tym pogodzić z faktem, że dla dużych i złożonych układów poziom wykrywalności uszkodzeń będzie z reguły niższy niż 100%. Nacisk na to, że sekwencja wektorów testowych powinna być jak najkrótsza, bierze się stąd, że testowanie dużych układów jest kosztowne. Wynika to z bardzo wysokiego kosztu nowoczesnych testerów. Nierzadkie są przypadki, gdy koszt testowania układu jest porównywalny z kosztem jego wyprodukowania. Wprowadziliśmy już pojęcie wyczerpującego testu funkcjonalnego i wiemy, że nie da się go wykonać w rozsądnym czasie. Dlatego powszechnie stosowana jest metoda zwana *testowaniem strukturalnym*. W tej metodzie nie badamy, czy układ prawidłowo wykonuje swą funkcję. Możemy nawet w ogóle nie interesować się, jaka ona jest. Zadaniem testowania strukturalnego jest sprawdzenie, czy w procesie produkcyjnym nie wystąpił defekt naruszający prawidłową strukturę układu. Przy założeniu, że modelem uszkodzenia jest model „SA0/SA1”, testowanie strukturalne polega na badaniu, czy w żadnym węźle układu nie występuje uszkodzenie tego typu. Zakłada się przy tym, że układ jest poprawnie zaprojektowany, tzn. brak defektów naruszających prawidłową strukturę układu jest równoznaczny z prawidłowym działaniem układu.

Proces testowania strukturalnego wygląda następująco. Wybiera się konkretny węzeł układu i podaje się na wejścia wektor testowy (lub sekwencję wektorów testowych) ustawiający w tym węźle konkretny stan (na przykład „0”). Następnie obserwując odpowiedź układu na wyjściu określa się, czy ten stan udało się ustawić, czy też nie. Tę samą procedurę powtarza się dla wszystkich węzłów, dla każdego z nich próbując ustawić stan zarówno „0”, jak i „1”. Jak widać, przy takiej procedurze testowania nie jest potrzebna znajomość funkcji

wykonywanej przez układ. Trzeba natomiast znać jego schemat logiczny, od niego bowiem zależy, jakie wektory testowe należy podać na wejścia i jakie powinny być prawidłowe odpowiedzi układu na wyjściach.

Zilustrujmy to bardzo prostym przykładem. Niech testowanym układem będzie połączenie dwóch bramek pokazane na rysunku 6-2.



Rysunek 6-2. Testowany układ kombinacyjny

Przypuśćmy, że chcemy zbadać, czy w węźle D, który jest wewnętrznym węzłem układu, nie występuje uszkodzenie typu „stałe zero". W tym celu staramy się w węźle D ustawić „1" - musimy podać wektor testowy, w którym A=„0" i B=„0", a stan C jest nieistotny. Aby móc zaobserwować stan węzła D na wyjściu E, trzeba na wejście C podać „1". Innymi słowy, wektor „00x" (x – stan nieistotny) zapewnia kontrolowalność węzła D dla ustawienia w nim „1", zaś wektor „xx1" zapewnia obserwowalność stanu węzła D. Stąd test węzła D ze względu na uszkodzenie „stałe zero" wymaga podania wektora testowego „001". Jeśli w węźle D występuje uszkodzenie „stałe zero", na wyjściu pojawi się „1", zaś w przypadku braku uszkodzenia na wyjściu pojawi się „0". Podobnie, dla testu węzła D ze względu na „stałą jedynkę" trzeba podać jeden z wektorów: „011" lub „101" lub „111".

Testowanie strukturalne ogromnie zmniejsza liczbę wektorów w porównaniu z wyczerpującym testem funkcjonalnym. Dla przetestowania strukturalnego omawianego układu wystarczają dwa wektory. W przypadku wyczerpującego testu funkcjonalnego potrzebne byłoby 2^3=8 wektorów. W przypadku rzeczywistych układów oszczędność jest daleko większa, bo często jeden wektor testowy testuje skutecznie więcej niż jeden węzeł. Na omówionej tu zasadzie działają generatory wektorów testowych – programy komputerowe określające optymalne sekwencje wektorów testowych dla układów o danym schemacie logicznym.

Testowanie strukturalne można stosować do układów cyfrowych realizowanych w różny sposób, nawet niekoniecznie w postaci układów scalonych. Istnieje jeszcze jeden sposób testowania układów cyfrowych, nadający się jednak tylko do testowania układów CMOS. Jest to testowanie prądowe (znane w literaturze anglojęzycznej jako „IDDQ testing"). Wykorzystuje się tu fakt, że prawidłowo skonstruowane i wykonane bramki CMOS pobierają prąd tylko w czasie przełączania. Gdy na wejściach panuje stan ustalony, bramki CMOS nie pobierają prądu (jeśli pominąć bardzo mały prąd podprogowy tranzystorów i prądy wsteczne złącz źródeł i drenów). Zdecydowana większość defektów, jakie mogą się pojawić w bramkach CMOS, prowadzi do tego, że istnieją takie kombinacje stanów na wejściach bramek, przy których bramki w stanie ustalonym pobierają prąd. Zatem, zamiast obserwować odpowiedzi układu na wyjściach, można badać pobór prądu. Po podaniu każdego wektora testowego należy odczekać, aż ustalą się stany na wyjściu i dokonać pomiaru prądu pobieranego ze źródła zasilania. Jeśli prąd ten jest o kilka rzędów wielkości większy od prądu w układzie działającym prawidłowo, to możemy mieć pewność, że w układzie występuje co najmniej jeden defekt.

Testowanie prądowe ma wiele zalet. Pozwala ono szybko i łatwo wykryć wiele uszkodzeń, które przy testowaniu strukturalnym wymagałyby podania bardzo długich sekwencji wektorów testowych. Pomiar prądu może być wykonywany wewnątrz układu, przez wprowadzenie do układu monitorów prądu – dość prostych układów, które śledzą pobór prądu w czasie pracy układu i sygnalizują przekroczenie jego progowej wartości oznaczającej wystąpienie uszkodzenia. Można więc testować układ w czasie jego normalnej pracy, co nie jest możliwe w przypadku testowania strukturalnego. Możliwe jest nawet budowanie układów samonaprawiających się. W takim układzie bloki funkcjonalne są zdublowane. W razie wykrycia przez monitor prądu, że w którymś bloku wystąpiło uszkodzenie, możliwa jest automatyczna rekonfiguracja układu przez odłączenie bloku uszkodzonego

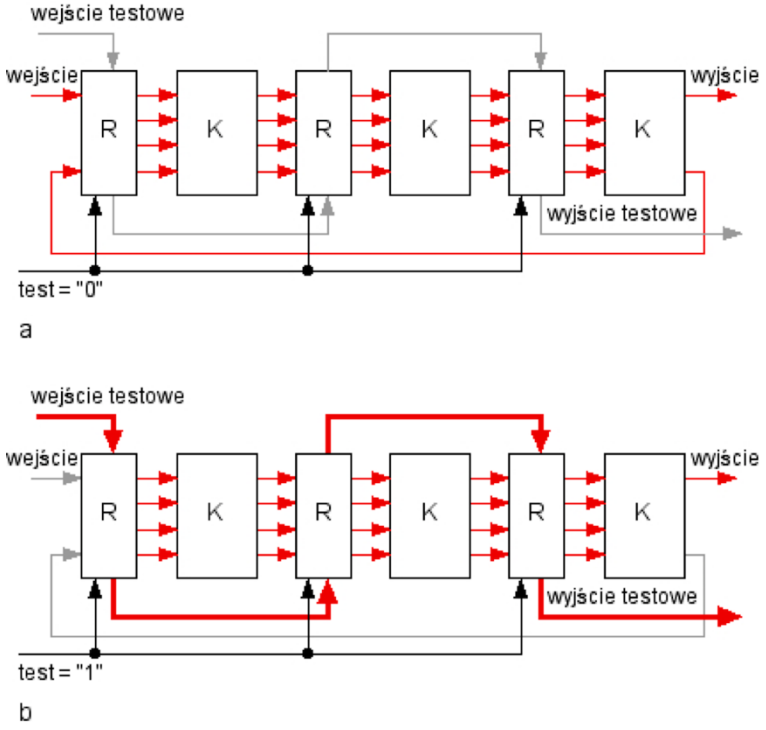
i podłączenie na jego miejsce bloku zapasowego. Możliwość ta jest jednak w praktyce bardzo rzadko wykorzystywana, bowiem układy scalone i bez tego cechują się bardzo wysoką niezawodnością, a zdublowanie wszystkich bloków oznacza podwojenie kosztu układu.

Testowanie prądowe i testowanie strukturalne nie zastępują się wzajemnie, lecz uzupełniają. Są bowiem takie uszkodzenia, których nie wykrywa testowanie prądowe, i takie, których wykrycie w testach strukturalnych wymaga bardzo długich sekwencji wektorów testowych. Ale nawet równoczesne zastosowanie obu sposobów testowania nie rozwiązuje do końca problemów testowania układów bardzo dużych i złożonych. W ich przypadku dla uzyskania dostatecznie wysokiego poziomu wykrywalności uszkodzeń potrzebne byłyby sekwencje wektorów testowych o długościach niemożliwych do zaakceptowania. Dlatego rozwinęły się metody projektowania układów łatwo testowalnych i samotestujących się. Jest o nich mowa w następnym punkcie.

6.2.2 Układy łatwo testowalne i samotestujące się

Zajmiemy się teraz odpowiedzią na pytanie: jak zaprojektować układ cyfrowy, by był on łatwo testowalny (czyli aby można było uzyskać wysoki poziom wykrywalności uszkodzeń przy umiarkowanej długości sekwencji wektorów testowych)?

Projektowanie układów łatwo testowalnych polega na wyborze rozwiązań poprawiających obserwowalność i kontrolowalność wewnętrznych węzłów. Głównym problemem są bloki sekwencyjne, bowiem w nich ustawienie zera lub jedynki w konkretnym węźle, a potem zaobserwowanie na wyjściu, jaki tam był stan, może wymagać podania bardzo długiej sekwencji wektorów testowych. Niekiedy poprawę kontrolowalności i obserwowalności można osiągnąć przez wprowadzenie zmian do schematu logicznego oraz ewentualnie dodatkowych wejść i wyjść wykorzystywanych tylko przy testowaniu. Dodatkowe wejścia i wyjścia są jednak rozwiązaniem podnoszącym znacznie koszt układu, bowiem pola montażowe zajmują bardzo dużą powierzchnię. Uniwersalnym, skutecznym i powszechnie stosowanym sposobem jest wyposażenie układu w łańcuch skanujący (zwany także ścieżką skanującą) – rysunek 6-3. Łańcuch skanujący jest to utworzony z przerzutników rejestr szeregowy, który umożliwia przekształcenie układu sekwencyjnego w kombinacyjny na czas testowania. Wykorzystuje się tu przerzutniki istniejące w układzie, nie ma potrzeby dodawania nowych. Przerzutniki te mają jednak specjalną konstrukcję umożliwiającą wykorzystanie ich w łańcuchu skanującym.

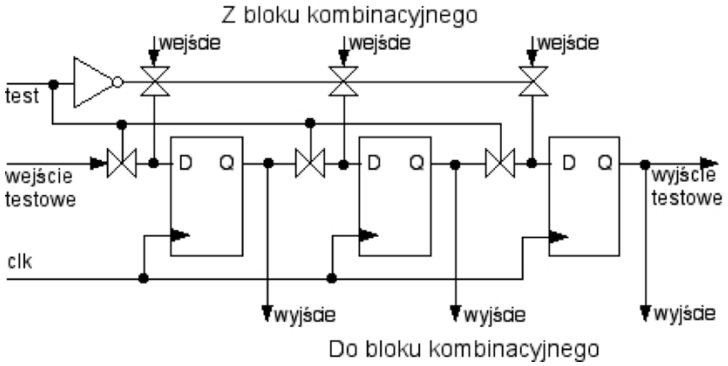


Rysunek 6-3. Układ sekwencyjny z łańcuchem skanującym. (a) - praca w trybie normalnym, (b) - praca w trybie testowym.

Na rysunku 6-3 kolorem szarym zaznaczono połączenia nieaktywne w danym trybie pracy, pogrubiona czerwona linia: szeregowy łańcuch skanujący.

Rysunek 6-3 ilustruje ideę łańcucha skanującego. Układ składa się z bloków kombinacyjnych K i z rejestrów R, zaś w każdym rejestrze znajduje się pewna liczba przerzutników typu D. Układ z łańcuchem skanującym ma dwa tryby pracy: zwykły i testowy. W zwykłym trybie pracy każdy z przerzutników w rejestrach działa niezależnie od pozostałych. Układ wykonuje swe zwykłe funkcje. W trybie testowym wszystkie przerzutniki w rejestrach zostają połączone w jeden szeregowy rejestr przesuwający zwany łańcuchem skanującym, zaznaczony na rysunku 6-3b pogrubioną czerwoną linią. Teraz możliwe jest szeregowe wpisanie z wejścia testowego do łańcucha skanującego wektora testowego, który z wyjść przerzutników podany będzie bezpośrednio na wewnętrzne węzły układu – wejścia bloków kombinacyjnych. Odpowiedzi są z wyjść wpisywane do przerzutników w łańcuchu skanującym, po czym można je szeregowo wyprowadzić na wyjście testowe i porównać z prawidłowymi. W czasie wyprowadzania odpowiedzi na poprzedni wektor testowy można równocześnie wpisywać następny wektor. Dzięki łańcuchowi skanującemu węzły wewnętrzne połączone z przerzutnikami stają się w trybie testowym łatwo kontrolowalne i obserwowalne.

Budowa przerzutników do łańcucha skanującego nie jest skomplikowana. Idea jest pokazana na rysunku 6-4. Przy stanie „1” sygnału „test” następuje wpisywanie szeregowe wektora testowego. Po zakończeniu tego procesu stan sygnału „test” zmienia się na „0” na czas tak długi, aby w blokach kombinacyjnych ustaliły się odpowiedzi na wyjściach. Odpowiedzi te zostają wpisane do przerzutników, po czym sygnał „test” ponownie otrzymuje wartość „1”. Można teraz szeregowo wyprowadzić odpowiedzi na wyjście testowe.



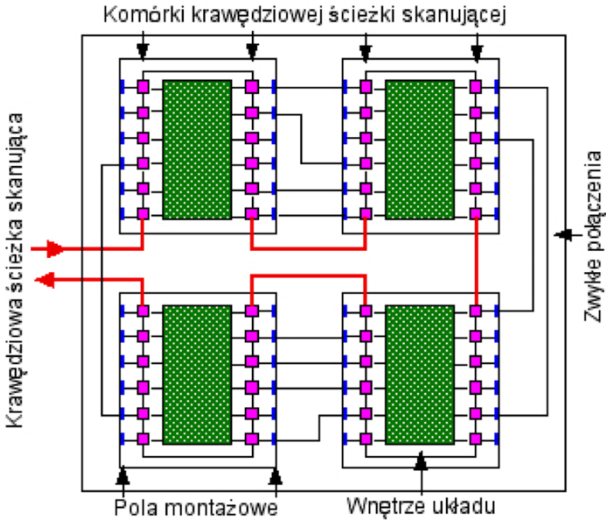
Rysunek 6-4. Idea łączenia przerzutników w łańcuch skanujący przy zastosowaniu bramek transmisyjnych sterowanych sygnałem "test"

W bibliotekach komórek standardowych dostarczanych przez producentów układów z reguły znajdziemy dwa rodzaje przerzutnika D – zwykły oraz przystosowany do tworzenia łańcuchów skanujących.

W dużych układach pojedynczy łańcuch skanujący byłby bardzo długi, przez co podawanie kolejnych wektorów testowych w trybie wpisywania szeregowego trwałoby dużo czasu. Dlatego w dużych układach można wprowadzić wiele niezależnych łańcuchów skanujących o umiarkowanej długości, które mogą działać w trybie testowania równocześnie.

Łańcuchy skanujące znalazły ważne zastosowanie także w testowaniu całych pakietów drukowanych, urządzeń i systemów. W przypadku wielowarstwowych płytek drukowanych wykonywanych nowoczesnymi technologiami mamy do czynienia z tym samym problemem, który występuje wewnątrz układów scalonych – brakiem dostępu do wewnętrznych węzłów. Co za tym idzie, testowanie pakietów drukowanych zawierających wiele układów scalonych jest równie trudne jak testowanie pojedynczych układów, a czasem nawet trudniejsze. Aby je ułatwić, wbudowuje się do wnętrza scalonych układów cyfrowych specjalne łańcuchy skanujące zwane brzegowymi lub krawędziowymi. Komórki tych łańcuchów znajdują się między wnętrzem układu, a polami

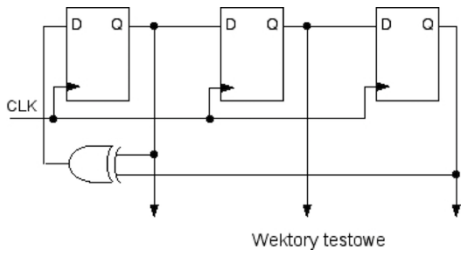
montażowymi. W normalnym trybie pracy są one „przezroczyste” dla sygnałów. W trybie testowania są one łączone w łańcuch skanujący obejmujący wszystkie układy na pakiecie. Łańcuch taki pozwala wpisywać szeregowo wektory testowe, które po wpisaniu trafiają bezpośrednio na wejścia układów scalonych, i po otrzymaniu odpowiedzi wyprowadzić je w trybie szeregowym. Łańcuch tego typu bywa nazywany krawędziową ścieżką skanującą (ang. „boundary scan path”). Tę ideę ilustruje rysunek 6-5.



Rysunek 6-5. Krawędziowa ścieżka skanująca dla czterech układów scalonych na pakiecie drukowanym

Krawędziowa ścieżka skanująca jest znormalizowana (norma IEEE 1149.1), dzięki czemu można swobodnie łączyć ze sobą układy różnych producentów. Większość katalogowych cyfrowych układów scalonych jest obecnie wyposażona w taką ścieżkę. Szczegółów technicznych nie będziemy tutaj omawiać.

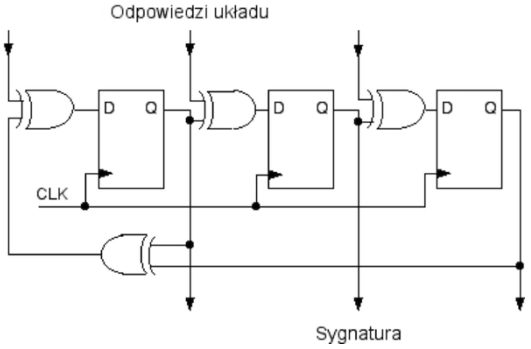
Testowanie staje się jeszcze prostsze, jeśli układ ma wbudowany mechanizm samotestowania. Samotestowanie można zrealizować wbudowując w układ generator wektorów testowych oraz układ analizujący prawidłowość odpowiedzi. Mogłoby się wydawać, że to jest do zrealizowania przez wprowadzenie do układu dwóch pamięci ROM: pamięci wektorów testowych i pamięci poprawnych odpowiedzi, oraz układu porównującego odpowiedzi otrzymane z prawidłowymi. Jest to jednak rozwiązanie niepraktyczne, bo potrzebne do tego pamięci z reguły zajmowałyby w układzie bardzo dużo miejsca. Stosowane zazwyczaj rozwiązanie polega na wprowadzeniu do układu generatora pseudolosowych sekwencji wektorów testowych oraz układu analizy odpowiedzi poddającego kolejne odpowiedzi kompresji prowadzącej do otrzymania t.zw. sygnatury - pojedynczego słowa binarnego, które ma jednoznacznie określoną wartość w przypadku układu sprawnego, natomiast przyjmuje inne wartości, jeśli w testowanym układzie jest uszkodzenie.



Rysunek 6-6. Trzybitowy generator wektorów pseudolosowych

Taki układ generuje w każdym takcie zegara nowy wektor różny od poprzedniego (pod warunkiem, że stany początkowe przerzutników to nie są same zera). Ciąg tych wektorów nie jest przypadkowy, zależy od stanów początkowych w przerzutnikach, ale ma cechy statystyczne ciągu losowego. Po wygenerowaniu 2^n-1 wektorów (*n* – liczba przerzutników) generowana sekwencja powtarza się. Jednak w testowaniu wystarcza zwykle wykorzystywanie tylko początkowej, znacznie krótszej sekwencji. W czasie testowania kolejne wektory

podawane są na wejścia testowanego układu, a odpowiedzi na wejście układu tworzącego sygnaturę. Taki układ może być zbudowany podobnie, jak generator wektorów pseudolosowych.



Rysunek 6-7. Układ tworzący trzybitową sygnaturę

W każdym takcie zegara tworzone jest na wyjściu układu nowe słowo binarne, które zależy od wszystkich poprzednich oraz od ostatniej odpowiedzi układu. Po zakończeniu testowania końcowe słowo jest sygnaturą. Porównywana jest ona z zapamiętaną sygnaturą poprawną (tj. otrzymaną z układu bez uszkodzeń).

Nic nie stoi na przeszkodzie, aby w dużym układzie stosować wszystkie omówione sposoby testowania jednocześnie, dobierając metodę testowania do charakteru i stopnia złożoności testowanego bloku.

Jak widać, testowanie jest na tyle złożonym zagadnieniem, że wymaga brania pod uwagę od samego początku procesu projektowania układu.

6.2.3 Układy, których poprawne działanie ma krytyczne znaczenie

W niektórych zastosowaniach błędne działanie układów scalonych i systemów z tymi układami może spowodować wielkie i nieodwracalne szkody: katastrofę, utratę życia ludzkiego, nienaprawialną szkodę w mieniu wielkiej wartości. Problemem jest wówczas nie tylko zbadanie czy układ po wyprodukowaniu działa prawidłowo, ale także kontrola prawidłowości działania układu bądź systemu w czasie pracy. W takim przypadku stosuje się redundancję z „głosowaniem”. Układ lub cały system jest powielony trzykrotnie, i wszystkie trzy układy lub systemy pracują równocześnie otrzymując te same dane. Jeśli trzy wyniki są zgodne, uznaje się, że są one prawidłowe. Jeśli jeden z wyników różni się od dwóch pozostałych, za prawidłowe uznawane są te dwa (bo prawdopodobieństwo wystąpienia identycznej awarii w tym samym momencie w dwóch identycznych systemach jest pomijalnie małe), a trzeci wynik jest odrzucany z równoczesnym sygnałem, że system dający błędny wynik jest uszkodzony i musi być przy najbliższej okazji wymieniony. Jest to sposób kosztowny, ale stosowany m.in. w technice lotniczej, kosmicznej i innych krytycznych zastosowaniach.

I na koniec warto dodać jeszcze jedno ważne spostrzeżenie. Troska o niezawodność zaczyna się na etapie projektu. Wiele tragicznych wydarzeń (np. katastrof lotniczych zawinionych przez układy elektronicznego sterowania samolotem) spowodowanych było nie awariami w trakcie działania systemów elektronicznych, lecz przyjęciu na samym początku projektowania założeń błędnych lub nie biorących pod uwagę wszystkich możliwych sytuacji i wydarzeń.

7 Rada na zakończenie

Aby lepiej wczuć się w problematykę cyfrowych układów scalonych, warto przerobić praktyczne ćwiczenia: symulacje logiczne w programie „Dsch” i projektowanie prostego układu w programie „Microwind”. Aby sprawnie posługiwać się tymi programami, możesz obejrzeć krótkie prezentacje wideo.