2019

Układy scalone

CZĘŚĆ I: MIKROELEKTRONIKA Z LOTU PTAKA

WIESŁAW KUŹMICZ

PÓŁPRZEWODNIK, UKŁAD SCALONY, MIKROPROCESOR, UKŁAD CYFROWY, UKŁAD ANALOGOWY, DIODA, TRANZYSTOR MOS, TRANZYSTOR BIPOLARNY, PROCES PRODUKCYJNY, KOSZT UKŁADU, UZYSK, ROZRZUT PRODUKCYJNY, MODEL BIZNESOWY

CZĘŚĆ PIERWSZA MATERIAŁÓW OMAWIAJĄCYCH UKŁADY SCALONE ZAWIERA WPROWADZENIE, A W TYM: RYS HISTORYCZNY MIKROELEKTRONIKI, OMÓWIENIE JEJ ZNACZENIA WE WSPÓŁCZESNEJ GOSPODARCE I CYWILIZACJI, PODSTAWY DZIAŁANIA TRANZYSTORÓW, BUDOWĘ BRAMEK LOGICZNYCH, OMÓWIENIE PROCESÓW PRODUKCYJNYCH MIKROELEKTRONIKI ORAZ ASPEKTY EKONOMICZNE PROJEKTOWANIA I PRODUKCJI UKŁADÓW SCALONYCH.

Spis treści

1 Wstęp: o czym tu będzie mowa

Droga Czytelniczko, Drogi Czytelniku: tutaj znajdziesz ogólne wprowadzenie do mikroelektroniki. Dowiesz się, do czego jest nam potrzebna, jak się zaczęła i rozwijała, jakie jest jej znaczenie we współczesnym świecie i jakie są „motory napędowe" jej rozwoju. Poznasz w zarysie, z jakich elementów składają się układy scalone i na jakiej zasadzie działają te elementy. Dowiesz się, jak powstają struktury układów scalonych, a także – od czego i jak zależy koszt układu scalonego. Jest to przygotowanie do głębszego poznania mikroelektroniki, w tym do praktycznego poznania podstaw projektowania układów scalonych.

Możesz zadać pytanie: a po co uczyć się, jak projektuje się układy scalone?
Otóż bez układów scalonych nie można dziś wyobrazić sobie żadnego wyrobu elektronicznego. Oprócz standardowych układów, zwanych katalogowymi, praktycznie w każdym urządzeniu elektronicznym znajdziemy dziś układy scalone zwane specjalizowanymi – zaprojektowane i wytwarzane specjalnie do tego urządzenia. Takich układów nigdzie nie kupimy gotowych. Ich zaprojektowanie jest zadaniem konstruktora urządzenia. Znajomość zasad projektowania specjalizowanych układów scalonych zalicza się dziś do podstawowych umiejętności inżyniera elektronika. Nawet konstruktor, który sam nie projektuje układów scalonych, powinien wiedzieć, jak to się robi, aby móc porozumieć się z projektantem, który zaprojektuje układy specjalizowane do konstruowanego przez niego urządzenia. Układy specjalizowane zwane są często układami ASIC (od angielskiego terminu „Application-Specific Integrated Circuit").

A czy taka wiedza się w Polsce w ogóle do czegoś przydaje? Przecież nie mamy fabryk układów scalonych! Rzeczywiście, w Polsce jest obecnie tylko jedna linia produkcyjna w Instytucie Technologii Elektronowej w Warszawie. Służy ona głównie do celów doświadczalnych i do wytwarzania niewielkich ilości nietypowych wyrobów półprzewodnikowych. Mimo to polski inżynier ma równie dobry i łatwy dostęp do możliwości wytwarzania specjalizowanych układów scalonych, jak inżynier we Francji, Niemczech, Japonii czy też USA. Czytaj dalej, a dowiesz się więcej o tym.

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

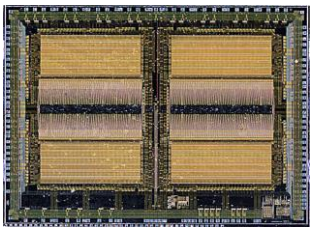
A co trzeba umieć, by bez kłopotów dalej poznawać układy scalone?
Bardzo przydatna byłaby znajomość - przynajmniej w podstawowym zakresie - zasad działania elementów półprzewodnikowych, ich charakterystyk i parametrów. Dobrze byłoby też orientować się w zasadach działania podstawowych układów elektronicznych, cyfrowych i analogowych, oraz w podstawach teorii układów logicznych. Ale nie martw się, jeśli są to zagadnienia Ci słabo znane lub w ogóle nieznane. Znajdziesz tu wprowadzenie, które nie powinno być bardzo trudne. Nie będzie zaawansowanej fizyki ani matematyki, a tylko to, co jest potrzebne, by poznać podstawowe pojęcia i rozumieć dalszy materiał. Oczywiście pogłębienie wiedzy będzie niezbędne, aby stać się w pełnym znaczeniu tego słowa specjalistą od układów scalonych.

2 Co to jest mikroelektronika i do czego jest nam potrzebna

Tu będzie mowa o tym, jak powstały pierwsze układy scalone, jak rozwijała się dziedzina techniki zwana mikroelektroniką i jakie jest jej obecne miejsce w gospodarce i w naszej cywilizacji.

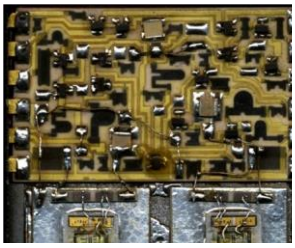
2.1 Co to jest układ scalony

Mikroelektronika jest to dziedzina techniki zajmująca się wytwarzaniem, projektowaniem i zastosowaniami układów scalonych.



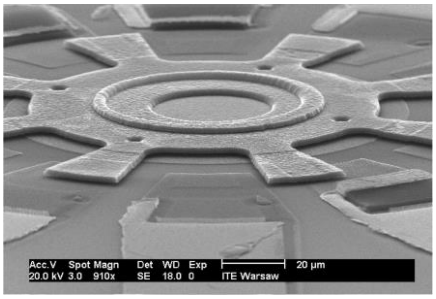
Rysunek 2-1. Mikrofotografia monolitycznego układu scalonego, wymiary: 10x7 mm

Układem scalonym nazywamy układ elektroniczny złożony z wielu elementów (tranzystorów i innych elementów) i połączeń elektrycznych między nimi, w sposób nierozdzielny umieszczonych na wspólnym podłożu. Są to przy dzisiejszym stanie technologii niemal wyłącznie monolityczne półprzewodnikowe układy scalone. Nazywamy je monolitycznymi i półprzewodnikowymi, ponieważ są to struktury wykonane w całości wewnątrz i na powierzchni płytki półprzewodnika, którym jest (z rzadkimi wyjątkami) krzem. Wszystkie elementy monolitycznego układu scalonego, a także połączenia elektryczne między tymi elementami, wytwarzane są równocześnie, w wyniku wykonania szeregu operacji technologicznych tworzących wewnątrz oraz na powierzchni płytki półprzewodnikowej obszary domieszkowane (tworzące struktury tranzystorów), a także obszary przewodzące i dielektryczne.



Rysunek 2-2. Hybrydowy układ scalony, wymiary: 5x5 cm

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



Rysunek 2-3. Mikrofotografia mikromechanizmu krzemowego

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

A co to jest półprzewodnik, obszar domieszkowany, przewodzący, dielektryczny, tranzystor? Nie wiesz? Czytaj dalej!

2.2 Trochę historii

2.2.1 Początki

W roku 1874 niemiecki fizyk Karl Ferdynand Braun zaobserwował, że styk metalowej igły z niektórymi kryształami (przykład: galena, czyli siarczek ołowiu) przewodzi prąd elektryczny asymetrycznie: opór elektryczny zależy od kierunku przepływu prądu. Była to pierwsza obserwacja odnosząca się do zjawisk występujących w półprzewodnikach. Zjawisko zaobserwowane przez Brauna znalazło zastosowanie praktyczne w postaci detektora kryształkowego: prymitywnej diody wykorzystywanej aż do drugiej wojny światowej w najprostszych radioodbiornikach. Do dziś stosowane są ostrzowe diody germanowe – współczesna wersja diody Brauna, w której galena zastąpiona została przez kryształ germanu – pierwiastka będącego półprzewodnikiem.

Zrozumienie zjawisk zachodzących w półprzewodnikach było możliwe dzięki rozwojowi fizyki w pierwszej połowie XX wieku: poznaniu budowy atomu, fizyce kwantowej, fizyce ciała stałego. Zjawiska na styku metal-półprzewodnik opisał teoretycznie niemiecki fizyk Walter Schottky w latach 20 XX wieku. Prace nad wykorzystaniem zjawisk w półprzewodnikach zostały zintensyfikowane w latach II Wojny Światowej w związku z rozwojem techniki radarowej. Wkrótce po wojnie fundamentalne prace teoretyczne i doświadczalne zespołu uczonych amerykańskich – Williama Shockleya, Johna Bardeena i Waltera Brattaina doprowadziły do powstania pierwszych tranzystorów: tranzystora ostrzowego (1947) i złączowego (1951). Tranzystor ostrzowy był elementem o niewielkiej przydatności i małej trwałości, więc szybko wyszedł z użycia, natomiast tranzystor złączowy, zwany obecnie tranzystorem bipolarnym, we współczesnych wersjach technologicznych jest wytwarzany i stosowany do dziś. Jednak – obok tranzystorów - niemniej ważnym wynikiem prac Shockleya, Bardeena i Brattaina był ścisły i spójny opis ilościowy zjawisk w diodzie półprzewodnikowej i tranzystorze bipolarnym, który stanowi do dziś, wraz z późniejszymi rozszerzeniami i uzupełnieniami, podstawę inżynierskiej wiedzy w dziedzinie techniki półprzewodnikowej.

Tranzystory bipolarne były podstawowymi elementami pierwszych układów scalonych, lecz dziś w mikroelektronice wykorzystywane są niemal wyłącznie tranzystory innego rodzaju: tranzystory unipolarne z izolowaną bramką zwane w skrócie tranzystorami MOSFET lub po prostu MOS. Ich historia rozpoczęła się nawet wcześniej, niż tranzystorów bipolarnych. Julius Lilienfeld, fizyk urodzony we Lwowie w zaborze austriackim, lecz działający w Niemczech, a następnie w USA, opatentował w latach 1925 i 1928 struktury będące pierwowzorami tranzystorów unipolarnych. Nie było jednak w tamtych czasach technologii umożliwiających wytworzenie takich tranzystorów. Pierwsze działające tranzystory MOSFET zawdzięczamy pracom prowadzonym w laboratoriach Bella w USA na przełomie lat 50 i 60 XX wieku.

I jeszcze jeden wynalazek, bez którego nie byłoby mikroelektroniki w jej współczesnej postaci: technologia wytwarzania monokryształów, czyli kryształów o idealnie regularnej budowie wewnętrznej. Wynalazcą metody wytwarzania monokryształów był polski chemik Jan Czochralski, urodzony w Kcyni w zaborze pruskim, działający najpierw w Niemczech, a od 1928 roku profesor Politechniki Warszawskiej. Pierwszy raz Czochralski zaobserwował zjawisko monokrystalizacji (ale nie półprzewodnika, lecz metalu – cyny) w roku 1916, a opublikował w 1918. Metodę Czochralskiego wykorzystano po raz pierwszy do produkcji monokryształów półprzewodnika (germanu) w roku 1950 w laboratoriach Bella w USA. Dziś metoda Czochralskiego używana jest powszechnie do produkcji monokryształów krzemu będących podstawowym materiałem mikroelektroniki.

A co to jest dioda, tranzystor bipolarny, tranzystor MOS? Jak to działa? Jeśli nie wiesz, czytaj dalej.

2.2.2 Miniaturyzacja układów elektronicznych, pierwsze układy scalone

W latach 50 i 60 XX wieku rozpoczął się szybki rozwój techniki wojskowej, lotniczej, rakietowej, i w końcu kosmicznej. Potrzebne były urządzenia elektroniczne o małych wymiarach i masie, niewrażliwe na wstrząsy i przeciążenia, nie wymagające wysokich napięć zasilających, nie pobierające dużej mocy i bardzo niezawodne. Próżniowe lampy elektronowe, które były podstawowymi elementami elektroniki od lat 20 XX wieku, nie spełniały żadnego z tych wymagań. Oczywiste było zwrócenie się ku tranzystorom. Zaczęły powstawać techniki montażu tranzystorów i innych elementów takie, jak obwody drukowane, montaż hybrydowy na podłożach ceramicznych czy trójwymiarowe mikromoduły. Prawdziwy przełom nastąpił z chwilą wynalezienia w końcu lat 50 XX wieku planarnej technologii produkcji tranzystorów bipolarnych. Stąd był już tylko jeden krok do powstania pierwszych monolitycznych układów scalonych. Za wynalazców monolitycznych układów scalonych uważani są Jack Kilby i Robert Noyce, którzy niezależnie od siebie zaprezentowali pierwsze struktury monolitycznych układów scalonych w 1958 roku. Początkowo były to układy zbudowane z kilku do kilkunastu tranzystorów bipolarnych i innych elementów pełniące bardzo proste funkcje. Dziś mówimy, że były to układy małej skali integracji. Jednak bardzo szybko znalazły one liczne zastosowania. Większość procesorów komputerów produkowanych w końcu lat 60 i w latach 70 XX wieku była zbudowana z dużej liczby takich układów.

W roku 1964 zaprezentowane zostały pierwsze układy scalone z tranzystorami MOS. Był to kolejny wielki krok naprzód. Struktury tranzystorów MOS zajmowały w układzie scalonym znacznie mniej miejsca, niż struktury tranzystorów bipolarnych, co umożliwiło produkcję układów o coraz bardziej złożonych funkcjach, zawierających na jednej płytce krzemowej na początku setki, a później tysiące i dziesiątki tysięcy tranzystorów. Dziś liczba tranzystorów w jednym układzie scalonym przekracza miliard – mówimy o układach VLSI od angielskiego określenia „very large scale of integration”.

Gdy w 1966 roku Robert Dennard przedstawił koncepcję pamięci półprzewodnikowej (dziś znanej jako pamięć dynamiczna – DRAM), a w roku 1970 pierwsze scalone pamięci półprzewodnikowe DRAM znalazły się w produkcji, możliwa stała się budowa urządzeń techniki cyfrowej zawierających wyłącznie półprzewodnikowe układy scalone. Odtąd aż do dziś trwa burzliwy rozwój mikroelektroniki i wszelkich wykorzystujących ją urządzeń i systemów.

Technologia planarna, pamięć dynamiczna? Wszystko się dalej wyjaśni.

2.2.3 Mikroprocesory, pamięci i „prawo Moore’a”

W końcu lat 60 XX wieku były już produkowane elektroniczne kalkulatory wykorzystujące układy scalone małej skali integracji. W roku 1971 w nowo powstałej firmie Intel zaprojektowano i wyprodukowano pierwszy układ scalony, który miał zastąpić kilka odrębnych układów stosowanych dotąd w kalkulatorach. Powstał układ o symbolu 4004, operujący na słowach czterobitowych, będący z punktu widzenia wewnętrznej architektury układem zasługującym na miano mikroprocesora. Jest on powszechnie uważany za pierwszy mikroprocesor, choć w rzeczywistości firma Texas Instruments miesiąc wcześniej wyprodukowała inny czterobitowy mikroprocesor TMS 1000 (zastosowany później w kalkulatorach tej firmy) i to właśnie Texas Instruments, a nie Intel, posiada amerykański patent na mikroprocesor. Wkrótce zarówno Intel, jak i inne firmy zaczęły wypuszczać na rynek coraz bardziej złożone układy mikroprocesorów. Pierwsze mikroprocesory 8-bitowe to 8008 (Intel, 1972) i jego rozbudowany następca 8080 (1974). Jednocześnie powstały mikroprocesory konkurencyjne, jak MC 6800 (Motorola, 1974), MOS 6502 (MOS Technology, 1975), Z80 (Zilog, 1976). Postępy technologii półprzewodnikowych, a w tym wprowadzenie do produkcji technologii CMOS, umożliwiły szybki rozwój architektur mikroprocesorów i wzrost ich szybkości działania przy równoczesnym spadku cen. Mikroprocesory

zaczęły być stosowane w sterowaniu urządzeniami przemysłowymi, w motoryzacji, w elektronice medycznej, a także w sprzęcie powszechnego użytku. Jednak prawdziwą rewolucję zapoczątkowały komputery osobiste – powstające najpierw jako projekty hobbystyczne, a nieco później już jako dojrzałe produkty rynkowe przeznaczone dla każdego, a nie tylko dla specjalistów. Pierwszym takim produktem był Apple II (1977), którego kolejne wersje produkowane były w wielkich ilościach do roku 1983. Jego głównym projektantem był Amerykanin polskiego pochodzenia Steve Wozniak. Projekty Wozniaka cechowały się techniczną wirtuozerią, co było jednym ze źródeł sukcesu Apple II. Pojawiły się wkrótce inne konkurencyjne komputery osobiste (Commodore, Atari, TRS-80 i in.). W roku 1981 wejście na rynek wielkiej firmy IBM z komputerem IBM PC zmieniło reguły gry. IBM zakupił system operacyjny od małej wówczas firmy Microsoft, a także opublikował specyfikację techniczną komputera, co zaowocowało powstaniem dziesiątków firm produkujących klony IBM PC oraz oprogramowanie. Ponadto komputer z IBM uzyskał w biznesie opinię „profesjonalnego", w odróżnieniu od Apple II uważanego za sprzęt do użytku domowego i dla hobbystów. Odpowiedzią Apple był komputer Macintosh (1984). W odróżnieniu od IBM PC, w którym wykorzystano 8-bitowy procesor Intela, Macintosh miał procesor Motoroli MC 68000 o mieszanej 16/32-bitowej architekturze i dużo wyższej sprawności obliczeniowej. Bogata jak na owe czasy grafika i „okienkowy" system operacyjny, koncepcyjnie wywodzący się z prac prowadzonych w latach 70 w laboratoriach firmy XEROX, stały się „znakiem firmowym" produktów Apple i umożliwiły przetrwanie i rozwój tej firmy mimo potężnej konkurencji kolejnych generacji komputerów wywodzących się od IBM PC, wyposażonych w końcu także w system Microsoft Windows wzorowany na systemie graficznym Macintosha.

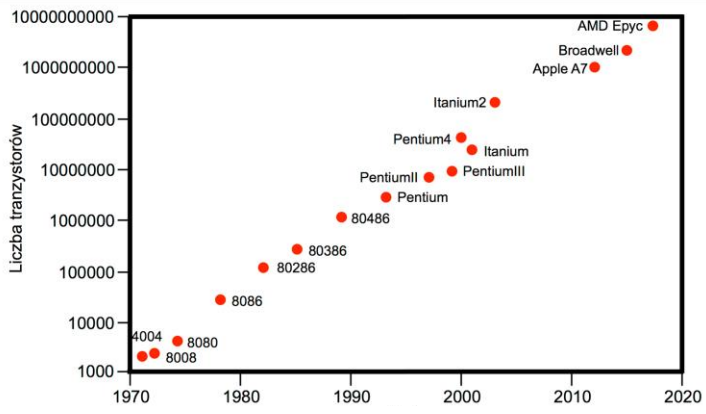
Błyskawiczny rozwój rynku komputerów osobistych spowodował, że mikroprocesory stały się jednym z dwóch głównych produktów mikroelektronicznych (drugim są pamięci półprzewodnikowe – o nich dalej). Od końca lat 80 XX wieku wyścig technologiczny w mikroelektronice dyktowany był głównie przez rosnące potrzeby kolejnych generacji mikroprocesorów. Komputery „IBM-kompatybilne" do dziś wykorzystują procesory o architekturze znanej jako „x86", wywodzącej się z 16-bitowego procesora Intela 8086 z lat 80 XX wieku. Do początku lat 90 XX wieku linia procesorów „68xxx" Motoroli skutecznie konkurowała z procesorami Intela. Potem konsorcjum Apple-IBM-Motorola wprowadziło na rynek procesory PowerPC o nowocześniejszej architekturze i większej wydajności od ówczesnych procesorów Intela. W latach 1994 – 2006 procesory PowerPC były stosowane w komputerach Apple, a także w wielu innych urządzeniach, lecz w końcu zaczęły przegrywać wyścig konkurencyjny z procesorami Intela z przyczyn ekonomicznych – Intel miał ogromne zasoby finansowe umożliwiające wprowadzanie mniej więcej co dwa lata nowych generacji technologii produkcji. Takich możliwości nie mieli producenci procesorów PowerPC (Motorola w roku 2004 w ogóle zrezygnowała z produkcji układów scalonych tworząc odrębną firmę Freescale Semiconductor). Intel jako dostawca procesorów do komputerów wygrał dzięki temu, że procesory o mniej wydajnej architekturze „x86", ale produkowane w nowocześniejszym procesie, były szybsze i pobierały mniej mocy od procesorów PowerPC. Od roku 2006 również komputery Apple mają procesory Intela. W ten sposób procesory „x86" stały się *de facto* standardem w świecie komputerów osobistych, a ich główny producent – Intel – praktycznie monopolistą. Ogromna skala produkcji tych procesorów dała firmie Intel gigantyczne przychody, dzięki którym Intel na wiele lat stał się liderem w rozwoju technologii produkcji, co odbywało się głównie przez zmniejszanie wymiarów tranzystorów, czemu jednak towarzyszyło wiele innych innowacji. W ostatnich latach pozycja Intela jako lidera technologicznego uległa jednak zachwianiu. Pojawiła się potężna konkurencja: urządzenia mobilne – smartfony i tablety, a do nich energooszczędne mikroprocesory ARM.

ARM jest firmą założoną w Wielkiej Brytanii w roku 1990. Nie jest i nigdy nie była producentem układów scalonych. Zajmuje się wyłącznie projektowaniem energooszczędnych (i z każdą generacją coraz bardziej wydajnych) mikroprocesorów ARM i sprzedażą licencji producentom tych mikroprocesorów. Licencjobiorcy

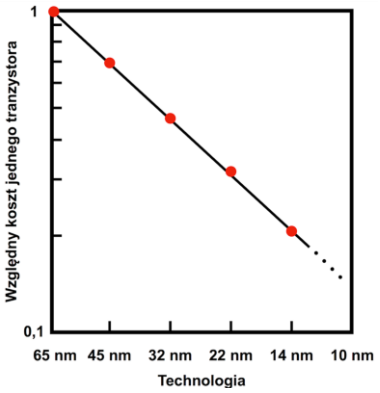
często wykorzystują rdzenie procesorów ARM we własnych bardziej złożonych układach tworzących systemy jednoukładowe („system on chip”). Niemal wszystkie produkowane na świecie smartfony i tablety wykorzystują procesory ARM lub bardziej złożone układy z rdzeniami ARM. ARMy są wykorzystywane także w wielu innych urządzeniach, jak odbiorniki GPS, konsole do gier, aparaty fotograficzne i kamery, telewizory cyfrowe, urządzenia elektromedyczne. Głównymi producentami układów scalonych dla tego olbrzymiego rynku są tajwańska firma TSMC (m.in. główny dostawca dla Apple) oraz koreański Samsung, ale procesory ARM są też w ofercie wielu innych firm, w tym m.in. firmy STMicroelectronics - najbardziej zaawansowanej technologicznie firmy europejskiej. Pod względem stopnia zaawansowania technologii produkcji TSMC i Samsung doścignęły, a pod pewnymi względami wyprzedzają Intela.

Równolegle do mikroprocesorów rozwijały się technologie produkcji pamięci półprzewodnikowych, przede wszystkim pamięci dynamicznych DRAM, które są powszechnie stosowane w pamięciach operacyjnych praktycznie wszystkich systemów cyfrowych – nie tylko komputerów, ale i tabletów, smartfonów i wielu innych urządzeń. Pamięci DRAM wymagają specyficznego procesu produkcyjnego, w którym wyspecjalizowały się głównie dwie firmy z Korei Południowej: Samsung i SK Hynix, trzecim z głównych producentów jest amerykańska firma Micron Technology. Pamięci DRAM są to pamięci ulotne – ich zawartość zanika po wyłączeniu zasilania. Odrębnym, coraz ważniejszym rodzajem pamięci są pamięci nieulotne zwane potocznie pamięciami „flash”. Przechowują one swą zawartość także, gdy nie są zasilane. To one znajdują się w popularnych kartach pamięci do aparatów fotograficznych i kamer, pamięciach USB (zwanych popularnie „pendrive”) oraz pamięciach masowych komputerów zwanych potocznie „dyskami SSD”, które coraz śmielej wypierają tradycyjne dyski magnetyczne z wirującym talerzem. Pamięci „flash” także wymagają szczególnej, niełatwej technologii produkcji. Tu też liderem jest Samsung, ale drugim producentem jest japońska Toshiba, ale i kilka innych firm, m.in. Western Digital, SK Hynix, Micron Technology i Intel.

Użytkownicy oczekują, że kolejne generacje komputerów, laptopów, tabletów, smartfonów, konsol do gier itp. będą coraz wydajniejsze, a producenci starają się zaspokajać to oczekiwanie, bo stagnacja oznacza utratę rynku. Konsumenci nie kupią nowego sprzętu, jeśli nie będzie wyraźne lepszy od tego, który już posiadają. To zmusza producentów układów scalonych – głównie mikroprocesorów i pamięci – do wytwarzania coraz wydajniejszych procesorów i coraz pojemniejszych i szybciej działających pamięci. W technologii CMOS wzrost szybkości działania układów cyfrowych osiąga się głównie przez zmniejszanie wymiarów tranzystorów. Krytycznym wymiarem jest długość kanału tranzystora, i kolejne generacje technologii określa się podając ten wymiar. W latach 60 XX wieku było to 10 mikrometrów, dziś (początek 2019 roku) najmniejsze tranzystory w układach produkowanych seryjnie mają kanały o długości 7 nanometrów, czyli ponad tysiąc razy mniej. Zmniejszanie wymiarów tranzystorów oznacza zarazem, że w układzie scalonym można ich zmieścić coraz więcej, a przy tym koszt pojedynczego tranzystora spada. Już w latach 60 XX wieku Gordon Moore, jeden z założycieli Intela, stwierdził, że co 18 miesięcy najkorzystniejsza z punktu widzenia kosztu układu scalonego liczba tranzystorów w układzie rośnie dwukrotnie. Ta tendencja, zwana „prawem Moore’a”, utrzymuje się od ponad 40 lat, z tą różnicą, że najkorzystniejsza ekonomicznie liczba tranzystorów w układzie podwaja się nie co 18 miesięcy, lecz co dwa lata.



Rysunek 2-4. Prawo Moore'a na przykładzie wybranych mikroprocesorów



Rysunek 2-5. Koszt jednego tranzystora w układzie scalonym odniesiony do kosztu w technologii 65 nm (źródło danych: Intel, 2012)

Procesy produkcyjne w każdej kolejnej generacji technologii są coraz kosztowniejsze. Dotyczy to zarówno kosztu opracowania i uruchomienia nowej technologii, jak i kosztu nowocześniejszej aparatury technologicznej oraz kosztu samego procesu produkcyjnego. W rezultacie koszt jednostki powierzchni (powiedzmy – jednego centymetra kwadratowego) wyprodukowanego układu scalonego rośnie. Jednak liczba tranzystorów mieszczących się na tej powierzchni rośnie szybciej, a więc koszt pojedynczego tranzystora nie rośnie, lecz maleje. Oznacza to, że układy produkowane w nowocześniejszej technologii są nie tylko lepsze (szybsze, o bogatszej i wydajniejszej architekturze), ale i z reguły tańsze.

A więcej o procesach produkcyjnych mikroelektroniki, technologii CMOS, pamięciach ulotnych i nieulotnych przeczytasz dalej.

2.2.4 I co dalej – czy już koniec rozwoju?

Zmniejszanie wymiarów tranzystorów ma oczywistą fizyczną granicę, są nią wymiary atomu. Ocenia się, że w praktyce najmniejsza możliwa długość kanału tranzystora to 3 nanometry. Mniejsze tranzystory można by w zasadzie zrobić, ale nie będą działać. Praktyczne bariery rozwoju pojawiły się już wcześniej. Mimo iż najbardziej zaawansowani producenci mają obecnie w próbnej produkcji układy z tranzystorami o długości kanału 7 nanometrów, a obiecują 5 nanometrów, to w gruncie rzeczy zmniejszanie wymiarów tranzystorów przestaje się opłacać. Koszty najnowszych technologii rosną zawrotnie, mniejsze tranzystory przestały być tańsze. Na dodatek praktycznie osiągalna szybkość działania układów cyfrowych jest ograniczona mocą pobieraną przez układ – tym większą, im jest on szybszy. Większa moc to więcej ciepła wydzielającego się w układzie, a już dość dawno osiągnięta została granica możliwości odprowadzania wydzielanego przez układ ciepła. Reguła zwana „prawem Moore'a", która przez dziesiątki lat była w rozwoju mikroelektroniki rodzajem drogowskazu, przestaje działać.

W ostatnich latach powstały nowe warianty technologii CMOS, w których tranzystory mają radykalnie zmodyfikowane struktury: technologia z tranzystorami zwanymi FinFET i technologia z tranzystorami FDSOI. Tranzystory FinFET i tranzystory FDSOI pozwalają budować układy bardziej energooszczędne. Obie technologie stają się coraz bardziej rozpowszechnione i pozwalają produkować układy, które przy danej szybkości działania pobierają mniej mocy, co jest korzystne zwłaszcza w sprzęcie mobilnym. Warto dodać, że technologia FDSOI jest wynalazkiem europejskim, a jej twórcą jest prof. Tomasz Skotnicki, przez wiele lat odpowiedzialny za rozwój technologiczny w firmie STMicroelectronics, absolwent Politechniki Warszawskiej i obecnie profesor tej uczelni.

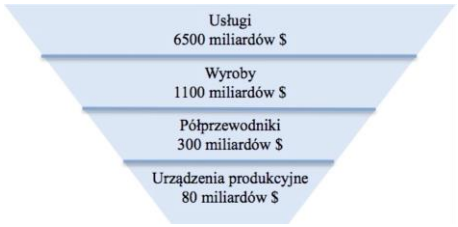
Nie ulega wątpliwości, że na dalszą metę postępy mikroelektroniki w jej obecnym kształcie, wykorzystującym krzem jako podstawowy materiał, a tranzystory MOS jako podstawowe elementy, nie będą już możliwe. Co dalej? Może nowe materiały, a może zupełnie nowe metody przetwarzania informacji? Czas pokaże.

Pobór mocy przez układy scalone, problemy i koszt najnowszych technologii, tranzystory FinFET i FDSOI – te tematy pojawią się dalej.

2.3 Mikroelektronika we współczesnym świecie

2.3.1 Rola mikroelektroniki w cywilizacji i gospodarce

Początkowo układy scalone były prymitywne technicznie, projektowanie ich odbywało się w dużym stopniu metodą prób i błędów, a produkcja była bardzo kosztowna. Mikroelektronikę w tych czasach uważano za niszową technologię do nielicznych zastosowań. Uważano, że tylko konieczność zapewnienia małych wymiarów i ciężaru sprzętu uzasadnia użycie w nim układów scalonych. Małe wymiary i ciężar układów scalonych są również dziś ich ważną zaletą. Istnieją tysiące zastosowań, w których właśnie małe wymiary i ciężar układów scalonych są niezastąpione. Wszyscy znamy wiele z tych zastosowań: elektroniczny zegarek, przenośny radiomagnetofon, kieszonkowy odtwarzacz plików mp3, telefon komórkowy, smartfon, laptop, karta kredytowa z mikroprocesorem, wszczepialny stymulator serca, komputer sterujący silnikiem samochodu, elektroniczny system sterowania lotem samolotu i wiele, wiele innych. Wszystkie te urządzenia i systemy nie dałyby się zrealizować, gdyby nie miniaturyzacja możliwa dzięki układom scalonym. Ale dziś, obok miniaturyzacji, dwie inne zalety układów scalonych są uważane za niezwykle ważne: wysoka niezawodność i niski koszt. Gdyby nie te dwie cechy układów scalonych, nie byłoby możliwe zbudowanie wielkich, obejmujących cały świat systemów przesyłania i przetwarzania informacji, jak zautomatyzowane sieci telefoniczne (stacjonarna, komórkowa, satelitarna), internet, sieci superkomputerowe, GPS. Te z kolei stały się środowiskiem technicznym, w którym możliwe było powstanie zupełnie nowych rodzajów usług. Gazety i książki czytamy w internecie lub z niego je pobieramy, w ten sam sposób słuchamy muzyki i oglądamy filmy, robimy zakupy w internetowych sklepach, załatwiamy przez sieć sprawy w banku i urzędach, komunikujemy się przez sieci społecznościowe, w blogach wyrażamy poglądy i prowadzimy dyskusje, wyszukujemy informacje w wyszukiwarkach i Wikipedii, programy telewizyjne oglądamy w wysokiej rozdzielczości, legitymujemy się paszportami biometrycznymi, lekarze dzięki internetowi i robotom wykonują na odległość skomplikowane operacje, uczeni oddaleni o tysiące kilometrów współpracują w badaniach mając wspólny dostęp do tych samych zbiorów danych, a dzięki sondom kosmicznym możemy oglądać obrazy z planet i komet. Rozpoczyna się właśnie epoka internetu rzeczy („Internet of Things" – IoT), gdy już nie tylko ludzie, ale maszyny i urządzenia będą połączone bezpośrednio z siecią, przesyłając i wymieniając informacje, dostarczając ludziom informacje i usługi. Można śmiało powiedzieć, że mikroelektronice zawdzięczamy współczesny kształt naszej cywilizacji.



Rysunek 2-6. Mikroelektronika w światowej gospodarce

Gospodarcze znaczenie mikroelektroniki ilustruje rysunek 2-6. Wartość rocznej produkcji wyrobów półprzewodnikowych (są to w przeważającej części układy scalone), nawet łącznie z wartością rocznej produkcji urządzeń dla przemysłu półprzewodnikowego (w sumie ok. 380 miliardów USD), jest w skali całej światowej gospodarki dość skromna. Jednak dzięki tej produkcji możliwe jest wytwarzanie wszelkich wyrobów elektronicznych lub zawierających elektronikę, a te wyroby z kolei wspierają wspomniany wyżej ogromny rynek usług. Szacuje się, że dzięki mikroelektronice funkcjonuje około 15% całej światowej gospodarki. Gdyby nagle na skutek jakiegoś kataklizmu przyrodniczego lub politycznego ustała produkcja układów scalonych u kilku największych światowych producentów (głównie na Dalekim Wschodzie), to w światowej gospodarce nastąpiłby wstrząs, przy którym kryzys lat 2007 – 2008 byłby zaledwie niewielkim wahnięciem koniunktury.

Na koniec dodajmy, że mikroelektronika ma też kluczowe znaczenie w technologiach wojskowych. Nie bez powodu dziedzinę tę intensywnie rozwijają kraje znajdujące się w stanie bezpośredniego zagrożenia militarnego: Korea Południowa, Tajwan, Izrael, lub mające ambicje militarnego panowania nad światem. Ten temat wykracza jednak poza zakres naszych rozważań.

2.3.2 Czynniki stymulujące rozwój mikroelektroniki

Jak wspomniano wyżej, dwie zalety układów scalonych są kluczowe: wysoka niezawodność i niski koszt. One właśnie stały się głównymi czynnikami napędowymi rozwoju mikroelektroniki.

Wzrost niezawodności systemów elektronicznych, jaki umożliwiła mikroelektronika, można dobrze zilustrować na przykładzie niezawodności komputerów. W latach 70 XX wieku średni czas bezawaryjnej pracy komputera wynosił typowo kilka godzin. Dziś średni czas bezawaryjnej pracy mikroprocesora, będącego pod względem mocy obliczeniowej odpowiednikiem komputera z lat siedemdziesiątych, wynosi kilkadziesiąt lat. Oznacza to wzrost niezawodności, mierzony czasem bezawaryjnej pracy, w stosunku 1:10 000, czyli o 4 rzędy wielkości. Przy takim poziomie niezawodności systemów elektronicznych, z jakim mieliśmy do czynienia przed początkiem rozwoju mikroelektroniki, nie byłoby żadnych szans realizacji tak wielkich systemów, jak te wymienione wyżej. Po prostu w tak wielkich systemach liczba uszkodzonych podzespołów i bloków byłaby w każdej chwili na tyle duża, że systemy te jako całość praktycznie nie nadawałyby się do użytku.

Komputery są także dobrym przykładem spadku kosztu urządzeń elektronicznych, jaki zawdzięczamy mikroelektronice. Koszt procesora komputera z połowy lat 70 XX wieku zawierał się w przedziale 10 000 USD - 1 000 000 USD, a dziś typowy koszt mikroprocesora zawiera się między 1 USD a 100 USD. Nastąpiła więc redukcja kosztu w stosunku 1:10 000, czyli także o 4 rzędy wielkości. Większość istniejących dziś i powszechnie używanych urządzeń elektronicznych mogłaby teoretycznie być produkowana już przed kilkudziesięciu laty, lecz byłyby one wówczas tak kosztowne, że ich produkcja nie miałaby ekonomicznego sensu, ponieważ nie znalazłyby nabywców lub użytkowników.

A skąd się bierze ten niski koszt układów scalonych i od czego zależy? I skąd się bierze niezawodność układów scalonych? Będzie już niedługo o tym mowa.

3 O tranzystorach, bramkach logicznych i układach elektronicznych

Tu będzie mowa o tym, z czego składają się układy scalone. Najpierw o tym, co to jest półprzewodnik i o mechanizmach przepływu prądu w półprzewodnikach, potem o tranzystorach – jak są zbudowane i jak działają. Dalej o podstawowych bramkach logicznych, a także parę słów o układach analogowych.

3.1 Tranzystor: co to takiego

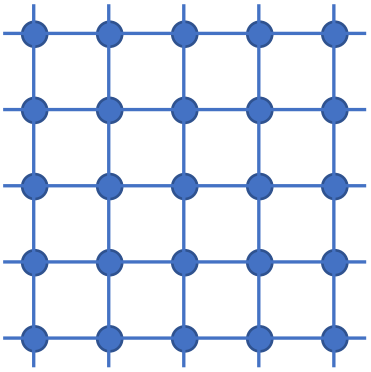
Jeśli rozumiesz mechanizm przepływu prądu w półprzewodniku, znasz zasady działania tranzystorów i innych elementów elektronicznych – możesz punkty 3.1.1 – 3.1.4 pominąć. Jeśli nie, znajdziesz w nich uproszczone, poglądowe omówienie tych zagadnień, bez wprowadzania nie całkiem łatwych pojęć mechaniki kwantowej i fizyki ciała stałego.

3.1.1 O elektronach i dziurach: jak płynie prąd w półprzewodniku

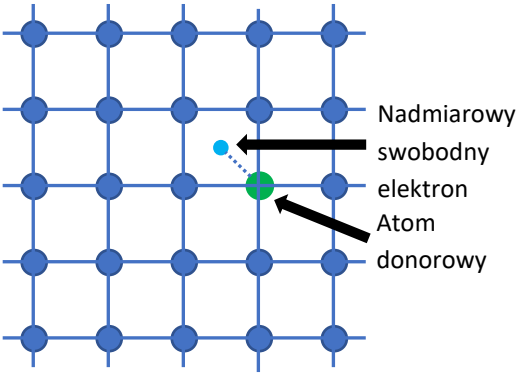
Prąd elektryczny w ciałach stałych jest to przepływ swobodnych elektronów, które są cząstkami elementarnymi niosącymi ładunek ujemny, od naładowanego ujemnie bieguna źródła prądu do naładowanego dodatnio bieguna źródła prądu. Z punktu widzenia przewodnictwa elektrycznego wszystkie ciała stałe można podzielić na trzy grupy: dielektryków (inaczej: izolatorów), przewodników i półprzewodników. W dielektrykach swobodnych elektronów praktycznie nie ma – wszystkie są związane silnymi więzami z atomami. W przewodnikach swobodnych elektronów jest dużo, z łatwością płynie w nich prąd elektryczny. Półprzewodniki w stanie czystym są zbliżone do dielektryków. Swobodnych elektronów jest w nich bardzo mało. Niektóre półprzewodniki przewodzą prąd bardzo słabo, inne praktycznie wcale. Jednak ilość swobodnych elektronów w półprzewodniku

może być powiększona, i to o wiele rzędów wielkości. Jeżeli temperatura półprzewodnika zostanie mniej lub bardziej podwyższona, pewna ilość elektronów zostanie uwolniona z więzów z atomami, i półprzewodnik będzie mógł przewodzić prąd. Jest to przewodnictwo zwane samoistnym. Inny sposób nadania półprzewodnikowi zdolności do przewodzenia prądu polega na dodaniu do półprzewodnika atomów specjalnej domieszki. Ten właśnie sposób służy do budowy struktur tranzystorów i innych elementów półprzewodnikowych układów scalonych. Jednak półprzewodnik różni się od dielektryków i przewodników nie tylko tym, że przewodnictwo elektryczne można w nim regulować w bardzo szerokim zakresie przez dodawanie domieszek, ale przede wszystkim tym, że w półprzewodniku obserwujemy dwa mechanizmy przepływu prądu – elektronowy i dziurowy.

W układach scalonych i większości innych struktur półprzewodnikowych materiałem wyjściowym jest półprzewodnik w postaci płytki monokrystalicznej. Monokryształ cechuje się regularnym rozmieszczeniem atomów; miejsca, w których się one znajdują, nazywamy węzłami sieci krystalicznej. W krzemie każdy atom w monokryształe związany jest czterema elektronami z czterema sąsiednimi atomami, co w uproszczeniu w dwóch wymiarach można przedstawić jak na rysunku 3-1. W idealnym monokryształe wszystkie elektrony są związane z atomami, swobodnych elektronów nie ma i dlatego taki monokryształ zachowuje się jak dielektryk. W wysokiej temperaturze energia drgań atomów powoduje uwalnianie niektórych elektronów – pojawia się przewodnictwo samoistne.

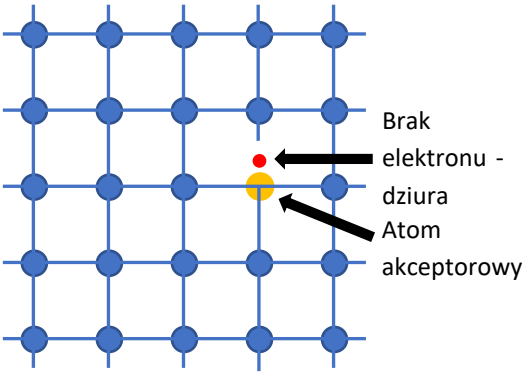


Rysunek 3-1. Idealny dwuwymiarowy monokryształ



Rysunek 3-2. Monokryształ z atomem donorowym i swobodnym elektronem

Jeżeli do monokryształu półprzewodnika dodane zostaną atomy pierwiastka pięciowartościowego, mającego pięć elektronów na zewnętrznej orbicie, i atomy te znajdą się w węzłach sieci krystalicznej zamiast atomów półprzewodnika, to cztery elektrony zostaną związane z sąsiednimi atomami półprzewodnika, a piąty pozostanie swobodny. Może on się przemieszczać w sieci krystalicznej półprzewodnika. Taki półprzewodnik może zatem przewodzić prąd. Mówimy, że jest to półprzewodnik domieszkowany atomami donorowymi. W przypadku krzemu mogą to być atomy fosforu lub arsenu.



Rysunek 3-3. Monokryształ z atomem akceptorowym i dziurą

Do monokryształu półprzewodnika można też dodać atomy pierwiastka trójwartościowego, który ma na zewnętrznej orbicie tylko trzy elektrony. Mówimy, że mamy do czynienia z półprzewodnikiem domieszkowanym atomami akceptorowymi. W przypadku krzemu mogą to być atomy boru. Gdy atom taki znajdzie się w węźle sieci krystalicznej, to do wiązania z jednym z sąsiednich atomów brakuje elektronu. Nazywamy takie puste miejsce dziurą. Obecność dziur powoduje, że półprzewodnik może przewodzić prąd, ponieważ do pustego miejsca – dziury – może się przemieścić elektron z sąsiedniego atomu. Wówczas w tym atomie powstaje dziura, którą może wypełnić kolejny

elektron. W ten sposób elektrony wędrują między dziurami zachowując się jak elektrony swobodne, a dziury przemieszczają się w przeciwnym kierunku.

Zarówno w półprzewodniku domieszkowanym donorami, jak i akceptorami, prąd elektryczny płynie dzięki ruchowi elektronów. Jednak badania w dziedzinie fizyki ciała stałego wykazały, że ruch dziur w półprzewodniku domieszkowanym akceptorami opisany jest dokładnie takimi samymi zależnościami matematycznymi, jak ruch elektronów w półprzewodniku domieszkowanym donorami. Wygodnie jest więc przyjąć, że mamy do dyspozycji dwa rodzaje domieszkowanych półprzewodników: półprzewodniki domieszkowane donorami, w których nośnikami ładunku przewodzącymi prąd są ujemne elektrony, i półprzewodniki domieszkowane akceptorami, w których nośnikami ładunku przewodzącymi prąd są dodatnie dziury.

Wprowadzimy teraz pojęcie koncentracji, może to być koncentracja atomów domieszki, koncentracja elektronów lub koncentracja dziur.

DEFINICJA

Koncentracją atomów, elektronów lub dziur w półprzewodniku nazywamy odpowiednio liczbę atomów, elektronów lub dziur w jednostce objętości. Zwyczajowo przyjęto, że tą jednostką jest centymetr sześcienny.

Fizyka ciała stałego dostarcza nam proste zależności między koncentracjami atomów domieszek, elektronów i dziur. Mianowicie, jeśli w półprzewodniku występują tylko atomy donorowe, to koncentracja swobodnych elektronów (w skrócie mówimy – koncentracja elektronów) jest praktycznie równa koncentracji atomów donorowych: $n = N_D$, gdzie n oznacza koncentrację elektronów, a N_D oznacza koncentrację atomów donorowych (w skrócie – koncentrację donorów). Jeśli w półprzewodniku występują tylko atomy akceptorowe, to koncentracja dziur jest praktycznie równa koncentracji atomów akceptorowych: $p = N_A$, gdzie p oznacza koncentrację dziur, a N_A oznacza koncentrację atomów akceptorowych (w skrócie – koncentrację akceptorów). Te proste zależności odnoszą się jednak do półprzewodników domieszkowanych jednorodnie (czyli takich, w których koncentracja domieszki jest w każdym punkcie taka sama) oraz znajdujących się w stanie zwanym stanem równowagi termodynamicznej, co oznacza, że temperatura półprzewodnika jest w każdym punkcie taka sama i nie oddziałuje nań żadne zewnętrzne pole elektryczne.

Fizyka ciała stałego dostarcza nam jeszcze jedną ważną zależność: w półprzewodniku znajdującym się w stanie równowagi termodynamicznej (ale niekoniecznie jednorodnym) iloczyn koncentracji elektronów i dziur jest stały:

Równanie 3-1

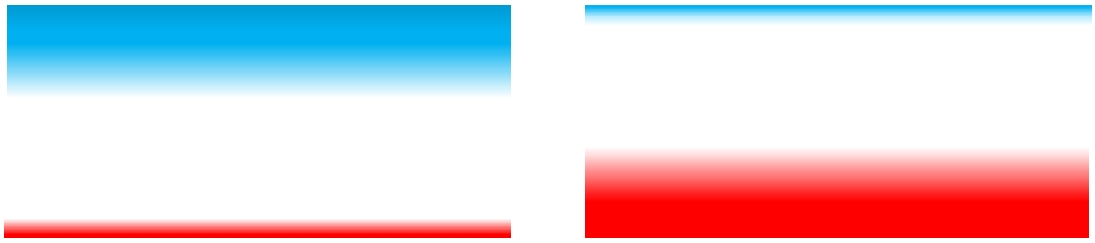
$$np = n_i^2$$

gdzie n_i^2 jest to kwadrat samoistnej koncentracji elektronów – takiej, jaka by istniała, gdyby w półprzewodniku nie było jakichkolwiek domieszek. Jest to wielkość bardzo silnie rosnąca ze wzrostem temperatury. Dla krzemu w temperaturze pokojowej (umownie przyjmuje się, że jest to 27° C) wartość n_i wynosi około 10^{10} cm^{-3}. Typowe koncentracje domieszek w krzemie zawierają się w przedziale 10^{14} cm^{-3} – 10^{20} cm^{-3}. Wynika z tego wszystkiego, że tam, gdzie jest dużo elektronów, tam jest bardzo mało dziur – i odwrotnie. Ta reguła przestaje obowiązywać w znacznie podwyższonej temperaturze, gdy pojawia się duża liczba elektronów i dziur tworzących

przewodnictwo samoistne, czyli gdy koncentracja samoistna n_i staje się porównywalna z koncentracją domieszek. Gdy dominuje przewodnictwo samoistne, domieszkowanie półprzewodnika nie ma już istotnego wpływu na koncentrację elektronów bądź dziur. Opisane dalej elementy półprzewodnikowe – diody, tranzystory – przestają prawidłowo działać.

Elektrony i dziury nazywamy łącznie nośnikami ładunku.

O półprzewodniku domieszkowanym donorami mówimy, że jest to półprzewodnik typu *n*, a o półprzewodniku domieszkowanym akceptorami mówimy, że jest to półprzewodnik typu *p*. Możemy jednorodnie domieszkowane półprzewodniki typu n i typu p zobrazować symbolicznie tak, jak na rysunku 3-4.



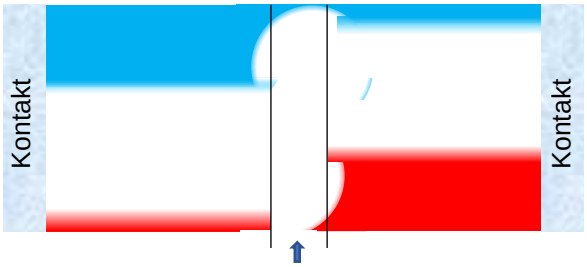
Półprzewodnik typu *n*: dużo elektronów, mało dziur

Półprzewodnik typu *p*: dużo dziur, mało elektronów

Rysunek 3-4. Symboliczne przedstawienie półprzewodników typu n i typu p

Już wkrótce jednak zobaczymy, że w półprzewodnikach domieszkowanych niejednorodnie lub nie będących w stanie równowagi termodynamicznej mogą istnieć sytuacje, w których w pewnym obszarze półprzewodnika typu *n* koncentracja dziur przeważa nad koncentracją elektronów, lub w półprzewodniku typu *p* nad koncentracją dziur przeważa koncentracja elektronów.

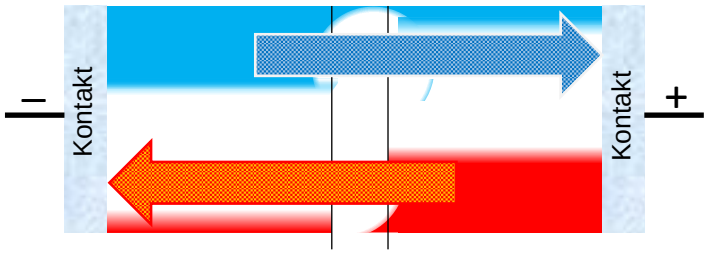
3.1.2 Jak działa dioda



Obszar zubożony (warstwa zaporowa)

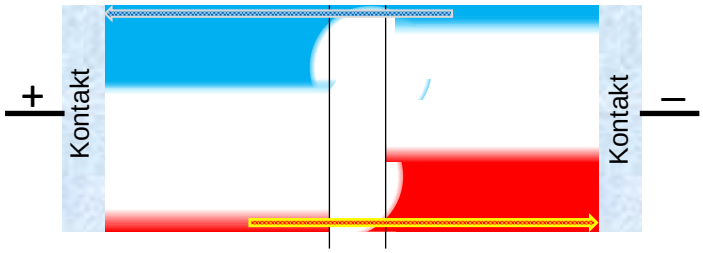
Rysunek 3-5. Złącze p-n

Zobaczymy teraz, co się stanie w półprzewodniku, w którym występują dwa obszary: jeden domieszkowany donorami, a drugi akceptorami. Taką strukturę symbolicznie pokazuje rysunek 3-5. Nazywamy ją złączem *p*-*n*. Pomiędzy obszarami typu *n* i typu *p* występuje obszar przejściowy. W tym obszarze koncentracja elektronów maleje w kierunku od obszaru typu *n* do obszaru typu *p*, a dziur – w przeciwnym kierunku. Obszar przejściowy nazywany jest obszarem zubożonym lub inaczej warstwą zaporową. W środkowej części tego obszaru bardzo mała jest zarówno koncentracja elektronów, jak i dziur, co wynika bezpośrednio z równania 3-1: w punkcie, w którym koncentracje elektronów i dziur zrównują się, obie muszą być równe koncentracji samoistnej n_i, a to jest bardzo mała liczba. Zatem obszar zubożony zachowuje się w pierwszym przybliżeniu tak, jak gdyby był obszarem dielektryka.



Rysunek 3-6. Złącze p-n spolaryzowane w kierunku zaporowym

Sytuacja jednak zmienia się, gdy do złącza *p-n* doprowadzone zostanie napięcie z zewnątrz. Jeżeli plus napięcia przyłożony będzie do obszaru *p*, a minus do obszaru *n*, strumień elektronów popłynie z obszaru *n* do obszaru *p* i przezeń do kontaktu, a dziur – w przeciwnym kierunku i do przeciwnego kontaktu (przypomnijmy, że dziura jest pojęciem umownym, a strumień dziur jest w rzeczywistości także strumieniem elektronów płynących w kierunku przeciwnym, niż dziury). Oba strumienie będą silne, bowiem obszar *n* jest obfitym źródłem elektronów, a obszar *p* – obfitym źródłem dziur. Ilustruje to symbolicznie rysunek 3-6. Zatem złącze *p-n* będzie całkiem dobrze przewodzić prąd. Mówimy, że zostało spolaryzowane w kierunku przewodzenia, a prąd nazywamy prądem przewodzenia złącza.



Rysunek 3-7. Złącze p-n spolaryzowane w kierunku zaporowym

Jeżeli kierunek zewnętrznego napięcia będzie przeciwny – plus do obszaru *n*, a minus do obszaru *p*, to także popłyną strumienie elektronów i dziur, ale będą one bardzo, bardzo słabiutkie, bo elektrony popłyną z obszaru *p*, gdzie jest ich znikomo mało, a dziury z obszaru *n*, gdzie także jest ich znikomo mało. Prąd popłynie, ale będzie bardzo słabiutki. Nazywamy go prądem wstecznym złącza. Stan taki ilustruje rysunek 3-7. Mówimy, że złącze *p-n* zostało spolaryzowane w kierunku zaporowym.

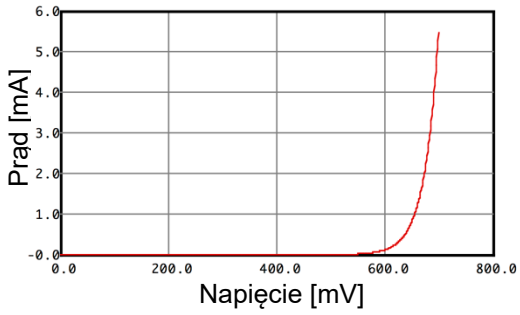
Jak widzimy, złącze p-n przepuszcza prąd elektryczny praktycznie tylko w jedną stronę. Element o takiej właściwości nazywamy diodą. W układach scalonych złącza *p-n* pełnią bardzo ważne role – między innymi wchodzą w skład struktur tranzystorów MOS i bipolarnych.

Zależność natężenia prądu płynącego przez złącze *p-n* od napięcia polaryzującego jest nieliniowa. Określa tę zależność funkcja wykładnicza:

Równanie 3-2

$$I = I_S\left(e^{\frac{qV}{kT}} - 1\right)$$

gdzie I jest prądem, V – napięciem, I_S jest stałą zwaną prądem nasycenia (jej wartość zależy m.in. od koncentracji domieszek w obszarach złącza i jest silnie rosnącą funkcją temperatury), q – ładunkiem elementarnym, k – stałą Boltzmanna, a T – temperaturą bezwzględną (w kelwinach). Często spotykane w mikroelektronice wyrażenie kT/q ma wymiar napięcia, a jego wartość w temperaturze pokojowej wynosi około 26 miliwoltów. Przyjmuje się, że w równaniu 3-2 napięcie polaryzujące w kierunku przewodzenia ma znak dodatni, a w kierunku zaporowym – ujemny. Rysunek 3-8 pokazuje przykładową zależność prądu od napięcia w diodzie krzemowej w temperaturze pokojowej przy polaryzacji w kierunku przewodzenia. Jak widać, dioda zaczyna dobrze przewodzić przy napięciu rzędu 0,6 – 0,7 V.



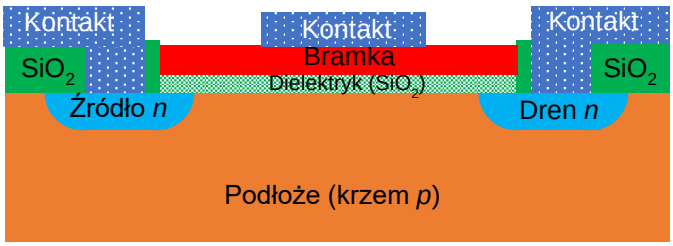
Rysunek 3-8. Przykładowa charakterystyka prądowo-napięciowa diody krzemowej w kierunku przewodzenia

Przy polaryzacji w kierunku zaporowym prąd płynący przez diodę dla napięć większych od $4\,kT/q$ jest praktycznie stały i równy I_S. Jeśli jednak napięcie w kierunku zaporowym jest dostatecznie duże, prąd płynący przez diodę zaczyna gwałtownie rosnąć ze wzrostem napięcia (nie pokazuje tego wzór 3-2). Jest to skutek zjawiska zwanego lawinowym powielaniem nośników ładunku – dziur i elektronów. Przy dostatecznie dużym napięciu w kierunku zaporowym w warstwie zubożonej natężenie pola elektrycznego osiąga dużą wartość. Elektrony i dziury zostają silnie przyspieszane w polu elektrycznym i uzyskują energie wystarczające do tego, by przy zderzeniu z atomami półprzewodnika uwalniać z tych atomów elektrony, czyli tworzyć nowe pary elektron-dziura. Te nośniki także są silnie przyspieszane i liczba elektronów i dziur w warstwie zubożonej gwałtownie rośnie, co powoduje silny wzrost natężenia prądu. Zjawisko to nazywamy lawinowym powielaniem nośników, a gwałtowny wzrost prądu – przebiciem lawinowym. Nie jest to zjawisko niszczące diodę, pod warunkiem, że prąd płynący w wyniku powielania lawinowego nie osiąga zbyt dużych wartości. Napięcie, przy którym występuje przebicie lawinowe, zależy od koncentracji domieszek po obu stronach złącza p-n. W złączach występujących w układach scalonych napięcie to wynosi typowo od kilku do kilkudziesięciu V.

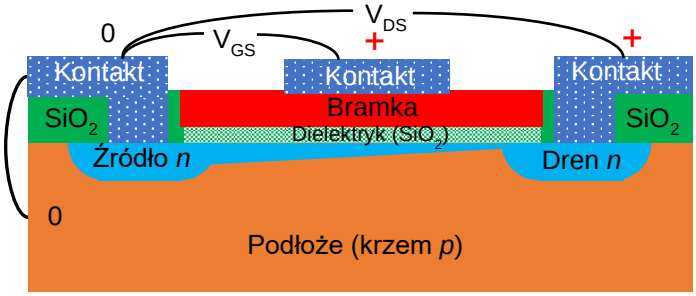
Złącze *p-n*, złożone z obszaru zachowującego się podobnie jak dielektryk umieszczonego pomiędzy dwoma obszarami przewodzącymi, jest zarazem kondensatorem o pewnej pojemności C_j zwanej pojemnością złączową. Jej wartość zależy od wartości przyłożonego do złącza napięcia polaryzującego. Wzór znajdziesz dalej (wyrażenie 3-10).

3.1.3 Jak działa tranzystor MOS

Tranzystor MOS jest podstawowym elementem współczesnych układów scalonych. Nieco uproszczoną strukturę tego tranzystora w przekroju pionowym (w głąb płytki półprzewodnikowej) pokazują rysunki poniżej.



Rysunek 3-9. Tranzystor MOS, do którego nie przyłożono żadnych napięć



Rysunek 3-10. Tranzystor MOS, w którym pod wpływem dodatniego napięcia bramki VGS powstał kanał przewodzący

Rysunek 3-9 pokazuje przekrój przez tranzystor MOS, do którego obszarów nie są przyłożone żadne napięcia. Podłożem tranzystora jest monokrystaliczna płytka krzemowa typu *p*. Do tej płytki wprowadzone zostały domieszki donorowe w taki sposób, że powstały dwa obszary typu *n* zwane źródłem i drenem. Na powierzchni płytki pomiędzy źródłem i drenem wytworzona została bardzo cienka warstwa dielektryka (dwutlenku krzemu SiO_2). Jest to dielektryk bramkowy. Na nim znajduje się bramka, jest to warstwa krzemu polikrystalicznego (krzem polikrystaliczny odpowiednio domieszkowany jest materiałem przewodzącym). Bramka jest izolowana od płytki podłożowej, źródła i drenu obszarami dielektrycznymi. Widoczne są też kontakty do źródła, drenu i bramki wykonane z metalu. Pomiędzy źródłem oraz drenem, a podłożem mamy złącza *p-n*. Jeżeli do drenu przyłożone zostanie dodatnie względem źródła napięcie, to żaden prąd ze źródła do drenu nie popłynie, bo między źródłem, a drenem jest obszar typu *p*, w którym swobodnych elektronów jest pomijalnie mało. Mówimy, że tranzystor jest wyłączony. Aby prąd mógł popłynąć, trzeba przyłożyć dodatnie napięcie V_{GS} do bramki (rysunek 3-10). Dodatni ładunek bramki odepchnie dziury, a przyciągnie elektrony, które utworzą pod bramką kanał przewodzący typu *n*. Dodatnie napięcie V_{DS} między drenem, a źródłem spowoduje przepływ

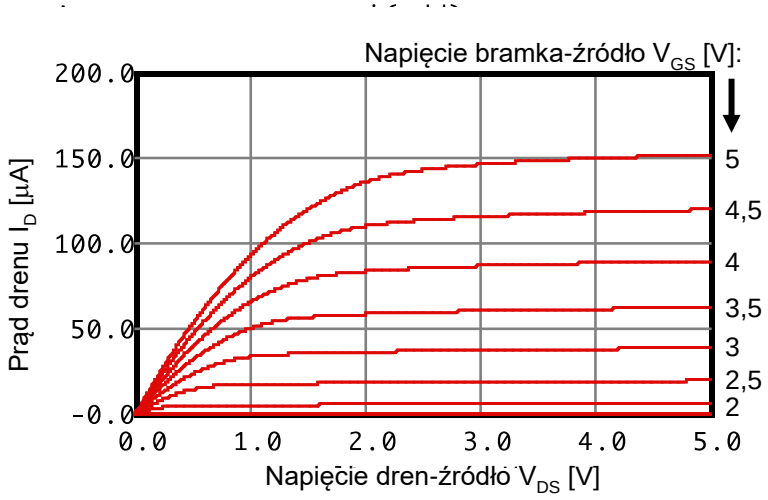
prądu: elektrony będą dopływać ze źródła i przez kanał płynąć do drenu. Mówimy, że tranzystor jest teraz włączony.

DEFINICJA

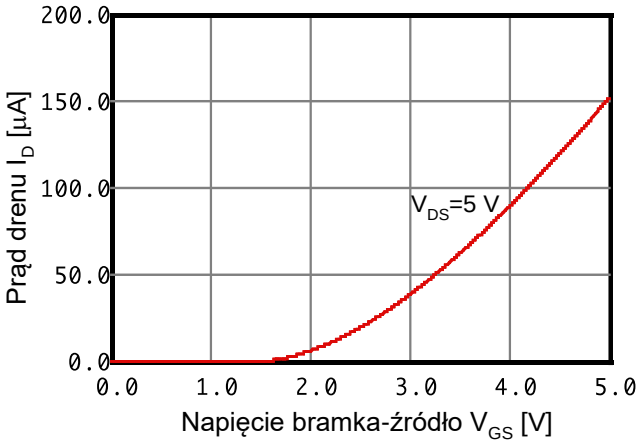
Napięcie bramki względem źródła, przy którym następuje włączenie tranzystora MOS, nazywamy napięciem progowym i zwyczajowo oznaczamy symbolem V_T.

Zarówno kanał, jak też źródło i dren są elektrycznie izolowane od podłoża, bowiem jest ono typu *p*, podczas gdy źródło, dren i kanał są obszarami o przewodnictwie typu *n*, a – jak już wiemy – między obszarami o przeciwnym typie przewodnictwa nie spolaryzowanymi lub spolaryzowanymi zaporowo istnieje obszar zubożony, przez który prąd praktycznie nie płynie. Rysunki 3-9 i 3-10 pokazują tranzystor MOS n-kanałowy (w skrócie: tranzystor nMOS). Napięcie progowe tranzystora n-kanałowego jest dodatnie, a – aby płynął prąd - dren polaryzuje się napięciem dodatnim względem źródła. Istnieją też i są równie ważne tranzystory p-kanałowe (w skrócie: tranzystory pMOS). W przekroju wyglądają one tak samo, z tym, że podłoże jest typu *n*, dren i źródło są typu *p*, a kanał powstaje pod wpływem ujemnego względem źródła napięcia bramki. W kanale tranzystora pMOS mamy do czynienia z przewodnictwem dziurowym, napięcie progowe takiego tranzystora jest ujemne, i ujemnym napięciem polaryzuje się dren względem źródła.

Zauważmy, że obszary źródła i drenu fizycznie niczym się nie różnią. O tym, który obszar jest drenem, a który źródłem, decyduje napięcie polaryzujące. W tranzystorze nMOS jako dren służy obszar spolaryzowany dodatnio względem źródła, w tranzystorze pMOS drenem jest obszar spolaryzowany ujemnie względem źródła.



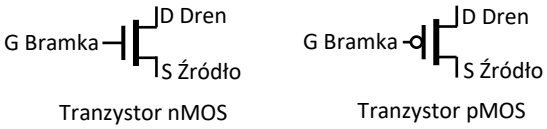
Rysunek 3-11. Przykładowa rodzina charakterystyk tranzystora MOS $I_D=f(V_{DS},V_{GS})$



Rysunek 3-12. Przykładowa charakterystyka $I_D=f(V_{GS})$ dla ustalonej wartości V_{DS}

Zależności prądu drenu od napięć bramki i drenu względem źródła są nieliniowe. Rysunek 3-11 pokazuje przykładową zależność prądu drenu I_D od napięcia dren-źródło V_{DS} dla różnych wartości napięcia bramka-źródło V_{GS}. Rysunek 3-12 pokazuje przykładową zależność prądu drenu I_D od napięcia bramka-źródło V_{GS} dla ustalonej, różnej od zera wartości napięcia dren-źródło V_{DS}. Z tej charakterystyki można oszacować wartość napięcia

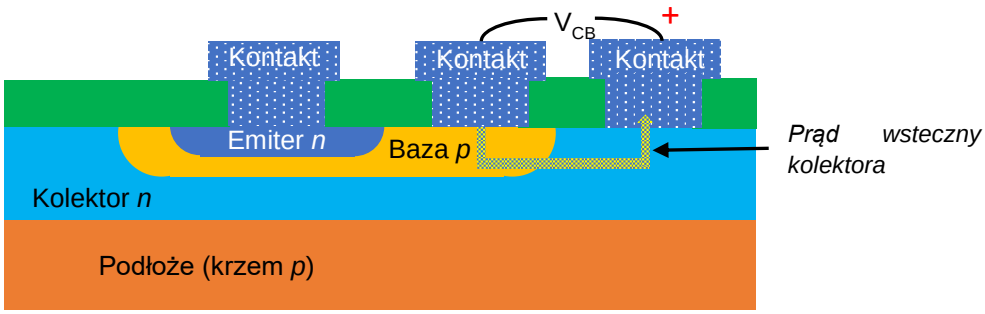
progowego tranzystora: około 1,5 V (tranzystory używane w dzisiejszych układach scalonych mają znacznie niższe napięcia progowe). Opis matematyczny tych zależności poznasz w punkcie 3.1.5.



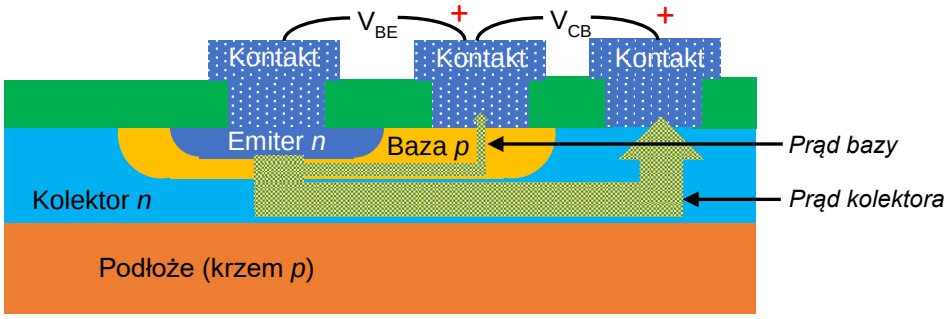
Rysunek 3-13. Symbole tranzystorów MOS

Rysunek 3-12 pokazuje symbole tranzystorów MOS używane w schematach elektrycznych układów elektronicznych. Litery S, D i G pochodzą od angielskich terminów: S – source, D – drain, G – gate.

3.1.4 Jak działa tranzystor bipolarny



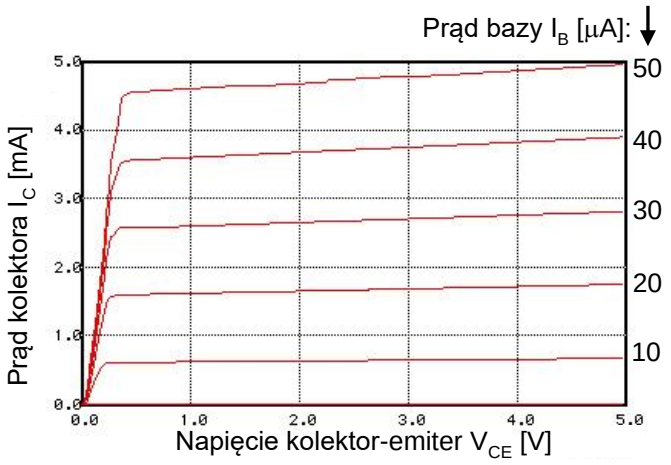
Rysunek 3-14. Tranzystor bipolarny, zaporowo spolaryzowane złącze kolektor-baza



Rysunek 3-15. Tranzystor bipolarny, złącze kolektor-baza spolaryzowane zaporowo, złącze emiter-baza w kierunku przewodzenia

Tranzystory bipolarne były wcześniej niż tranzystory MOS używane w układach scalonych, dziś w dużym stopniu straciły swoje znaczenie. Mają jednak nadal w mikroelektronice swoje obszary zastosowań, więc wypada je tu omówić.

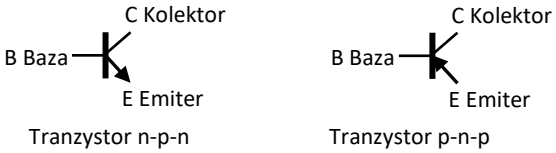
Rysunek 3-14 ilustruje w pewnym uproszczeniu strukturę tranzystora bipolarnego. Tranzystor składa się z trzech obszarów: kolektora typu *n*, bazy typu *p* i emitera typu *n*. Pomiędzy kolektorem i bazą oraz pomiędzy bazą i emiterem są złącza *p-n*. Jeśli złącze kolektor-baza jest spolaryzowane zaporowo (jak na rysunku 3-14), to płynie prąd wsteczny tego złącza. Jak już wiemy, jest on bardzo mały. Jeśli teraz złącze emiter-baza zostanie spolaryzowane w kierunku przewodzenia, jak na rysunku 3-15, z emitera do bazy popłynie silny prąd. W bazie rozdzieli się on na dwa. Przeważająca część elektronów płynących z emitera zostanie przechwycona przez warstwę zaporową złącza kolektor-baza i popłynie do kontaktu kolektora, bardzo niewielka część tych elektronów (typowo około 1/100 lub mniej) popłynie do kontaktu bazy. W ten sposób napięcie V_{BE} między bazą, a emiterem steruje przepływem prądu kolektora I_C. Gdy napięcie emiter-baza V_{BE} polaryzuje złącze emitera w kierunku przewodzenia, a napięcie kolektor-baza V_{CB} polaryzuje złącze kolektora w kierunku zaporowym, mówimy że tranzystor bipolarny pracuje w warunkach polaryzacji normalnej. Prąd wstrzykiwany przez emiter do bazy jest prądem spolaryzowanego w kierunku przewodzenia złącza *p-n*, toteż ten prąd, a więc i prąd kolektora, jest uzależniony od napięcia V_{BE} tak samo, jak prąd diody (rysunek 3-8). W warunkach polaryzacji normalnej prąd bazy I_B jest w dobrym przybliżeniu proporcjonalny do prądu kolektora.



Rysunek 3-16. Przykładowa rodzina charakterystyk tranzystora bipolarnego $I_C=f(V_{CE}, I_B)$

Rysunek 3-16 pokazuje zależność prądu kolektora I_C od napięcia kolektor-emiter V_{CE}, gdy jako parametr traktujemy prąd bazy I_B. W taki sposób zwykle przedstawiane są charaterystyki tranzystora bipolarnego.

Rysunki 3-14 i 3-15 przedstawiają tranzystor *n-p-n* (emiter i kolektor typu *n*, baza typu *p*). Istnieją także tranzystory przeciwnego typu: *p-n-p*. W układach scalonych ich struktury wyglądają nieco inaczej, odwrotne są także znaki napięć polaryzujących, ale zasada działania jest taka sama.



Rysunek 3-17. Symbole tranzystorów bipolarnych

Opis matematyczny charakterystyk tranzystora bipolarnego poznasz w punkcie 3.1.6. Rysunek 3-17 pokazuje symbole tranzystorów bipolarnych *n-p-n* i *p-n-p* używane w schematach elektrycznych układów elektronicznych. Litery E, B i C pochodzą od angielskich terminów: E – emitter, B – base, C – collector.

3.1.5 Model matematyczny tranzystora MOS

Modelem matematycznym elementu elektronicznego nazywamy zespół równań opisujących zależności wiążące prądy i napięcia w elemencie, czyli równania opisujące na przykład takie charakterystyki, jak pokazane na rysunkach 3-8, 3-11, 3-12, 3-16. Omówimy teraz najprostszy model matematyczny tranzystora MOS, jaki będzie używany w dalszych rozważaniach. W układach CMOS wykorzystywane są tranzystory pMOS i tranzystory nMOS. Oba rodzaje tranzystorów są typu wzbogacanego, czyli do bramki trzeba przyłożyć napięcie, aby utworzył się kanał między źródłem i drenem i tranzystor zaczął przewodzić (takie tranzystory były omówione w punkcie 3.1.3). W przypadku tranzystorów nMOS jest to napięcie dodatnie względem źródła, a w przypadku tranzystorów pMOS - ujemne. Gdy analizujemy działanie tranzystora MOS w bramkach logicznych, można w uproszczeniu powiedzieć, że tranzystor nMOS jest włączany napięciem dodatnim względem źródła, a pMOS – ujemnym, przy czym w każdym przypadku napięcie to powinno być większe co do wartości bezwzględnej od napięcia progowego tranzystora V_T.

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

- w zakresie zwanym podprogowym:

Równanie 3-3

$$I_D = 0 \qquad \text{dla } V_{GS} < V_T$$

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

Równanie 3-4

$$I_D = \mu C_{ox} \frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right] \qquad \text{dla } V_{GS} \geq V_T \text{ oraz } V_{DS} \leq V_{DSsat}$$

i w zakresie zwanym zakresem nasycenia:

Równanie 3-5

$$I_D = \mu C_{ox} \frac{W}{L} \frac{(V_{GS}-V_T)^2}{2} \qquad \text{dla } V_{GS} \geq V_T \text{ oraz } V_{DS} \geq V_{DSsat}$$

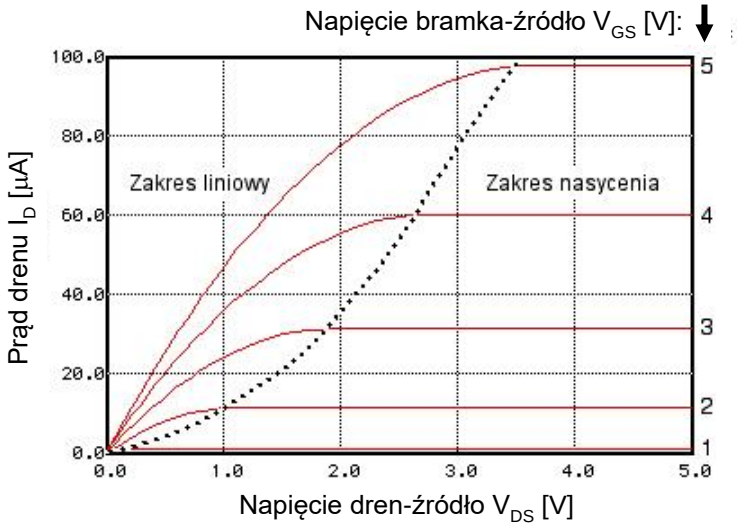
W tych wzorach $V_{DSsat} = V_{GS} - V_T$, W jest szerokością kanału, L – jego długością, μ jest ruchliwością nośników ładunku (elektronów w tranzystorze nMOS, dziur w tranzystorze pMOS) w kanale, C_{ox} – pojemnością dielektryku bramkowego na jednostkę powierzchni.

UWAGA

Wzory 3-3 do 3-5 opisują charakterystyki tranzystora nMOS. ***W przypadku tranzystora pMOS można stosować te same wzory podstawiając do nich wartość bezwzględną napięcia progowego (które - jak wiemy - jest dla tranzystorów pMOS ujemne).***

Wyjaśnimy teraz sens niektórych wielkości występujących w równaniach 3-4 i 3-5. Ruchliwością nośników nazywamy współczynnik proporcjonalności między natężeniem pola elektrycznego w półprzewodniku, a prędkością, z jaką poruszają się nośniki ładunku w tym polu. Pojemność dielektryku bramkowego na jednostkę powierzchni C_{ox} określona jest przez grubość warstwy tego dielektryku t_{ox} i jego przenikalność dielektryczną ε_{ox}: $C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$. Wymiary kanału: W i L dokładnie zdefiniujemy dalej, omawiając realne, nieuproszczone struktury tranzystorów w układach scalonych.

Przykładowa rodzina charakterystyk opisana zależnościami (3-3) – (3-5) wygląda tak:



Rysunek 3-18. Kształt charakterystyk tranzystora opisanych równaniami 3-4 i 3-4. Linia przerywana oddziela zakres liniowy od zakresu nasycenia

Wzory (3-3) – (3-5) stanowią podstawę najprostszego modelu matematycznego tranzystora MOS używanego w symulatorach układów elektronicznych, zwanego modelem poziomu 1 („Level 1"). (To określenie pochodzi z najstarszych wersji symulatora układów elektronicznych SPICE, w których dostępne były trzy modele tranzystora MOS o różnym stopniu komplikacji i różnej dokładności rozróżniane wartością parametru o nazwie „level" i wartościach równych 1, 2 i 3).

Wzory (3-3) – (3-5) opisują charakterystyki tranzystora w bardzo uproszczony sposób. Nie są w nich uwzględnione liczne zjawiska fizyczne wpływające bardzo poważnie na kształt charakterystyk takich tranzystorów MOS, jakie występują we współczesnych układach CMOS. Wzory te będziemy jednak stosować do

prostych obliczeń ilustrujących działanie bramek logicznych i układów analogowych, ponieważ umożliwiają one łatwe wyprowadzenie podstawowych zależności ilustrujących jakościowo właściwości tych bramek i układów. Niemniej trzeba pamiętać, że do symulacji układów elektronicznych w praktycznych pracach projektowych wzory (3-3) – (3-5) i oparty na nich model „level 1" w żadnym przypadku nie wystarczają, a otrzymane przy ich użyciu wyniki będą z reguły odległe od rzeczywistości. Modelami stosowanymi dziś najczęściej w projektowaniu układów CMOS są modele o nazwach BSIM3, BSIM4, EKV, PSP. Występują one w różnych symulatorach z różnymi wartościami parametru „level". Są to modele skomplikowane od strony matematycznej, ale dobrze oddające charakterystyki tranzystorów o bardzo małych długościach kanału. Producenci układów scalonych podają wartości parametrów tych modeli dla typowych struktur tranzystorów wytwarzanych w dostępnych u nich procesach technologicznych.

Dokładność wzorów (3-3) – (3-5) można nieco poprawić uwzględniając dwa ważne zjawiska występujące w tranzystorach MOS: zależność napięcia progowego V_T od napięcia polaryzacji podłoża względem źródła V_{BS} („efekt polaryzacji podłoża") i zjawisko zależności elektrycznej długości kanału od napięcia dren-źródło V_{DS} („efekt skracania kanału"). Na rysunku 3-10 obszar źródła tranzystora jest elektrycznie połączony z podłożem, jednak w realnych układach często tak nie jest. Źródło może być spolaryzowane zaporowo względem podłoża napięciem V_{BS}. Taka polaryzacja wpływa na wartość napięcia progowego (podnosi jego wartość bezwzględną, czyli napięcie progowe tranzystora nMOS staje się bardziej dodatnie, a pMOS – bardziej ujemne). Efekt ten w przybliżeniu opisuje równanie 3-6:

Równanie 3-6

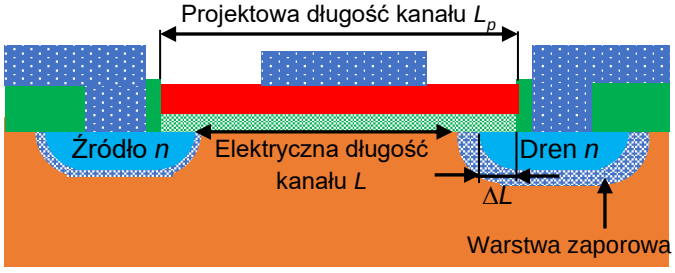
$$V_T = V_{T0} + \gamma\left(\sqrt{|2\varphi_F| - V_{BS}} - \sqrt{|2\varphi_F|}\right)$$

w którym V_{T0} jest napięciem progowym przy braku polaryzacji podłoża, a parametry γ oraz φ_F zależą od szczegółów technologii tranzystora. Ich wartości dla konkretnej technologii podaje producent układów.

Efekt skracania kanału polega na tym, że rzeczywista „elektryczna" długość kanału jest mniejsza od odległości między źródłem i drenem L_p, jaką zaprojektował projektant, z dwóch powodów: po pierwsze, obszary domieszkowane źródła i drenu zachodzą pod obszar bramki na pewną odległość ΔL, a po drugie warstwa zaporowa złącza drenu wnika na pewną odległość w obszar kanału efektywnie skracając go. Ten drugi efekt powoduje, że długość kanału maleje ze wzrostem napięcia V_{DS}, zaś prąd drenu wzrasta. W rezultacie w zakresie nasycenia prąd drenu nie jest stały (jak wynikałoby ze wzoru 3-5 i rysunku 3-18), lecz wzrasta (ten wzrost widać na rysunku 3-11, który pokazuje charakterystyki rzeczywistego tranzystora). Oba te efekty łącznie można uwzględnić opisując rzeczywistą „elektryczną" długość kanału L zależnością:

Równanie 3-7

$$\frac{1}{L} = \frac{1}{L_p - 2\Delta L}(1 + \lambda V_{DS})$$



Rysunek 3-19. Projektowa i elektryczna długość kanału tranzystora MOS

W tym wzorze λ jest parametrem empirycznym, wyznaczanym tak, by charakterystyki tranzystora w zakresie nasycenia miały nachylenie zgodne z rzeczywiście obserwowanym. Wartość tego parametru podaje producent układów. Wielkości występujące w zależności 3-7 ilustruje rysunek 3-19.

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

UWAGA

W dalszych rozważaniach będziemy najczęściej przy wyprowadzaniu wzorów pomijali zarówno wpływ napięcia polaryzacji podłoża na napięcie progowe, jak i wpływ napięcia drenu na długość kanału tranzystora. Jest to równoznaczne z założeniem, że parametry γ, λ oraz ΔL mają wartości równe zeru. Dzięki temu uzyskamy proste i łatwe do interpretacji zależności, trzeba jednak pamiętać, że w większości przypadków będą one dawać ilościowo wyniki dalekie od rzeczywistości.

Wzór 3-3 przewiduje, że dla napięcia bramki mniejszego od progowego prąd drenu jest dokładnie równy zeru. Tak jednak w rzeczywistości nie jest. Gdy napięcie bramki staje się mniejsze od progowego, prąd drenu nie spada dokładnie do zera, lecz wykazuje zależność od napięcia bramki o charakterze wykładniczym. Prąd ten, zwany prądem podprogowym, dla napięć wyraźnie mniejszych od progowego można przybliżyć wyrażeniem

Równanie 3-8

$$I_D = I_t \frac{W}{L} e^{\frac{q(V_{GS}-V_T)}{nkT}} \left(1 - e^{\frac{qV_{DS}}{kT}}\right)$$

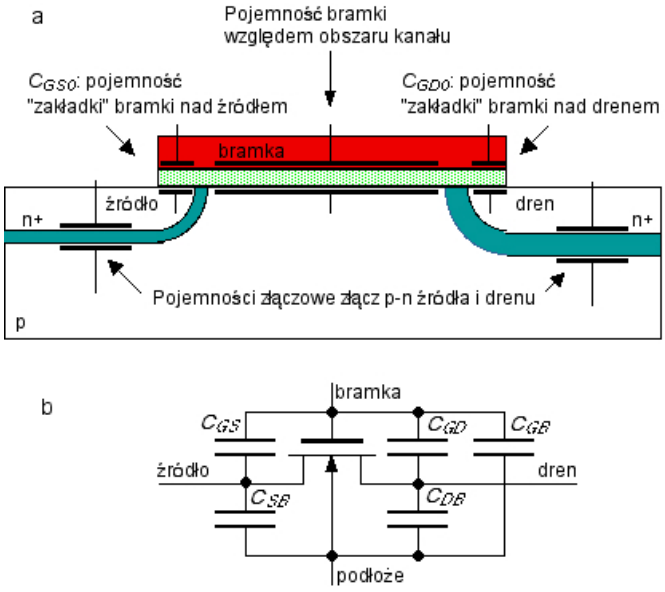
w którym I_t oraz n są parametrami zależnymi od konstrukcji tranzystora. Współczynnik n jest liczbą zawartą między 1, a 2. Prąd podprogowy nie ma bezpośredniego wpływu na działanie większości typów bramek logicznych, jednak nie jest całkiem bez znaczenia, bowiem jego obecność zwiększa prąd, jaki pobierają ze źródła zasilania układy logiczne. Praca tranzystorów w zakresie podprogowym, czyli gdy prąd drenu jest wykładniczą funkcją napięcia bramki, bywa stosowana w niektórych układach analogowych.

Charakterystyki prądowo-napięciowe złącz *p-n* źródła i drenu są z dostateczną dokładnością opisane zależnością

Równanie 3-9

$$I = I_S \left(e^{\frac{qV}{n_j kT}} - 1\right)$$

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



Rysunek 3-20. Pojemności w strukturze tranzystora MOS: (a) w strukturze fizycznej, (b) reprezentacja w schemacie

Do projektowania układów cyfrowych i niektórych rodzajów układów analogowych potrzebna jest, oprócz znajomości charakterystyk prądowo-napięciowych, także znajomość pojemności występujących w tranzystorze MOS. Są to pojemności typu metal-dielektryk-półprzewodnik związane z bramką tranzystora, oraz pojemności złączowe związane ze złączami *p-n* obszarów źródła oraz drenu. Wszystkie te pojemności ilustruje rysunek 3-20.

Jak widać, możemy wyróżnić trzy rodzaje pojemności:
- pojemności "zakładek" bramki nad źródłem i drenem C_{GS0} i C_{GD0},
- pojemności złącz *p-n* źródła i drenu,
- pojemność bramki względem obszaru kanału.

Pojemności C_{GS0} i C_{GD0} można uważać za niezależne od napięć polaryzujących tranzystor. Są one proporcjonalne do szerokości kanału tranzystora i dlatego w modelach tranzystorów są wyrażane jako pojemność na jednostkę długości (a nie powierzchni). Pojemności złącz *p-n* źródła i drenu C_j są opisywane prostą zależnością

Równanie 3-10

$$C_j = \frac{C_{j0}}{\left[1-\left(\frac{V}{V_D}\right)\right]^m}$$

w której V jest napięciem polaryzującym złącze (ujemnym w przypadku polaryzacji zaporowej, dodatnim w przeciwnym przypadku), V_D jest napięciem zwanym napięciem dyfuzyjnym (jego wartość zależy od koncentracji domieszek w obszarach złącza), C_{j0} jest pojemnością złącza niespolaryzowanego, zaś m jest wykładnikiem o wartości zależnej od rozkładu domieszek w złączu. W bardziej dokładnych obliczeniach pojemności złączowe są rozdzielane na pojemności dna złącza (które jest złączem płaskim) i pojemności obszarów bocznych (które nie są płaskie). Obie składowe całkowitej pojemności są opisywane tym samym wzorem 3-10, ale wartości parametrów C_{j0}, V_D i m są różne ze względu na różnice w kształcie obszarów płaskich i bocznych oraz różnice w rozkładach domieszek. Parametr C_{j0} jest zwykle podawany na jednostkę powierzchni złącza w przypadku obszarów dna i na jednostkę obwodu (czyli długości) w przypadku obszarów bocznych.

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

kanału musi być w jakiejś proporcji podzielona na dwie: bramka-źródło i bramka-dren. W prostym modelu („level 1“) przyjęto dość arbitralnie następujące założenia:
- całkowita pojemność bramki jest równa $C_g = WLC_{ox}$
- w zakresie podprogowym pojemność ta jest równa pojemności bramka-podłoże C_{GB}; pojemności bramka-dren i bramka-źródło są równe zeru,
- w zakresie liniowym pojemność ta jest dzielona po połowie między pojemność bramka-dren C_{GD} i pojemność bramka-źródło C_{GS}, zaś pojemność bramka-podłoże jest równa zeru,
- w zakresie nasycenia pojemność ta jest przypisywana pojemności bramka-źródło C_{GS}, przy czym przyjmuje się, że jest zmniejszona do wartości równej $2WLC_{ox}/3$; pojemności bramka-dren i bramka-podłoże są równe zeru.

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

W rozważaniach układowych wszystkie pojemności są reprezentowane przez pojemności bramka-źródło C_{GS}, bramka-dren C_{GD}, bramka-podłoże C_{GB}, źródło-podłoże C_{SB} i dren-podłoże C_{DB} (patrz rysunek 3-20b). Pojemności te mają następujące składowe:
- pojemność C_{GS} jest sumą pojemności „zakładki“ C_{GS0} i uzależnionej od napięcia części pojemności bramki C_g,
- pojemność C_{GD} jest sumą pojemności „zakładki“ C_{GD0} i uzależnionej od napięcia części pojemności bramki C_g,
- pojemność C_{GB} jest uzależnioną od napięcia częścią pojemności bramki C_g, ma wartość równą zeru gdy tranzystor przewodzi (istnieje kanał między źródłem i drenem),
- pojemność C_{SB} jest równa pojemności złączowej źródła,
- pojemność C_{DB} jest równa pojemności złączowej drenu.

3.1.6 Model matematyczny tranzystora bipolarnego

We współczesnej mikroelektronice królują układy CMOS. W układzie scalonym CMOS tranzystory bipolarne występują zwykle tylko jako elementy pasożytnicze (dokładniej o elementach pasożytniczych będzie mowa dalej), ale niektóre z nich mogą być wykorzystane jako aktywne elementy w układzie. Istnieją także technologie BiCMOS, w których na równi można się posługiwać tranzystorami MOS i bipolarnymi (nie mówimy o nich tutaj). Jak zobaczymy, w układach analogowych w wielu przypadkach tranzystory bipolarne są korzystniejsze od tranzystorów MOS. Warto więc przypomnieć sobie także ich charakterystyki i parametry. Omówimy je w uproszczeniu, tylko w takim zakresie, jaki będzie potrzebny w dalszych rozważaniach. Najprostszym opisem charakterystyk prądowo-napięciowych tranzystora bipolarnego jest model Ebersa-Molla. W tym modelu ogólna zależność prądu kolektora I_C od napięć emiter-baza V_{BE} i kolektor-baza V_{CB} dana jest wzorem

Równanie 3-11

$$I_C = I_{ES0}\left(e^{\frac{qV_{BE}}{kT}} - 1\right) - I_{CS0}\left(e^{\frac{qV_{CB}}{kT}} - 1\right)$$

gdzie I_{ES0} i I_{CS0} są stałymi zależnymi od koncentracji domieszek i powierzchni złącz p-n tranzystora.

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

W układach analogowych tranzystory bipolarne pracują prawie zawsze w zakresie napięć nazwanym wcześniej polaryzacją normalną: złącze kolektor-baza jest polaryzowane zaporowo, a złącze emiter-baza w kierunku przewodzenia. Dla takich warunków polaryzacji model Ebersa-Molla można poważnie uprościć. Dla $V_{BE} \gg \frac{kT}{q}$ (ten warunek jest zawsze spełniony w typowych warunkach polaryzacji krzemowego tranzystora bipolarnego) oraz dla $V_{CB} \leq 0$ pierwszy składnik we wzorze 3-11 ma wartość o wiele rzędów wielkości większą od drugiego, i równocześnie $e^{\frac{qV_{BE}}{kT}} \gg 1$, co pozwala sprowadzić wzór 3-11 do prostej i bardzo użytecznej postaci

Równanie 3-12

$$I_C = I_{ES0} e^{\frac{qV_{BE}}{kT}}$$

Ta zależność jest podstawą wielu rozwiązań układowych w analogowych układach bipolarnych. Wzór ten opisuje rzeczywistą charakterystykę tranzystora z dużą dokładnością w szerokim zakresie prądów kolektora (kilka dekad). Odstępstwa obserwowane są dopiero w zakresie dużych gęstości prądu kolektora. W tym zakresie prąd kolektora rośnie z napięciem V_{BE} wolniej niż przewiduje wzór 3-12.

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

Równanie 3-13

$$I_{ES0} = J_{ES0} A_E$$

zaś gęstość tego prądu, oznaczona J_{ES0}, zależy od elektrycznej grubości bazy tranzystora - jest do niej w przybliżeniu odwrotnie proporcjonalna. W tym ukryta jest zależność prądu kolektora od napięcia kolektor-baza V_{CB}, które nie występuje jawnie w zależności 3-12. Gdy napięcie V_{CB} (polaryzujące złącze kolektorowe w kierunku zaporowym) wzrasta, elektryczna grubość bazy maleje (bo poszerza się warstwa zaporowa złącza kolektor-baza i głębiej wnika w bazę) i współczynnik J_{ES0} wzrasta. Współczynnik J_{ES0} jest też silnie zależny od temperatury, o czym będzie mowa dalej.

Wyznaczając z zależności 3-12 napięcie V_{BE} otrzymujemy inną często wykorzystywaną zależność

Równanie 3-14

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

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

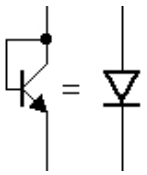
Równanie 3-15

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

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

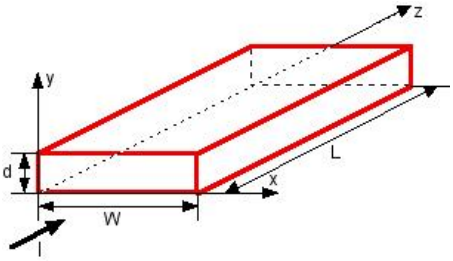
3.1.7 Nie tylko tranzystory: inne elementy układów scalonych

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



Rysunek 3-21: Tranzystor bipolarny w połączeniu diodowym

Jako diody wykorzystywane są zwykle struktury tranzystorów bipolarnych w połączeniu zwanym diodowym - kolektor zwarty z bazą. Prąd takiej diody jest sumą prądu kolektora i prądu bazy tranzystora, z tym że prąd bazy jest h_{FE} - razy mniejszy, a ponieważ h_{FE} ma zwykle wartość rzędu 50 - 200, prąd bazy jest w pierwszym przybliżeniu do pominięcia. Zatem charakterystyka prądowo-napięciowa takiej diody wynika bezpośrednio z charakterystyki $I_C = f(V_{BE})$ tranzystora, jest więc opisana wzorem 3-12.



Rysunek 3-22. Ilustracja do definicji rezystancji warstwowej

Rezystory w układach CMOS są wykonywane jako ścieżki z krzemu polikrystalicznego położone na warstwie dielektrycznej. Możliwy do uzyskania zakres rezystancji jest ograniczony. Aby dokładniej omówić rezystory, poznamy najpierw pojęcie rezystancji warstwowej. Rezystancja warstwowa (zwana także rezystancją powierzchniową lub gwarowo rezystancją "na kwadrat") jest pojęciem wygodnym do charakteryzowania rezystancji obszarów, które są niejednorodne w kierunku prostopadłym do kierunku przepływu prądu, na przykład z powodu nierównomiernego rozkładu domieszek. Taki obszar pokazany jest na rysunku poniżej. Może to być na przykład ścieżka domieszkowanego polikrzemu, w którym koncentracja domieszki maleje w kierunku od powierzchni w głąb. Rozważmy przepływ prądu przez prostopadłościan pokazany na rysunku 3-22. Jego rezystywność zmienia się wzdłuż osi y, zaś prąd płynie w kierunku osi z. Konduktancję tego prostopadłościanu dla prądu I można obliczyć całkując konduktywność σ w granicach od 0 do d. Konduktancja warstwy o nieskończenie małej grubości dy jest równa

Równanie 3-16

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

zatem rezystancja prostopadłościanu dla prądu I wynosi

Równanie 3-17

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

co można zapisać w postaci

Równanie 3-18

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

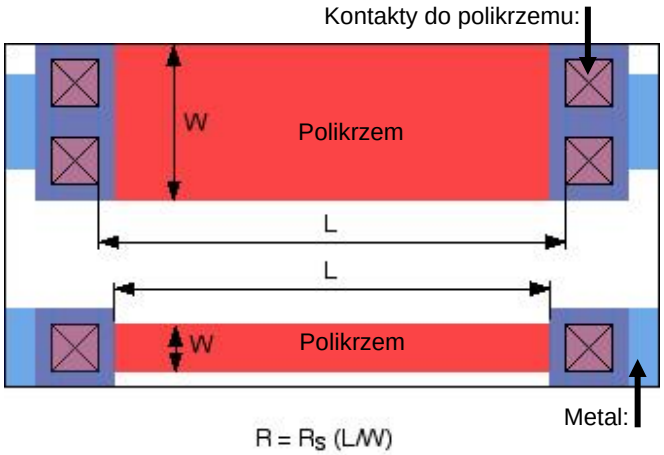
gdzie R_S jest rezystancją warstwową:

Równanie 3-19

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

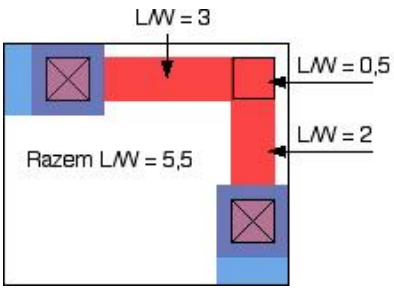
Jest ona potocznie nazywana rezystancją „na kwadrat", ponieważ jest to rezystancja obszaru o długości L równej szerokości W, czyli - patrząc z góry - kwadratu. Mianem rezystancji warstwowej jest om, ale dla zaznaczenia charakteru tej wielkości używa się często oznaczenia $\Omega/\square$ („om na kwadrat").

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

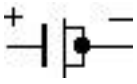


Rysunek 3-23. Rezystory w postaci ścieżek polikrzemowych, widok z góry

Rysunek 3-23 pokazuje rezystory w postaci prostych ścieżek polikrzemowych. Jeśli ścieżka rezystora zawiera zagięcia pod kątem prostym, to każde takie zagięcie traktuje się jako fragment ścieżki o stosunku L/W równym 0,5 (rysunek 3-24). W przypadku rezystorów o małych rezystancjach, istotna jest rezystancja kontaktów, która może wynosić nawet kilkadziesiąt omów. Wartość rezystancji kontaktów podaje producent układów.



Rysunek 3-24. Rezystor z zagięciem



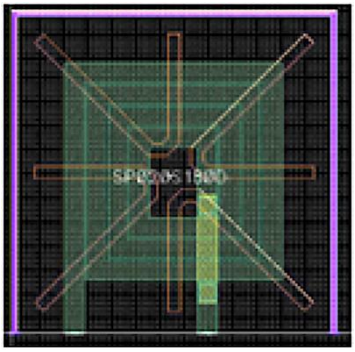
Rysunek 3-25. Tranzystor nMOS użyty jako kondensator

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

złączowe złącz *p-n*. Maksymalne pojemności takich kondensatorów są ograniczone w najlepszym razie do pojedynczych pikofaradów. Pojemność kondensatora jest proporcjonalna do powierzchni A jego okładek: $C = A\frac{\varepsilon_{ox}}{t_{ox}}$. W układzie scalonym duża powierzchnia to duży koszt.

Najprostszym kondensatorem jest kondensator metal-dielektryk-metal. W starszych technologiach CMOS standardem były dwie warstwy metalu rozdzielone dielektrykiem, służące do wykonywania połączeń w układzie. W bardziej zaawansowanych technologiach takich warstw może być nawet kilkanaście. Można tworzyć z nich stosy metal-dielektryk-metal-dielektryk… itd., co po odpowiednim połączeniu pozwala utworzyć kondensator o stosunkowo dużej pojemności. Jednak nawet wtedy pojemność jest ograniczona w najlepszym razie do kilkunastu pikofaradów.

Indukcyjności do niedawna były uważane za elementy niemożliwe do wykonania w układach scalonych. Ten stan rzeczy zaczął ulegać zmianom, gdy częstotliwości pracy układów scalonych sięgnęły kilku, a nawet kilkudziesięciu GHz. Przy tych częstotliwościach potrzebne są indukcyjności bardzo małe, o wartościach rzędu kilku nanohenrów. Można je wykonać jako płaskie spirale w warstwach metalu. Dobroć takich indukcyjności jest niewielka. Wynika to z sąsiedztwa przewodzącego podłoża krzemowego. Cewka indukuje w nim prądy wirowe, które są przyczyną strat energii pola magnetycznego cewki. Można temu w pewnym stopniu zapobiegać wprowadzając w podłożu pod cewką obszary utrudniające przepływ prądów wirowych. Jednym ze sposobów jest wprowadzenie do podłoża, do wąskich prostokątnych obszarów skierowanych prostopadle do kierunku prądów wirowych, domieszki przeciwnego typu niż podłoże. Tworzą się w ten sposób złącza *p-n* skutecznie przecinające drogę prądów wirowych (rysunek 3-26).



Rysunek 3-26. Tak wygląda cewka w układzie scalonym

3.1.8 Elementy i sprzężenia pasożytnicze

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

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

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

wielu elementów i jest obszarem przewodzącym o dość znacznej rezystywności. Sprzężenia pasożytnicze przez podłoże będą omówione przy omawianiu analogowych układów CMOS.

3.2 Funkcje logiczne i bramki logiczne

3.2.1 Z czego zbudowany jest system cyfrowy

Każdy system cyfrowy, od najprostszych do najbardziej złożonych, zbudowany jest z dwóch rodzajów układów logicznych: układów zwanych kombinacyjnymi i układów zwanych sekwencyjnymi. Układy logiczne realizują funkcje logiczne algebry Boole'a, w której zmienne mogą przybierać tylko dwie wartości: 0 oraz 1. Układy logiczne mają wejścia i wyjścia, na których pojawiają się zera i jedynki reprezentowane przez napięcia (jak – dowiesz się dalej).

DEFINICJA

Układem kombinacyjnym nazywamy układ logiczny, w którym stany logiczne na wyjściach zależą tylko od aktualnych stanów na wejściach, natomiast nie zależą od stanów poprzednich.

Przykładem układu kombinacyjnego może służyć układ sumatora mający dwa wejścia i wyjście. Na wejścia podawane są dwie liczby binarne, na wyjściu pojawia się ich suma. Zależy ona tylko od tego, jakie są aktualne wartości liczb binarnych na wejściach. Innymi słowy, układy kombinacyjne nie zawierają pamięci stanów poprzednich.

DEFINICJA

Układem sekwencyjnym nazywamy układ logiczny, w którym stany logiczne na wyjściach zależą nie tylko od aktualnych stanów na wejściach, ale i od stanów poprzednich, które są zapamiętane wewnątrz układu i są nazywane stanami wewnętrznymi.

Przykładem układu sekwencyjnego może służyć układ sumujący długi ciąg kolejno podawanych na wejście liczb. Stan jego wyjścia zależy od liczby aktualnie podanej na wejście oraz od sumy wszystkich poprzednio podawanych liczb, która musi być w układzie zapamiętana. Innymi słowy, układy sekwencyjne muszą zawierać pamięć.

Zarówno układy kombinacyjne, jak i sekwencyjne zbudowane są z bramek logicznych. Bramki logiczne są to proste na ogół układy elektroniczne realizujące podstawowe operacje algebry Boole'a (zwane też funkcjami logicznymi) na pojedynczych bitach. Oznaczmy dane wejściowe przez A i B, a wynik przez Q. Zależności między danymi wejściowymi, a wynikiem pokażemy w postaci tablic zwanych tablicami prawdy. Możemy wówczas zapisać te podstawowe operacje następująco:

- funkcja NOT (negacja): $Q = .NOT.A$ (inny zapis: $Q = \bar{A}$)
Tablica prawdy dla negacji jest bardzo prosta:

Dana A	Wynik Q
0	1
1	0

- funkcja OR (suma logiczna): $Q = A.OR.B$ (inny zapis: $Q = A + B$)
Tablica prawdy dla tej funkcji:

Dana A	Dana B	Wynik Q
0	0	0
0	1	1
1	0	1
1	1	1

Jak widać, na wyjściu jest stan „1”, gdy na jednym lub obu wejściach jest stan „1”.

- funkcja AND (iloczyn logiczny): $Q = A.AND.B$ (inny zapis: $Q = A * B$)
Tablica prawdy dla tej funkcji:

Dana A	Dana B	Wynik Q
0	0	0
0	1	0
1	0	0
1	1	1

Jak widać, na wyjściu jest stan „1” tylko wtedy, gdy obu wejściach jest stan „1”.

Z bramek OR i NOT lub z bramek AND i NOT można zbudować każdą dowolnie skomplikowaną kombinacyjną funkcję logiczną.

W mikroelektronice wygodniej jest budować z tranzystorów funkcje OR i AND zanegowane, zwane NOR i NAND. Tablice prawdy dla tych funkcji otrzymujemy zmieniając zera na jedynki i jedynki na zera w kolumnie „wynik”.

NOR:

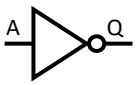
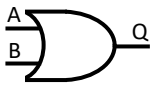
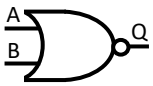
Dana A	Dana B	Wynik Q
0	0	1
0	1	0
1	0	0
1	1	0

NAND:

Dana A	Dana B	Wynik Q
0	0	1
0	1	1
1	0	1
1	1	0

Z samych bramek NOR lub z samych bramek NAND można zbudować każdą dowolnie skomplikowaną kombinacyjną funkcję logiczną.

W schematach układów logicznych stosowane są następujące symbole bramek logicznych:

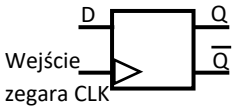
NOT	OR	AND	NOR	NAND
				

Oprócz tych pięciu rodzajów bramek spotkamy jeszcze w układach kombinacyjnych inne bramki – o nich później.

Do budowy układów kombinacyjnych wystarczają bramki opisane wyżej, natomiast do układów sekwencyjnych potrzebne są także przerzutniki – bramki, które tworzą pamięć układów sekwencyjnych. W układach cyfrowych CMOS stosowane są niemal wyłącznie przerzutniki typu D (od angielskiego słowa „delay"). Przerzutniki D wymagają taktowania – sygnału zwanego zegarowym. Jest to ciąg regularnie zmieniających się w czasie zer i jedynek. Działanie przerzutnika D można opisać następująco: w chwili zmiany stanu sygnału zegarowego na wejściu zegarowym przerzutnik pobiera stan z wejścia i zapamiętuje go, jednocześnie na wyjściu pojawia się stan zapamiętany poprzednio (oraz zwykle także jego negacja). Opisać to można następującą tablicą (zwaną w przypadku przerzutników tablicą przejść), gdzie podane są stany w chwili t oraz w następnej chwili t+1 (czas mierzony taktami zegara):

Wejście D(t)	Wyjście Q(t)	Wyjście Q(t+1)	Wyjście .NOT.Q(t+1)
0	0	0	1
0	1	0	1
1	0	1	0
1	1	1	0

Symbol logiczny przerzutnika D jest następujący:



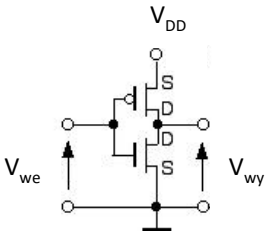
W systemach cyfrowych oprócz przerzutników w układach sekwencyjnych znajdziemy też inny rodzaj układów przechowujących informacje. Są to pamięci – bloki komórek, z których każda zapamiętuje jeden bit, do którego dociera się poprze adres. Dalej poznamy także budowę i działanie komórek pamięci i całych pamięci.

3.2.2 Jak z tranzystorów buduje się bramki logiczne

Gdy z tranzystorów MOS budujemy bramki logiczne, traktujemy tranzystory jako sterowane wyłączniki: tranzystor nMOS jest włączany napięciem dodatnim bramki względem źródła (większym od napięcia progowego), a wyłączany napięciem równym zeru, zaś tranzystor pMOS jest włączany napięciem bramki ujemnym względem źródła (większym co do wartości bezwzględnej od napięcia progowego), a wyłączany napięciem równym zeru.
W tym punkcie omówione będą bramki zwane statycznymi. O bramkach nie należących do tej kategorii będzie mowa później.

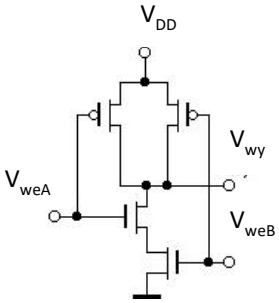
DEFINICJA

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.

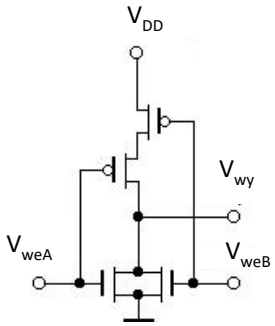


Rysunek 3-27. Inwerter CMOS

Bramka NOT zwana też inwerterem zbudowana jest z dwóch połączonych szeregowo tranzystorów: nMOS i pMOS – rysunek 3-27. Dreny obu tranzystorów są połączone z wyjściem. Bramki obu tranzystorów są połączone ze sobą i sterowane przez sygnał wejściowy. Zasada działania jest bardzo prosta. Gdy wejście jest w stanie „0", czyli napięcie wejściowe jest równe lub bliskie zeru, tranzystor nMOS jest wyłączony (nie przewodzi), zaś tranzystor pMOS jest włączony (przewodzi). Wyjście jest połączone przez tranzystor pMOS ze źródłem zasilania, napięcie na wyjściu jest równe V_{DD}, czyli wyjście jest w stanie „1". Odwrotna sytuacja powstaje, gdy wejście jest w stanie „1", czyli napięcie wejściowe jest równe lub bliskie V_{DD}. Włączony jest wówczas tranzystor nMOS, zaś tranzystor pMOS jest wyłączony. Wyjście jest uziemione przez tranzystor nMOS, a więc napięcie na nim jest równe zeru, co oznacza logiczne „0".



Rysunek 3-28. Bramka NAND



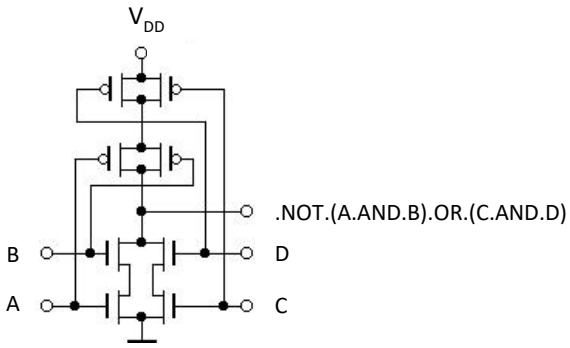
Rysunek 3-29. Bramka NOR

Bramki wykonujące funkcje NOR i NAND tworzy się przez równoległe i szeregowe połączenia tranzystorów. Rysunek 3-28 przedstawia bramkę NAND, a rysunek 3-29 – bramkę NOR.

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

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

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



Rysunek 3-30. Przykład bramki AND-OR-INVERT

Przy pomocy połączeń równoległych i szeregowych można zbudować bramki wykonujące funkcje bardziej złożone, niż NOR i NAND. W tym celu zauważmy, że szeregowo łączone tranzystory nMOS można uznać za realizację funkcji AND, a równolegle łączone tranzystory nMOS za realizację funkcji OR. Zatem łącząc równolegle dwa łańcuchy tranzystorów połączonych szeregowo otrzymamy funkcję (A.AND.B).OR.(C.AND.D). Dla zbudowania kompletnej bramki dodajemy tranzystory pMOS w następujący sposób: każdemu połączeniu szeregowemu tranzystorów nMOS odpowiada połączenie równoległe pMOS, i odwrotnie. Zatem łączymy szeregowo dwie pary równolegle połączonych tranzystorów pMOS. Otrzymujemy w rezultacie schemat jak na rys. 3-30. Bramka realizuje funkcję .NOT.((A.AND.B).OR.(C.AND.D)). Tak zbudowane bramki nazywane bywają bramkami AND-OR-INVERT (w skrócie AOI). Zamieniając połączenia szeregowe na równoległe, a równoległe na szeregowe otrzymamy bramkę realizującą funkcję .NOT.((A.OR.B).AND.(C.OR.D)). Bramki o takiej strukturze nazywane bywają bramkami OR-AND-INVERT (w skrócie OAI). Jak widzimy, w układach CMOS funkcje bardziej złożone, niż NOR i NAND, można zrealizować w postaci pojedynczej bramki.

Teraz już wiemy, jak z tranzystorów budowane są bramki – łączy się odpowiednio, szeregowo i równolegle, tranzystory nMOS i pMOS. Dalej poznamy także inne rodzaje bramek kombinacyjnych, budowę przerzutników oraz komórek pamięci różnych rodzajów. Dowiemy się też, od czego zależą i jakie są właściwości oraz parametry bramek i jak projektant dobiera wymiary tranzystorów w bramkach.

3.3 Nie tylko bramki – świat jest analogowy

3.3.1 Analogowy układ elektroniczny: co to jest i do czego służy

Sygnały w realnym, fizycznym świecie nie mają postaci cyfrowej. Dlatego potrzebne są także analogowe układy elektroniczne. Weźmy dla przykładu dźwięk. Gdy mówimy do mikrofonu, wytwarza on napięcie elektryczne zmieniające się w czasie w taki sam sposób, w jaki zmienia się ciśnienie powietrza pobudzane do drgań naszym głosem. To napięcie z mikrofonu można wzmacniać (czyli zwiększać jego amplitudę) i kształtować na inne sposoby, przesyłać i w końcu ponownie zamienić na dźwięk, doprowadzając do głośnika, którego membrana pobudzi do drgań otaczające powietrze. To przykład systemu w pełni analogowego. Do odbioru sygnałów ze świata fizycznego i przetworzenia na sygnał elektryczny potrzebne są różnego rodzaju czujniki (inaczej – sensory). Mikrofon jest jednym z nich. Istnieją czujniki wielkości fizycznych, na przykład temperatury, ciśnienia, przyspieszenia, drgań (czyli zmian ciśnienia w czasie), natężenia światła itp., a także wielkości chemicznych, na przykład stężenia jakiegoś rodzaju jonów w cieczy lub jakiegoś gazu w atmosferze. Wspólną cechą tych wszystkich czujników jest to, że dostarczają sygnały elektryczne (zwykle w postaci napięć) zmieniające się w czasie w sposób ciągły i mogące przybierać dowolne wartości napięcia w pewnym zakresie. Zatem analogowe układy elektroniczne działają na sygnałach (zwykle napięciach) zmieniających się w czasie w sposób ciągły. W sposób ciągły zmieniają się też napięcia i prądy w tranzystorach, które pracują w tych układach. Tranzystorów nie traktuje się więc jak sterowanych wyłączników.

3.3.2 Rodzaje układów analogowych

Wszystkie układy analogowe można podzielić na dwie duże grupy: układy liniowe i układy nieliniowe. Układy liniowe to te, w których zależność pomiędzy sygnałem wejściowym, a wyjściowym można opisać funkcją liniową. Głównym reprezentantem tej klasy układów są wzmacniacze, czyli układy służące do zwiększania amplitudy

zmiennego sygnału wejściowego. Wzmacniacze mogą być układami szerokopasmowymi, to znaczy wzmacniającymi jednakowo dobrze sygnały o różnych częstotliwościach zawartych w szerokim zakresie, lub wzmacniaczami selektywnymi, które wzmacniają sygnały w wąskim pasmie częstotliwości.

Układy nieliniowe to takie, w których zależność między sygnałami na wejściu i na wyjściu nie jest liniowa. Przykładem mogą służyć układy mnożące, czyli takie, które mają dwa wejścia, na które podawane są dwa różne sygnały, a na wyjściu pojawia się sygnał proporcjonalny do iloczynu sygnałów wejściowych. Jeśli na oba wejścia podany jest ten sam sygnał, to na wyjściu otrzymamy sygnał proporcjonalny do kwadratu sygnału wejściowego. Przy użyciu układów mnożących można też zrealizować wiele innych funkcji nieliniowych.

3.3.3 Pomiędzy światem cyfrowym i analogowym

Powszechnie dziś stosowane jest cyfrowe przetwarzanie sygnałów analogowych. Aby to było możliwe, sygnał analogowy poddawany jest próbkowaniu. Służą do tego układy przetworników analogowo-cyfrowych. Układ taki w kolejnych chwilach czasu bada wartość chwilową napięcia sygnału analogowego i wyraża tę wartość w postaci liczby. Ciąg takich liczb jest cyfrową reprezentacją sygnału analogowego. Im częściej próbkowany jest sygnał analogowy, tym dokładniejsza jest taka reprezentacja. Następnie można poddać ten ciąg liczb rozmaitym operacjom matematycznym w układzie cyfrowym zwanym procesorem sygnałowym - jest to specjalna odmiana mikroprocesora. Analogowy sygnał w postaci cyfrowej można, jak wiemy, przesyłać tak, jak każdy inny ciąg liczb binarnych, można go utrwalić (na przykład w postaci zapisu na płycie CD lub DVD), i w końcu można go przekształcić ponownie na sygnał analogowy. Służy do tego przetwornik cyfrowo-analogowy. Na jego wejście podawane są kolejne liczby będące próbkami sygnału analogowego, na ich podstawie przetwornik odtwarza na wyjściu sygnał w postaci analogowej.

W każdym przetworniku analogowo-cyfrowym musi znajdować się co najmniej jeden komparator napięcia. Jest to układ, który ma dwa wejścia dla sygnałów analogowych i wyjście cyfrowe. Układ porównuje napięcia na wejściach analogowych i w zależności od tego, które z nich jest większe, na wyjściu pojawia się „0” lub „1”.

Warto dodać, że wiele dużych i złożonych układów scalonych zawiera w sobie zarówno bloki analogowe, jak i cyfrowe.

Układy analogowe także będą omawiane dalej.

4 Jak się wytwarza układy scalone i ile to kosztuje

Teraz opowiemy, jak powstają układy scalone: omówione będą główne procesy produkcyjne i sekwencja tych procesów pozwalająca wytworzyć strukturę układu scalonego. Koszt układu – jak i od czego zależy – także jest tu omówiony. Szczególna uwaga będzie zwrócona na uzysk – jaka część spośród wyprodukowanych układów działa prawidłowo, dlaczego nie wszystkie działają, jak to wpływa na koszt i jak starać się, by uzysk był maksymalny, a koszt układu jak najmniejszy. Na koniec będzie parę słów o tym, jaki jest stan mikroelektroniki w Polsce.

4.1 Procesy produkcyjne mikroelektroniki

Aby móc wytworzyć dowolną strukturę układu scalonego, trzeba mieć możliwość wytworzenia obszarów domieszkowanych w półprzewodniku (z nich budowane są struktury tranzystorów i innych elementów), obszarów dielektrycznych (służą one do izolowania elementów oraz ścieżek przewodzących, które nie powinny być połączone elektrycznie) oraz obszarów przewodzących (które tworzą sieć połączeń elektrycznych między elementami układu). Zatem podstawowe operacje technologiczne w mikroelektronice to operacje wytwarzania obszarów domieszkowanych, dielektrycznych i przewodzących. Obszary te muszą mieć właściwe kształty,

wymiary i położenie w układzie. Uzyskuje się to przy zastosowaniu fotolitografii oraz operacji selektywnego trawienia.

4.1.1 Podłoża układów scalonych

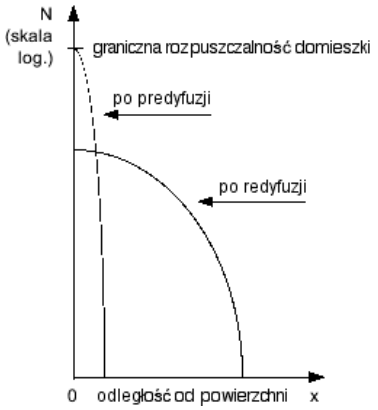
Podstawowym materiałem, z którego wytwarzane są półprzewodnikowe układy scalone, jest krzem (Si). Jest to jeden z najbardziej rozpowszechnionych w przyrodzie pierwiastków. Do wytwarzania wyrobów półprzewodnikowych potrzebny jest krzem w postaci monokrystalicznych płytek, płytki te mają grubość około 1 mm i średnicę równą 20 cm lub 30 cm. Płytki te powstają w wyniku cięcia monokryształów krzemu mających postać walców o średnicy równej średnicy płytek i długości sięgającej kilkudziesięciu centymetrów. Walce te są to monokryształy wyprodukowane metodą Czochralskiego. Po pocięciu monokryształu na płytki podłożowe płytki te poddaje się bardzo starannemu polerowaniu i trawieniu powierzchni, aby były idealnie płaskie i pozbawione wszelkich zanieczyszczeń.

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

4.1.2 Wytwarzanie warstw domieszkowanych

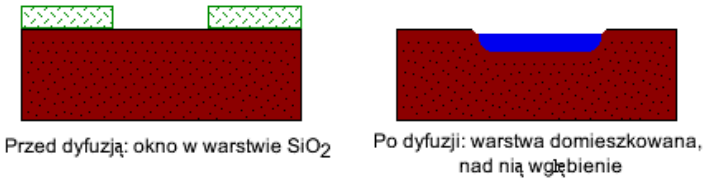
Pod pojęciem warstw domieszkowanych rozumiane są warstwy w płytce podłożowej, w których występują celowo wprowadzone domieszki donorowe lub akceptorowe. Domieszkami donorowymi są pierwiastki z piątej grupy układu okresowego, przede wszystkim fosfor (P) i arsen (As), zaś jako domieszka akceptorowa wykorzystywany jest zwykle bor (B), chociaż akceptorami są także inne pierwiastki z trzeciej grupy układu okresowego (np. aluminium (glin, Al)). W przemyśle półprzewodnikowym stosowane są dziś trzy sposoby wytwarzania warstw półprzewodnika domieszkowanych domieszkami donorowymi lub akceptorowymi: epitaksja, dyfuzja oraz implantacja jonów.

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



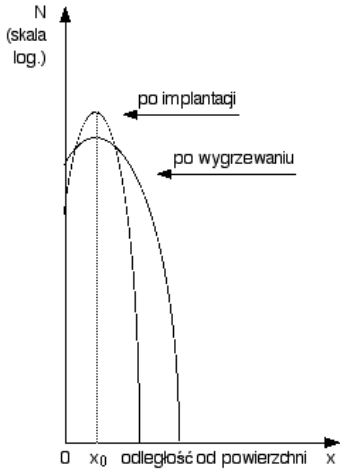
Rysunek 4-1. Rozkłady koncentracji domieszek w procesie dyfuzji

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



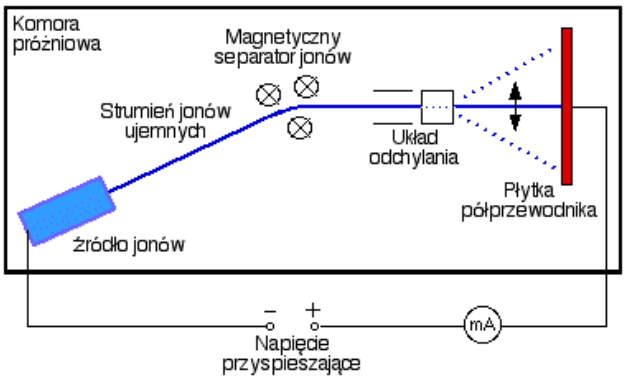
Rysunek 4-2. Dyfuzja selektywna - do obszaru określonego przez okno w warstwie SiO_2

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



Rysunek 4-3. Rozkłady domieszek w procesie implantacji jonów

Implantacja jonów polega na „wstrzeliwaniu” w płytkę podłożową jonów domieszki, którym nadana została duża energia kinetyczna przez rozpędzenie w silnym polu elektrycznym. Implantacja jonów odbywa się w próżni w temperaturze otoczenia. Jony wytracają energię w zderzeniach z atomami półprzewodnika i lokują się w jego sieci krystalicznej. Procesem implantacji można precyzyjnie sterować: zasięg jonów zależy od ich energii (czyli od napięcia przyspieszającego), a dawkę (czyli liczbę implantowanych jonów) można ustalić całkując w czasie prąd jonów płynący między ich źródłem, a płytką podłożową. Dzięki łatwej i precyzyjnej regulacji zasięgu i dawki jonów implantacja jonów pozwala wykonywać warstwy płytkie i słabo domieszkowane. Jest we współczesnej mikroelektronice najczęściej stosowanym sposobem wprowadzania domieszek do wnętrza płytek półprzewodnikowych.



Rysunek 4-4. Implantator jonów

Implantacja jonów odbywa się w implantatorze, którego budowa przypomina akcelerator cząstek elementarnych. Przyspieszane są jednak nie cząstki, lecz jony pierwiastka domieszkującego półprzewodnik. Jony ze źródła jonów są w poprzecznym polu magnetycznym "oczyszczane" (jony różnych pierwiastków mają różną masę, toteż odchylane są pod różnymi kątami) i przyspieszane w silnym polu elektrycznym. Układ elektrostatycznego odchylania odchyla strumień jonów i „przemiata" nim po całej płytce półprzewodnikowej. Jony przyspieszone silnym polem elektrycznym uderzają w płytkę z dużą energią kinetyczną i dzięki temu przemieszczają się w głąb, wytracając stopniowo energię w zderzeniach z atomami sieci krystalicznej półprzewodnika. Równocześnie ich ładunek ulega zobojętnieniu, np. jony ujemne oddają elektron, który przez zewnętrzny obwód odpływa do źródła napięcia przyspieszającego. W zderzeniach jonów z atomami półprzewodnika część tych atomów zostaje przemieszczona, regularna budowa monokryształu zostaje w mniejszym lub większym stopniu zaburzona - powstają defekty sieci krystalicznej. Implantowane atomy domieszki zajmują przypadkowe położenia, wiele z nich ląduje w położeniach międzywęzłowych, gdzie nie wykazują właściwości domieszki donorowej czy też akceptorowej. Po implantacji poddaje się płytkę wygrzewaniu, co powoduje odbudowę regularnej sieci monokryształu, zaś atomy domieszki lokują się w węzłach sieci krystalicznej. Równocześnie rozkład domieszki ulega redyfuzji. Implantację można wykonać poprzez bardzo cienką warstwę SiO_2. Wówczas maksimum rozkładu domieszek może znaleźć się bliżej powierzchni, a nawet wewnątrz warstwy tlenku. Implantacja przez tlenek zmniejsza energię jonów bombardujących monokryształ i tym samym zmniejsza liczbę powstających defektów. Implantację, podobnie jak dyfuzję, wykonuje się zazwyczaj przez okno wykonane techniką fotolitografii.

Regulując czasy i temperatury predyfuzji i redyfuzji (w przypadku dyfuzji) lub energię, dawkę jonów oraz czas i temperaturę wygrzewania po implantacji (w przypadku implantacji jonów) można dość dokładnie kontrolować końcowy rozkład domieszek. Jednak w całym cyklu produkcyjnym występuje wiele operacji wysokotemperaturowych, a w każdej z nich zachodzi proces redyfuzji wcześniej wprowadzonych domieszek. Dlatego zgranie warunków wszystkich operacji w całym procesie produkcji układów jest trudne, wymaga symulacji komputerowych i eksperymentów. Raz ustawiony proces nie jest już później zmieniany. Konstruktor układu scalonego nie może wymagać zmiany warunków procesu, musi dostosować projekt do istniejącego procesu produkcyjnego.

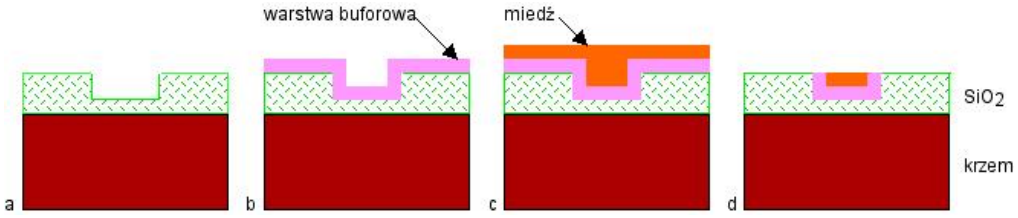
4.1.3 Warstwy dielektryczne

Dominujące w mikroelektronice warstwy SiO_2 można wytwarzać przez utlenianie krzemu, co wykonuje się umieszczając płytkę krzemową w atmosferze utleniającej w wysokiej temperaturze. Jest to jednak możliwe tylko w miejscach, w których powierzchnia krzemu jest odsłonięta. W każdym innym przypadku warstwę SiO_2 uzyskuje się przez osadzanie SiO_2 z par powstających w wyniku reakcji chemicznej w fazie gazowej (ang. Chemical Vapor Deposition, CVD). Utlenianie krzemu daje warstwy o najlepszych właściwościach elektrycznych, dlatego w ten sposób wytwarza się warstwy dielektryku bramkowego w tranzystorach MOS. Inne rodzaje warstw dielektrycznych uzyskuje się przez osadzanie różnymi metodami, np. Si_3N_4 podobnie jak SiO_2 metodą CVD. W najbardziej zaawansowanych technologiach spotyka się też inne dielektryki, niż czysty SiO_2. Przykładowo, dodatek pierwiastka ziem rzadkich – hafnu – do warstwy SiO_2 pod bramką tranzystora MOS zwiększa przenikalność dielektryczną tej warstwy, co poprawia parametry tranzystora.

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

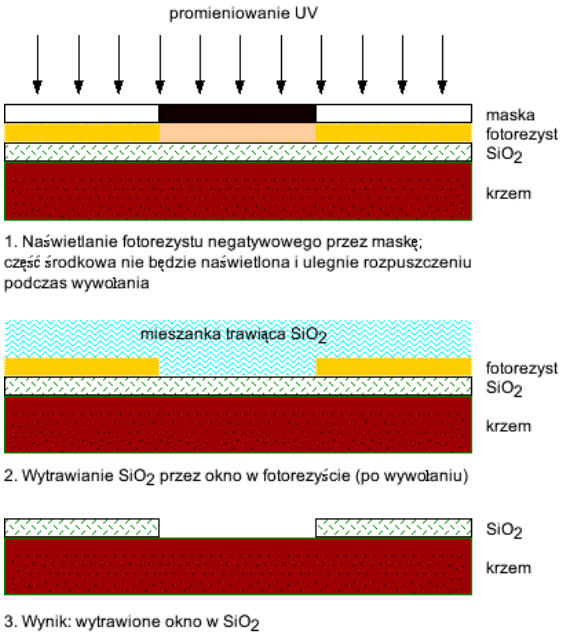
4.1.4 Warstwy przewodzące

W pierwszych latach rozwoju mikroelektroniki do wytwarzania sieci połączeń w układach scalonych stosowane było aluminium. Warstwę aluminium otrzymuje się w prosty sposób przez naparowanie w próżni ze źródła par aluminium. Równie ważne są warstwy polikrzemu, z których w wielu wariantach technologii wytwarza się bramki tranzystorów MOS. Polikrzem jest osadzany w wyniku wydzielania się atomów krzemu przez rozkład silanu (SiH_4) w wysokiej temperaturze. Lepsze od aluminium właściwości mają warstwy miedziane. Wytwarzanie warstw przewodzących z miedzi jest procesem daleko bardziej skomplikowanym. Proces ten, zwany damasceńskim, składa się z kilku etapów (rysunek 4-5). Najpierw przy użyciu fotolitografii i trawienia wykonuje się rowki w warstwie SiO_2 - w nich będą ścieżki przewodzące (rysunek 4-5a). Następnie na płytce osadzana jest warstwa buforowa, ma ona dobrą przyczepność do SiO_2, nie przepuszcza atomów miedzi w kierunku płytki podłożowej i jest przewodząca (rysunek 4-5b). Płytka pokrywana jest elektrolitycznie warstwą miedzi (rysunek 4-5c). W ostatnim kroku wykonywane jest mechaniczno - chemiczne polerowanie płytki (ang. Chemical - Mechanical Polishing, CMP), po którym miedź pozostaje tylko w głębi rowków.



Rysunek 4-5. Proces wytwarzania miedzianej ścieżki przewodzącej

4.1.5 „Rzeźbienie w krzemie", czyli fotolitografia



Rysunek 4-6. Proces wytwarzania okna w warstwie SiO_2 przy pomocy fotolitografii

Służy do nadawania obszarom domieszkowanym, dielektrycznym i przewodzącym wymaganych położeń, kształtów i wymiarów. Wykorzystywana jest tu wrażliwość niektórych związków chemicznych na promieniowanie elektromagnetyczne. Istnieją substancje, które pod wpływem tego promieniowania (w mikroelektronice - ultrafioletu) ulegają utwardzeniu - tracą rozpuszczalność w określonych rozpuszczalnikach. Takie substancje nazywamy fotorezystami negatywowymi. Inne substancje pod wpływem promieniowania stają się łatwo rozpuszczalne. Są to fotorezysty pozytywowe. Rysunek 4-6 pokazuje zasadę fotolitografii na przykładzie wykonywania okna w warstwie SiO_2. Płytka jest pokrywana warstwą fotorezystu negatywowego, następnie naświetlana przez maskę mającą obszary przezroczyste i nieprzezroczyste, naświetlona płytka jest następnie zanurzona w rozpuszczalniku rozpuszczającym nienaświetlony fotorezyst, po czym mieszanka trawiąca SiO_2, a nie naruszająca fotorezystu, wytrawia okno w warstwie SiO_2. W podobny sposób wytrawia się w warstwie aluminium ścieżki przewodzące. W przypadku procesu implantacji jonów okna w warstwie fotorezystu bezpośrednio określają obszary, do których wprowadzone będą jony

domieszki – warstwa fotorezystu o dostatecznej grubości zatrzymuje jony nie dopuszczając ich do powierzchni półprzewodnika.

Rysunek 4-6 pokazuje najprostszy proces fotolitograficzny, w którym płytka jest naświetlana bezpośrednio przez maskę (fotolitografia kontaktowa). We współczesnej mikroelektronice stosowana jest fotolitografia projekcyjna – obraz maski jest wyświetlany na płytkę przez obiektyw, na podobnej zasadzie, jak w zwykłym rzutniku do przezroczy lub powiększalniku fotograficznym. Urządzenie do naświetlania zwane jest potocznie steperem, ponieważ naświetla nie całą płytkę równocześnie, lecz kolejne układy na płytce („step and repeat" – „zrób krok i powtórz").

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

4.2 Jak powstaje i ile kosztuje układ scalony

4.2.1 Układy CMOS w technologii LOCOS

W układach CMOS elementami czynnymi, jak już wiemy, są dwa typy tranzystorów: nMOS i cMOS. Dla tranzystorów o dwóch różnych typach kanału potrzebne są obszary podłoża o dwóch typach przewodnictwa. Obecnie produkuje się niemal wyłącznie układy, w których podłożem jest płytka półprzewodnikowa typu *p* (jest to podłoże dla tranzystorów nMOS), a w niej wytwarza się wyspy o przewodnictwie typu *n* (są one podłożem dla tranzystorów pMOS). Między wyspami, a podłożem powstają złącza *p-n*.

UWAGA

Wyspy typu n dla zapewnienia izolacji od podłoża typu p muszą być względem podłoża spolaryzowane zaporowo. Podłoże jest z reguły połączone z ujemnym biegunem napięcia zasilania układu. Obszary wysp typu n najczęściej połączone są z dodatnim biegunem napięcia zasilania, ale w zasadzie mogą być spolaryzowane dowolnym napięciem dodatnim w stosunku do podłoża.

W normalnych warunkach polaryzacji wszystkie obszary tranzystora MOS (źródło, dren i kanał) są spolaryzowane zaporowo względem podłoża, na którym są wykonane, toteż w jednym obszarze podłoża (lub wyspy w przypadku tranzystorów pMOS) można umieścić wiele tranzystorów, i będą one wzajemnie od siebie odizolowane.

Prześledzimy teraz etapy powstawania struktury układu scalonego CMOS. Ilustrowane one będą przekrojami przez tę strukturę. Będzie to układ wytwarzany w technologii zwanej LOCOS (ang. LOCal Oxidation of Silicon). Lokalnie wytwarzany tlenek polowy oddziela obszary aktywne, w których wykonywane są tranzystory. Technologia ta królowała w mikroelektronice przez dziesiątki lat. Obecnie istnieją też technologie bardziej zaawansowane, opowiemy o nich dalej.

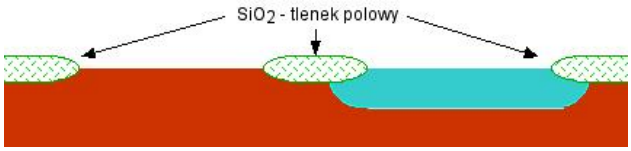
Podłożem jest płytka krzemowa typu *p* o grubości około 1 mm (skala pionowa przekrojów nie jest zachowana).

Rysunek 4-7. Podłoże

Etap 1: Wytworzenie wysp typu n. Wyspa typu n powstaje w wyniku procesu fotolitografii i następującego po nim procesu domieszkowania (implantacji jonów donorowych).

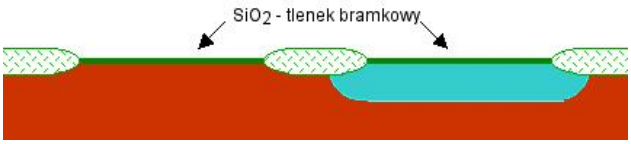
Rysunek 4-8. Podłoże typu p z wyspą typu n

Etap 2: Wytworzenie obszarów grubego (0.5 - 1 μm) tlenku zwanego polowym, pomiędzy nimi obszary zwane aktywnymi. Płytka jest pokrywana azotkiem krzemu (Si_3N_4), który następnie jest usuwany w procesie fotolitografii znad obszarów, gdzie będzie tlenek polowy. Następnie płytka jest utleniania. Obszary SiO_2 powstają tam, gdzie usunięto azotek. Na koniec azotek jest usuwany chemicznie, pozostaje tlenek i odsłonięte obszary aktywne. W obszarach aktywnych powstaną tranzystory.



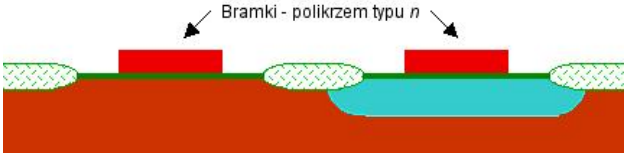
Rysunek 4-9. Tlenek polowy i obszary aktywne

Etap 3: Wytworzenie tlenku bramkowego. Powstaje w wyniku utleniania odsłoniętych powierzchni obszarów aktywnych. Grubość tlenku bramkowego jest bardzo mała. Będzie to dielektryk pod bramkami tranzystorów.



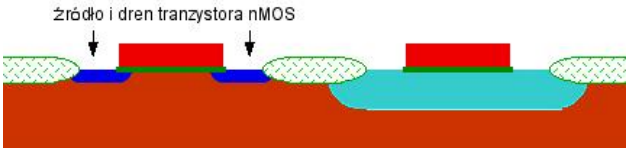
Rysunek 4-10. Tlenek bramkowy w obszarach aktywnych

Etap 4: Wytworzenie bramek tranzystorów. Bramki powstają przez osadzenie warstwy polikrzemu domieszkowanego atomami donorowymi oraz proces fotolitografii.



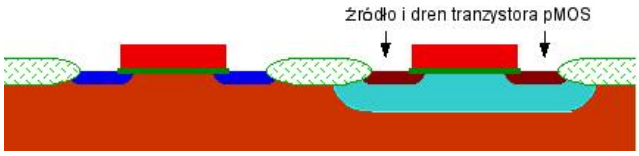
Rysunek 4-11. Bramki na dielektryku w obszarach aktywnych

Etap 5: Wytworzenie źródeł i drenów tranzystorów nMOS. Źródła i dreny tranzystorów nMOS powstają w wyniku implantacji jonów donorowych. Przed tym procesem wykonywana jest fotolitografiia, której celem jest zasłonięcie fotorezystem obszarów aktywnych na wyspach, gdzie będą tranzystory pMOS. W procesie implantacji obszary źródeł i drenów powstają tam, gdzie nie ma tlenku polowego ani polikrzemu - warstwy te są na tyle grube, że nie przepuszczają jonów domieszki.



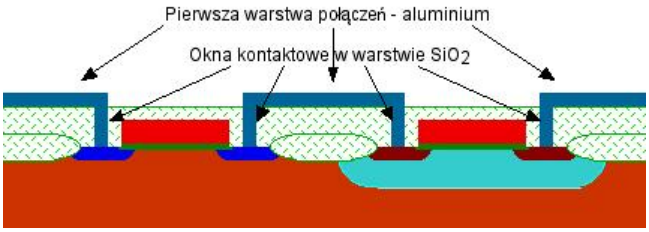
Rysunek 4-12. Źródła i dreny tranzystorów nMOS

Etap 6: Wytworzenie źródeł i drenów tranzystorów pMOS. Źródła i dreny tranzystorów pMOS powstają w wyniku implantacji jonów akceptorowych. Przed tym procesem wykonywana jest fotolitografiia, której celem jest zasłonięcie już wykonanych tranzystorów nMOS. W procesie implantacji obszary źródeł i drenów powstają tam, gdzie nie ma tlenku polowego ani polikrzemu - warstwy te są na tyle grube, że nie przepuszczają jonów domieszki. Po tym etapie wytworzone są już tranzystory obu typów.



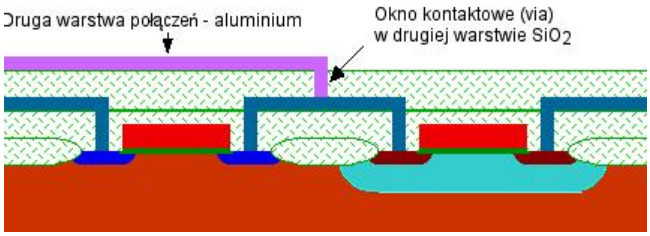
Rysunek 4-13. Źródła i dreny tranzystorów pMOS

Etap 7: Wytworzenie pierwszego poziomu połączeń. Aby wykonać połączenia elektryczne, pokrywa się płytkę dielektrykiem, po czym wykonuje się fotolitografię i wytrawienie okien kontaktowych w tym dielektryku. Następnie płytkę pokrywa się warstwą metalu i wykonuje kolejną fotolitografię, w wyniku której powstają ścieżki połączeń. Na przekroju nie pokazano połączeń do bramek tranzystorów. W tym przekroju nie są one widoczne, ponieważ nie wykonuje się ich nad kanałami tranzystorów.



Rysunek 4-14. Pierwszy poziom połączeń.

Etap 8: Wytworzenie drugiego poziomu połączeń. Kolejny poziom połączeń elektrycznych wykonuje się pokrywając poprzednie połączenia drugą warstwą dielektryka, następnie wykonuje się w niej okna kontaktowe (zwane potocznie via) do ścieżek połączeń pierwszej warstwy, osadza kolejną warstwę metalu i wykonuje kolejną fotolitografię, w wyniku której powstają ścieżki połączeń drugiego poziomu. Ścieżki drugiego poziomu mogą kontaktować się tylko ze ścieżkami pierwszego poziomu, nie ma bezpośrednich kontaktów między ścieżkami drugiego poziomu, a źródłami, drenami i bramkami tranzystorów. W nowszych procesach technologicznych poziomów połączeń jest zwykle więcej niż dwa, nawet do kilkunastu.

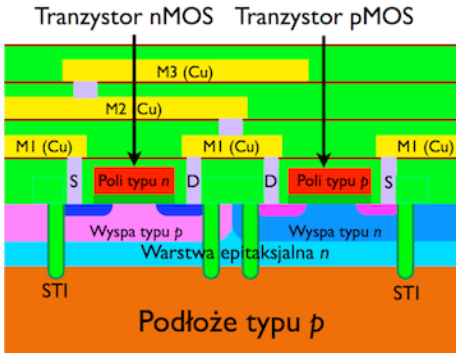


Rysunek 4-15. Drugi poziom połączeń

Na gotowy układ nakładana jest warstwa szkliwa ochronnego, w którym fotolitograficznie wytwarza się duże okna nad obszarami metalu, do których będą dołączone zewnętrzne wyprowadzenia. Szkliwo ochronne i okna w nim nie są pokazane na rysunkach 4-7 do 4-15. Cały proces produkcyjny liczy od 200 do 500 i więcej operacji, takich jak utlenianie, nakładanie różnych warstw, operacje fotolitograficzne (nakładanie fotorezystu, naświetlanie, wywoływanie, trawienie, usuwanie fotorezystu), operacje domieszkowania (implantacja, wygrzewanie poimplantacyjne), czyszczenie i mycie płytek pomiędzy poszczególnymi operacjami itp.

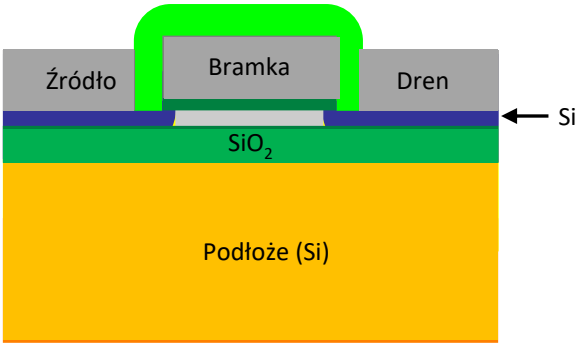
4.2.2 Układy CMOS w technologii STI

W nowszych technologiach obszary aktywne nie są od siebie oddzielane obszarami tlenku polowego, lecz rowkami trawionymi w głąb krzemu, a następnie wypełnianymi dielektrykiem (SiO_2). Umożliwia to zmniejszenie powierzchni zajmowanej przez układ. Układy w technologii STI mają też zwykle wiele innych ulepszeń, na przykład więcej niż dwie warstwy połączeń. Przykładowy przekrój układu w technologii STI pokazuje rysunek 4-16. Widać tam również, że w układzie są dwa rodzaje wysp i warstwa epitaksjalna. Kontakty do źródeł i drenów (a także bramek) wykonane są w ten sposób, że najpierw okna kontaktowe są wypełniane metalem takim, jak na przykład molibden, a dopiero potem wykonywane są miedziane ścieżki połączeń.



Rysunek 4-16. Przekrój przez układ w technologii STI z trzema warstwami połączeń miedzianych

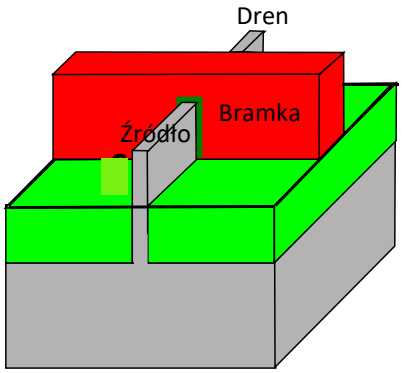
Skrót STI pochodzi od angielskiego terminu „Shallow Trench Isolation". Spotyka się też określenie DTI („Deep Trench Isolation") w przypadku, gdy rowki izolujące obszary aktywne są głębsze.



Rysunek 4-17. Tranzystor MOS w technologii FDSOI

4.2.3 Najnowsze technologie CMOS: FDSOI, FinFET

Gdy długości kanałów tranzystorów zmniejszono poniżej 100 nanometrów, okazało się, że takie bardzo „krótkie" tranzystory nie dają się całkowicie wyłączyć. Nawet przy braku napięcia na bramce płynie niepomijalny prąd drenu – jest to prąd podprogowy. Zjawisko to można w pewnym stopniu ograniczyć wykonując kanał tranzystora w bardzo cienkiej warstwie krzemu (kilka nanometrów). Rysunek 4-17 pokazuje przekrój przez tranzystor nMOS w technologii FDSOI (skrót od ang. „Fully Depleted Silicon on Insulator"). Obszar kanału tranzystora wykonany jest w bardzo cienkiej (kilka nanometrów) warstwie krzemu monokrystalicznego odizolowanej od podłoża warstwą dielektryka. Przy braku polaryzacji bramki obszar kanału jest całkowicie zubożony w nośniki ładunku. Podobna idea zrealizowana jest w technologii FinFET w inny sposób (rysunek 4-18). Kanał tranzystora znajduje się w wąskim, pionowym pasku krzemu (ang. „Fin") otoczonym warstwą dielektryka i bramką.

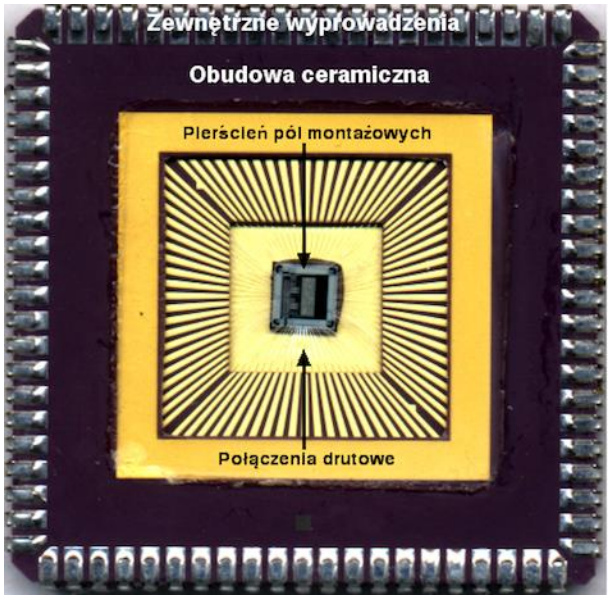


Rysunek 4-18. Tranzystor MOS w technologii FinFET

4.2.4 Montaż i obudowy

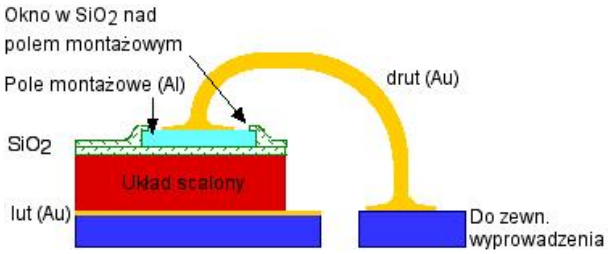
Układy scalone po wyprodukowaniu i wstępnych testach (zwanych testami ostrzowymi – o nich dalej) są montowane w obudowach. Obudowy umożliwiają elektryczne i mechaniczne połączenie układu scalonego z urządzeniem, w którym układ ma działać, zapewniają ochronę układu przed uszkodzeniami mechanicznymi i szkodliwymi wpływami środowiska (np. wilgocią) oraz umożliwiają odprowadzenie ciepła wydzielającego się w czasie pracy układu. Obudowy wykonywane są z tworzywa sztucznego lub z ceramiki. Obudowy z tworzyw są najtańsze w produkcji wielkoseryjnej, stosuje się je więc do montażu układów katalogowych produkowanych masowo. Istotną wadą obudów z tworzyw sztucznych jest znaczna różnica współczynnika rozszerzalności cieplnej tworzywa i płytki półprzewodnikowej. Ogranicza to zakres temperatur, w jakich mogą pracować układy zamknięte w takich obudowach. Obudowy te nie zapewniają także idealnej szczelności, zwłaszcza gdy poddawane są częstym zmianom temperatury w szerokim zakresie. Obudowy ceramiczne są znacznie droższe, ale mają wiele zalet. Znacznie lepiej chronią układ przed szkodliwymi wpływami zewnętrznymi, umożliwiają pracę układu w szerszym zakresie temperatur oraz lepsze odprowadzanie ciepła. Dlatego układy zamknięte w takich obudowach cechują się zwykle wyższą niezawodnością.

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



Rysunek 4-19. Układ scalony w obudowie ceramicznej bez pokrywki

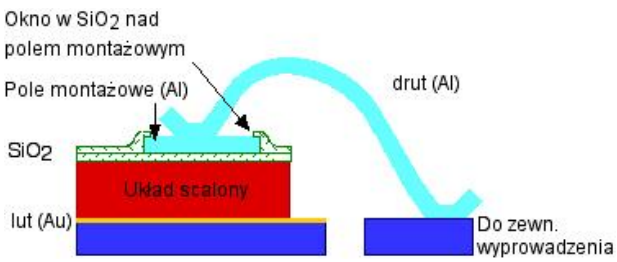
Obudowa ceramiczna składa się z dwóch części: podstawy i pokrywki. Podstawa jest niemal kompletną obudową. Zawiera wnękę, w której umieszczony będzie układ, oraz komplet odpowiednio ukształtowanych zewnętrznych wyprowadzeń. Układ jest mocowany (lutowany lutem złotym lub klejony) do przeznaczonego na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



Rysunek 4-20. Połączenie wykonane metodą termokompresji

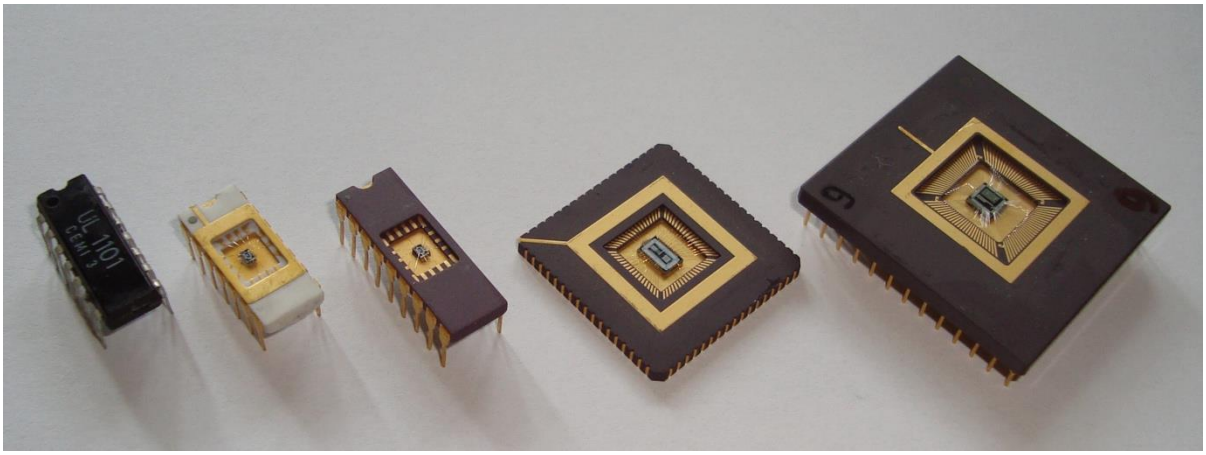
Połączenia między układem, a zewnętrznymi wyprowadzeniami wykonuje się drutem złotym metodą termokompresji lub niekiedy drutem aluminiowym metodą ultrakompresji. Metoda termokompresji polega na tym, że drut z uformowaną na końcu kulką jest specjalnym narzędziem dociskany do miejsca, w którym ma nastąpić elektryczne połączenie, a wszystko to dzieje się w podwyższonej temperaturze. Pod wpływem nacisku i wysokiej temperatury kulka ulega deformacji i zarazem trwale łączy się z metalem, do którego jest dociskana. W metodzie ultrakompresji zamiast wysokiej temperatury stosuje się drgania ultradźwiękowe, a drut jest dociskany narzędziem o kształcie klina. Oba sposoby wykonania połączeń pokazują rysunki 4-20 i 4-21.

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



Rysunek 4-21. Połączenie wykonane metodą ultrakompresji

zminiaturyzowane elementy są zmontowane metodą montażu powierzchniowego. Obecnie powszechnie stosowane są luty bezołowiowe, których temperatura topnienia jest wyższa, niż dawniej używanych lutów ołowiowo-cynowych. Demontaż układu wlutowanego lutem bezołowiowym jest szczególnie trudny.



Rysunek 4-22. Przykłady układów w obudowach

Rysunek 4-22 pokazuje zdjęcia układów w różnego rodzaju obudowach. Z obudów ceramicznych zdjęte zostały pokrywki dla pokazania wnętrza. Pierwsza z lewej: obudowa typu DIL14 z tworzywa, do montażu przewlekanego. Druga: obudowa typu DIL14 ceramiczna, z pokrywką metalową, do montażu przewlekanego. Trzecia: obudowa typu DIL18 ceramiczna, z pokrywką ceramiczną, do montażu przewlekanego. Czwarta: obudowa typu PLCC68 ceramiczna, z pokrywką ceramiczną, do montażu w podstawce lub powierzchniowego. Piąta: obudowa typu PGA100 ceramiczna, z pokrywką ceramiczną, do montażu w podstawce lub przewlekanego.

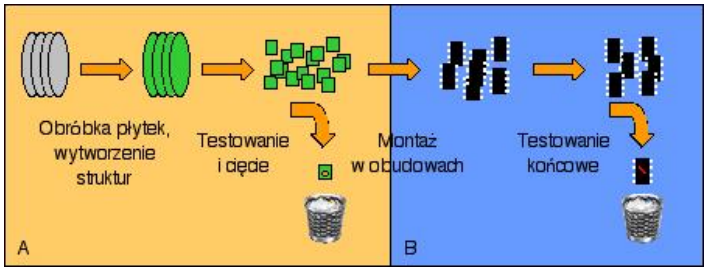
Stosowane bywa także montowanie układu bez obudowy. Ten sposób jest wykorzystywany wtedy, gdy nawet najmniejsze z istniejących obudów zajęłyby zbyt dużo miejsca (np. w zegarkach, kartach płatniczych itp.) lub gdy tradycyjne połączenia drutowe między układem i wyprowadzeniami byłyby zbyt długie (układy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

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

4.2.5 Od czego i jak zależy koszt układu scalonego

Całkowity koszt jednego egzemplarza układu scalonego jest sumą dwóch składników:
- K_{prod} – koszt wytworzenia jednego egzemplarza układu, jego montażu i testowania,
- K_{proj} – część kosztu zaprojektowania układu i przygotowania jego produkcji przypadająca na jeden wyprodukowany egzemplarz układu.

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



Rysunek 4-23. Wytwarzanie układu scalonego: faza A - wytworzenie struktur, testy ostrzowe i cięcie, faza B - montaż i testy końcowe

Proces wytworzenia gotowego układu scalonego można podzielić na dwie fazy, zilustrowane symbolicznie na rysunku 4-23. W pierwszej fazie (A na rysunku 4-23) płytki półprzewodnikowe przechodzą proces obróbki, w wyniku którego powstają na nich struktury układów scalonych. Wszystkie wytworzone struktury są kolejno poddawane testom zwanym testami ostrzowymi. Kontakt z elektrycznymi wyprowadzeniami każdej struktury na płytce zapewniają sprężyste ostrza. Automatyczny tester doprowadza do wejść sygnały testowe i bada prawidłowość sygnałów wyjściowych układu. Układy niesprawne są znakowane farbą, co pozwala je później odsortować i odrzucić. Dopiero po wykonaniu testów ostrzowych płytka jest cięta na poszczególne struktury. Struktury zakwalifikowane jako sprawne przechodzą do drugiej fazy (B na rysunku 4-23). Struktury są montowane w obudowach, a po montażu następują testy końcowe. W tych testach niektóre układy okazują się niesprawne mimo przejścia przez testy ostrzowe. Zmontowane układy mogą okazać się niesprawne z dwóch powodów. Po pierwsze, testy ostrzowe nie zawsze pozwalają zbadać dostatecznie dokładnie działanie układu. Po drugie, układ może ulec uszkodzeniu w trakcie montażu lub też montaż może być wykonany wadliwie.

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

Równanie 4-1

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

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

Równanie 4-2

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

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

Równanie 4-3

$$u_m = \frac{N_{su}}{N_{so}}$$

Ostatecznie, z początkowej liczby N_u pozostanie N_{su} sprawnych układów. Stosunek N_{su}/N_u nazwiemy uzyskiem ostatecznym u. Jest on iloczynem uzysków produkcyjnego i montażowego i w praktyce zawsze jest mniejszy od jedności.

Równanie 4-4

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

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

Możemy już teraz policzyć, ile kosztuje wyprodukowanie jednego sprawnego układu scalonego. Koszt k_s należy pomnożyć przez liczbę układów poddanych montażowi i testom końcowym N_{so}, zsumować z kosztem K_A, a następnie sumę tę podzielić przez liczbę gotowych sprawnych układów N_{su}. Łącząc równania 4-1 do 4-4 otrzymujemy następującą zależność określającą całkowity koszt K_{prod} wytworzenia jednego sprawnego układu:

Równanie 4-5

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

Ze wzoru 4-5 widać, że:
- koszt wytworzenia jednego egzemplarza układu scalonego składa się z dwóch składników: jeden z nich (k_s) jest w przybliżeniu proporcjonalny do liczby zewnętrznych wyprowadzeń układu, drugi do jego powierzchni A_u, przy czym proporcje tych składników mogą być różne,
- układ jest tym droższy, im niższe są uzyski: produkcyjny u_p i montażu u_m.

4.2.6 Defekty, rozrzuty produkcyjne, a uzysk i koszt

Jednym z kluczowych czynników w kształtowaniu się kosztu układu jest uzysk. Stąd pytanie: dlaczego pewna część wyprodukowanych układów scalonych okazuje się być wadliwa (nie spełnia wymagań technicznych) i w testach ostrzowych lub końcowych zostaje odrzucona?

Wyprodukowany układ może w ogóle nie działać (nie wykonywać swej funkcji).

DEFINICJA

O układzie, który w ogóle nie działa, mówimy, że wystąpiło w nim uszkodzenie katastroficzne.

Może być i tak, że układ działa, ale jego parametry nie mieszczą się w dopuszczalnych granicach zwanych granicami tolerancji.

DEFINICJA

O układzie, który działa, ale jego parametry nie mieszczą się w granicach tolerancji, mówimy, że wystąpiło w nim uszkodzenie parametryczne.

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

Przyczyną uszkodzeń parametrycznych są zwykle nadmierne rozrzuty produkcyjne. Te omówimy dokładniej, bowiem ich występowanie ma bezpośredni i silny wpływ na to, jak projektuje się układy scalone. W mikroelektronice - jak w każdej innej dziedzinie techniki - we wszystkich operacjach technologicznych występują nieuchronnie zaburzenia powodujące, że wyniki operacji nigdy nie są dokładnie zgodne z zamierzonymi. Zaburzenia te nazywamy rozrzutami produkcyjnymi. Procesy produkcyjne mikroelektroniki są niezwykle subtelne, toteż nawet znikomo małe zaburzenia tych procesów prowadzą do dużych rozrzutów parametrów i charakterystyk elementów układów scalonych. Nie należą do rzadkości rozrzuty na poziomie na przykład $\pm$50% wartości nominalnej danego parametru - rzecz nie do pomyślenia w innych dziedzinach techniki, np. w mechanice. Sztuka projektowania układów scalonych polega między innymi na tym, by z elementów o bardzo dużych rozrzutach parametrów zbudować układ, który nie tylko będzie działał, ale którego parametry użytkowe będą utrzymane w wąskich granicach tolerancji.

Jest to możliwe, jeśli poznamy dokładniej naturę rozrzutów produkcyjnych. Można je podzielić na dwa rodzaje: rozrzuty globalne i rozrzuty lokalne.

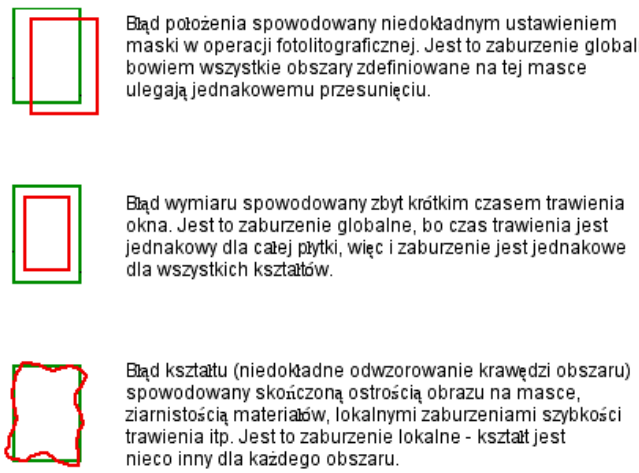
DEFINICJA

Rozrzutami globalnymi nazywamy takie rozrzuty produkcyjne, które jednakowo oddziałują na wszystkie elementy w układzie scalonym.

DEFINICJA

Rozrzutami lokalnymi nazywamy takie rozrzuty produkcyjne, które dla każdego elementu mają inną wielkość.

Innymi słowy, gdyby istniały tylko rozrzuty globalne, to w danym układzie elementy identycznie zaprojektowane miałyby zawsze identyczne parametry i charakterystyki (które jednak miałyby różne wartości w układach pochodzących z różnych płytek i różnych serii produkcyjnych). Natomiast rozrzuty lokalne powodują istnienie różnic pomiędzy elementami, które powinny być identyczne. Różnicę między rozrzutami globalnymi i lokalnymi wygodnie jest pokazać na przykładzie fotolitografii. Obszary uzyskane w wyniku procesu fotolitografii nigdy nie są identyczne z zaprojektowanymi. Trzy rodzaje zaburzeń prowadzących do powstawania różnic między kształtem zaprojektowanym, a uzyskanym, pokazuje rysunek 4-24. Podział rozrzutów na lokalne i globalne nie dotyczy wyłącznie fotolitografii. Praktycznie wszystkie rozrzuty produkcyjne mają składową globalną i składową lokalną. Przykładowo, podczas procesów wysokotemperaturowych kilkadziesiąt płytek znajduje się w piecu w różnych miejscach, każda z nich w nieco innej temperaturze. Jest to rozrzut globalny. Różnice temperatury występują jednak także w obrębie każdej płytki. Jest to rozrzut lokalny.



Rysunek 4-24. Trzy rodzaje zaburzeń obserwowanych w procesach fotolitografii. Zielony kontur: kształt obszaru według projektu, czerwony kontur: kształt rzeczywiście otrzymany

UWAGA

Cechą charakterystyczną mikroelektroniki jest to, że chociaż rozrzuty globalne są bardzo duże, to równocześnie rozrzuty lokalne są małe.

Innymi słowy, *nie można* liczyć na to, że wyprodukowane elementy będą miały parametry zawsze bardzo bliskie nominalnym, ale *można* liczyć na to, że para elementów zaprojektowanych jako identyczne i znajdujących się tuż obok siebie w tym samym układzie scalonym będzie miała prawie identyczne parametry. Tę własność powszechnie wykorzystuje się w projektowaniu układów scalonych, a zwłaszcza układów analogowych.

4.2.7 Pracochłonność i koszt projektu układu scalonego

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

Równanie 4-6

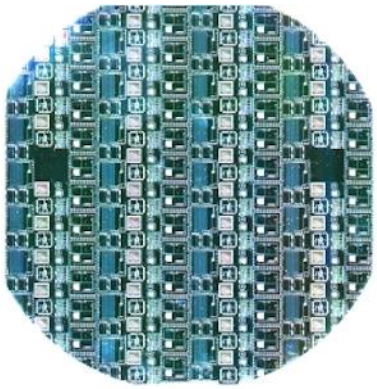
$$K_{calk} = K_{prod} + \frac{K_{proj}}{S}$$

Innymi słowy, koszt zaprojektowania układu i przygotowania jego produkcji rozkłada się na wszystkie wyprodukowane egzemplarze układu.

Koszt zaprojektowania układu jest proporcjonalny do potrzebnych do tego nakładów pracy. Te mogą być bardzo wysokie. Kiedyś, gdy układy scalone liczyły nie więcej niż kilkadziesiąt elementów, cały proces projektowania był wykonywany przez człowieka (ale przy wykorzystaniu wspomagających programów komputerowych):

projektant opracowywał schemat układu (logiczny i/lub elektryczny), wykonywał wszystkie obliczenia projektowe, projektował poszczególne elementy, rozmieszczał je w układzie i projektował połączenia. Ten sposób projektowania nazywany jest projektowaniem w stylu *full custom* (brak udanego polskiego terminu). Ma on i dziś pewien obszar zastosowań, o czym będzie mowa dalej. Pracochłonność tego sposobu projektowania można w dużym przybliżeniu oszacować następująco: doświadczony projektant mający do dyspozycji współczesne oprogramowanie wspomagające zużywa, średnio biorąc, jedną godzinę pracy na jeden tranzystor. Czy to dużo? Jeżeli układ liczy 1000 tranzystorów, to pracochłonność projektu wynosi około 7 osobo-miesięcy. Ale dla mikroprocesora mającego 40 milionów tranzystorów (np. klasy Pentium 4) otrzymujemy pracochłonność rzędu 20 000 osobo-lat! Dalej poznamy sposoby wspomaganego i zautomatyzowanego projektowania, które pozwalają zmniejszyć pracochłonność o wiele rzędów wielkości. Niemniej, projekt dużego i złożonego układu może być bardzo kosztowny.

Koszt przygotowania produkcji to w praktyce koszt wykonania kompletu masek, które służą w procesach fotolitografii do określenia kształtów, wymiarów i położenia wszystkich elementów i połączeń w układzie. Dla starszych, mniej zaawansowanych technologii (gdzie minimalny wymiar w układzie jest rzędu 1 mikrometra lub nieco mniej) koszt wykonania takiego kompletu wynosi kilkanaście do kilkudziesięciu tysięcy dolarów. Dla najnowocześniejszych procesów technologicznych potrzebne są maski o znacznie bardziej skomplikowanej strukturze i jest ich znacznie więcej. Koszt wykonania kompletu takich masek może sięgać nawet miliona dolarów. Jest oczywiste, że bardzo poważnie ograniczałoby to ekonomiczny sens projektowania układów, które miałyby być wyprodukowane w niezbyt wielkiej liczbie egzemplarzy. Problem ten dotyczy w szczególności układów specjalizowanych, o których była mowa we wstępie, a w jeszcze większym stopniu układów prototypowych, do testów i badań, które potrzebne są zwykle w liczbie zaledwie kilkudziesięciu egzemplarzy.



Rysunek 4-25. Płytka wieloprojektowa. Na płytce znajduje się 7 różnych układów CMOS. Zdjęcie płytki wyprodukowanej na linii doświadczalnej Instytutu Technologii Elektronowej w Warszawie

Istnieje jednak rozwiązanie tego problemu w postaci płytek wieloprojektowych (Multi-Project Wafer, w skrócie MPW). Płytką wieloprojektową nazywamy płytkę, na której wytwarza się równocześnie wiele różnych układów według różnych projektów. Całkowity koszt wykonania kompletu masek rozkłada się wtedy na wiele projektów. W jednej partii produkcyjnej wytwarzane są w ten sposób układy dla bardzo wielu różnych odbiorców. Dzięki takiej organizacji produkcji, oraz zautomatyzowanym metodom projektowania, układy scalone mogą być opłacalne również wtedy, gdy potrzeba ich niewiele - kilkaset lub kilka tysięcy egzemplarzy, a nawet pojedyncze sztuki.

W przypadku układów produkowanych metodą płytek wieloprojektowych nieco odmiennie, niż przy zwykłej produkcji seryjnej, wygląda testowanie. Płytki wieloprojektowe nie są testowane ostrzowo. Testowane są dopiero gotowe, zmontowane w obudowach układy. Oznacza to, że produkcja płytek wieloprojektowych ma sens tylko wtedy, gdy uzysk produkcyjny u_p jest bliski 1. Gdyby tak nie było, do montażu trafiałoby dużo niesprawnych układów.

4.2.8 Modele biznesowe mikroelektroniki

We współczesnej mikroelektronice wyróżnić można dwa modele biznesowe:
- model „projektowanie i produkcja pod jednym dachem”, gdy producent sam projektuje produkowane układy (Integrated Device Manufacturer, w skrócie IDM),

- model, w którym projekt powstaje w firmie nie posiadającej linii produkcyjnych („*fabless*"), a produkcja jest zlecana producentowi wyspecjalizowanemu w wytwarzaniu układów na zamówienie („*silicon foundry*", dosłownie „odlewnia krzemu").

Pierwszy model jest charakterystyczny dla producentów układów standardowych. Obecnie na świecie pozostało niewielu takich producentów, najbardziej znanym jest Intel. Drugi model jest związany z układami specjalizowanymi (ASIC), których projekty powstają w firmach produkujących sprzęt finalny lub są zlecane do wykonania wyspecjalizowanym firmom projektowym. Od pewnego czasu ten model został zaadaptowany również przez niektóre firmy dostarczające na rynek układy standardowe. Firmy te zajmują się tylko projektowaniem, zaś produkcję zlecają producentom typu *silicon foundry*. Wyprodukowane układy są dostarczane jako produkt firmy, która je zaprojektowała, a odbiorca zazwyczaj nie wie, kto był faktycznym producentem. Producenci działający wg modelu IDM optymalizują swoje procesy technologiczne i organizację produkcji pod kątem wielkoseryjnej produkcji własnych wyrobów. Producenci typu *silicon foundry* muszą wykazać znacznie większą elastyczność, aby móc obsługiwać klientów zlecających różne rodzaje układów i zapewnić ekonomicznie uzasadnioną produkcję również w przypadku wytwarzania prototypów i niewielkich serii produkcyjnych. Produkcja prototypowa i małoseryjna jest realizowana przy wykorzystaniu płytek wieloprojektowych. Model *fabless/silicon foundry* umożliwia dostęp do projektowania i produkcji układów specjalizowanych także firmom małym i średnim, o niewielkim budżecie. Zatem ten model znakomicie wspiera innowacyjność. Większość firm typu *silicon foundry* nie potrzebuje najbardziej zaawansowanych technologii i nimi nie dysponuje, bo z technicznego punktu widzenia większość układów specjalizowanych ich nie wymaga. Są i tacy producenci, którzy równocześnie produkują własne układy i oferują usługi typu *silicon foundry*, jednak w tym przypadku usługi te są oferowane firmom zamawiającym układy produkowane w wielkich seriach (tu przykładem może służyć Samsung, który produkuje własne układy procesorów, a równocześnie w tej samej technologii produkuje procesory dla firmy Apple).

Producenci układów na ogół nie chcą się zajmować kontaktami z setkami drobnych klientów zamawiających niewielkie ilości układów specjalizowanych. Dlatego powstały instytucje pośredniczące. Mają one zawarte umowy z producentami, zbierają projekty od klientów, sprawdzają czy projekty te są prawidłowe (w sensie zgodności z wymaganiami producenta), łączą zebrane projekty ze sobą i wysyłają do producenta. Producent wykonuje maski wieloprojektowe, wytwarza płytki i odsyła. Instytucja pośrednicząca organizuje montaż: wysyła płytki do firmy, która tnie płytki na poszczególne struktury i montuje w obudowach. Po kilku miesiącach od wysłania projektu klient otrzymuje zamówione układy. Zamawia się zazwyczaj najpierw prototyp układu (10 - 50 egzemplarzy). Gdy prototypowe układy spełniają wymagania, zamawia się potrzebną liczbę układów. Jeśli jest ona niewielka, zostanie wyprodukowana w technice płytek wieloprojektowych. Przy dużej serii (zwykle od kilkudziesięciu tysięcy układów wzwyż) producent może uznać, że celowe jest wykonanie indywidualnego kompletu masek dla zamówionego układu i przeznaczenie na ten układ całych płytek lub nawet całych partii produkcyjnych.

W Europie najważniejszą instytucją pośredniczącą między klientami zamawiającymi układy specjalizowane, a ich producentami jest EUROPRACTICE (http://www.europractice-ic.com).

4.2.9 Mikroelektronika w polskich warunkach

W Polsce pierwsze próby wytwarzania diod i bipolarnych tranzystorów germanowych podejmowane były już w latach 50 XX wieku. Na początku lat 60 powstała w Warszawie fabryka TEWA, produkująca aż do połowy lat 70 różne rodzaje tranzystorów germanowych. Na początku lat 70 powołane zostało Naukowo-Produkcyjne Centrum Półprzewodników CEMI, w skład którego weszła TEWA, Instytut Technologii Elektronowej i kilka innych zakładów produkcyjnych. Licencje i linie produkcyjne zakupione we Francji i Japonii pozwoliły uruchomić

produkcję układów scalonych. Przez jakiś czas CEMI było jednym z większych producentów układów scalonych do sprzętu powszechnego użytku w krajach „bloku wschodniego". Słabością CEMI było to, że produkowano tam wyłącznie układy licencyjne lub samodzielnie kopiowane układy katalogowe różnych zagranicznych producentów. W dodatku kryzys lat 80 XX wieku spowodował wstrzymanie wszelkich inwestycji. W rezultacie na początku lat 90 CEMI mogło oferować wyłącznie układy będące odpowiednikami układów produkowanych przez wielkie firmy światowe, a na dodatek produkowane na przestarzałym i wyeksploatowanym sprzęcie. Taka produkcja nie miała żadnych szans konkurencyjności na wolnym rynku, a priorytety ówczesnych władz nie uwzględniały potrzeby utrzymania w Polsce przemysłu półprzewodnikowego, toteż CEMI uległo likwidacji. Przetrwał tylko Instytut Technologii Elektronowej, i wykorzystując część urządzeń produkcyjnych ze zlikwidowanych zakładów CEMI utworzył istniejący do dziś zakład doświadczalny w Piasecznie pod Warszawą. Zakład ten, nieco rozbudowany, nie jest jednak fabryką w komercyjnym sensie tego słowa, zajmuje się badaniami nad technologiami krzemowymi i wytwarzaniem małych ilości różnych nietypowych wyrobów (w tym mikrosystemów krzemowych). Jego niewielka załoga stara się podtrzymywać polskie kompetencje w zakresie technologii krzemowych. Już w nowym tysiącleciu zbudowany został w Warszawie ośrodek badawczy CEZAMAT, pomyślany jako miejsce zaawansowanych badań łączących technologie półprzewodnikowe z biotechnologiami, technologiami mikrosystemów różnych rodzajów itp. Są też w Polsce mniejsze ośrodki, powiązane z wyższymi uczelniami, zajmujące się pracami technologicznymi (Łódź, Wrocław, Rzeszów). Komercyjnej produkcji układów scalonych jednak nie ma.

Nie oznacza to jednak, że nie ma w Polsce mikroelektroniki. Projektowanie układów scalonych jest dziedziną badań i nauczania na wielu polskich uczelniach (AGH w Krakowie, politechniki w Łodzi, Gdańsku, Poznaniu, Warszawie). Laboratoria projektowania wyposażone są w najnowocześniejszy sprzęt komputerowy i oprogramowanie do projektowania, a poziom nauczania dorównuje światowym standardom. A ponieważ polski inżynier elektronik ma równie dobry dostęp do firm typu „*silicon foundry*", jak inżynierowie z innych krajów uprzemysłowionych (o czym będzie mowa dalej), każdego roku powstają nie tylko projekty, ale też zamawiane i badane są prototypy bardzo zaawansowanych układów do przeróżnych zastosowań, w tym układy projektowane dla najnowocześniejszych technologii. Dzięki temu powstało i funkcjonuje wiele firm typu „*fabless*", są wśród nich niewielkie firmy czysto polskie, ale także oddziały firm zagranicznych łącznie z centrami projektowymi światowych liderów. Firmy takie można znaleźć w Warszawie, na Wybrzeżu, w Krakowie, na Śląsku. Wykształceni w Polsce projektanci znajdują też pracę w zagranicznych centrach projektowania.

Zakończmy więc pierwszą część materiałów dotyczących układów scalonych takim stwierdzeniem: mimo iż nie ma w Polsce obecnie fabryk układów scalonych, mikroelektronika i jej zastosowania są dla polskiego inżyniera równie dostępne, jak dla jego zagranicznych kolegów.