

Flash to nie zawsze znaczy to samo...



Nie ma chyba przesady w stwierdzeniu, że pamięć typu Flash jest „wyczekana“ przez elektroników. Stosunkowo szybki zapis danych, na ogół wystarczający czas odczytu, a przede wszystkim nieulotność, to cechy które decydują o jej atrakcyjności, która pomimo pojawienia się ferroelektrycznych pamięci nieulotnych, ciągle nie maleje. Korzystamy z nich na co dzień, ale czy wiemy jak działają?

...czyli co we Flashach piszczy?

Wydawać by się mogło, że wynalezienie czegoś, czego świat jeszcze na oczy nie widział, wymaga po prostu niestandardowego myślenia. Wiele wynalazków powstało właśnie dzięki temu, że ktoś nie sugerując się aktualną wiedzą, po prostu pomyślał odmiennie. Tymczasem, jeśli przyjrzymy się obudowie najnowocześniejszych pamięci półprzewodnikowych, to okaże się, że mają one wspólną cechę z pamięciami magnetycznymi, które „rodziły się“ razem z pierwszymi komputerami, gdzieś na początku lat 50. ubiegłego stulecia. Tą cechą jest matrycowa budowa zastosowana w obu rozwiązaniach. W pamięciach magnetycznych informacja była pamiętana dzięki koralikom z materiału ferromagnetycznego, nawlekanych na specjalną siatkę wykonaną z cienkich drucików. Stąd nazwa takich pamięci - pamięci ferrytowe (rys. 1). Krawiec to określenie, które najbardziej pasowałoby do osoby zatrudnionej przy produkcji takich elementów, jako że jej praca polegała na „szytciu“ matryc za pomocą najwykleszej igły. Można powiedzieć, że najważniejszym blokiem współczesnych pamięci Flash jest matryca, tyle że do jej ręcznego wykonania byłiby zdolni chyba tylko dwaj wielcy mis-



Rys. 1

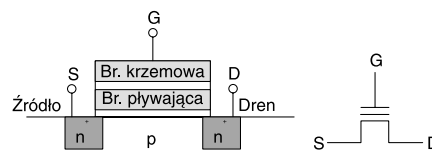
trzowie-konstruktorzy, jakim są Trurl i Klapaucjusz z lemowskiej Cyberiady.

Przyjrzyjmy się zatem, jak jest zbudowana współczesna pamięć Flash? Jej pojedynczą komórkę przedstawiono na rys. 2. Widzimy tu przekrój poprzeczny tranzystora MOS, nieznacznie zmodyfikowanego w stosunku do typowego elementu. Aby tranzystor taki był zdolny do przechowywania informacji, zastosowano w nim tzw. pływającą bramkę umieszczoną pomiędzy bramką sterującą a podłożem typu p. To właśnie w pływającej bramce gromadzony jest ładunek (elektrony), określający, czy zapamiętano wartość 0 czy 1. Odczytanie zawartości takiej komórki polega na przyłożeniu napięcia pomiędzy elektrody G i S, a następnie zmierzeniu prądu płynącego między elektrodami D i S. Ładunek zgromadzony w pływającej bramce będzie ekranował półprzewodnik pod bramką tak, że nie zostanie zaindukowany w tranzystorze kanał po podaniu napięcia odczytującego na elektrodę G. Tym samym ze źródła do drenu nie będzie mógł przepływać prąd IDS.

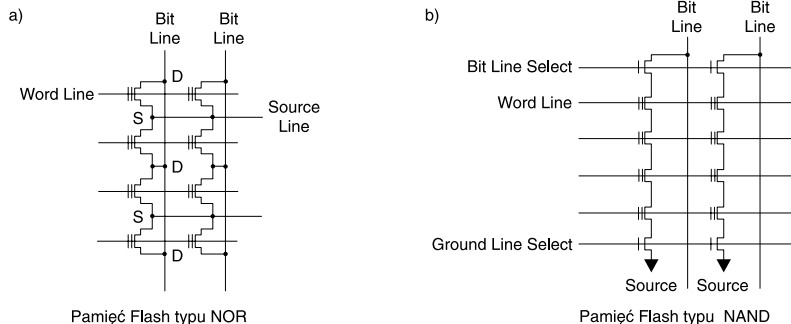
To, czego zazwyczaj nie wiemy: NOR czy NAND?

Kupując pamięci Flash, mało kto zadaje sobie pytanie, czy są wykonane w technice NOR, czy NAND. Wielu kupujących zapewne nawet nie wie, że są takie rodzaje pamięci. Tymczasem okazuje się, że każde z tych rozwiązań charakteryzuje się parametrami, które w pewnych zastosowaniach mogą być korzystne, w innych zaś mogą wręcz dyskwalifikować układ pamięci. O co więc chodzi?

Aby wyjaśnić różnice, musimy odwołać się do cechy, o której mówiliśmy na wstępie, czyli do matrycowej struktury pamięci. Pojedyncze tranzystory MOS, stanowiące elementarne komórki pamięci, łączy się w specjalną sieć. Dzięki temu, możliwe staje się sensowne zorganizowanie zapamiętywanych danych. Jak się jednak okazuje, matrycę można zbudować na dwa sposoby. Ilustruje to rys. 3. W pamięci typu NOR wiele tranzystorów tworzących kolumnę matrycy ma dołączone dreny do wspólnej linii bitu, natomiast źródła są odpowiednio dołączone do linii *Source Line* (rys. 3a). Pozostaje jeszcze połączenie elektrod G, które tworzą wyprowadzenie *Word Line*. Odczyt pamięci jest realizowany poprzez ustawienie wysokiego poziomego napięcia na linii *Word Line* oraz dołączenie odpowiedniej linii *Bit Line* do specjalnego wzmacniacza pomiarowego. Jeśli tak zaadresowana komórka jest zaprogramowana, wzmacniacz nie wykryje przepływu prądu przez tę linię. W praktyce odczytuje się 8, 16, 32 lub więcej bitów jednocześnie, zależnie od organizacji pamięci. Wpisanie zerowej wartości do komórki jest realizowane poprzez symultaniczne podanie impulsów napięciowych na linię *Bit Line* - impuls 5V, oraz *Word Line* - impuls 10 V. Powoduje to „wstrzelenie“ niezbędnej



Rys. 2



Rys. 3

dawki elektronów do obszaru pływającej bramki. Tak zgromadzony ładunek pozostaje później w pułapce (odizolowana bramka pływająca), nawet po odłączeniu wszelkich napięć od układu. Wewnętrzny sterownik pamięci dba o to, aby operacja zakończyła się sukcesem, kilkakrotnie powtarzając powyższą procedurę. Kasowanie, czyli wpisanie logicznej jedynki, jest realizowane poprzez podanie dodatniego napięcia (+5 V) na linię *Source Line* równocześnie z ujemnymi impulsami (-8 V) podawanymi na linię *Word Line*. Następuje wówczas uwolnienie elektronów z pułapki i ich odpływ z pływającej bramki. Ta operacja może być dla pewności również powtarzana kilkakrotnie.

W przypadku pamięci o strukturze NAND, matryca jest zbudowana jak to pokazano na rys. 3b. Odczyt następuje w wyniku połączenia linii *Source Line* do potencjału zerowego i podaniu na linię *Word Line* napięcia odczytu. Włączenie odpowiedniej grupy komórek jest realizowane poprzez wyprowadzenia *Bit Line Select*. Wyprowadzenia *Bit Line* są dołączone do wzmacniaczy pomiarowych, lecz o tym, które będą czytane, decyduje *Source Line*. Trzeba zauważyć, że przez cały czas pozostają w stanie włączenia nawet te tranzystory, które nie będą w danym cyklu odczytywane. Do detekcji prądu płynącego przez wszystkie linie należące do jednego słowa pamięci konieczne jest kilkakrotne przełączenie tranzystorów. Odbija się to niestety na czasie dostępu, jakim charakteryzują się układy tego rodzaju. Jest on typowo równy ok. 25 μ s. W celu zaprogramowania pojedynczej komórki linia *Word Line* jest dołączona do bardzo wysokiego napięcia (15 do 20 V), natomiast *Bit Line* znajduje się na potencjale zero. Niewybrane komórki przepuszczają prąd programujący bez żadnego wpływu na ich późniejszy stan.

Kasowanie sektora odbywa się po połączeniu z zerowym potencjałem wszystkich linii *Word Line* należących do tego sektora oraz doprowadzeniu wysokiego napięcia (15 do 20 V) do podłoża wszystkich kasowanych tranzystorów. Wstrzykiwanie elektronów do odpowiednich obszarów zarówno podczas kasowania, jak i programowania komórek pamięci typu NAND odbywa się na zasadzie tunelowania Fowlera-Nordheima (rys. 4). Konieczność stosowania wyższych napięć programujących w porównaniu z pamięciami NOR (rys. 5) powoduje, że w pamięciach NAND niezbędne staje się wbudowywanie w strukturę układu bardziej wydajnych pomp ładunkowych. Stanowi to dodatkową trudność dla konstruktorów, którzy muszą ten problem rozwiązywać, nikt bowiem obecnie nie zainteresuje się układami, które wymagałyby doprowadzania dodatkowych napięć podczas programowania lub kasowania. Wyższe napięcia stanowią również przeszkodę w realizacji wewnętrznych kluczy przełączających napięcia programujące. Powyższe prob-

lemy są wprawdzie do pokonania w technologii 0,25 μ m, jednak pojawiają się przy tym inne kłopoty, związane chociażby z nieuniknionym wzrostem zakłóceń. Zaoszczędzona powierzchnia struktury półprzewodnikowej, jaką udaje się uzyskać w matrycach typu NAND w porównaniu z NOR, jest niestety zajmowana przez większe powierzchniowo pompy ładunkowe i inne niezbędne bloki elektroniczne pamięci.

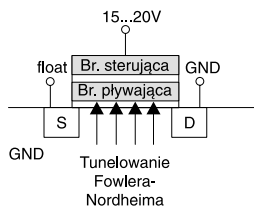
Wymiary i wynikające z nich koszty

Rozmiary komórek pamięci podawane są w umownych jednostkach, zależnych od zastosowanej technologii, np. 0,18 μ m lub 0,15 μ m. Jednostki te są oznaczane dużą literą F. Podstawowa komórka pamięci Flash wykonana w konfiguracji NOR zajmuje powierzchnię 10 F². Jej topografię przedstawiono na rys. 6a. Konfiguracja NAND jest bardziej oszczędna pod tym względem (rys. 6b). Podstawowa komórka zajmuje w tym przypadku powierzchnię jedynie 6 F², co stwarza możliwość - przynajmniej teoretycznie - uzyskania większej gęstości upakowania komórek. Na końcową wielkość powierzchni, jak już wspomniano wcześniej, wpływają dodatkowo pozostałe elementy pamięci. Trudno jest więc jednoznacznie odpowiedzieć, która konfiguracja będzie wymagała mniejszej powierzchni struktury półprzewodnikowej.

Zanim coś zrobisz, dwa razy się zastanów

Zapewne niestandardowe myślenie konstruktorów doprowadziło do tego, że wykonywane są również

Tab. 1. Zestawienie podstawowych właściwości pamięci Flash typu NOR i NAND			
Operacja	NOR	NAND	Komentarz
Odczyt swobodny	35...150 ns	25 μ s	NOR - prawdziwy dostęp swobodny NAND - wolny dostęp stronicowy
Odczyt sekwencyjny	20...150 ns	50 ns	NOR - prawdziwy dostęp swobodny NAND - odczyt strony z bufora
Czas programowania	5 μ s/bajt swobodny	0,5 μ s/bajt swobodny	NOR - prawdziwe programowanie swobodne NAND - programowanie stronicowane lub sekwencyjne stronicowane
Czas kasowania	0,7 s/128 kB	2 ms/16 kB	NOR - wstępne programowanie wszystkich komórek NAND - szybkie kasowanie
Bezбłędność	100%	98%	NOR - wszystkie adresy mają gwarantowany poprawny odczyt NAND - Błędne bloki niewykorzystywane do przechowywania danych
Wytrzymałość	100000 bez uszkodzenia komórek	250000 z błędami odczytu	NOR - 100% pewność odczytu każdej komórki w 100000 cykli NAND - spodziewane błędy odczytu niektórych komórek



Programowanie pamięci Flash typu NAND

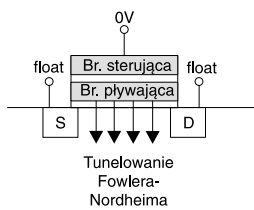
Rys. 4

układy potrafiące zapamiętać więcej niż jeden bit informacji w pojedynczej komórce. Przecież to niemożliwe! A jednak. Jak pamiętamy, stan logiczny pojedynczej komórki pamięci jest zdefiniowany poprzez ustalenie pewnej wartości ładunku zgromadzonego w pływającej bramce. W klasycznym wariantcie jest to połowa ładunku jaki może być wstrzyknięty do bramki. Jeśli więc ładunku jest mniej niż połowa, przyjmujemy, że komórka reprezentuje stan 1, w przeciwnym razie stan 0 (rys. 7a). Nietrudno zauważyć, że dokładniejszy pomiar prądu I_{DS} oraz bardziej precyzyjne kontrolowanie wstrzykiwanego ładunku pozwoliłoby przyjmować, że dana komórka może się znajdować w stanach np. 00, 01, 10 i 11 (rys. 7b).

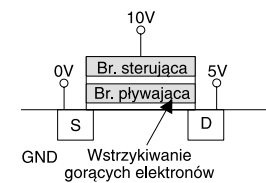
Powyższe rozważania nie są bynajmniej tylko hipotetyczne. Zgodnie z tym pomysłem wytwarza się pamięci Flash z tak zwanymi komórkami *Multi Bit*. Nasuwa się pytanie: która konfiguracja lepiej nadaje się do realizacji komórek *Multi Bit*? Okazuje się, że większą liczbę bitów przypadającą na jedną komórkę udaje się uzyskać w pamięciach NOR. Ponadto zasada pomiaru prądu I_{DS} stosowana w konfiguracji NAND praktycznie wyklucza uzyskanie nawet czterech bitów na komórkę. Wyjątkowo korzystny współczynnik \$/MB, jakim cechują się pamięci *Multi Bit*, powoduje, że wszystkie laboratoria pracują intensywnie nad unowocześnianiem tej technologii. Powinniśmy się więc spodziewać nowych typów pamięci oraz obniżania ich cen. Na rys. 8 przedstawione są zależności powierzchni równoważnych komórek w funkcji zastosowanej technologii i konfiguracji.

Różnice nie tylko w budowie

W tab. 1 zestawiono ważniejsze różnice związane z działaniem obu rodzajów pamięci. Jak widać, każdy typ ma swoje zalety i wady. Jedną z większych niedogodności wszystkich pamięci Flash jest niemożność



Kasowanie pamięci Flash typu NAND



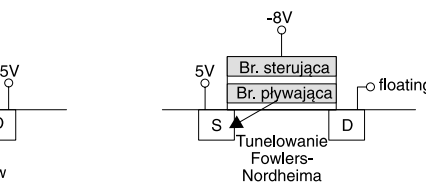
Programowanie pamięci Flash typu NOR

Rys. 5

skasowania pojedynczego bajtu. Operacja ta zawsze dotyczy całego bloku (lub sektora - nazewnictwo zależy od producenta). Jeśli zachodzi potrzeba wymazania fragmentu bloku, to niestety musi być on wcześniej przeniesiony w inne miejsce. Jeśli się tego nie zrobi, to dane, które powinny pozostać w pamięci, zostaną skasowane. Wynika z tego, że im mniejszy będzie blok, tym sprawniej będzie można kasować dane. W pamięciach NOR blok ma wielkość od 16 do 128 kB. W NAND-ach bloki mają pojemność od 8 do 16 kB.

Organizowanie danych w pamięciach Flash związane jest z techniką stronicowania. Koncepcje stronicowania, jak można się spodziewać, są odmienne dla pamięci z matrycami NOR i NAND. Pamięć NOR w trybie stronicowym grupuje do czterech adresów należących do jednej strony. Pierwszy dostęp do nich trwa znacznie dłużej w porównaniu z kolejnymi. W pamięciach NAND każdy sektor jest podzielony na strony zazwyczaj o pojemności 512 bajtów. Czas dostępu do danych w obszarze jednej strony podczas odczytu jest krótki, niestety jest to okupione długim czasem przeskoku do innych stron.

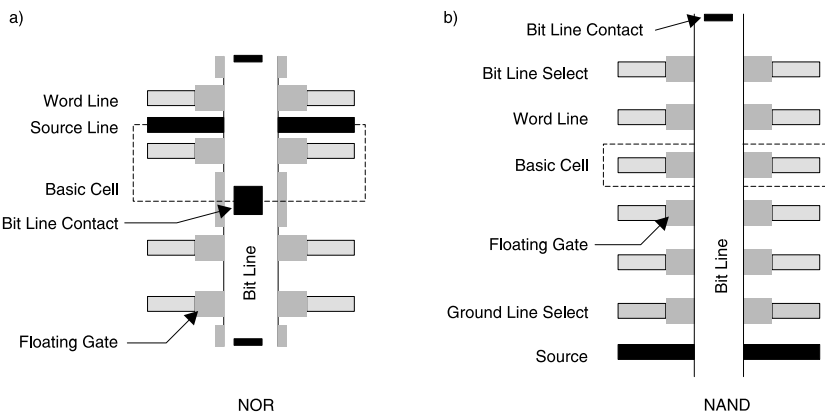
Stosunkowo duża łatwość kasowania i programowania Flashy powoduje, że są one chętnie wykorzystywane jako pamięci programu w sys-



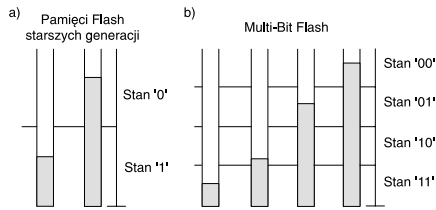
Kasowanie pamięci Flash typu NOR

temach mikroprocesorowych. Warunkiem koniecznym do takich zastosowań jest zapewnienie całkowite swobodnego dostępu do danych podczas operacji odczytu, bez względu na ich umieszczenie w pamięci. Trudno przecież sobie wyobrazić, że procesor, np. po pobraniu rozkazu skoku, będzie czekał aż pamięć przełączy się na pobranie danej z odległego adresu. Warunek ten spełniają na pewno pamięci NOR. Można w nich uzyskać czas dostępu nawet poniżej 35 ns. W pamięciach pracujących w trybie stronicowym czas dostępu do pierwszej strony wynosi ok. 150 ns, do następnych zaś 20 ns. A jak jest z pamięciami NAND? Zapewniają one wprawdzie dość szybki dostęp do danych zawartych w jednej stronie - ok. 50 ns, lecz przełączenie strony trwa aż 25 μs.

Różnice pomiędzy omawianymi rodzajami pamięci ujawniają się również w przypadku operacji programowania i kasowania. NOR Flash daje się programować w zupełnie dowolnej kolejności, potrzebując do tego ok. 10 μs na bajt lub słowo. W przypadku komórek typu *Multi Bit* czas ten wydłuża się, lecz w takich pamięciach stosuje się specjalny bufor zezwalający na równoległe programowanie wielu bajtów lub słów. Dzięki temu zabiegowi uzyskuje się czas programowania równy ok. 10 μs na bajt lub słowo, który -



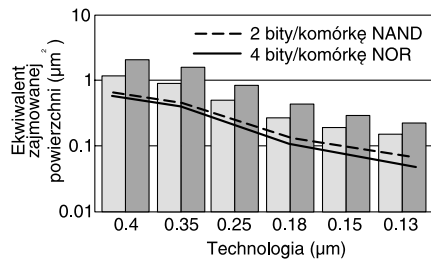
Rys. 6



Rys. 7

jak się przewiduje - już wkrótce powinien zmaleć do wartości poniżej 2 μ s. Pamięci NAND mogą być programowane tylko całymi stronami, zapis jednego bajtu nie jest możliwy. Programowanie strony trwa ok. 200 μ s, co przy typowej pojemności strony równej 512 bajtów oraz uwzględnieniu niezbędnych operacji pośrednich daje ok. 0,5 μ s na bajt. W przypadku komórek *Multi Bit* programowanie strony zajmuje 900 μ s, a bajtu 2 μ s.

Jak pamiętamy, blok danych w pamięci NOR ma rozmiar od 16 do 128 kB. Do wykasowania typowego bloku 64 kB potrzeba aż 0,7 s. Wynika to z faktu, że przed wykasowaniem każdej komórki bloku musi być ona wcześniej zapisana wartością 0. Operacja ta jest niezbędna w celu zagwarantowania jednakowego ładunku we wszystkich pływających bramkach przed rozpoczęciem kasowania. W nowoczesnych pamięciach stosuje się odpowiednie we-



Rys. 8

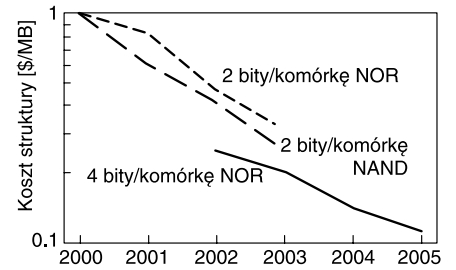
wewnętrzne kontrolery programowania i kasowania odciażające mikroprocesor. Pamięci NAND posiadają blok o rozmiarze 16 kB. Jego kasowanie trwa ok. 2 ms. W tym przypadku nie trzeba wstępnie wyrównywać ładunku w pływających bramkach.

Konfiguracja NOR gwarantuje, że 100% komórek pamięci wytrzyma 100000 cykli programowania/kasowania. Nie oznacza to oczywiście, że 100001-sze kasowanie/programowanie zniszczy nam całkowicie pamięć. Od tego momentu musimy być jednak przygotowani na zwiększone prawdopodobieństwo pojawiania się błędnych odczytów.

Znacznie gorzej jest z pamięciami NAND. Te z założenia gwarantują wykorzystywanie typowo 98% komórek pamięci. Co gorsza, pojawienie się błędnych bloków może się ujawnić dopiero podczas normalnej pracy układu. Konieczne staje się więc markowanie zepsutych bloków jako „bad” i późniejsze ich pomijanie. Jak na ironię, producenci gwarantują dłuższą żywotność tych pamięci w porównaniu z NOR-ami, oceniając ją na 250000 cykli zapisu/kasowania.

Trendy

Jak ziliustrowano to na rys. 9, w przypadku pamięci NAND i NOR, w których zastosowano rozwiązanie *Multi Bit* - 2 bity na komórkę, przeliczniki \$/MB są do siebie zbliżone z nieznaczną przewagą NAND. Wyraźnie konkurencyjne mogą być 4-bitowe pamięci NOR. Dla zwykłych użytkowników pocieszający jest na pewno silnie malejący przebieg prezentowanych funkcji. Wykres ten nie oddaje jednak rzeczywistych kosztów ponoszonych podczas budowy urządzeń wykorzystujących



Rys. 9

pamięci Flash. Wiele aplikacji wykorzystujących pamięci NAND wymaga stosowania w nich dodatkowych kontrolerów, podczas gdy NOR-y można bezpośrednio dołączać do większości mikroprocesorów. Położenie nacisku na rozwój 4-bitowych pamięci NOR spowoduje, że staną się one układami nadającymi się do większości zastosowań. Przyszłość należy więc raczej do pamięci NOR, chyba że pojawi się jakaś kolejna odmiana. Wszystko wskazuje na to, że stanie się to faktem. Znany producent - STMicroelectronics opracował i wprowadza do produkcji pamięci tzw. *LightFlash*, o których z pewnością będziemy jeszcze pisać.

Podsumowanie

Wierzę, że doczekam czasów, w których będą dostępne odmiany pamięci nieulotnych (może będzie to jakiś zupełnie nowy typ) będące dokładnymi odpowiednikami statycznych pamięci RAM. Warto w tym miejscu przypomnieć, że właściwie układy takie już istnieją, choć nie są bardzo powszechne. Przykładem mogą być pamięci RAM z podtrzymaniem danych za pomocą umieszczonej w obudowie pamięci baterii litowej (opracowane przez dawnego Dallasa). Interesujące są również pamięci FRAM, będące bardzo pomysłowym połączeniem pamięci SRAM i Flash, oferowane przez firmę Ramtron.

Jarosław Doliński, AVT
 jaroslaw.dolinski@ep.com.pl

Dodatkowe informacje

Artykuł powstał w oparciu o materiały firmy STMicroelectronics.

Więcej informacji można znaleźć na stronach:
<http://developer.intel.com/design/flcomp/isfbgmd.htm>
<http://www.howstuffworks.com/flash-memory1.htm>
<http://www.commsdesign.com/main/2000/05/0005feat2.htm>
http://www.st.com/stonline/prodpres/memory/flash/fl_light.htm