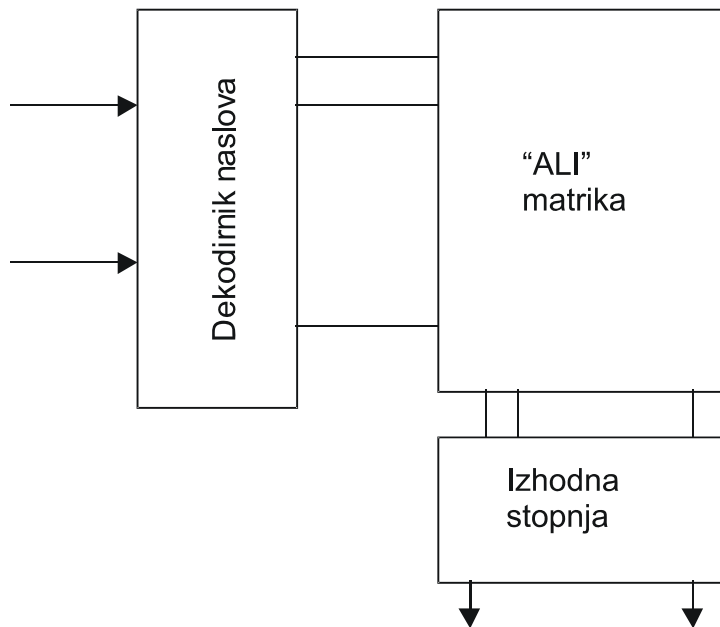
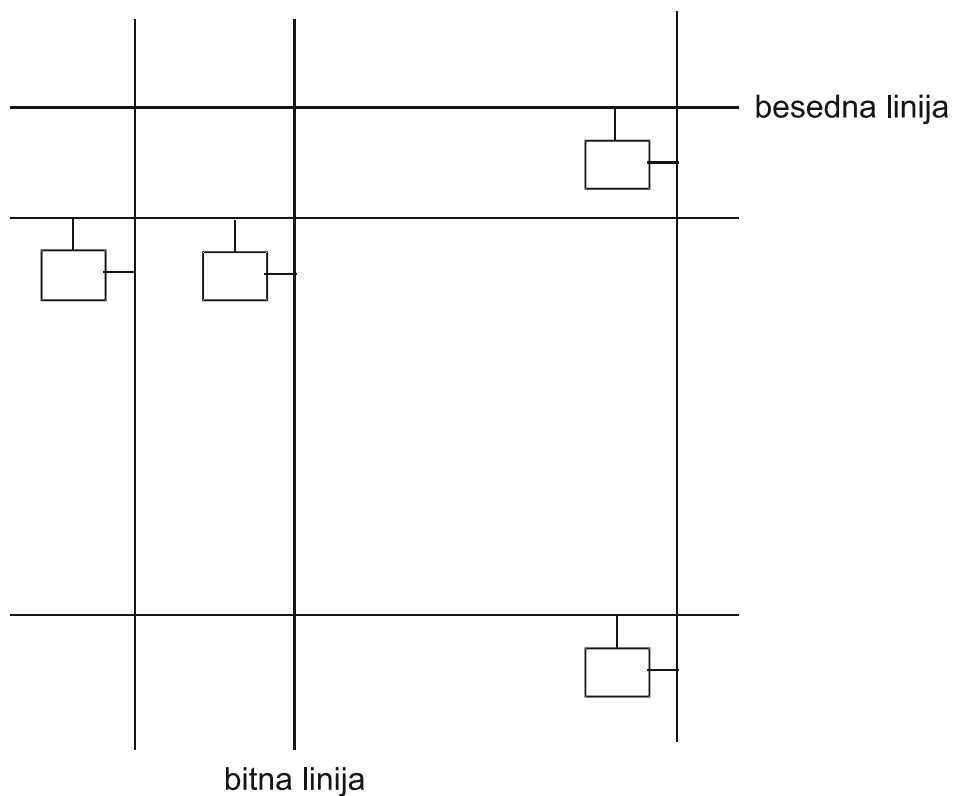


Bralni pomnilniki

Bralni pomnilnik lahko predstavimo kot strukturo treh sklopov, ki jo prikazuje slika.



Najbolj nas zanima nas sklop, imenovan matrika "ALI". V njem ločimo besedne linije, bitne linije in vezne elemente.



Vezni element (VE) predstavlja pomnilno celico bralnega pomnilnika. Vezni element v matriki je lahko zelo različen in določa tip bralnega pomnilnika. Zadoščati mora nekaj zahtevam, ki so razvidne iz naslednje slike. Idealno bi bilo, da bi vezni elementi izkazovali različno upornost, odvisno od smeri prevajanja toka. Če so vezni elementi prisotni v vseh sečiščih besednih in bitnih linij, naj neaktivni (ali aktivni) vezni elementi izkazujejo zelo visoko upornost. Obratno naj aktivni elementi izkazujejo čim manjšo upornost in kapacitivnost.

Veznemu elementu je dodana serijsko vezana "varovalka", ki jo

- ob programiranju raztalimo, ali pa
- v primeru MOS transistorja posebna topologija vezana na gate elektrodo.

Varovalka predstavlja logično stanje v celici.

Zakaj VE ne more biti kar ohmska upornost? Besedna linija bi aktivirala bitno linijo, ta pa ostale besedne linije. Zato mora VE prevajati samo v eno smer. Vezni element matrike ALI je torej lahko dioda ali transistor (bipolarni ali MOS).

Na začetku so se uporabljale diode. Varovalka pri diodi pomeni "1". Bitne linije so na nizki napetosti preko "pull-down" uporov. Izbrana besedna linija dvigne bitno linijo preko diode na "1". Zato so v začetnem stanju same enice. Pri diodah bitna linija ne more aktivirati drugih besednih linij zaradi zapore.

Začetnim poskusom z diodami so hitro sledile realizacije s tranzistorji. Bitna linija je vezana preko "pull-up" transistorja na napajanje, zato je na njej "1". Z naslovom izberemo eno besedno linijo. Napetost na njej povzroči, da začne transistor prevajati in s tem povzroči določen upad napetosti na b-liniji (ne prav velik zaradi velike kapacitivnosti bitne linije). Če transistorja v celici ni, na bitni liniji ne pride do upada napetosti.

Kot vidimo, potrebujemo za pomnenje enega bita le en sam tranzistor in element, ki ga aktivira. V primeru standardnih ROM vezij ne potrebujemo niti tega elementa, saj z masko zagotovimo povezavo besedne linije z veznimi elementi (mask ROM).

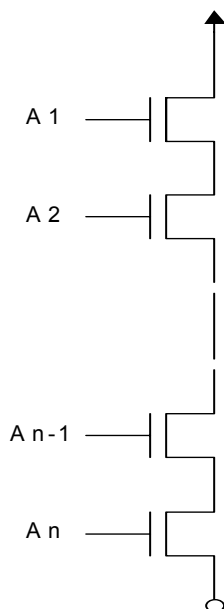
Bralne pomnilnike delimo v več podskupin. Najbolj enostavna je izvedba bralnih pomnilnikov, kjer vsebino - podatke določimo že z masko v procesu proizvodnje (t.i. mask ROM). Vendar tudi te najbolj enostavne izvedbe niso enolične. V literaturi, ki obravnava bralne pomnilnike kot strukture, izhajamo vedno iz zelo poenostavljenega modela trobločne strukture, ki smo ga uporabili tudi že mi. Dejanska izvedba pa se seveda loči od takega idealiziranega modela in je tudi močno odvisna od uporabljene tehnologije (unipolarna, bipolarna).

Pri mask ROM-u proizvajalec naredi masko glede na kupčevo specifikacijo vsebine. Ker to ni poceni, je ta tip pomnilnika smiseln samo za zelo velike serije.

Prvi bralni pomnilniki, pri katerih je imel uporabnik možnost določanja vsebine, so bili PROM (programmable ROM). Tipično so bili bipolarni; to pomeni, da je bil VE bipolarni transistor. K transistorju je bila zaporedno vezana varovalka, ki jo je uporabnik prežgal z dovolj velikim tokom, tako da je priključil dovolj veliko napetost. Seveda je bilo možno le enkratno programiranje (OTP – one time programmable). Takih izvedb se danes več ne izdeluje. Danes so PROM-i (tj. OTP) večinoma EPROM-i v cenejšem ohišju brez kvarčnega okna.

Dekodirnik naslova

Z naraščanjem števila naslovnih bitov (linij) eksponentno narašča število besednih linij. Že pri bralnem pomnilniku 1k x 8 bi potrebovali 1024 »IN« operatorjev z 10 vhodi. Tudi manj zahteven bralni pomnilnik 256 x 4 bi zahteval 256 »IN« operatorjev z 8 vhodi. V MOS tehnologiji je možno enostavno realizirati »IN« (pa tudi »ALI«) operatorje kot zaporedno (oz. vzporedno) vezavo tranzistorjev:

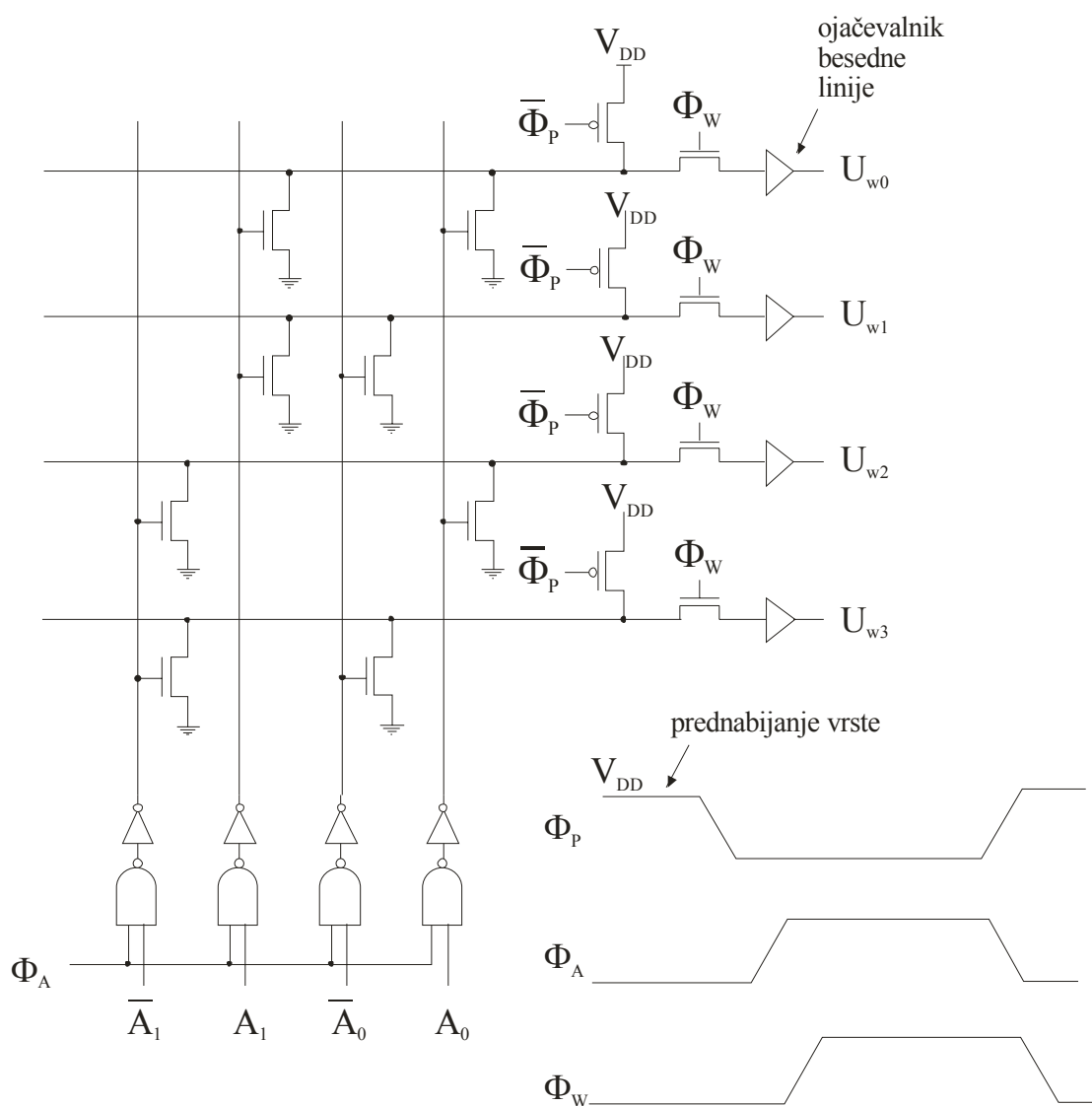


Pogoj je le, da število vhodov ni preveliko. Veliko matriko zato razrežemo (lahko tudi večkrat) na dele. Vsebino preberemo iz vseh delov in jo zatem z izbiro (multipleksor) na osnovi preostalih naslovnih bitov vodimo na izhod. Tako imamo v 70. letih standardne bralne pomnilnike 1k (256 x 4 ali 128 x 8), 2k (256 x 8 ali 512 x 4) in 4k (512 x 8 ali 1k x 4). Shematski prikaz izvedbe 256 x 4 prikazuje slika:

X-dekodirnik je 12/4K, Y-dekodirnik pa 5/32.

Osnovna ideja je v tem, da se celotna matrika razdeli v bloke s skupnimi dekodirniki naslovov in izbiralniki, ki izbirajo prav tako na osnovi dela naslova.

Če pogledamo dekodirnik natančneje, vidimo, da je sinhroniziran s posebnimi urinimi signali. Razlog je v tehniki t.i. prednabijanja (precharging), ki se uporablja zaradi manjše porabe moči.



Bralni pomnilniki, pri katerih je stanje stikal (MOS tranzistorjev; navadno NMOS zaradi doseganja višje hitrosti ali pa celo bipolarni tranzistor) določeno s kovinsko masko – povezavami (mask ROM), se množično uporabljajo kot kodirniki in prekodirniki. Tipičen primer so: prekodirniki iz BCD v sedem segmentno kodo, prekodirniki iz podmnožice ASCII v kodo za matrične tiskalnike pa tudi programski pomnilniki računalnikov, ki se izdelujejo masovno za krmiljenje določenega procesa.

Danes prevladuje uporaba bralnih pomnilnikov, pri katerih je možen ponoven vpis vsebine. Naslednja slika prikazuje dve taki stikali.

EPROM

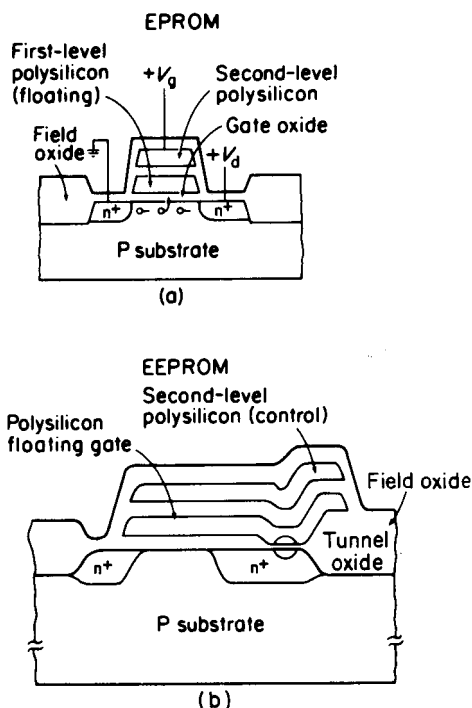
Stikalo na sliki (a) se imenuje EPROM (pogosto tudi UVRAM ali UV-EPROM). E pomeni izbrisljiv ("Erasable"). EPROM ima plavajočo plast (floating gate, plavajoča vrata) iz polisilicija (polysilicon), tj. dodatno krmilno elektrodo, ki je na vse strani izolirana z dielektrikom (oksidom, debelim med 50 in 100 nm). Struktura izkazuje efekt kondenzatorja. Zato EPROM transistor pogosto narišemo z dodatno črto.

Vpis vsebine (programiranje) poteka tako, da se na krmilno (gate) elektrodo in ponor dovede dovolj visoka pozitivna napetost (navadno 12.5V ali 21V) proti substratu in izvoru (source), kar povzroči preboj dielektrika - plazoviti (ali vroči) preboj elektronov ("avalanche injection", "hot electron injection") iz substrata proti drain in gate. Elektroni morajo dobiti dovolj energije, da preskočijo 3.2eV energijsko bariero med substratom in SiO₂. Majhen neg. naboj (10⁶ elektronov) pa ostane v plavajoči plasti tudi, ko ni več napetosti. Nima kam odteči, ker je naokrog močan izolator (SiO₂). "Kondenzator" se ne prazni, ker so elektroni ujeti v potencialno luknjo (časovna konstanta nekaj sto let). Negativen naboj na plavajoči plasti onemogoča, da bi transistor prevajal, ko je izbran. Naboj v bistvu znatno poveča pragovno napetost MOS transistorja. EPROM programator priskrbi določene napetosti, ki jih konkreten EPROM potrebuje. Programiranje npr. 512K*8 traja kako minuto. Proces programiranja celice je samo-omejujoč, tj. ni bojazni, da bi celico "preveč" sprogramirali. Elektroni v plavajoči plasti, ki jih je vedno več, vedno bolj omejujejo prihod novih elektronov iz substrata. Vpis je možen potem, ko vsem stikalom zberemo vsebino.

Brisanje: EPROM je treba vzeti iz vezja in ga s posebno napravo (EPROM eraser) osvetljevati z UV svetlobo določene valovne dolžine (253.7nm) 10-30 minut, da se vsebina izbriše (fototok). Ujeti elektroni dobijo dovolj energije (3.2eV), da preskočijo iz plavajoče plasti proti gate in substratu. Ob obsevanju z UV svetlobo naboj, ujet v tej plasti, izgine. Ker ta frekvenca ne prodre skozi večino vrst plastike ali stekla, ima EPROM čip na vrhu kvarčno okence. Čip mora biti pri brisanju dovolj blizu UV-izvoru (nekaj cm). Treba je izbrisati celoten čip (ne gre po delih) in ga nato sprogramirati ter vložiti nazaj v vezje. EPROM pa lahko "preveč" zberemo; pri tem postane plavajoča plast tako nabita, da sploh ne more več sprejeti elektronov. Na sončni svetlobi se izbriše v kakem tednu (z neprozorno nalepko čez okence to lahko preprečimo). Max. število brisanj in pisanj proizvajalci neradi deklarirajo, je pa vsaj 100.

Pri izbrisani celici transistor prevaja. Ko je izbran, se bitna linija (ki je običajno prednabita) prazni skozenj in s tem ji pade napetost. Bitna linija je preko stikalnega transistorja s tipalno linijo povezana s tipalnim ojačevalnikom ("sense amplifier"), ki primerja to napetost z referenčno napetostjo. Če je manjša od referenčne, to običajno tolmači kot logično 1, in obratno. Logična 1 je torej, kadar transistor prevaja. Zato ima običajno prazen EPROM same "1".

Hiba EPROM celice je potreba po višji napetosti za programiranje, brisanje celotne vsebine ter dejstvo, da je treba čip vzeti iz vezja.



Te hibe odpravlja EEPROM (b), ki je bil razvito kasneje. Med plavajočo elektrodo in drain je področje zelo tankega oksida (5 do 10nm).

Z uporabo ustrezne polaritete napetosti (brišemo z napetostjo nasprotne polaritete kot pri programiranju) med drain in gate je možno v in iz plavajoče plasti premakniti elektrone na osnovi tunelskega efekta (hladen preboj elektronov oz. Fowler-Nordheim tuneliranje), pri katerem elektroni prebijejo energijsko bariero Si-SiO₂ z manj kot 3.2eV. Potrebna napetost za programiranje je zato manjša kot pri EPROMu.

Pri pozitivni napetosti med ponorom in vrati tunelirajo elektroni skozi tanko plast oksida iz plavajoče plasti v ponor (programiranje). S tem se zmanjša pragovna napetost in transistor prevaja (kar pri EEPROM pomeni logično 0). Pri brisanju je napetost nasprotne polaritete in elektroni tunelirajo iz ponora v plavajočo plast. Pragovna nap. se poveča, transistor ne prevaja (logična 1). Vpis vsebine (odstranitev naboja iz plavajoče plasti) je možen le v celico, ki je bila prej izbrisana (naboj v plavajoči plasti).

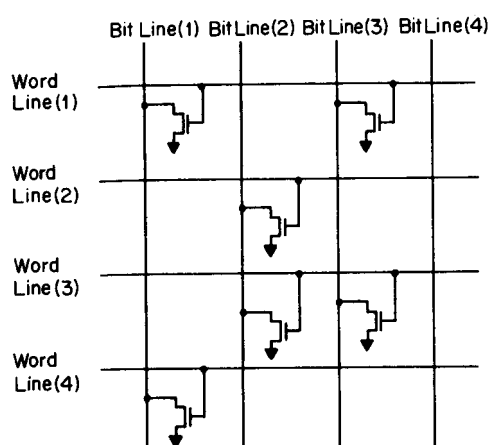
Brisati in reprogramirati ga je možno le po 1-byte naenkrat, vendar v celoti. Zaradi možnosti brisanja in programiranja po bytih potrebuje vsaka celica poleg pomnilnega še t.i. izbiralni transistor oz. stikalo. EEPROM celico sestavljata 2 transistorja: izbiralni in pomnilni. Zato je površina EEPROM celice ustrezno večja kot površina EPROM celice.

EEPROM-a ni treba vzeti iz vezja, če ga hočemo reprogramirati. Ne rabimo niti dodatne opreme; briše se z lokalizirano uporabo električnega polja za vsako celico.

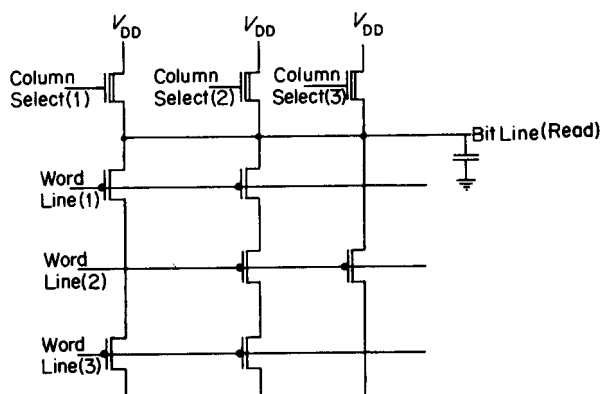
EEPROM ni zamenjava za RAM, kajti:

- pisanje je počasno $\sim 30 \mu\text{s}/\text{byte}$
- št. vpisov je omejeno ~ 10000 , ker je plast dielektrika zelo tanka in se obrabi.

Po določenem številu brisanj in vpisov se ujame vedno več elektronov v plasti oksida, kar začne zmanjševati razliko med stanji stikala in končno prepreči nadaljnjo uporabo.



(a)

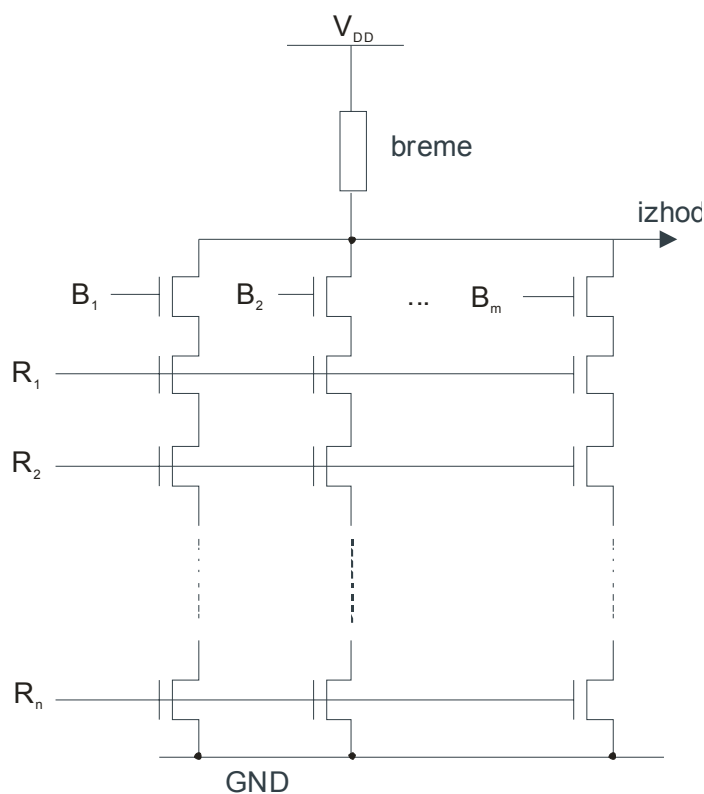


V večini obravnav se bralni pomnilnik prikaže kot struktura s tremi gradniki (prva slika). Matrika je običajno tipa "ALI" (vzporedna vezava nMOS transistorjev v resnici predstavlja funkcijo NOR). Transistorji so vezani vzporedno na bitno linijo – vsak s ponorom (to se da narisati tudi drugače kot na sliki a). Vendar imamo vsaj pri bralnih pomnilnikih, kjer sloni programiranje na maski (mask ROM), na voljo dve izvedbi te matrike. S programiranjem z masko razumemo, da ustvarjamo povezave v času proizvodnje integriranih

vezij. Najbolj enostavno je programiranje matrike "ALI", ki jo prikazuje prejšnja slika (a). V zadnjem procesnem koraku z masko realiziramo povezavo med besedno linijo in gate elektrodo. Vsa stikala morajo imeti že v predhodnih procesnih korakih realizirano povezavo z bitno linijo, kar zavzema prostor na integriranem vezju.

Drugačna je izvedba na sliki (b) (črke b sicer ni videti). Z izbiro besedne linije razklenemo vsa stikala, vezana na to besedno linijo, ki so normalno sklenjena. S krmiljenjem stikala na bitno linijo (na sliki narisano kot povezava), se signal prenese na bitno linijo. Proces izdelave takega bralnega pomnilnika zahteva, da se v verigo povežejo tranzistorji v zgodnjih procesih proizvodnje integriranega vezja, ne zahteva povezave na bitno linijo in zato zahteva seveda bistveno manj kontaktov na tranzistor. Daljša kot je veriga, večji je prihranek na prostoru, toda tudi zakasnitev branja, ki eksponencialno narašča s številom tranzistorjev v verigi. Naraščajoča odvisnost časa branja in padajoča vrednost števila bitov na enoto površine dasta optimum pri določenem številu zaporedno vezanih tranzistorjev. Ta številka je zavidljivo visoka (približno 64 tranzistorjev!).

Naslednji korak je bila združitev serijske in paralelne izvedbe v enem vezju, ki združuje prednosti obeh realizacij. Prikazuje jo slika:



Krajše vzporedne veje omogočajo hitrejši dostop po eni strani in še vedno ohranjajo manjše potrebe po površini silicija.

Posebno vejo bralnih pomnilnikov predstavljajo pomnilniki, ki v celici hranijo več bitov. Da lahko hranijo dva bita, potrebujejo štiri stanja. To so dosegli na

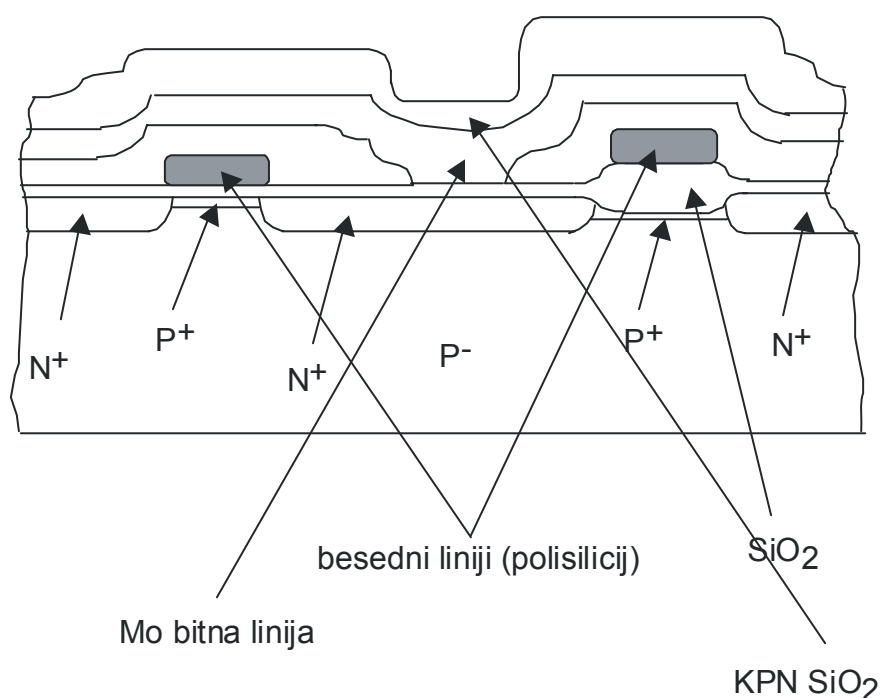
dva načina. Prvi je uporaba gate elektrod s spremenljivim razmerjem širina/dolžina. Tako dosežemo lahko štiri nivoje toka skozi MOS tranzistor (odvisno od tega katera gate elektroda je maskirana). Posebni diferencialni tokovni ojačevalniki lahko na osnovi primerjave z referenčnimi tokovi določijo vrednost vpisanih bitov. Drugi sloni na uporabi različnih ionskih implantov in s tem na različni napetosti praga MOS tranzistorja. Linearno naraščajoča napetost na besedni liniji aktivira najprej celice s tranzistorji, ki imajo najnižji prag in nato postopoma tranzistorje z višjim pragom. Z meritvijo časa, v katerem se aktivira tranzistor v celici, lahko ugotovimo vrednost vpisanih bitov.

Tipi programiranja celic

Pri bralnih pomnilnikih tipa mask ROM je zelo pomemben čas med določitvijo vsebine in izvedbo vezja. Nanj najbolj vpliva dejstvo, v katerem procesnem koraku se izvrši programiranje. Poznamo tri načine programiranja: z debelino oksida (FOX sloj), modifikacijo napetosti praga z implantacijo in programiranje s kontaktom skozi odprtino.

1 Programiranje z debelino oksida

Izvajamo ga v zgodnjih korakih proizvodnje integriranega vezja. Na napetost praga vplivamo z debelino oksida med gate elektrodo in kanalom. Poenostavljeno shemo takega vezja prikazuje slika:



Na zelo tanko plast silicijevega oksida nanesimo vzorec SiN_4 , ki določa z odprtinami mesta, kjer bo v nadaljevanju lahko zrastel debel silicijev oksid. Ker je kanal med source in drain P⁺ dopiran, napetost na besedni liniji

zadošča, da pri tankem oksidu kanal prične prevajati, pri debelem oksidu pa ostane neprevoden.

Odlika te rešitve je v dejstvu, da imamo na dve celici en priključek bitne linije, izvedene kot metalizacija z Mo. Ker lahko na besednih linijah zagotovimo kvaliteten krmilni signal, je uporabljen za besedne linije polisilicij. Vidimo, da koraku v katerem določimo "vsebino", sledi še veliko število korakov. Programiranje je izvedeno zelo zgodaj. Izvedba zagotavlja majhno površino celice. To je razvidno iz primerjav velikosti ROM, DRAM in SRAM celice v $2\mu\text{m}$ tehnologiji. Zahtevajo po vrsti $62\mu\text{m}^2$, $85\mu\text{m}^2$ in $550\mu\text{m}^2$. Dodatno pa DRAM in SRAM zahtevata več kot dva sloja za povezave in torej bistveno večje število korakov.

2 Programiranje z določanjem praga napetosti na osnovi implantacije

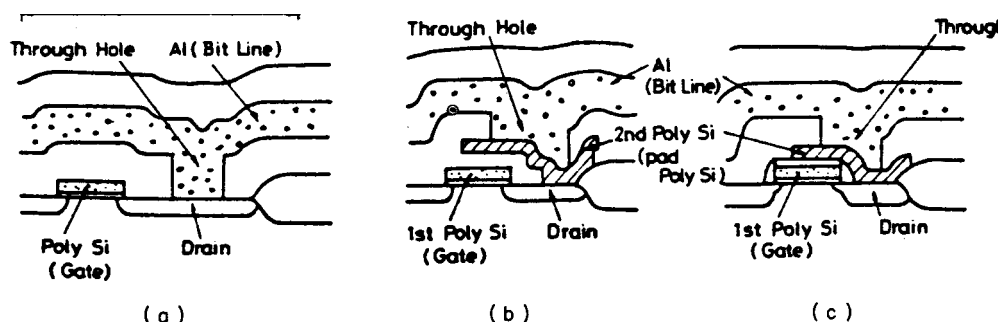
Tehnika uporablja močno implantacijo bora v n-kanalni tranzistor. Ta zaradi tega neha prevajati. Implantacijo je možno izvršiti skozi polisilicijevo gate elektrodo, ki pripada besedni liniji. Korak je možno izvršiti bolj proti koncu postopka proizvodnje.

Tehnika ima poleg prednosti, da ohranja visoko gostoto, tudi dve slabosti. Prva je v tem, da močna implantacija lahko poškoduje gate oksid, druga pa v visoki parazitni kapacitivnosti spoja med source (drain) in področjem kanala, ki nastane zaradi implantacije borovih ionov v področje kanala. To seveda zveča celotno kapacitivnost besednih linij in s tem poslabša čas dostopa.

3 Metoda programiranja z odpiranjem lukenj

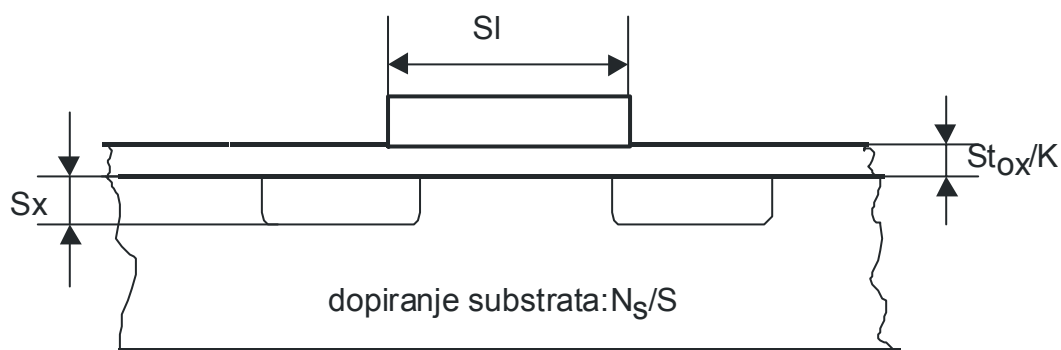
Možna je v zadnjem koraku nanašanja metalne plasti za povezave. Zahteva, da se odprejo skozi zaščitni oksid luknje do drain elektrode in potem nanese vzorec povezav. Potrebno je ustvariti kontakt z vsako programirano drain elektrodo. Dostopne morajo torej biti vse, kar povečuje potrebno površino in s tem zmanjšuje število celic in povečuje ceno na celico.

Nov zagon je ta tehnika doživela z uvedbo druge plasti polisilicija. Sedaj za vsako celico v drugi plasti nanese skrit kontakt, ki delno prekriva tudi področje gate elektrode. Do tega skritega kontakta iz polisilicija potem v zadnjem procesnem koraku izjedkamo luknjo in nanese bitno linijo iz Al. Slika prikazuje razlike med klasičnim procesom direktne metalizacije drain elektrode in procesoma, kjer metaliziramo skriti kontakt polisilicija.



Zmanjševanje velikosti tranzistorja

Izboljšave procesov (maske, nadzor nad dopiranjem, uvedba postopkov brez mask) omogočajo doseganje vedno manjših mer (geometrije) delov osnovnih gradnikov. Začetki so bili blizu $10\mu\text{m}$, zatem so na prelomu 80. in 90. let napovedale svoj prihod že "submikronske" tehnologije, da se danes približujemo že $0,1\mu\text{m}$. Večina avtorjev soglaša, da se bo ta trend ob tem, kar vemo o tehnologiji že danes, nadaljeval še vsaj 10 let. Za MOS tranzistor lahko pravila pomanjšanja prikažemo na sliki:



Seveda pa tako zmanjšanje dimenzij zahteva tudi zmanjšanje napetosti, pri kateri deluje vezje, če naj bo učinek popoln. Vidimo, da razen pri zmanjšanju debeline oksida nastopa povsod samo faktor S . Skrajšanje kanala med source in drain pomeni povečanje hitrosti delovanja takega tranzistorja. Prav dodaten faktor K , ki se pojavlja pri debelini oksida, kaže na to, da je zmanjšanje možno izvršiti le v določenih mejah, potem pa je potrebno na novo določiti celotna snovalska pravila.

Nekaj lastnosti silicijevega oksida

Silicijev oksid je dober izolator in ga zato lahko uporabljamo za horizontalno plast, ki ločuje plasti, ki prevajajo, kot tudi za vertikalno izolacijo sosednjih tranzistorjev. Vendar debel oksid, ki nastane na površini silicija, potem ko vanj izjedkamo odprtine, predstavlja potencialno nevarnost za lom predvsem tankih prevodnih kovinskih plasti, ki jih nanese preko praga, ki nastane.

Silicijev oksid nastaja na najbolj enostaven način z oksidacijo silicija v atmosferi bogati z kisikom pri povišani temperaturi. Ta postopek daje najbolj kvaliteten oksid, ki tudi pri zelo tanki plasti ne kaže defektov.

Pogosto potrebujemo oksid na mestu, ki je bilo pred tem dopirano. Če ne želimo porušiti zaželenih lastnosti dopiranja, je potrebno oksid nanesti v kemičnem procesu pri dovolj nizki temperaturi. Vendar se v tem postopku lahko v oksid vgradijo nečistoče, prisotne v plinski mešanici, nastanejo lahko poleg oksida tudi še druge spojine silicija. Govorimo o defektih oksidne plasti in je zato takšna plast primerna zgolj za izolacijo.

Kadar je lahko oksidna plast debela, kot je v primeru vertikalne ločitve tranzistorjev, taki defekti niso kritični. Bistveno drugačne pa so razmere v primeru izredno tankih oksidnih plasti, ki smo jih srečali npr. pri EEPROM vezjih.

Pomemben napredek je bila tehnika, s katero so zaščitili področja, na katerih so v naslednjih korakih procesa oblikovali tranzistorje. Zaščito so izvršili s SiN_4 . Nezaželen efekt tega procesa (LOCOS) je bil le v tem, da je oksid, ki je zrastel na siliciju, segel tudi delno pod SiN_4 . Z modifikacijami so uspeli vpliv tega širjenja bistveno omejiti.

Poseben postopek je izolacija z jarkom. Ta zahteva, da je relativno globoko na površino silicija npr. z reaktivnim ionskim jedkanjem izkoplje jarek in vanj kemično nanese oksid. Ta tehnika omogoča tudi tvorbo kondenzatorjev visoke kapacitivnosti na majhni površini.

Za povezave se je v začetku izključno uporabljala kovina (Al, redko tudi Mo). Bistven napredek je prineslo močno dopiranje silicija (polisilicij), ki je omogočilo najprej eno dodatno plast povezav, zatem pa še drugo in tretjo. Predvsem uvedba tretje plasti je omogočila izvedbo pomnilniškega kondenzatorja nad stikalnim tranzistorjem, kar je ena izmed tehnologij, ki je v 90. letih zagotovila rast kapacitete DRAM pomnilnikov.

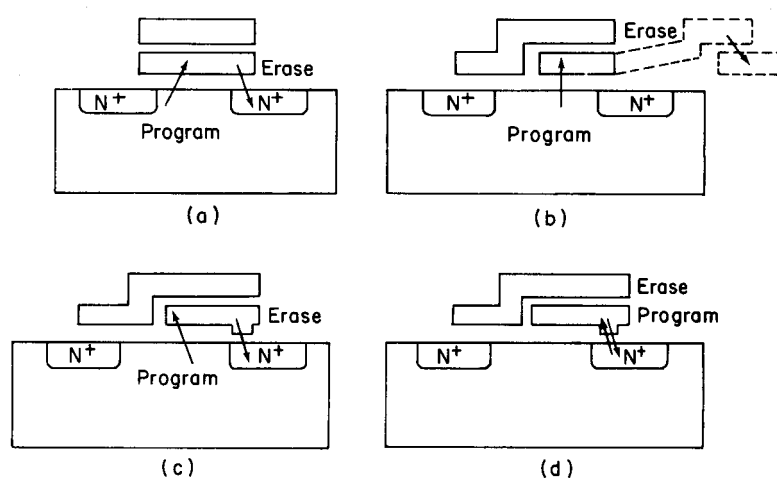
Flash pomnilniki

Samo poimenovanje izvira iz dejstva, da je bilo možno hitro (bliskovito) elektriško zbrisati vsebino celotnega pomnilnika in jo potem ponovno selektivno vpisati - za razliko od EEPROMa, kjer brišemo vsak byte posebej.

Snovalci pa so kmalu ponudili tudi pomnilnike, ki jih je bilo možno brisati po (vnaprej določenih) delih (blokkih, sekcijah) in si tako prihraniti nepotreben vpis. To omogoča notranja organizacija čipa. Flash (rečemo tudi Flash EEPROM) je torej mnogo hitrejši kot običajen EEPROM, ker piše po blokkih (običajno 512 bytov) namesto 1 byta naenkrat.

Možen je tudi naslednji korak, to je brisanje posameznih celic, ki pa bi bistveno zmanjšal razpoložljivo površino za celice in se ne uporablja.

FLASH pomnilniki so se razvijali skozi več faz. Prikazane so na sliki (levo ponor, desno izvor):



Najbolj preprosta je rešitev (a), pri kateri uporabimo tanjši oksid pod plavajočo elektrodo (tanjši, kot je uporabljen pri EPROM). Programiranje poteka z vročim injektiranjem elektronov iz področja kanala blizu drain, brisanje pa s tuneliranjem elektronov iz plavajoče elektrode skozi tanek oksid med to elektrodo in source elektrodo. Prednost te rešitve je majhna velikost celice, slabost pa sorazmerno velik tok, potreben za programiranje.

Številni pomnilniki uporabljajo koncept prikazan na slikah (b), (c) in (d). Površino nad kanalom si delita dve elektrodi; plavajoča in krmilna. Zato se ta način imenuje "split gate". To je izpeljanka dvotransistorske EEPROM celice, kjer imamo izbiralni (select) in pomnilni (storage) transistor, in hkrati enotransistorske EPROM celice. Flash celica sicer zavzame večjo površino od EPROM celice (en transistor), vendar znatno manjšo od EEPROM celice (dva transistorja). Flash celica se smatra za enotransistorsko, tako kot EPROM.

Za razliko od vročega injektiranja elektronov, ki je samoomejujoče, pri tuneliranju tega efekta ni in lahko pride do pozitivnega naboja plavajoče elektrode (pobegne preveč elektronov). Preprečiti moramo brisanje že zbranih celic, kar dosežemo s krmilno elektrodo. Ta blokira tranzistor vedno, razen v času branja. Posamezne rešitve imajo seveda ponovno prednosti in slabosti. Rešitev (b) zahteva sorazmerno visoko napetost za brisanje, ki jo moramo uporabiti med dvema polisilicijevima plastema, ločenima s sorazmerno debelim oksidom. Rešitev (c) zahteva za brisanje manjšo napetost (tanek oksid), uporablja pa za obe operaciji isti spoj, kar obremeni oksid in zmanjša število ponovitev. Rešitev (d) uporablja za brisanje in vpis hladno injektiranje (tuneliranje). Potreben nizek tok lahko zagotovimo že z 5V napajalno napetostjo. Hladno injektiranje zahteva daljši čas, kar pri več ponovitvah tudi kvarno vpliva na oksid.

FLASH pomnilniki so pri vpisu počasnejši od SRAM pomnilnikov, imajo pa pred njimi bistveno prednost, da se vsebina ne izgubi. Idealni so torej v vseh aplikacijah, kjer je potrebna pogostost vpisa zanemarljiva v primeri z branjem.

Struktura pomnilne celice Flash je v osnovi enaka kot pri EPROM-u. Sta pa dve bistveni razliki med obema. Prvič, Flash ima tanjšo oksidno plast (10 nm ali manj), da je možno tuneliranje elektronov iz plavajoče plasti med brisanjem. Drugič, ima globlji izvor, da je možno hitrejše brisanje.

Današnji pomnilniki Flash vsebujejo v čipu končni avtomat, ki avtomatizira operaciji pisanja in brisanja. V starejših pomnilnikih Flash je moral to početi računalnik.

Pri bralnih pomnilnikih in FLASH pomnilnikih se pojavlja potreba po generatorjih dodatnih napetosti. To so napetost substrata proti masi, napetosti za programiranje in pri submikronskih tehnologijah tudi napetosti, ki so nižje od 5V napajalne napetosti. Napajalno napetost uporabljajo le vezja (tranzistorji) na periferiji, v samih pomnilnih celicah pa zaradi kratke dolžine kanala med source in drain uporabljamo nižjo napetost, da ne prihaja do vročega injektiranja elektronov. Dodatne napetosti so generirane interno («charge-pumps«).

Kakor pri ostalih polprevodnikih, je tudi pri FLASH pomnilnikih velikost čipa glavni dejavnik pri ceni. Kot smo že omenili pri bralnih pomnilnikih, je možno matriko celic realizirati kot "negirano in" (NAND) ali "negirano ali" (NOR) izvedbo. Pri FLASH pomnilnikih se zato običajno uporabljajo tovrstne alternativne arhitekture matrike. Predvsem v NAND izvedbi se javlja ponovno problem optimalnega števila tranzistorjev v verigi.

V izvedbah z NOR matriko se zaradi velike kapacitivnosti bitnih linij uporablja rešitev z diferencialnim izhodnim ojačevalnikom. Eno vejo napaja tok izbrane pomnilniške celice, drugo pa tok referenčne celice.

Pisanje pri NOR Flash običajno uporablja plazovit preboj elektronov. Pri tem uporabimo napetost 12V na krmilni elektrodi, 6V na ponoru, izvor pa vežemo na maso. Ker je ponor pozitiven proti izvoru, elektroni začnejo teči proti ponoru (to pomeni električni tok proti izvoru), pri tem pa pridobijo nekaj kinetične energije. Ker pa je na krmilni elektrodi še višja napetost, jih potegne preko oksida v plavajočo plast. Posledica je povečanje pragovne napetosti transistorja, ki ne prevaja, tudi če je izbran. Tipalni ojačevalniki zaznajo, da ni toka, kar pomeni vrednost "0".

Brisanje pri NOR Flash poteka na osnovi tuneliranja. Na izvoru je 12V, ponor je na masi, krmilna elektroda pa je plavajoča (brez napetosti). Elektroni odtečejo iz plavajoče plasti proti izvoru. Transistor prevaja, kadar je izbran, kar pomeni "1" v celici.

NAND Flash pomnilniki imajo namesto enega kontakta na dve celici (NOR) en kontakt na več celic (npr. 16), ki so vezane serijsko preko dveh izbiralnih transistorjev (GSL: ground select, in SSL: bit-line select). Brana celica ima

besedno linijo na masi, ostalih 15 celic le prevaja signal. V stanju »0« ima celica pozitivno pragovno napetost (programirana celica), v »1« pa negativno (zbrisana celica). Operacija beri-skozi je počasna (10us), zato je NAND flash primeren kot masovni pomnilnik.

Programiranje poteka s tuneliranjem iz kanala v plavajočo plast, pri tem je besedna linija na napetosti 18-20V, substrat pa je na masi. Brisanje poteka s tuneliranjem iz plavajoče plasti v kanal, kar povzroči negativno pragovno napetost, pri tem je besedna linija na masi, substrat pa na 20V (s pomočjo charge-pumps, ki zavzamejo precej prostora na čipu).

Primer NAND Flasha je AMD-64 Mb, ki ima 1024 blokov, vsak blok ima 16 strani (pages) po 512 bytov (blok je torej 8KB). Ukazni register in vhodno/izhodno vodilo sta 8-bitna. Vsebina ukaznega registra je vhod v končni avtomat, ki nadzoruje operacije branja, programiranja in brisanja. Programiranje poteka na nivoju strani, brisanje pa na nivoju blokov. Za oboje ima čip ustrezen algoritem. Napajanje je 3V, za ostale napetosti so »charge-pumpe«. Pomnilnik ima 50ns dostop do zaporednih bytov.

NAND Flash v primerjavi z ROM Flash:

- Hitrejše branje in pisanje
- Večja gostota
- Nižja cena
- Toda: sekvenčni dostop do podatkov. Zato primeren za mass-storage (smart media, memory stick, multimedia card, usb drive)

NAND ni primeren kot zamenjava za programski rom (mikroprocesorji zahtevajo dostop do programa na nivoju bytov).

Večina NAND Flash pomnilnikov uporablja t.i. upravljanje s slabimi bloki (bad-block management). Pri procesu izdelave prihaja do defektov. Če bi želeli pomnilnik 100% brez napak, bi bil proces bistveno dražji. Tako pa je ceneje dodati mehanizem, s pomočjo katerega pomnilnik teh blokov ne uporablja, ostale pa preslika tako, da navzven izgleda vse v redu.

FLASH pomnilniki potrebujejo interno večje število različnih napetosti. V tabeli vidimo potrebne napetosti pri brisanju in programiranju pomnilnika tipa NOR Flash:

	G	S	D
Izbris	0 V	12 V	plavajoč
	-10 V	5 V	plavajoč
	0 V *	plavajoč	plavajoč
Vpis	12 V	0 V	5 V

* in pozitivna napetost na substratu

Tudi Flash pomnilnik ni zamenjava za RAM, kajti RAM mora biti naslovljiv na bajtnem nivoju, ne na nivoju bloka. Dostop do Flasha je običajno preko SW gonilnika.

Flash pomnilnik se uporablja v

- USB pogonih
 - računalnikih (BIOS),
 - mobilnih telefonih,
 - digitalnih kamerah (CompactFlash, SmartMedia, Memory Stick),
 - LAN stikalih,
 - karticah PCMCIA za prenosnike,
 - vgradnih sistemih (namesto diska),
 - igralnih konzolah
- itd.