Izvori disipacije snage

Postoje tri glavna izvora disipacije snage kod digitalnih CMOS kola. Imajući to u vidu potrošnja se može izraziti sledećom jednačinom

$$P_{total} = p_t * (C_L * V * V_{DD} * f_{CLK}) + I_{sc} * V_{DD} + I_{leakage} * V_{DD}$$

Prvi član, $p_t * (C_L * V * V_{DD} * f_{CLK})$, se odnosi na komutiranu komponentu snage, gde je:
C_L - komutirana kapacitivnost; f_{CLK} - taktna frekvencija; p_t - faktor aktivnosti; V - naponska promena impulsa na izlazu kola, ista kao V_{DD}; V_{DD} - napon napajanja
Drugi član posledica je kratkog spoja koji se javlja između PMOS i NMOS tranzistora kada su oni simultano aktivni, tj. struja protiče od napajanja ka masi kada oba tranzistora istovremeno provode.
Treći član se javlja zbog struje curenja, $I_{leakage}$, i zavisan je od tehnologije.

Treba istaći da dominantan uticaj ima prvi član, što znači da ako želimo da minimiziramo potrošnju, a da pri tome zadržimo neophodnu funcionalnost, treba da minimiziramo p_t, C_L, V_{DD}, f_{CLK}. (Kod CMOS digitalnih kola disipaciju snage čine dinamička (prvi član) i statičke komponente (drugi i treći član). S obzirom da je dinamička potrošnja proporcionalna sa kvadratom V_{DD}, a statička proporcionalna sa V_{DD}, evidentno je da smanjenje V_{DD}-a vodi ka najefikasnijoj redukciji potrošnje energije. Skaliranjem napona napajanja, *treshold voltage* (V_{th}) tranzistora treba takođe da se skalira kako bi se sačuvale performanse. Na žalost, ovo skaliranje dovodi do povećanja struje curenja (*leakage current*) koja postaje dominantni faktor od uticaja na performanse kod *low-voltage high-performance circuit designs*.)

U cilju efikasnijeg sagledavanja ove problematike uvešćemo proizvod *power-delay* koji se može interpretirati kao iznos potrošene energije pri svakom komutacionom događaju (ili prelazu). Ovaj proizvod je posebno koristan kod upoređivanja različitih stilova realizacije kola sa tačke gledišta disipacije. Pri ovome se usvaja da je kod disipacije snage važna samo prva komponenta, pa shodno tome

$$energija\_po\_prelazu = \frac{P_{total}}{f_{CLK}} = C_{efektivno} * V_{DD}^2$$

gde je $C_{efektivno} = p_t * C_L$.

Projektovanje kola i tehnološka razmatranja

Postoji veći broj opcija dostupnih kod izbora osnovnih kola i topologija za implementaciju različitih logičkih i aritmetičkih funkcija. Odluke o izboru se donose između:

- statičke $\leftrightarrow$ dinamičke implementacije
- pass-gate $\leftrightarrow$ konvencionalnih CMOS logičkih stilova
- sinhroni $\leftrightarrow$ asinhroni tajming.

Drugi nivo izbora se odnosi na različite arhitekturno/strukturne izbore za implementaciju date logičke funkcije. Na primer za realizaciju sabirača se može koristiti jedna od sledećih topologija:
a) *ripple-carry*; b) *carry*-select; c) *carry-lookahead*.

Arhitekturni pristupi za smanjenje potrošnje

U nekoj gruboj aproksimaciji snaga koja se disipira na nekom logičkom bloku data je sledećom relacijom

$$P_{dis} = C_{\log} * V_{DD}^2 * f_{CLK}$$

gde je:

$C_{\log}$ - efektivno komutirana kapacitivnost logičkog kola

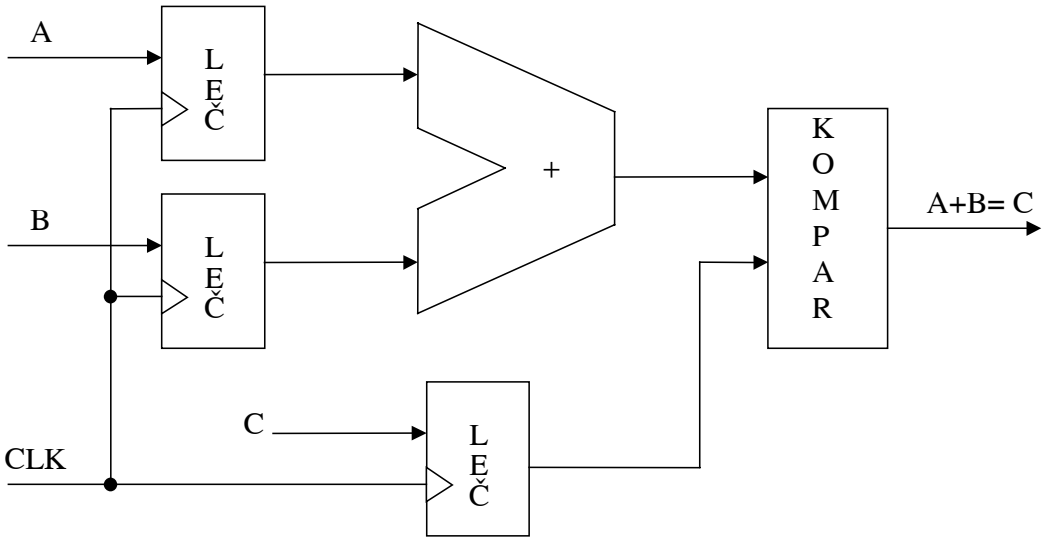
V_{DD} - napajanje logičkog kola

f_{CLK} - taktna frekvencija rada kola

Sa ciljlem da ukažemo kako arhitekturni pristupi utiču na potrošnju energije analiziraćemo strukturu deo staze podataka jednog mikroprocesora koju čine sabirač i komparator implementirani u 2 μm CMOS tehnologiji. Ako, u najgorem slučaju, pri naponu napajanja V_{DD} = 5V, kašnjenje signala kroz sabirač, komparator i leč aproksimativno iznosi 25 ns, tada u najboljem slučaju, taktna frekvencija rada dela staze podataka sa Slike 1 može biti 40 MHz. Snagu koja se disipira u tom slučaju obeležićemo sa P_{ref} i ona iznosi

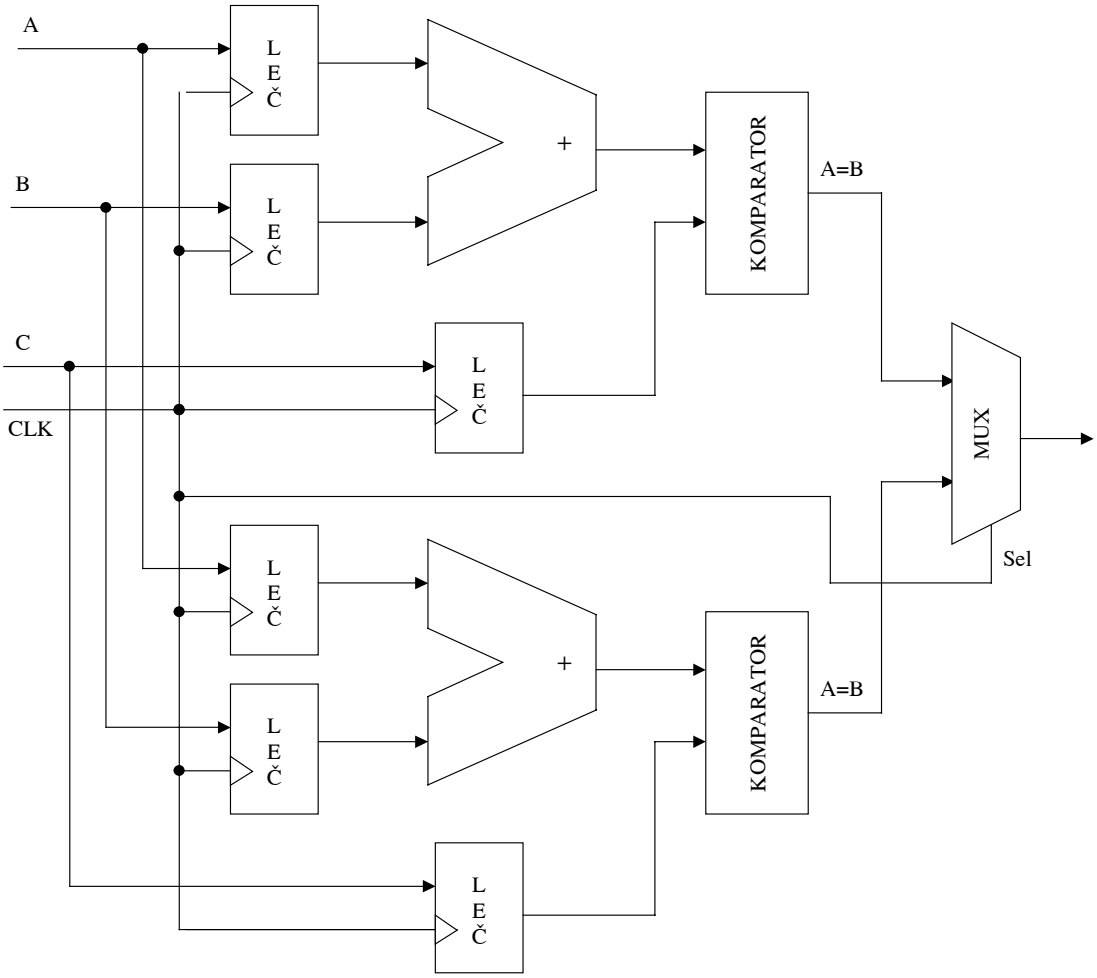
$$P_{ref} = C_{ref} * V_{ref}^2 * f_{ref}$$

gde je u konkretnom slučaju V_{ref} = 5V a f_{ref} = 1/25 ns.



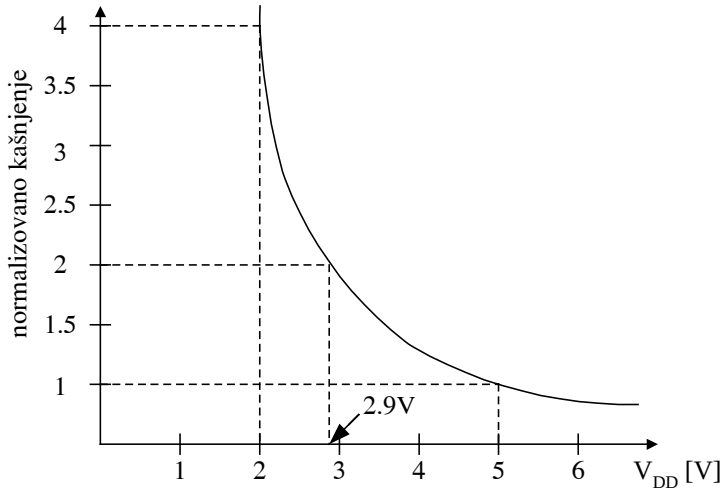
Slika 1 Deo staze podataka

Jedan od načina da se sačuva propusnost (*throughput*) a da se pri tome redukuje napon napajanja je da se koristi paralelna arhitektura. Na Slici 2 prikazana je struktura kola koja ispunjava zahteve u pogledu propusnosti a bazira se na korišćenju paralelizma. Kao što se vidi sa Slike 2 koriste se dva identična sabirača i komparatora.



Slika 2 Paralelna implementacija dela staze podataka

Pošto zahtevi za brzinom rada sabirača, komparatora i leča su sada smanjeni sa 25 ns na 50 ns, napon napajanja se može redukovati sa 5V na 2.9V (vidi Sliku 3).



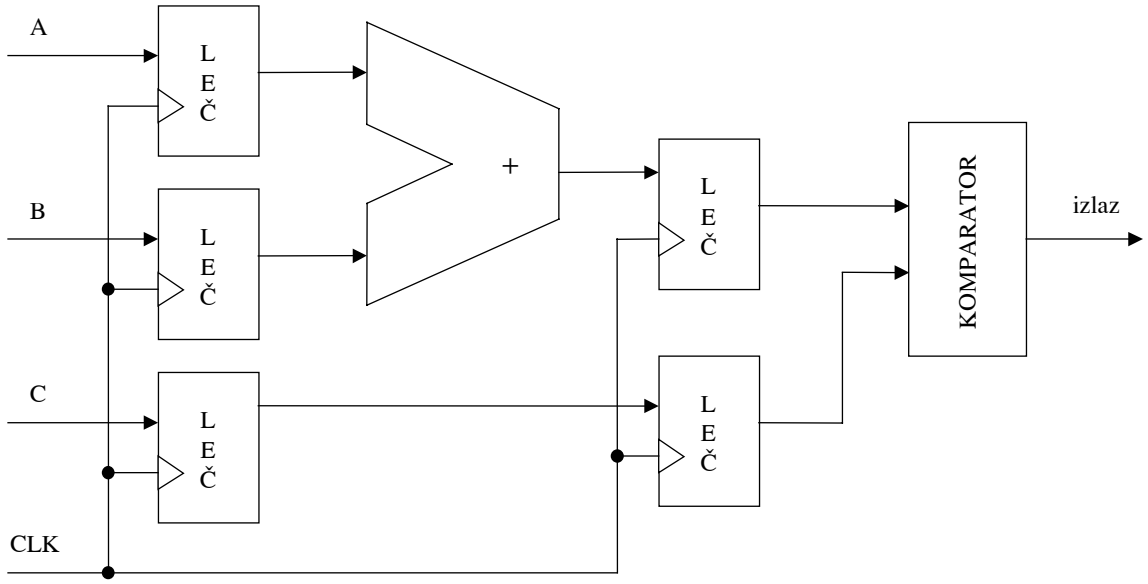
Slika 3 Karakteristike kašnjenja logičkih CMOS kola (gruba aproksimacija)

Kapacitivnost staze podataka je povećana zbog potrebe rutiranja za faktor 2.15 (idealno bi bilo za faktor 2) tako da je ukupna disipirana snaga sada

$$P_{par} = C_{par} * V_{par}^2 * f_{par}$$
$$= (2.15 \cdot C_{ref}) * (0.58 \cdot V_{ref})^2 * \left(\frac{f_{ref}}{2} \right)$$
$$= 0.36 \cdot P_{ref}$$

Pristup koji se bazira na redukovanju napajanja koristeći paralelizam ima za posledicu povećanje površine čipa i kao rešenje je dobro kada ograničenja sa aspekta površine čipa ne postoje. U opštem slučaju paralelizam uslovljava ekstra rutiranje koje mora da se minimizira.

Drugo moguće rešenje se zasniva na korišćenju protočne arhitekture, kako je to prikazano na Slici 4. Sa dva dodatna protočna leča, kritični put postaje max[T_{sab}, T_{comp}] i obezbeđuje da sabirač i komparator rade pri manjim brzinama.



Slika 4 Protočna implementacija dela staze podataka

U rešenju sa Slike 4 kašnjenja kroz sabirač i komparator su jednaka, što obezbeđuje da se napajanje redukuje sa 5V na 2.9V a da pri tome propusnost ostane nepromenjena. Ipak dodavanjem novih protočnih lečeva (u odnosu na Sliku 1) povećava se površina čipa, a kao posledica aproksimativno efektivne kapacitivnosti je za faktor 1.15. Snaga koja se troši od strane protožne obrade podataka sa Slike 4 iznosi sada

$$P_{pro} = C_{pro} * V_{pro}^2 * f_{pro}$$
$$= (1.15 \cdot C_{ref}) * (0.58 \cdot V_{ref})^2 * f_{ref}$$
$$= 0.39 \cdot P_{ref}$$

Sa ovom arhitekturom snaga se smanjuje za faktor od aproksimativno 2.5 u odnosu na Sliku 1, ali treba naglasiti da površina dela staze podataka nije drastično povećana.

Kao logički se sada nameće zaključak do kakvih se rezultata dolazi implementacijom arhitektura koje se zasnivaju na kombinovanim korišćenjem tehnika protočnosti i paralelizma. Strateški posmatrano, protočno-paralelnu arhitekturu (udvostručeni paralelizam i udvostručena protočnost) karakteriše:

- redukcija kritičnog puta a time i povećanje brzine rada za faktor 4
- smanjenje napona napajanja sve dok se kašnjenje ne poveća za faktor 4 (vidi Sliku 3).

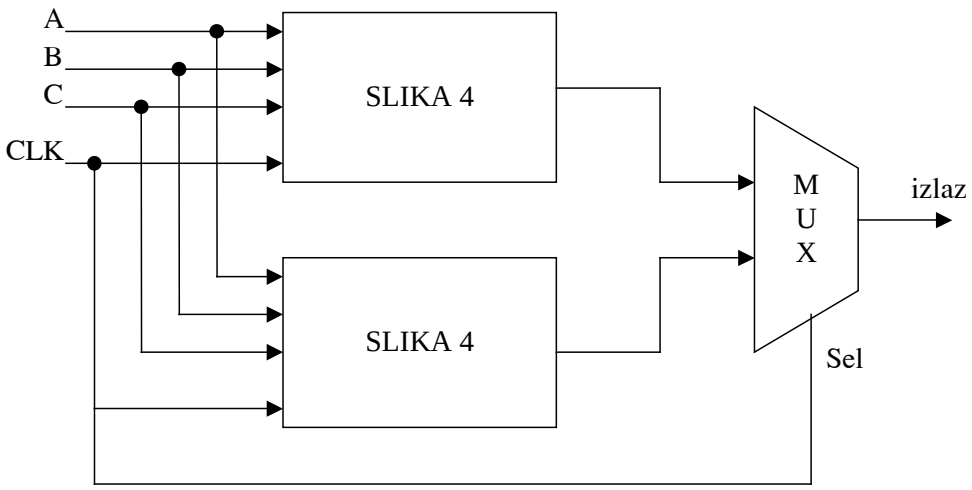
Potrošnja snage u ovom slučaju iznosiće

$$P_{parpro} = Cpar_{pro} * V_{parpro}^2 * f_{parpro}$$

$$= \left(2.5 \cdot C_{ref}\right) * \left(0.4 \cdot V_{ref}\right)^2 * \left(\frac{f_{ref}}{2}\right)$$

$$= 0.2 \cdot P_{ref}$$

Paralelno-protočna implementacija rezultira redukciji snage od 5 puta.



Slika 5 Paralelno-protočna staza podataka

Uporedni rezultati koji se odnose na različite arhitekture opisane na primeru jednostavne staze podataka koju čine sabirač i komparator su date na Slici 6.

Tip arhitekture	Napon	Površina	Snaga
jednostavna staza podataka (Slika 1)	5V	1	1
protočna staza podataka (Slika 4)	2.9V	1.3	0.39
paralelna staza podataka (Slika 2)	2.9V	3.4	0.36
paralelno-protočna staza podataka (Slika 5)	2.0V	3.7	0.2

Slika 6 Uporedni pregled karakteristika različitih staza podataka

Optimalni izbor napona napajanja

Već smo uočili da se iznos kašnjenja kroz digitalna kola povećava kako se smanjuje napon napajanja. Pri ovome smo naglasili da kada se napon napajanja smanjuje kompenzovanje povećanog kašnjenja kola se može uspešno rešiti korišćenjem paralelnih arhitektura. Ali treba pri ovome naglasiti da kada se napon napajanja približi *threshold* vrednosti kašnjenje kola se drastično povećava. Projektantima kola se sada postavlja sledeći zadatak: Kako odrediti optimalnu vrednost napona napajanja pri kojoj dodatno ugrađeni paralelizam ne daje više očekivane efekte. Da bi odredili vrednost napona koristićemo se modelom koji je definisan sledećom jednačinom:

$$Power(N) = N \cdot C_{ref} \cdot V^2 \cdot \frac{f_{ref}}{N} + C_{ip} \cdot V^2 \cdot \frac{f_{ref}}{N} + C_{\text{int}erface} \cdot V^2 \cdot f_{ref}$$

gde je:

N - broj paralelnih procesora,

C_{ref} - kapacitivnost jednog (single) procesora,

C_{ip} - interprocesorski komunikacioni *overhead* kao rezultat uvedenog paralelizma (upravljanje i rutiranje)

$C_{\text{int}erface}$ - *overhead* interfejsa koji se ne smanjuje sa brzinom ako se u arhitekturu uvede više paralelizma

U opštem slučaju C_{ip} i $C_{\text{int}erface}$ su funkcije N, a poboljšanja sa tačke gledišta potrošnje u odnosu na referentni slučaj (bez uvođenja paralelizma) se mogu izraziti kao

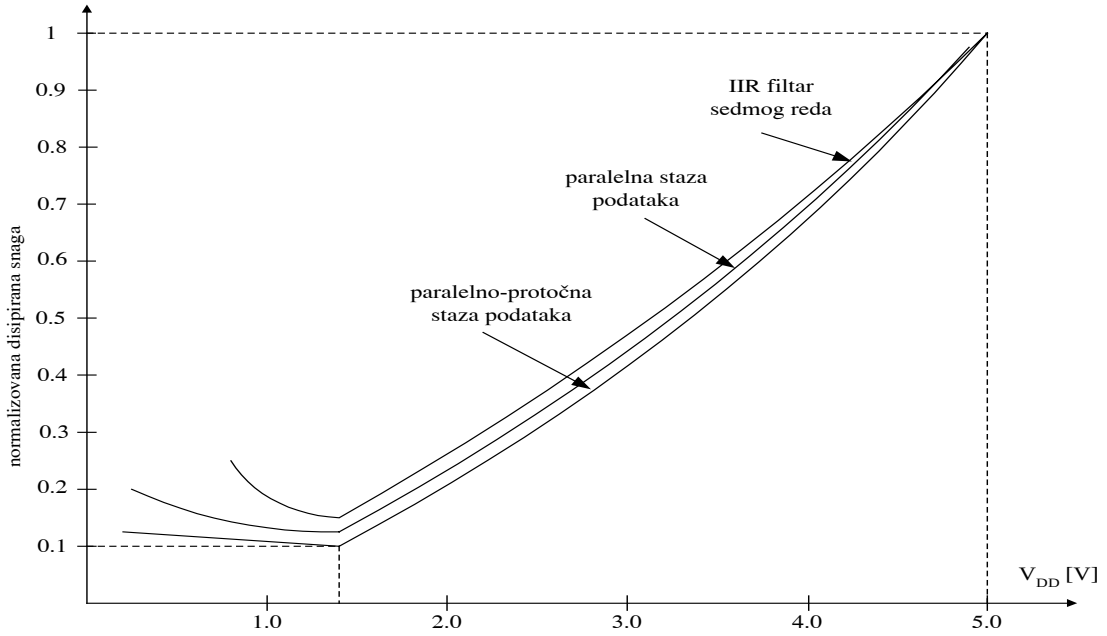
$$P_{normalized} = \left(1 + \frac{C_{ip}(N)}{N \cdot C_{ref}} + \frac{C_{\text{int}erface}(N)}{C_{ref}}\right)\left(\frac{V}{V_{ref}}\right)^2$$

Pri malim naponima napajanja (blizu *threshold*-a)broj procesora (a time i *overhead*) obično raste brže nego što član V^2 opada što dovodi da se sa daljim smanjenjem napona napajanja potrošnja povećava.

Drugo ograničenje koje se odnosi na smanjenje napona napajanja dolazi od *noise-margin* ograničenja, pri čemu mora biti ispunjen sledeći uslov

$$V_{noise-m\arg in} \le V_{optimal} \le V_{critical}$$

Na Slici 7 prikazan je dijagram disipirane snage normalizovan na 1 za V_{DD} = 5V u funkciji V_{DD} -a za različita arhitekturna rešenja kod 2.0 μm CMOS tehnologije.



Slika 7 Optimalni napon rada

Analizom Slike 7 se može zaključiti da je optimalni napon relativno nezavisan i za 2.0 μm CMOS tehnologiju iznosi 1.5V.

Na Slici 8 prikazane su normalizovane vrednosti površina/snaga za različite napone napajanja arhitekture sa Slike 7.

napon	arhitektura		
	paralelna površina/snaga (Slika 2)	protočno-paralelna površina/snaga (Slika 5)	IIR filtar sedmog reda površina/snaga
5	1/1	1/1	1/1
2	6/0.19	3.7/0.2	2.6/0.23
1.5	11/0.13	7/0.12	7/0.14
1.4	15/0.14	10/0.11	dostiže se rekurzivno usko grlo

Slika 8 Normalizovana površina/disipacija za različite napone napajanja

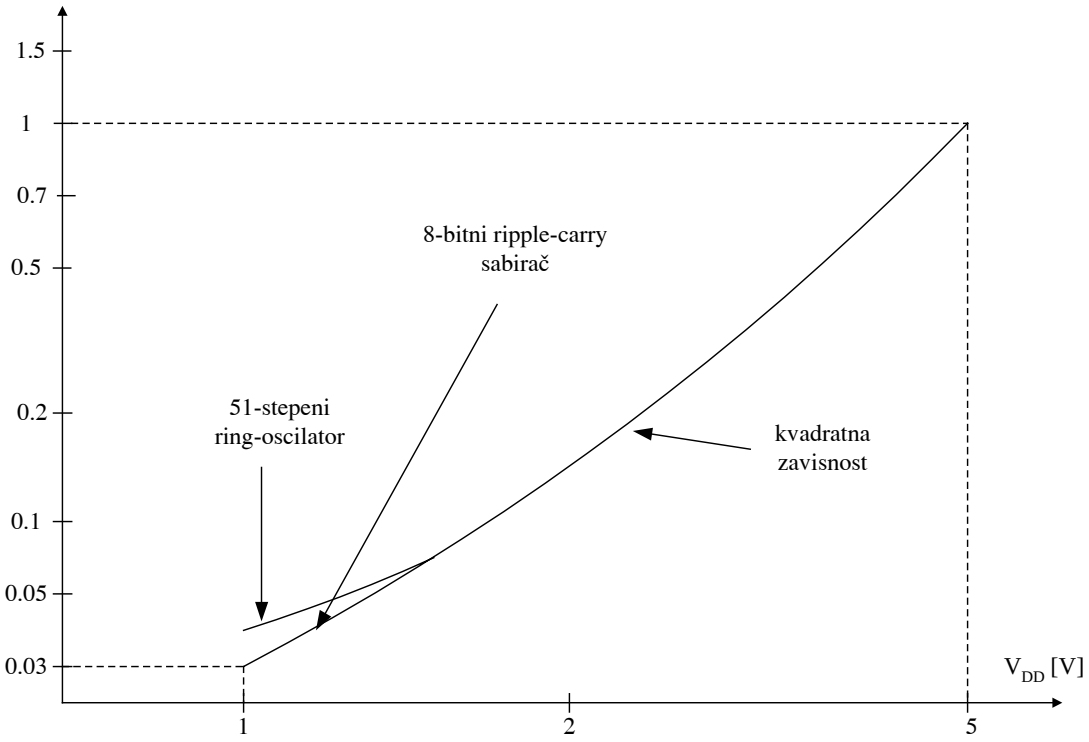
Strategija za smanjenje potrošnje

U zavisnosti od tipa dgitalnih kola, sinhroni ili asinhroni, koriste se različite strategije. Logika koja se bazira na sinhronim kolima koristi registre (lečeve) koji se ubacuju između stepena koji vrše izračunavanje. Stepeni se realizuju kao kombinaciona logika a pamćenje informacije u lečevima se vrši nailaskom svakog taktnog impulsa. Da bi se smanjila potrošnja logike zasnovane na sinhronom dizajnu

neophodno je minimizirati komutatorske aktivnosti. Smanjenje ovih aktivnosti se obično izvodi tako što se komutatorske aktivnosti izvršne jedinice ne izvode kada kolo ne obavlja željene aktivnosti. Ovo je važan aspekt jer logički moduli mogu komutirati i trošiti energiju čak i slučaju kada se oni aktivno ne koriste. To znači da dizajn sinhronih kola treba da se bazira na specijalnim kolima kao i rešenjima kojima se detektuje neaktivnost nekog logičkog modula a zatim uključuju *power-down* kola pomoću kojih se redukuje napon napajanja tog dela logike.
Kod asinhronih digitalnih kola situacija je nešto drugačija jer se njihov princip rada bazira na konceptu *power-down* kada ta kola nisu aktivna.

Skaliranje napona napajanja

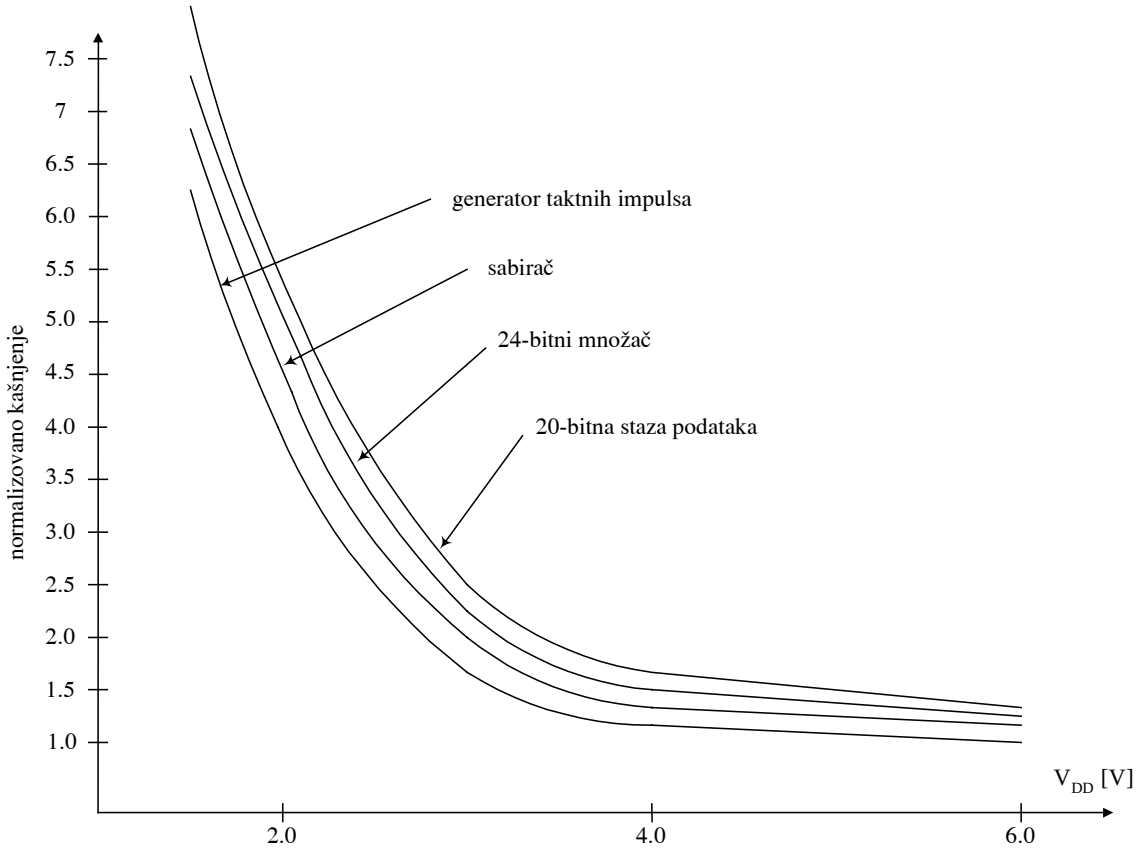
Energija koja se troši po prelazu, ili *power-delay* proizvod, kod CMOS kola proporcionalna je sa V^2. Na Slici 9 prikazan je dijagram normalizovanog *power-delay* proizvoda za dva kola (*ring*-oscilator i 8-bitni *ripple-carry* sabirač).



Slika 9 *Power-delay* proizvod za dva različita kola

Napomena: Slika 9 je data u log-log razmeri

Kao što se vidi sa Slike 9 smanjenje napona napajanja kod date logičke familije ima za efekat kvadratno poboljšanje *power-delay* proizvoda.
Na žalost, smanjenje napona napajanja ima svoju cenu. Na Slici 10 prikazan je efekat koji ima smanjenje V_{DD}-a na kašnjenje za različita logička kola (složenost logičkih kola u zavisnosti od funkcije koju ona obavljaju se menja od 56 do 44000 tranzistora, ali je bitno to što sve krive ukazuju na to da postoji identična zavisnost).



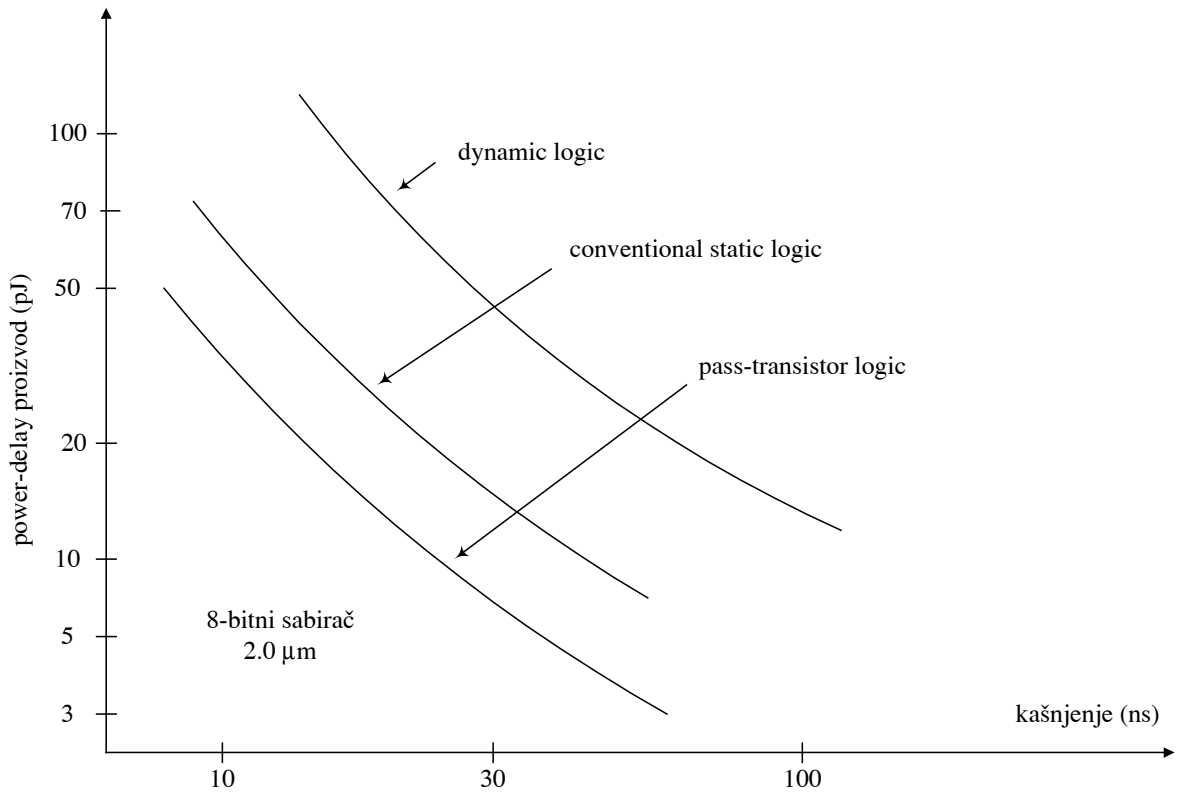
Slika 10 Karakteristike kašnjenja za različita logička kola za 2.0 μm CMOS tehnologiju

Napomena: Složenost kola u odnosu na broj tranzistora je sledeća - 20-bitna staza podataka ima ugrađeno 44802 tranzistora, 24-bitni množač ima 20432 tranzistora, sabirač ima 256 a generator taktnih impulsa 56 tranzistora.

Prosto rečeno, smanjenjem V_{DD}-a povećava se kašnjenje koje je drastično kada se približavamo *threshold* naponu. Tačna analiza kašnjenja je složena zbog nelinearne karakteristike CMOS gejta, ali za grubu aproksimaciju prvog reda važi da je

$$T_d = \frac{C_L * V_{DD}}{I} = \frac{C_L * V_{DD}}{\mu \cdot C_{OX} \cdot \left(\frac{W}{L}\right) \cdot (V_{DD} - V_T)^2}$$

Dok *power-delay* produkt za 8-bitni sabirač izveden u 2.0 μm CMOS tehnologiji za različite stilove (*pass-tranzistor logic, conventional static, dynamic logic*) je prikazan na Slici 11.



Slika 11 *Power-delay* proizvod za različite stilove

U fazi optimizacije arhitekture za *low-power* dizajn treba tretirati V_{DD} kao promenljivu veličinu a pri tome menjati arhitekturu tako da se održi konstantna propusnost.