- Napajanje i energija u integrisanim kolima
  - Napajanje mora biti dovedeno do svake aktivne komponente, mnogo pinova i slojeva štampe se troše na napajanje (+XV i 0V)
  - Gledano za ceo računarski sistem
    - Šta je maksimalna količina energije koja je neophodna?
      - Napajanje mora biti u stanju da to obezbedi
    - Kolika je stalna potrošnja?
      - Naziva se i TDP - thermal design power
      - Napajanje mora biti jače od TDP, a hlađenje mora biti u stanju da ohladi sistem kada troši TDP ili više
      - Procesori mogu da smanje takt kada temperatura postane previsoka; takođe mogu se i kompletno isključiti da spreče pregorevanje
    - Energetska efikasnost
      - Energija je generalno bolja mera od snage (1W = 1J/1s) jer uključuje i potrebno vreme za obavljanje zadatka
      - Poređenje efikasnosti dva procesora/sistema radimo preko utrošene energije za neki zadatak; račun za struju zavisi od utrošene energije, dužina rada bez punjenja baterije takođe

- Napajanje i energija u integrisanim kolima
  - Gledano za ceo računarski sistem
    - Potrošnja snage se najčešće koristi kao ograničenje, a ne kao mera efikasnosti
  - Unutar mikroprocesora
    - Najviše energije se troši na prebacivanje tranzistora iz jednog stanja u drugo, tzv. dinamička energija:

      $$E_{din} \sim C * V^2$$

      C - kapacitivno opterećenje tranzistora, V - napon

      Ovo odgovara jednom impulsu (0->1->0 ili 1->0->1), za jedan prelaz bi bilo polovina toga
    - Snaga za jedan tranzistor:

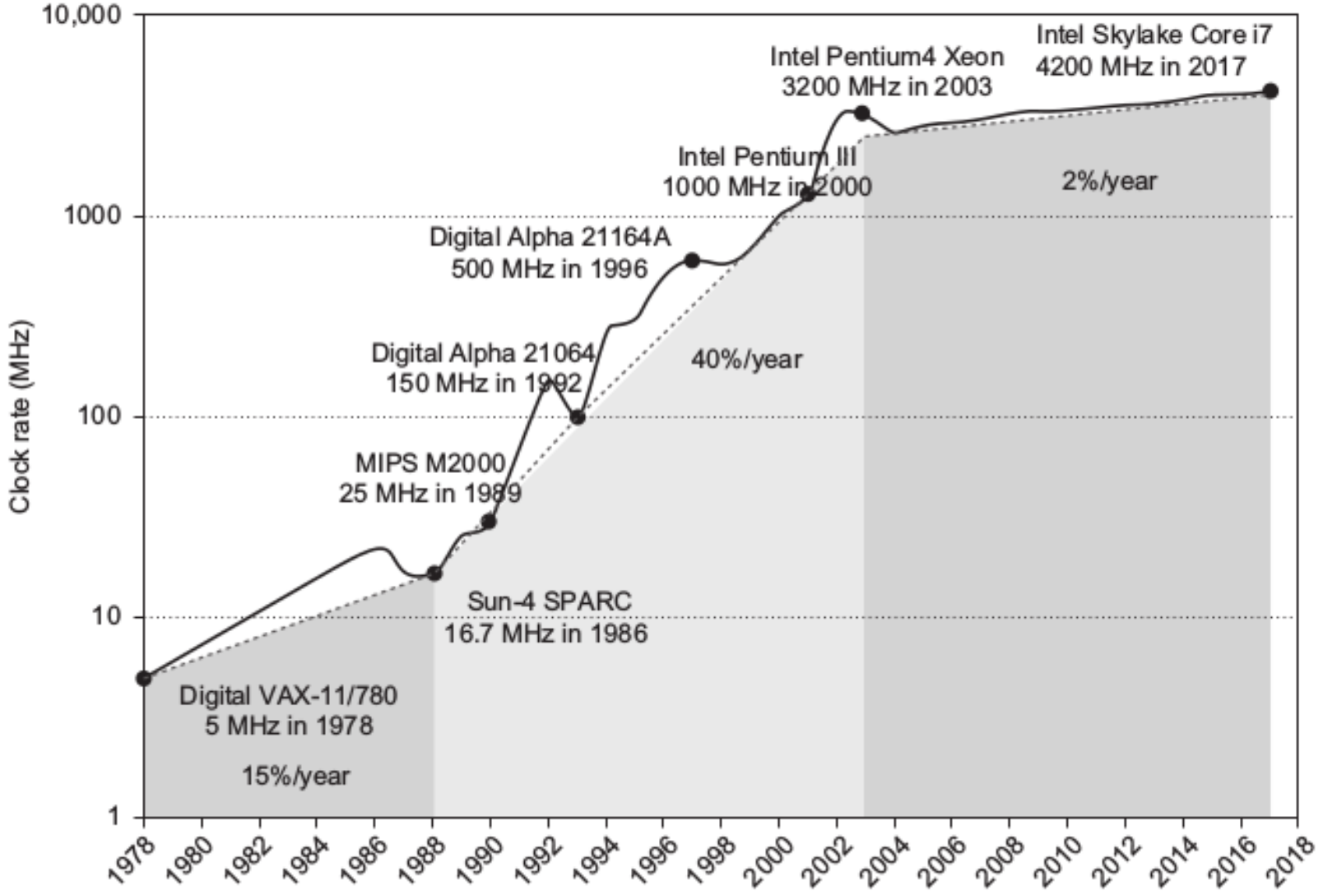
      $$P_{din} \sim 1/2 * C * V^2 * F$$

      F - frekvencija

      Za izvršenje zadatka, smanjenje frekvencije smanjuje snagu, ali ne i potrebnu energiju

- Napajanje i energija u integrisanim kolima
  - Unutar mikroprocesora
    - Dinamička snaga i energija se mogu smanjiti smanjenjem napona; napon se od 80-tih do danas smanjivao sa +5V na nešto ispod 1V
    - Kapacitivno opterećenje je funkcija broja tranzistora povezanih na izlaz posmatranog i tehnologije izrade (određuje kapacitivnosty vodova i pojedinačnih tranzistora)
    - Iako se napon smanjivao i tehnologija izrade poboljšavala, broj tranzistora je rastao više: prvi mikroprocesori su trošili manje od 1W, 80386 je trošio oko 2W, današnji procesori znaju da troše i više od 100W

- Napajanje i energija u integrisanim kolima
  - Unutar mikroprocesora
    - Kao što je spištanje radnog napona skoro dostiglo limit, tako je i podizanje frekvencije



- Napajanje i energija u integrisanim kolima
  - Tehnike poboljšanja energetske efikasnosti
    - Isključenje kloka segmentima koji se ne koriste
    - Dinamička promena frekvencije (DVFS - Dynamic voltage-frequency scaling)
      - Smanjenje frekvencije u periodima smanjenog opterećenja
    - Dizajniranje za tipičnu upotrebu
      - Prenosni uređaji se često ne koriste, pa se njihove komponente prave tako da mogu raditi u režimima smanjene potrošnje i efikasnosti (procesor, memorija, diskovi)
    - Overclock
      - Turbo mode, Intel, 2008 - povećanje frekvencija u kraćim intervalima, dok temperatura ne postane kritična (često kombinovano sa isključivanjem jezgara koje se ne koriste)
  - Statička snaga je takođe bitna
    - Nastaje zbog struje curenja tranzistora kada je isključen

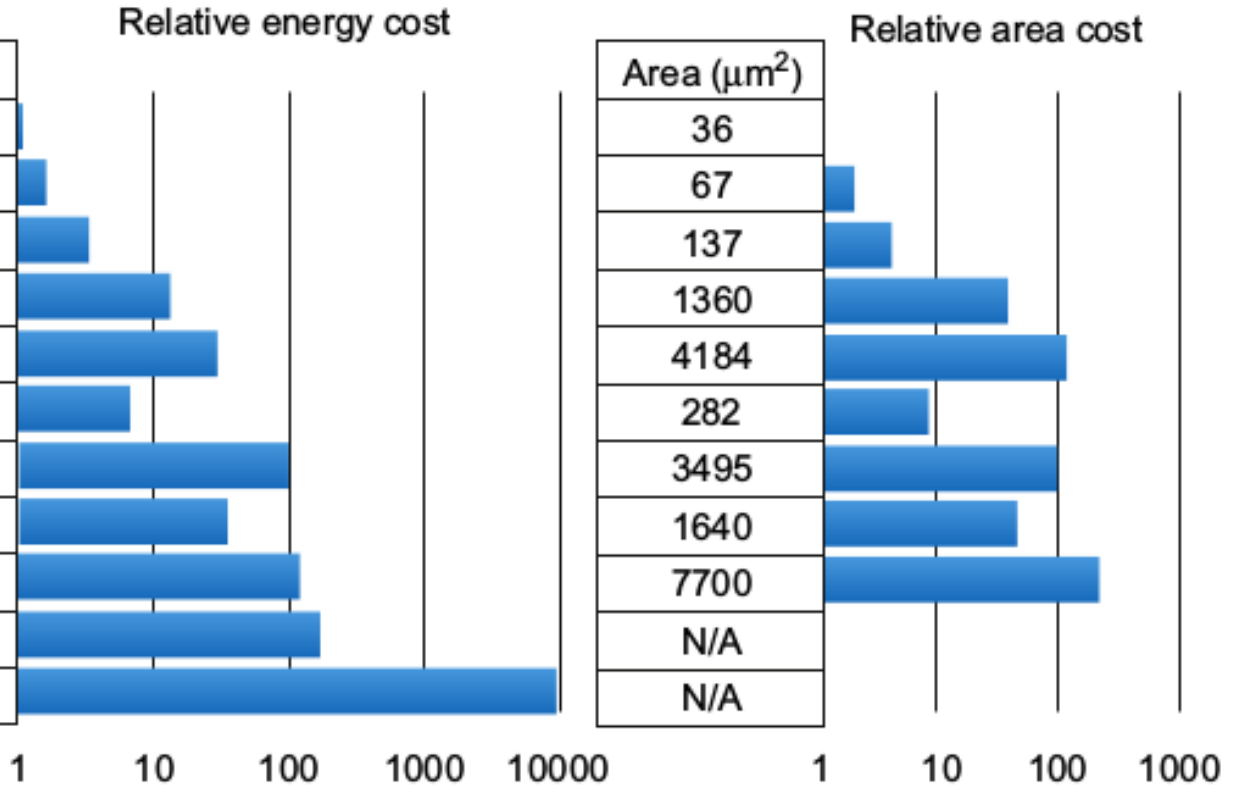
      $$P_{stat} \sim I_{stat} * V$$

    - Veća je kod manjih tranzistora i proporcionalna broju tranzistora
    - Neaktivnim delovima se može isključiti napajanje, ne samo klok
  - Performanse po J ili W su zamenile preformanse po mm^2

- Napajanje i energija u integrisanim kolima
  - Uticaj na arhitekturu računara
    - Danas je jedan od glavnih ciljeva što bolje iskorištenje energije
    - dark silicon - zbog energetskih ograničenja, dešava se da je nemoguće da svi delovi čipa budu istovremeno pod napajanjem - u svakom trenutnutku neki deo čipa mora biti isključen
      - tehnički, može se sve uključiti, ali tada TDP raste preko sigurnih granica i temperatura čipa bzo prelazi dozvoljene granice
      - neke procene kažu da će za proizvodne procese ispod 10nm procenat čipa koji će morati biti isključen dostizati i do 50%

- Napajanje i energija u integrisanim kolima
  - Uticaj na arhitekturu računara
    - Prisup za 32 bita DRAM-a troši ~6500 puta više energije nego 32-bitno sabiranje
    - SRAM je ~150 puta energetski efikasniji od DRAM-a
    - Uvođenje domenski specifičnih (delova) procesora i memorije - manja potrošnja za neke zadatke

Operation:	Energy (pJ)	Area (μm^2)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FB Add	0.4	1360
32b FB Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FB Mult	1.1	1640
32b FB Mult	3.7	7700
32b SRAM Read (8KB)	5	N/A
32b DRAM Read	640	N/A



Energy numbers are from Mark Horowitz *Computing's Energy problem (and what we can do about it)*. ISSCC 2014

- Ekonomski faktori integrisanih kola
  - Pošto se proizvodni troškovi smanjuju tokom vremena, troškovi proizvodnje komponenti padaju i bez unapređenja tehnologije
    - Learning curve - svako sledeće izvođenje nekog zadatka zahteva manje vremena nego prošli put (odnosi se na učenje, proizvodnju, itd)
    - veći procenat ispravnih čipova na vaferu (yield) -> manji troškovi po komadu
    - veća količina proizvedenih čipova -> manji troškovi po komadu (duplo veća količina ~ 10% manji troškovi)
  - Standardne komponente (DRAM, Flash, monitori, tastature, ...) - proizvodi ih više proizvođača, veća je konkurencija, manja je razlika između troškova i cene, a i sami troškovi su manji

- Ekonomski faktori integrisanih kola
  - Troškovi proizvodnje integrisanih kola
    - Značajan deo troškova za prenosne uređaje zbog SoC pristupa
    - Značajno kod klastera zbog broja čipova
    - Trošak za proizvodnju jednog integrisanog kola je srazmeran sumi troškova proizvodnje silicijumske pločice, testiranja, pakovanja i finalnog testiranja, a obrnuto srazmeran ukupnom procentu ispravnih čipova
    - Broj čipova po vaferu zavisi od dimenzija pojedinačne pločice i dimenzija samog vafera
    - Procenat ispravnih pločica na vaferu zavisi od raspodele defekata
      - za 2017-tu, za 28nm proces 0.012-0.016 defekata po cm^2, a za 16nm proces 0.016-0.047
      - Ovo je i tekući Intelov problem za 10nm tehnologiju

- Ekonomski faktori integrisanih kola
  - Troškovi proizvodnje integrisanih kola
    - Procenat ispravnih zavisi i od veličine pojedinačne pločice
      - Većina mikroporcesora zauzima 1-2.25cm^2
      - embedded mikroprocesori 0.05cm^2, a kontroleri i manje od 0.01cm^2
      - GPU i do 8cm^2
    - Uvođenje redundanse u čipove -> povećanje yield-a
      - SRAM/DRAM imaju dodatne memorijske ćelije
      - GPU imaju dodatna jezgra (ponekad i procesori)
    - U 2017 za 300mm vafer, 28nm tehnologija košta 4-5000$, 16nm - 7000$
      - za 16nm, 1cm^2 čip bi bio 16$, a 2.25cm^2 čip bi bio 58$
    - Za manje obimnu proizvodnju (manje od milion komada), cena izrade maski za osvetljavanje postaje značajan faktor
      - Za današnje proizvodne procese, za 10 slojeva na silicijumu, maske su 4 miliona $ za 16nm, a 1.5 miliona $ za 28nm
      - Mogućnost stavljanja različitih čipova na isti vafer

- Ekonomski faktori integrisanih kola
  - Cena i troškovi proizvodnje
    - Razlika između troškova i cene se smanjuje
      - Koristi se za R&D, marketing, održavanje, kancelarije, itd.
    - U cenu računara se već neko vreme mora uračunati i cena korišćenja (napajanje, hlađenje, ...), naročito kod sistema sa više hiljada i desetina hiljada jedinica
      - Green TOP 500 - rangiranje po energetskoj efikasnosti, jun 2018:

        Shoubu system B - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc., Advanced Center for Computing and Communication, RIKEN, Japan

        Suiren2 - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc., High Energy Accelerator Research Organization /KEK, Japan

        Sakura - ZettaScaler-2.2, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc., PEZY Computing K.K., Japan

        DGX SaturnV Volta - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , NVIDIA Corporation, United States

        Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory, United States

- Pouzdanost integrisanih kola
  - Dugo vremena, integrisana kola su bila najpouzdaniji deo računara - kvarovi u toku rada na samim čipovima su bili veoma retki
    - Sa prelaskom na 16nm i manje dimenzije, to polako prestaje da bude tako
    - Pitanje je i na kom nivou detalja je kvar - nije isto ako je otkazalo celo jezgro, ili ako je otkazao bit memorije

- Pouzdanost integrisanih kola
  - Kvantifikovanje pouzdanosti
    - Pouzdanost modula
      - MTTF - mean time to failure, srednje vreme ispravnog rada između dva kvara
      - MTTR - mean time to repair, srednje vreme popravke
      - MTBF - mean time between failures = MTTF+MTTR, najčešće korišetna mera pouzdanosti
    - Dostupnost modula
      - MTTF/MTBF
    - Što više modula sistem ima, veća je šansa da sistem kao celina bude u kvaru (verovatnoće pojedinih modula se sabiraju)
    - I ovo se može popraviti uvođenjem redundanse
      - višestrukim ponavljanjem operacija (npr. više pokušaja čitanja sektora diska)
      - višestrukim komponentama za istu funkciju (npr. diskovi u nekom od RAID režima sa duplikacijom)

- Performanse računara
  - Računar X je brži od računara Y - šta to znači?
    - Za korisnika telefona je bitno da se nešto izvrši za što kraće vreme (vreme odziva, vreme izvršavanja)
    - Za korisnika klastera je bitno da propusna moć bude što veća (količina posla u jedinici vremena)
  - Smatraćemo da je brži ≡ kraće vreme izvršavanja

$$n = \frac{Vreme\ izvršavanja\ Y}{Vreme\ izvršavanja\ X} = \frac{Performanse\ X}{Performanse\ Y}$$

    - X je n puta brži od Y
  - Postoje i drugačije mere performansi
  - Najjednostavnija mera vremena je ukupno proteklo vreme za zadatak, od startovanja do završetka, koje uključuje pristup diskovima, memoriji, overhead OS-a, ukratko - sve

- Performanse računara
  - Kada imamo multiprogramski pristup, može se koristiti i procesorsko vreme - vreme koje se stvarno troši na zadatak, a koje isključuje sve ostale aktivnosti procesora (međutim, ono što korisnik vidi je ukupno vreme)
  - Idealna mera bi bio korisnik koji najčešće koristi isti set programa (workload) - postavi se za drugi računar sa identičnom instalacijom i on oceni da li je drugi računar brži ili sporiji
    - Što je prilično teško naći i utvrditi...

- Performanse računara
  - Benchmark programi
    - Idealni su oni koji su realni (stvarno korišteni) programi
    - Manje idealni su oni koji su “pojednostavljenje” realnih programa:
      - Jezgra - mali delovi realnih programa
      - Sintetički benchmark testovi - “lažni” programi, napravljeni da oponašaju izvršavanje realnih programa - npr. Dhrystone
      - Uglavnom su diskreditovani (ali i dalje često korišteni)
      - Intel je ove godine imao par kontroverznih rezultata testiranja:

        - Juna 2018 su uživo pokazani impresivni test rezultati novog 28-jezgarnog procesora...
          - ... jedino što su “zaboravili” da pomenu da je procesor bio overklokovan na 5GHz, uz korićenje snažnog industrijskog hlađenja
        - Oktobra 2018 su objavljeni uporedni rezultati AMD i Intel procesora koji su značajno bili na strani Intel-a. Ispostavilo se da su sintetički benčmarci imali ogroman udeo tačno onih instrukcija koje Intel izvršava brže, a mnogo manje ostalih
      - Dieselgate, Volkswagen 2015-2016