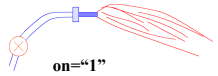
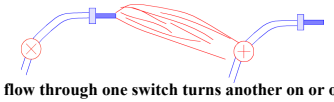
Lecture 28
Introduction to Parallel Processing

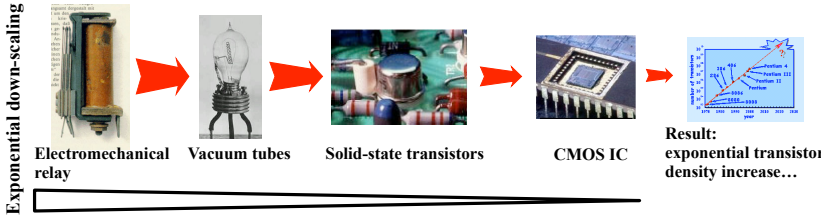
A little history... Zuse's paradigm

- Konrad Zuse (1938) Z3 machine
  - Use binary numbers to encode information
  - Represent binary digits as on/off state of a current switch





The flow through one switch turns another on or off.



A little history... programs

- Stored program model has been around for a long time...

First Draft of a Report
on the EDVAC

by

John von Neumann

Contract No. W–670–ORD–4926

Between the

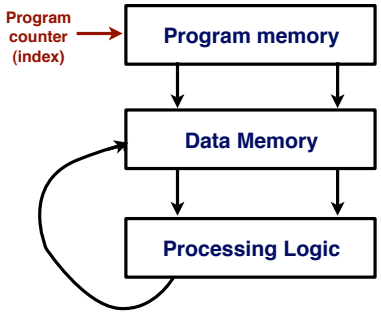
United States Army Ordnance Department

and the

University of Pennsylvania

Moore School of Electrical Engineering
University of Pennsylvania

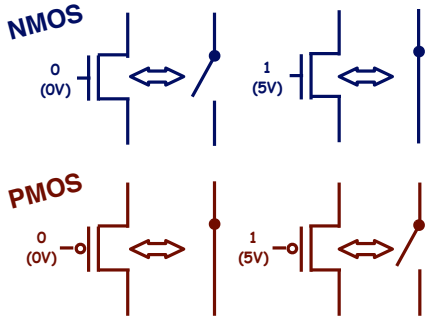
June 30, 1945

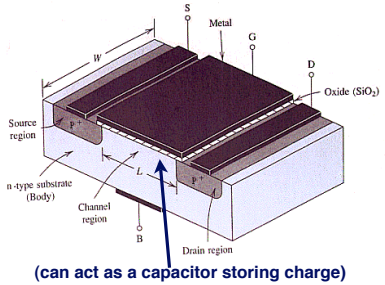


Transistors used to manipulate/store 1s & 0s

Switch-level representation

Cross-sectional view





(can act as a capacitor storing charge)

Using above diagrams as context, note that (with NMOS) if we (i) apply a suitable voltage to the gate & (ii) then apply a suitable voltage between source and drain, current will flow.

Moore's Law

- "Cramming more components onto integrated circuits."

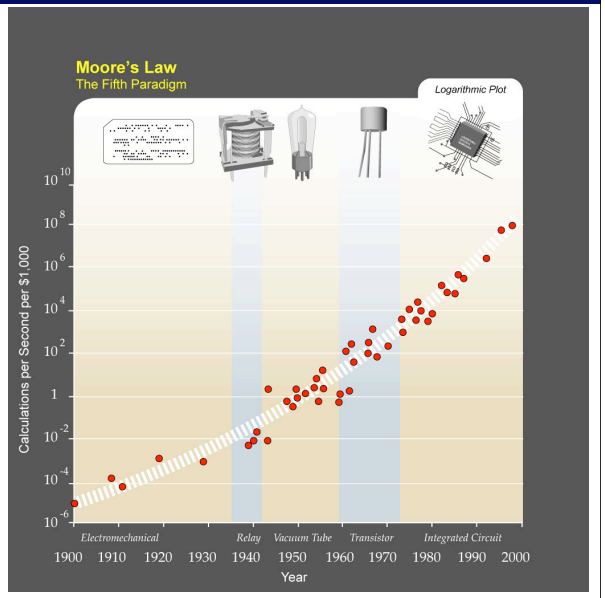
  - G.E. Moore, Electronics 1965

- Observation: DRAM transistor density doubles annually
  - Became known as "Moore's Law"
  - Actually, a bit off:
    - Density doubles every 18 months
    - (in 1965 they only had 4 data points!)

- Corollaries:
  - Cost per transistor halves annually (18 months)
  - Power per transistor decreases with scaling
  - Speed increases with scaling
  - Reliability increases with scaling
    - Of course, it depends on how small you try to make things
      - » (I.e. no exponential lasts forever)

Remember these!

Moore's Law



Moore's Law

- Moore's Curve is a self-fulfilling prophecy
  - 2X every 2 years means ~3% per month
    - I.e. ((1 X 1.03) * 1.03)*1.03… 24 times = ~2
  - Can use 3% per month to judge performance features
  - If feature adds 9 months to schedule…it should add at least 30% to performance
    - ($1.03^9 = 1.30$ ⇨ 30%)

A bit on device performance...

- One way to think about switching time:
  - Charge is carried by electrons
  - Carrier velocity is proportional to the lateral E-field between source and drain
    - i.e. $v = mE$
      - m = carrier mobility (and can be though of as a constant)
  - Electric field defined as: $E = V_{ds}/L$
  - Time for charge to cross channel = length/speed
    - (i.e. meters / (meters/s) = seconds)
    - $= L/v$
    - $= L/(mE)$
    - $= L/(m*(V_{ds}/L))$
    - $= L^2/(mV_{ds})$

Thus, to make a device faster, we want to either increase V_{ds} or decrease feature sizes (i.e. L)

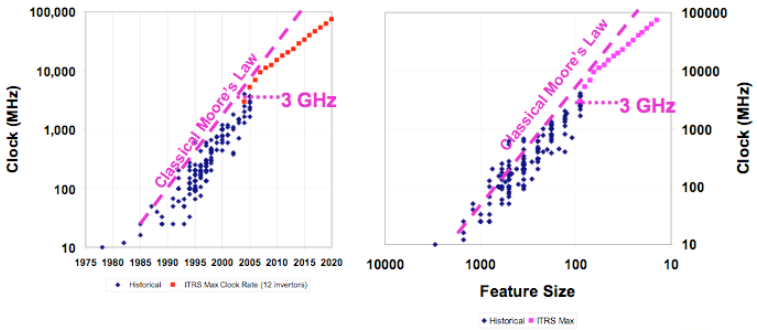
Some more important relationships

- What about power (i.e. heat)?
  - First, need to quickly discuss equation for capacitance:
    - $C_L = (e_{ox}WL)/d$
      - e_{ox} = dielectric, WL = parallel plate area, d = distance between gate and substrate
  - Then, dynamic power becomes:
    - $P_{dyn} = C_L V_{dd}^2 f_{0\text{-}1}$
      - Dynamic power is a function of the frequency of 0 to 1 or 1 to 0 transitions (as this involves the movement of charge)
        - » Note frequency in this context is NOT clock frequency
      - Note that as W and L scale, C_L decreases which in turn will cause a decrease in P_{dyn}.
      - Note that while an increase in V_{dd} will *decrease* switching time, it will also cause a quadratic *increase* in dynamic power.

A funny thing happened on the way to 45 nm

•Speed increases with scaling...

Remember these!

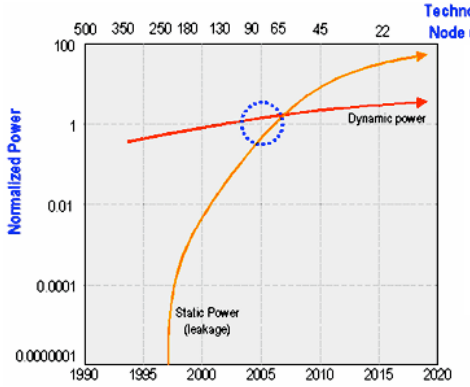


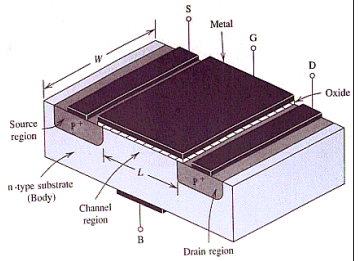
2005 projection was for 5.2 GHz - and we didn't make it in production. Further, we're still stuck at 3+ GHz in production.

A funny thing happened on the way to 45 nm

•Power decreases with scaling...

Remember these!





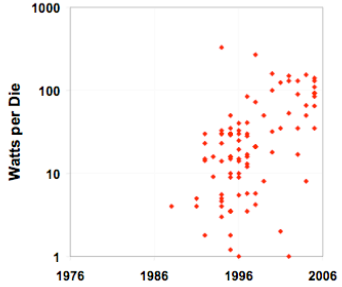
As transistors get smaller, they're harder to turn off.

A funny thing happened on the way to 45 nm

- Speed increases with scaling...
- Power decreases with scaling...

Remember these!

Why the clock flattening? POWER!!!!

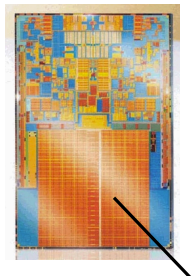


Recall...

- $P_{dyn} = C_L V_{dd}^2 f_{0-1}$
  - Dynamic power is a function of the frequency of 0 to 1 or 1 to 0 transitions (as this involves the movement of charge)
    - » Note frequency in this context is NOT clock frequency
  -

Other reasons too, but this should give you a good feel for technology...

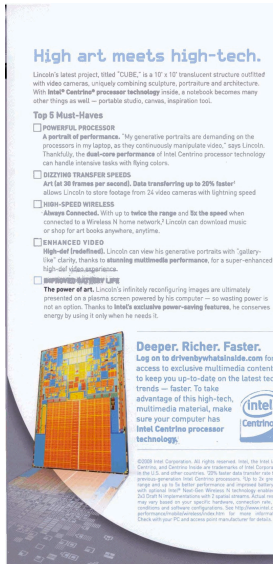
Transistors used for memory too...



Saw earlier, that transistors used for *on chip* memory too...

Problem: Program, data = bigger + power problem isn't going away...

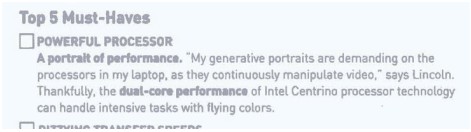
- **Why? Put faster memory closer to processing logic...**
  - **SRAM (logic): density +25%, speed +20%**
  - **DRAM (memory): density +60%, speed +4%**
  - **Disk (magnetic): density +25%, speed +4%**



Solution?

Motivation:
Processor complexity is good enough, transistor sizes scale, we can slow processors down, manage power, and get performance from...

Parallelism



(i.e. 1 processor, 1 ns clock cycle vs. 2 processors, 2 ns clock cycle)