Digital VLSI design

# Lecture 18: Scaling

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
**DELHI**

# Mid Sem Performance

- ECE 314:
  - Average: 53.21 (out of 100)
  - Highest: 85
  - Lowest: 09

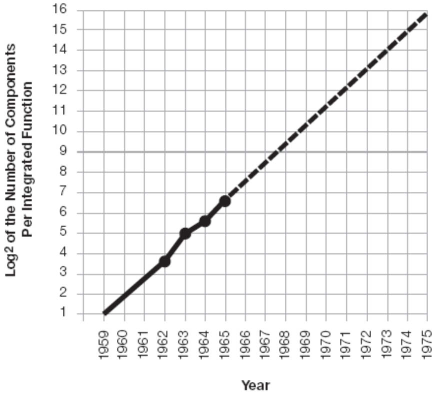
- ECE 514:
  - Average: 79.87 (out of 125)
  - Highest: 115
  - Lowest: 58

# Outline

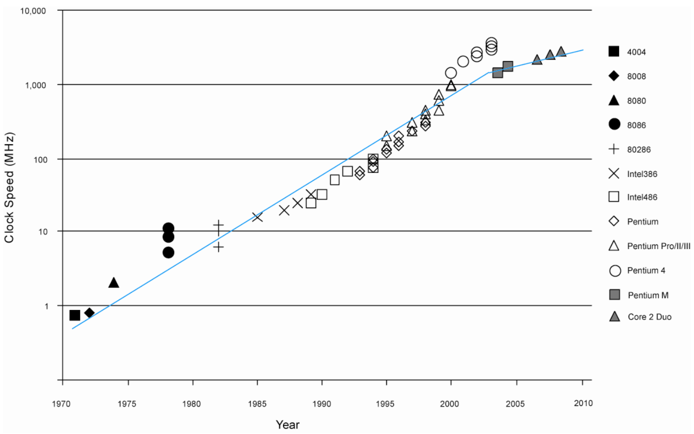
- Scaling
  - Transistors
  - Interconnect
  - Future Challenges
- Economics

# Moore’s Law

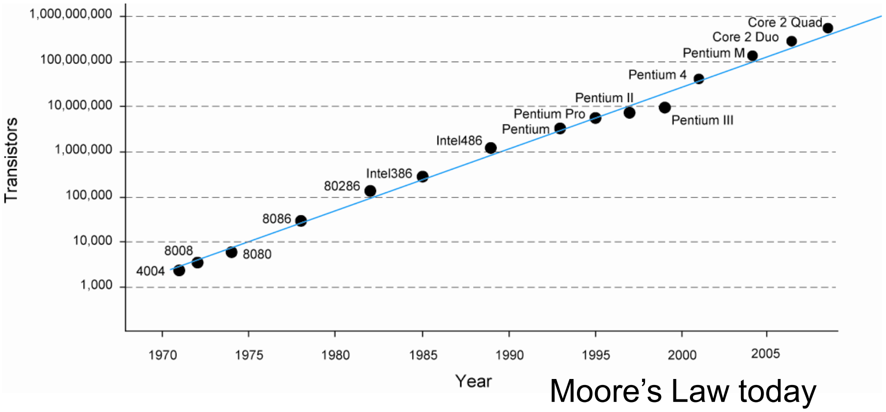
- Recall that Moore’s Law has been driving CMOS



[Moore65]



Corollary: clock speeds have improved



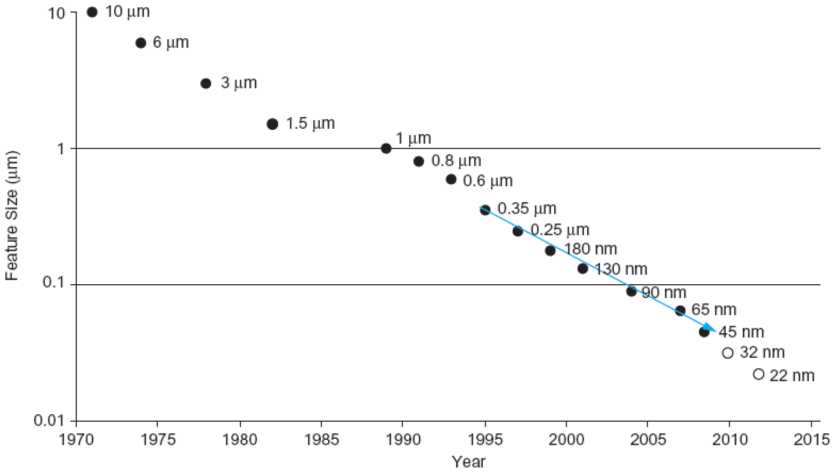
Moore’s Law today

# Why?

- Why more transistors per IC?
  - Smaller transistors
  - Larger dice
- Why faster computers?
  - Smaller, faster transistors
  - Better microarchitecture (more IPC)
  - Fewer gate delays per cycle

# Scaling

- The only constant in VLSI is constant change
- Feature size shrinks by 30% every 2-3 years
  - Transistors become cheaper
  - Transistors become faster and lower power
  - Wires do not improve (and may get worse)
- Scale factor S
  - Typically $S = \sqrt{2}$
  - Technology nodes



# Dennard Scaling

- Proposed by Dennard in 1974
- Also known as *constant field* scaling
  - Electric fields remain the same as features scale
- Scaling assumptions
  - All dimensions (x, y, z => W, L, $t_{ox}$)
  - Voltage ($V_{DD}$)
  - Doping levels

# Device Scaling

| Parameter | Sensitivity | Dennard Scaling |
|---|---|---|
| L: Length | | 1/S |
| W: Width | | 1/S |
| $t_{ox}$: gate oxide thickness | | 1/S |
| $V_{DD}$: supply voltage | | 1/S |
| $V_t$: threshold voltage | | 1/S |
| NA: substrate doping | | S |
| $\beta$ | $W/(Lt_{ox})$ | S |
| $I_{on}$: ON current | $\beta(V_{DD}-V_t)^2$ | 1/S |
| R: effective resistance | $V_{DD}/I_{on}$ | 1 |
| C: gate capacitance | $WL/t_{ox}$ | 1/S |
| $\tau$: gate delay | RC | 1/S |
| f: clock frequency | $1/\tau$ | S |
| E: switching energy / gate | $CV_{DD}^2$ | $1/S^3$ |
| P: switching power / gate | Ef | $1/S^2$ |
| A: area per gate | WL | $1/S^2$ |
| Switching power density | P/A | 1 |
| Switching current density | $I_{on}/A$ | S |

# Observations

- Gates get faster with scaling (good)
- Dynamic power goes down with scaling (good)
- Current density goes up with scaling (bad)

# Example

- Gate capacitance is typically about 1 fF/$\mu$m
- The typical FO4 inverter delay for a process of feature size *f* (in nm) is about 0.5*f* ps
- Estimate the ON resistance of a unit (4/2 $\lambda$) transistor.

# Real Scaling

- $t_{ox}$ scaling has slowed since 65 nm
  - Limited by gate tunneling current
  - Gates are only about 4 atomic layers thick!
  - High-k dielectrics have helped continued scaling of effective oxide thickness
- $V_{DD}$ scaling has slowed since 65 nm
  - SRAM cell stability at low voltage is challenging
- Dennard scaling predicts cost, speed, power all improve
  - Below 65 nm, some designers find they must choose just two of the three

# Wire Scaling

- Wire cross-section
  - w, s, t all scale
- Wire length
  - Local / scaled interconnect
  - Global interconnect
    - Die size scaled by $D_c \approx 1.1$

# Interconnect Scaling

| Parameter | Sensitivity | Scale Factor |
|---|---|---|
| w: width | | 1/S |
| s: spacing | | 1/S |
| t: thickness | | 1/S |
| h: height | | 1/S |
| $D_c$: die size | | $D_c$ |
| $R_w$: wire resistance/unit length | 1/wt | $S^2$ |
| $C_{wf}$: fringing capacitance / unit length | t/s | 1 |
| $C_{wp}$: parallel plate capacitance / unit length | w/h | 1 |
| $C_w$: total wire capacitance / unit length | $C_{wf} + C_{wp}$ | 1 |
| $t_{wu}$: unrepeated RC delay / unit length | $R_wC_w$ | $S^2$ |
| $t_{wr}$: repeated RC delay / unit length | sqrt($RCR_wC_w$) | sqrt(S) |
| Crosstalk noise | w/h | 1 |
| $E_w$: energy per bit / unit length | $C_wV_{DD}^2$ | $1/S^2$ |

# Interconnect Delay

| Parameter | Sensitivity | Local / Semiglobal | Global |
|---|---|---|---|
| l: length | | 1/S | $D_c$ |
| Unrepeated wire RC delay | $l^2t_{wu}$ | 1 | $S^2D_c^2$ |
| Repeated wire delay | $lt_{wr}$ | sqrt(1/S) | $D_c$sqrt(S) |
| Energy per bit | $lE_w$ | $1/S^3$ | $D^c/S^2$ |

# Observations

- Capacitance per micron is remaining constant
  - About 0.2 fF/μm
  - Roughly 1/5 of gate capacitance
- Local wires are getting faster
  - Not quite tracking transistor improvement
  - But not a major problem
- Global wires are getting slower
  - No longer possible to cross chip in one cycle

# Quiz V

- **A 32 bit off-chip bus operating at 1.5V and 1GHz clock rate is driving a capacitance of 3pF/bit. Each bit is estimated to have a toggling probability of 0.25 at each clock cycle. What is the power dissipation in operating the bus?**