Advanced VLSI Design : 2021-22
Lecture 14-B
CMOS Scaling

By Dr. Sanjay Vidhyadharan

Technology Scaling

- Currently, technology scaling has a threefold objective:
  - Reduce the gate delay by 30% (43% increase in frequency)
  - Double the transistor density
  - Saving 50% of power (at 43% increase in frequency)
- How is scaling achieved?
  - All the device dimensions (lateral and vertical) are reduced by $1/\alpha$
  - Concentration densities are increased by α
  - Device voltages reduced by $1/\alpha$ (not in all scaling methods)
  - Typically $1/\alpha = 0.7$ (30% reduction in the dimensions)

Technology Scaling

- The scaling variables are:
  - Supply voltage: $V_{dd} \rightarrow V_{dd} / \alpha$
  - Gate length: $L \rightarrow L / \alpha$
  - Gate width: $W \rightarrow W / \alpha$
  - Gate-oxide thickness: $t_{ox} \rightarrow t_{ox} / \alpha$
  - Junction depth: $X_j \rightarrow X_j / \alpha$
  - Substrate doping: $N_A \rightarrow N_A \times \alpha$

This is called **constant field** scaling because the electric field across the gate-oxide does not change when the technology is scaled

If the power supply voltage is maintained constant the scaling is called **constant voltage**. In this case, the electric field across the gate-oxide increases as the technology is scaled down.

Due to gate-oxide breakdown, below 0.8µm only "constant field" scaling is used.

Technology Scaling

Some consequences 30% scaling in the constant field regime ($\alpha = 1.43$, $1/\alpha = 0.7$):

- Device/die area:

$$W \times L \rightarrow (1/\alpha)^2 = 0.49$$

  - In practice, microprocessor die size grows about 25% per technology generation! This is a result of added functionality.

- Transistor density:

$$(\text{unit area}) / (W \times L) \rightarrow \alpha^2 = 2.04$$

  - In practice, memory density has been scaling as expected. (not true for microprocessors…)

Technology Scaling

- Gate capacitance:

$$W \times L / t_{ox} \rightarrow 1/\alpha = 0.7$$

- Drain current:

$$(W/L) \times (V^2/t_{ox}) \rightarrow 1/\alpha = 0.7$$

- Gate delay:

$$(C \times V) / I \rightarrow 1/\alpha = 0.7$$

$$\text{Frequency} \rightarrow \alpha = 1.43$$

  - In practice, microprocessor frequency has doubled every technology generation (2 to 3 years)! This faster increase rate is due to two factors:
    - the number of gate delays in a clock cycle decreases with time (the designs become highly pipelined)
    - advanced circuit techniques reduce the average gate delay beyond 30% per generation.

Technology Scaling

- Power:

$$C \times V^2 \times f \rightarrow (1/\alpha)^2 = 0.49$$

- Power density:

$$1/t_{ox} \times V^2 \times f \rightarrow 1$$

- Active capacitance/unit-area:

•

Power dissipation is a function of the operation frequency, the power supply voltage and of the circuit size (number of devices). If we normalize the power density to $V^2 \times f$ we obtain the active capacitance per unit area for a given circuit. This parameter can be compared with the oxide capacitance per unit area:

$$1/t_{ox} \rightarrow \alpha = 1.43$$

  - In practice, for microprocessors, the active capacitance/unit-area only increases between 30% and 35%. Thus, the twofold improvement in logic density between technologies is not achieved.

Technology Scaling

- Interconnects scaling:
  - Higher densities are only possible if the interconnects also scale.
  - Reduced width → increased resistance
  - Denser interconnects → higher capacitance
  - To account for increased parasitics and integration complexity more interconnection layers are added:
    - thinner and tighter layers → local interconnections
    - thicker and sparser layers → global interconnections and power

  Interconnects are scaling as expected

Technology Scaling

Quantity	Sensitivity	Constant Field	Constant Voltage
Scaling Parameters			
Length	L	$1/S$	$1/S$
Width	W	$1/S$	$1/S$
Gate Oxide Thickness	t_{ox}	$1/S$	$1/S$
Supply Voltage	V_{dd}	$1/S$	1
Threshold Voltage	V_{T0}	$1/S$	1
Doping Density	N_A, N_D	S	S^2
Device Characteristics			
Area (A)	WL	$1/S^2$	$1/S^2$
β	W/Lt_{ox}	S	S
D-S Current (I_{DS})	$\beta(V_{dd} - v_T)^2$	$1/S$	S
Gate Capacitance (C_g)	WL/t_{ox}	$1/S$	$1/S$
Transistor On-Resistance (R_{tr})	V_{dd}/I_{DS}	1	S
Intrinsic Gate Delay (τ)	$R_{tr}C_g$	$1/S$	$1/S$
Clock Frequency	f	f	f
Power Dissipation (P)	$I_{DS}V_{dd}$	$1/S^2$	S
Power Dissipation Density (P/A)	P/A	1	S^3

Technology Scaling

Lithography:

Optics technology	Technology node
248nm mercury-xenon lamp	180 - 250nm
248nm krypton-fluoride laser	130 - 180nm
193nm argon-fluoride laser	100 - 130nm
157nm fluorine laser	70 - 100nm
13.4nm extreme UV	50 - 70nm

Technology Scaling

Lithography:

- Electron Beam Lithography (EBL)
  - Patterns are derived directly from digital data
  - The process can be direct: no masks
  - Pattern changes can be implemented quickly
  - However:
    - Equipment cost is high
    - Large amount of time required to access all the points on the wafer

Thank you