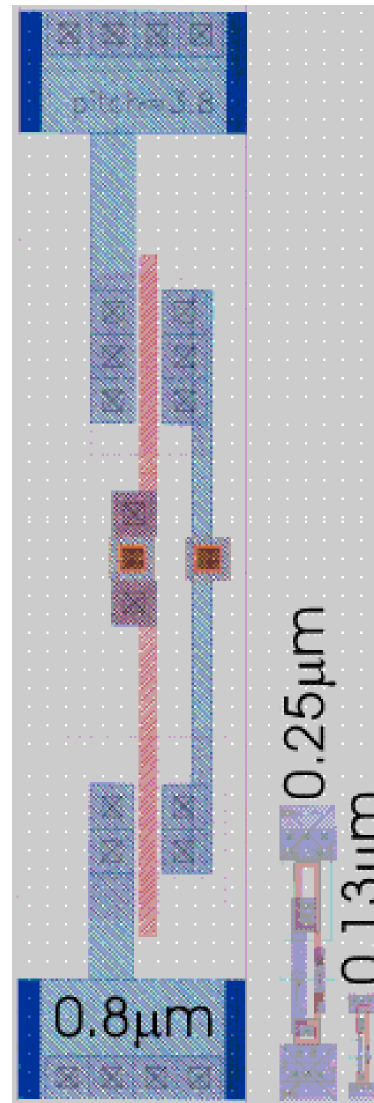
Outline

- Introduction – *"Is there a limit?"*
- Transistors – *"CMOS building blocks"*
- Parasitics I – *"The [un]desirables"*
- Parasitics II – *"Building a full MOS model"*
- The CMOS inverter – *"A masterpiece"*
- Technology scaling – *"Smaller, Faster and Cooler"*
- Technology – *"Building an inverter"*
- Gates I – *"Just like LEGO"*
- The pass gate – *"An useful complement"*
- Gates II – "A portfolio"
- Sequential circuits – *"Time also counts!"*
- DLLs and PLLs – " A brief introduction"
- Storage elements – *"A bit in memory"*

Technology scaling

- Scaling objectives
- Scaling variables
- Scaling consequences:
  - Device area
  - Transistor density
  - Gate capacitance
  - Drain current
  - Gate delay
  - Power
  - Power density
  - Interconnects

Scaling, why is it done?



Technology scaling

- Technology scaling has a threefold objective:
  - Increase the transistor density
  - Reduce the gate delay
  - Reduce the power consumption
- At present, between two technology generations, the objectives are:
  - Doubling of the transistor density;
  - Reduction of the gate delay by 30% (43% increase in frequency);
  - Reduction of the power by 50% (at 43% increase in frequency);

Technology scaling

- 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.

Scaling consequences

Some consequences of 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…)

Scaling consequences

- 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 highly pipelined architectures (“less gates per clock cycle”)

Scaling consequences

- 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.

Scaling consequences

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

Interconnects are scaling as expected

Scaling consequences

Parameter	Constant Field	Constant Voltage	
Supply voltage (V_{dd})	$1/\alpha$	1	Scaling Variables
Length (L)	$1/\alpha$	$1/\alpha$	
Width (W)	$1/\alpha$	$1/\alpha$	
Gate-oxide thickness (t_{ox})	$1/\alpha$	$1/\alpha$	
Junction depth (X_j)	$1/\alpha$	$1/\alpha$	
Substrate doping (N_A)	α	α	
Electric field across gate oxide (E)	1	α	Device Repercussion
Depletion layer thickness	$1/\alpha$	$1/\alpha$	
Gate area (Die area)	$1/\alpha^2$	$1/\alpha^2$	
Gate capacitance (load) (C)	$1/\alpha$	$1/\alpha$	
Drain-current (I_{dss})	$1/\alpha$	α	
Transconductance (g_m)	1	α	
Gate delay	$1/\alpha$	$1/\alpha^2$	Circuit Repercussion
Current density	α	α^3	
DC & Dynamic power dissipation	$1/\alpha^2$	α	
Power density	1	α^3	
Power-Delay product	$1/\alpha^3$	$1/\alpha$	