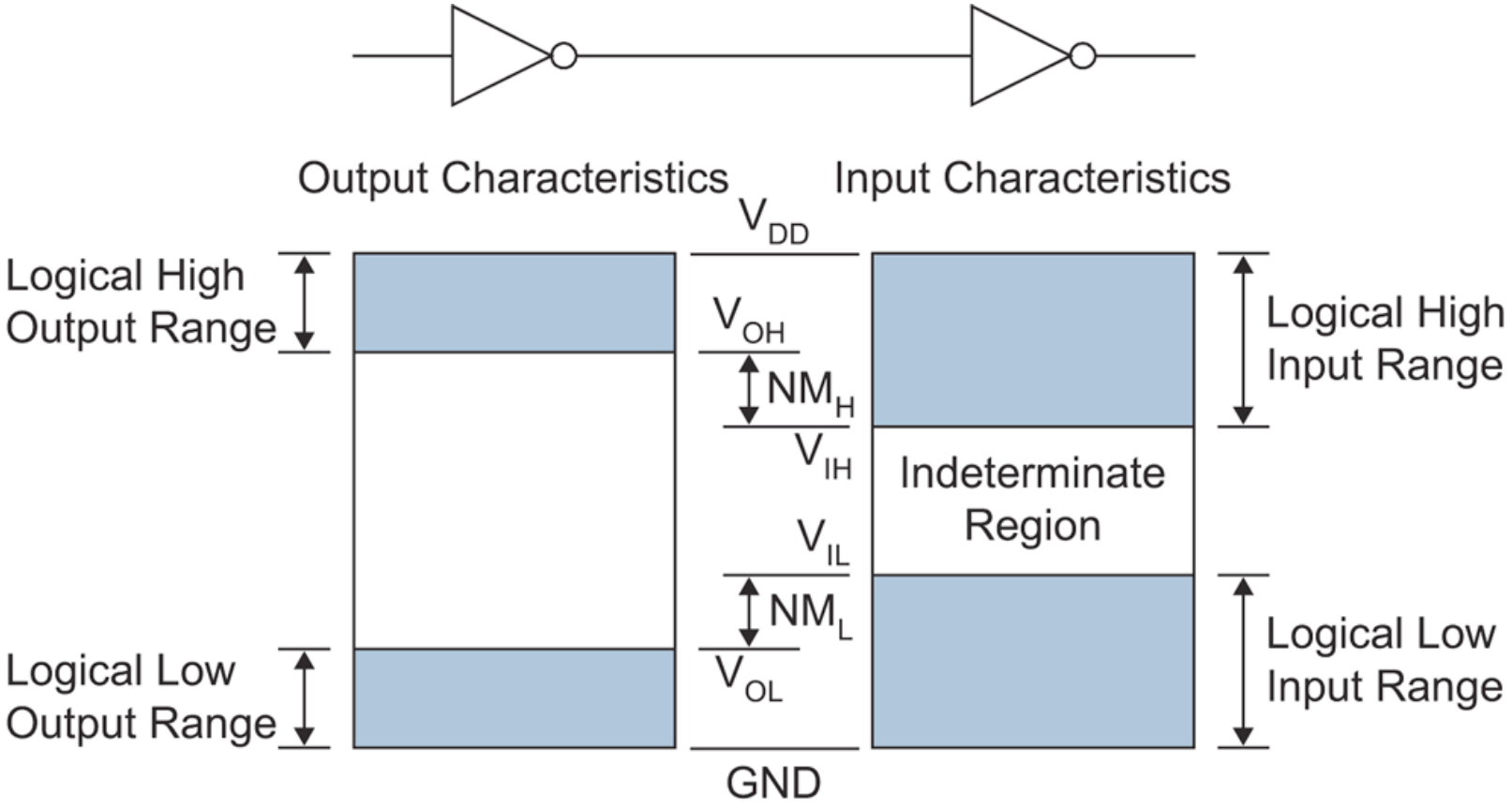
Transistor Sizing
and
Technology Scaling

Transistor Sizing
or
T – Sizing

Noise Margin

- **Noise margin**
  - is the amount of noise that a CMOS circuit could withstand without compromising the operation of circuit
  - does make sure that any signal which is logic '**1**'with finite noise added to it, is still recognized as logic '**1**' & not logic '**0**'
  - is a parameter closely related to the input-output voltage characteristics
  - allows us to determine the allowable noise voltage on the input of a gate so that the output will not be affected

Noise Margin



Noise Margin

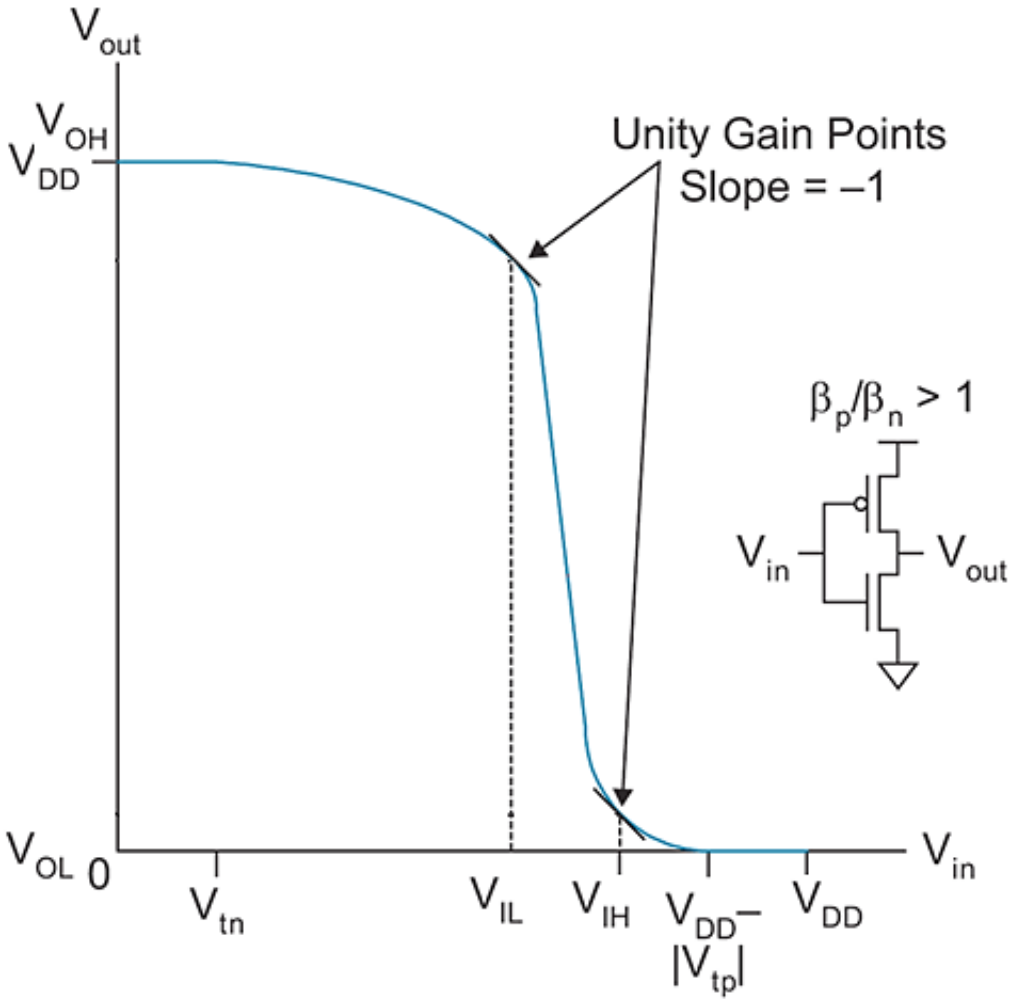
- Noise margin is specified in terms of 2 parameters:
  - LOW Noise Margin, $\mathbf{NM_L}$
  - HIGH Noise Margin, $\mathbf{NM_H}$

$$NM_L = |V_{ILmax} - V_{OLmax}|$$
$$NM_H = |V_{OHmin} - V_{IHmin}|$$

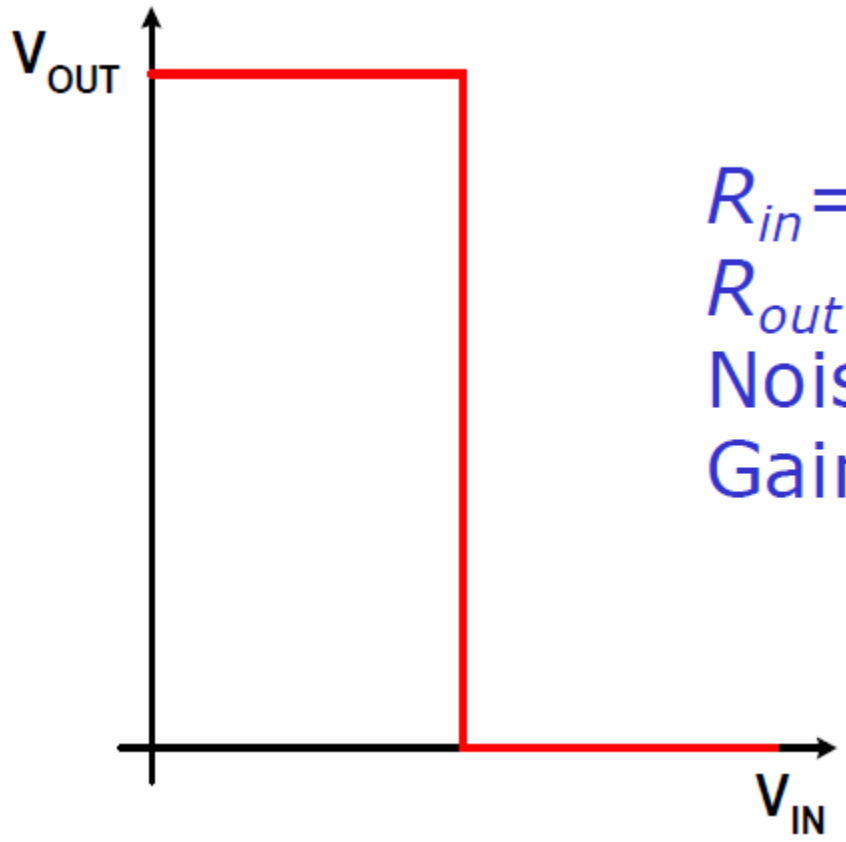
To optimize both noise margins NM_L and NM_H generally it is desirable to have $V_{IH} = V_{IL}$ & this to be a value that is midway in the logic swing, that is $V_{DD}/2$.

Or, $\mathbf{V_{inv,th} = V_{DD}/2}$

Noise Margin



The Ideal Gate



$R_{in}=\infty$

$R_{out}=0$

Noise Margin$=V_{DD}/2$

Gain $= \infty$

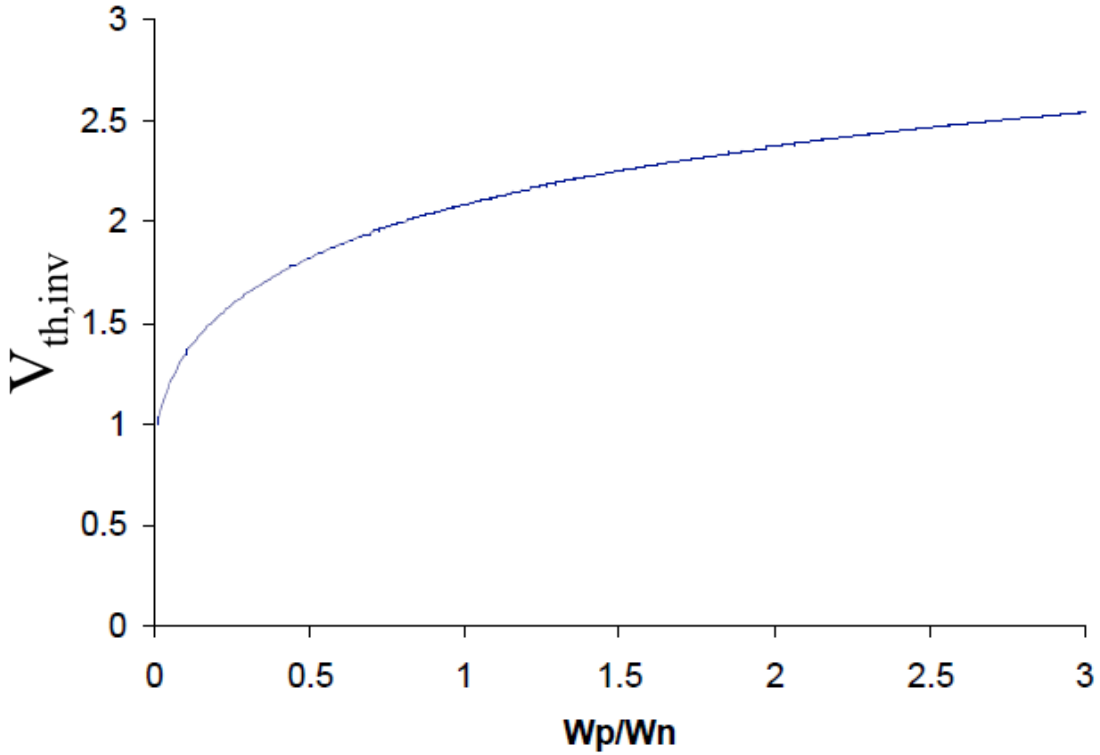
- We know that, for a CMOS inverter:

$$V_{inv} = \frac{V_{DD} + V_{tp} + V_{tn}\sqrt{\frac{\beta_n}{\beta_p}}}{1 + \sqrt{\frac{\beta_n}{\beta_p}}}$$

- To make $V_{th,inv} = V_{DD}/2$, $\beta_n = \beta_p \rightarrow \mu_n C_{ox} W_n/L_n = \mu_p C_{ox} W_p/L_p$
- But $\mu_n = 2.7\mu_p$
- Therefore, $\mathbf{W_p = 2.7W_n}$
- That is, PMOS is made 2.7 times wider than NMOS.

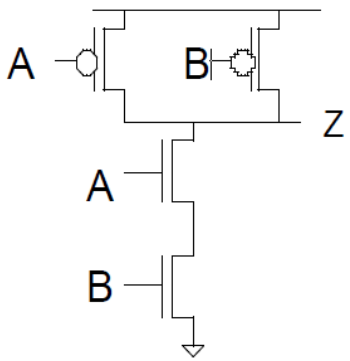
- This trick of making the PMOS larger to compensate for the lower hole mobility is universally used in silicon design.

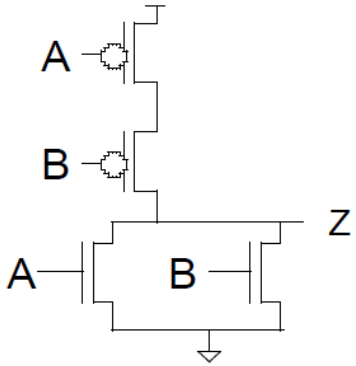
- However, making the PMOS 2.7 times larger than the NMOS leads to large circuits that occupy large areas of silicon for little extra benefit.



- If we examine Fig. above for V_{DD} = 5V, we see that the rate of change of $V_{th,inv}$ w.r.t. W_p/W_n is large for $W_p/W_n < 1$, but is smaller for $W_p/W_n > 1$. Because of the diminishing improvement in $V_{th,inv}$ for increasing W_p/W_n there is little reward for making W_p/W_n large. In practice, real designs use a compromise value for W_p/W_n – typically in the range 1.5 to 2.0
- The adroit selection of transistor dimensions is known as **T-sizing**.

T – Sizing Standard Gates





- In NAND gate above, there are two parallel PMOSs between V_{DD} and Z. If one PMOS is OFF and the other is ON, then the pull-up strength of the gate will be the same as that of an inverter with a similarly sized PMOS

- However, the NMOSs, both must be ON at once to pull down Z. There are two NMOSs, hence two ON resistances added in series between V_{SS} and Z. Hence, the pull-down strength of a NAND gate will be less than that of an inverter with a similarly sized NMOS. This problem is solved by T-sizing

Stacks and Stack Depth

- The N-block and P-blocks of CMOS gates are often referred to as ***stacks***
- Each gate can have one or more circuit paths connecting the o/p to one or other of the supply rails, & the number of transistors that appears in any path is referred to as the ***stack depth*.**

Examples:

- The NAND gate has an N-stack depth of 2 and a P-stack depth of 1
- The NOR gate has an N-stack depth of 1 and a P-stack depth of 2

- If we want to make a NAND gate have the same transfer characteristic as an inverter with W_{ip} and W_{in}, then the MOSFETs in the NAND gate will have $W_{np} = W_{ip}$ and $W_{nn} = 2W_{in}$

$$\mathbf{W_{np} = S_p W_{ip}} \quad \text{and} \quad \mathbf{W_{nn} = S_n W_{in}}$$

- where S_p is the P-stack depth and S_n is the N-stack depth

T – Sizing Compound Gates

- The method of T-sizing discussed so far is reliable and easy to apply for simple, single logic function gates, but is less straightforward to apply to compound gates.

Example:

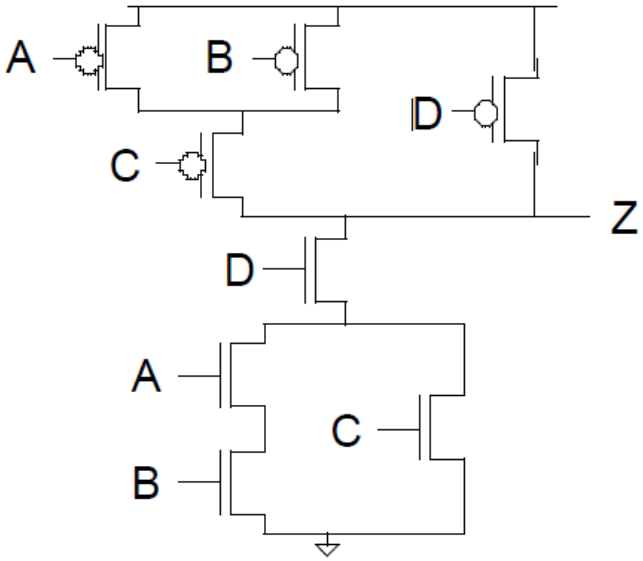
$$Z = \overline{(A + B).C + D}$$

PMOS paths	NMOS paths	Stack depth (S)	New W_p	New W_n

T – Sizing Compound Gates

- The method of T-sizing discussed so far is reliable and easy to use for simple, single logic function gates, but is less straightforward to apply to compound gates.

Example:



$$Z = \overline{(A.B + C).D}$$

The Problem

PMOS paths	NMOS paths	Stack depth (S)	New W_p	New W_n
A-C		2	$2W_{ip}$	
B-C		2	$2W_{ip}$	
D		1	W_{ip}	
	B-A-D	3		$3W_{in}$
	C-D	2		$2W_{in}$

- This procedure generates an anomaly – NMOS **D** is required to have two different sizes. This is clearly impossible to achieve.

The Solution

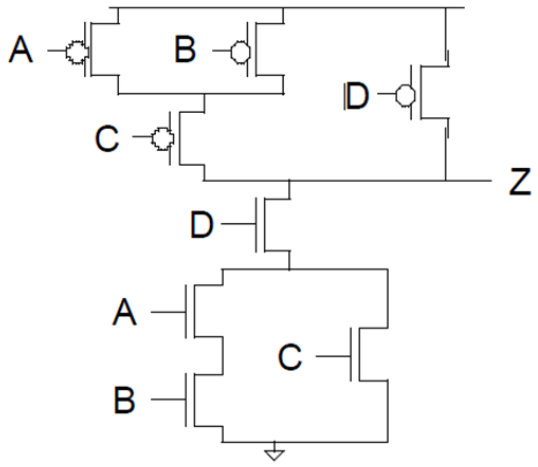
- A variety of solutions to this problem have been proposed, but the problem is not strictly tractable.

Here are a few:

1. T-size the deepest stack first.
2. T-size the least deep stack first

} Linear Scaling

3. Use sub-linear scaling – e.g. multiply by $\sqrt{S}$ instead of S



N- or PMOS (by input label)	Method 1 Linear – deepest stack first		Method 2 Linear – least deep stack first		Method 3 Sub-linear (√) – deepest stack first	
	New W_p	New W_n	New W_p	New W_n	New W_p	New W_n
A	$2W_{ip}$	$3W_{in}$	$2W_{ip}$	$4W_{in}$	$1.4W_{ip}$	$1.7W_{in}$
B	$2W_{ip}$	$3W_{in}$	$2W_{ip}$	$4W_{in}$	$1.4W_{ip}$	$1.7W_{in}$
C	$2W_{ip}$	$1.5W_{in}$	$2W_{ip}$	$2W_{in}$	$1.4W_{ip}$	$1.22W_{in}$
D	W_{ip}	$3W_{in}$	W_{ip}	$2W_{in}$	W_{ip}	$1.7W_{in}$

Assignments

- Design gates with the following logic functions and T-size them using all linear and sub-linear methods:

$$Z = \overline{(A.B + C.D).E}$$

$$Z = \overline{(A + B).C + D}$$

Technology Scaling

Typical Scaling Scenario

- 1974: 5μm Technology, V_{DD}= 10V
- 1984: 1μm Technology, V_{DD}= 5V
- 1994: 0.35μm Technology, V_{DD}= 3.5V
- 2004: 90nm Technology, V_{DD}= 1V
- 2014: 15nm Technology, V_{DD}= 0.8V
- 2018: 10nm Technology, V_{DD}= 0.5V

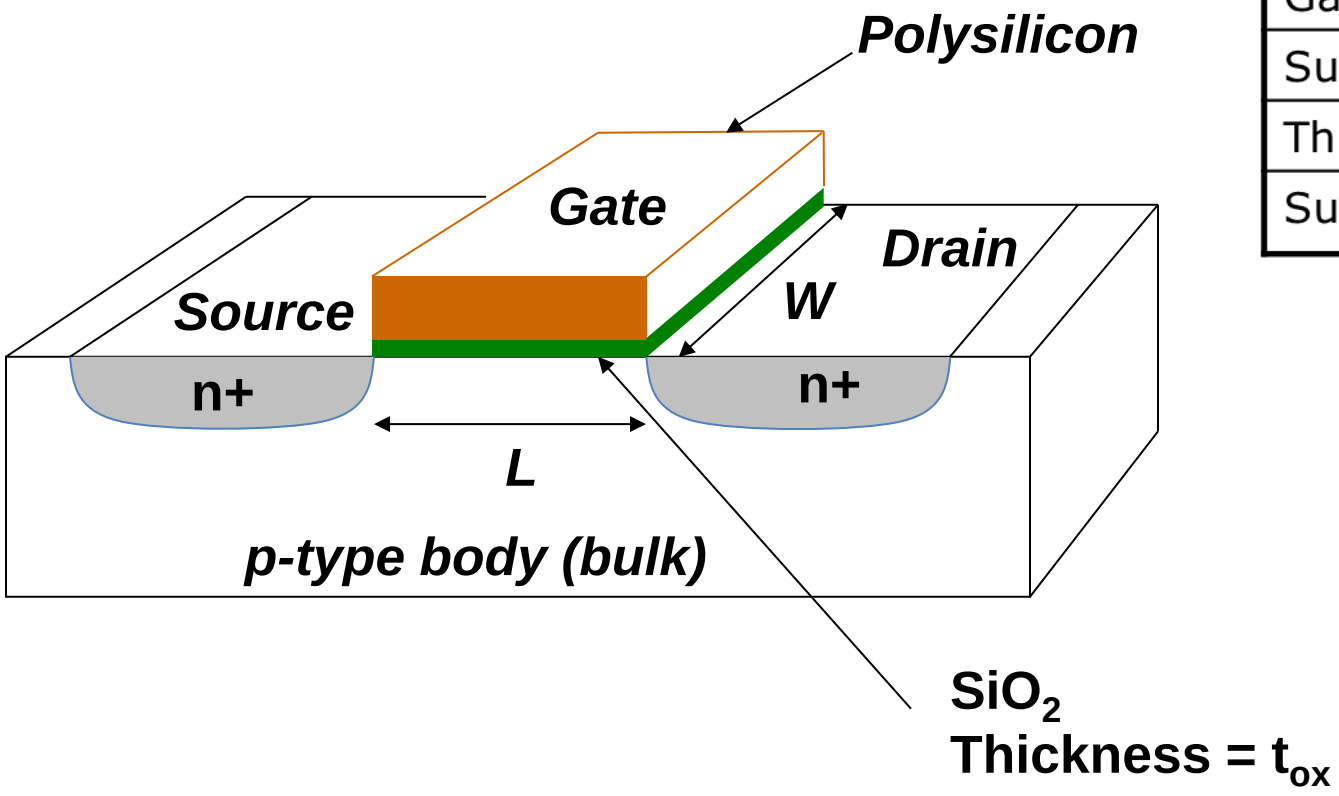
Technology Scaling

- 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% (at 43% increase in frequency)
  - Reduction of the power by 50% (at 43% increase in frequency)

Technology Scaling

Device Parameter
Length, L
Width, W
Gate oxide thickness, t_{ox}
Supply voltage, V_{DD}
Threshold voltages, V_{tn}, V_{tp}
Substrate doping, N_A

Bulk nMOSFET



Why odd numbers?

0.25µ, 0.18µ, 0.13µ, 90nm and 65nm technologies, continuing on to 15nm and 10nm, 7nm, 5nm

Technology Scaling

- A scaling factor (*S*) reduces device dimensions as 1/*S*.
- Successive generations of technology have used a scaling $\boldsymbol{S = \sqrt{2}}$, doubling the number of transistors per unit area. This produced 0.25µ, 0.18µ, 0.13µ, 90nm and 65nm technologies, continuing on to 7nm and 5nm …
- A 5% gate shrink (*S* = 1.05) is commonly applied to boost speed as the process matures.

N. H. E. Weste and D. Harris, *CMOS VLSI Design, Third Edition*, Boston: Pearson Addison-Wesley, 2005, Section 4.9.1.

Technology Scaling

1. Constant Field Scaling
2. Constant Voltage Scaling
3. Lateral Scaling

Constant Electric Field Scaling

Device Parameter	Scaling
Length, L	$1/S$
Width, W	$1/S$
Gate oxide thickness, t_{ox}	$1/S$
Supply voltage, V_{DD}	$1/S$
Threshold voltages, V_{tn}, V_{tp}	$1/S$
Substrate doping, N_A	S

Constant Electric Field Scaling (Cont.)

Device Characteristic		Scaling
β	$W / (L\ t_{ox})$	S
Current, I_{ds}	$\beta\ (V_{DD} - V_t)^2$	$1/S$
Resistance, R	V_{DD} / I_{ds}	1
Gate capacitance, C	$W\ L / t_{ox}$	$1/S$
Gate delay, τ	RC	$1/S$
Clock frequency, f	$1/\tau$	S
Dynamic power per gate, P	$CV^2 f$	$1/S^2$
Chip area, A		$1/S^2$
Power density	P/A	1
Current density	I_{ds}/A	S

Table 4.12 Influence of scaling on MOS device characteristics

Parameter	Sensitivity	Constant Field	Lateral	Constant Voltage
Scaling Parameters				
Length: L		$1/S$	$1/S$	$1/S$
Width: W		$1/S$	1	$1/S$
Gate oxide thickness: t_{ox}		$1/S$	1	$1/S$
Supply voltage: V_{DD}		$1/S$	1	1
Threshold voltage: V_{tn}, V_{tp}		$1/S$	1	$1/S$
Substrate doping: N_A		S	1	S
Device Characteristics				
β	$\frac{W}{L}\frac{1}{t_{ox}}$	S	S	S
Current: I_{ds}	$\beta(V_{DD} - V_t)^2$	$1/S$	S	$1/S$
Resistance: R	$\frac{V_{DD}}{I_{ds}}$	1	$1/S$	1
Gate capacitance: C	$\frac{WL}{t_{ox}}$	$1/S$	$1/S$	$1/S$
Gate delay: τ	RC	$1/S$	$1/S^2$	$1/S$
Clock frequency: f	$1/\tau$	S	S^2	S
Dynamic power dissipation (per gate): P	CV^2f	$1/S^2$	S	$1/S$
Chip area: A		$1/S^2$	1	$1/S^2$
Power density	P/A	1	S	1
Current density	I_{ds}/A	S	S	S