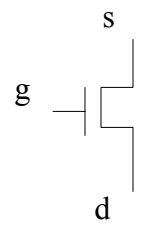
Principles of VLSI Review

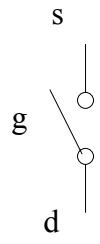
- Static CMOS Logic Design
- Delay Models
- Power Consumption
- CMOS Technology Scaling

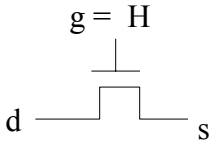
CMOS Transistors (N-type)

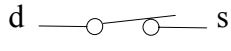


N-type (NMOS) transistor - can think of it as a switch.

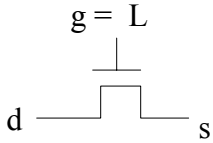
g: gate, d: drain, s: source

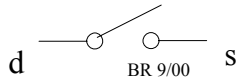






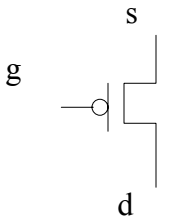
When g= H, d is connected to s (current flows between s, d because switch is closed.





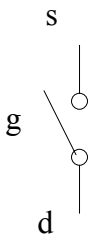
When g= L, d is disconnected from s (current does not flow between s, d because switch is open.

CMOS Transistors (P-type)

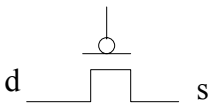


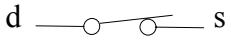
P-type (PMOS) transistor - can think of it as a switch.

g: gate, d: drain, s: source



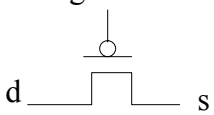
g = L

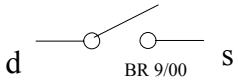




When g= L, d is connected to s (current flows between s, d because switch is closed.

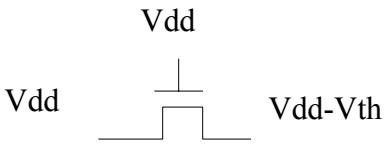
g = H

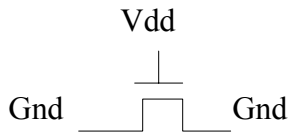


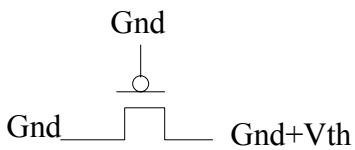


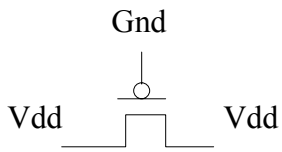
When g= H, d is disconnected from s (current does not flow between s, d because switch is open.

Reality





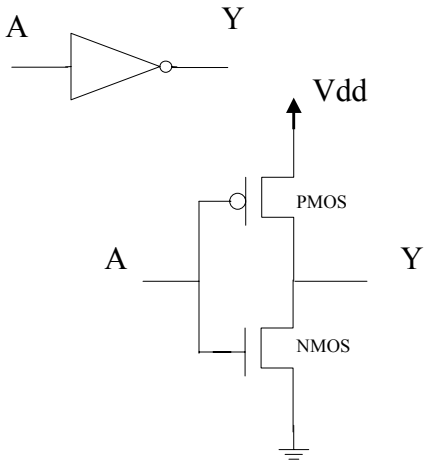




Vth = threshold voltage

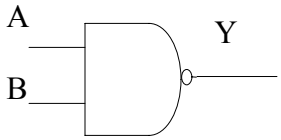
Want NMOS pulling to GND, PMOS pulling to Vdd

Inverter gate - takes 2 Transistors

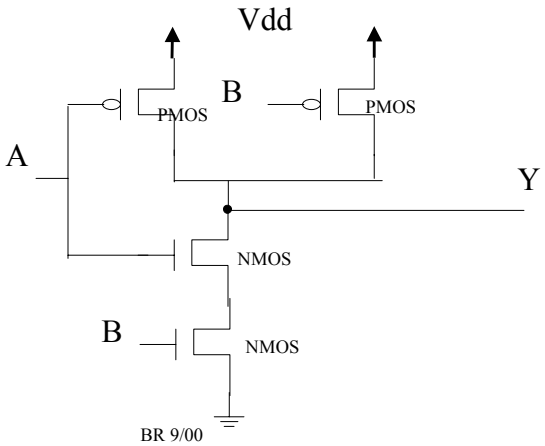


A	Y
L	H
H	L

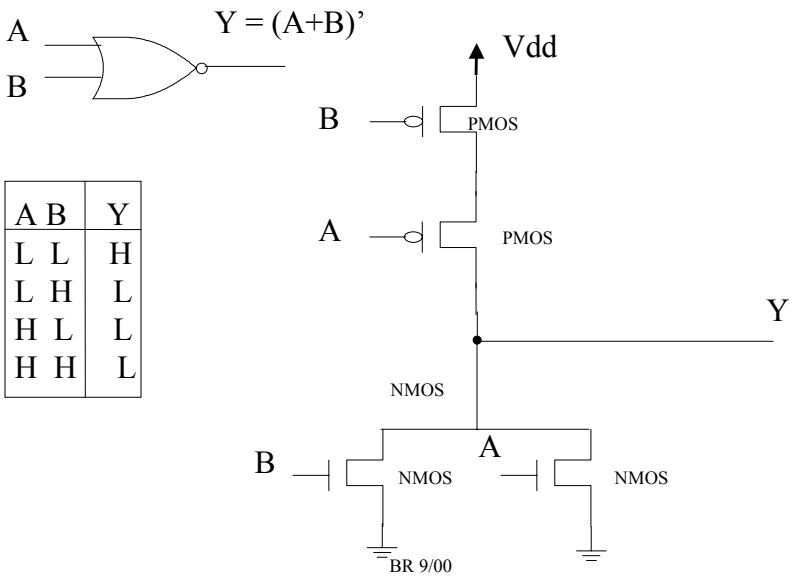
NAND gate - takes 4 Transistors



A	B	Y
L	L	H
L	H	H
H	L	H
H	H	L

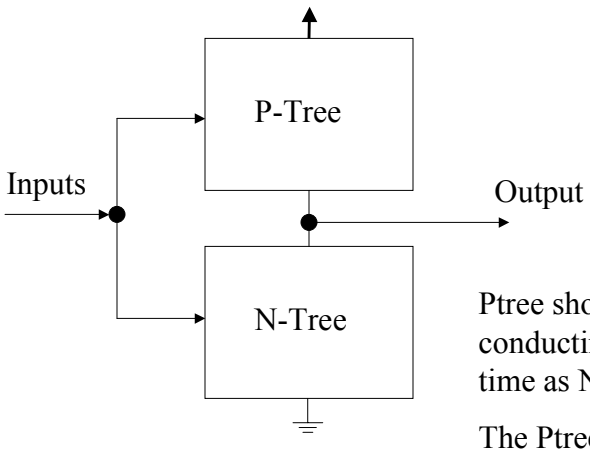


NOR gate - takes 4 Transistors



A	B	Y
L	L	H
L	H	L
H	L	L
H	H	L

General Static CMOS Gate



Ptree should never be conducting at the same time as Ntree.

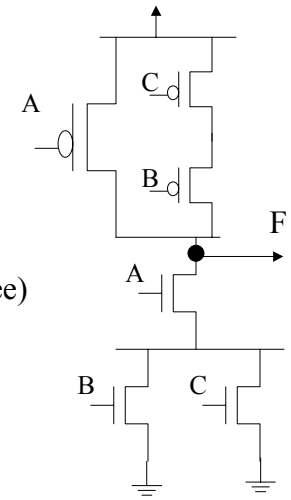
The Ptree and Ntree are logic *duals* of each other.

Complex Gate

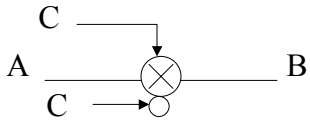
F = Not (A and (B or C)) (Ntree)

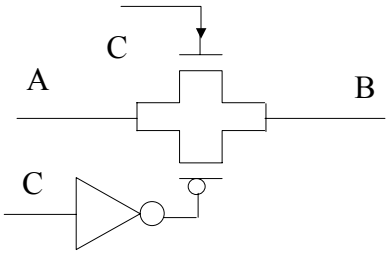
F Dual = Not (A' or (B' and C')) (Ptree)

Dual: Replace Ands with Ors, Ors with Ands, complement inputs.



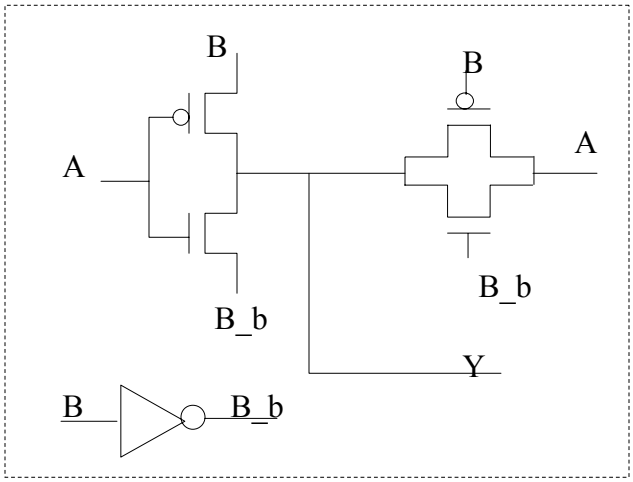
Transmission Gate





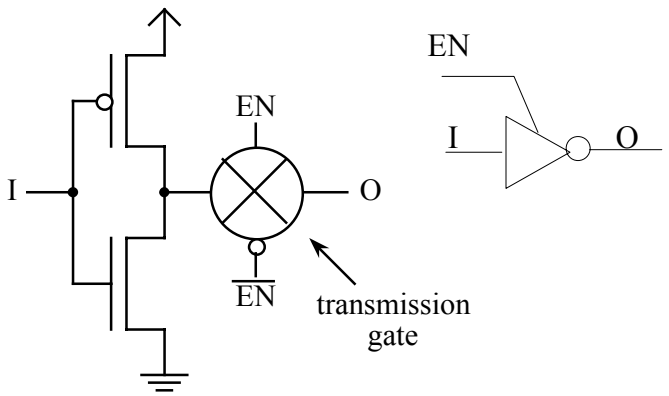
Full Vdd to Gnd swing between A, B

xor2 implementation

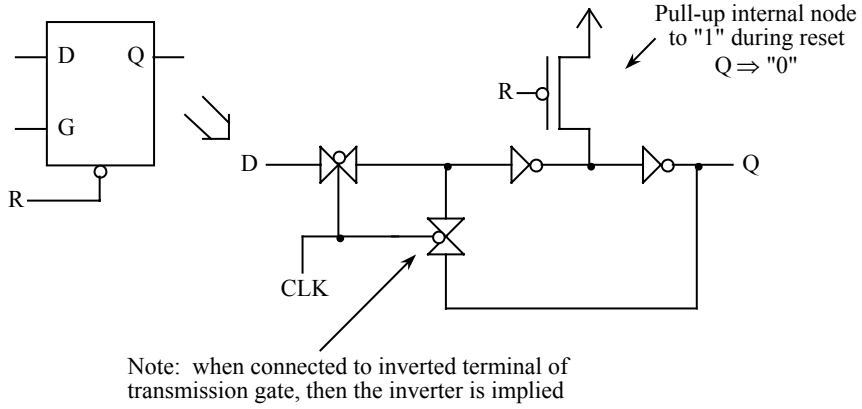


six transistors

Tristate Inverter



D-Latch

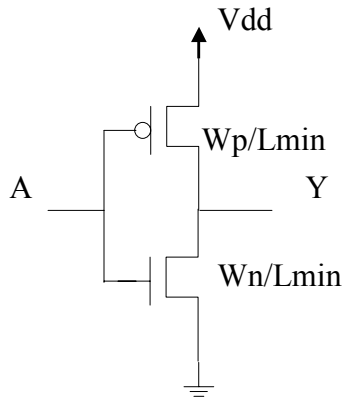


Lmin, Lambda

Typically for digital circuits, gate length of transistors are all the same and will be the minimum gate length allowed for that technology. Typically the number used to describe the technology is the minimum gate length.

1.2U Lmin = 1.2U
0.8 Lmin = 0.8U
0.25 Lmin = 0.25U

A scaling factor called *lambda* is often use to specify W/L values in a technology independent manner. Typically, Lambda = ½ Lmin (not always, we will this in our class).

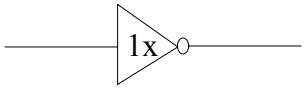


Inverter Switching Point

- The inverter switching point is determined by the ratio of the *B*n/*B*p, where *B*n is the gain of the N-tree and *B*p is the gain of the P-tree.
  - If *B*n/*B*p = 1, then switching point is Vdd/2
- If the W/L of both N and P transistors are equal, then *B*n/*B*p = $\mu_n/\ \mu_p$ = electron mobility/hole mobility
  - This ratio usually between 2 and 3
  - This means that ratio of the Wptree / Wntree needs to be between 2 and 3 in order for Bn/Bp to be approximately 1
- In this class will use W_{ptree}/W_{ntree} to be 2

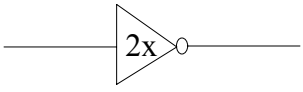
Gate Sizes

Assume minimum sized inverter is Wp/Wn = 2/1, L= Lmin . Wn = Lmin, Wp = 2* Lmin. This becomes the 1X inverter for this technology.

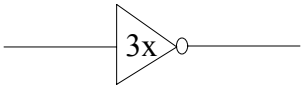


1X inverter

In a standard cell library, will have different sized gates to choose from for increasing loads

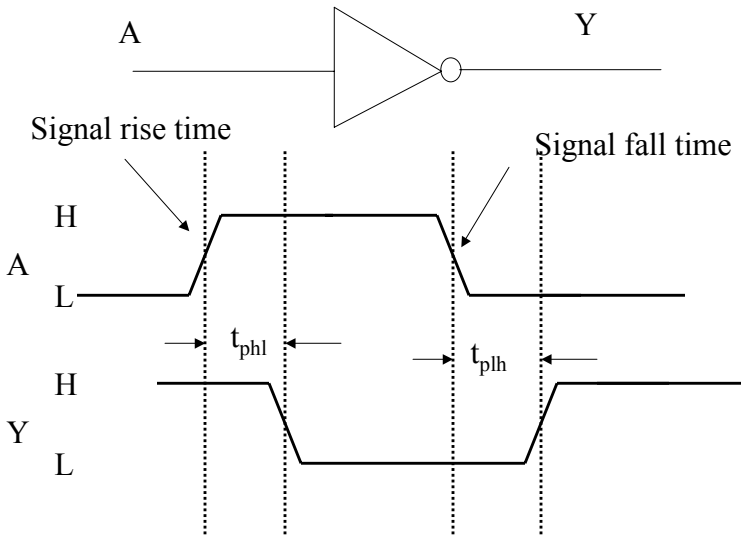


Wp/Wn still = 2/1, but Wp, Wn are double the size used in 1X inverter.

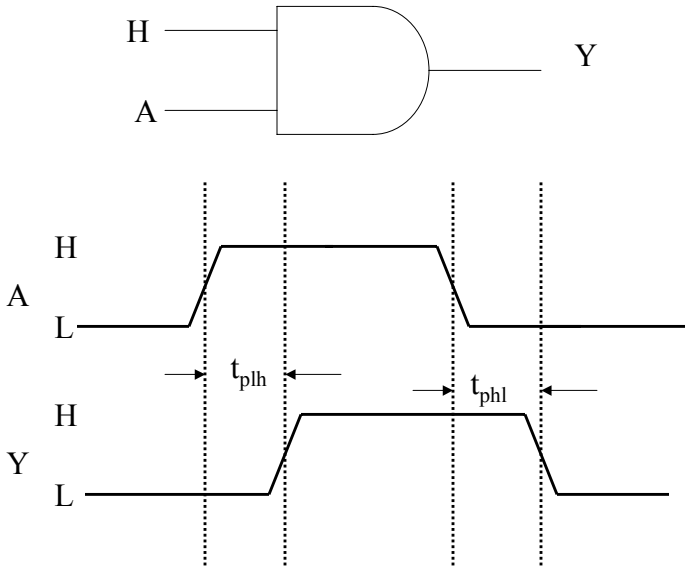


For most gates, diminishing returns after about 4x.

Propagation Delay (inverting)



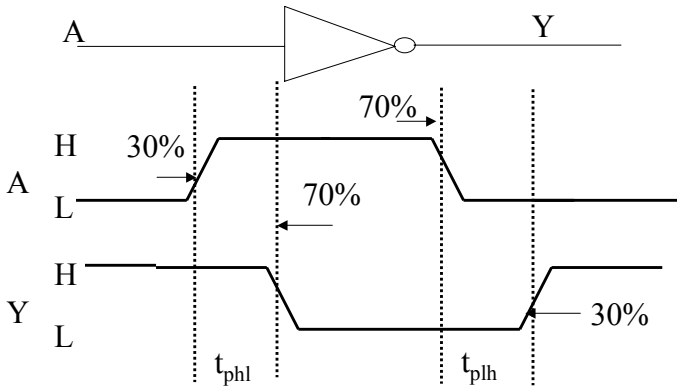
Propagation Delay (non inverting)



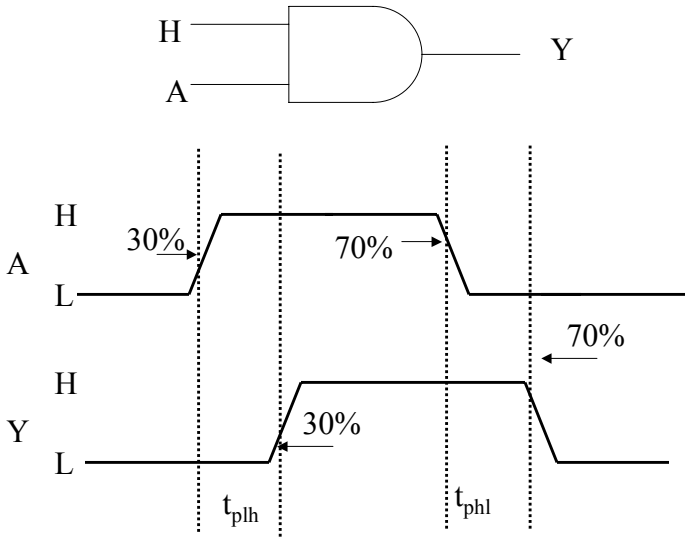
Where to measure delay?

If use 50% point (input) to 50% point (output), can produce negative delays (slow input slope, fast output slope).

A better way is to use the 30% and 70% points on the signals.



Non-Inverting Gate



What affects gate delay?

- Environment
  - Vdd – increasing VDD decreases delay
  - Temperature – decreasing Temp decreases delay
  - Fabrication effects – Fast NMOS, Fast PMOS, Slow NMOS, Slow PMOS
- Usually measure delay for at least three cases
  - Best - high Vdd, low Temp, Fast N, Fast P
    (use high Vdd, Low temp only if these conditions can adversely affect the circuit operation, e.g. timing margin for hold time is affected by fastest paths, not slowest paths).
  - Worst - low Vdd, high Temp, Slow N, Slow P - used for determining timing margin purposes.
  - Nominal - nominal Vdd, Room Temp (25 C), nominal N,P

What else affects gate delay?

Input slew and output load both affect timing. For a FIXED input slope, FIXED environment, a simple timing model is:

delay = Tnoload + K* Cload

Tnoload is the delay of the gate with no external load.

K is different for TPLH, TPHL since it represents the channel resistance. Same equation is used for Slew values.

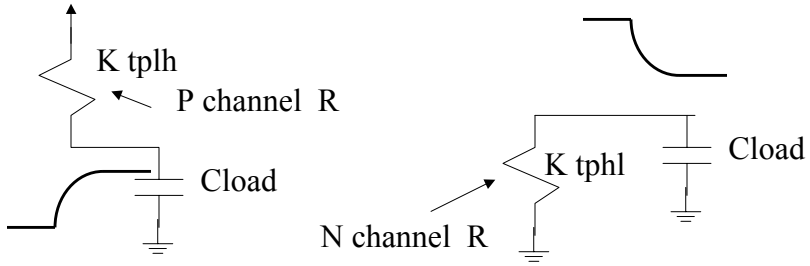


Table Lookup for Different Input Slews

Tplh Table (delay)

Slope	K
S_1	K_1
S_2	K_2
S_2	K_3
S_3	K_4
S_4	K_5

Tplh Table (slew)

Slope	K
S_1	K_1
S_2	K_2
S_2	K_3
S_3	K_4
S_4	K_5

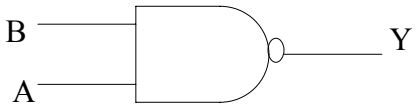
Also need tables for delay, slew for Tphl.

2D Table Based on Cload

	Cload		
Slope	1X CL	5X CL	15x CL
S_1	K_{1_1x}	K_{1_5x}	K_{1_15x}
S_2	K_{2_1x}	K_{2_5x}	K_{2_15x}
S_3	K_{3_1x}	K_{3_5x}	K_{3_15x}
S_4	K_{4_1x}	K_{4_5x}	K_{4_15x}
S_5	K_{5_1x}	K_{5_5x}	K_{5_15x}

Would need four tables for Tplh (delay), Tplh (slew), Tphl (delay), Tphl(slew)

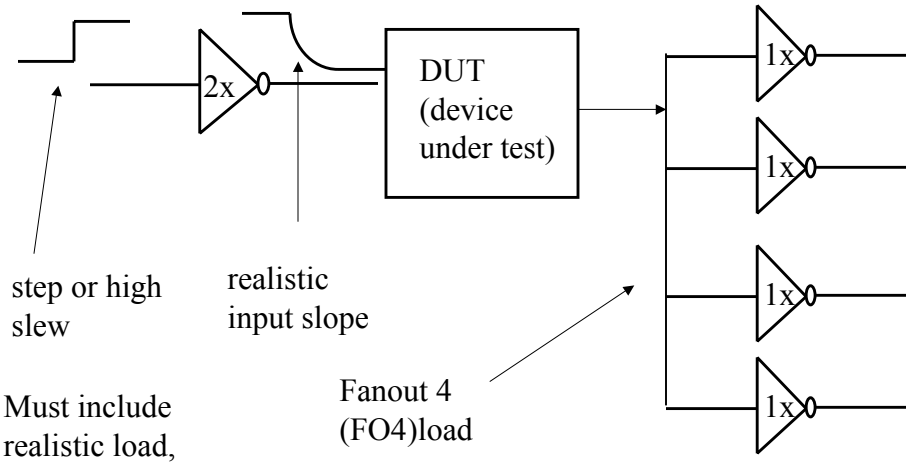
Multiple Input Gates



Timings are A2Y Tplh, A2Y Tphl, B2Y Tplh B2Y Tphl

Would need delay tables for each of these timings.

Timing Characterization Setup



Must include
realistic load,
realistic input slew

Power Consumption In CMOS

- Static Consumption
  - Ideally, static CMOS gates with P-Tree/N-trees that are duals of each other should have negligible power consumption
  - Unfortunately, as devices have shrunk, sub-threshold current has increased and static power consumption has increased to significant values, even for static CMOS
- Dynamic Power Consumption
  - Power consumed during switching (due to signal transitions)

Dynamic Power Consumption

Pd = Vdd**2 * Cap (switched) * Frequency

Switched capacitance is the capacitance that charges/discharges during a clock cycle or a computation. Includes MOSFET capacitances, routing capacitance.

Note that the biggest influence is Vdd. Power consumption will be reduced by the square of Vdd if Frequency is constant.

Direct Path Current

There is also a '*Direct Path'* power component (short circuit current) that is due to both Ntree, Ptree being momentarily on during switching.

Influenced by Rise/Fall time of input signals.

P direct path = Vdd * I*peak* (Trise + Tfall/2)/Frequency

I*peak* is the saturation current of the Ptree and Ntree and is proportional to the size of the transistors.

We will ignore this contribution in this class.

System Power Consumption

System Pd = Static + Dynamic

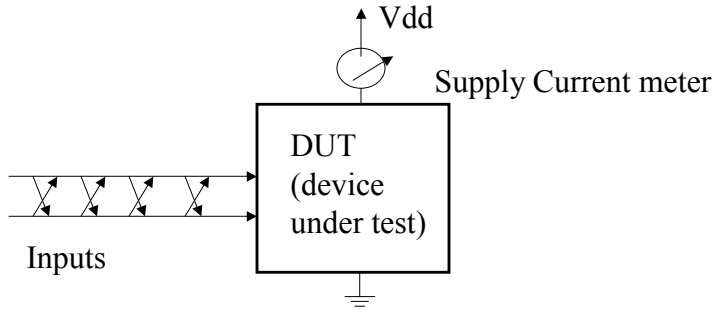
Static = proportional to number of transistors in system (will be a constant)

Dynamic = Capacitance switched at clock frequency (clock net, FF clock loads) +

Capacitance switched at compute rates (combinational gates)

Not all combinational gates switch with each calculation. Their toggle rates are typically anywhere from 5% to 15% of the clock frequency. Logic synthesis techniques can be used to reduce the toggle rate of a system.

Power Characterization

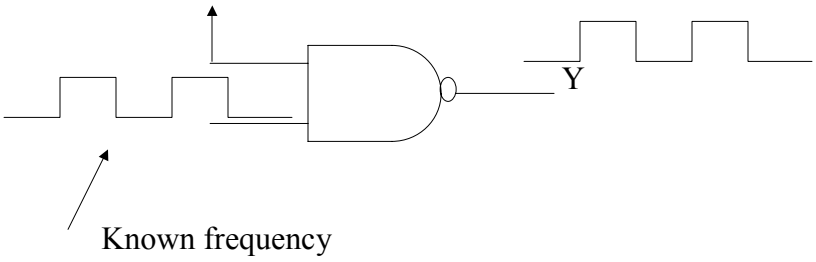


Change inputs at a known frequency with typical input sequence. Measure Supply current. System Capacitance is:

$$Pd = Vdd^{**2} * C_{sys} * Freq$$
$$Vdd * Iss = Vdd^{**2} * C_{sys} * Freq$$
$$C_{sys} = Iss / (Vdd * Freq)$$

System Capacitance for high performance microprocessors a few nano-Farads.

Gate Power Characterization



May have output of gate loaded by a typical FO4 load.

Measure Iss for just the gate, compute effective switched capacitance of gate. Usually in the 10's to low 100's of femto-farads.

CMOS Technology Scaling

- See page 146-149 of Rabaey text
- Full Scaling – Vdd scaled by S, Dimensions by S
- General Scaling - Vdd scaled by U, Dimensions by S
- Fixed-Voltage Scaling - only Dimensions by S
- Overall Capacitance scales by 1/S, all scaling models
- Delay by 1/S (short channel devices, all scaling models)