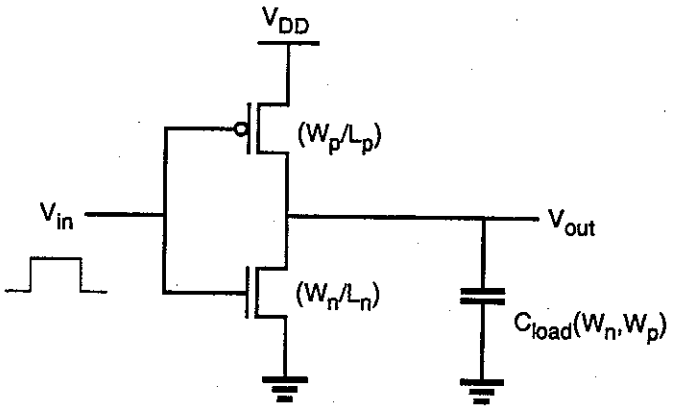
Unit 2
MOS Inverters

CMOS Design With Delay Constraints: *Design for Performance*

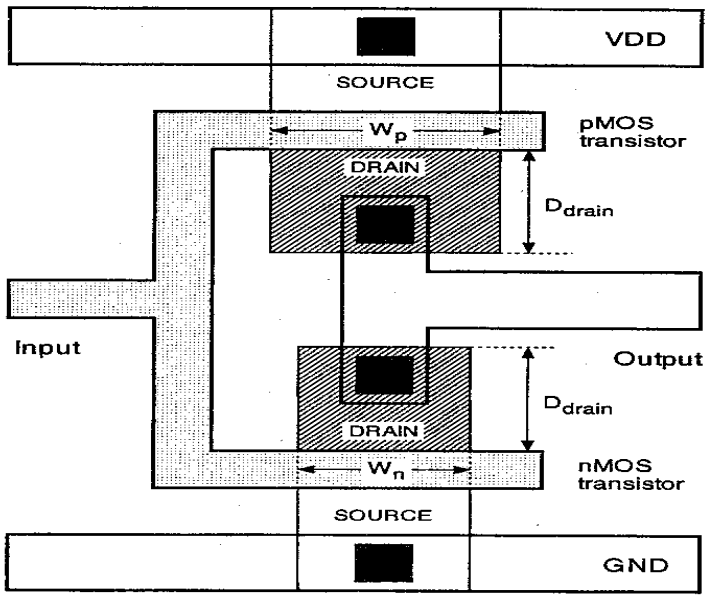
- The propagation delay equations on chart 4-5 can be rearranged to solve for W/L, as shown below, where we substituted $C_{ox}\mu_n(W_n/L_n)$ for k_n and similarly for k_p
- These equations can then be used to "size" a CMOS circuit to achieve a desired minimum rising or falling propagation delay assuming C_{load} and other parameters are known
  - After determining the desired W/L values, we can obtain the device widths W based on the technology minimum design device lengths L
- Other constraints such as rise time/fall time or rise/fall symmetry may also need to be considered in addition to rise and fall delay

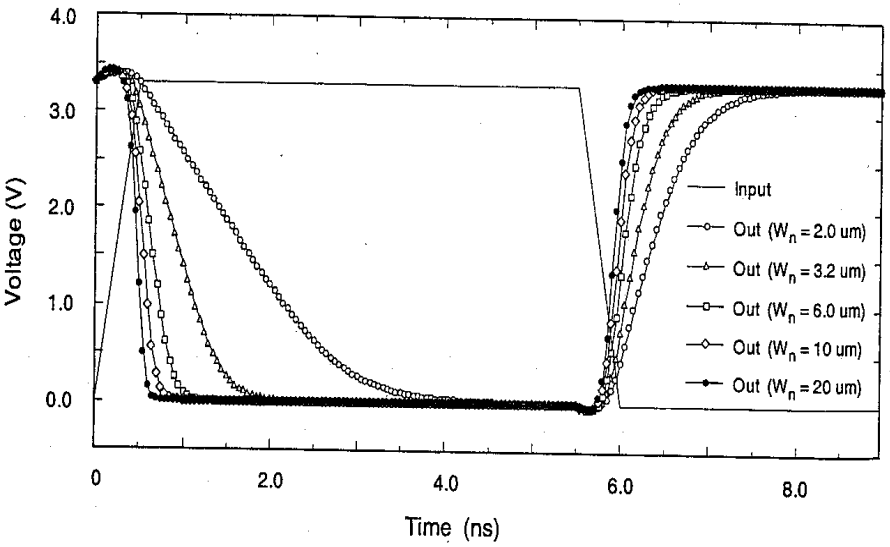


$$\left(\frac{W_n}{L_n}\right)=\frac{C_{load}}{\tau^*_{PHL}\,\mu_n\,C_{ox}\left(V_{DD}-V_{T,n}\right)}\left[\frac{2\,V_{T,n}}{V_{DD}-V_{T,n}}+\ln\left(\frac{4\left(V_{DD}-V_{T,n}\right)}{V_{DD}}-1\right)\right]$$

$$\left(\frac{W_p}{L_p}\right)=\frac{C_{load}}{\tau^*_{PLH}\,\mu_p\,C_{ox}\left(V_{DD}-\left|V_{T,p}\right|\right)}\left[\frac{2\left|V_{T,p}\right|}{V_{DD}-\left|V_{T,p}\right|}+\ln\left(\frac{4\left(V_{DD}-\left|V_{T,p}\right|\right)}{V_{DD}}-1\right)\right]$$

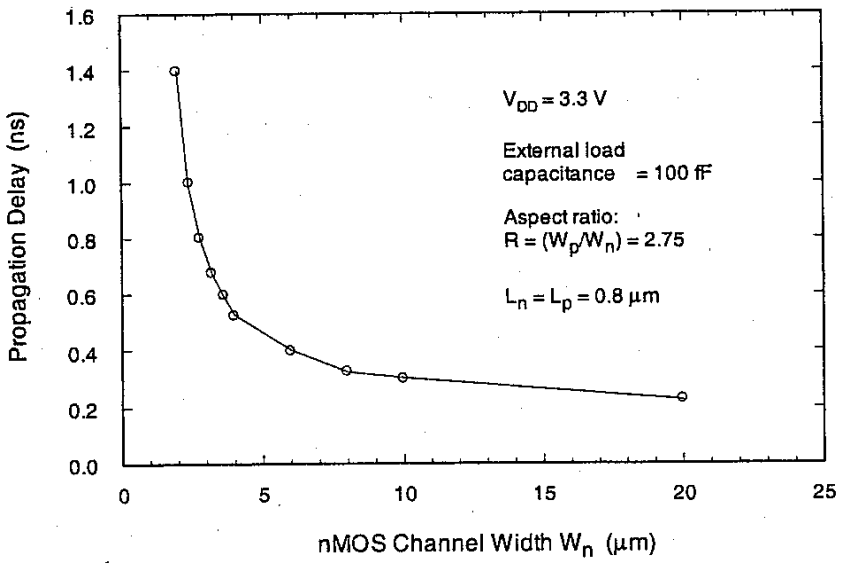
Computing Intrinsic Transistor Capacitance

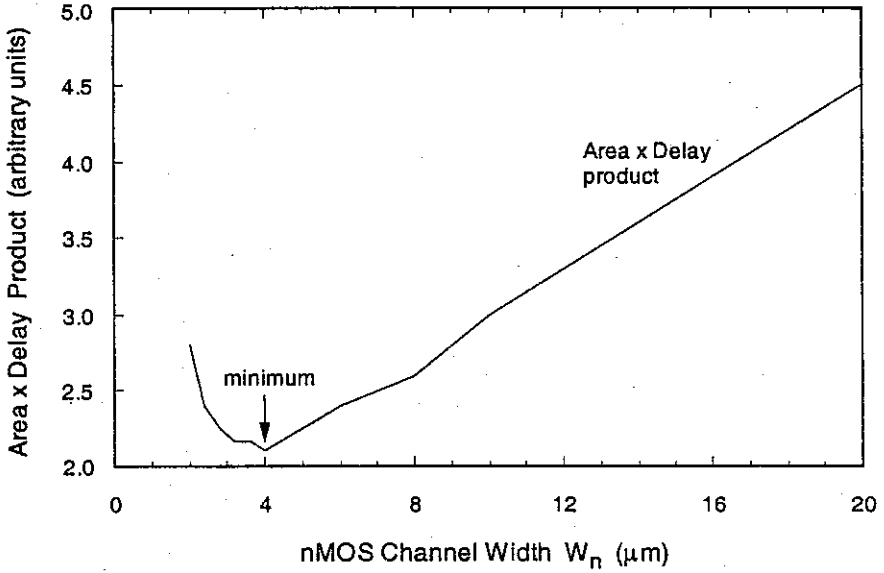




- Intrinsic PN junction capacitance of the driving circuit must be added to the load capacitance C_{load}
- Consider the inverter example at left:
  - Area and perimeter of the PMOS and NMOS transistors are calculated from the layout and inserted into the circuit model
    - NMOS drain area = $W_n \times D_{drain}$
    - PMOS drain area = $W_p \times D_{drain}$
    - NMOS drain perimeter = $2 (W_n + D_{drain})$
    - PMOS drain perimeter = $2 (W_p + D_{drain})$
- SPICE simulations were done (bottom left) for a fixed extrinsic load of 100fF with increasing transistor width (Wp/Wn = 2.75)
  - Results show diminishing returns beyond a certain Wn (say about 6 um) due to effect of the increasing drain capacitance on the overall capacitive load

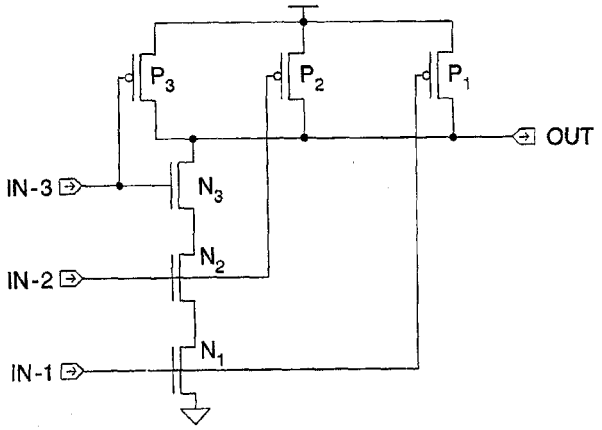
Area x Delay Figure of Merit

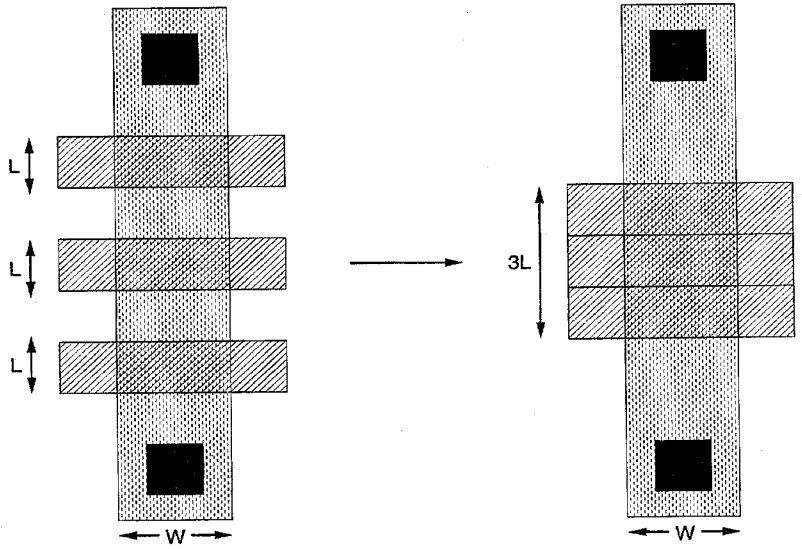




- Increasing device width shows diminishing returns on propagation delay time (inverter circuit of chart 4-24)
- Define a figure of merit as area x delay for the inverter circuit
  - Increasing device width Wn shows a minimum in area x delay product
- Unconstrained increase in transistor width in order to improve circuit delay is often a poor tradeoff due to the high cost of silicon real estate on the wafer!!

Transistors in Series: CMOS NAND





- Several devices in series each with effective channel length L_{eff} can be viewed as a single device of channel length equal to the combined channel lengths of the separate series devices
  - e.g. 3 input NAND: a single device of channel length equal to $3L_{eff}$ could be used to model the behavior of three series devices each with L_{eff} channel length, assuming there is no skew in the increasing gate voltage of the three N pull-down devices.
  - The source/drain junctions between the three devices essentially are assumed as simple zero resistance connections
  - During saturation transient, the bottom two devices will be in their linear region and only the top device will be pinched off.

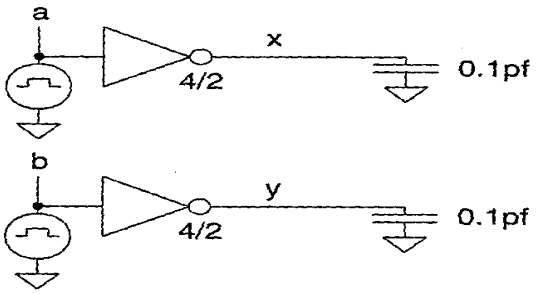
Delay Dependence on Input Rise/Fall Time

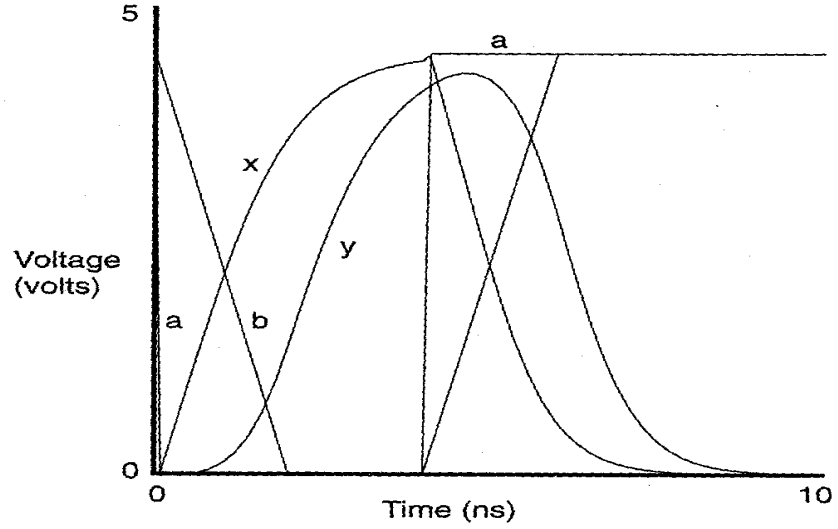
- For non-abrupt input signals, circuit delays show some dependency upon the input rise/fall time
  - Case a with input a and output x shows minimum rising and falling delays
  - Case b with input b and output y shows added delay due to the delay in getting the input to the switching voltage.
  - Empirical relationship to include input rise/fall time on output fall/rise delay are given:

$$t_{df} = [t_{df}^2(\text{step input}) + (t_r/2)^2]^{1/2}$$

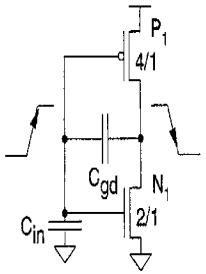
$$t_{dr} = [t_{dr}^2(\text{step input}) + (t_f/2)^2]^{1/2}$$

- For CMOS the affect of input rise(fall) time on the output fall(rise) time will be less severe than the impact on the falling(rising) delay.

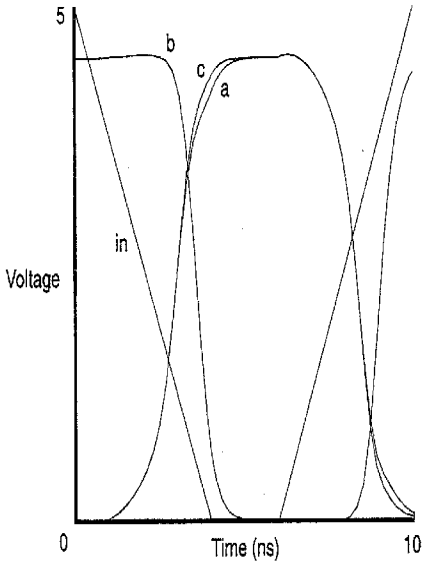




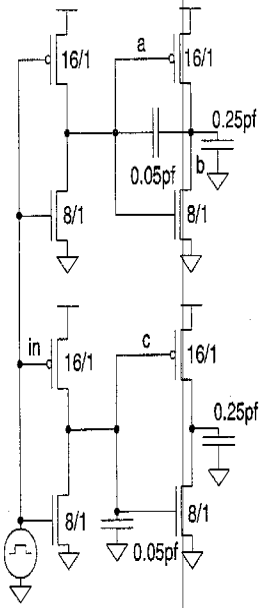
Bootstrapping Effect on Inverter Delay



(a)



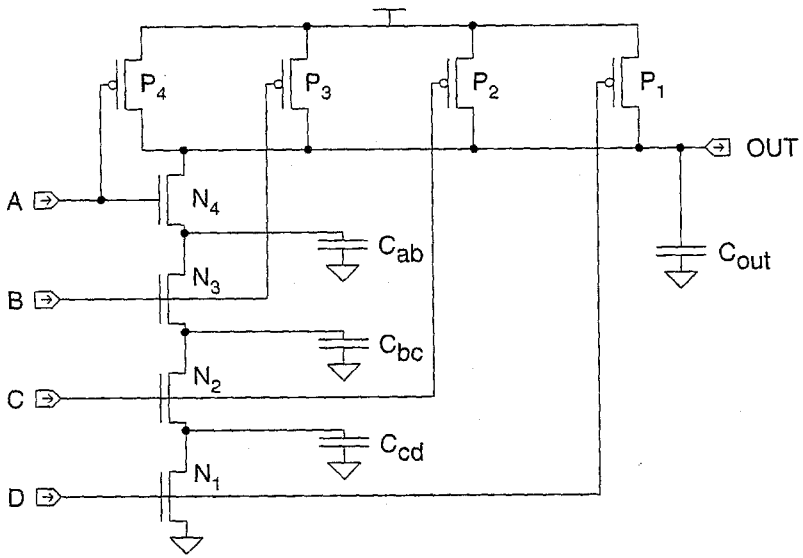
(b)



(c)

- Gate-to-drain capacitance Cgd in a CMOS inverter (or other MOS logic ckt) causes feedback of the transient signal from the output to the input gate
  - called ***Bootstrapping*** or ***Miller Effect***
  - as input rises and output falls, Cgd couples back a portion of output transient to the input, thus slowing the input rising waveform
  - SPICE simulation at left shows impact on input node ‘a’ due to a 0.05pF bootstrap capacitor versus no impact on input node ‘c’ inverter with no bootstrap capacitor
  - Small effect in most small inverters and logic circuits
- Voltage doubling circuits and certain large swing drivers use intentionally designed bootstrap capacitors to provide overdrive to the gate of pullup devices

Modeling Parasitic Capacitances: 4 input NAND

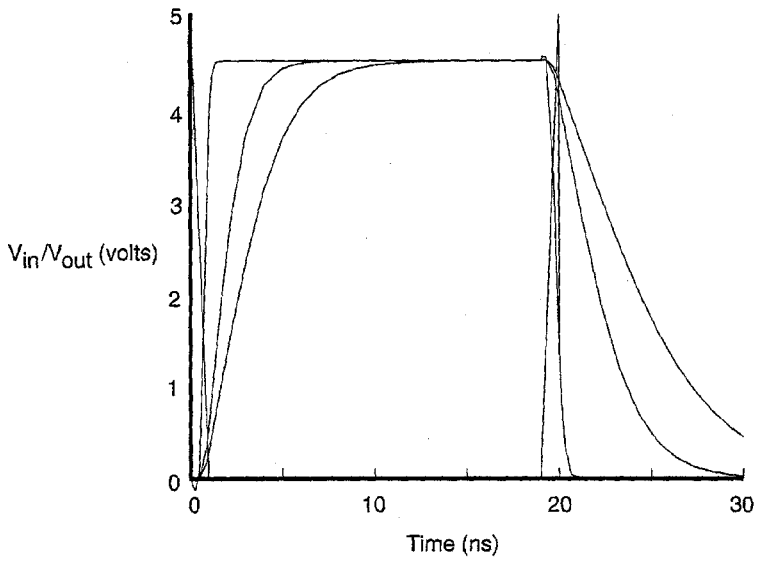


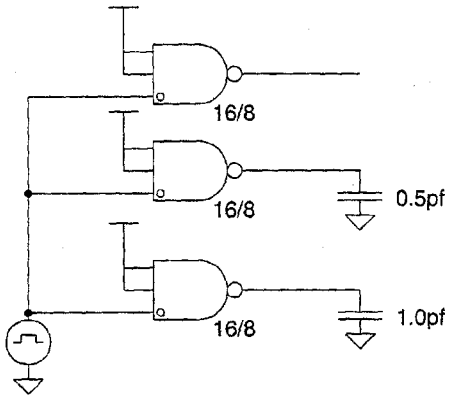
- Capacitances Cab, Cbc, Ccd exist at internal nodes of series-connected devices and add to delay of circuit
  - must be discharged to ground along with Cout through N1, N2, N3 series devices when all inputs go high
  - must be charged through N2, N3, N4 and P1 when input D goes low
- Modeling approaches (Simple RC Delay):

  (Rn1 + Rn2 + Rn3 + Rn4) x (Cout + Cab + Cbc + Ccd)
  - not very accurate
- Modeling approaches (Elmore ladder delay):

Rn1 Ccd + (Rn1 + Rn2) Cbc + (Rn1 + Rn2 + Rn3) Cab + (Rn1 + Rn2 + Rn3 + Rn4) Cout
  - more accurate
- Penfield-Rubenstein Slope Delay Model factors the input rise (fall) time into the above

Effect of Loading Capacitance on Gate Delay



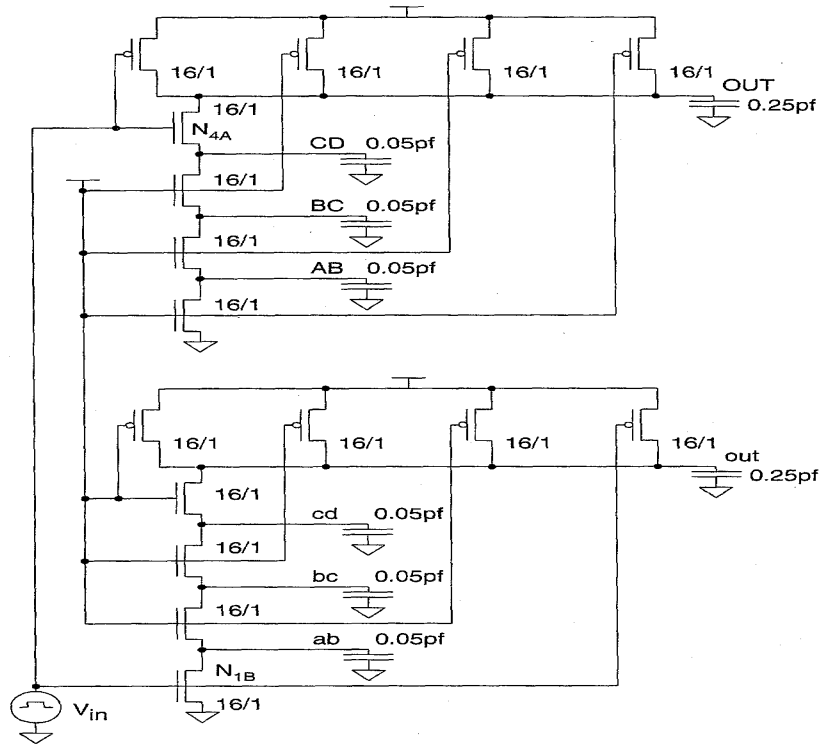


- Delay equations are often written to factor the impact of the fan-out and load capacitance to the circuit delay

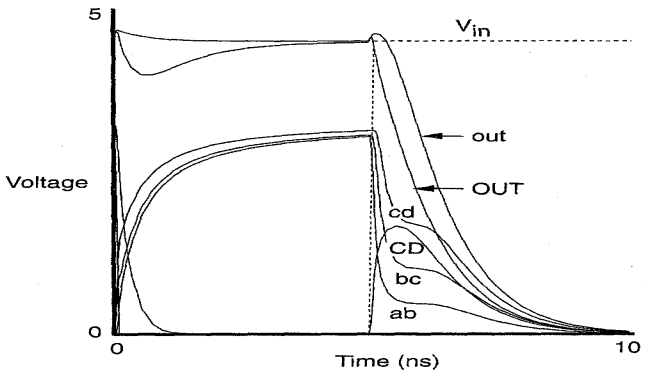
$$t_d = t_{d\_intrinsic} + (k1 \times C_L) + (k2 \times FO)$$

  - where CL is the load capacitance, FO is the fan-out, and td_intrinsic is the unloaded delay of the circuit
- Tables of delay versus load condition are built up from simulation models and used for path delay prediction.

Body Effect on Delay: 4 input NAND



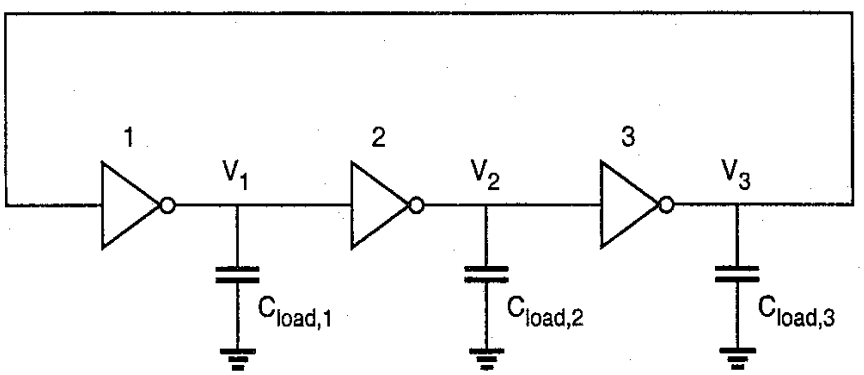
(a)



(b)

- In a logic gate with devices in series causing source voltages above ground (for a NAND) or below Vdd (for a NOR), the circuit response is slowed due to the body effect on increasing threshold voltage Vtn (or |Vtp|).
  - If only the bottom series N device is switched, nodes ab, bc, and cd are sitting at Vdd – Vtn prior to the switching
  - Each node must be discharged to ground successively prior to discharging Cout through the 4 series N devices
    - See figure at left

CMOS Ring Oscillator Circuit

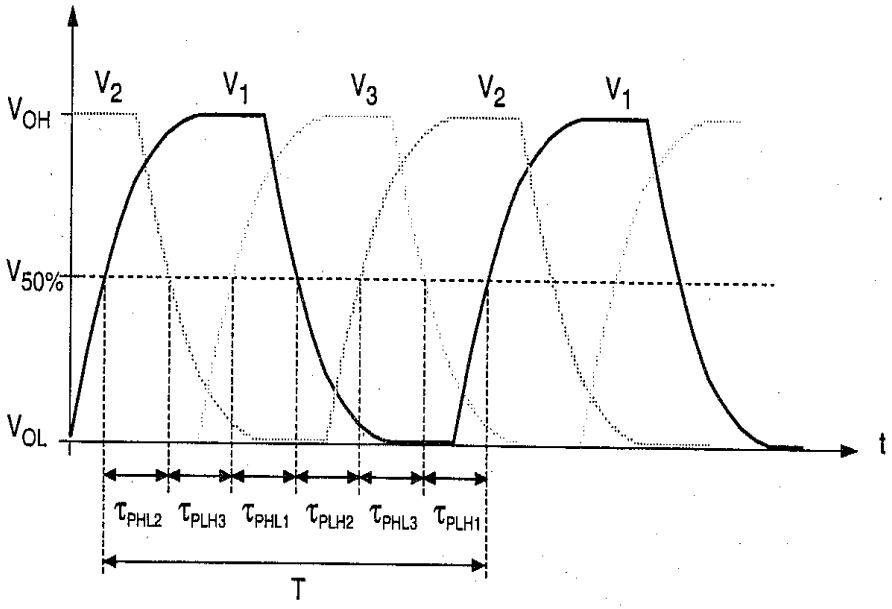


- An odd number of inverter circuits connected serially with output brought back to input will be astable and can be used an an oscillator (called a ring oscillator)
- Ring oscillators are typically used to characterize a new technology as to its intrinsic device performance
- Frequency and stage are related as follows:

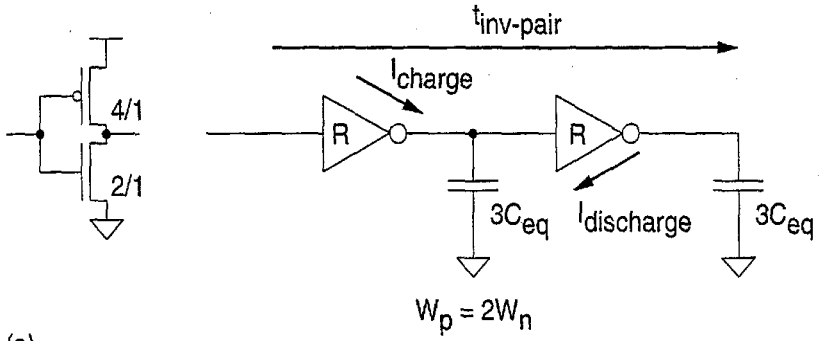
  $\mathbf{f = 1/T = 1/(2n\tau_P)}$

  where n is the number of stages and

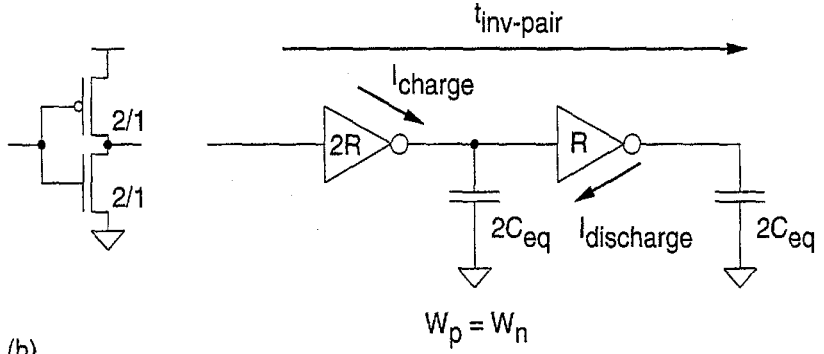
  τ_P is the stage delay



CMOS Gate Transistor Sizing



(a)



(b)

- Symmetrical inverter design (case a):
  - P mobility = ½ x N mobility
  - Wp = 2 x Wn
  - Input gate capacitance = 3 x Ceq where Ceq is the pull-down device gate capac.

Pair delay = tfall + trise = R3Ceq + 2(R/2)3Ceq
= 6RCeq

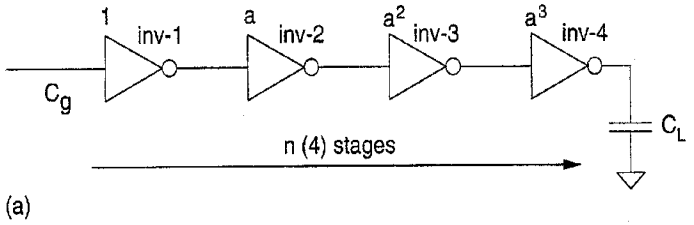
Non-symmetrical inverter design (case b):

- Wp = Wn
- Input gate capacitance = 2 x Ceq

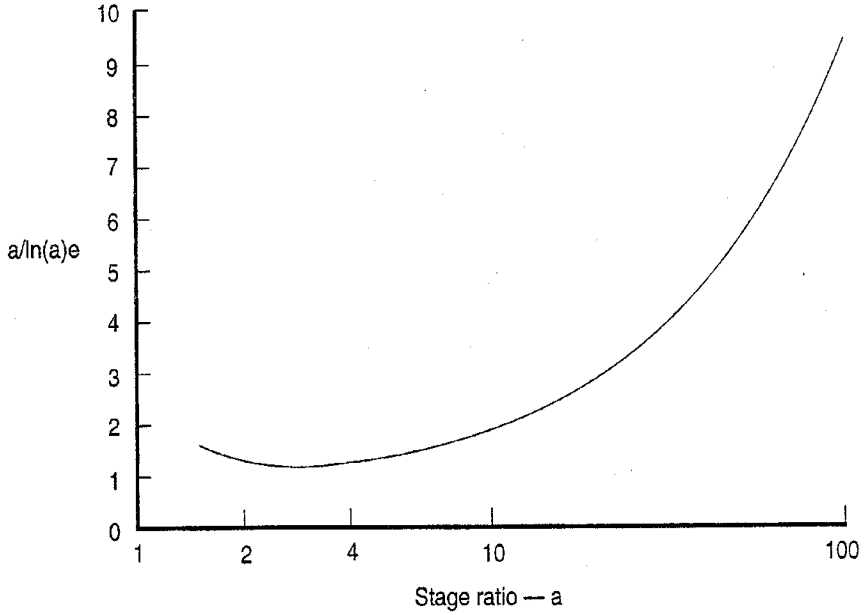
Pair delay = tfall + trise = R2Ceq + 2R2Ceq
= 6RCeq

In the simple case where the load is comprised mainly of input gate capacitance no impact to the total delay of the pair of inverters was observed by using non-symmetrical Wn=Wp

Driving Large Capacitive Loads: Stage Ratio



(a)

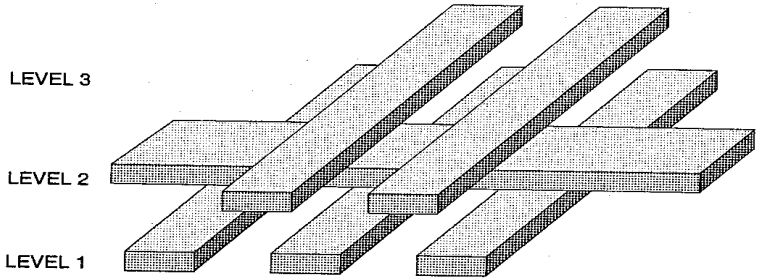


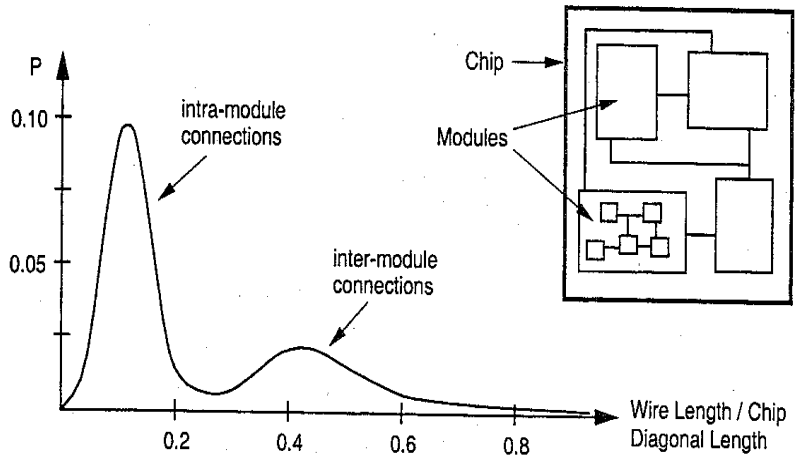
(b)

- For driving large load capacitance C_L, can use N buffer drivers in series, each with stage ratio $C_{out}/C_{in} = a$
  - Input capacitance Cg
  - Delay per stage = $\mathbf{at_d}$ given that the delay of a minimum size stage driving another minimum size stage is $\mathbf{t_d}$
  - Let $\mathbf{R = C_L/C_g = a^N}$
  - Then the total stage delay is given by **Total Delay = $\mathbf{Nat_d}$ = $\mathbf{at_d(\ln R/ \ln a)}$**
  - Setting derivative of total delay w/r **a** equal to zero yields optimum stage ratio **a = e**
- If we allow inclusion of inverter output drain capacitance term in the analysis, the optimum stage ratio is given by $\mathbf{a_{opt} = e^{(k + aopt)/aopt}}$ where $k = C_{drain}/C_{gate}$

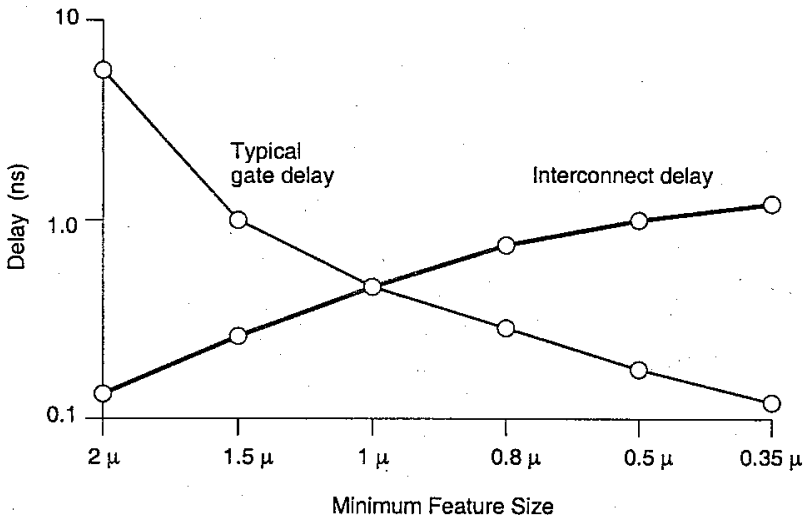
Increasing Importance of Interconnect Delay

- IC's are going to 6-7 levels of metal interconnect in advanced technologies
- Chart at bottom left shows typical distribution of wire length on a processor chip or an ASIC

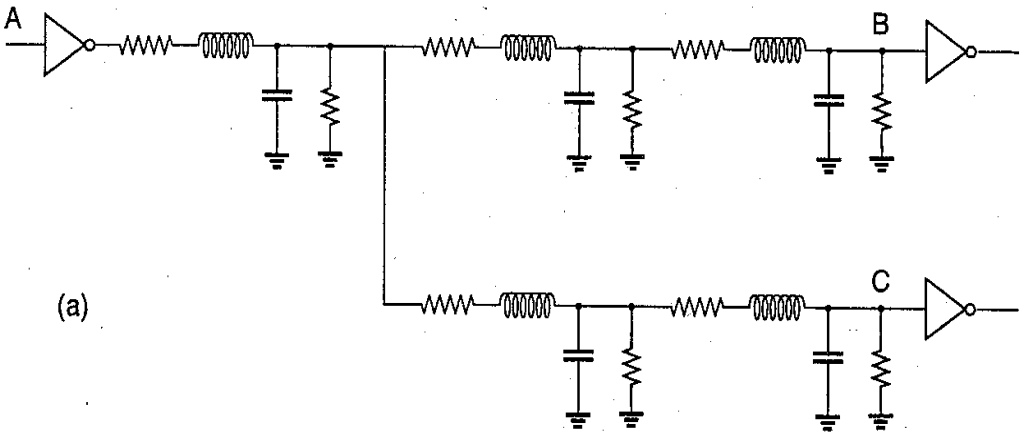




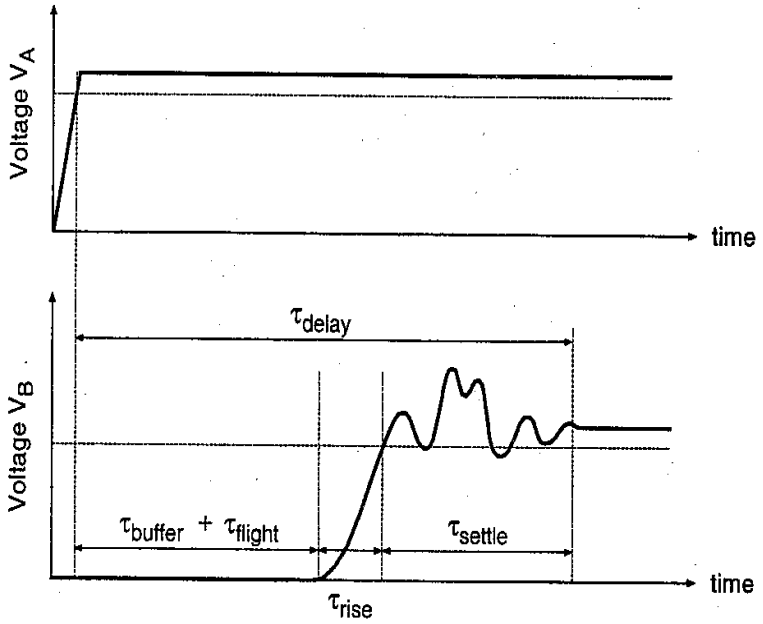
- As feature size drops, interconnect delay often exceeds gate delay
  - Chart below shows that for very long wires, interconnect delay has exceeded gate delay above 1um feature size
- Interconnect delay is becoming the most serious performance problem to be solved in future IC design



Interconnect Delay with Inductive Effects



(a)

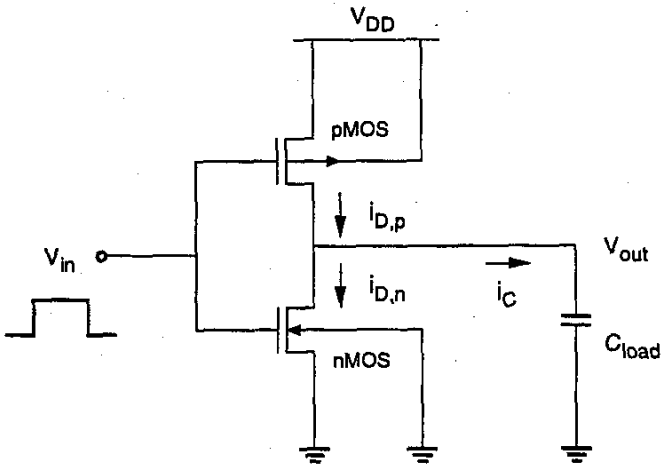


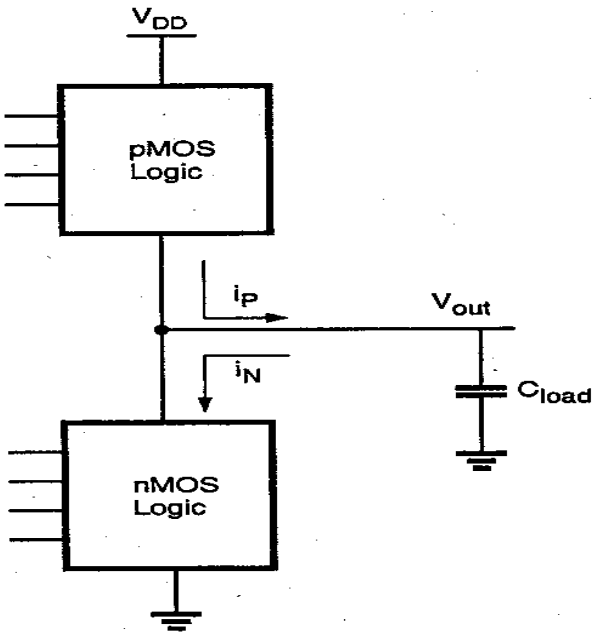
(b)

- For the design of critical performance nets (such as clock distribution) on a processor chip, inductance must be taken into consideration

- Simulation result in (b) shows the effect of ringing on a rising transition due to reflections at a discontinuity on an inductive net
  - Additional delay due to settling time is incurred if such ringing can not be eliminated by proper transmission line design techniques

Power Dissipation in a CMOS Inverter: Summary





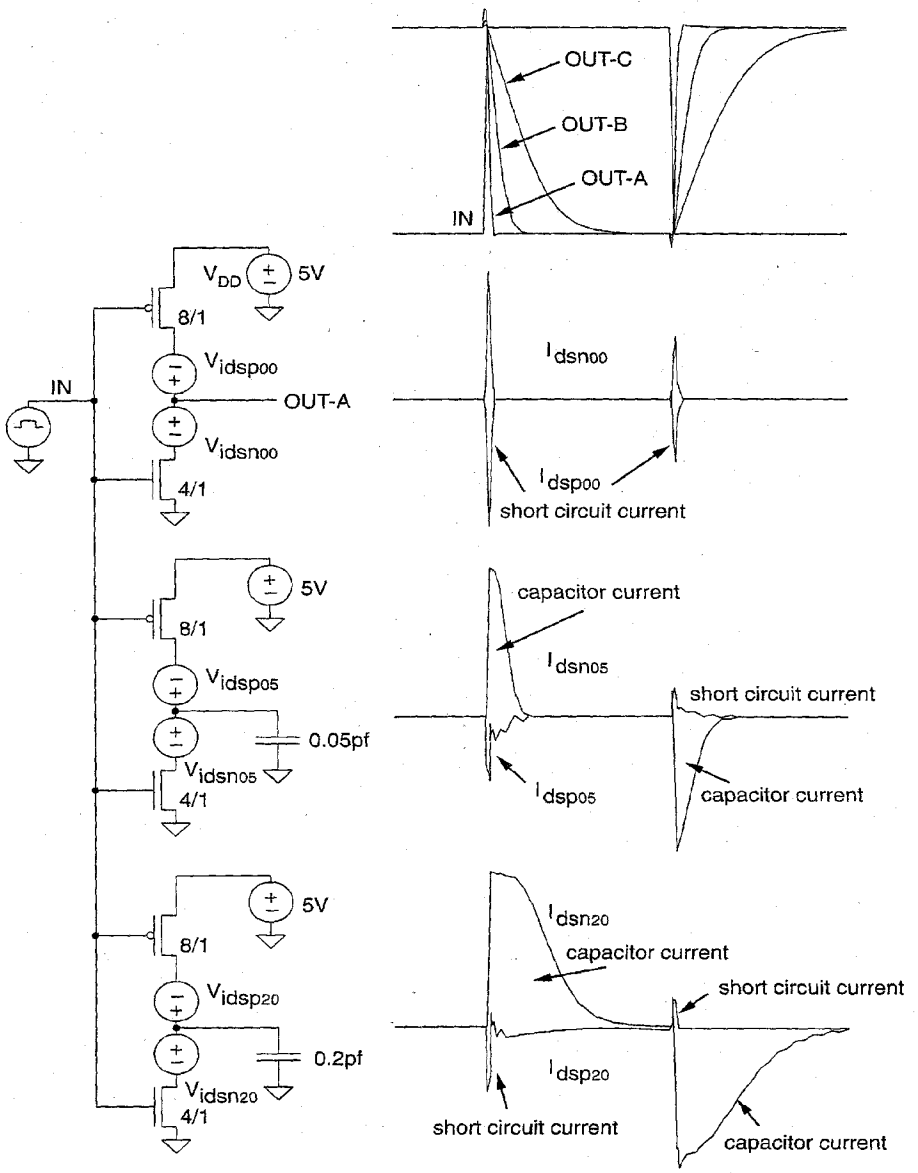
- For complementary CMOS circuits where no dc current flows, average dynamic power is given by

$$P_{ave} = C_L V_{DD}^2 f$$

where C_L represents the total load capacitance, V_{DD} is the power supply, and f is the frequency of the signal transition

  - above formula applies to a simple CMOS inverter or to complex, combinational CMOS logic
  - applies only to dynamic (capacitive) power
  - dc power and/or short-circuit power must be computed separately

Average Dynamic Power in CMOS Inverter



- Average dynamic power derivation:
  - On negative going input, pull-up device charges the load capacitance. On positive going input, pull-down device discharges the load into ground.
  - Average power given by

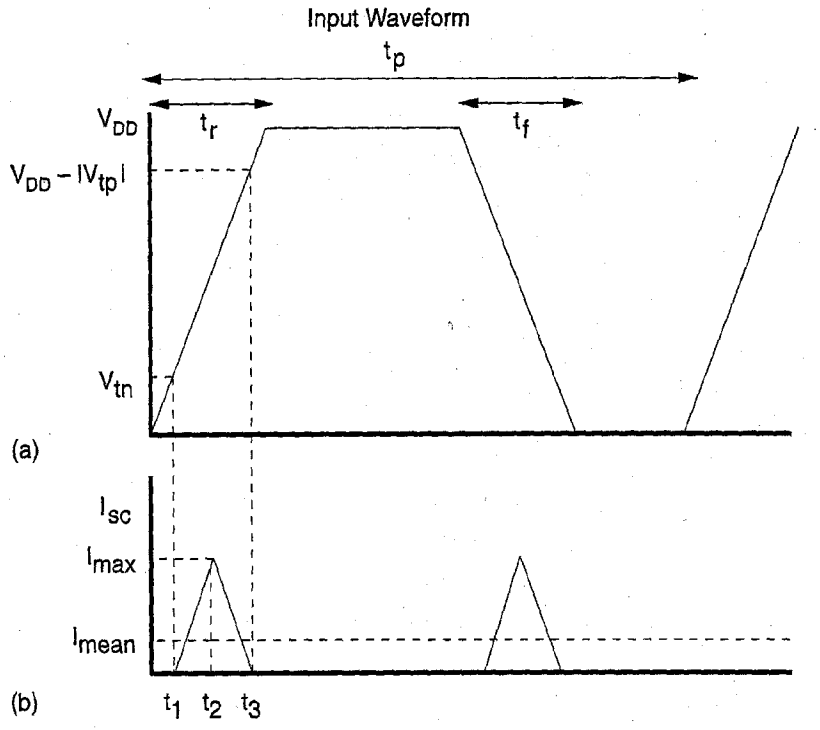
$$P_{ave} = (1/T) f C_L (dv_{out}/dt) (V_{dd} - v_{out}) dt + (1/T) f (-1) C_L (dv_{out}/dt) v_{out}\, dt$$ where the first integral is taken from 0 to T/2 and the second integral is from T/2 to T

- completion of the integral yields

$$P_{ave} = C_L V_{dd}^2 f$$ where f = 1/T

- Note that the dynamic power is independent of the typical device parameters, but is simply a function of power supply, load capacitance and frequency of the switching!

CMOS Short-Circuit Power Dissipation

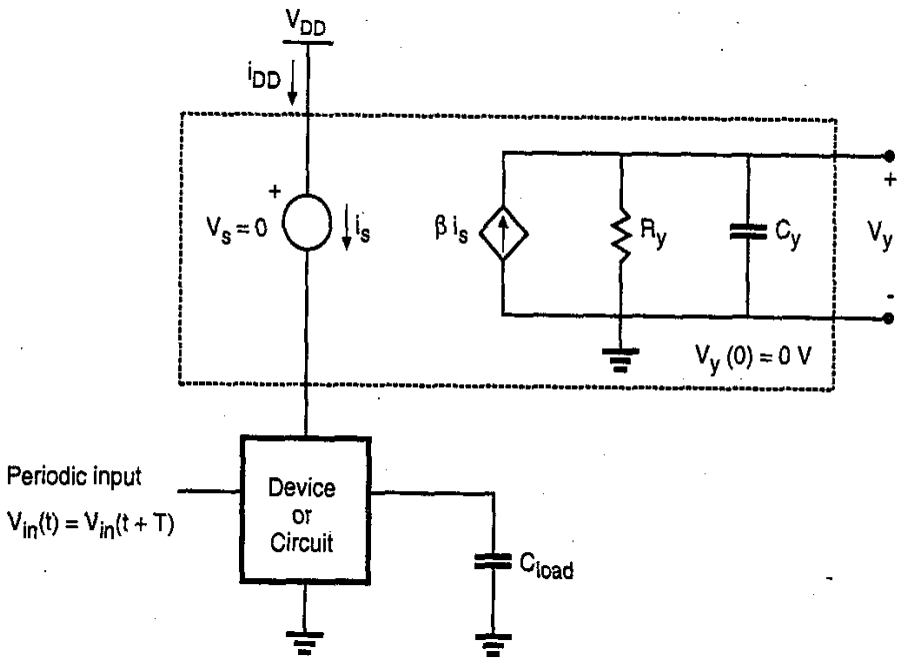


- The total power in a CMOS circuit is given by $\mathbf{P_{total} = P_d + P_{sc} + P_s}$ where Pd is the dynamic average power (previous chart), Psc is the short circuit power, and Ps is the static power due to ratio circuit current, junction leakage, and subthreshold Ioff leakage current
- Short circuit current flows during the brief transient when the pull down and pull up devices both conduct at the same time where one (or both) of the devices are in saturation
- For a balanced CMOS inverter with $\beta n = \beta p$, and $V_{tn} = |V_{tp}|$, the short circuit power can be expressed by

$$\mathbf{P_{sc} = (\beta/12)(V_{dd} - 2V_t)^3 (t_{rf}/t_p)}$$

where t_p is the period of the input waveform and t_{rf} is the total risetime (or falltime) $tr = t_f = t_{rf}$

Power Meter for use in SPICE Simulation



- Add a zero value voltage source Vs in series with V_{DD} and circuit in question
  - i_S is the current through Vs
- Add current source βi_S, resistor Ry, and capacitor Cy in parallel, as shown
- Integrating the current in the power circuit $C_y(dV_y/dt) = \beta i_S - V_y/R_y$ yields the solution

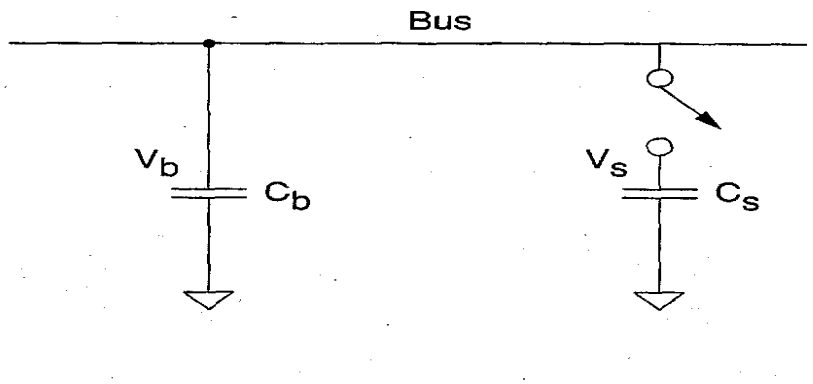
  $$\mathbf{V_y(T) = (V_{DD}/T)\int_0^T i_{DD}(\tau)d\tau}$$

  where $\beta = V_{DD}C_y/T$
  - $V_y(T)$ will be the average power dissipated over the period T and can be plotted or printed out during the SPICE simulation

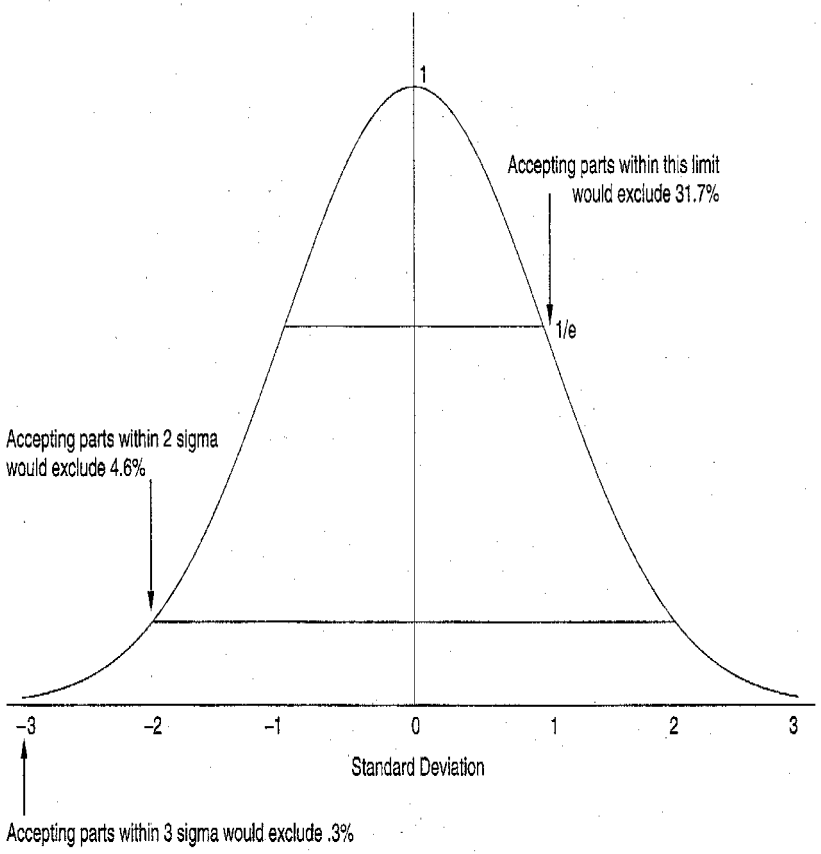
Charge Sharing Principle

- At time t=0-, switch is open and each capacitor contains some initial charge
- At time t=0+, the switch is closed and the charge redistributes across both capacitors
- Conserve the total charge:
  - Sum up initial charge Qt = Qb + Qs = CbVb + CsVs
  - Final charge is given by Qt = (Cb + Cs)Vf
  - Therefore, **Vf = (CbVb + CsVs)/(Cb + Cs)**
- If Vb = Vdd and Vs = 0, then

  Vf = Vdd Cb/(Cb + Cs) (which is similar to the equation for a resistor divider)
- Charge sharing plays an important role in many dynamic circuits, especially pulsed DOMINO and NORA logic as well as in DRAM operation.



Process Variation: Normal Distribution



- CMOS and other MOSFET circuit design requires designing around tolerances in the technology and process, the supply voltage Vdd, and the temperature.
  - Process parameter distributions are typically normal (Gaussian) where operation out to the 3 sigma point is usually a requirement
  - Statistical models are often derived with Gaussian or log normal distributions for each process parameter such as Tox, Xj, Vt, W, L, and the various mask dimensional images
  - Rejecting product outside the +/- 3 sigma limits only excludes 0.3% of the product
  - Power supply and temperature are normally given uniform distributions
- Definition of the design space involves identifying those corners of the multi-dimensional space where critical circuit performance, power, and operability exist

Definition of Design Window Corners

- Worst Case Design Methodology:
  - Identify corners of the design space where the circuit is slowest, or power is highest, or circuit ratio effects are critical
- Slow Circuit:
  - Vdd is low (say 10%), temperature is high, n and p transistors are slow caused by thick tox, high Vt, long L, and narrow W
- Fast Circuit/High Power:
  - Vdd is high, temperature is low, n and p transistors are fast
- Ratio circuit down level worst case:
  - Vdd is high, temperature is low, n device is slow, p device is fast

TABLE 4.11 CMOS Digital System Checks (Commercial)

PROCESS	TEMP	VOLTAGE	TESTS
Fast-n/fast-p	0°C	5.5V (3.6V)	Power dissipation (DC), clock races, hold time constraints
Slow-n/slow-p	125°C	4.5V (3.0V)	Circuit speed, setup time constraints
Slow-n/fast-p	0°C	5.5V (3.6V)	Pseudo-nMOS noise margin, level shifters, memory write/ read, ratioed circuits
Fast-n/slow-p	0°C	5.5V (3.6V)	Memories, ratioed circuits, level shifters

MOSFET Device Technology Scaling

PARAMETER	SCALING MODEL		
	Constant field	Constant voltage	Lateral
Length (L)	$1/\alpha$	$1/\alpha$	$1/\alpha$
Width (W)	$1/\alpha$	$1/\alpha$	1
Supply voltage (V)	$1/\alpha$	1	1
Gate-oxide thickness (t_{ox})	$1/\alpha$	$1/\alpha$	1
Current ($I = (W/L)(1/t_{ox})V^2$)	$1/\alpha$	α	α
Transconductance (g_m)	1	α	α
Junction depth (X_j)	$1/\alpha$	$1/\alpha$	1
Substrate doping (N_A)	α	α	1
Electric Field across gate oxide (E)	1	α	1
Depletion layer thickness (d)	$1/\alpha$	$1/\alpha$	1
Load Capacitance ($C = WL/t_{ox}$)	$1/\alpha$	$1/\alpha$	$1/\alpha$
Gate Delay (VC/I)	$1/\alpha$	$1/\alpha^2$	$1/\alpha^2$
	RESULTANT INFLUENCE		
DC power dissipation (P_s)	$1/\alpha^2$	α	α
Dynamic power dissipation (P_d)	$1/\alpha^2$	α	α
Power-delay product	$1/\alpha^3$	$1/\alpha$	$1/\alpha$
Gate Area ($A = WL$)	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha$
Power Density (VI/A)	1	α^3	α^2
Current Density	α	α^3	α^2

- Bob Dennard of IBM Watson Research Labs developed scaling theory for reducing device dimensions, power supply voltage and junction depths, while maintaining roughly constant electric fields
- Scaling theory is the basis for the SIA's NTRS (National Technology Roadmap for Semiconductors) which has been the roadmap for the industry for many technology generations
  - Moore's Law (Gordon Moore of Intel) has quantified the reduction in dimensions and increase in density and performance
    - 4X increase in DRAM and logic density every generation (2-3 years)
    - 2X increase in logic device performance every generation (2-3 yrs)