BROWN

Low Power VLSI System Design

Lectures 2 & 3: Sources of Power and Logic-Level Power Estimation

Prof. R. Iris Bahar
EN2912
September 11 & 13, 2017

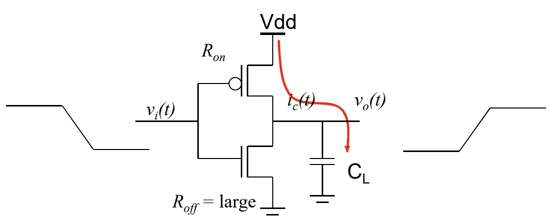
BROWN

Class webpage

- The webpage is now up and running:
  http://www.brown.edu/Departments/Engineering/Courses/engn2912
- Last week's lecture has been posted

BROWN

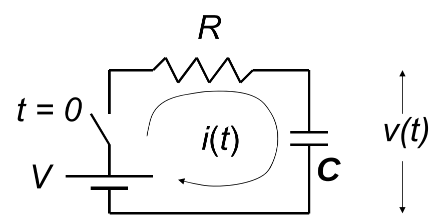
Power of a Transition: P_{tran}



What goes into charging/discharging the capacitance C_L?

BROWN

Charging of a Capacitor



Charge on capacitor, $q(t)$ = $C\,v(t)$

Current, $i(t)$ = $dq(t)/dt$ = $C\,dv(t)/dt$

BROWN

$$i(t) = C\frac{dv(t)}{dt} = \frac{[V - v(t)]}{R}$$
$$\frac{dv(t)}{dt} = \frac{V - v(t)}{RC}$$
$$\int \frac{dv(t)}{V - v(t)} = \int \frac{dt}{RC}$$
$$\ln[V - v(t)] = \frac{-t}{RC} + A$$

Initial condition, $t = 0,\ v(t) = 0 \rightarrow A = \ln V$

$$v(t) = V[1 - e^{\frac{-t}{RC}}]$$

BROWN

$$v(t) = V[1 - e^{\frac{-t}{RC}}]$$
$$i(t) = C\frac{dv(t)}{dt} = \frac{V}{R}e^{\frac{-t}{RC}}$$

BROWN

Total Energy Per Charging Transition from Power Supply

$$E_{trans} = \int_0^\infty Vi(t)dt = \int_0^\infty \frac{V^2}{R} e^{\left(\frac{-t}{RC}\right)} dt$$
$$E_{trans} = CV^2$$

BROWN

Energy Dissipated per Transition in Resistance

$$E = R\int_0^\infty i^2(t)dt = R\frac{V^2}{R^2}\int_0^\infty e^{\frac{-2t}{RC}} dt$$
$$E = \tfrac{1}{2}CV^2$$

Energy Stored in Charged Capacitor

$$E = \int_0^\infty v(t)i(t)dt = \int_0^\infty V\left[1 - e^{\frac{-t}{RC}}\right]\frac{V}{R}e^{\frac{-t}{RC}}dt$$

$$E = \tfrac{1}{2}CV^2$$

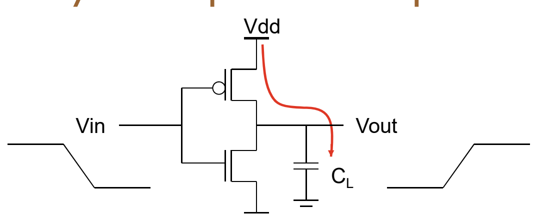
Transition Power

- Gate output rising transition
  - Energy dissipated in pMOS transistor = $CV^2/2$
  - Energy stored in capacitor = $CV^2/2$
- Gate output falling transition
  - Energy dissipated in nMOS transistor = $CV^2/2$
- Energy dissipated per transition = $CV^2/2$
- Power dissipation:

$$P_{trans} = E_{trans}\, \alpha\, f_{ck} = \alpha\, f_{ck}\, CV^2/2$$

$$\alpha = \text{activity factor}$$

Dynamic power dissipation



Energy/transition = $C_L * V_{DD}^2 * P_{0\to1}$

$f_{0\to1}$

P_{dyn} = Energy/transition * f = $C_L * V_{DD}^2 * P_{0\to1} * f$

$P_{dyn} = C_{EFF} * V_{DD}^2 * f$ where $C_{EFF} = P_{0\to1}\, C_L$

Data dependent → a function of switching activity!

Lowering dynamic power

Capacitance:
Function of fan-out, wire length, transistor sizes

Supply Voltage:
Has been dropping with successive generations

$P_{dyn} = C_L\, V_{DD}^2\, P_{0\to1}\, f$

Activity factor:
How often, on average, do wires switch?

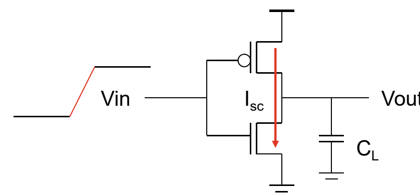
Clock frequency:
Recent scaling back…

Components of Power

- Dynamic
  - Signal transitions
    - Logic activity
    - Glitches
  - Short-circuit
- Static
  - Leakage

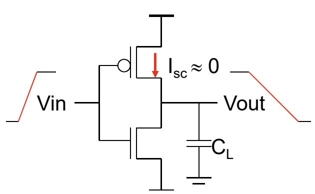
$$P_{total} = P_{dyn} + P_{stat}$$
$$= P_{tran} + P_{sc} + P_{stat}$$

Short circuit power dissipation

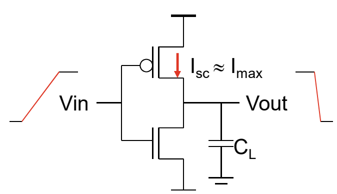


Finite slope of the input signal causes a direct current path between V_{DD} and GND for a short period of time during switching when both the NMOS and PMOS transistors are conducting.

Impact of C_L on P_{sc}

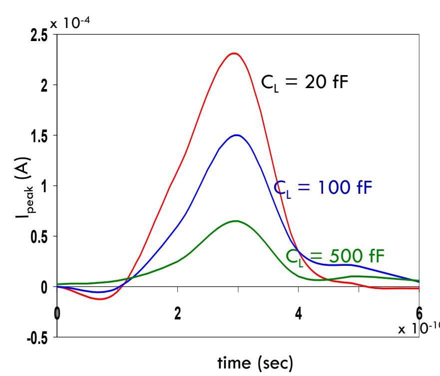


Large capacitive load
BUT BOTH DEVICES ARE ON FOR A SHORT TIME
Output fall time significantly larger than input rise time.



Small capacitive load
BOTH DEVICES ARE ON FOR A LONG TIME
Output fall time substantially smaller than the input rise time.

I_{peak} as a function of C_L



500 psec input slope

When load capacitance is small, I_{peak} is large.

Short circuit dissipation is minimized by matching the rise/fall times of the input and output signals - slope engineering.

BROWN

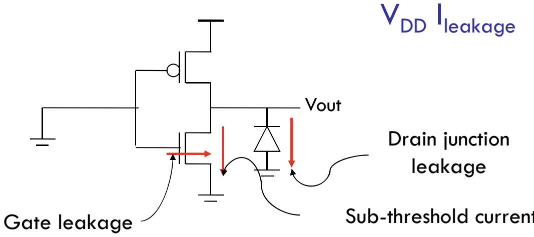
Summary: Short-Circuit Power

- Short-circuit power is consumed by each transition (increases with input transition time).
- Reduction requires that gate output transition should not be faster than the input transition (faster gates can consume more short-circuit power).
- Increasing the output load capacitance reduces short-circuit power.
- Has become less of an issue as voltages scale down.

BROWN

Leakage (static) power dissipation

$V_{DD}\,I_{leakage}$



Sub-threshold current is the dominant factor.

All increase exponentially with temperature!

BROWN

Leakage (static) power dissipation

- Reducing V_{DD} reduces dynamic power dissipation
- BUT… reduced V_{DD} also means reduced performance
- Improve performance by reducing V_{th}
- BUT… reduced V_{th} means increased leakage current

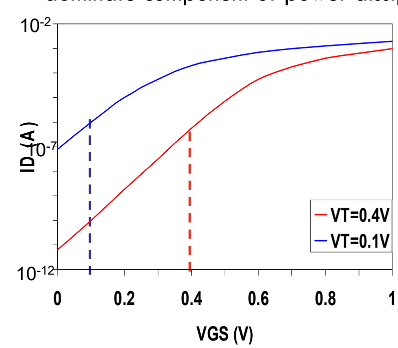
$$I_{leakage} \approx I_0 e^{\frac{V_{GS}-V_{th}}{nVT}}$$

- I_0 is a function of gate oxide, mobility, and size of device
- VT is the thermal voltage (26mV at T=300K)
- Even when V_{gs}=0 V, $I_{leakage}$ is non-zero. The closer V_{th} is to zero, the larger the leakage current at V_{gs} = 0 V.

BROWN

Leakage as a function of V_{TH}

- Continued scaling of supply voltage and the subsequent scaling of threshold voltage will make subthreshold conduction a dominate component of power dissipation.



- Notice 90mV/decade V_T roll-off
- Each 255mV increase in V_T gives 3 orders of magnitude reduction in leakage (but adversely affects performance)

Summary: Leakage Power

- Leakage power as a fraction of the total power increases as clock frequency drops.
  - *Turning supply off in unused parts can save power.*
- For a gate it is a small fraction of the total power; it can be significant for very large circuits.
- Scaling down features requires lowering the threshold voltage, which increases leakage power
  - roughly doubles with each shrinking.

Problem: A Design Example

- A battery-operated 65nm digital CMOS device is found to consume equal amounts (P) of dynamic power and leakage power while the short-circuit power is negligible. The energy consumed by a computing task, that takes T seconds, is 2PT.
- Compare two power reduction strategies for extending the battery life:
  - Clock frequency is reduced to half, keeping all other parameters constant.
  - Supply voltage is reduced to half. This slows the gates down and forces the clock frequency to be lowered to half of its original (full voltage) value. Assume that leakage current is held unchanged by modifying the design of transistors.

Solution: Part A. Clock Frequency Reduction

- Reducing the clock frequency will reduce dynamic power to P / 2, keep the static power the same as P, and double the execution time of the task.
- Energy consumption for the task will be,

  Energy = (P / 2 + P) 2T = 3PT

  which is greater than the original 2PT.

Solution: Part B. Supply Voltage Reduction

- When the supply voltage and clock frequency are reduced to half their values, dynamic power is reduced to P / 8 and static power to P / 2. The time of task is doubled and the total energy consumption is,

  Energy = (P / 8 + P / 2) 2T = 5PT / 4 =1.25PT
- The voltage reduction strategy reduces energy consumption while a simple frequency reduction consumes more energy.

Probabilistic Analysis

- Dynamic power dissipation:

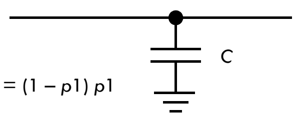
$P_{trans} = E_{trans}\, \alpha\, f_{ck} = \alpha\, f_{ck}\, CV^2/2$

α = *activity factor*

- *How do we compute this activity factor probabilistically?*
- View signals as a random processes

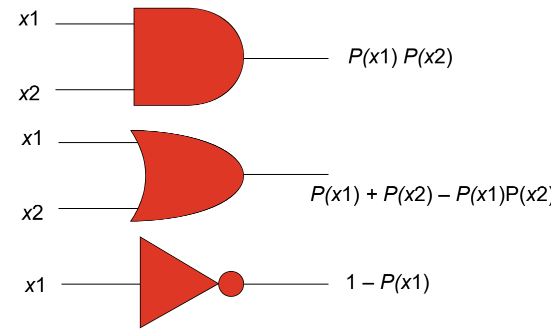
Prob{s(t) = 1} = $p1$
$p0 = 1 - p1$

0→1 static transition probability = $(1 - p1)\,p1$



Power, $P = (1 - p1)\,p1\,CV^2 f_{ck}$

Signal Probabilities

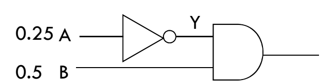


Transition Probabilities for Basic Gates

	$P_{0\to1} = P_{out=0} \times P_{out=1}$
NOR	$(1 - (1 - P_A)(1 - P_B)) \times (1 - P_A)(1 - P_B)$
OR	$(1 - P_A)(1 - P_B) \times (1 - (1 - P_A)(1 - P_B))$
NAND	$P_AP_B \times (1 - P_AP_B)$
AND	$(1 - P_AP_B) \times P_AP_B$
XOR	$(1 - (P_A + P_B - 2P_AP_B)) \times (P_A + P_B - 2P_AP_B)$

Transition Prob. for Basic Gates

	$P_{0\to1} = P_{out=0} \times P_{out=1}$
AND	$(1 - P_AP_B) \times P_AP_B$



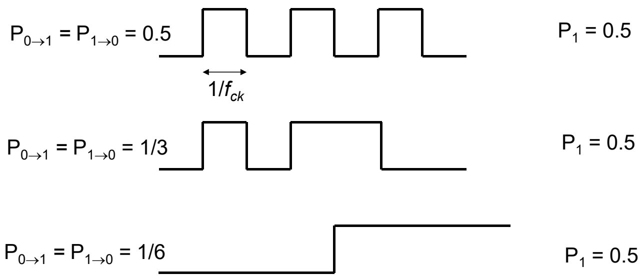
For Y: $P_1(Y) = 0.75$

$P_{0\to1}(Y) = P_0(A) \times P_1(A) = (1 - P_1(A))\,P_1(A)$
$= 0.75 \times 0.25 = 0.1875$

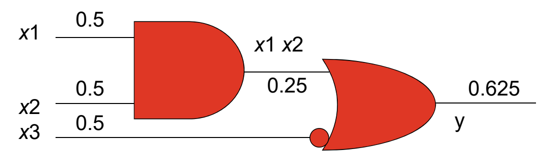
For Z: $P_1(Z) = P_1(Y)\,P_1(B) = 0.75 \times 0.5 = 0.375$

$P_{0\to1}(Z) = P_0(Z) \times P_1(Z) = (1 - P_1(Y)\,P_1(B)) \times P_1(Y)\,P_1(B)$
$= (1 - (0.75 \times 0.5)) \times (0.75 \times 0.5) = 0.234$

Transition and Signal Probabilities



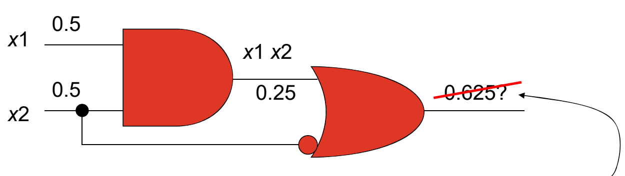
Signal Probabilities



X1	*X2*	*X3*	*Y*
0	0	0	1
0	0	1	0
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	1

P(y) = *x1x2 + (1 – x3) – x1x2(1 – x3)*
= *x1x2 + (1 – x3) – x1x2 + x1x2x3*
= 1 - *x3 + x1x2x3*
= 0.625

Correlated Signal Probabilities

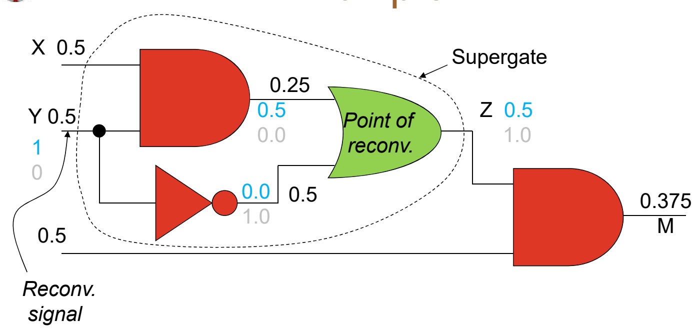


X1	*X2*	*Y*
0	0	1
0	1	0
1	0	1
1	1	1

P(y) = *x1x2 + (1 – x2) – x1x2(1 – x2)*
= *x1x2 + (1 – x2) – x1x2 + x1x2x2*
= (1 – *x2*) + *x1x2*
= *0.5 + 0.25* = 0.75 (correct value)

Propagating probabilities may produce an incorrect result if there is reconvergent fanout

Example



Z = xy + y' ➔ P(Z) = .5(.5) + .5 = .75
M = (xy + y')(w)
M = y(xw) + y'(w) ➔ P(M) = .5(.5)(.5) + (.5)(.5) = .375

Transition Density

- $T \quad = 2\, p1\, (1 - p1) = p0 \;\; p01 + p1 \;\; p10$

  $= 2\, p10 \;\; p01 \;/\; (p10 + p01)$

  $= 2\, p1 \;\; p10 = 2\, p0 \;\; p01$

- How do you compute p01 or p10 ?
  - You need to know the switching activity (transition density) of the primary inputs and then you can propagate this forward.

Transition Density

- For some output $y=f(x_i, \ldots, x_n)$, the signal activity (i.e. transition density) of y, T(y), can be computed as:

$$T(y) = \sum_{i=1}^{n} P\left(\frac{\partial y}{\partial x_i}\right) T(x_i)$$

- In other words, we need to sum the contribution of switching activity at y due to each input x_i.

Boolean Difference

$$\text{Boolean diff}(Y, X_i) = \frac{\partial Y}{\partial X_i} = Y(X_i=1) \oplus Y(X_i=0)$$

- Boolean diff(Y, X_i) = 1 means that a path is sensitized from input X_i to output Y.
- Prob(Boolean diff(Y, X_i) = 1) is the probability of transmitting a toggle from X_i to Y.
- Probability of Boolean difference is determined from the probabilities of cofactors of Y with respect to X_i.

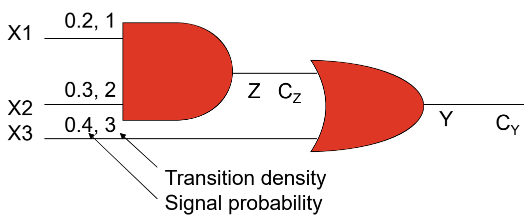
Dynamic Power Computation

- For each primary input, determine signal probability and transition density for given vectors.
- For each internal node and primary output Y, find the transition density T(Y), using supergate partitioning and the Boolean difference formula.
- Compute power,

$$\text{Power} = \sum_{\text{all } Y} 0.5 C_Y V^2 T(Y)$$

where C_Y is the capacitance of node Y and V is supply voltage.

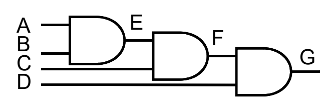
Transition Density and Power



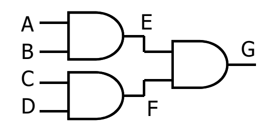
- What are the signal probabilities for Z and Y?
- What are the transition densities for Z and Y?
- What is the total power dissipated for the circuit?

Problem 1

- For equiprobable inputs analyze the 0→1 transition probabilities of all gates in the two implementations of a four-input AND gate shown below.
- Assuming that the gates have zero delays, which implementation will dissipate less average dynamic power?



Chain structure



Tree structure

Problem 1 Solution

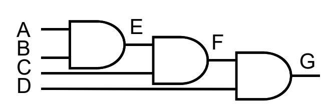
Given the primary input probabilities, P(A) = P(B) = P(C) = P(D) = 0.5, signal and transition (0→1) probabilities are as follows:

Signal name	Chain		Tree	
	Prob(sig.= 1)	Prob(0→1)	Prob(sig.=1)	Prob(0→1)
E	0.2500	0.1875	0.2500	0.1875
F	0.1250	0.1094	0.2500	0.1875
G	0.0625	0.0586	0.0625	0.0586
Total transitions/vector		0.3555		0.4336

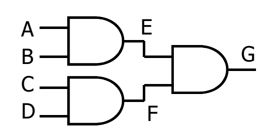
Note that it's just signal F that matters here
The tree implementation dissipates 100×(0.4336 – 0.3555)/0.3555 = 22% more average dynamic power.
This advantage of the chain structure may be somewhat reduced because of glitches caused by unbalanced path delays.

Problem 2

- Assume that the two-input AND gates in Problem 1 each has one unit of delay. Find input vector pairs for each implementation that will dissipate the peak dynamic power.
- Which implementation dissipates less peak dynamic power?



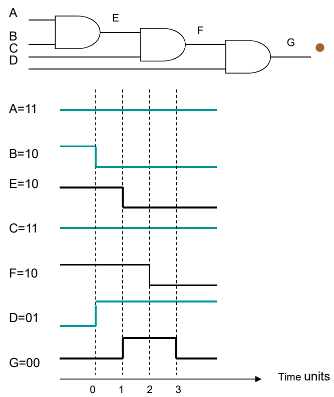
Chain structure



Tree structure

BROWN

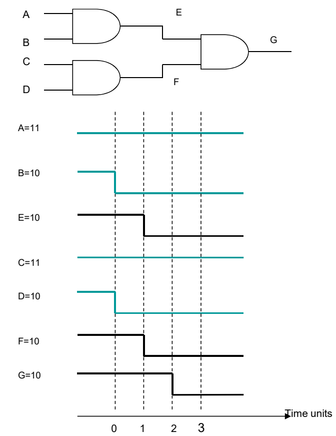
Problem 2 Solution



- For the chain structure, a vector pair
  {A B C D} = {1110},{1011}
  will produce four gate transitions as shown below.

BROWN

Problem 2 Solution (Cont.)



- The tree structure has balanced delay paths, so it cannot make more than 3 gate transitions.
  {ABCD} = {1111},{1010}
  will produce three transitions
- The chain dissipates
  100(4 – 3)/3 = 33% higher peak power than the tree.

BROWN

Power Calculation

- Power can be estimated if transition density is known for all signals.
- Calculation of transition density requires
  - Signal probabilities
  - Transition densities for primary inputs; computed from vector statistics

BROWN

Observation

- Numerical computation of signal probabilities is accurate for fanout-free circuits.
- What about circuits with gates having multiple fanouts?

Remedies

- Use Shannon's expansion theorem to compute signal probabilities.
- Use Boolean difference formula to compute transition densities.

EN2912
Lecture 1-58

Shannon's Expansion Theorem

- Consider:
  - Boolean variables, $x_1, x_2, \ldots, x_n$
  - Boolean function, $f(x_1, x_2, \ldots, x_n)$
- Then $f = x_i\, f(x_i = 1) + x_i'\, f(x_i = 0)$
- Where
  - x_i' is the complement of x_i
  - Cofactors, $f(x_i = k) = f(x_1, x_2, \ldots, x_i = k, \ldots, x_n)$, k = 0 or 1

EN2912
Lecture 1-59

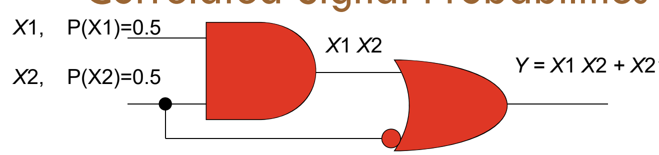
Expansion About Two Inputs

$f = x_i x_k\, f(x_i=1, x_k=1)\ +$
$\quad x_i x_k'\, f(x_i=1, x_k=0) +$
$\quad x_i' x_k\, f(x_i=0, x_k=1) +$
$\quad x_i' x_k'\, f(x_i=0, x_k=0)$

- In general, a Boolean function can be expanded about any number of input variables.
- Expansion about n variables will have 2^n terms.

EN2912
Lecture 1-60

Correlated Signal Probabilities



X1	X2	Y
0	0	1
0	1	0
1	0	1
1	1	1

Shannon expansion about the reconverging input, *X2*:

Y = X2 Y(X2=1) + X2' Y(X2=0)
  = X2 (X1) + X2' (1)

P(Y) = P(X2) (0.5) + P(1-X2) (1)
  = 0.5 (0.5) + (1-0.5) (1)
  = 0.75

EN2912
Lecture 1-61

BROWN

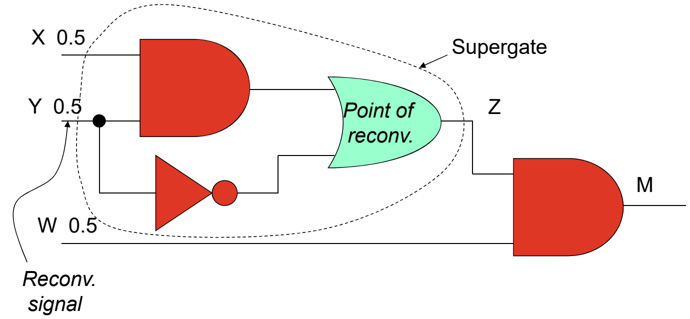
Correlated Signals

- When the output function is expanded about all reconverging input variables,
  - All cofactors correspond to fanout-free circuits.
  - Signal probabilities for cofactor outputs can be calculated without error.
  - A weighted sum of cofactor probabilities gives the correct probability of the output.
- For two reconverging inputs:

  $f = x_i x_k\ f(x_i=1, x_k=1) + x_i x_k'\ f(x_i=1, x_k=0)$
  $\quad + x_i'\ x_k\ f(x_i=0, x_k=1) + x_i' x_k'\ f(x_i=0, x_k=0)$

BROWN

Example

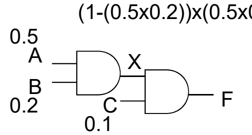


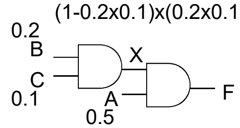
Z = xy + y'　　　　M = (xy + y')(w)

BROWN

Input Ordering

(1-(0.5x0.2))x(0.5x0.2)=0.09　　　(1-0.2x0.1)x(0.2x0.1)=0.0196





- Beneficial to postpone the introduction of signals with a high transition rate
- Note that transition activity at node F is the same for both implementations.

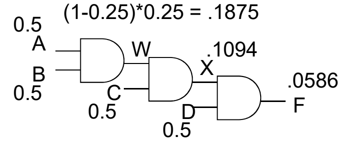
BROWN

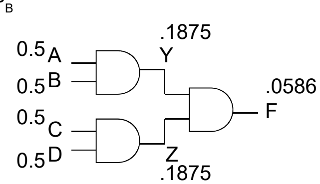
Logic Restructuring Revisited

- Logic restructuring: Changing the topology of a logic network to reduce transitions

AND: $P_{0\rightarrow 1} = P_0 \times P_1 = (1 - P_A P_B) \times P_A P_B$

(1-0.25)*0.25 = .1875

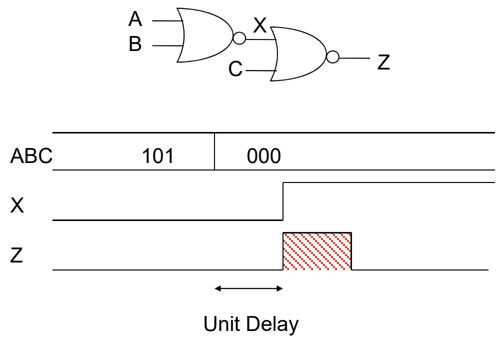




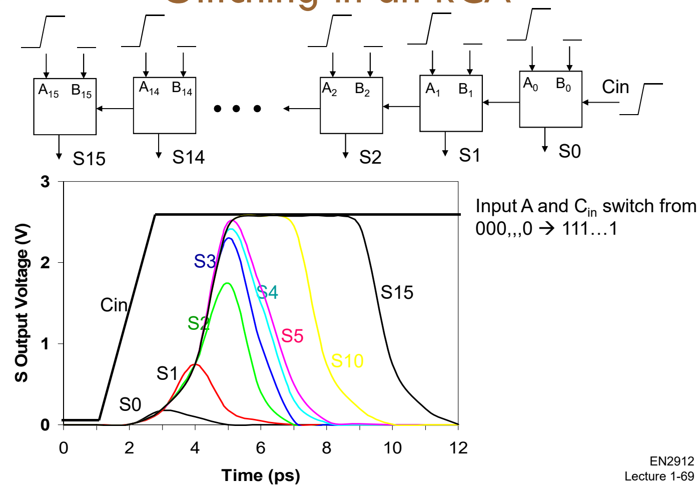
- Chain implementation has a lower overall switching activity than the tree implementation for random inputs
- Ignores glitching effects

Glitching in Static CMOS Networks

- Gates have a nonzero propagation delay resulting in spurious transitions or glitches (dynamic hazards).

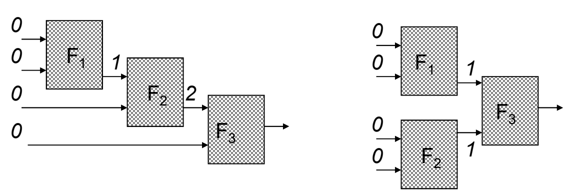


Glitching in an RCA



Balanced Delay Paths to Reduce Glitching

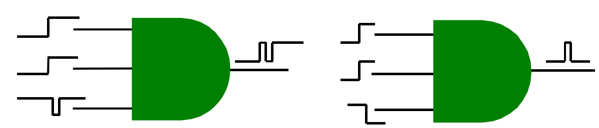
- Glitching is due to a mismatch in the path lengths in the logic network; if all input signals of a gate change simultaneously, no glitching occurs



- So equalize the lengths of timing paths through logic

Theorem 1

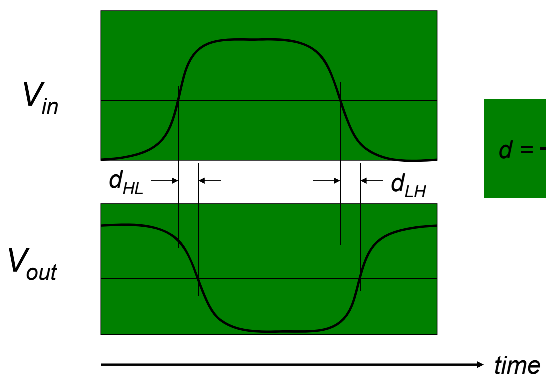
- For correct operation with minimum energy consumption, a Boolean gate must produce no more than one event per transition.



Output logic state changes
One transition is necessary

Output logic state unchanged
No transition is necessary

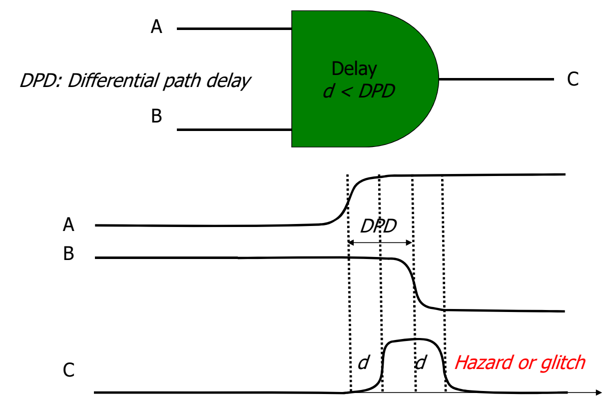
Inertial Delay of an Inverter



$$d = \frac{d_{HL}+d_{LH}}{2}$$

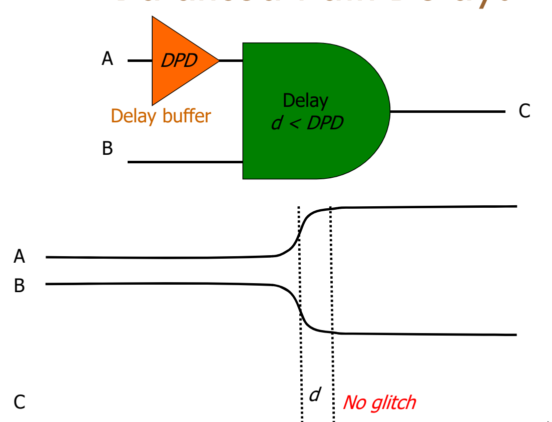
EN2912
Lecture 1-74

Multi-Input Gate



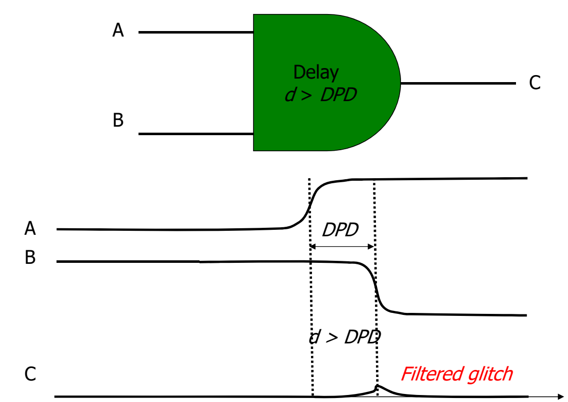
EN2912
Lecture 1-75

Balanced Path Delays



EN2912
Lecture 1-76

Glitch Filtering by Inertia

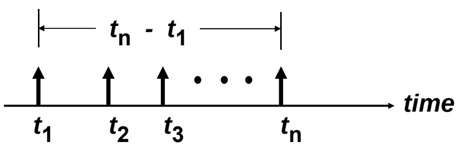


EN2912
Lecture 1-77

Theorem 2

- Consider a gate with n inputs. Given that events occur at the input of the gate with inertial delay d at times, $t_1 \leq . . . \leq t_n$, the number of events at the gate output cannot exceed

$$\min\left(n,\ 1 + \left\lfloor \frac{t_n - t_1}{d} \right\rfloor\right)$$



Minimum Transient Design

- Minimum transient energy condition for a Boolean gate:

$$| t_i - t_j | < d$$

Where
t_i and t_j are arrival times of input events and
d is the inertial delay of gate