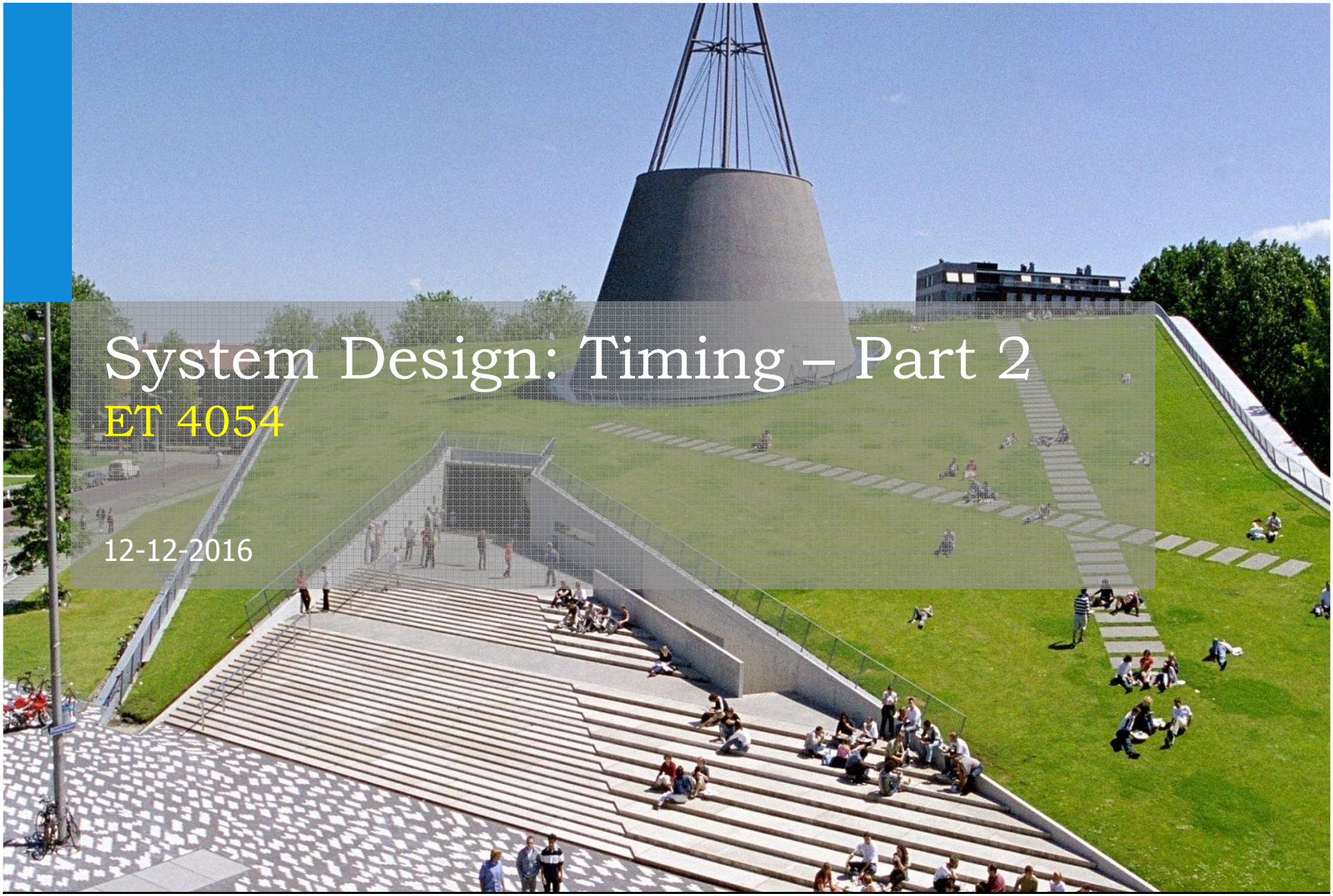


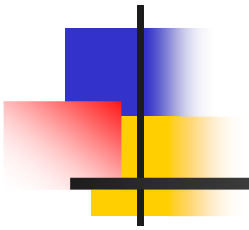
System Design: Timing – Part 2

ET 4054

12-12-2016

Overview

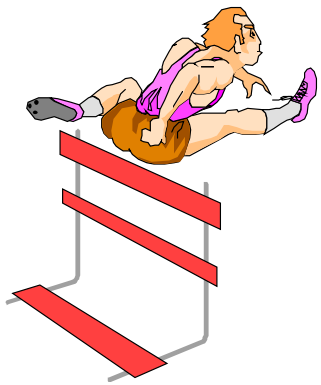
- Design Constraints
  - Power, Area, Frequency, CMOS Scaling
- Timing
  - Timing Metrics, Paths, Variability and Delay
- Deterministic Timing Analysis (Static Timing Analysis)
  - Models, Interconnect, Networks, Clock Distribution
- Statistical Timing Analysis
  - Probability, Spatial Correlations, MAX function
- Design Flow
  - Synthesis, Transformation, Definitions, Constrains



Deterministic Timing Analysis

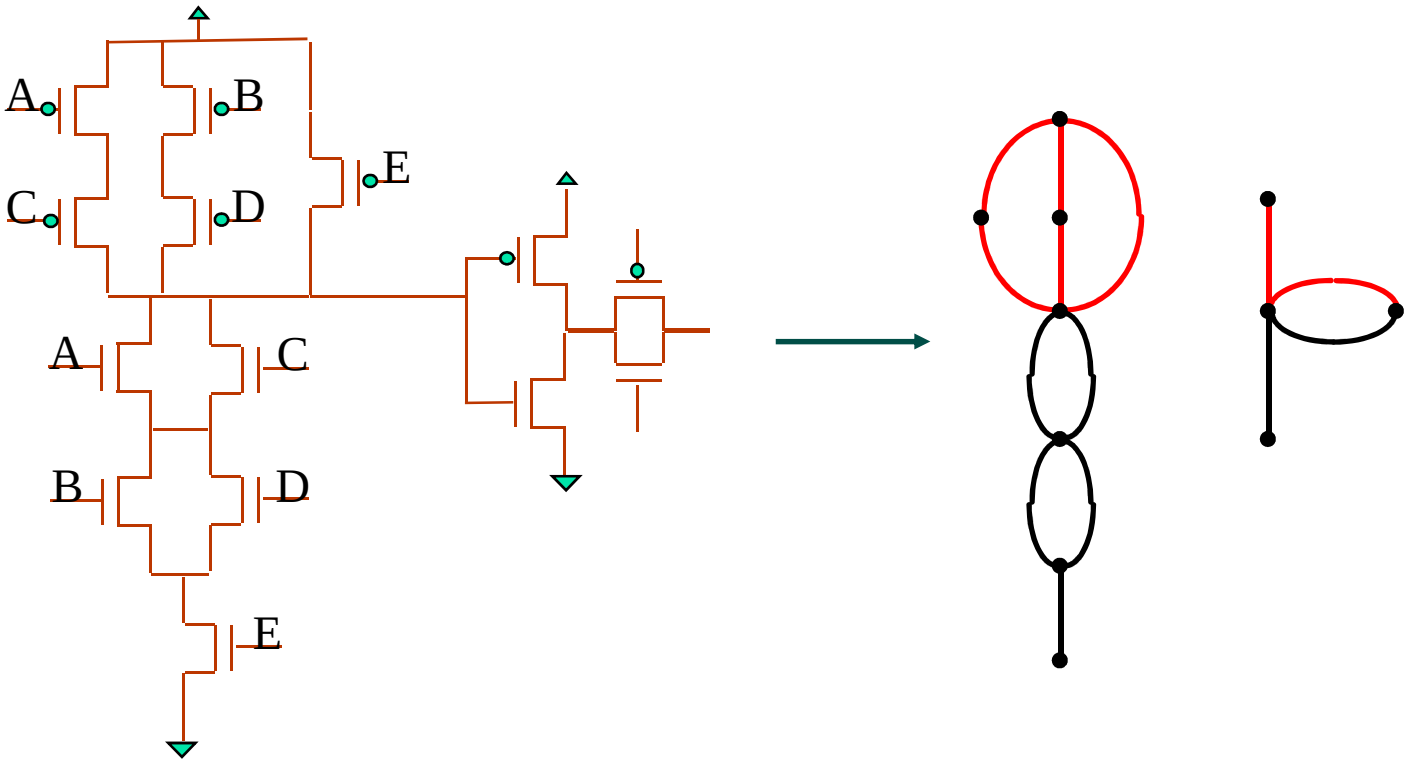
Static Timing Analysis

- Problem
  - Given a transistor level description of combinational circuit, find arrival times at all gate outputs (pattern independent)
- Solution
  - Clump transistors together into fundamental gates **("channel-connected components")**
  - Propagate timing information (low **and** high polarities from primary inputs (PI's) to primary outputs (PO's) **("critical path method (CPM)")**
  - Overcome miscellaneous hurdles along the way



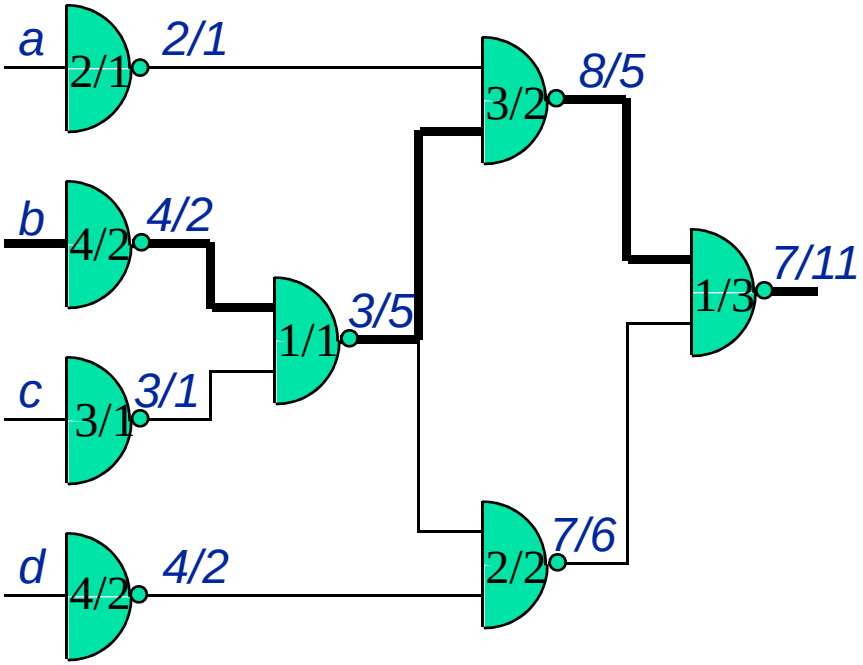
Channel-Connected Components

- A set of transistors interconnected by drain/source nodes



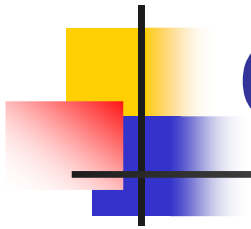
The Critical Path Method (CPM)

- Respond to minor changes in event-driven manner



Propagate changes
by propagating events

```
incremental_analyze(g)
GATE g;
{
if (arr_time@output(g) changes) {
    arr_time = newvalue;
    for (i ∈ fanout(g))
      incremental_analyze(i);
}
```

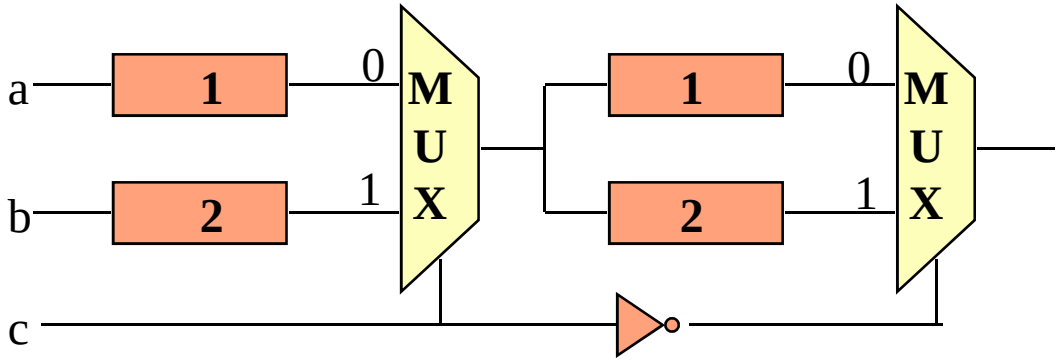


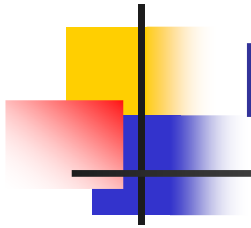
Questions

- What simplifications are made here?
- What are the implications?

False Paths (Briefly!)

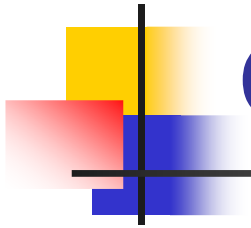
- Classic example





False Paths II

- No efficient algorithms exist to automatically flag false paths
- Knowledge like “this block will not run in two modes at the same time” is often crucial to determine false paths
- So: you need to specify them by hand…



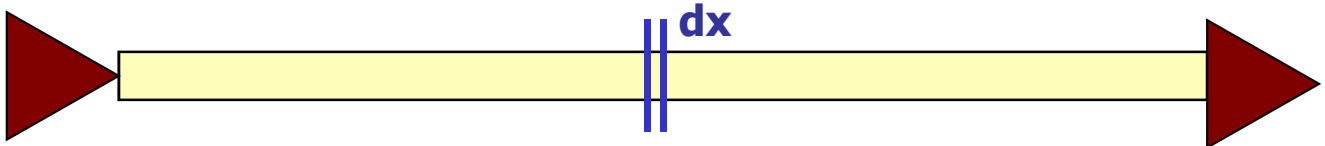
Gate Delay Models

- Build delay models for individual gates (current source/voltage source models)
  - In reality, Delay = f(widths, transition times, loads,…)
  - Similar idea used in standard cell characterization:

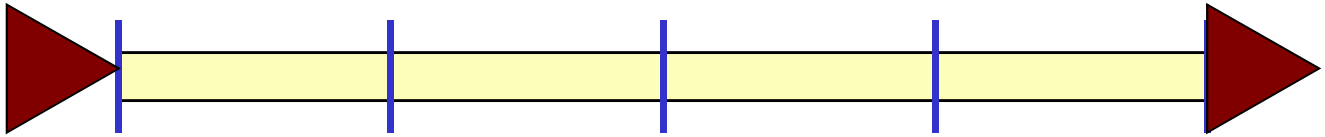
    Delay = f (transition times, load)
- Table lookup models: storage/accuracy tradeoff (e.g. .lib format)
- Fast circuit simulation – used in many delay calculators

Interconnect Modeling

- Precise model requires transmission line analysis

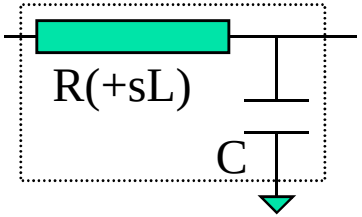


- Break up wire into segments (distributed model)

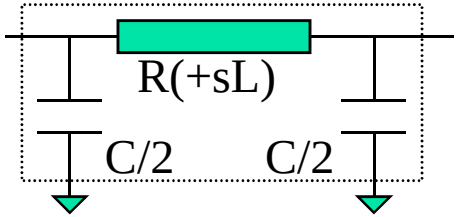


- Each segment can be modeled as

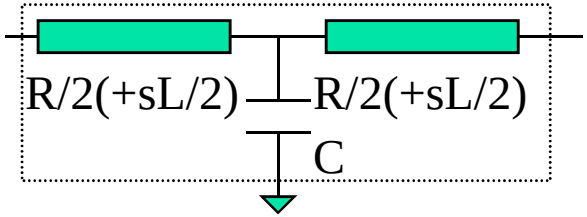
L-model



π-model



T-model



- Example: 5% delay accuracy requires 4 π-segments or 64 L-segments
- π-model has fewer circuit nodes than T-model
- Other issues (crosstalk etc.) modeled using coupling caps

Interconnect Modeling

Assume: Wire modeled by N equal-length segments

$$\tau_{DN} = \left(\frac{L}{N}\right)^2 (rc + 2rc + \ldots + Nrc) = (rcL^2)\frac{N(N+1)}{2N^2} = RC\frac{N+1}{2N}$$

For large values of N:

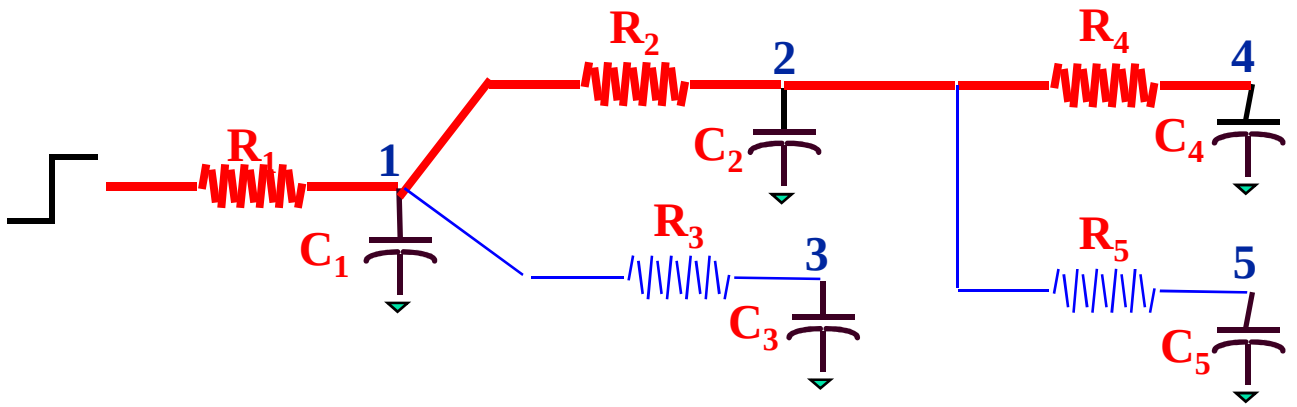
$$\tau_{DN} = \frac{RC}{2} = \frac{rcL^2}{2}$$

- Quadratic function of length
- Distributed model - ½ of the delay predicted by lumped RC model

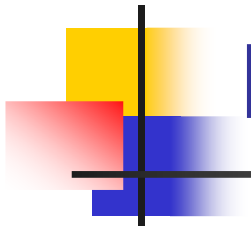
Elmore Delay Computations

- For an RC tree:

$$t_d = \Sigma_{i \text{ on path}} R_i \; C_{downstream,i}$$



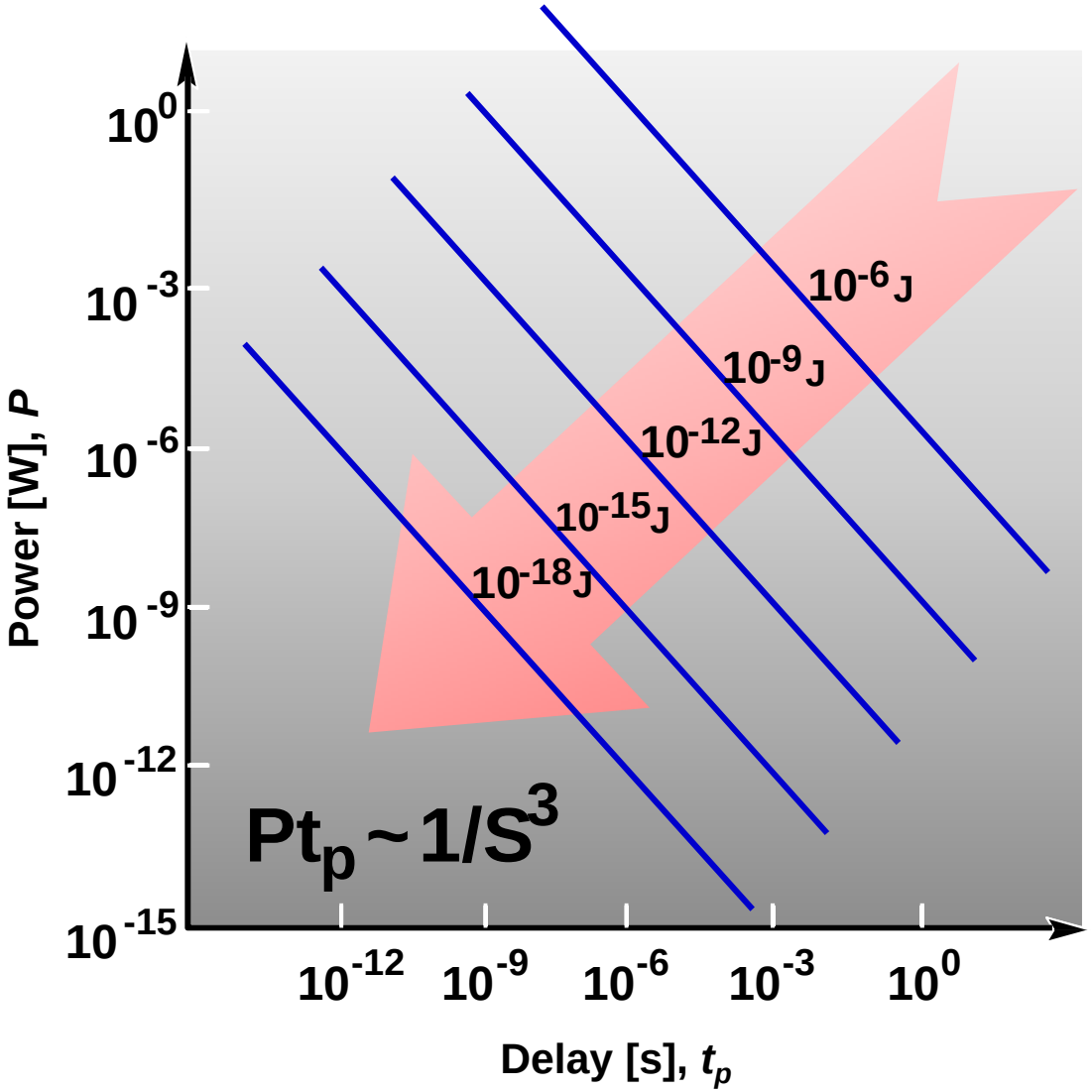
$$t_{d,4} = R_1 (C_1+C_2+C_3+C_4+C_5) + R_2 (C_2+C_4+C_5) + R_4C_4$$



Notes

- Elmore Delay inaccurate for the very resistive wires of modern technologies!
- Crosstalk has significant impact as well in modern technologies, as the wires are so close together.

Logic Scaling

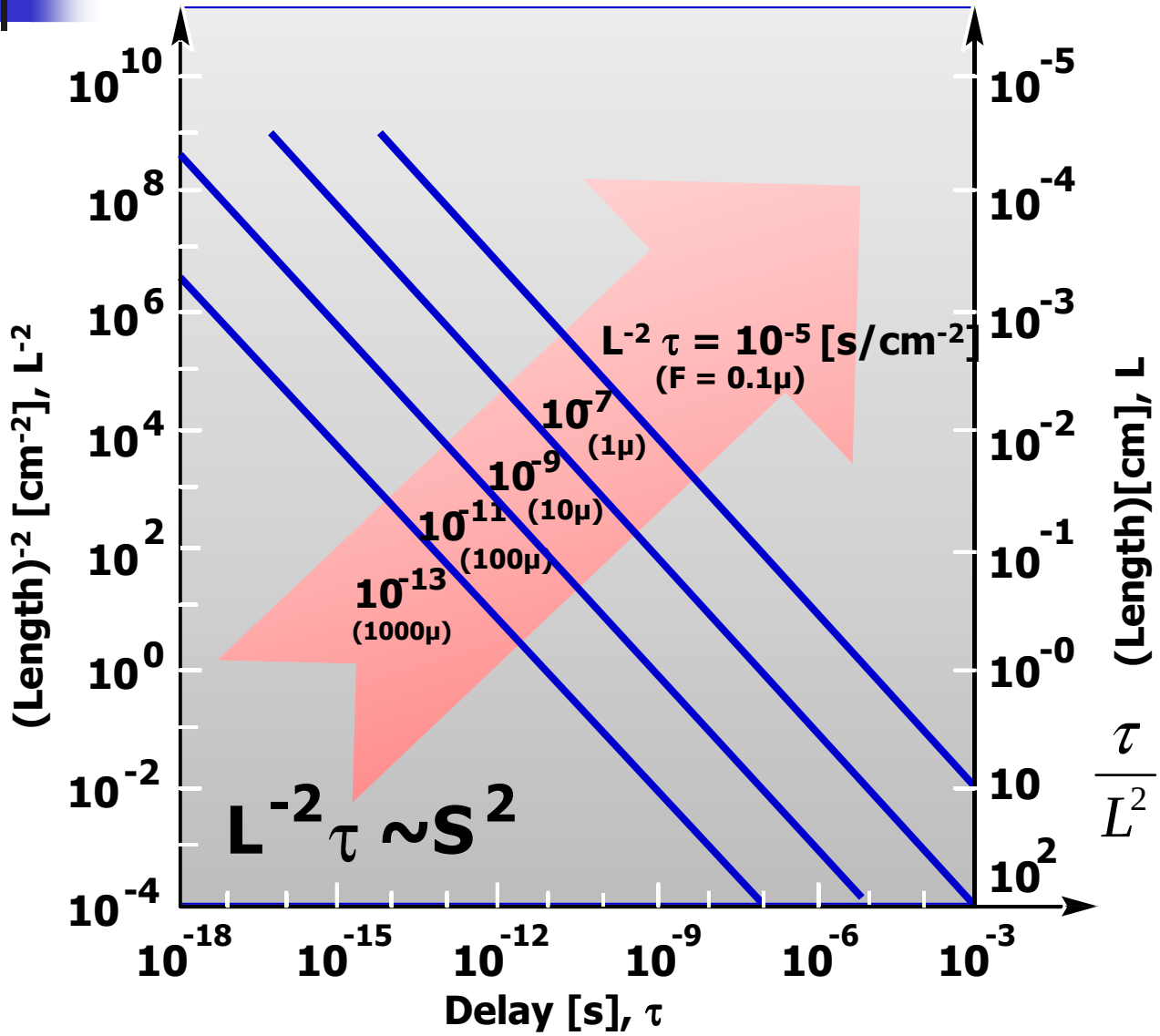


Logic Scaling

Constant Field Scaling: S = U

Parameter	Relation	General Scaling
W, L, t_{ox}		$1/S$
V_{DD}, V_T		$1/U$
C_{gate}	$C_{ox} W L$	$1/S$
I_{sat}	$C_{ox} W V$	$1/U$
R_{on}	V / I_{sat}	1
Power / Device	$I_{sat} V$	$1/U^2$

Interconnect Scaling

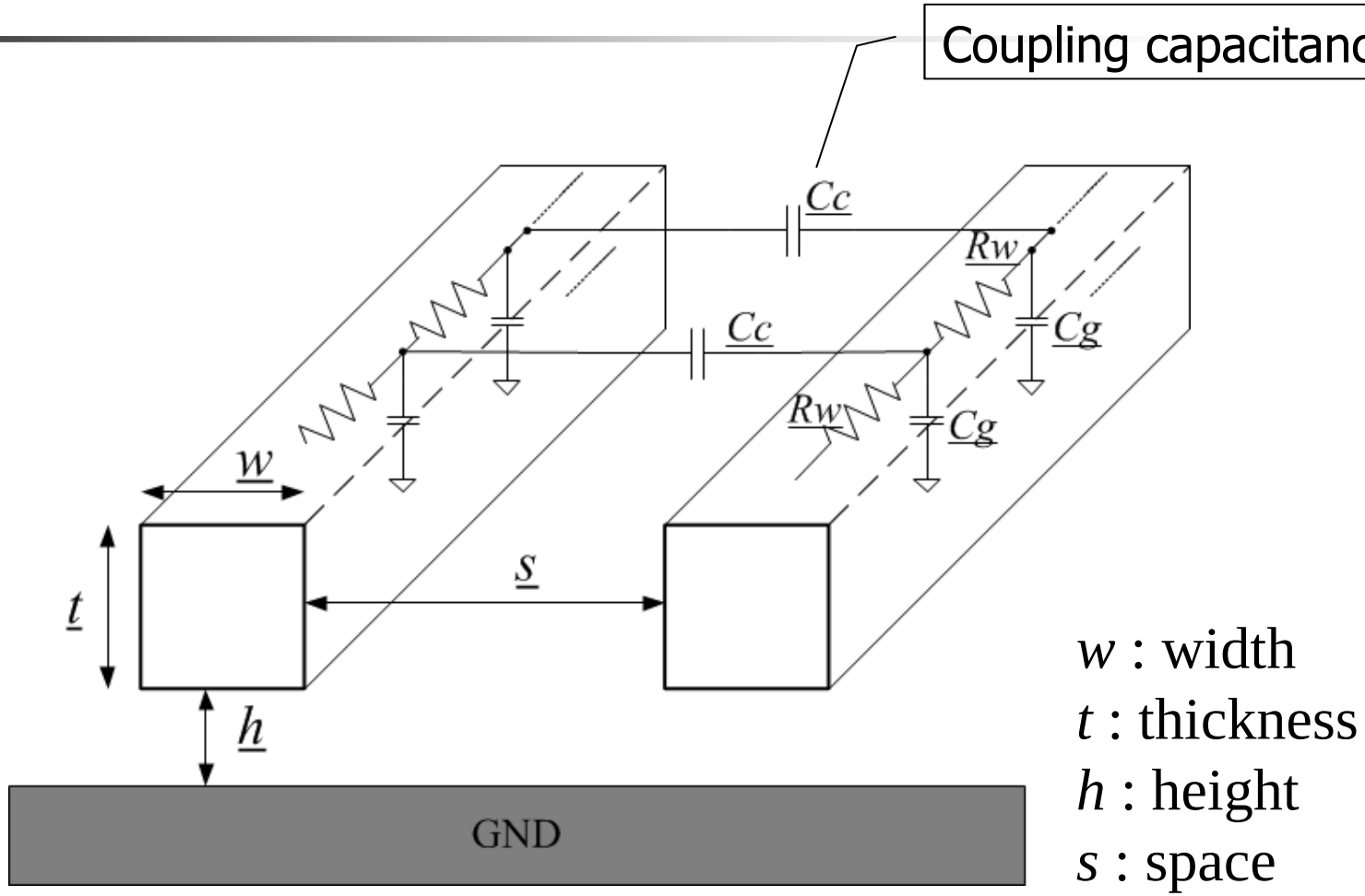


$$\frac{\tau}{L^2} = r_w c_w = \frac{\rho\varepsilon}{HT} \propto S^2$$

Idealized Wire Scaling Model

Parameter	Relation	Local Wire	Constant Length	Global Wire
W, H, T		*1/S*	*1/S*	*1/S*
L		*1/S*	*1*	$1/S_C$
C	*LW/T*	*1/S*	*1*	$1/S_C$
R	*L/WH*	*S*	S^2	S^2/S_C
$t_p \sim CR$	L^2/HT	*1*	$\underline{S^2}$	S^2/S_C^2
E	CV^2	$1/SU^2$	$1/U^2$	$1/(S_C U^2)$

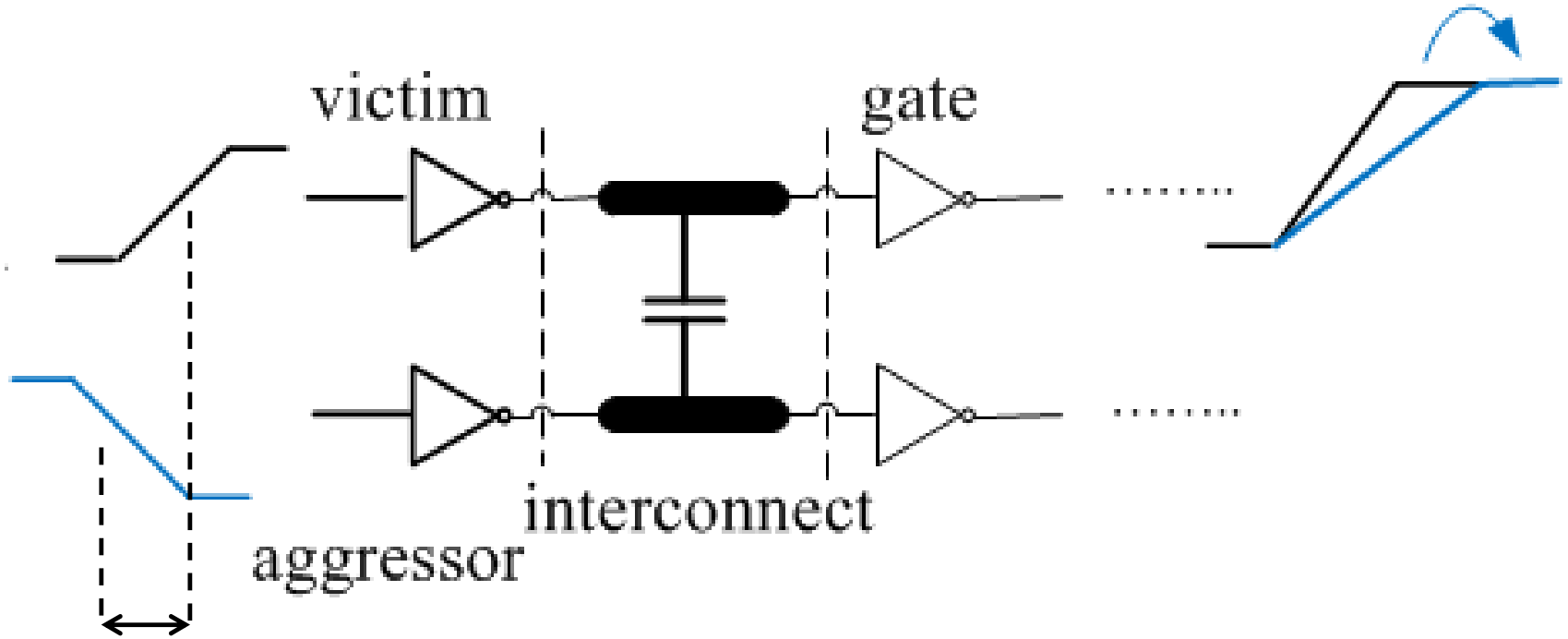
Crosstalk



w : width
t : thickness
h : height
s : space

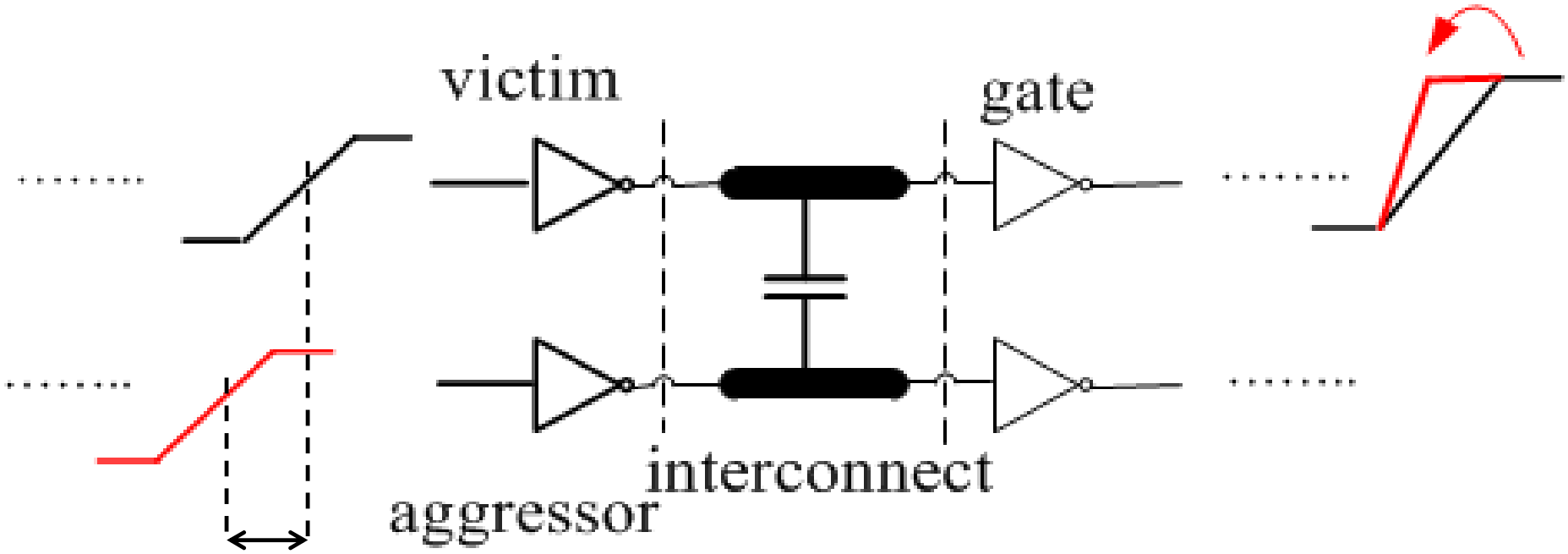
Interconnect Structure and RC model

The Impact of Crosstalk on Delay



Input skew: arrival time difference

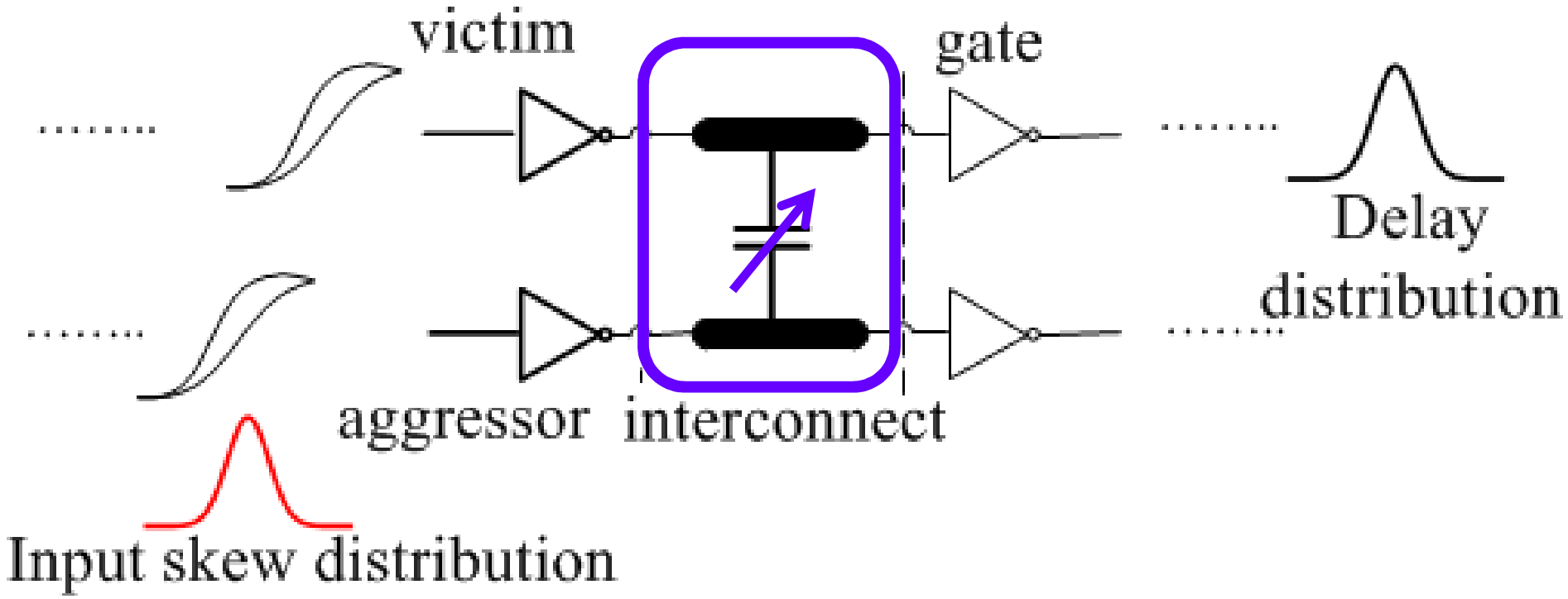
The Impact of Crosstalk on Delay

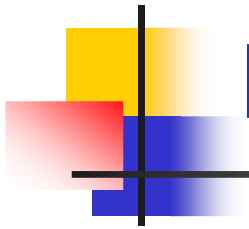


Input skew: arrival time difference

Crosstalk and Variability

- Increasing crosstalk effects
- Increasing process variation (length, Vth, width, thickness)

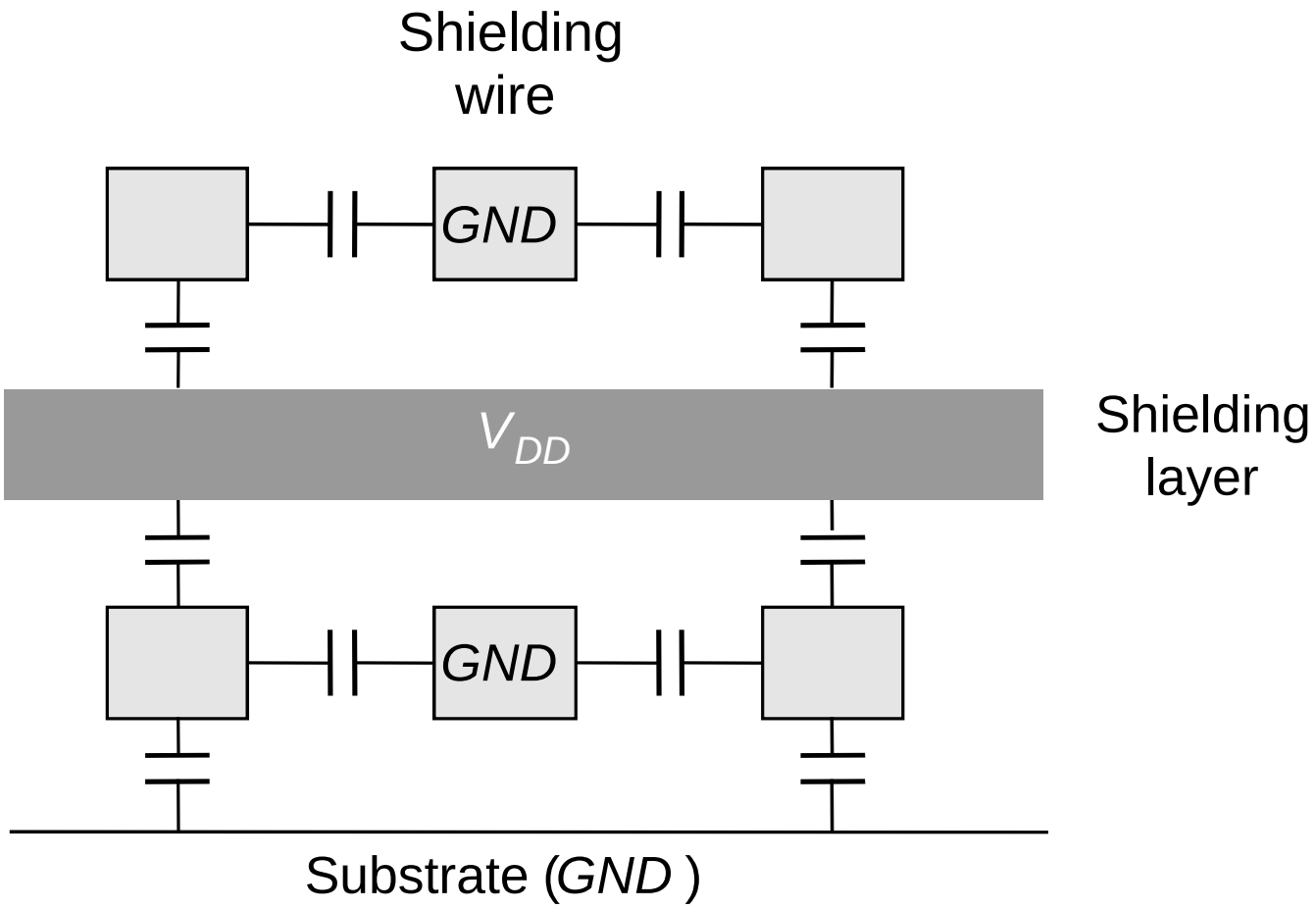


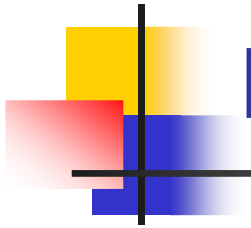


Dealing with Capacitive Crosstalk

- Avoid floating nodes
- Protect sensitive nodes
- Differential signaling
- Do not run wires together for a long distance
- Use shielding wires
- Use shielding layers

Shielding

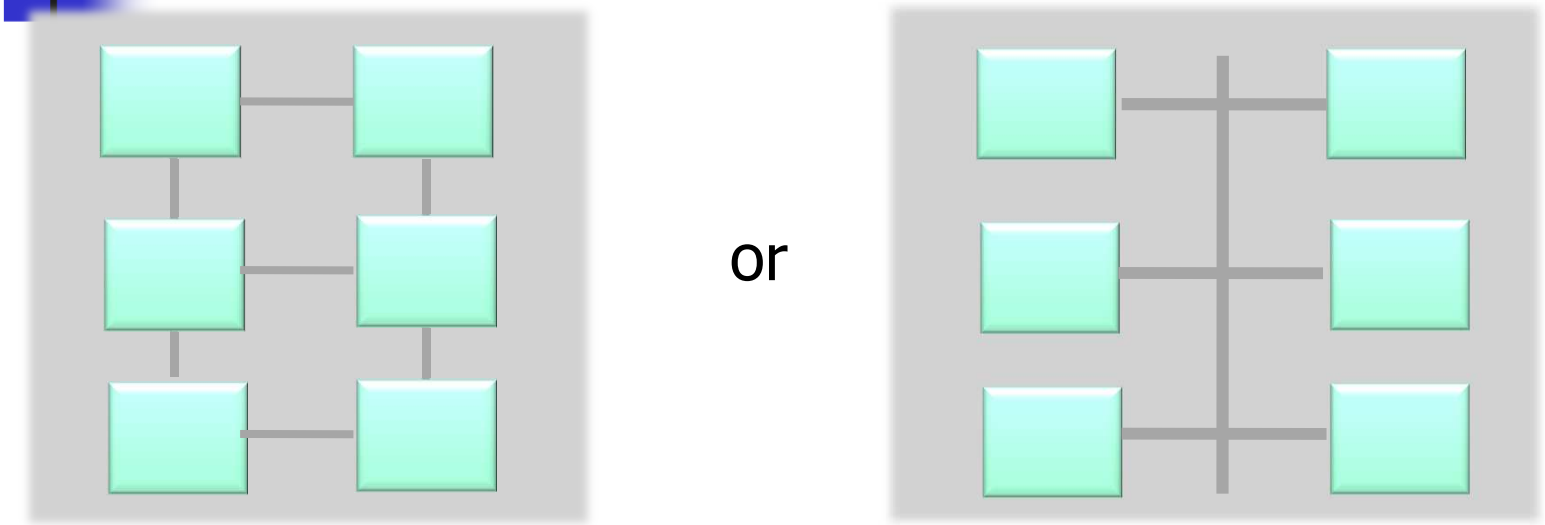




Reducing Interconnect Power/Energy

- Same philosophy as with logic: reduce capacitance, voltage (or voltage swing) and/or activity
- A major difference: sending a bit(s) from one point to another is fundamentally a **communications /networking problem**, and it helps to consider it as such.
- Abstraction layers are different:
  - For computation: device, gate, logic, micro-architecture
  - For communication: wire, link, network, transport
- Helps to organize along abstraction layers, well understood in the networking world: the OSI (open system interconnection) protocol stack
- Some exciting possibilities for the future: 3D-integration, novel interconnect materials, optical or wireless I/O

Network-on-a-Chip (NoC)

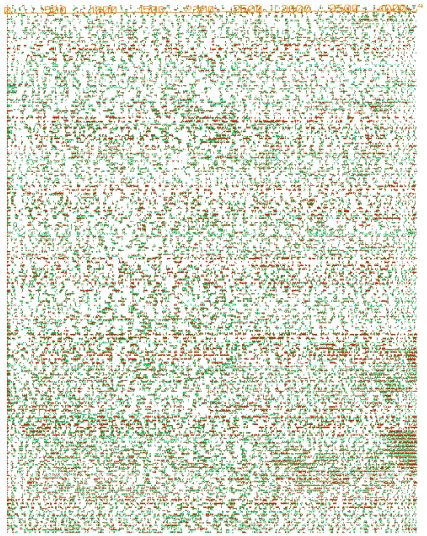


- Point-to-point or time-multiplexed bus network
- Dedicated networks with reserved links preferable for high traffic channels – but: limited connectivity, area overhead
- Flexibility an increasing requirement in multi (many) – core chip implementations

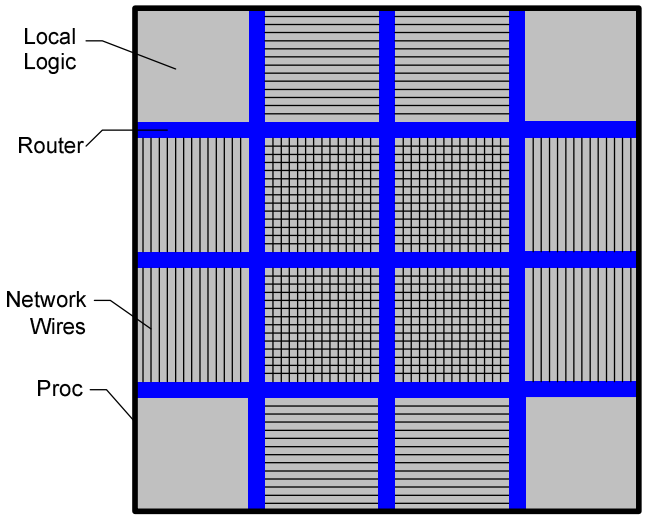
The Network Trade-off's

Trades-off flexibility, latency, energy and area-efficiency through:

- Locality - eliminate *global* structures
- Hierarchy - expose locality in communication requirements
- Concurrency/Multiplexing (function/area) – optimal reuse of resources



Dedicated wiring

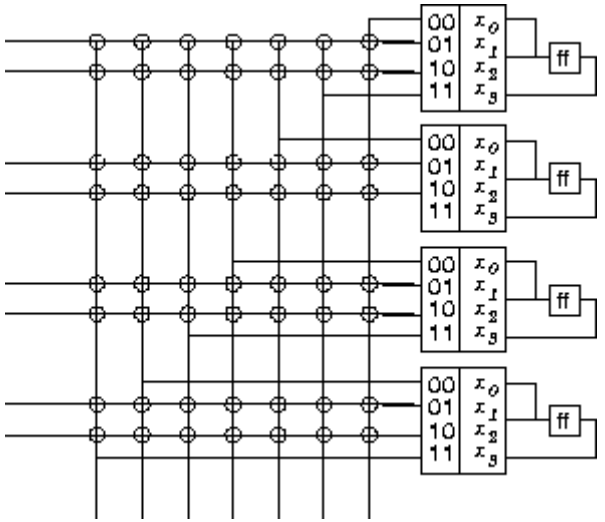


Network-on-a-Chip

[Courtesy: B. Dally, Stanford]

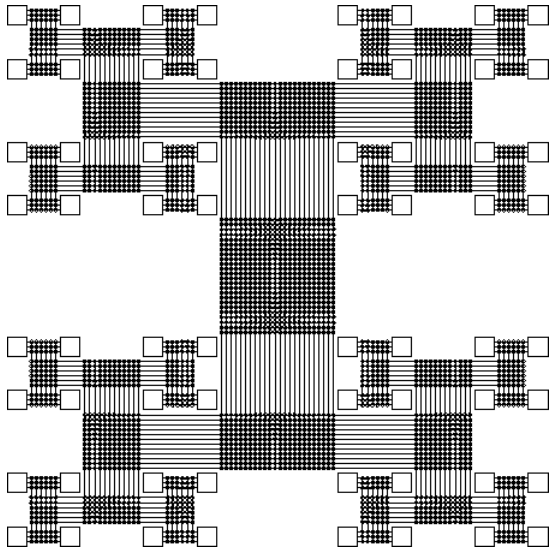
Networking Topology

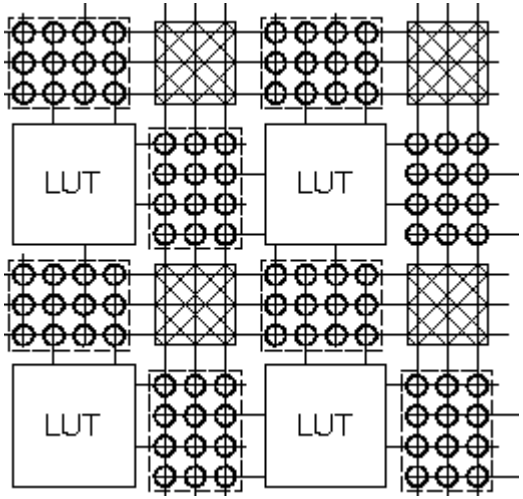
- Homogeneous
  - Bus, Star, Tree, Ring, Crossbar, Mesh (granularity), …
- Heterogeneous
  - Hierarchy



Crossbar

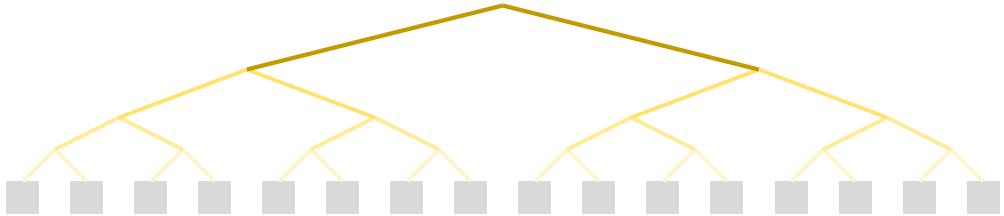
Tree



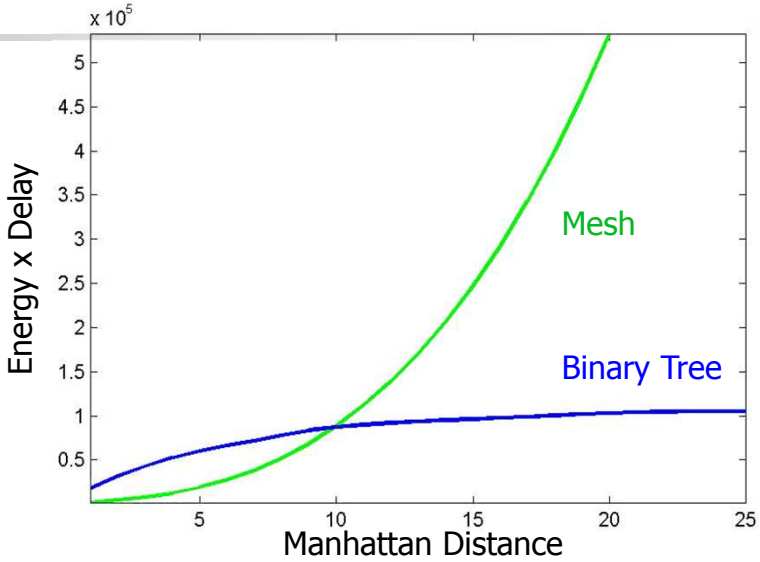


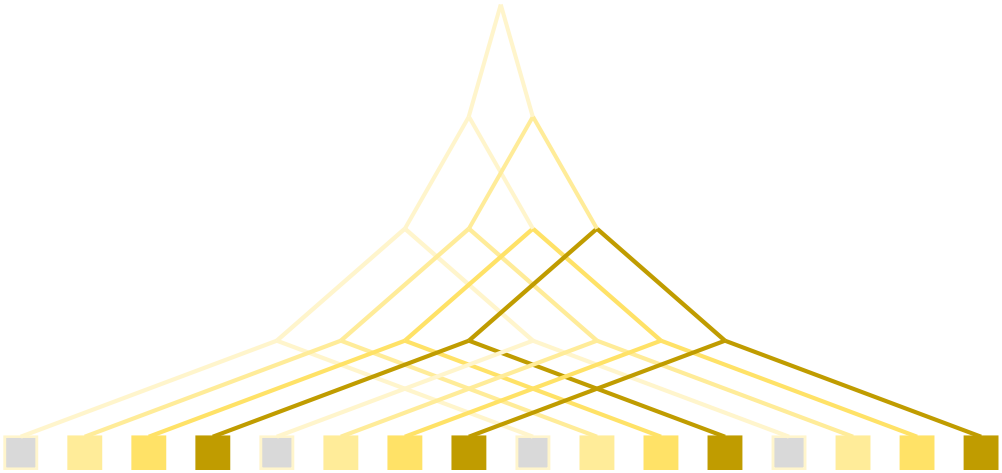
Mesh (FPGA)

Network Topology Exploration

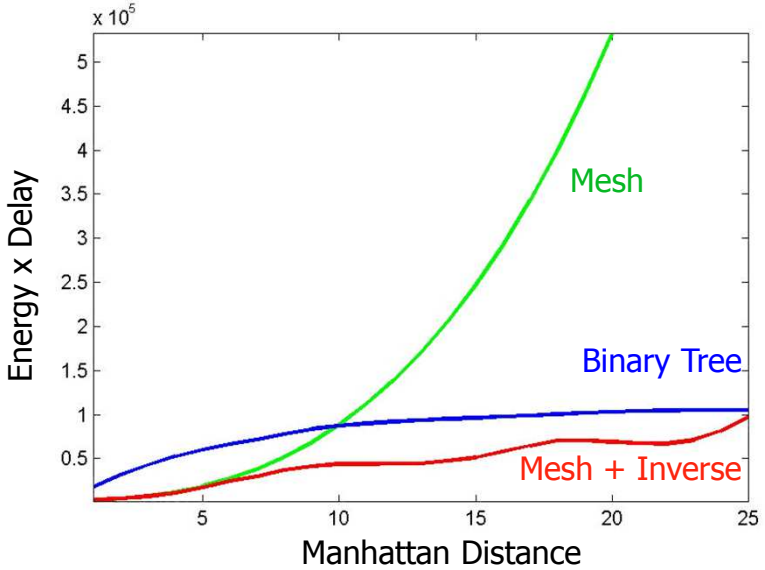


Short connections in tree are redundant



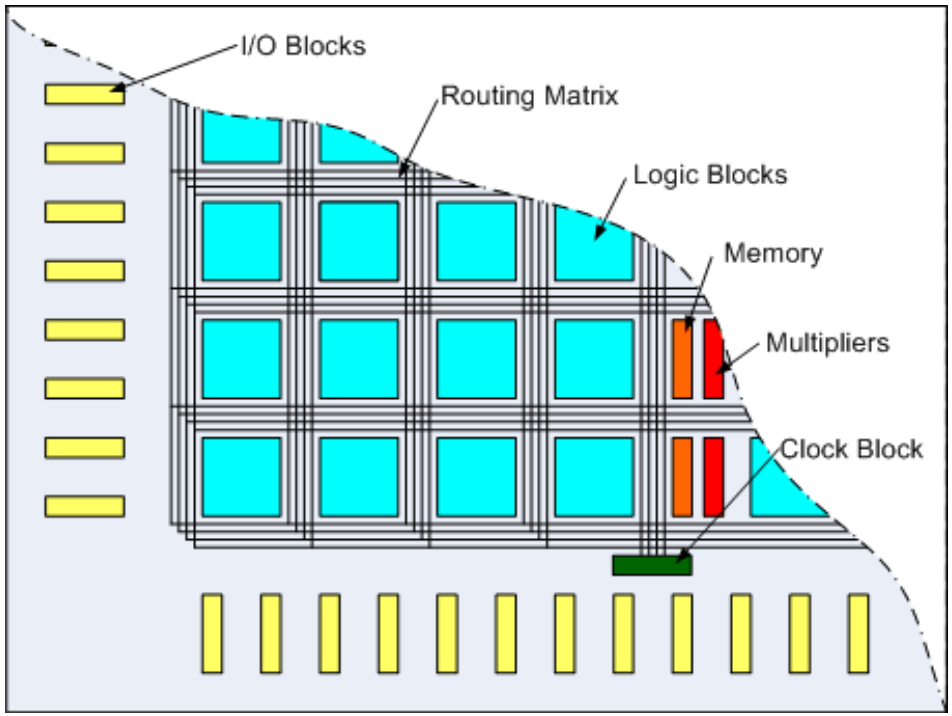


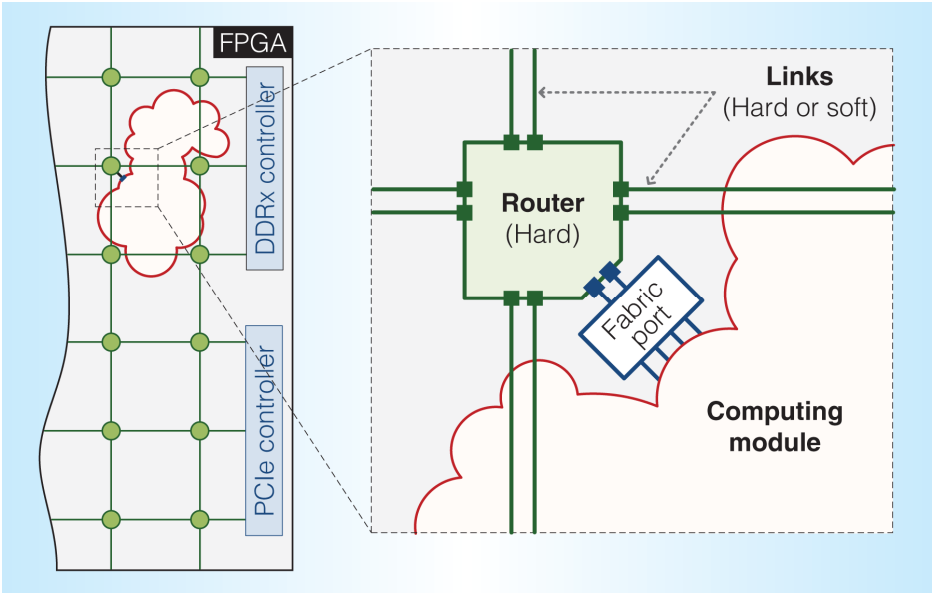
Inverse clustering complements mesh



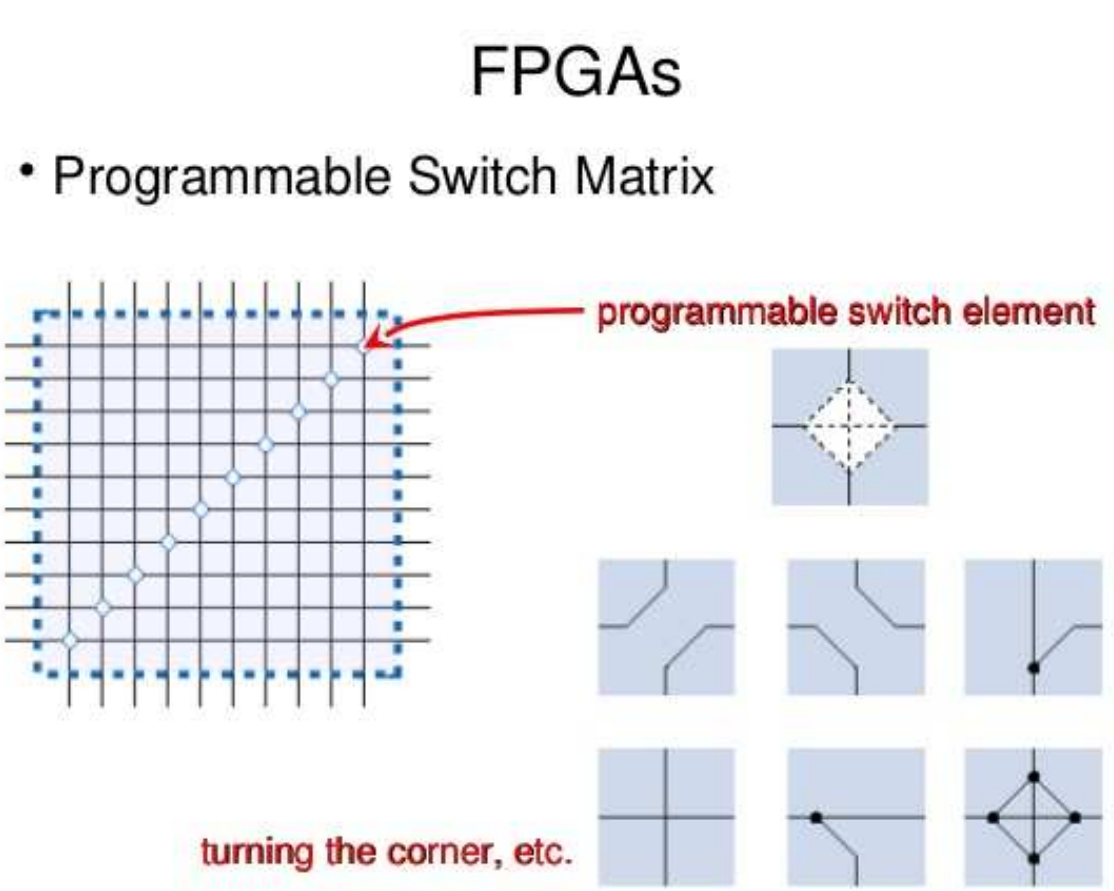
[Ref: V. George, Springer'01]

FPGA Architecture



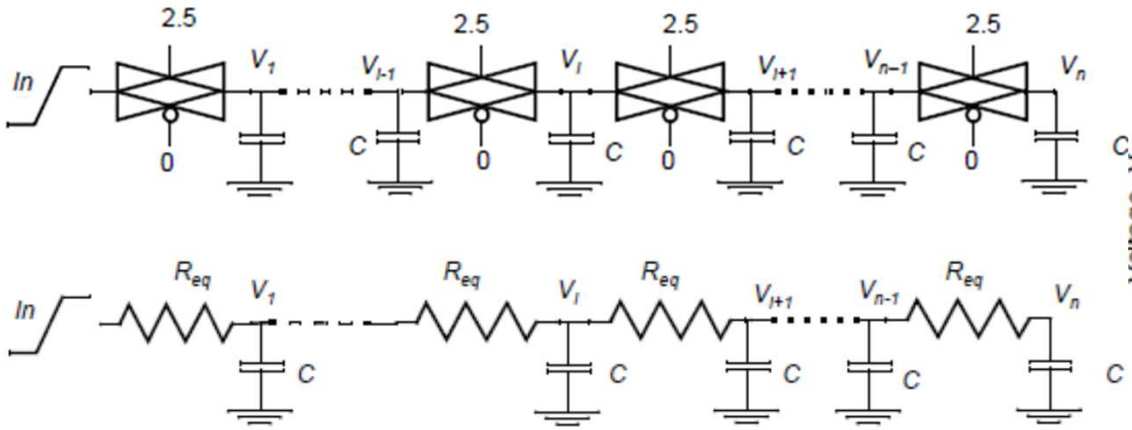


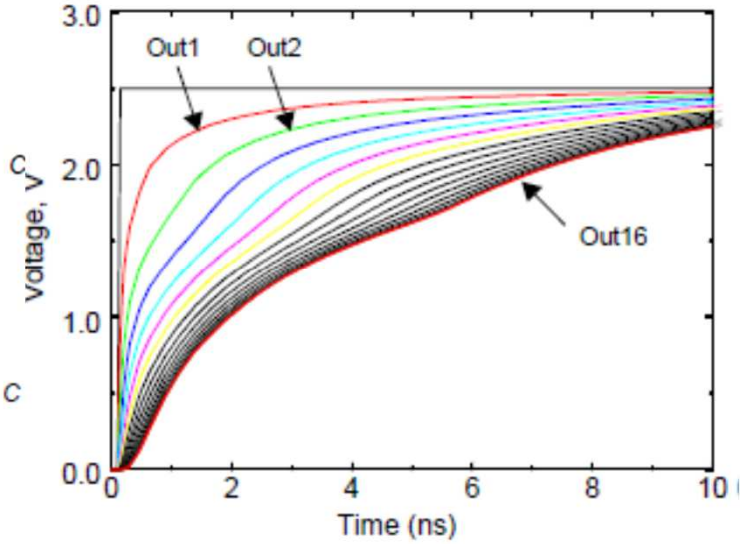
FPGA PSM and Delay



- The switches in the switch matrix are small (pass-) transistors.
- In FPGA's, the switch matrices in the connections add considerable resistance and hence delay!

(Pass-) Transistor Logic Delay





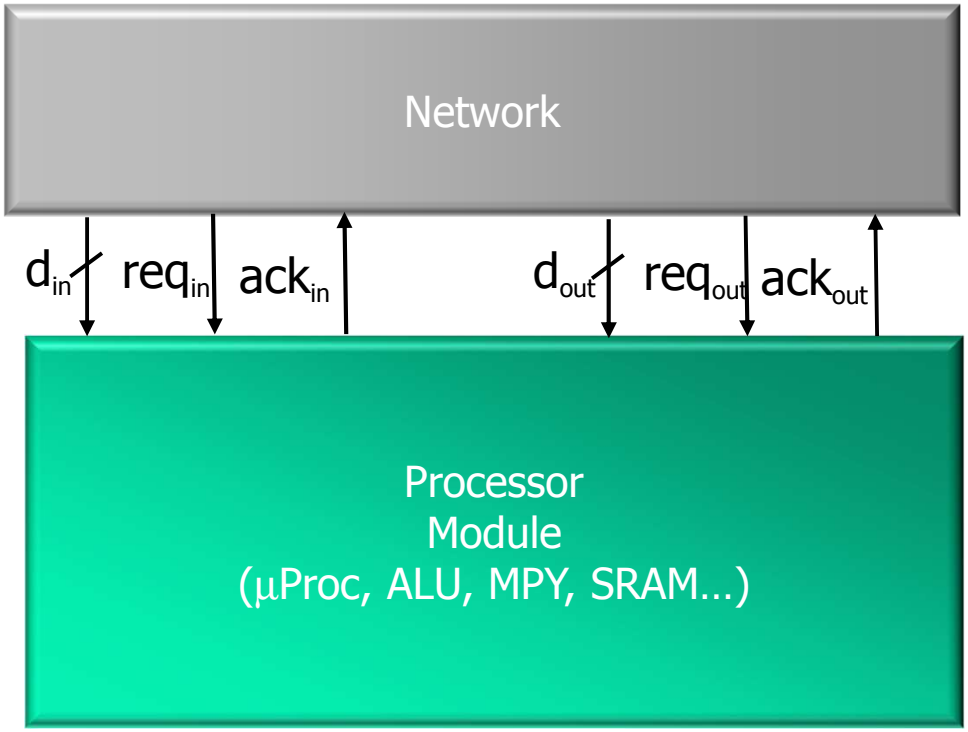
$$t_p(V_n) = 0.69\sum_{k=0}^{n} kCR_{eq} = 0.69CR_{eq}\frac{n(n+1)}{2}$$

- **Propagation delay is proportional to n^2!**
- **Insert buffers**

$$m_{opt} = 1.7\sqrt{\frac{t_{pbuf}}{CR_{eq}}}$$

- **In current technologies, m_{opt} is typically 3 or 4**

Signaling Protocols



Globally Asynchronous

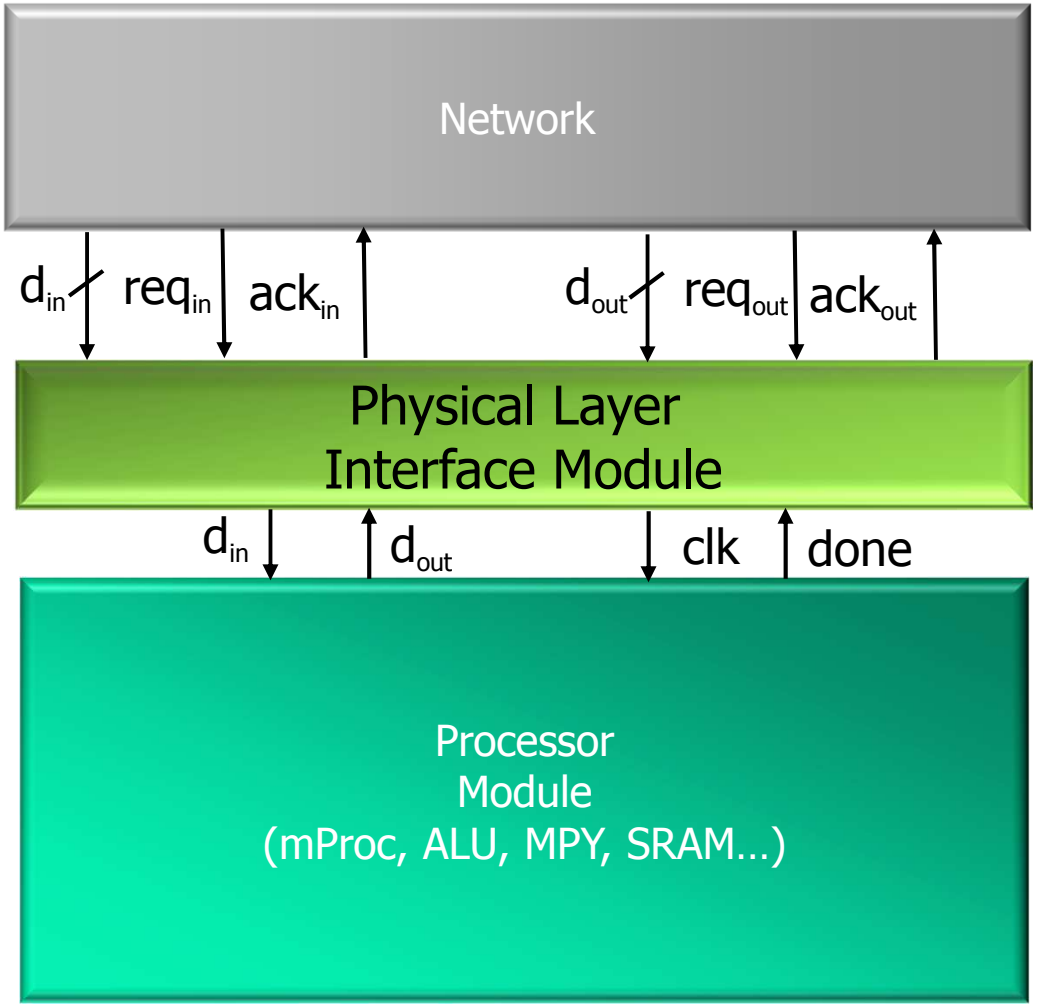
self-timed handshaking protocol

Locally Synchronous

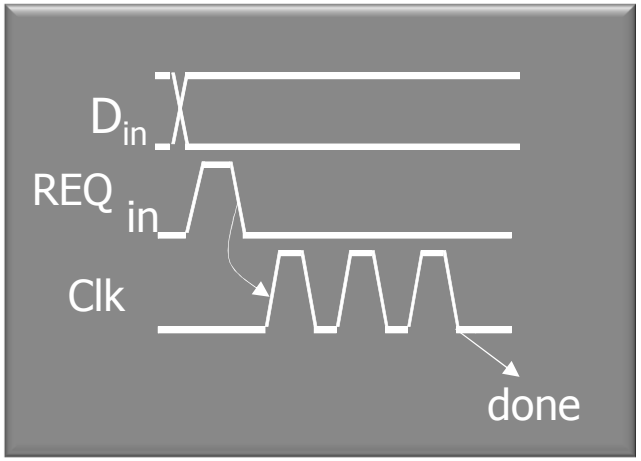
Allows individual modules to dynamically trade-off performance for energy-efficiency

- 1 cm chip: signal need 66 ps from one side to the other (transmission line)
- (*rc* effects dominate) – 500 ps → 2 GHz clock
- Pipelining by inserting clocked buffer elements
  - Complicates timing, links timing of glob. interconnect and loc. computation
  - Hampers introduction of power reduction techniques

Signaling Protocols



Globally Asynchronous



Locally synchronous

Future: Exploring the Unknown – Alternative Computational Models

Humans

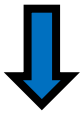
- Brain performs amazingly well
  - Under very low SNR conditions
  - Adapts effectively to failure and changing conditions

Concurrency

Brain-inspired computing

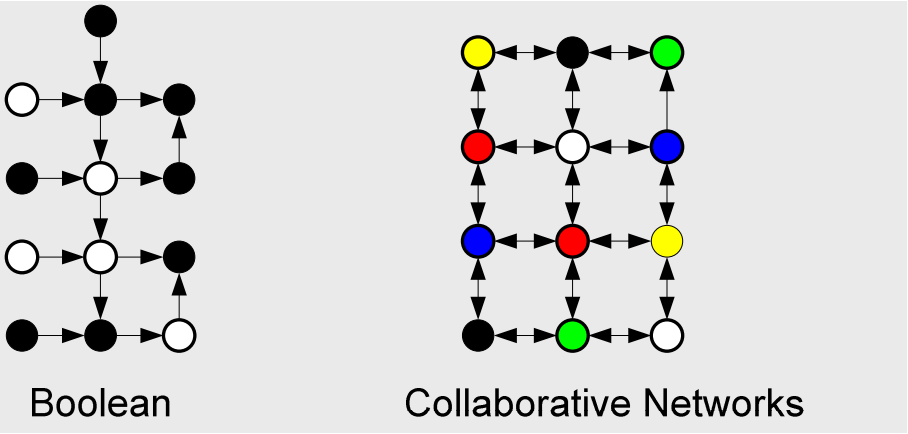
Ants

Instead of a system based on a small number of very reliable and complex components



A complex reliable system can emerge from the communication between huge numbers of simple nodes

Future: Collaborative Networks



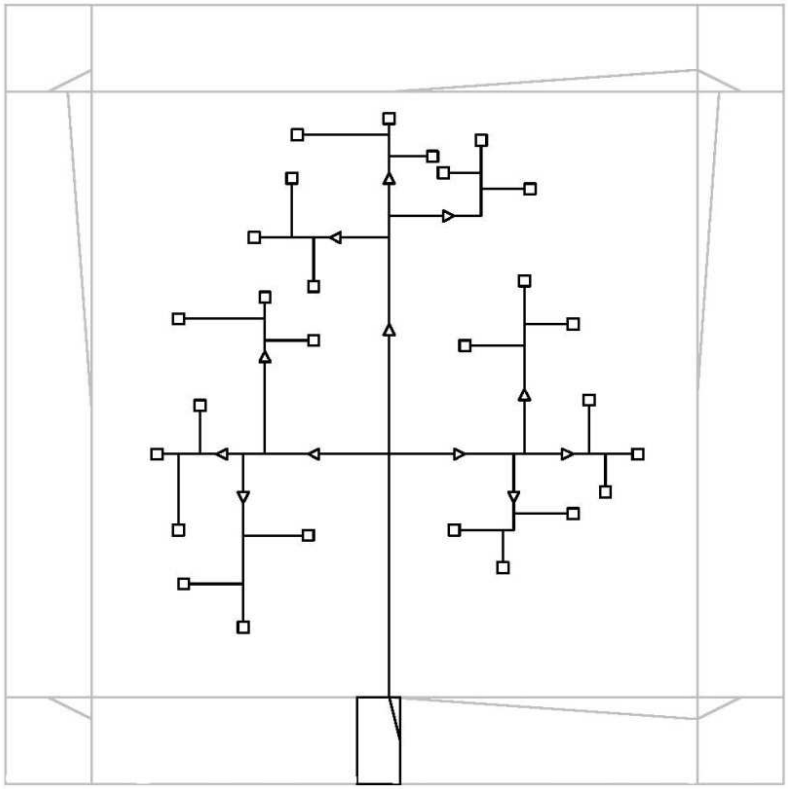
Metcalfe's Law
to the rescue of
Moore's Law!

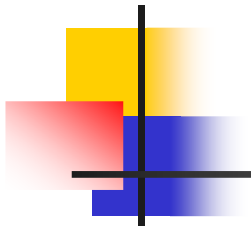
Bio-inspired

- Networks are intrinsically robust → exploit it!
- Massive ensemble of cheap, unreliable components
- Network Properties:
  - Local information exchange → global resiliency
  - Randomized topology & functionality → fits nano properties
  - Distributed nature → lacks an "Achilles heel"

What about Clock Distribution ?

- Clock easily the most energy-consuming signal of a chip
  - Largest length
  - Largest fan-out
  - Most activity ($\alpha = 1$)
- Skew control adding major overhead
  - Intermediate clock repeaters
  - De-skewing elements
- Opportunities
  - Reduced swing
  - Alternative clock distribution schemes
  - Avoiding a global clock altogether





Arguments for Sleep Mode Management

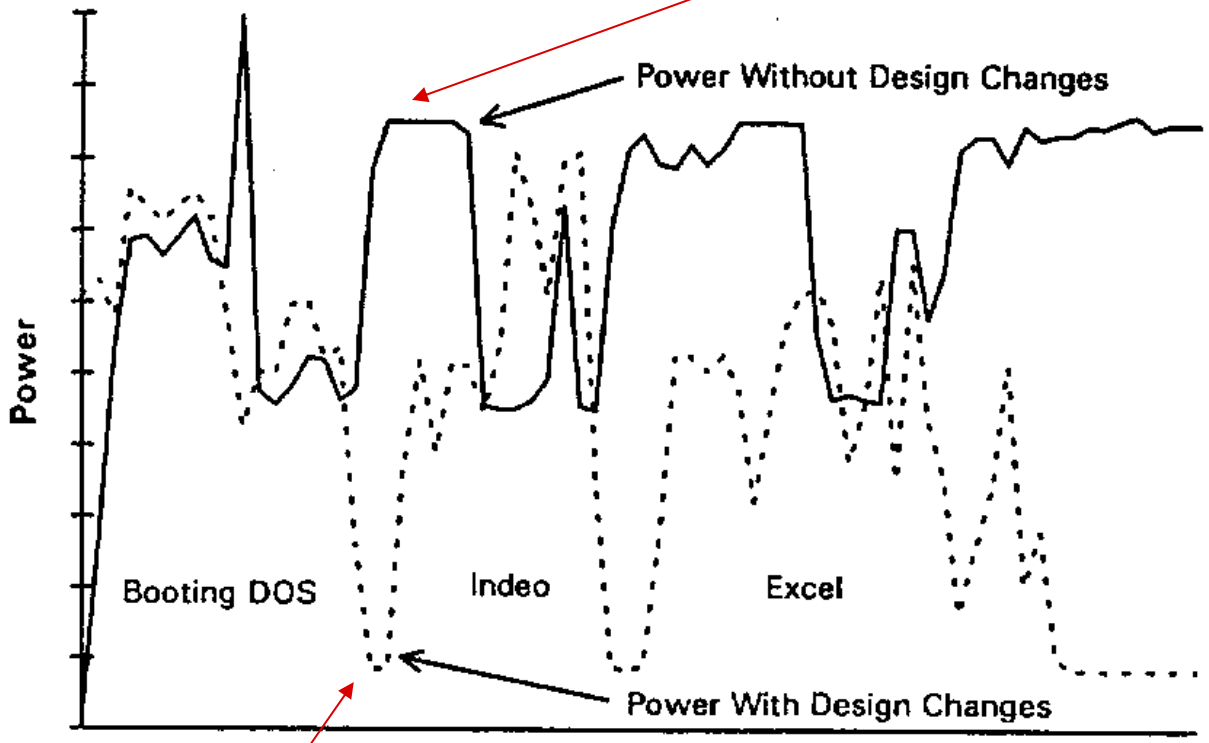
- Many computational applications operate in burst modes, interchanging active and non-active modes
  - General purposes computers, cell phones, interfaces, embedded processors, consumer applications, ...

- Prime concept: Power dissipation in standby should absolutely minimum, if not zero
- Sleep mode management has gained importance with increasing leakage

Standby Power - Was Not A Concern In Earlier Days

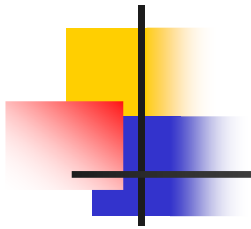
Pentium-1: 15 Watt (5V - 66MHz)
Pentium-2: 8 Watt (3.3V- 133 MHz)

Processor in idle mode!



Floating Point Unit and Cache powered down when not in use

[Source: Intel]

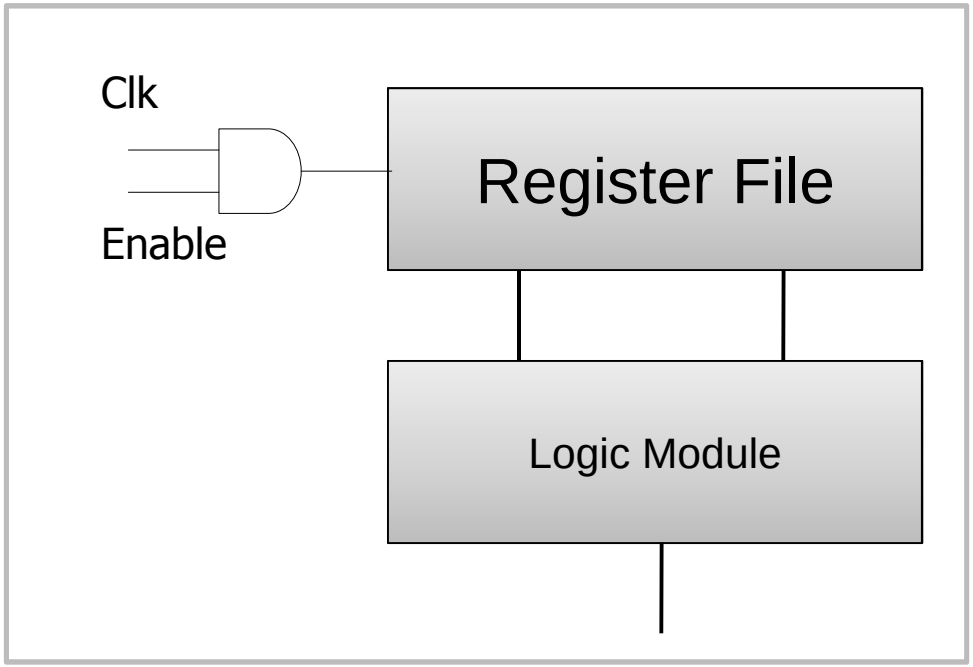


Dynamic Power - Clock Gating

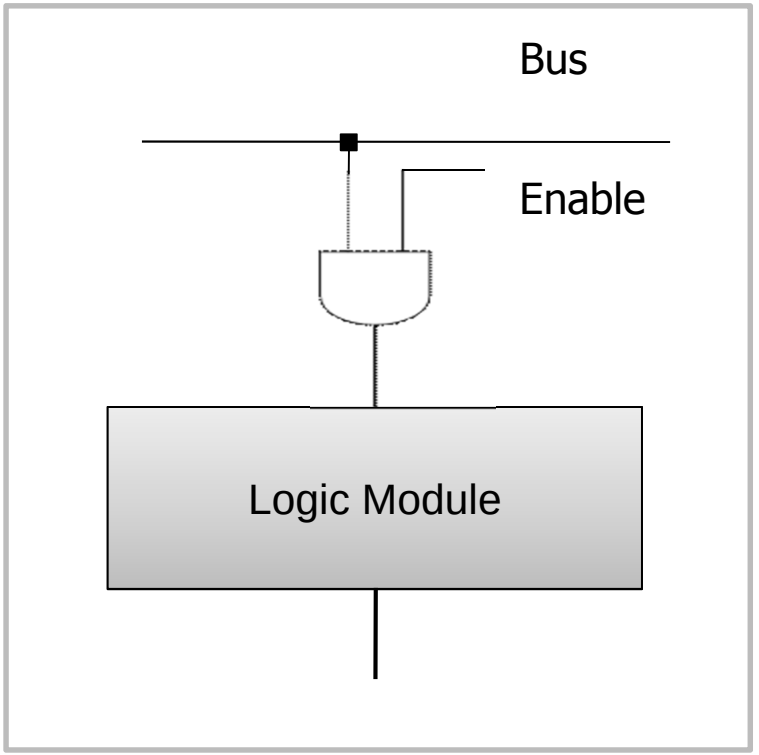
- Turn off clocks to idle modules
  - Ensure that spurious activity is set to zero
- Must ensure that data inputs to module are in stable mode
- Can be done at different levels of system hierarchy

Clock Gating

Turning off the clock for non-active components

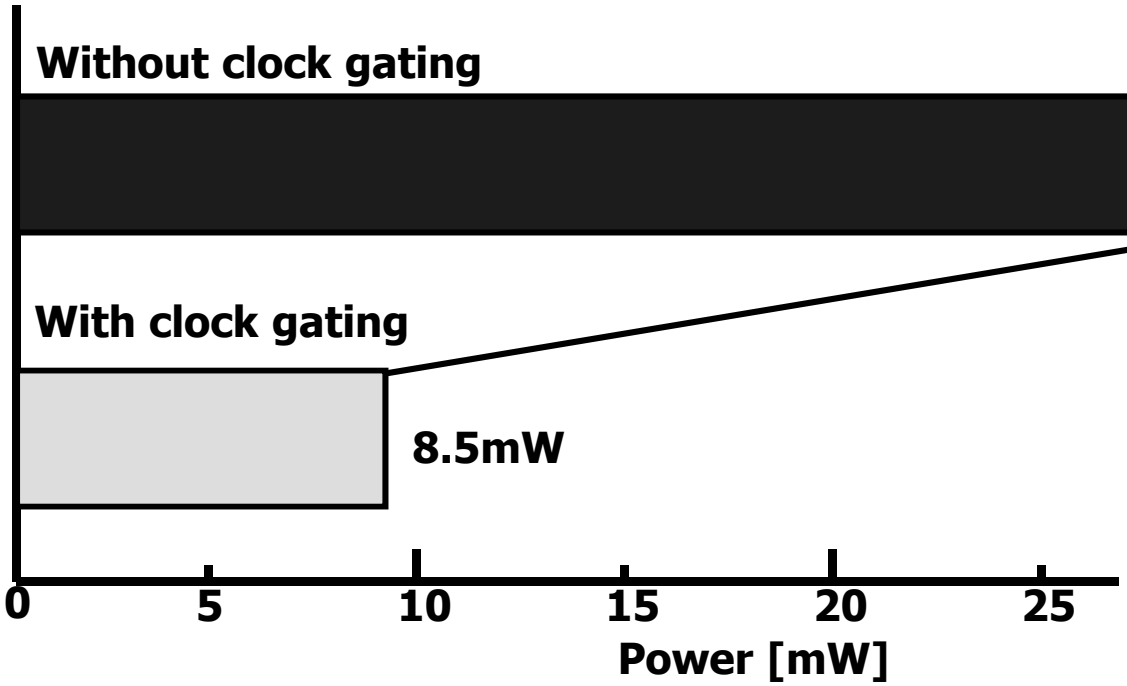


Disconnecting the inputs

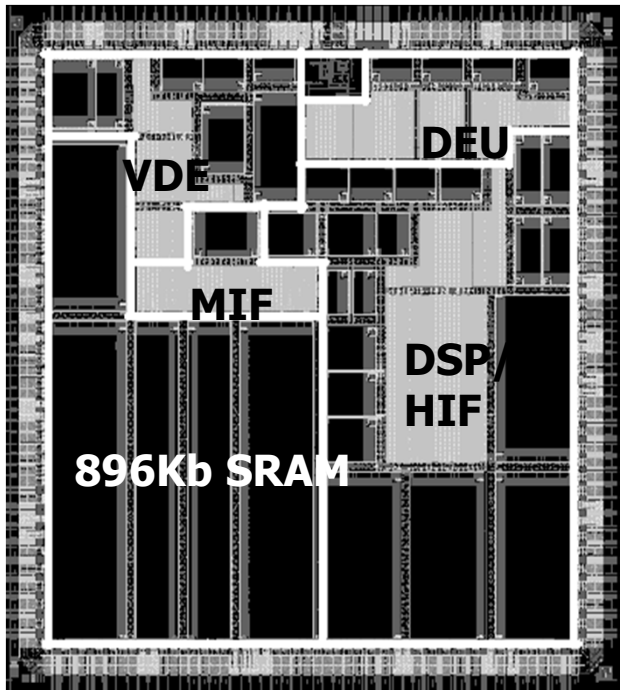


Gated clock signal suffers from additional gate delay!

Clock-gating Efficiently Reduces Power



MPEG4 decoder



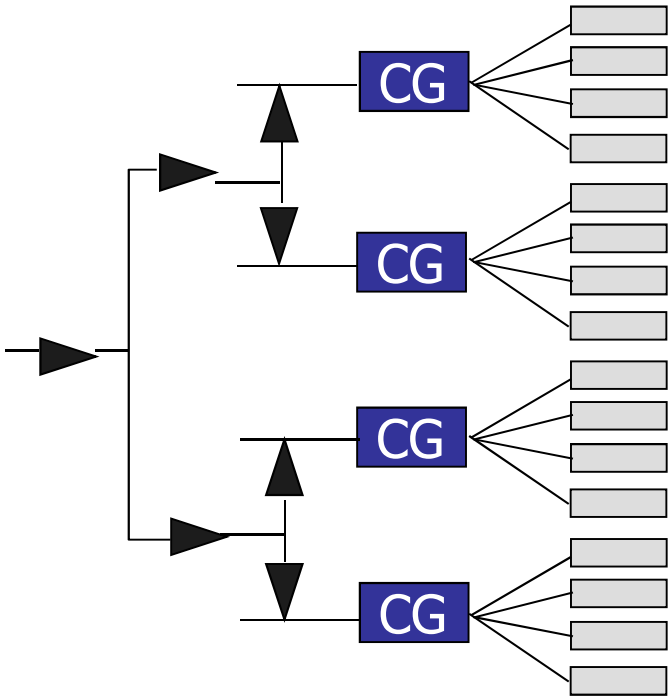
90% of F/F's clock-gated.

70% power reduction by clock-gating alone.

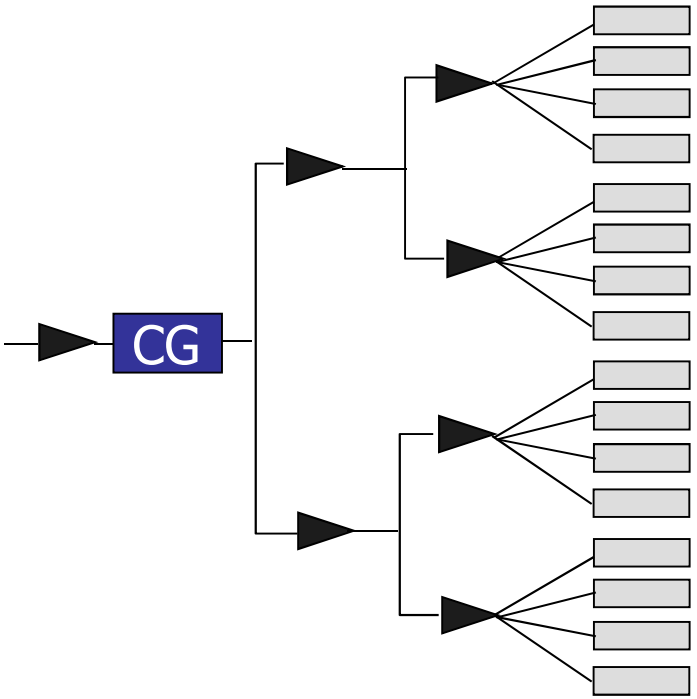
Clock-gating reduced significance in the leakage dominant generation!

Clock Gating

- Challenges to skew management and clock distribution (load on clock network varies dynamically)
- State-of-the-art design tools are starting to do a better job
  - For example, physically aware clock-gating inserts gaters in clock-tree based on timing constraints and physical layout



Power savings



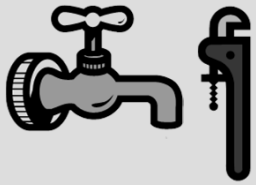
Simpler skew management, less area

Trade-Off between Sleep-Modes and Sleep-Time

Typical operation modes

Active mode
normal processing

Standby mode
fast resume
high passive power



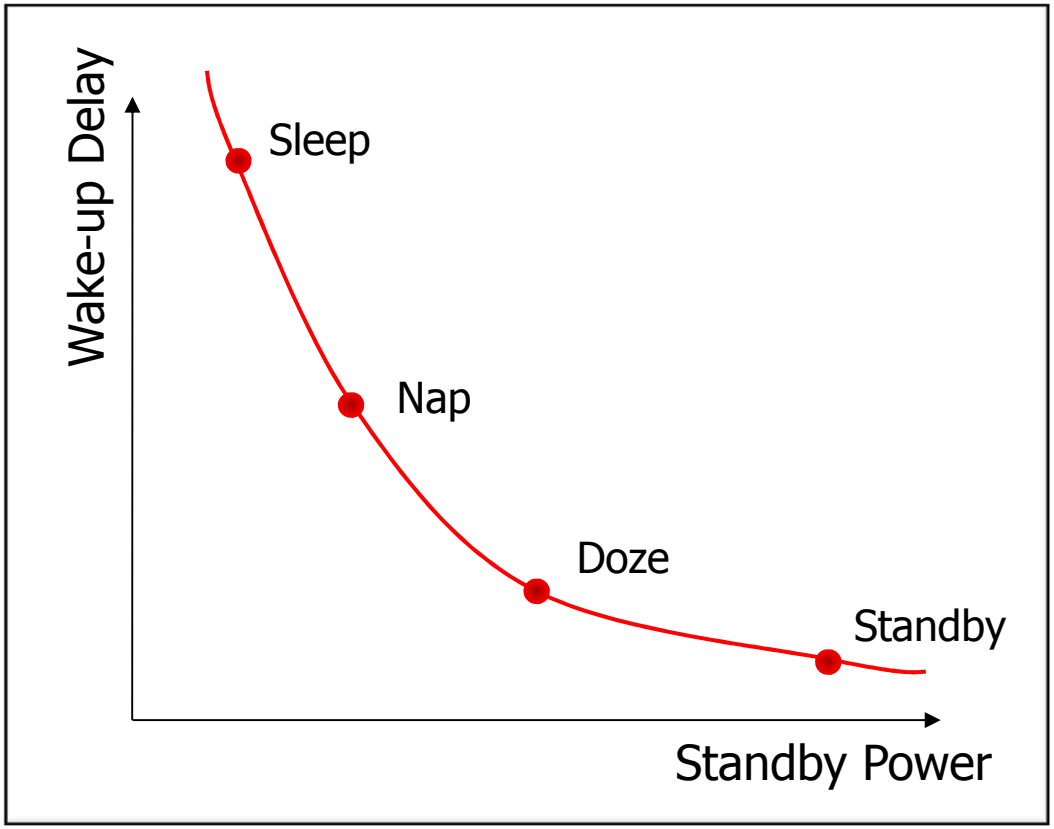
Sleep mode
slower resume
low passive power

Resume time from clock gating determined by the time it takes to turn on the clock distribution network

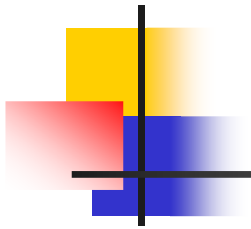
Standby Options:

- Just gate the clock to the module in question
- Turn off phased-locked loop(s)
- Turn off clock completely

The Standby Design Exploration Space

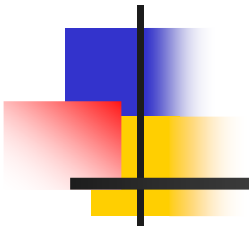


Trade-off between different operational modes
Should blend smoothly with run-time optimizations

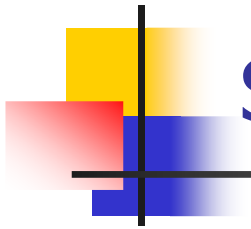


Overview

- Design Constraints
  - Power, Area, Frequency, CMOS Scaling
- Timing
  - Timing Metrics, Paths, Variability and Delay
- Deterministic Timing Analysis (Static Timing Analysis)
  - Models, Interconnect, Networks, Clock Distribution
- Statistical Timing Analysis
  - Probability, Spatial Correlations, MAX function
- Design Flow
  - Synthesis, Transformation, Definitions, Constrains



Statistical Timing Analysis



Static Timing Analysis

Pro's of non-statistical STA

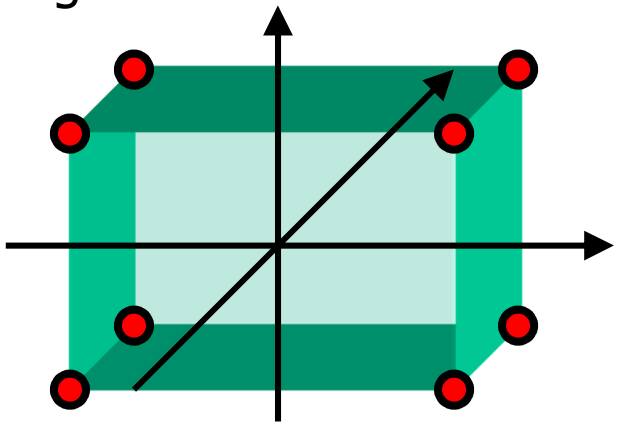
- Run-time linear in circuit size
- Conservative result
- Typically uses some fairly simple libraries (e.g. delay and slew)
- Easy to extend for use in optimization

Con's of non-statistical STA

- Cannot easily handle within-die correlation, especially if spatial correlation is included
- Needs many corners to handle all possible cases
- With significant random variations, to be conservative at all times, it is too pessimistic to result in competitive products
- Slower than linear time

Traditional Corner-Based Analysis

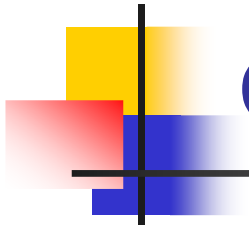
- Given a set of parameters $p_1, p_2, \ldots, p_k$
  - Each parameter varies between $[p_{i,min}, p_{i,max}]$
  - The variational region forms a multidimensional box



  - Corner-based analysis performs simulations at each corner
- Typically parameters correspond to process parameters, temperature and voltage

Corner Checks

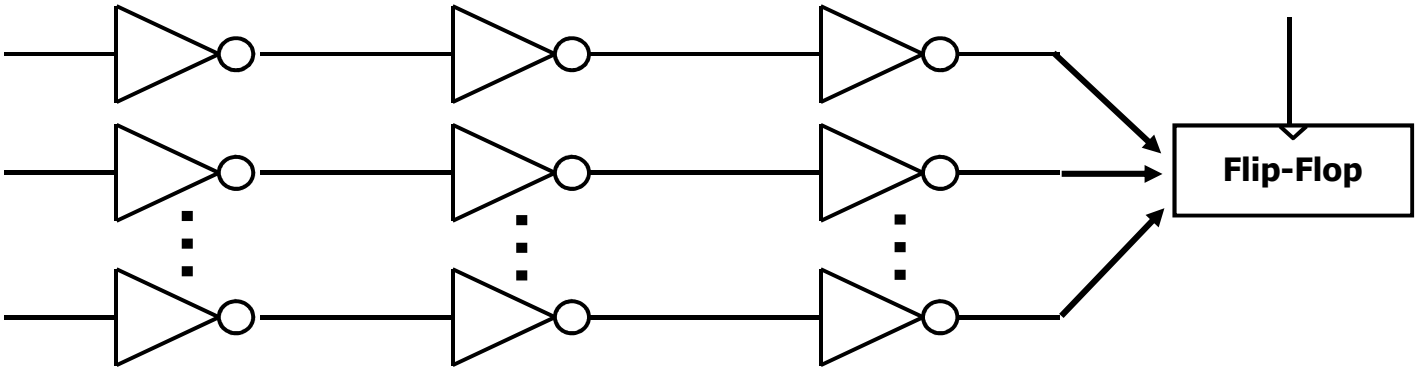
Corner					Purpose
nMOS	pMOS	Wire	V_{DD}	Temp	
T	T	T	S	S	Timing specifications (binned parts)
S	S	S	S	S	Timing specifications (conservative)
F	F	F	F	F	Race conditions, hold time constraints, pulse collapse, noise
S	S	?	F	S	Dynamic power
F	F	F	F	S	Subthreshold leakage noise and power, overall noise analysis
S	S	F	S	S	Races of gates against wires
F	F	S	F	F	Races of wires against gates
S	F	T	F	F	Pseudo-nMOS and ratioed circuits noise margins, memory read/write, race of pMOS against nMOS
F	S	T	F	F	Ratioed circuits, memory read/write, race of nMOS against pMOS



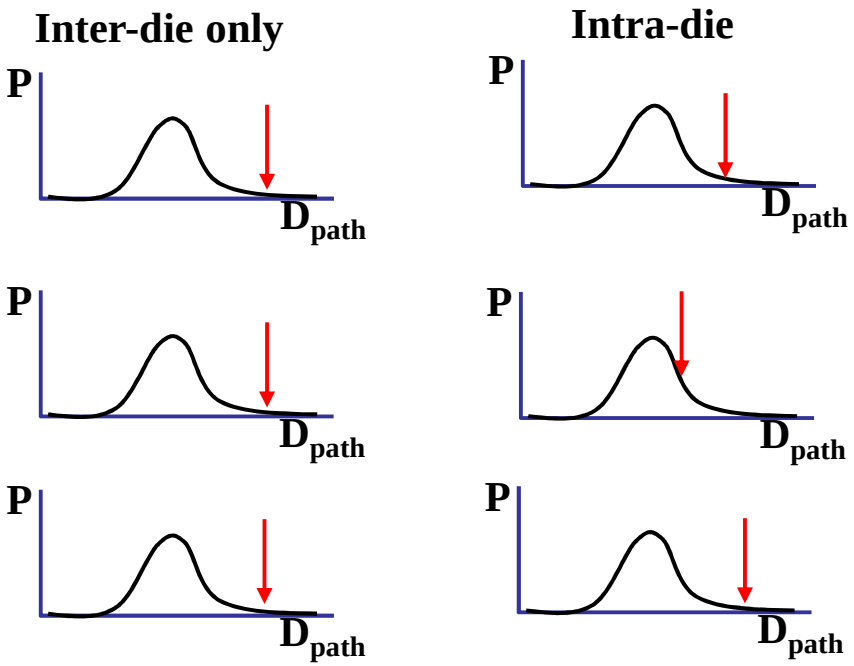
Corner-Based Analysis: Problems

- The number of corners we need to examine grows exponentially with the number of parameters
- It is only conservative if there is a monotone relationship between the parameter and the delay; otherwise the worst behavior may be somewhere in between!
- For very advanced technologies, monotonicity is no longer true for all parameters!

Intra-Die vs. Inter-Die Variation

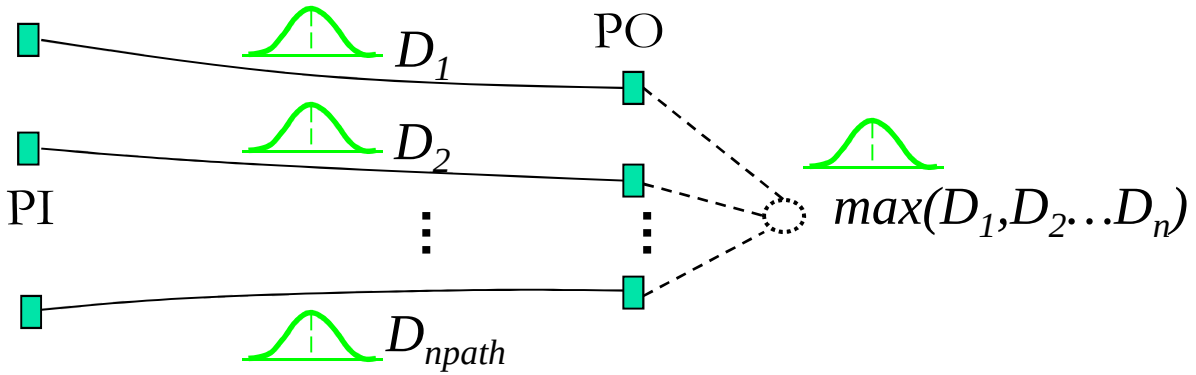


- Circuit delay is maximum of individual path delays
- Ignoring intra-die variation under-estimates circuit delay



Statistical Static Timing Analysis

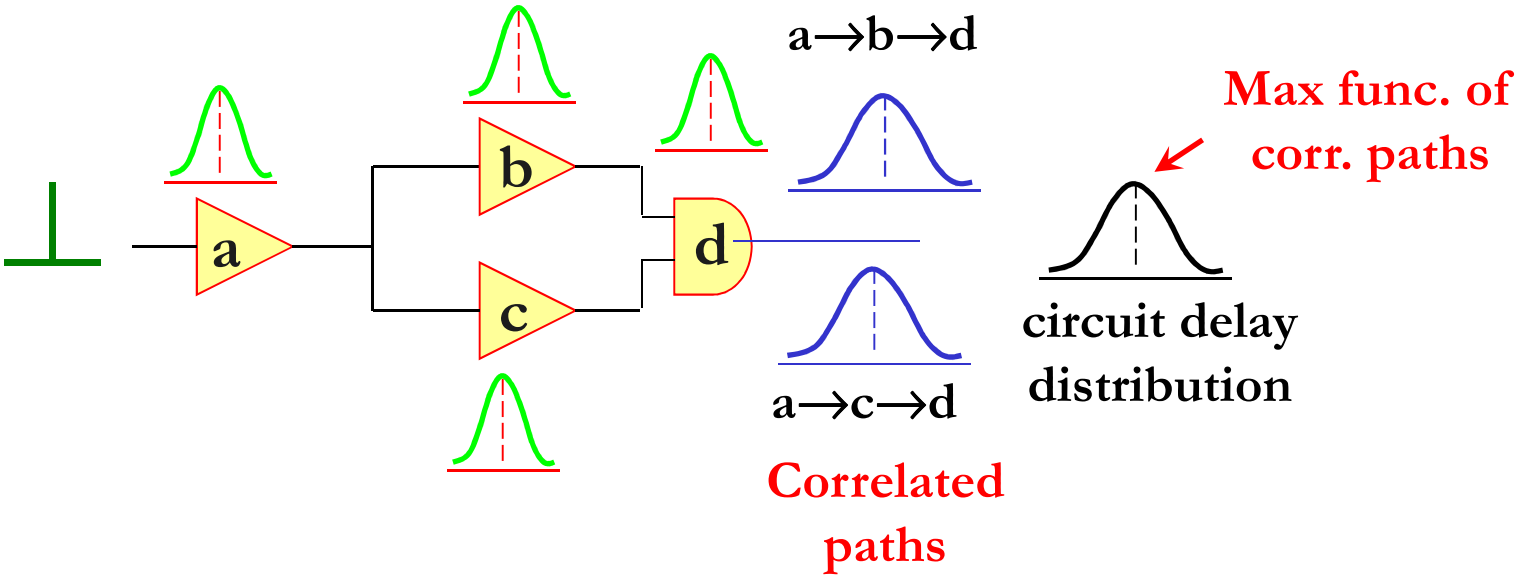
- Path-based analysis: find variability along a single path (sums gate and wire delays on specific paths)
  - Path selection important!



- Block-based analysis: calculate arrival times for each node (forward and backward from the clocked elements):
  - Statistical max (or min) operation that also considers correlation!

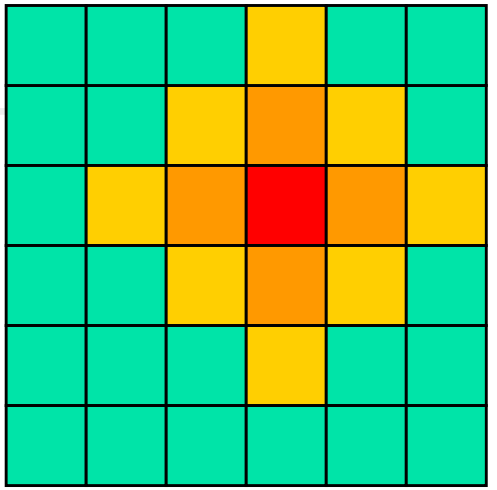
Difficulties in Statistical Timing Analysis

- Path correlation due to reconvergent fan-outs

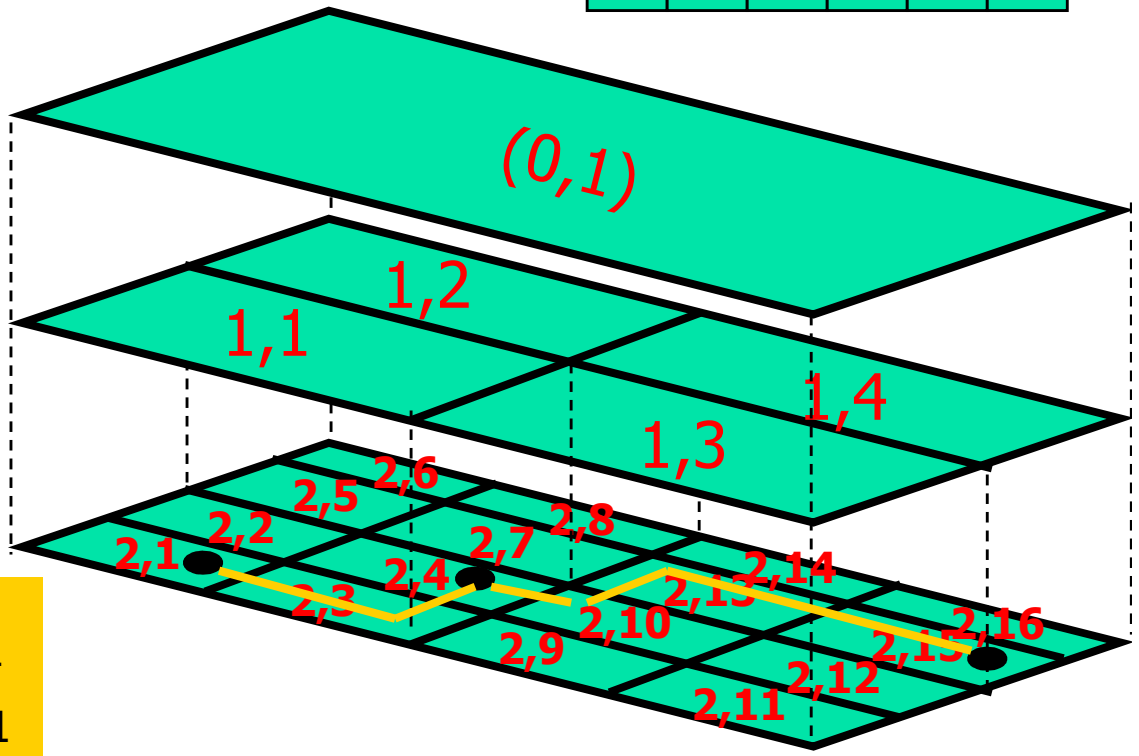


- Spatial correlations between nearby gates

Modeling Spatial Correlations



- Chip area divided into rectangles
  - Nearby squares are correlated
- Alternative hierarchical model

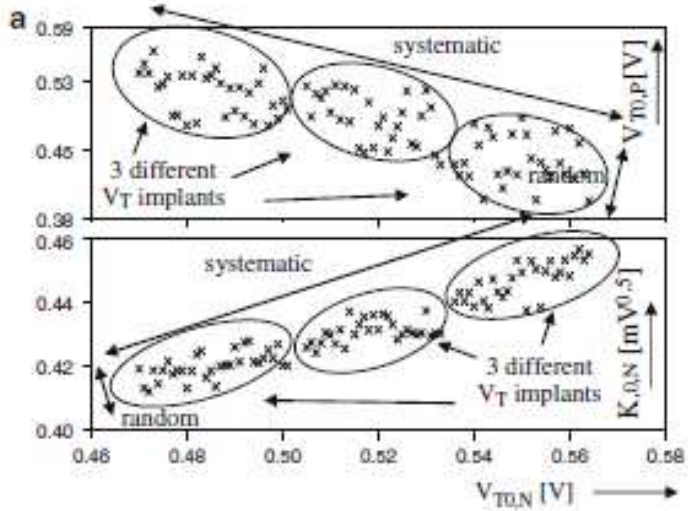


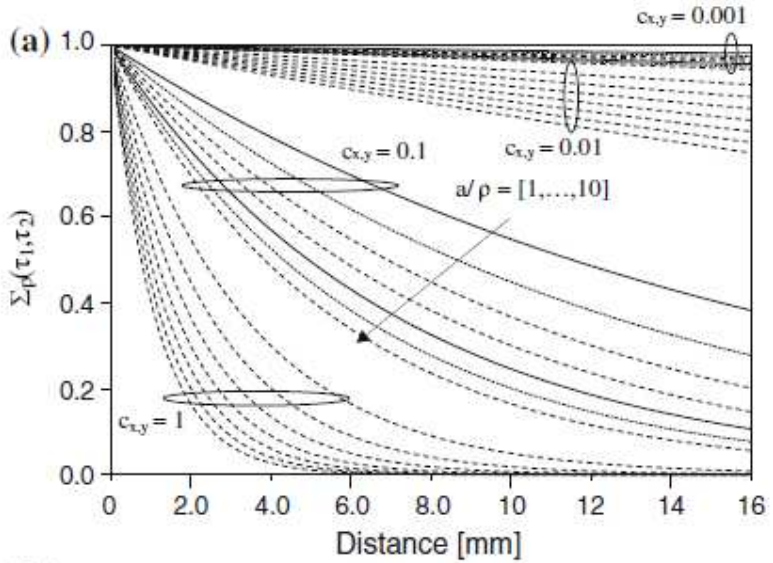
$$\Delta L_{g1} = \Delta L_{2,1} + \Delta L_{1,1} + \Delta L_{0,1}$$
$$\Delta L_{g2} = \Delta L_{2,4} + \Delta L_{1,1} + \Delta L_{0,1}$$
$$\Delta L_{g3} = \Delta L_{2,15} + \Delta L_{1,4} + \Delta L_{0,1}$$

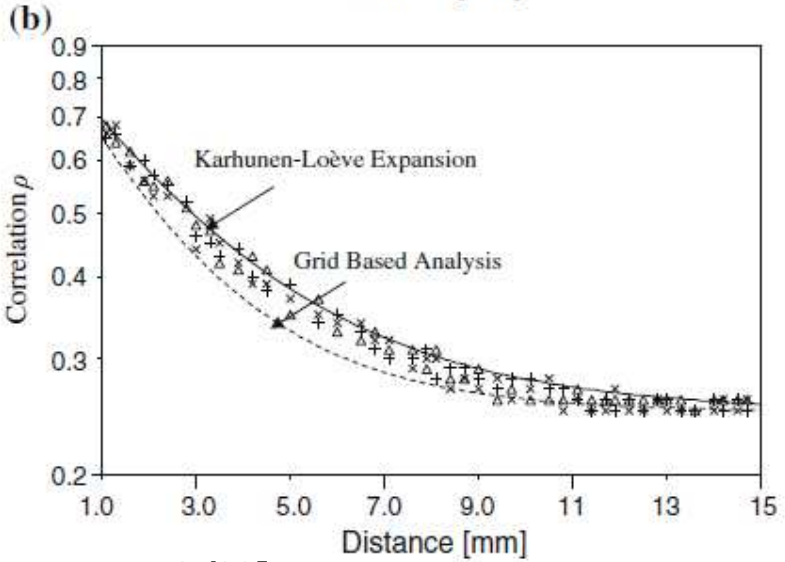
Modeling Spatial Correlations

Table 4.1 MOST key parameters in 0.18 CMOS technology at $V_{BS} = 0$V (a) $I_{DS,lin}$ at $V_{GS} = 1.8$V and $V_{DS} = 0.1$V c. $I_{DS,lin}$ at $V_{GS} = -1.8$V and $V_{DS} = -0.1$V (b) $I_{DS,sat}$ at $V_{GS} = 1.8$V and $V_{DS} = 1.8$V d. $I_{DS,sat}$ at $V_{GS} = -1.8$V and $V_{DS} = -1.8$V

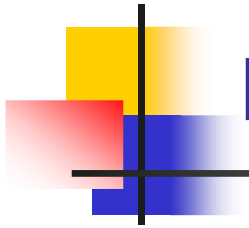
	$W/L = 10/0.18$			$W/L = 10/0.18$		
p	μ	σ	p	μ	σ	Unit
$V_{T0,N}$	516.92	10.44	$V_{T0,P}$	481.148	10.103	mV
$K_{0,N}$	422.53	10.34	$K_{0,P}$	518.538	13.109	$mV^{1/2}$
K_N	446.967	8.461	K_P	451.971	17.434	$mV^{1/2}$
β_N	26.334	1.290	β_P	6.775	0.261	mA/V^2
$W_{eff,N}$	10.034	0.010	$W_{eff,P}$	10.034	0.010	µm
$L_{eff,N}$	0.108	0.005	$L_{eff,P}$	0.143	0.005	µm
$I_{DS,lin}$ [a]	1.354	0.018	$I_{DS,lin}$ [c]	0.402	0.018	mA
$I_{DS,sat}$ [b]	6.035	0.226	$I_{DS,sat}$ [d]	2.914	0.226	mA







[Ref: A. Zjajo, TVLSI'09]



Problem Statement

- Find the PDF/CDF for circuit delay distribution:

$$D_{max} = \max(D_1, D_2, \ldots, D_{npaths})$$

where D_i: delay distribution of i^{th} path in the circuit

- Assume normal distributions on process parameter values
  - Why?
  - Is this reasonable? If not, what is?
- Parameter correlations
  - L_{eff} shows high spatial correlations
  - T_{ox}, N_d are largely uncorrelated

Some Basics of Probability

- Mean, variance

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx,$$

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

  - Mean - expected value of variable X.
  - Variance - expected value of squared deviation from the mean of X.

- Covariance

$$\operatorname{cov}(X,Y) = \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right],$$

  - Covariance - a measure of how much two random variable change together.

- Correlation coefficient

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

  - Number that quantifies correlation.

- Independence

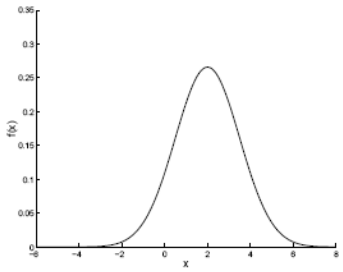
$$f(x,y) = f(x)f(y) \Rightarrow E(XY) = E(X)E(Y)$$

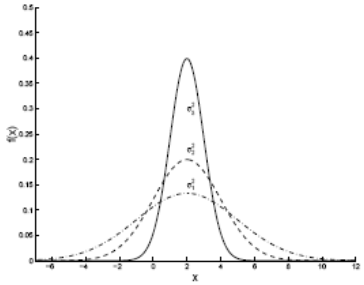
  - Independent if the realization of one does not affect the probability distribution of the other.

Commonly-Encountered Distributions

- Gaussian or normal distribution $N(\mu,\sigma^2)$
  - In one variable
    - PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty.$$





- PDF - function that describes the local density of probability

Figure 1.1: Gaussian or Normal pdf, $N(2, 1.5^2)$ Gaussian pdf with different variances $(\sigma_1^2 = 3^2, \sigma_2^2 = 2^2, \sigma_3^2 = 1)$

- CDF

$$Pr\{X \le x_a\} = \begin{cases} 0.5 - erf(\frac{\mu - x_a}{\sigma}) & for \quad x_a \le \mu \\ 0.5 + erf(\frac{x_a - \mu}{\sigma}) & for \quad x_a \ge \mu \end{cases}$$

$$erf(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp^{-y^2/2} dy$$

CDF of *X* evaluated at *x* - probability that *X* will take a value ≤ *x*.

- For a Gaussian, independence is identical to uncorrelatedness

$$f(xy) = f(x)f(y) \quad \Leftrightarrow \quad E(XY) = E(X)E(Y) \Leftrightarrow \rho = 0.$$

[http://users.isr.ist.utl.pt/~mir/pub/probability.pdf]

Commonly-Encountered Distributions (contd.)

- Multivariate Gaussian N(μ,Σ)
  - PDF $$f_X(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - m_X)^T \Sigma^{-1}(x - m_X)\right\}$$

$$m_X = E\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} m_{X_1} \\ m_{X_2} \\ \vdots \\ m_{X_n} \end{bmatrix}.$$

  - A random vector is said to be *k*-variate normally distributed if every linear combination of its *k* components has a univariate normal distribution.

$$\Sigma_X = \Sigma_X^T = \begin{bmatrix} E(X_1 - m_{X_1})^2 & E(X_1 - m_{X_1})(X_2 - m_{X_2}) & \dots & E(X_1 - m_{X_1})(X_n - m_{X_n}) \\ E(X_2 - m_{X_2})(X_1 - m_{X_1}) & E(X_2 - m_{X_2})^2 & \dots & E(X_2 - m_{X_2})(X_n - m_{X_n}) \\ \vdots & & & \vdots \\ E(X_n - m_{X_n})(X_1 - m_{X_1}) & \dots & \dots & E(X_n - m_{X_n})^2 \end{bmatrix}.$$

Commonly-Encountered Distributions (contd.)

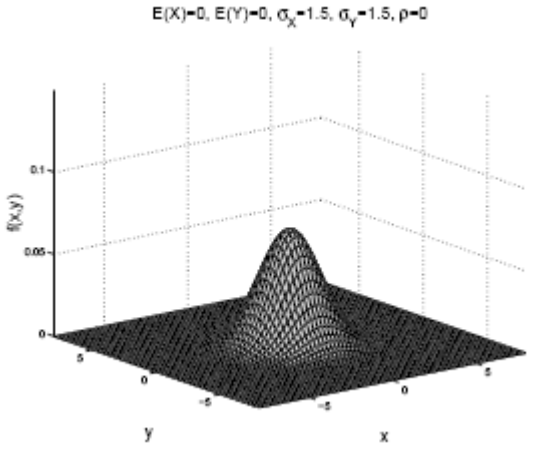
- Bivariate Gaussian
  - PDF

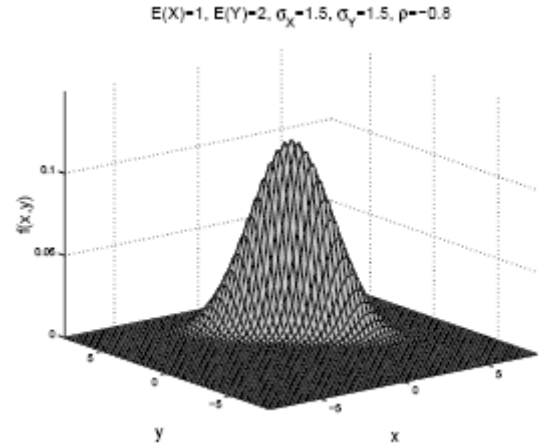
$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-m_X)^2}{\sigma_X^2}-\frac{2\rho(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y}+\frac{(y-m_Y)^2}{\sigma_Y^2}\right)\right]$$

  - The sum of two Gaussians is a Gaussian
  - Z= X+Y, where X, Y are uncorrelated Gaussians

$$\sigma_Z^2=\sigma_X^2+\sigma_Y^2$$

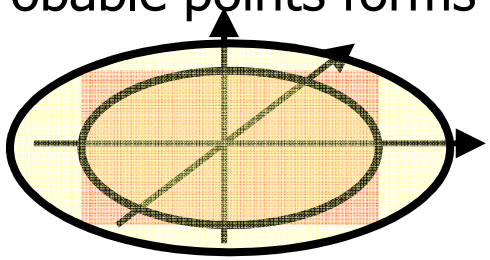
$$\sigma_{X+Y}=\sqrt{\sigma_X^2+\sigma_Y^2+2\rho\sigma_X\sigma_Y},$$



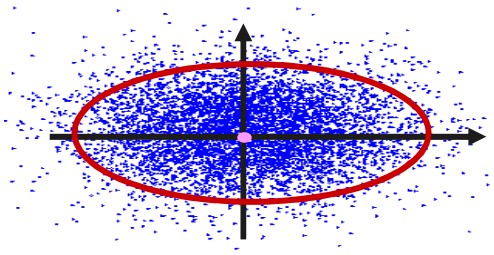


The Ellipsoid

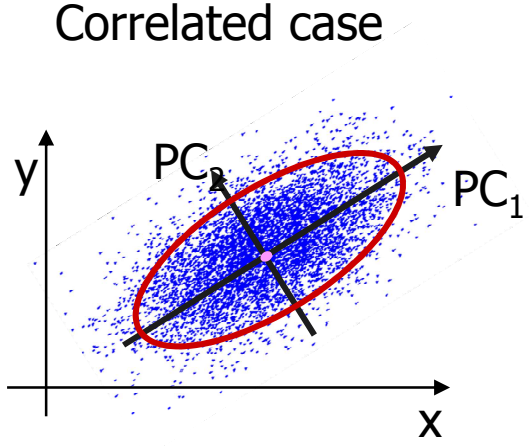
- Locus of equiprobable points for most distributions is not a box
  - Gaussian: locus of equiprobable points forms an ellipsoid



  - Ellipse centered at (m_X, m_Y), axes along the eigenvectors of the covariance matrix
  - Uncorrelated case



Correlated case



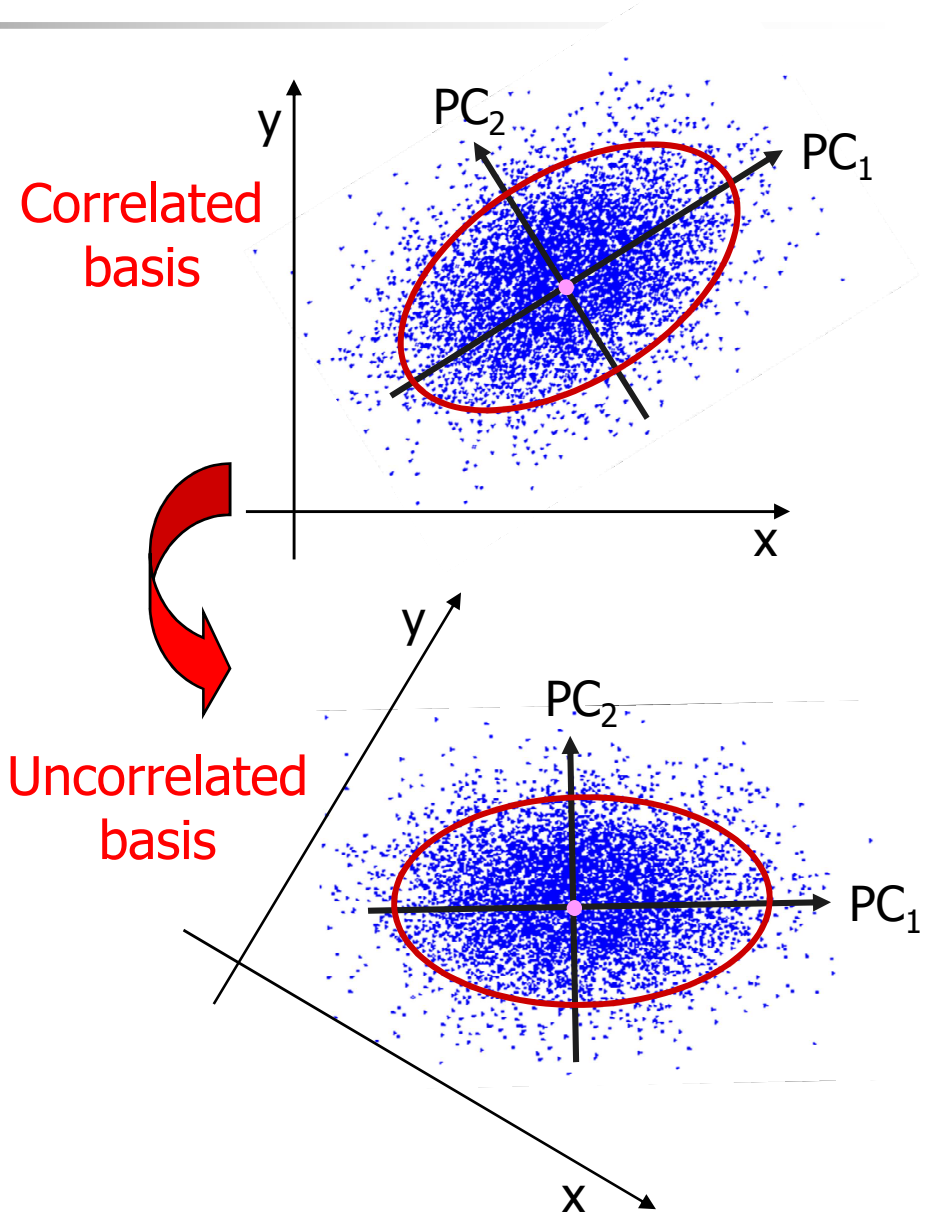
Idea of Orthogonal Transformations

Orthogonal transformation (OT) - a linear transformation on an inner product space (preserve lengths of vectors and angles between them)

PCA – converting correlated into linearly uncorrelated variables using OT

- 1st principal component (PC) has the largest σ; each succeeding PC has the highest σ possible under the constraint that it is orthogonal to the preceding components.

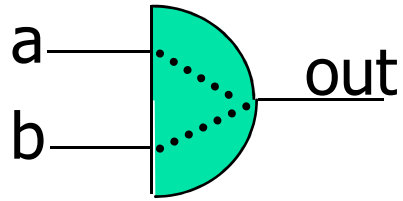
- the resulting vectors are an uncorrelated orthogonal set.



Berkelaar's Method

[Ber97a,Ber97b,JB00]

- Types of operations in STA
  - SUM: $T_{a\rightarrow out} = T_a + d_{a\rightarrow out}$; $T_{b\rightarrow out} = T_b + d_{b\rightarrow out}$
  - MAX: $T_{out} = \max(T_{a\rightarrow out}, T_{b\rightarrow out})$



- Gate delay modeled as a Gaussian
  - SUM is easy: sum of Gaussians = Gaussian
    - S = A+B: $\mu_S = \mu_A + \mu_B$, $\sigma_s^2 = \sigma_A^2 + \sigma_B^2$
  - MAX of Gaussians is not a Gaussian
  - Approach: approximate max by a Gaussian
    - Analytic expressions for mean, variance in [JB00]

Note: Calculating "MIN" for early mode analysis is analogous to calculating "MAX" since MIN(f) = MAX(-f)

Approach: Approximate MAX by a Gaussian

$$\mu_C = Ex_C = \frac{\sqrt{\sigma_A^2 + \sigma_B^2}}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)^2} + \mu_A \phi\left(\frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right) + \mu_B \phi\left(\frac{\mu_B - \mu_A}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)$$

$$\phi(x) = \int_{-\infty}^{x} e^{-\frac{1}{2}u^2} du$$

$$Ex_C^2 = (\mu_A + \mu_B)\frac{\sqrt{\sigma_A^2 + \sigma_B^2}}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)^2} +$$

$$(\sigma_A^2 + \mu_A^2)\phi\left(\frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right) +$$

$$(\sigma_B^2 + \mu_B^2)\phi\left(\frac{\mu_B - \mu_A}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right)$$

$$\sigma_C^2 = Ex_C^2 - \mu_C^2$$

Why I didn't write the precise expression in the last slide… (proof)

Appendix A

We will now derive the mean and standard deviation of a stochastic variable C which is the maximum of two normal distributed statistically independent stochastic variables A and B. In order to derive this mean and standard deviation we will change the bases of the double integration:

$$\int_{-\infty}^{\infty} x f_A(x) \int_{-\infty}^{y} f_B(y)\,dy\,dx \tag{19}$$

which part of the calculation of $\mu_C = Ex_C$ as follows:

$$\frac{x-\mu_A}{\sigma_A} = \frac{u\sigma_B}{\sqrt{\sigma_A^2+\sigma_B^2}} - \frac{v\sigma_A}{\sqrt{\sigma_A^2+\sigma_B^2}} \tag{20}$$

which gives:

$$x = \frac{u\sigma_A\sigma_B}{\sqrt{\sigma_A^2+\sigma_B^2}} - \frac{v\sigma_A^2}{\sqrt{\sigma_A^2+\sigma_B^2}} + \mu_A \tag{21}$$

and:

$$\frac{y-\mu_B}{\sigma_B} = \frac{u\sigma_A}{\sqrt{\sigma_A^2+\sigma_B^2}} + \frac{v\sigma_B}{\sqrt{\sigma_A^2+\sigma_B^2}} \tag{22}$$

which gives:

$$y = \frac{u\sigma_A\sigma_B}{\sqrt{\sigma_A^2+\sigma_B^2}} + \frac{v\sigma_B^2}{\sqrt{\sigma_A^2+\sigma_B^2}} + \mu_B \tag{23}$$

For this change of base we calculate:

$$\left|\frac{\delta(x,y)}{\delta(v,u)}\right| = \begin{vmatrix} \frac{\sigma_A\sigma_B}{\sqrt{\sigma_A^2+\sigma_B^2}} & -\frac{\sigma_A^2}{\sqrt{\sigma_A^2+\sigma_B^2}} \\ \frac{\sigma_A\sigma_B}{\sqrt{\sigma_A^2+\sigma_B^2}} & \frac{\sigma_B^2}{\sqrt{\sigma_A^2+\sigma_B^2}} \end{vmatrix} = \frac{\sigma_A\sigma_B^3+\sigma_A^3\sigma_B}{\sigma_A^2+\sigma_B^2} = \sigma_A\sigma_B \tag{24}$$

The mean of the stochastic variable C then becomes:

$$\mu_C = \int_{-\infty}^{\infty} x f_C(x)\,dx \tag{25}$$

$$= \int_{-\infty}^{\infty} x f_A(x) F_B(x)\,dx + \int_{-\infty}^{\infty} x F_A(x) f_B(x)\,dx$$

$$= \frac{1}{\sigma_A\sigma_B\sqrt{2\pi}\sqrt{2\pi}} \int_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} \int_{-\infty}^{\infty} \left(\frac{u\sigma_A\sigma_B}{\sqrt{\sigma_A^2+\sigma_B^2}} - \frac{v\sigma_A^2}{\sqrt{\sigma_A^2+\sigma_B^2}} + \mu_A \right) e^{-\frac{1}{2}u^2+v^2)} \sigma_A\sigma_B\,dv\,du + \ldots$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} \int_{-\infty}^{\infty} \left(\frac{u\sigma_A\sigma_B}{\sqrt{\sigma_A^2+\sigma_B^2}} - \frac{v\sigma_A^2}{\sqrt{\sigma_A^2+\sigma_B^2}} + \mu_A \right) e^{-\frac{1}{2}u^2+v^2)}\,dv\,du + \ldots$$

$$= \left[\frac{1}{\sqrt{2\pi}} \frac{\sigma_A^2}{\sqrt{\sigma_A^2+\sigma_B^2}} e^{-\frac{1}{2}v^2} \right]_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} + \frac{\mu_A}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} e^{-\frac{1}{2}v^2}\,dv + \frac{\mu_B}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\mu_B-\mu_A}{\sqrt{\sigma_A^2+\sigma_B^2}}} e^{-\frac{1}{2}v^2}\,dv +$$

$$\left[\frac{1}{\sqrt{2\pi}} \frac{\sigma_B^2}{\sqrt{\sigma_A^2+\sigma_B^2}} e^{-\frac{1}{2}v^2} \right]_{-\infty}^{\frac{\mu_B-\mu_A}{\sqrt{\sigma_A^2+\sigma_B^2}}}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} \left(-\frac{v\sigma_A^2}{\sqrt{\sigma_A^2+\sigma_B^2}} + \mu_A \right) e^{-\frac{1}{2}v^2}\,dv + \ldots$$

$$= \frac{\sqrt{\sigma_A^2+\sigma_B^2}}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)^2} + \mu_A\phi\left(\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}\right) + \mu_B\phi\left(\frac{\mu_B-\mu_A}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)$$

in which $\phi(x)$ is given by:

$$\phi(x) = \int_{-\infty}^{x} e^{-\frac{1}{2}u^2}\,du \tag{26}$$

Note that in some lines of equation 25 we have only given one half of the equation explicitly. The other half is depicted by triple dots, and is similar to the first half of the equation. We will now calculate the standard deviation of stochastic variable C in two steps. The first step is the calculation of Ex_C^2:

$$Ex_C^2 = \int_{-\infty}^{\infty} x^2 f_C(x)\,dx \tag{27}$$

$$= \int_{-\infty}^{\infty} x^2 f_A(x) F_B(x)\,dx + \int_{-\infty}^{\infty} x^2 F_A(x) f_B(x)\,dx$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} \int_{-\infty}^{\infty} \left(\frac{u\sigma_A\sigma_B}{\sqrt{\sigma_A^2+\sigma_B^2}} - \frac{v\sigma_A^2}{\sqrt{\sigma_A^2+\sigma_B^2}} + \mu_A \right)^2 e^{-\frac{1}{2}u^2+v^2)}\,dv\,du + \ldots$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} \int_{-\infty}^{\infty} \left(\frac{u^2\sigma_A^2\sigma_B^2}{\sigma_A^2+\sigma_B^2} + \frac{v^2\sigma_A^4}{\sigma_A^2+\sigma_B^2} - \frac{2v\mu_A\sigma_A^2}{\sqrt{\sigma_A^2+\sigma_B^2}} + \mu_A^2 \right) e^{-\frac{1}{2}u^2+v^2)}\,dv\,du + \ldots$$

$$= \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{-u\sigma_A^2\sigma_B^2}{\sigma_A^2+\sigma_B^2} e^{-\frac{1}{2}u^2+v^2)}\,dv \right]_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} +$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} \left(\frac{v^2\sigma_A^4}{\sigma_A^2+\sigma_B^2} - \frac{2v\mu_A\sigma_A^2}{\sqrt{\sigma_A^2+\sigma_B^2}} + \mu_A^2 \right) e^{-\frac{1}{2}v^2}\,dv +$$

$$\frac{1}{2\pi} \int_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} \int_{-\infty}^{\infty} \frac{\sigma_A^2\sigma_B^2}{\sigma_A^2+\sigma_B^2} e^{-\frac{1}{2}u^2+v^2)}\,dv\,du + \ldots$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} \left(\frac{v^2\sigma_A^4}{\sigma_A^2+\sigma_B^2} - \frac{2v\mu_A\sigma_A^2}{\sqrt{\sigma_A^2+\sigma_B^2}} + \mu_A^2 + \frac{\sigma_A^2\sigma_B^2}{\sigma_A^2+\sigma_B^2} \right) e^{-\frac{1}{2}v^2}\,dv + \ldots$$

$$= \left[-\frac{1}{\sqrt{2\pi}} \frac{v\sigma_A^4}{\sigma_A^2+\sigma_B^2} e^{-\frac{1}{2}v^2} \right]_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} +$$

$$\left[\frac{2\mu_A\sigma_A^2}{\sqrt{\sigma_A^2+\sigma_B^2}} e^{-\frac{1}{2}v^2} \right]_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} +$$

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}} \left(\frac{\sigma_A^4+\sigma_A^2\sigma_B^2}{\sigma_A^2+\sigma_B^2} + \mu_A^2 \right) e^{-\frac{1}{2}v^2}\,dv + \ldots$$

$$= (\sigma_A^2+\mu_A^2)\phi\left(\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}\right) +$$

$$\frac{e^{-\frac{1}{2}\left(\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)^2}}{\sqrt{2\pi}\sqrt{\sigma_A^2+\sigma_B^2}} \left(2\mu_A\sigma_A^2 - \frac{\sigma_A^4(\mu_A-\mu_B)}{\sigma_A^2+\sigma_B^2} \right) + \ldots$$

$$= \frac{e^{-\frac{1}{2}\left(\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)^2}}{\sqrt{2\pi}\sqrt{\sigma_A^2+\sigma_B^2}} \left(\frac{\mu_A\sigma_A^4+2\mu_A\sigma_A^2\sigma_B^2+\mu_B\sigma_A^2}{\sigma_A^2+\sigma_B^2} \right) +$$

$$(\sigma_A^2+\mu_A^2)\phi\left(\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}\right) + \ldots$$

$$= (\sigma_A^2+\mu_A^2)\phi\left(\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}\right) +$$

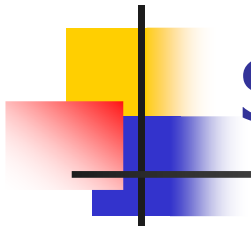
$$(\sigma_B^2+\mu_B^2)\phi\left(\frac{\mu_B-\mu_A}{\sqrt{\sigma_A^2+\sigma_B^2}}\right) +$$

$$(\mu_A+\mu_B)\frac{\sqrt{\sigma_A^2+\sigma_B^2}}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\mu_A-\mu_B}{\sqrt{\sigma_A^2+\sigma_B^2}}\right)^2}$$

Note that in equation 27 we also have given only one half of the equation explicitly, with the other half which is similar to the first, depicted by triple dots. We can now calculate the standard deviation of stochastic variable C with the following equation:

$$\sigma_C^2 = Ex_C^2 - \mu_C^2 \tag{28}$$

We have now expressed μ_C and σ_C as functions of just μ_A, μ_B, σ_A and σ_B.



Statistical Static Timing Analysis

Not considering statistics can result in > 30% delay errors

SSTA Con's

- Complex, especially with realistic (non-Gaussian) distributions
- Difficult to extend to an optimization flow
- Required data likely to be time-varying and hence unreliable
- If the fab change statistical properties of the process, design have to be re-evaluated

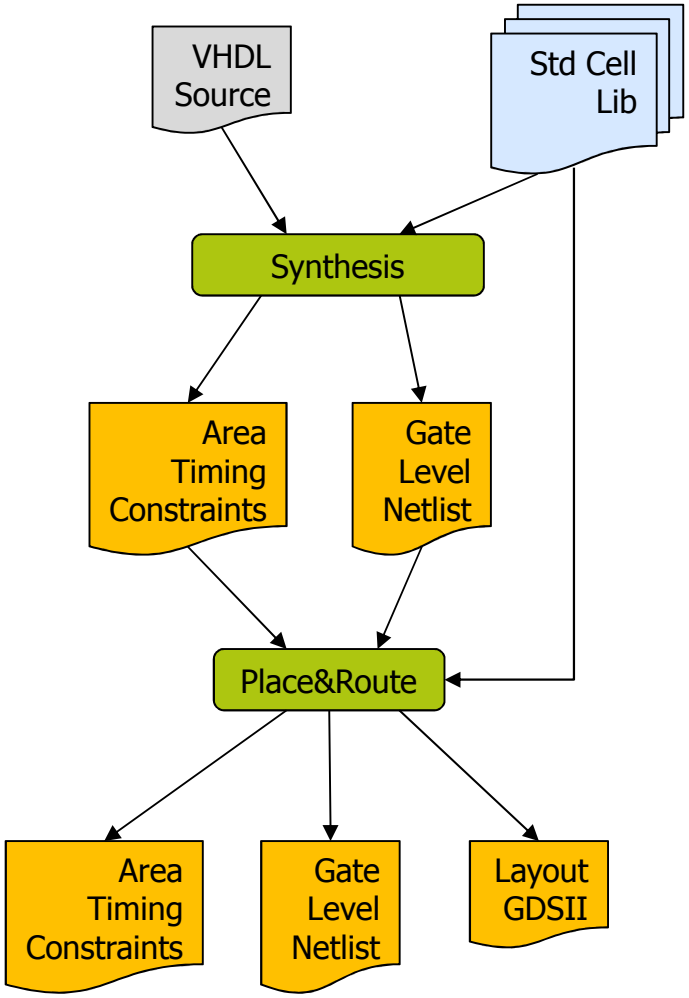
Active research field

- Efficient SSTA solvers (e.g., PWL-RDE solver, see Q.Tang (TCAD,'14))
- An enhanced deterministic STA that also takes into account sensitivities and correlation

Overview

- Design Constraints
  - Power, Area, Frequency, CMOS Scaling
- Timing
  - Timing Metrics, Paths, Variability and Delay
- Deterministic Timing Analysis (Static Timing Analysis)
  - Models, Interconnect, Networks, Clock Distribution
- Statistical Timing Analysis
  - Probability, Spatial Correlations, MAX function
- Design Flow
  - Synthesis, Transformation, Definitions, Constrains

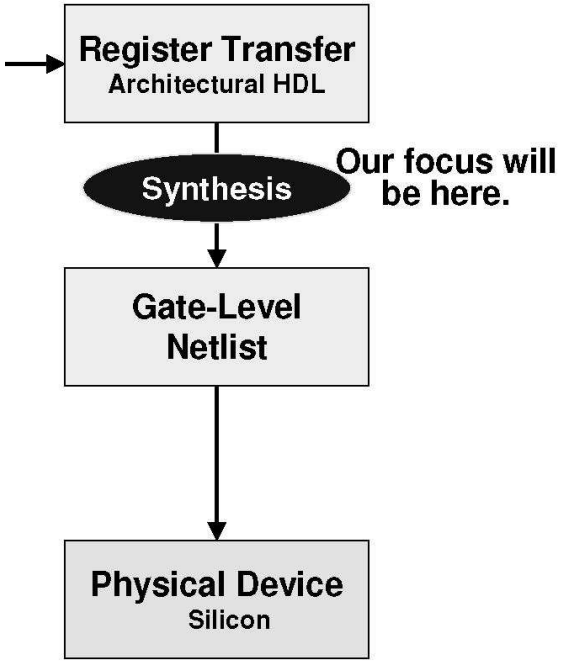
Design Flow



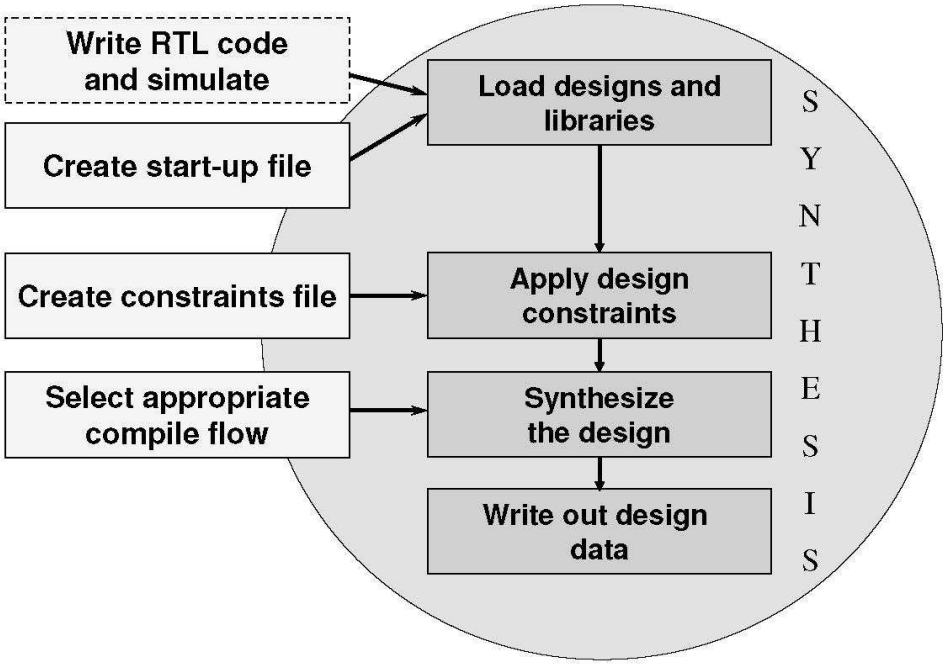
- Tools
  - Synthesis: Synopsys Design Compiler
  - Place & Route: Cadence SOC Encounter 8.1
- Process
  - UMC L90 SP
- Standard cell library
  - Faraday:
    - fsd0a_a_generic_core
    - fod0a_b25_t25_generic_io

Synthesis Flow

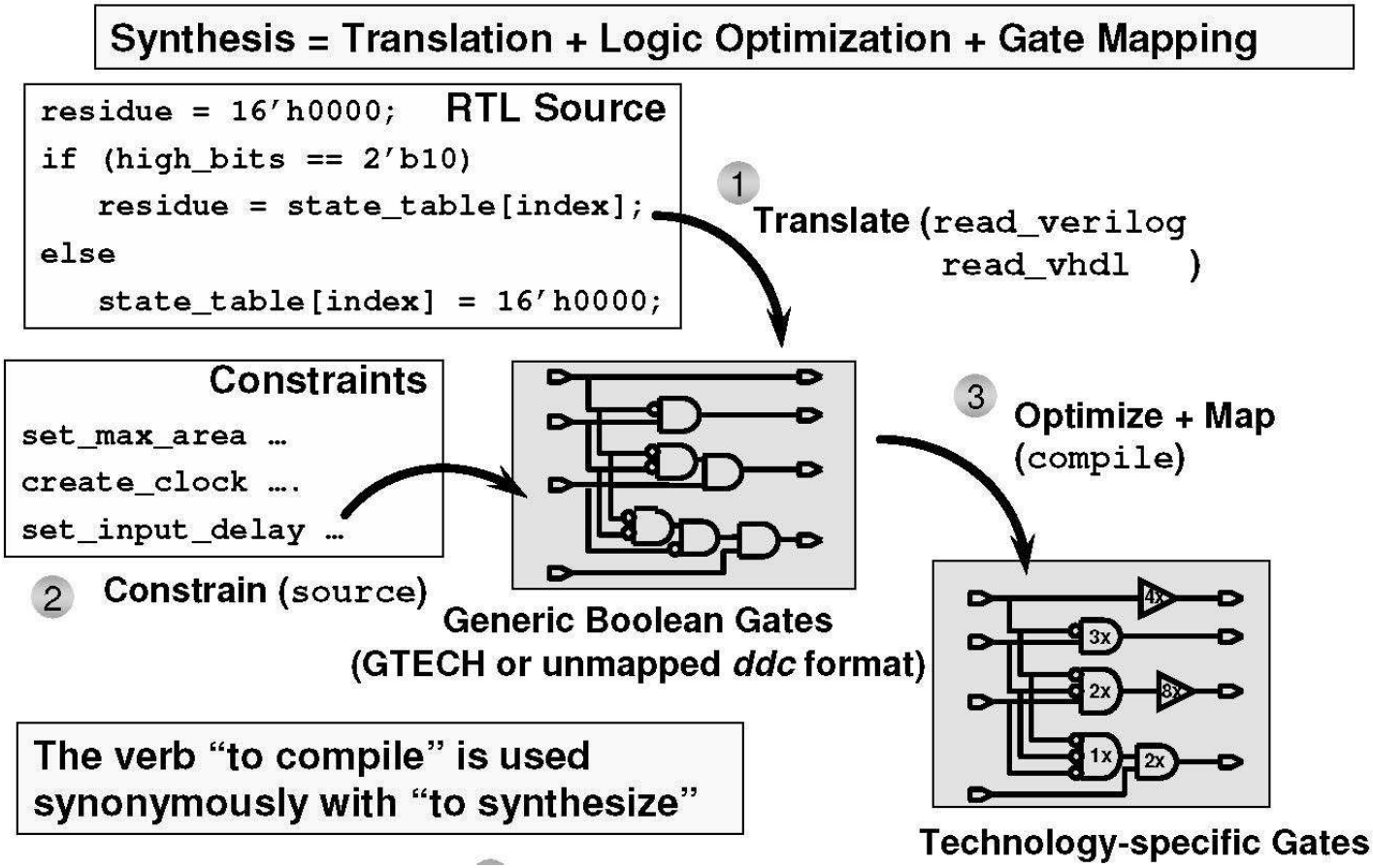
Asic Synthesis Flow



Design Compiler Flow



Synthesis Transformations

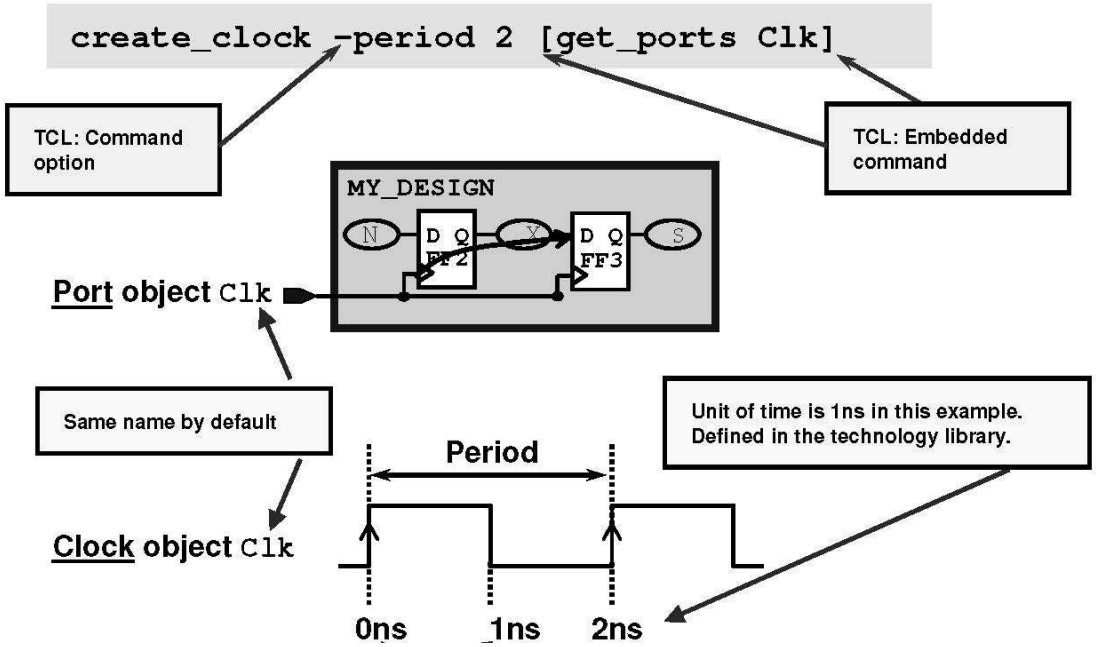


Default Clock Behavior

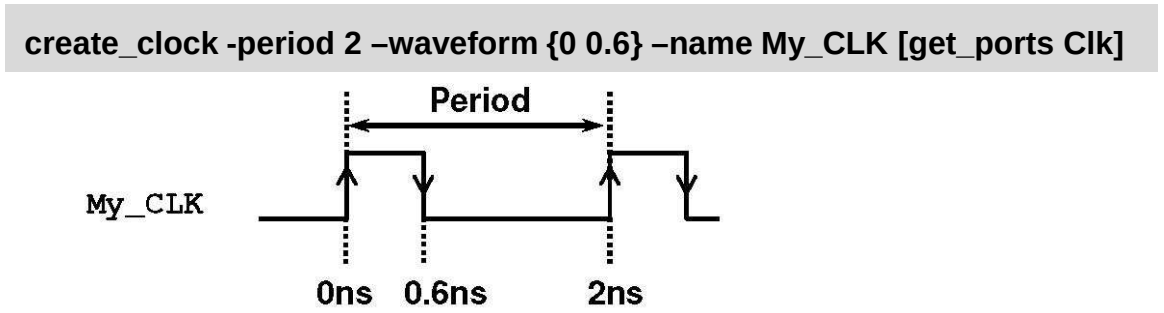
- **Defining the clock in a single-clock design constrains all timing paths between registers for single-cycle, setup time**
- **By default the clock rises at 0ns and has a 50% duty cycle**
- **By default DC will not "buffer up" the clock network, even when connected to many clock/enable pins of flip-flops/latches**
  - The clock network is treated as "ideal" - infinite drive capability
    - Zero rise/fall transition times
    - Zero skew
    - Zero insertion delay or latency
  - Estimated skew, latency and transition times can, and should be modeled for a more accurate representation of clock behavior

Defining a Clock

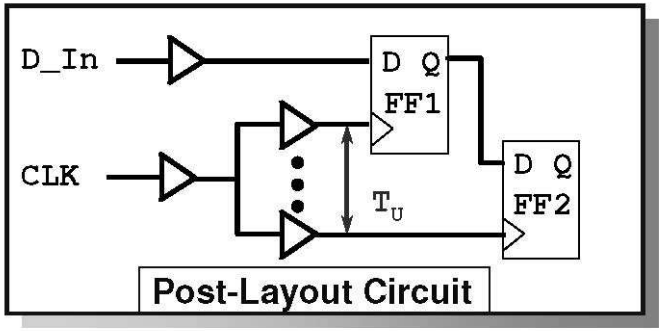
Default clock with 50% duty cycle

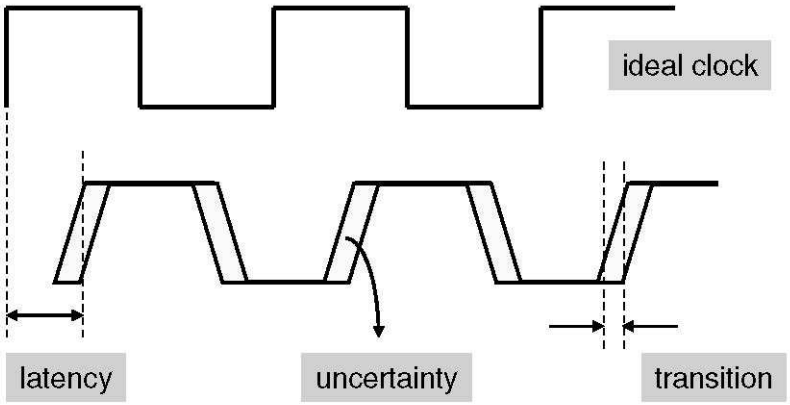


Clock with specified duty cycle



Modeling Clock

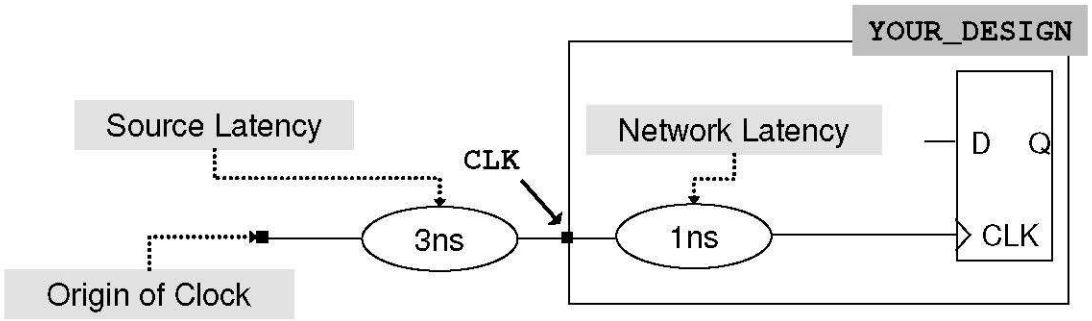




[jitter 0.05 + skew 0.3 + margin 0.15]

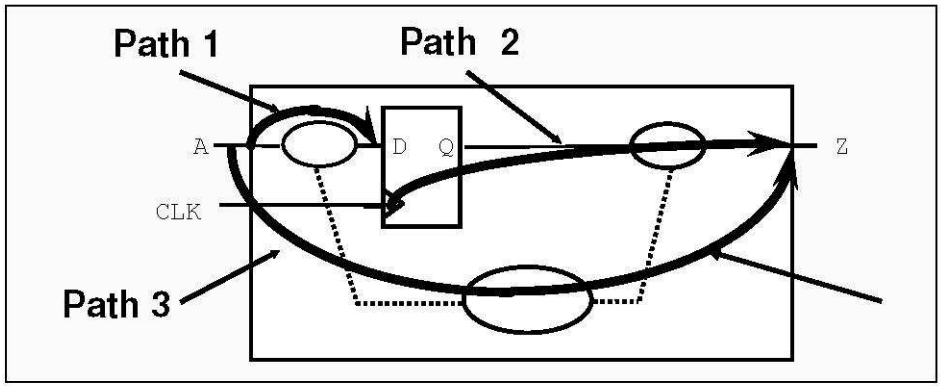
set_clock_uncertainty –setup Tu [get_clocks CLK]

Pre-Layout: clock skew + jitter + margin



reset_design
create_clock -p 5 -n MCLK Clk
set_clock_uncertainty 0.5 MCLK
set_clock_transition 0.08 MCLK
set_clock_latency -source –max 4 MCLK
set_clock_latency –max 2 MCLK

Specifying Setup-Timing Constraints

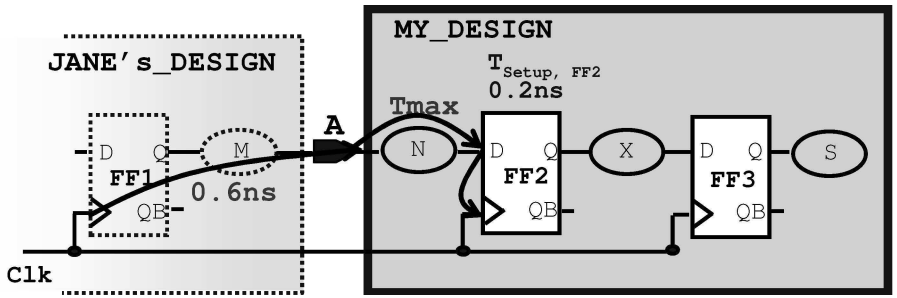


- **Objective: Define setup timing constraints for all paths within a sequential design**
  - All input logic paths (starting at input ports)
  - The internal (register to register) paths
  - All output paths (ending at output ports)

Constraining Input Paths

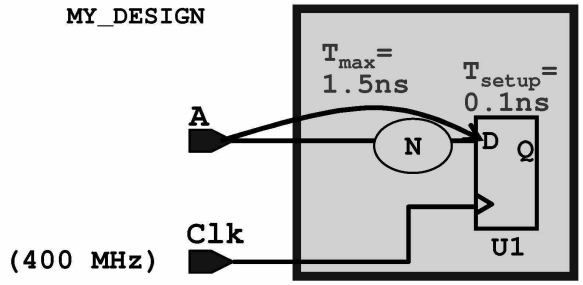
The user must specify the latest arrival time of the data at input A

What is T_{max} for N ?



create_clock –period 2 [get_ports Clk]

set_input_delay –max 0.6 –clock Clk [get_ports A]



The maximum delay for path N = 1.5 ns

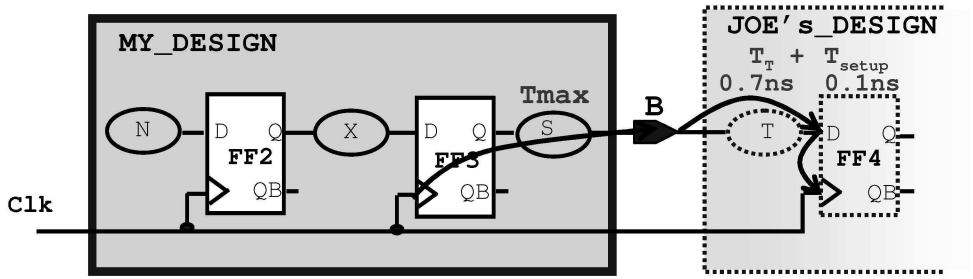
create_clock –period 2.5 [get_ports Clk]

set_input_delay –max 0.9 –clock Clk [get_ports A]

Constraining Output Paths

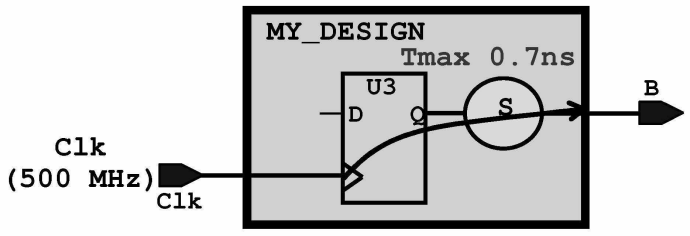
The user must specify the latest arrival time of the data at output B

What is T_{max} through S ?



create_clock –period 2 [get_ports Clk]

set_output_delay –max 0.8 –clock Clk [get_ports B]

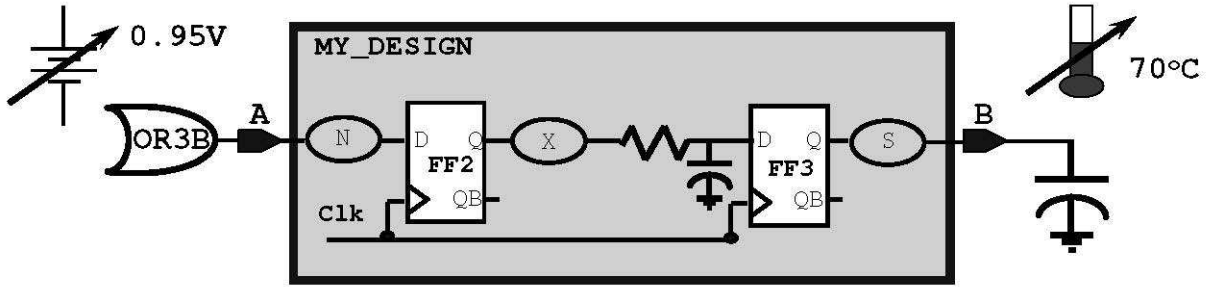


The maximum delay to port B = 0.7 ns

create_clock –period 2 [get_ports Clk]

set_output_delay –max 1.3 –clock Clk [get_ports B]

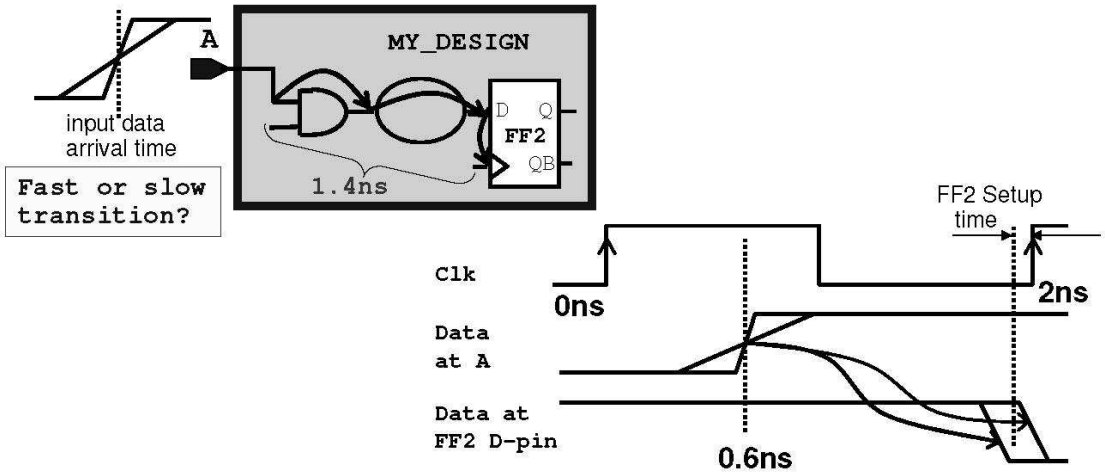
Environmental attributes



- Input drivers and transition times
- Capacitive output loads
- Process/Voltage/Temperature (PVT) operating conditions
- Interconnect parasitic RCs

Input drivers and transition times

- **Rise and fall transition times on an input port affect the cell delay of the input gate**



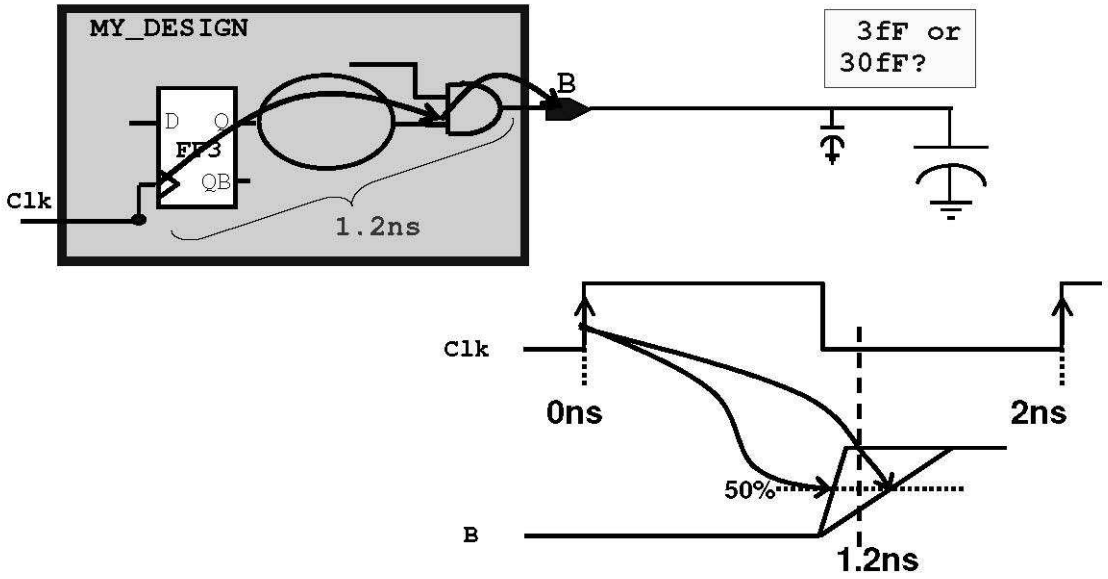
set_input_transition 0.6 [get_ports A]

set_driving_cell –lib_cell OR3B [get_ports A]

set_driving_cell –lib_cell FD1 –pin Qn [get_ports A]

Capacitive output loads

Capacitive loading on an output port affects the transition time and thereby the cell delay of the output driver



set_load [expr 30.0/1000] [get_ports B]

Wrap-Up

- Design Constraints
  - Power, Area, Frequency, CMOS Scaling
- Timing
  - Timing Metrics, Paths, Variability and Delay
- Deterministic Timing Analysis (Static Timing Analysis)
  - Models, Interconnect, Networks, Clock Distribution
- Statistical Timing Analysis
  - Probability, Spatial Correlations, MAX function
- Design Flow
  - Synthesis, Transformation, Definitions, Constrains