Custom Clock Distribution Network Design Fundamentals

Fuding Ge

Clock Uncertainty

- Jitter (temporal, dynamic uncertainty due to PLL, generally not considered in the design)
- Skew (spatial, static uncertainty)
  - Systematic: uneven load, routing,…
  - Statistic: process variation, temperature, noise, power supply,…
- Total clock inaccuracy (jitter & skew) is generally about 10%

The goal of clock distribution network is to deliver clock signal with minimum skew at the cost of reasonable power, area, delay.

Skew Source

- Non-uniform load
- Temperature Gradient
- Threshold Voltage Fluctuation
- Transistor channel length Tolerance
- Gate oxide thickness Tolerance
- Wire thickness variation

Design Target

- Delay (Power-Area-Delay Trade off)
- Skew (As small as possible)
- Duty cycle (Keep constant: rising edge delay = falling edge delay, not rising time = falling time)
- Slew rate (large means more power, large noise, and small delay. The Slew rate should be kept reasonable fast)

Noise Reduction

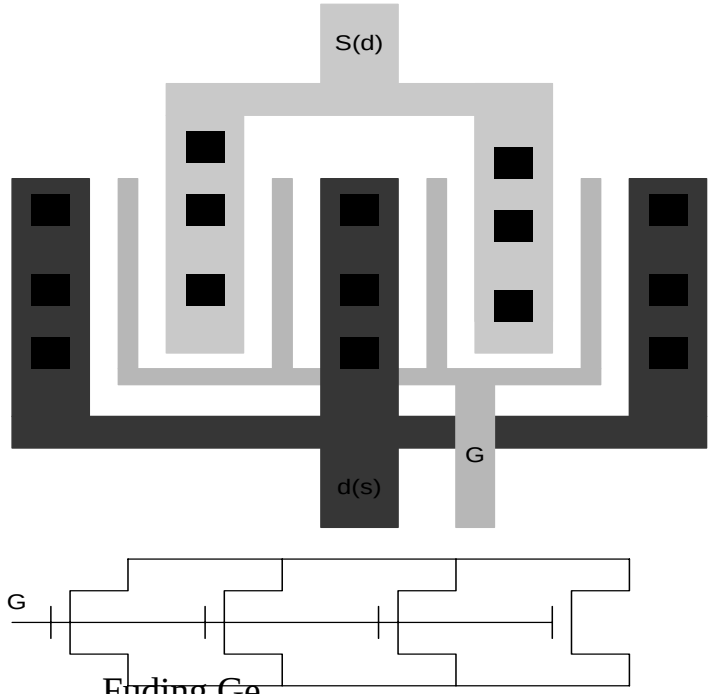
- Shielding the clock signal path:
  - Minimum width vcc and vss line in each side of the clock routing metal line (same metal layer).
  - Minimum crossover signal lines, shielding (different metal layers).
- Decoupling Capacitor:
  - Size of decoupling capacitor = 3 – 5 (load capacitor) for each buffer (inverter).

Matching

- The goal is to reduce statistic skew.
- Using the same number of inversions for every branch.
- Using same gate for clock gating (both NAND or NOR).
- Using the same size inverters if possible.
- Same transistor orientation with each other.
- Clock choppers should be close to the clock signal destination (usually bistable elements).
- The average polysilicon density in a region should be kept nearly constant.

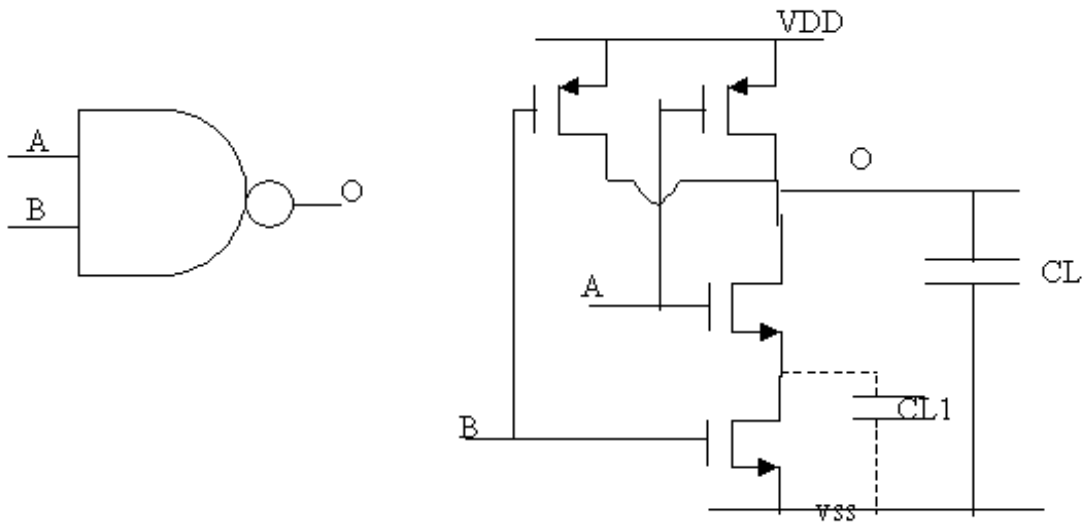
Layout Issue: Matching

- Mirror: Sometime the layout is to mirror one block to become another block. Make sure when the transistor flips, no mismatch occur: break it into even number fingers



Equivalent Pin Ordering (Low Power Idea)

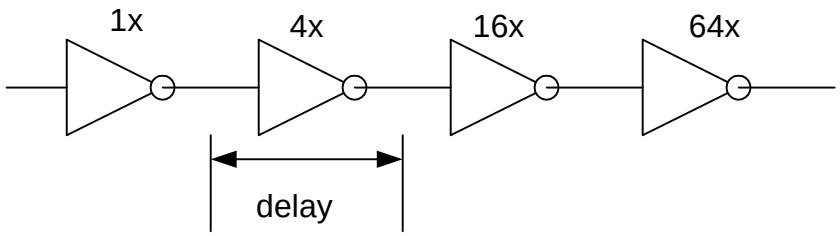
- Transitions involve transistors closer to the output node have **less delay** and consume **less energy !!!**
  - Clock signal always connect to the transistors closer to output node;
  - Example: NAND, Clock connect to pin A !



Fanout, Delay and Power

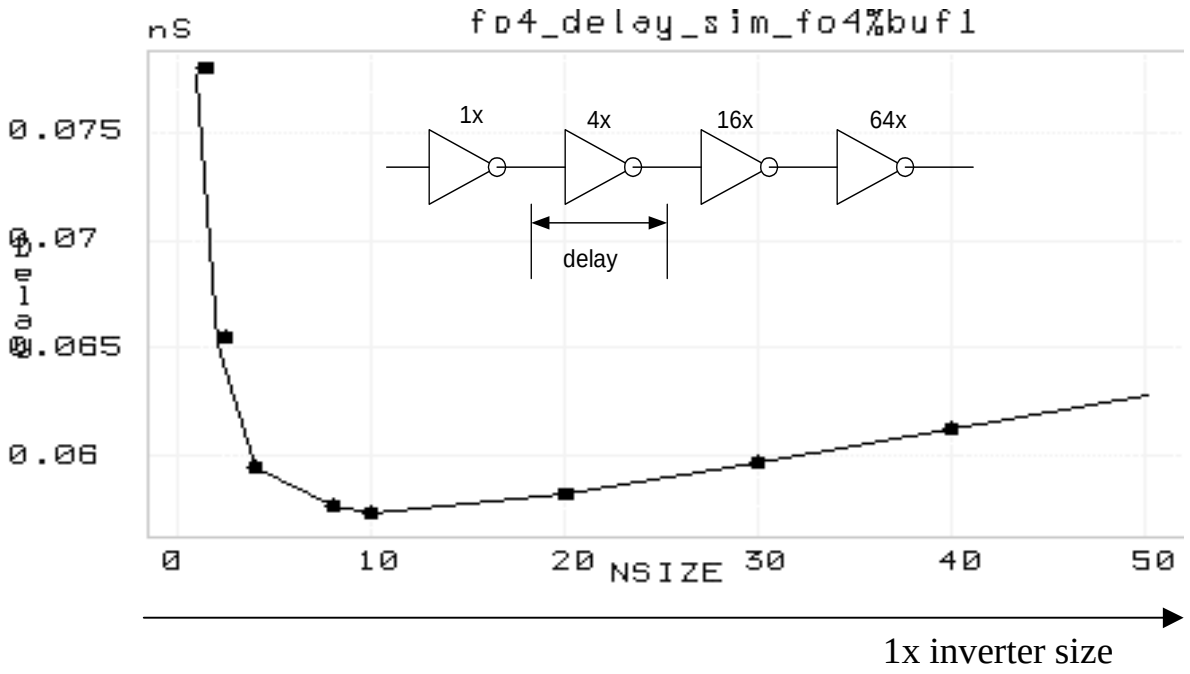
- Fanout = 3- 4 for minimum delay and power
- Fanout = 7-8 for minimum power and reasonable delay
- Fanout = or < 2, both power and delay increase dramatically.
- The last drive usually need a large slew rate to reduce skew (Fanout = 3 is good)

FO4 Inverter Delay



- FO4 (fan-out of 4) inverter delay is a process-independent unit of delay.
- Divide speed of circuit by speed of FO4 delay to get a metric pretty stable over process, voltage and temperature.
- Cycle time can be expressed with FO4 inverter delay to show the design aggressiveness.

FO4 Inverter Delay



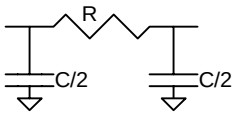
This figure shows the FO4 delay VS the transistor size. 0.13 μm CMOS technology.

Interconnect Modeling Parameters

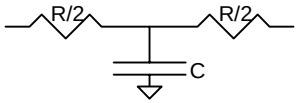
- Width and length of the wire and which metal layer it is
- The metal layer above and below
- Space between the interconnect and its neighboring line in the same metal level (space).
- Miller Coefficient models the line-line capacitance

Line-Line Capacitance is typically 60-80% of the total wire capacitance for a minimum wire at minimum pitch in a fully occupied wiring environment.

Interconnect Modeling

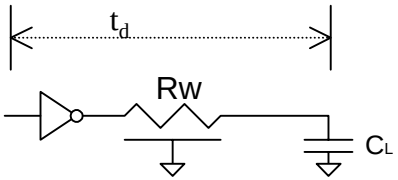


π Model

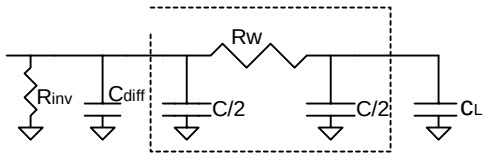


T model

It is recommended to model long wire using 4 π–segments in simulation and one for hand calculation.



Original Circuit



RC Model

The Elmore delay of the model is:

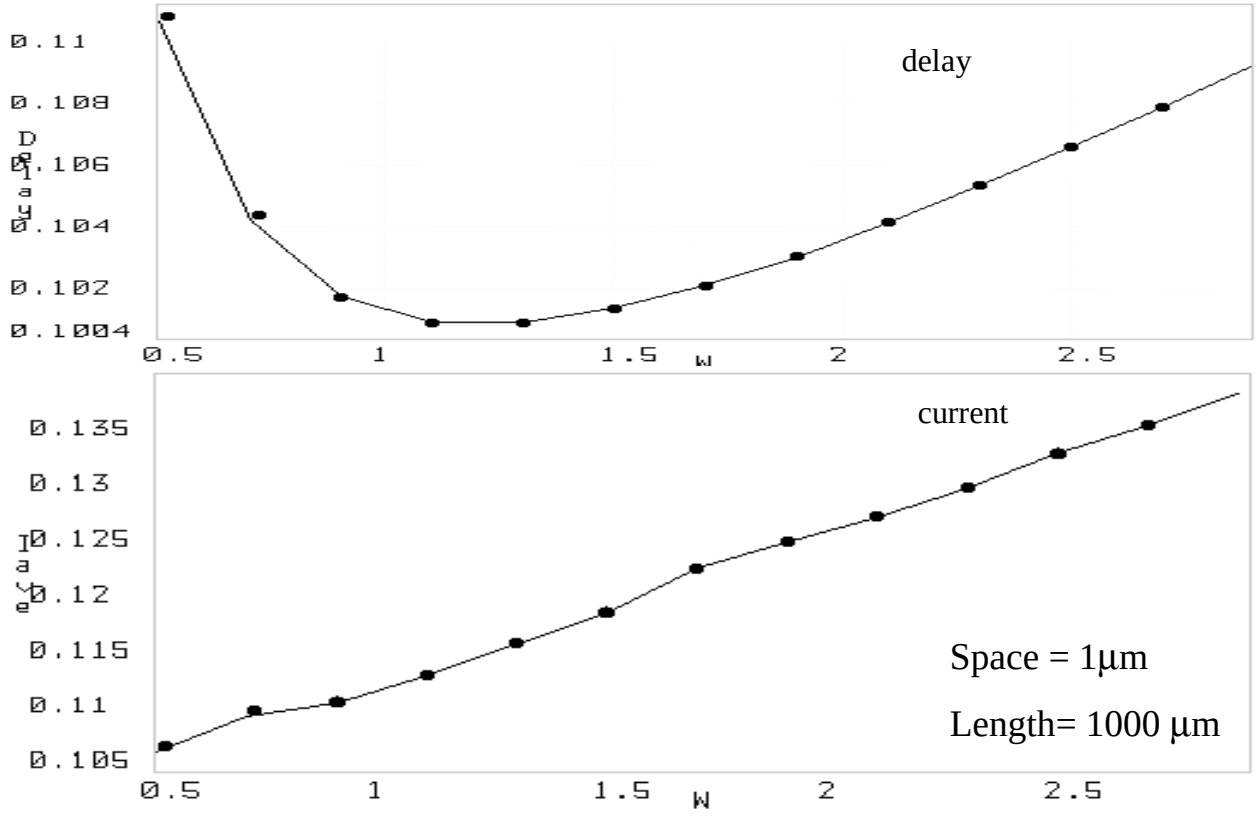
$$t_d = R_{inv}(C_{diff} + C/2) + (R_{inv} + R_w)(C/2 + C_L)$$
$$= R_{inv}(C_{diff} + C_L + C) + R_w C/2 + R_w C_L$$

1st term: inverter driving its own diffusion, the load and the wire capacitance;

2nd term: quadratic delay of the wire self loading;

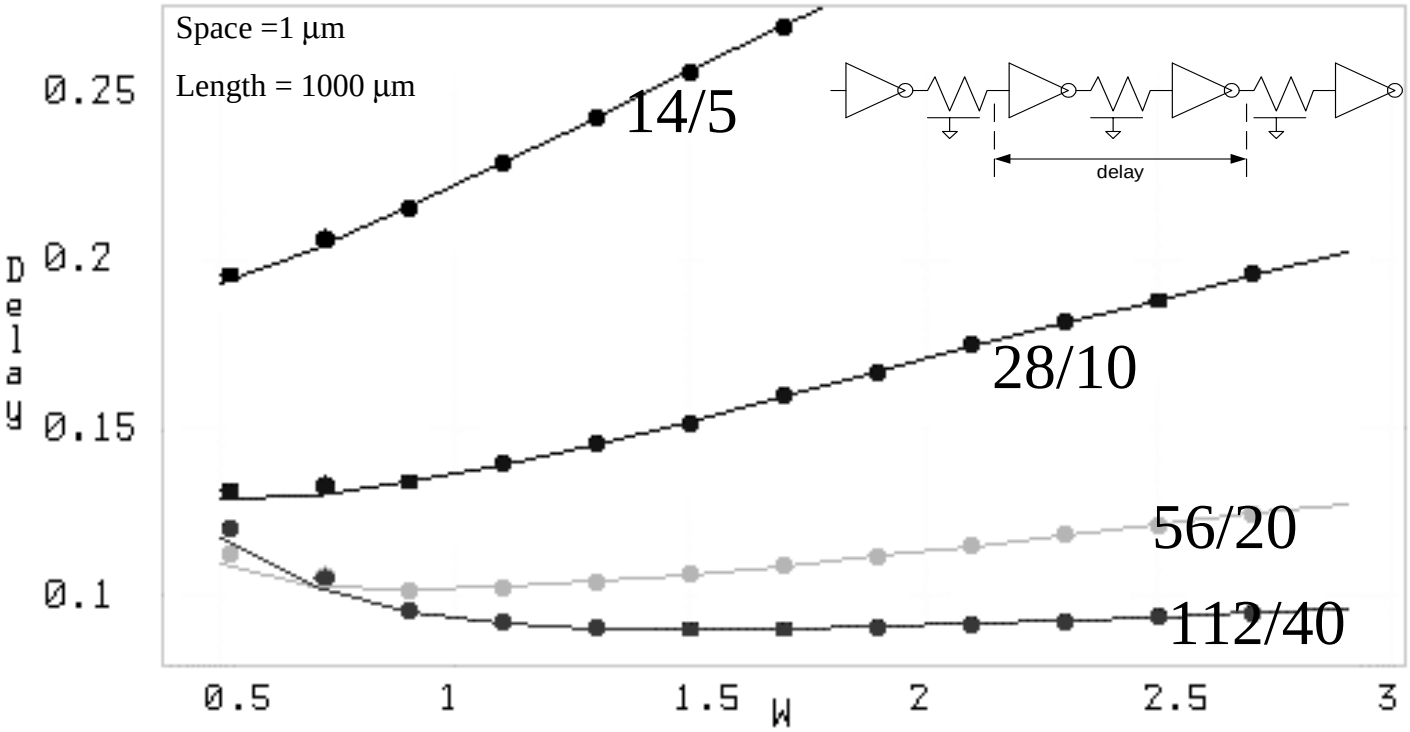
3rd term: extra delay contributed by wire resistance discharging the load capacitance.

Interconnect Delay and Power VS Width



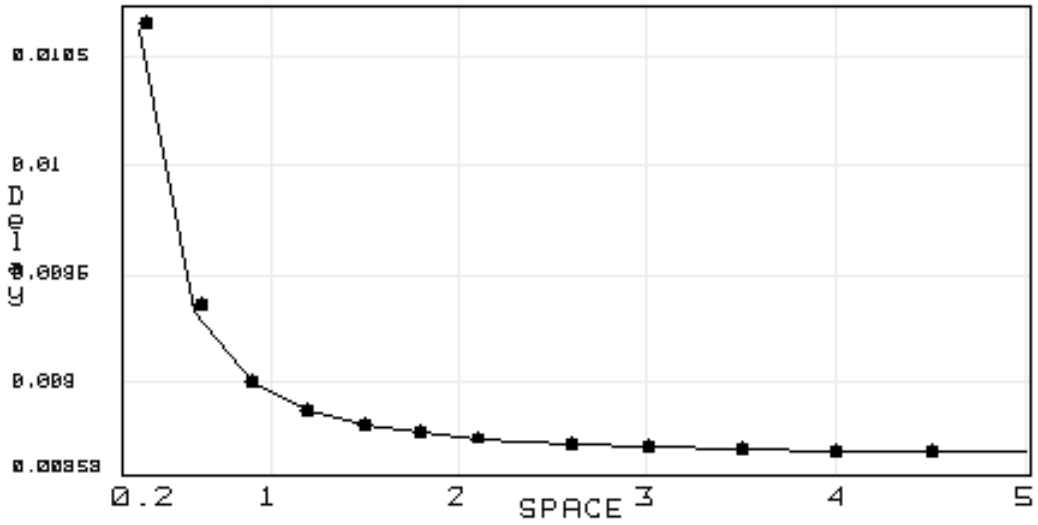
For a given driver size and wire space, there is a interconnect width that the delay shows minimum value. The reason is that as the width increases, the resistance decreases, while the capacitance increases.

Minimum Interconnect Delay VS Inverter Repeater Size



The best interconnect width is driver size dependent !!!

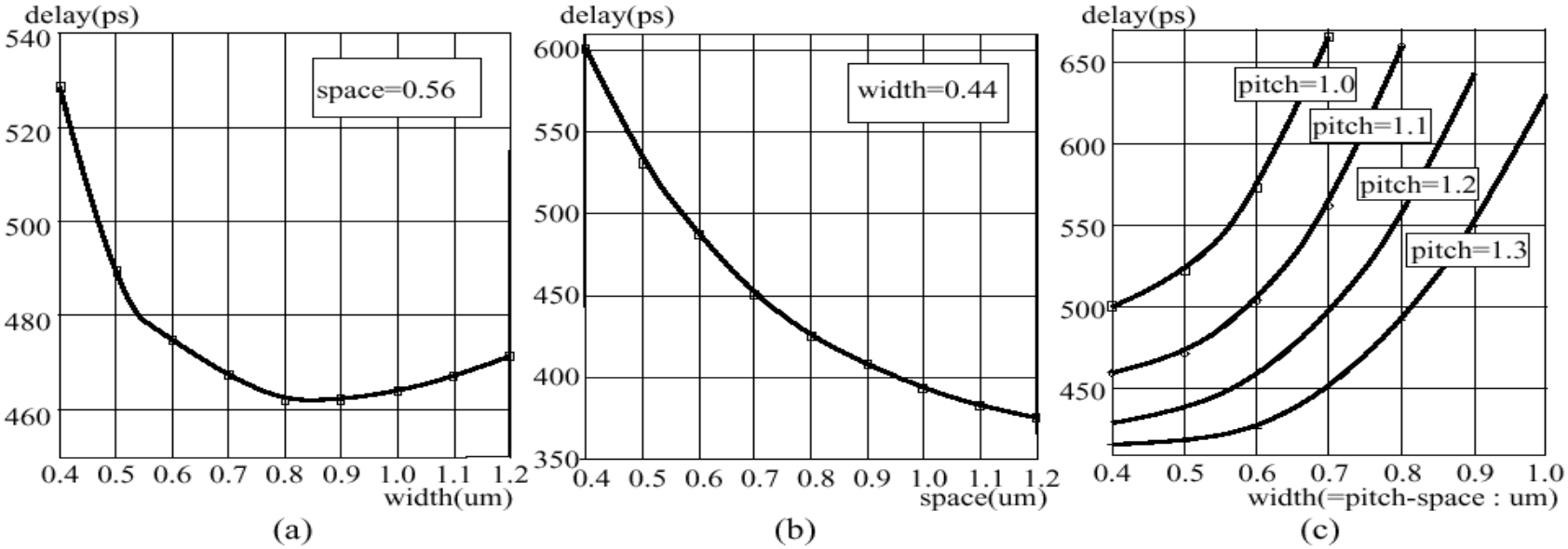
Interconnect Delay VS Wire Spacing



Increasing the inter-wire space reduces the cross coupling delay.

It is typically better first to increase wire spacing rather than width when trying to minimize interconnect delay. This is more important because of non-uniform scaling of metal layers. The thickness of metal layers is not being scaled down as fast as the width and length dimensions.

Interconnect Delay VS Wire Pitch



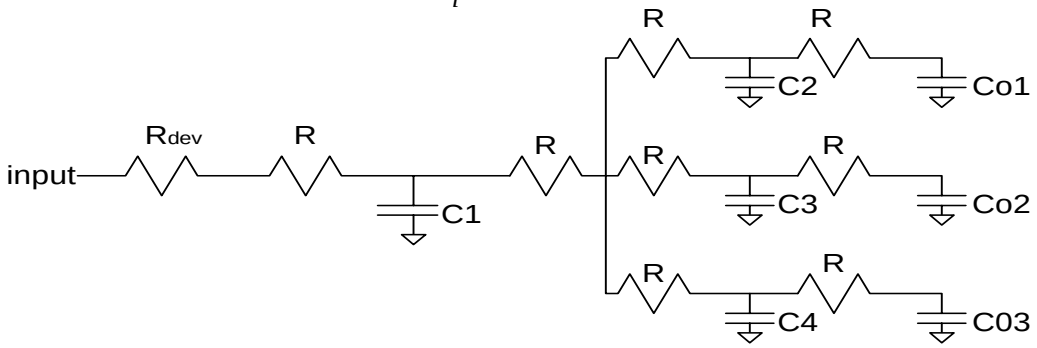
For the same pitch, increasing the width (and, therefore decreasing the space) tends to increase the interconnect delay, as shown in Fig.c above. Therefore, for a given pitch, it is better to increase the wire space than the wire width.

Reference: J. Yim, S. Bae, and C. Kyung “A Floorplan-based Planning Methodology for Power and Clock Distribution in ASICs”, Proc. Design Automation Conference, PP. 766-71, 1999

Elmore Delay

Equation: $$D_i = \sum_{k \in P_i} R_k C_k$$

Example:



The delay from input to Co1 is:

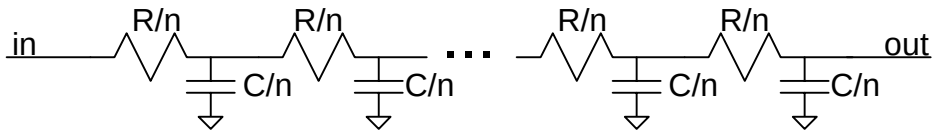
$$D_{Co1} = (Rdev + R)C1 + (Rdev + 2R)(C3 + Co2 + C4 + Co3)$$
$$+ (Rdev + 3R)C2 + (rdev + 4R)Co1$$

Elmore delay of a branched network can be described as the RC delay of the unique path from input to output where all side branches are replaced by their lumped capacitive value

Interconnect Delay Modeling Using Elmore Delay

Assume the total resistance and capacitance of the interconnect is R and C respectively. We model the wire using n segments of RC network. Then let n to be infinity to simulate the distribution manner of the wire.

Using Elmore delay model:

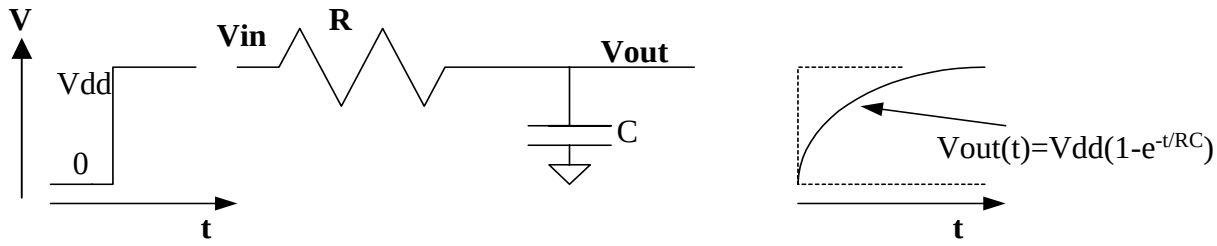


$$D = \frac{C}{n}\left(\frac{R}{n} + \frac{2R}{n} + \frac{2R}{n} + \cdots + \frac{nR}{n}\right) = \frac{n(n+1)}{2n^2}RC$$

$$\lim_{n \to \infty} \frac{n(n+1)}{2n^2}RC = \frac{1}{2}RC$$

The delay of a distributed RC line is equal to the the delay of a lumped RC with half the resistance !!!

RC Delay and Rise (Fall) Time Constant



When Vout is 50% Vdd:

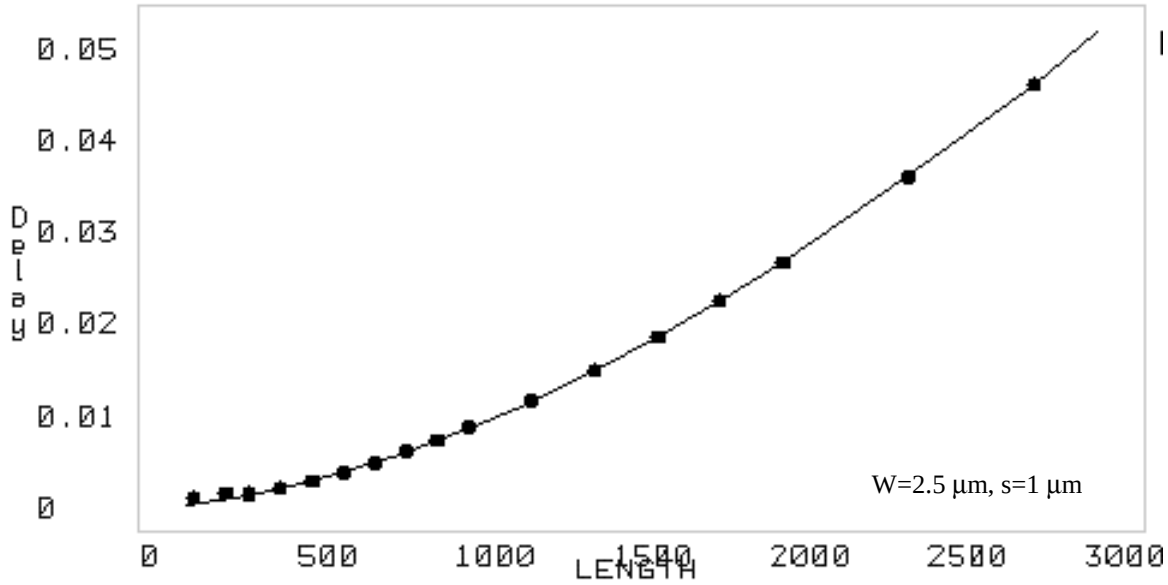
$$Vdd\ (1 - e^{-t/RC}) = \frac{1}{2} Vdd$$

$$t_d = RC \ln 2 = 0.69\, RC$$

The rise time from 10% Vdd to 90% Vdd is

$$t_r = RC \ln 0.9 - RC \ln 0.1 = RC \ln 9 = 2.2\, RC$$

Delay of Long Wire

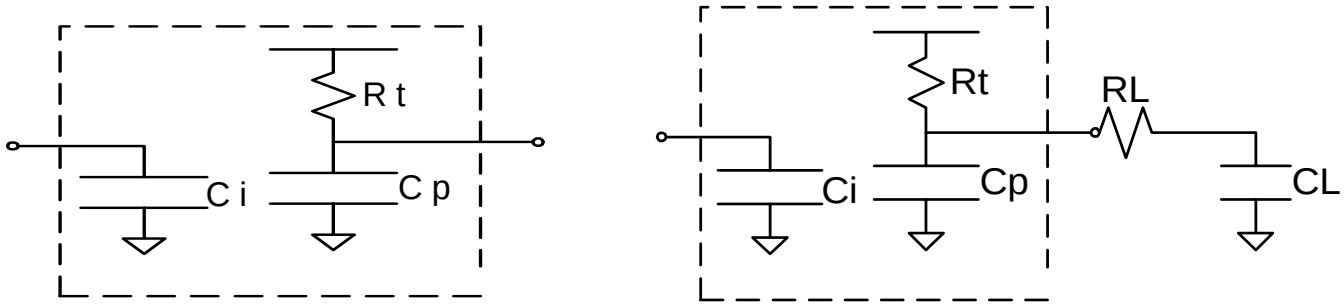


The delay of a long wire is proportional to the square of its length because both the resistance and capacitance increase linearly with length.

How to Minimize Long Wire Delay ?

- One way to minimize the delay of a long wire is to break the wire into segments and insert repeaters between segments. In this way the delay can be made linear with length.
- How many segments and how large the repeater ???
- We answer these questions step by step.

Device Modeling



•We model the driver as shown above, Ci is the input capacitance (gate capacitance), Rt is the on resistance and Co is the parasitic capacitance from source to drain. RtCo product is the intrinsic delay of the MOS transistor which is the delay of an inverter driving its own gate (about 12 ps for 0.35 channel length).

•When driving a RC network, the 50% delay for an step input is:

$$t_d = 0.69[RtCp + (Rt + RL)CL]$$

Logic Effort: General Gate Delay Model

•Logic effort L_e is the ratio of the input capacitance of a gate to the input capacitance of normal skew inverter with the same drive strength. It describes the relative ability of gate topology to deliver current (defined to be 1 for an inverter)

•Electric effort (fanout) E_e is the ratio of output to input capacitance CL/Ci.

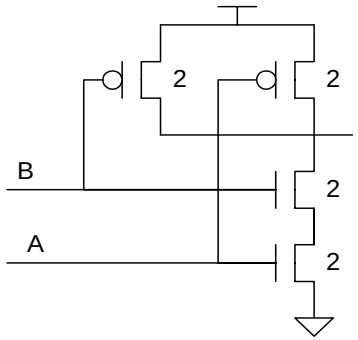
•t_0 is the intrinsic delay of the normal skew inverter and p=Cp/Ci

We have: $$t_d = t_0 L_e (E_e + p)$$

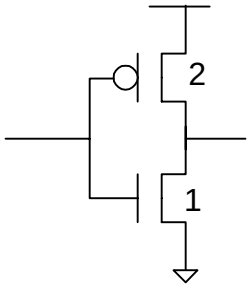
where $t_0 = 0.69RtCi$

t_0 is a process technology parameter, independent of the size of the inverter.

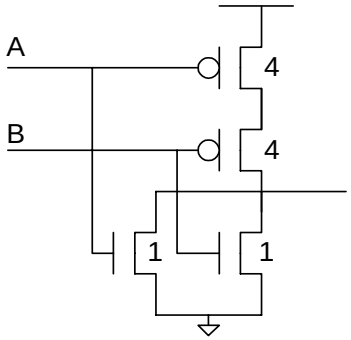
Logic Efforts of Some Gates



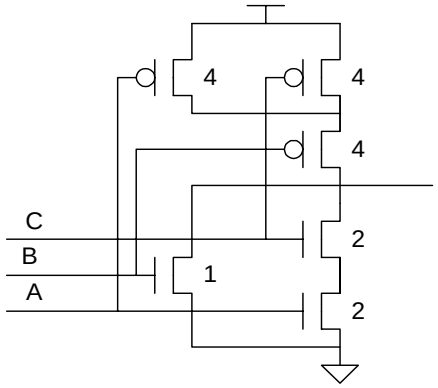
NAND, L_e=4/3



Inverter, L_e=1



NOR, L_e=5/3



AOI, L_e=2 for A,C; 5/3 for B

$L_e=(C_{gate}R_{gate})/(C_{inv}R_{inv})$. One find L_e by either making the input capacitance of the gate equal to the inverter, and looking the resistance ratio, or by making the resistance equal and looking the capacitance.

Long Wire Delay Driving by Inverter Repeaters

- Source/drain diffusion capacitance is negligible.
- The size of the inverter repeater are n times of an unit inverter whose on resistance and input capacitance are Rt and Ci respectively, so its on resistance is Rt/n and input capacitance is nCi.
- The resistance and capacitance per unit length of the wire are Rw and Cw respectively, and the wire length is L.
- The long wire is broken into S segments.

The delay through the entire wire, using Elmore delay equation, is:

$$t_w = t_0 S[(\frac{Rt}{n} + \frac{RwL}{2S})\frac{CwL}{S} + (\frac{Rt}{n} + \frac{RwL}{S})nCi\ ]$$

Optimal Inverter Size and Number of Segments

From $\partial t_w / \partial n = 0$ and $\partial t_w / \partial S = 0$

We have:

$$n = \sqrt{\frac{CwRt}{CiRw}} = \sqrt{\frac{Cwt_0}{Ci^2 Rw}}$$

$$S = L\sqrt{\frac{CwRw}{2CiRt}} = L\sqrt{\frac{CwRw}{2t_0}}$$

At these condition, the inverter delay is equal the wire RC delay which is RtCt

Optimal segment length: $l_s = \sqrt{\frac{CiRt}{CwRw \quad / 2}} = \sqrt{\frac{2t_0}{CwRw}}$

Wire delay: $t_w = L\sqrt{CwRwCiRt}(2+\sqrt{2}) = 3.41L\sqrt{CwRwCiRt}$

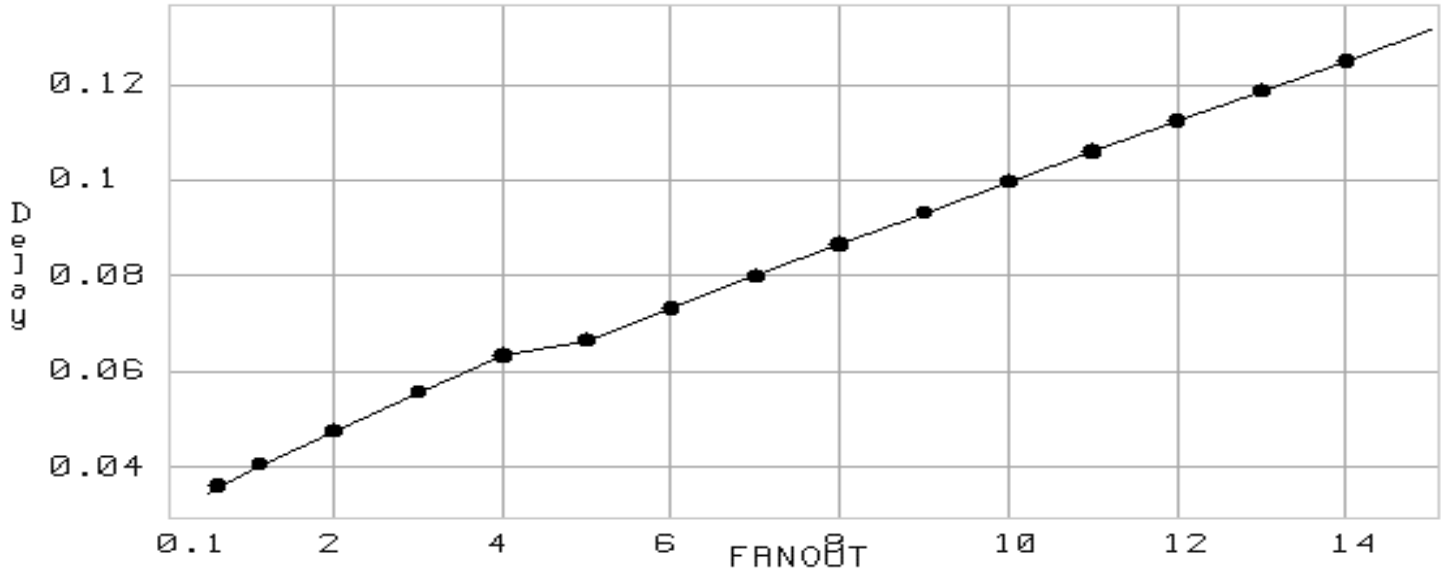
Note that CiRt is the intrinsic delay of the inverter driver. The optimal inverter size is independent of wire length; it is only a function of the physical parameters of wire and transistors.

Inverter Driver Characterization: t_0

For a inverter, it's logic effort Le is 1, we have:

$$t_d = t_0(E_e + p) = 0.69RtCi(E_e + p)$$

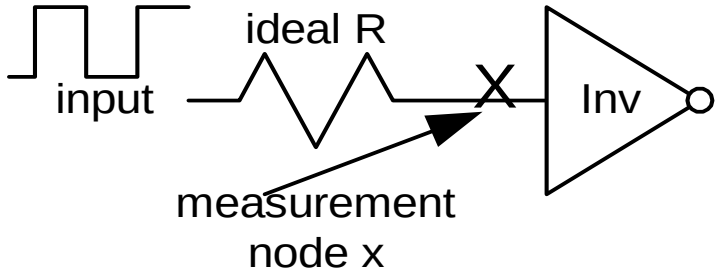
We see that delay is a linear function of fan-out and the intrinsic delay t_0 is the slope.

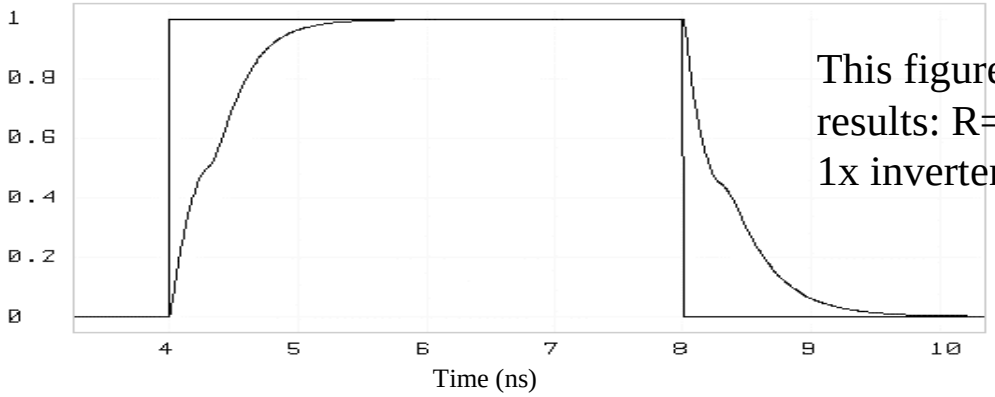


This is the results for 0.13 μm technology, t_0 is only about 7 ps!!

Inverter Driver Characterization: C_i

The input capacitance of the inverter can be measured using the following test circuit. The middle point delay from input to node x is 0.693RCi.





This figure shows the simulation results: R=100Ω and the load is 100 1x inverters.

The result should be compared with $Ci = \varepsilon_{ox}WL/T_{ox}$, where ε_{ox} is the dielectric constant of the gate oxide (about 35 aF/μm for SiO_2).

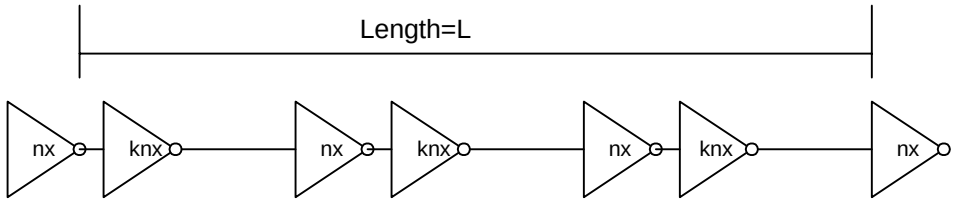
Optimal for Delay-Power Product

The issues with speed-optimal design are large driver size, high power dissipation and noise generation.

It has been shown that when the inverter size is scaled down by a factor of 0.7 from the speed-optimal size and the repeating length is enlarged by 1.43 (=1/0.7), the delay-power product reaches minimum.

Reference: Hongjiang Song, “A theoretical design basis for low power and small area VLSI interconnect repeaters”,

Buffer Repeater Optimal for Speed



The delay of the interconnect is minimized by choosing the drivers size n, k and segments number S. The delay is:

$$t_w = t_0 S[(\frac{Rt}{nk} + \frac{RwL}{2S})\frac{CwL}{S} + (\frac{Rt}{nk} + \frac{RwL}{S})nCi + kRtCt\ ]$$

Take the partial derivatives with respect to k, n and S, we get:

$$k = \sqrt{1 + \frac{CwL}{nSCi}} \qquad n = \sqrt{\frac{CwRt}{kCiRw}} \qquad S = L\sqrt{\frac{CwRw}{2(k + 1/k)CiRt}}$$

Buffer Repeater Size

$$k = \sqrt{2 + \sqrt{5}} = 2.06$$

The ratio of inverter size is independent of wire and transistor characteristics.

$$t_w = 3.64 L \sqrt{CwRwCiRt}$$

Inverter or Buffer Repeater ?

No polarity problem for buffer repeaters and longer repeat length for buffer repeater.

But:

Delay is 3.64/3.41=1.067, %7 delay penalty for buffer repeater;

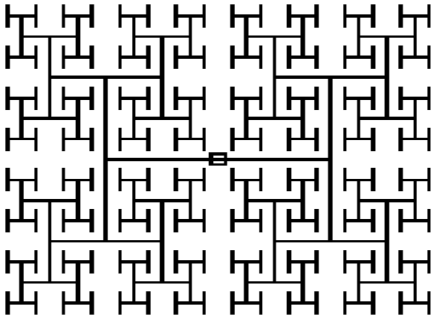
Area penalty is about 1.33;

Power is about 1.17 times of the inverter repeater.

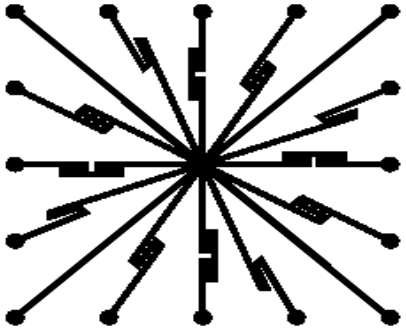
Critical Signals VS Non-critical Signals

- Critical Signals: Signals need to be synchronous with signal out of the chip
  - Skew must be small, global skew
- Non-Critical: synchronous inside the chip
  - Signals between different blocks
  - Skew can be larger than the critical signals, local skew

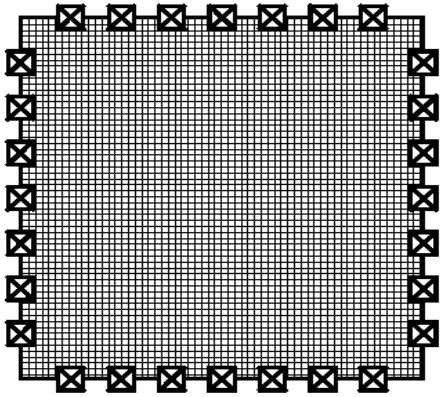
Clock Distribution Topologies: H-Tree



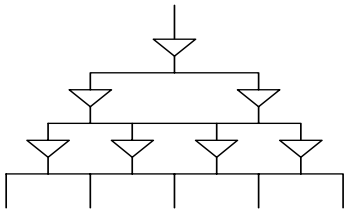
H-Tree: Non-uniform load distribution leads to skew. Best wiring efficiency. Poor automatic clock routing



Serpentine: Great amount of wiring resource. Easy to implement with non-even load

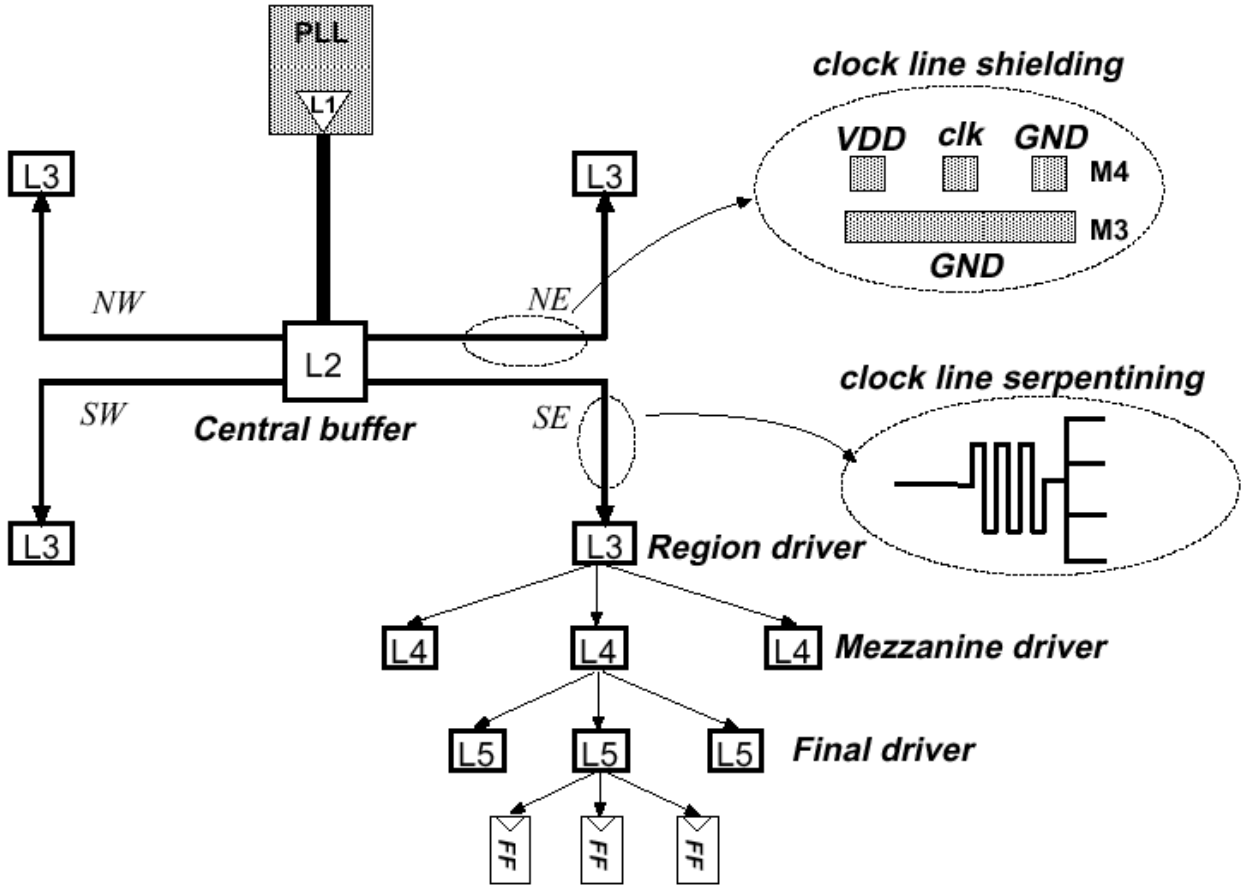


Grid: Relatively independent of the actual distribution of clock load, very robust. Excellent for automatic routing



A mesh distribution reduces the skew by tying the outputs of their drivers together in a 2-dimensional mesh. It can range anywhere from a global grid mesh to a localized net at the end of a H-tree. Skew can be very small. Large static current (so large power)

Hierarchical Clock Distribution Scheme Example



Reference: J. Yim, S. Bae, and C. Kyung "A Floorplan-based Planning Methodology for Power and Clock Distribution in ASICs", Proc. Design Automation Conference, PP. 766-71, 1999

Rules of Thumb

- Distribute a single global clock, then locally derive the multiple phases near where they are necessary.
- The number of buffering stages in the clock system should be minimized to reduce the skew introduced by process variation
- Clock buffer stages should be scattered across the chip to avoid large RC effects by reducing interconnect lengths.
- An unbalanced clock buffering system is inevitable, clock skew ranges typically 5%.
- Enough decoupling capacitance should be used to reduce VCC reduction and ground bounce.

Error Analysis

- Assume the clock tree has n stages of inversion, and every stage can introduce S_i ps skew, the standard deviation is:

$$\sigma = \sqrt{\sum_{i=1}^{n} s_i^2}$$

Assume S_i = 5 ps,

If n = 6 , σ=12 ps

If n = 8, σ =14 ps