Solutions to the metal interconnect design and analysis bottleneck

Eli Chiprout
research staff member

Austin Research Laboratory
Research Division
IBM

Interconnect functions

- Enable communication of signals among devices
- Distribute a driving electric voltage to every device (Vdd)
- Distribute a ground return for the entire chip (Gnd)
- Provide a driving clock (sometime more than one) in synchronous design to the entire chip (Clk)

Shift of views

- "***A VLSI circuit can be viewed as a dense system of metal conductors in a stratified dielectric medium over a silicon substrate***" -- recent DEC journal paper.

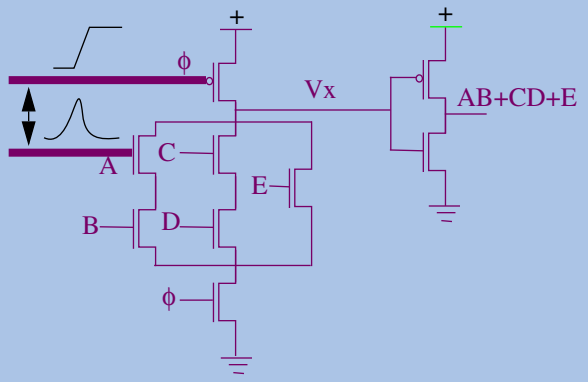
Change in engineering approach

	OLD	NEW
Layout	**place transistors and connect them**	**place signal paths and fit useful circuitry in between**
Models	**complex transistor models, C for interconnect**	**complex interconnect models, thousands of nodes with few transistor nodes**
Design	**functionality**	**clean signal transmission**
Hierarchy	**digital**	**none --> analog**

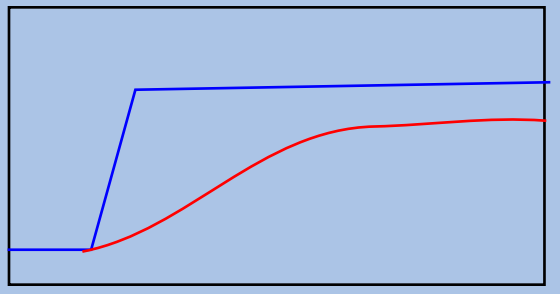
- Analogy: message passing among individuals. Not only does the sender have to create a proper signal and send it, but the medium of transmission must be appropriate.
- Both signal integrity from point A to B across a channel must be maintained (no loss, distortion, etc.) as well as freedom from interference of other channels (coupling).

Desirable properties

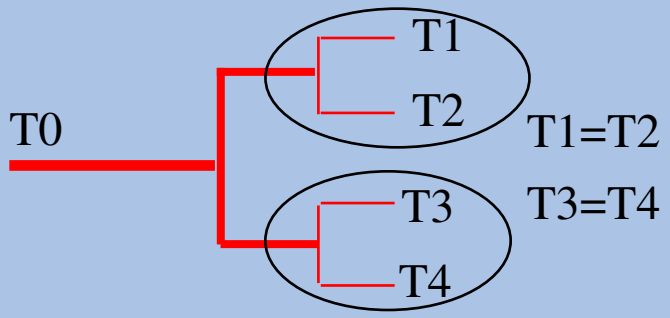
- Ideal lines should be
  - noise free



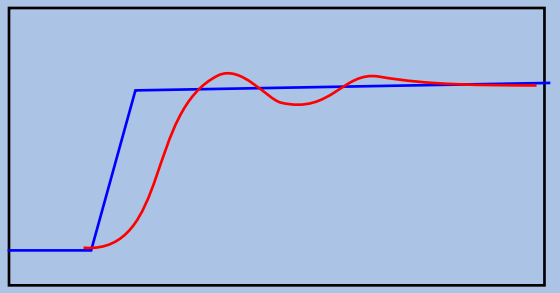
  - loss free



  - skew free (Clk)

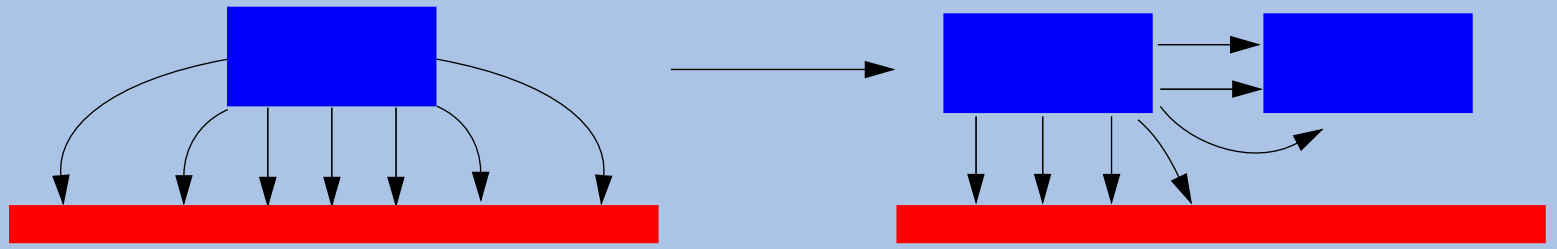


  - low inductance (usually)



- General trends in scaling and frequency are degrading all of these measures.
- Interconnect has become a significant portion of signal delay time and is expected to become the dominant portion on the critical path.

Fundamental properties



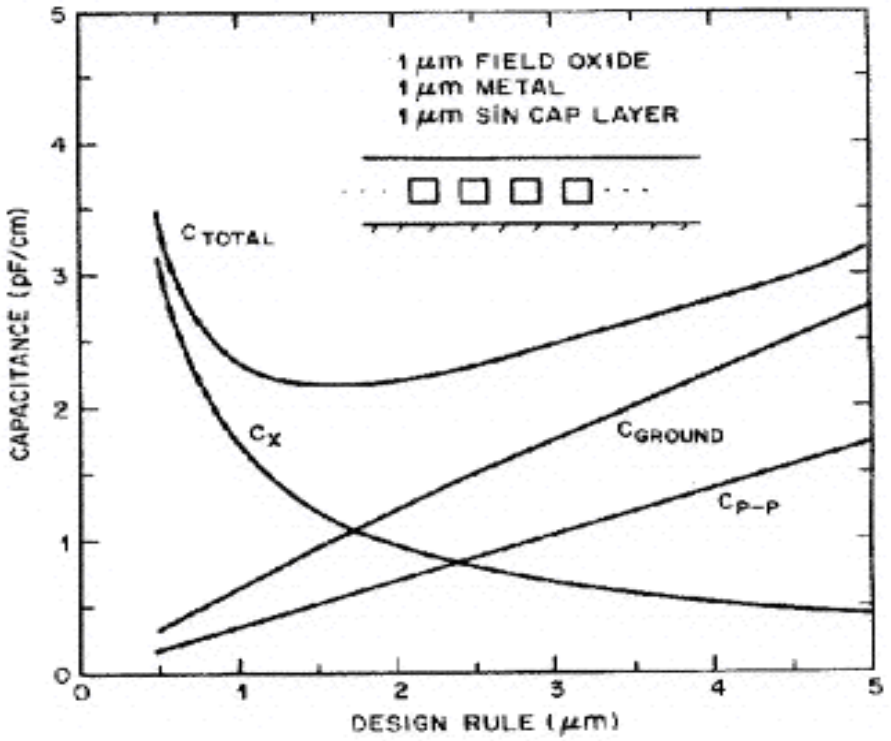
- The closer another object is the more field lines will terminate there, the more coupling (mainly capacitive, some inductive).
- Stable ground or power suplies should preferably be close to signal lines
- Object with active signals should be isolated from the immediate neighborhood.
- Resistance per length of a line is inversely proportional to cross sectional area:

$$R = \frac{\rho l}{A}$$

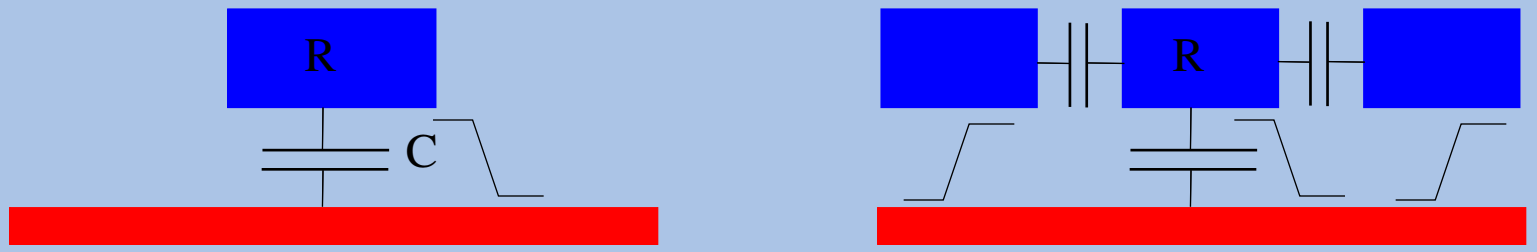
- Total resistance and capacitance varies linearly with the length of the line
- Time delay of the line is directly related to its capacitance and resistance: “RC” delay
- Noise is increased with the proximity to switching neighbors (capacitive and inductive coupling)

Fundamental properties

Impact of Interwire Capacitance



Fundamental properties



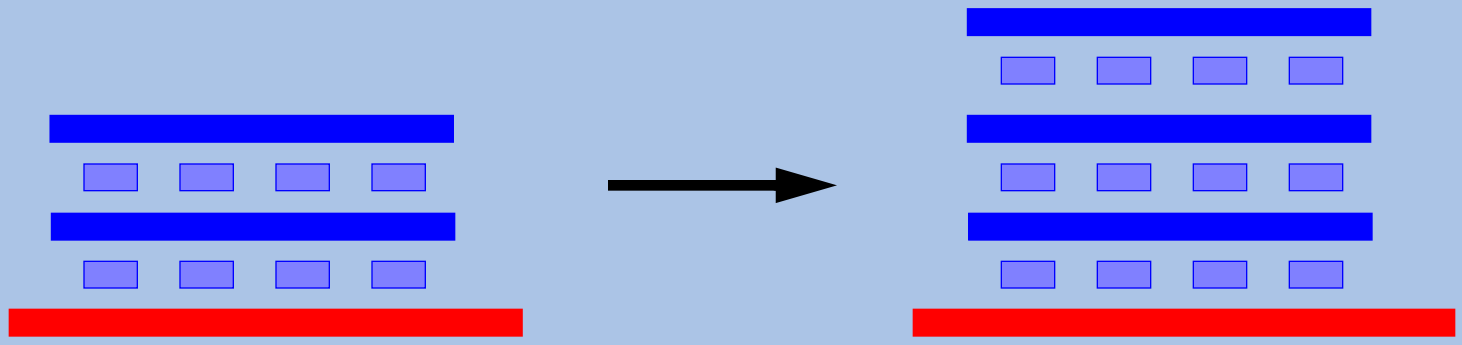
- Delay with dominant cap to ground depends only on the signal and the line's RC delay: static timing analysis reasonably accurate
- Delay with coupling to other lines becomes pattern dependent and highly inaccurate in static timing (but this method is all we have when millions of transistors!)
- Design is becoming limited by coupling interaction not lithography!

M2 wire with width 0.4u, spacing=0.4u and thickness=0.4u

Wn:Wp	wire length	data pattern	wire delay	difference
13.3u:30.6u	0.2 mm	- + -	13.0 ps	+76% -60%
13.3u:30.6u	0.2 mm	0 + 0	7.4 ps	
13.3u:30.6u	0.2 mm	+ + +	3.0 ps	
26.5u:61.0u	0.2 mm	- + -	7.4 ps	+ 72% - 53%
26.5u:61.0u	0.2 mm	0 + 0	4.3 ps	
26.5u:61.0u	0.2 mm	+ + +	2.0 ps	

Trends on chip: layers

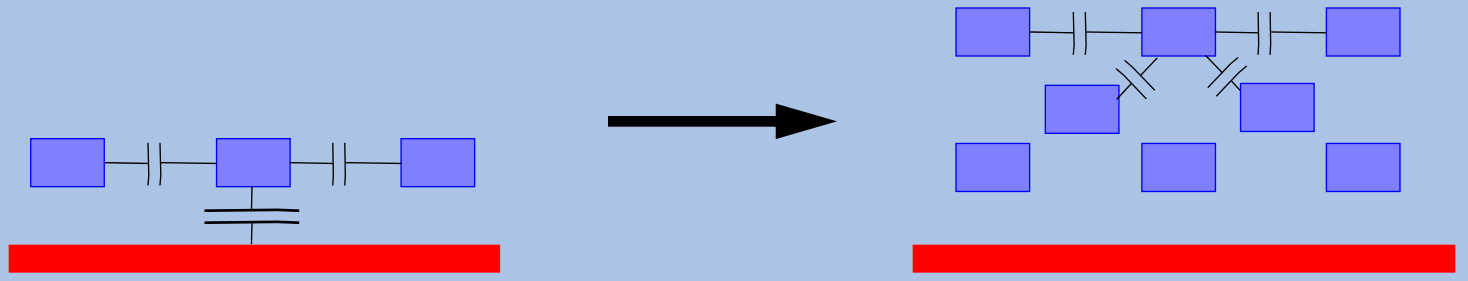
- Increasing number of metal layers:



- Needed due to tight integration of sub-micron devices
- Provides more wiring freedom, denser circuitry, greater area usage
- Results in:
  - greater susceptibility to switching activity in neighboring line

Trends on chip: layers

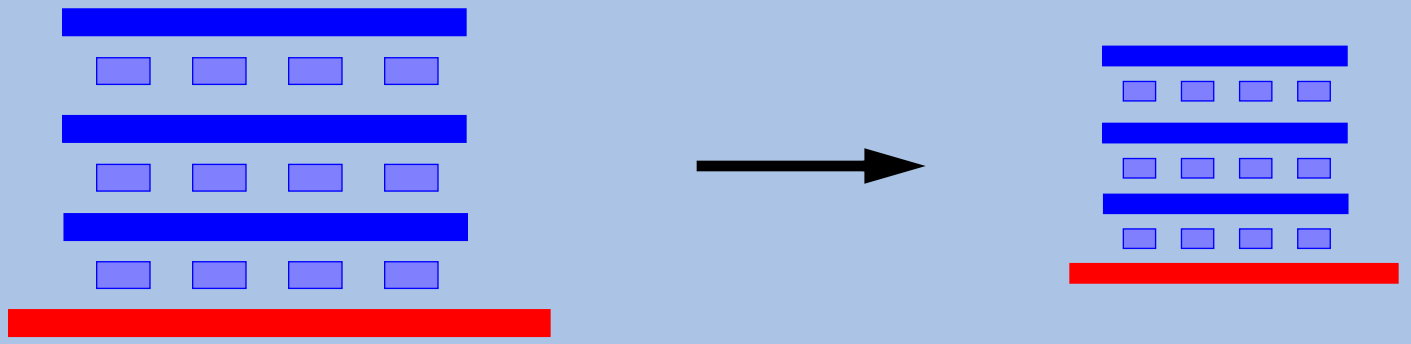
- Increasing number of metal layers:



- Dominant stable ground allows field lines to terminate there and ensures less noise coupling (less capacitance) from neighboring lines
- multilevel interconnection scheme lacks a stable reference and one reference is not necessarily more dominant than another
- Ground must often be explicitly provided by running a power supply nearby and distancing other signal lines
- Need: more accurate 3D capacitive extraction/modeling and for more layers (CPU intensive)

Trends on chip: scaling

- Downward scaling (by S) of cross-section dimensions (but not length):

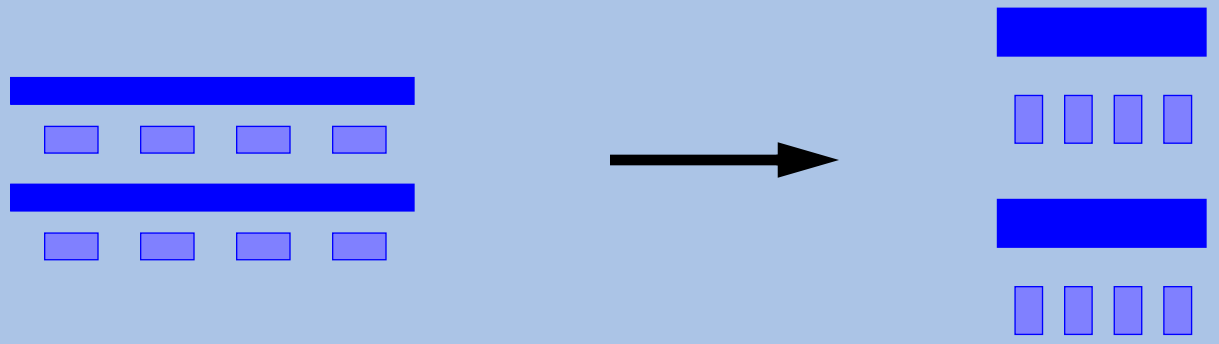


- Results in:

  - increased resistance of lines $O(S^{1.5}\text{-}S^{2})$.
- Lines become more lossy, distribution of supply or clock on long global lines becomes increasingly difficult
- Length of global interconnect lines remains approx. same because the size of the die is static or growing with technology --> more sever problem
- Increased RC delay. Global > $O(S^{2})$.

Trends on chip: aspect ratio

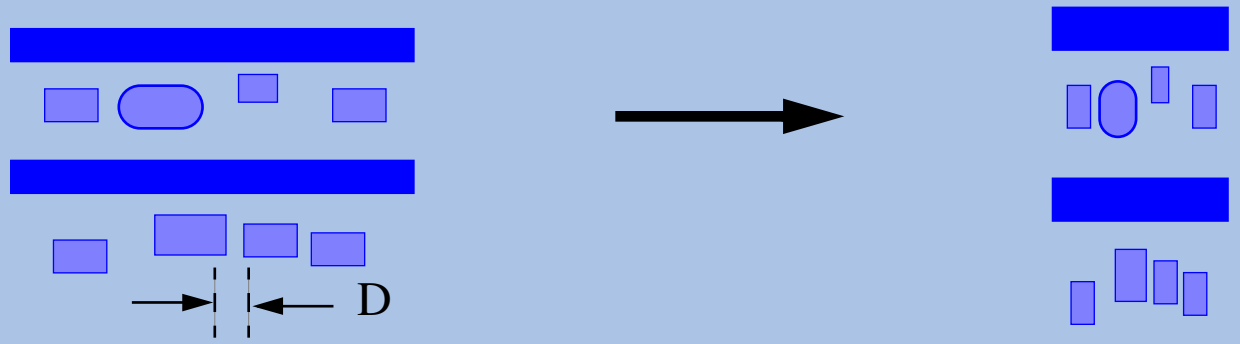
- Increasing height to width ratio to minimize resistance (for reduced resistance while decreasing feature sizes)



- Results in:
  - increased “side-wall” capacitance to neighboring lines
  - increased coupling noise from neighboring signal lines
- Need:
  - more accurate modeling tools and techniques
  - floating caps in simulation (expensive in timing simulator)

Trends on chip: variability

- Increasing impact of process variability on timing and skew



- Results in:

  - increased $\frac{\partial C_{ij}}{\partial D}$ with decreased D

  - increased sensitivity to corner rounding

  - sensitivity to layout densities, interaction between neighbouring geometries

- Need:

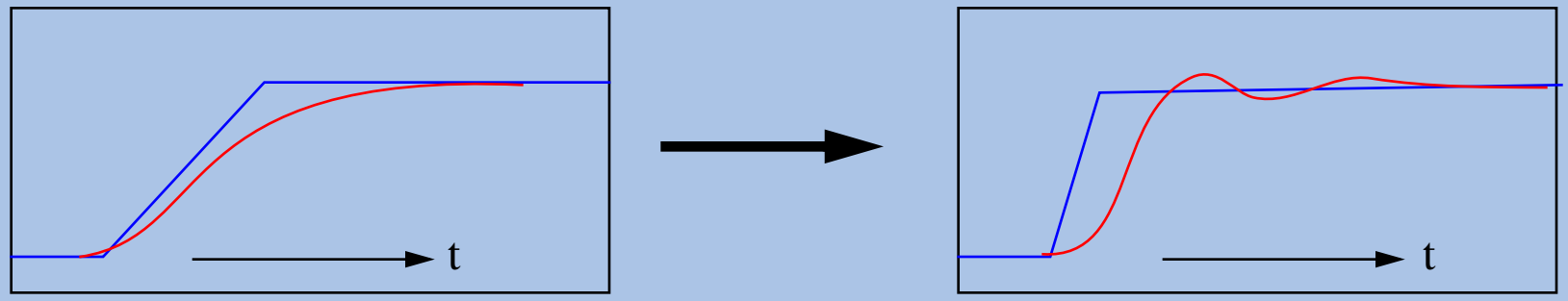
  - statistical simulation techniques

  - ACLV modeling techniques

  - modeling that includes the impact of layout density on models

Trends on chip: switching times

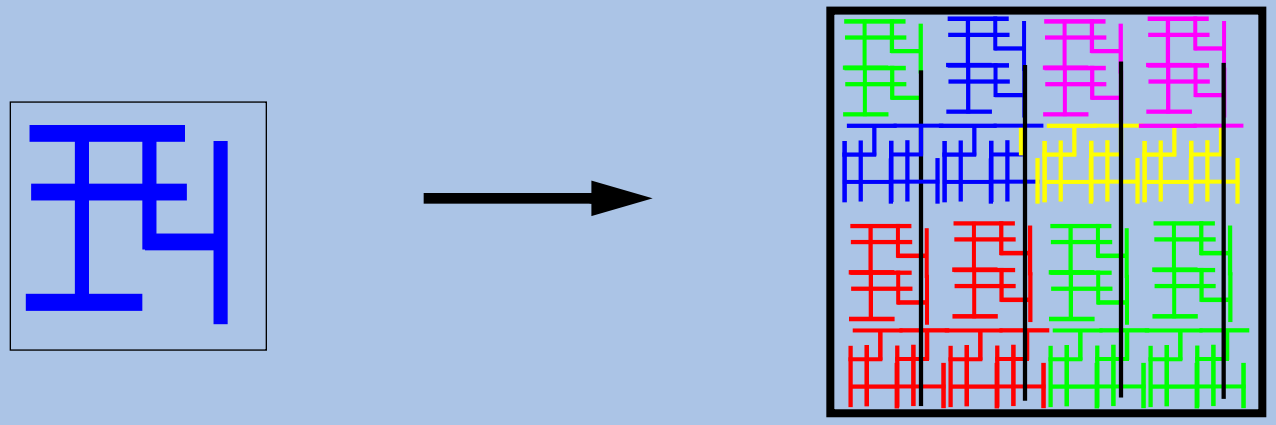
- Decreasing rise/fall times



- Results in:
  - increased impact of inductance in low resistance lines
  - ringing and signal bounce
  - larger, more accurate models (in numbers of nodes)
  - wrong estimate from RC delay used in design estimation and PD
- Need:
  - thinner, less inductive lines (more resistive)
  - proper matching at terminations
  - model reduction

Trends on chip: complexity/length

- Increasing numbers of interconnects, more complex geometries:



- Results in:
  - complex full chip modeling for verification
  - intensive CPU time and memory requirements
- Need:
  - algorithms that speed up extraction of large and dense geometries
  - algorithms that reduce the massive amounts of data necessary for analysis and design

Impact of improper interconnect

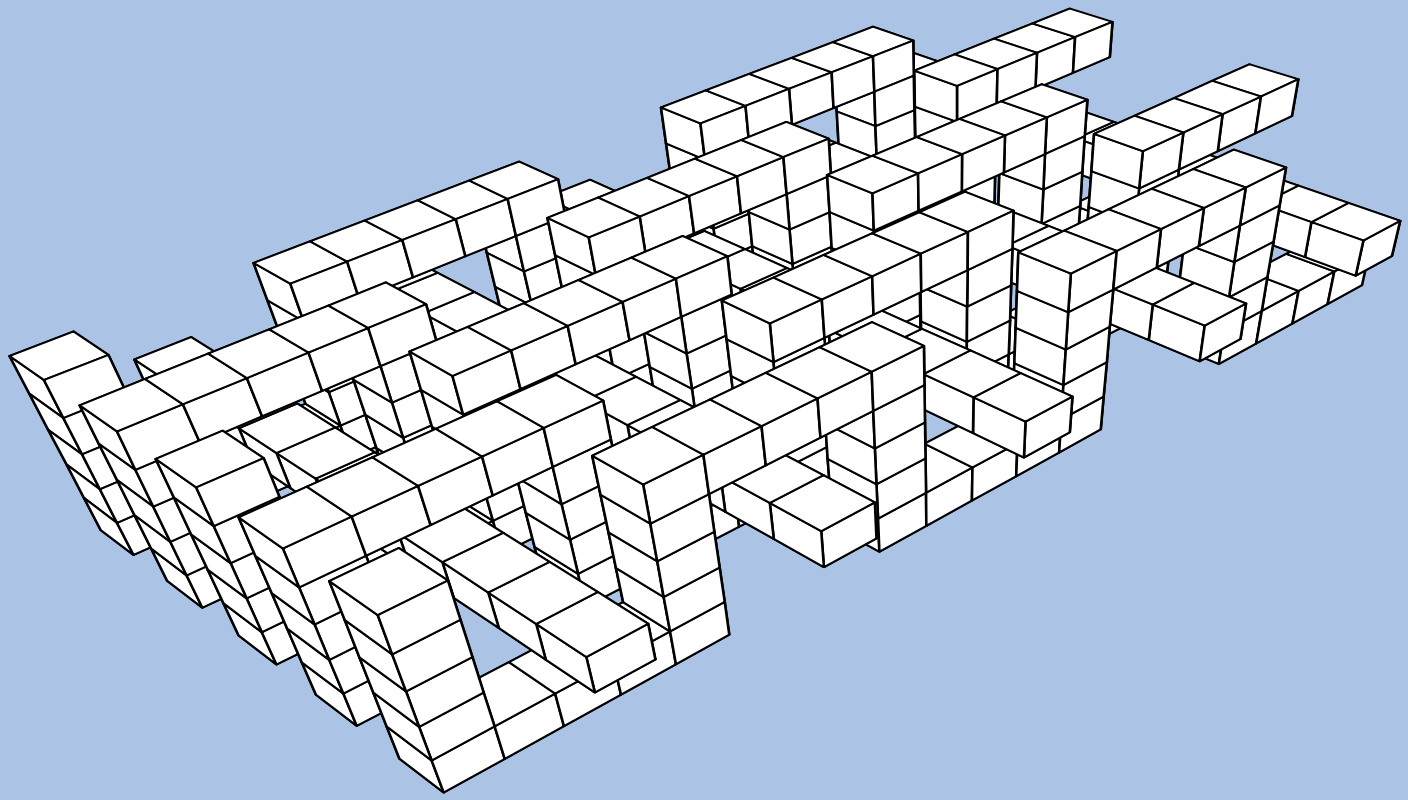
- Reduction of reliability (electromigration)
- Lowering of yield (sensitive to process)
- Loss of hierarchy and functionality (coupling, noise)
- Unpredictability -->more reliance on CAD analysis tools (time to market)
- Degradation of performance (timing)

- Delay of interconnect now on par with gate delay
- Gordon Moore - chairman of Intel - has begun to express doubts about the projections of Moore's law: logic density of silicon integrated circuits follows the curve (bits per square inch) = 2^((n - 1962)); that is, the amount of information storable in one square inch of silicon is roughly doubled yearly
- main reason: impact of interconnect
- Solutions must be a converging of CAD solutions and design solutions where CAD is not feasible

Design/CAD challenges

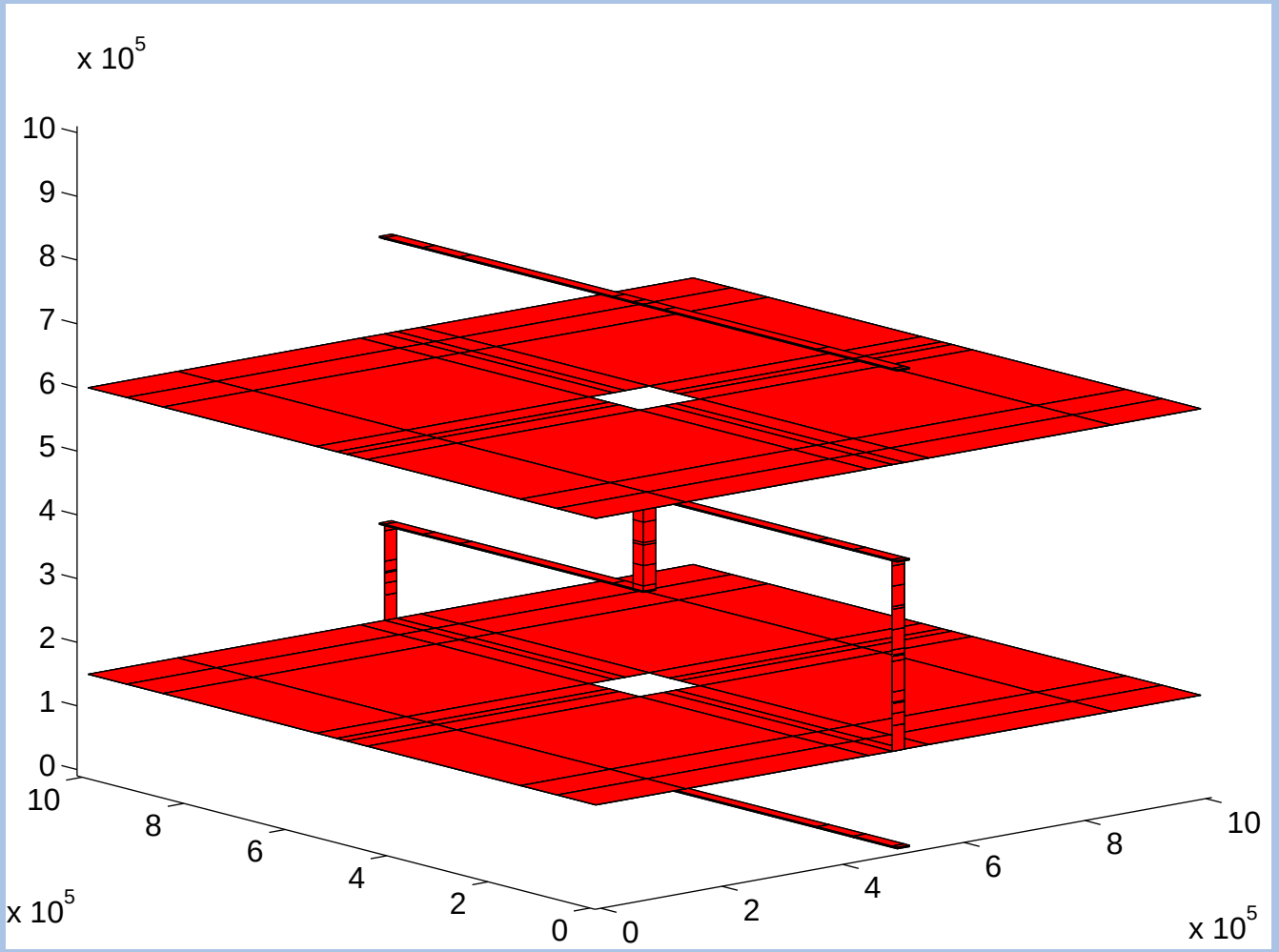
- Either:
  1. design and then analyze and correct (only if small number of iterations, minimal sections of the design need correction)
  2. design correctly by construction with analysis on the fly (need different levels of analysis with CPU/accuracy trade-off)
- Pessimistic approach is often the fastest but the costliest in terms of area, wire usage and circuit optimality (brute force designs are quick as are pessimistic analyses)
- Detailed approach consumes both CPU and design resources and compromise time to market but achieves compact designs with high yield and reliability.

Modeling challenges

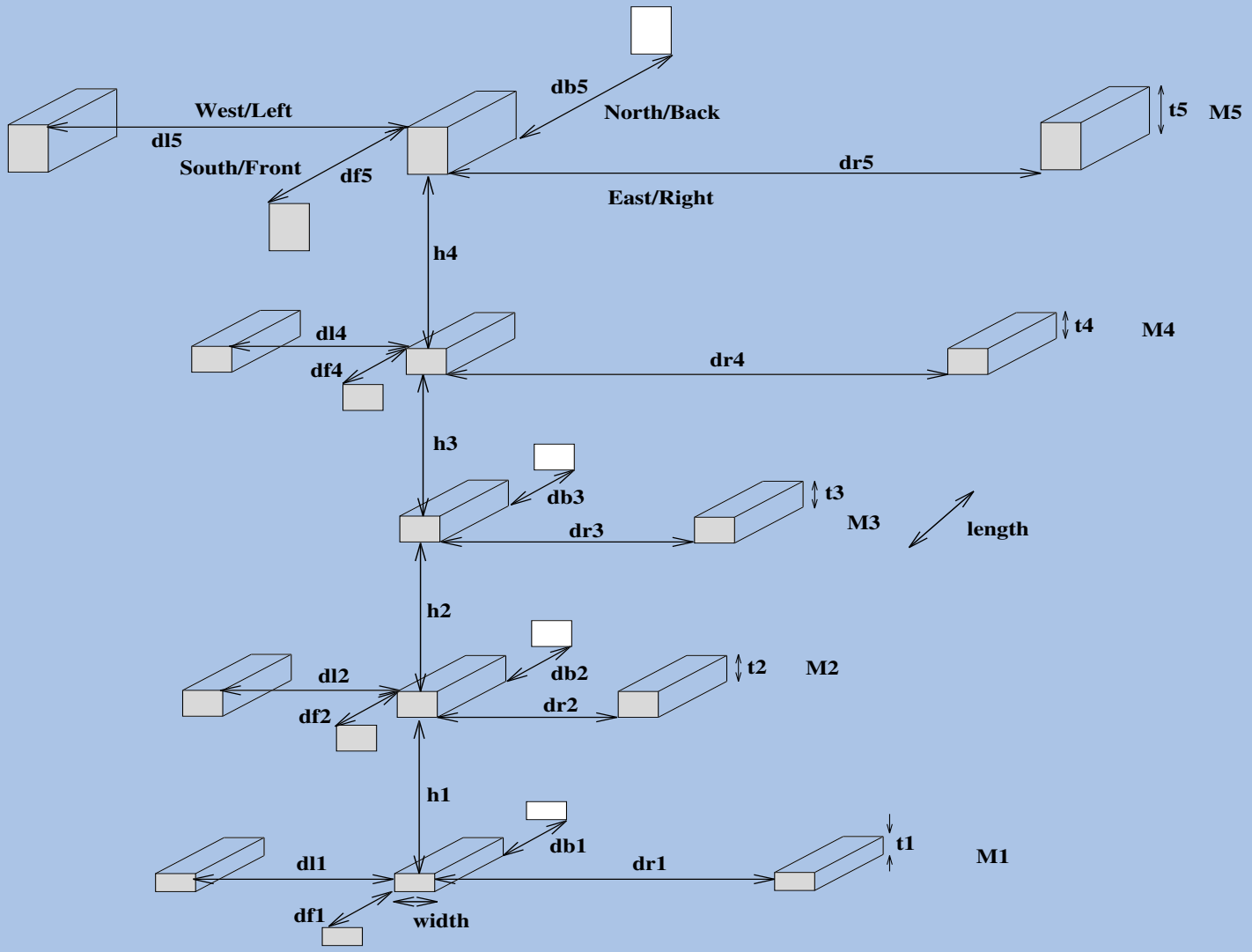


courtesy MIT

Modeling challenges

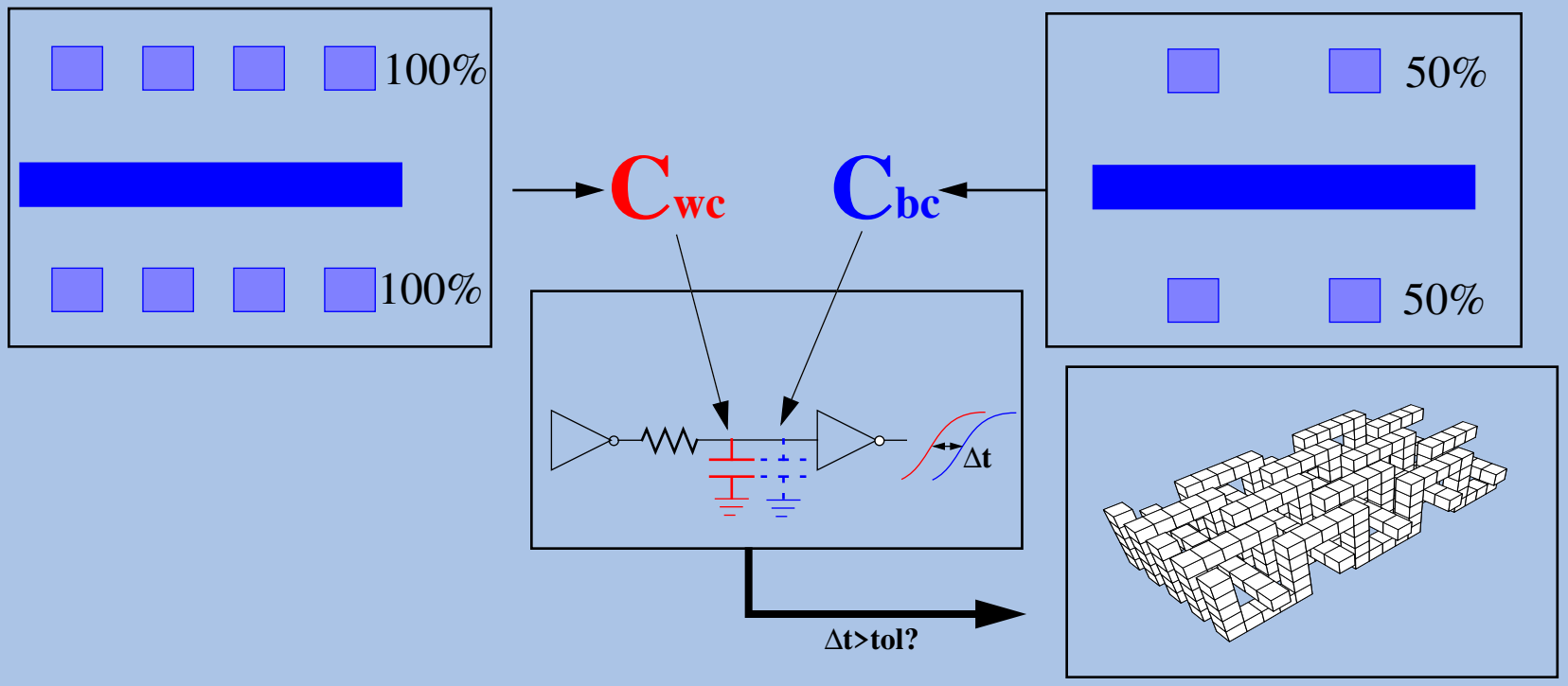


Modeling challenges



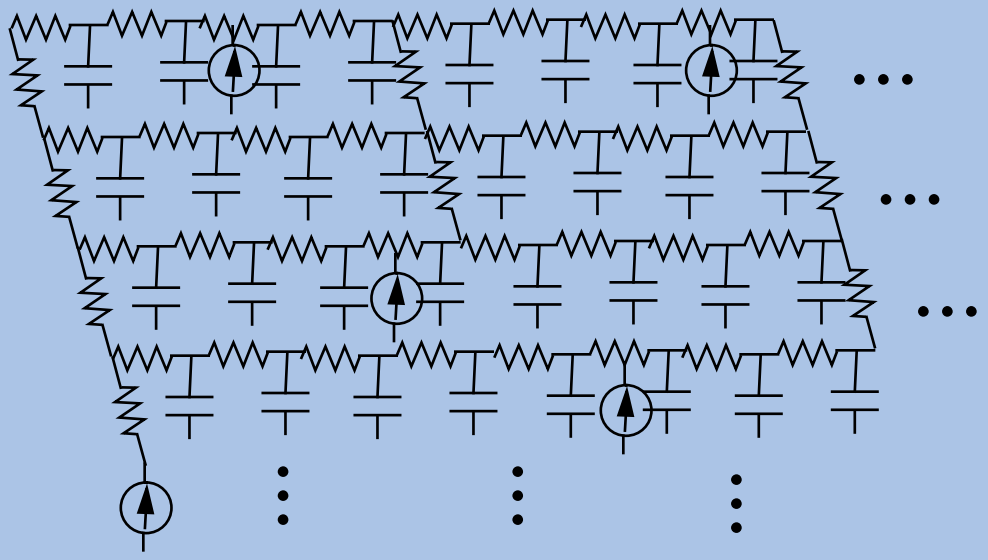
“Smart” extraction

- Calculate the capacitance estimates using 2D methods with worst case and best case scenarios, then estimate impact on timing variability
- If variation is high, extract net in 3D --> generally only 15% of nets



Simulation challenges: power analysis

- Macros feeding into large power grid.
- Mutiple independent sources (inputs) and one large linear network



Model reduction for power analysis

- Lumped linear state equations:

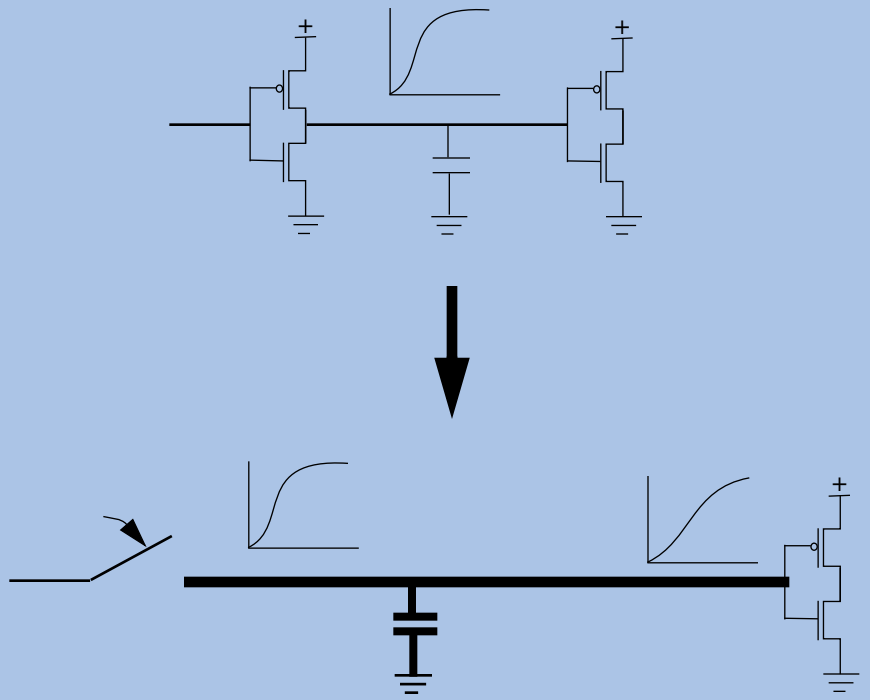
$$C\dot{x} = -Gx + Bu(s)$$

$$y = L^T x$$

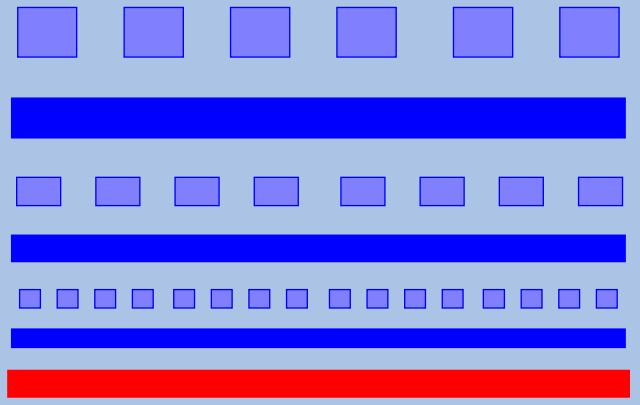
- where $u(s) = \begin{bmatrix} u_1 & 0 & u_2 & \dots & u_n \end{bmatrix}$ and $u_j = \Sigma \frac{\exp(-k_i s)}{s^2}$
- Transfer function: $H(s) = L^T (G + sC)^{-1} B(s)$
- Taylor series: $H(s) = H_0 + H_1 s + H_2 s^2 + \dots$
- Taylor series of model $\hat{H}(s) = H_0 + H_1 s + H_2 s^2 + \dots + H_{2q-1} s^{2q-1}$
- $\hat{H}(s) = \sum \frac{k_i}{s - p_i}$ --> one decomposition but multiple pade approximations!

Simulation challenges: dynamic models

- Model evaluation is the most CPU intensive process in circuit simulation (>80%)
- In interconnect dominated environment must reduce the sophistication of device model when possible.
- Make models less/more accurate dependent on load.

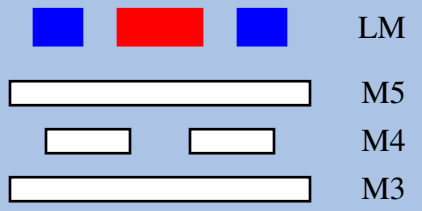


Technology design: hierarchical wiring

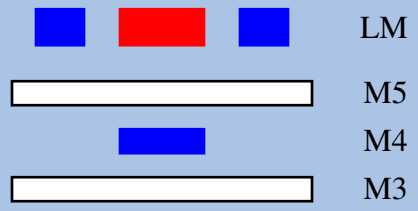


- capacitance is a function of geometry, not of size
- resistance is a function of size
- RC is lowered upwards in the levels
- local interconnections on lower level, global on higher
- skin effect at higher level? resistance will increase further.

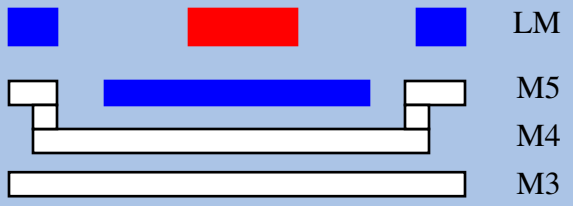
Evolution of Backend Design



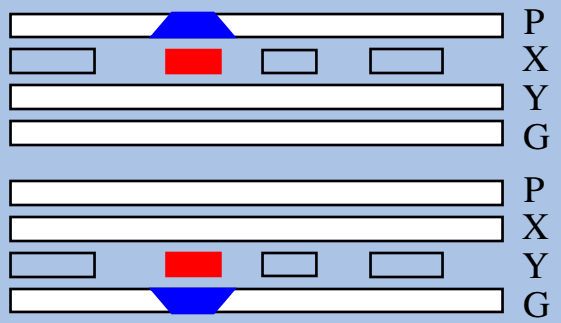
1993 (~100 MHz)



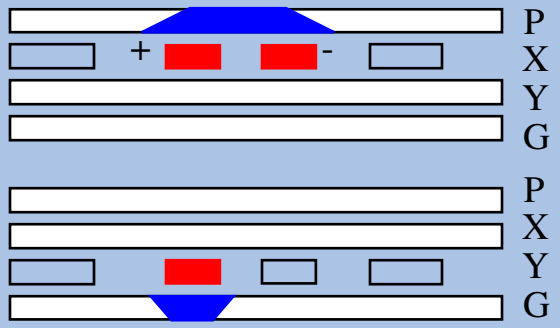
1994 (~200 MHz)



1995 (~400 MHz)



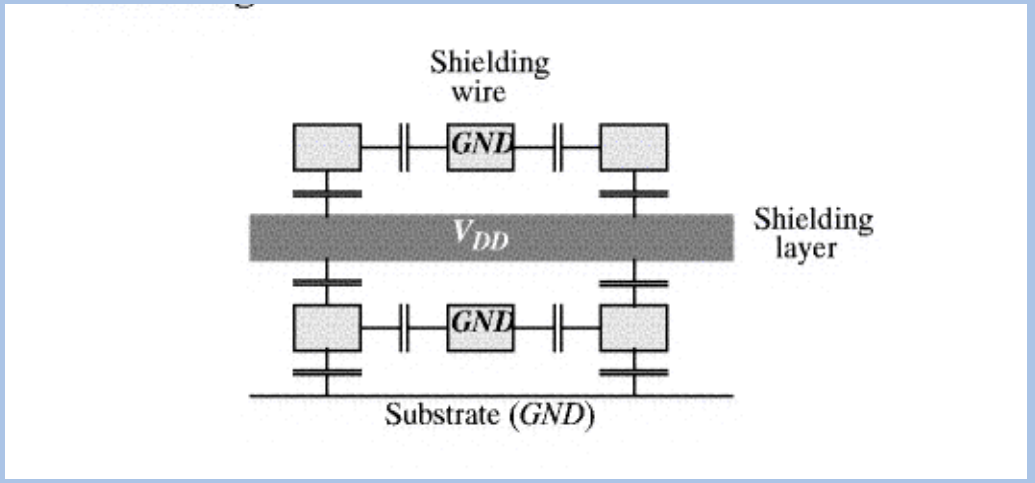
1996 (~600 MHz)



2000 (~1 GHz)

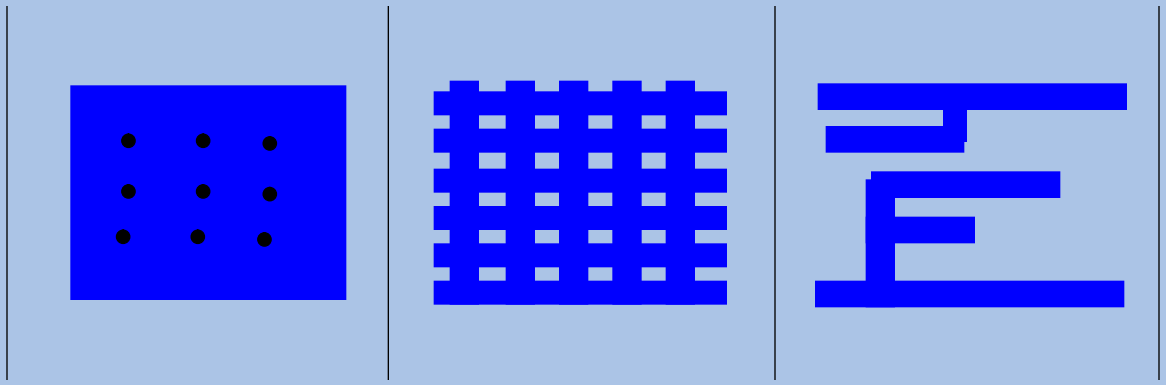
Design Isolation with Metal Planes

- DEC first to begin using them on chip
- Are becoming generally acceptable solution because
  - many metal layers available
  - solves coupling problem quickly without too much complicated analysis
  - if power sources lie at periphery of chip planes allow for small IR drop
- Drawbacks
  - must punch through plane for inter-layer routing
  - removes an entire wiring plane which increases area usage, die size
  - needs special wiring level thickness



Courtesy Rabaey (Berkeley)

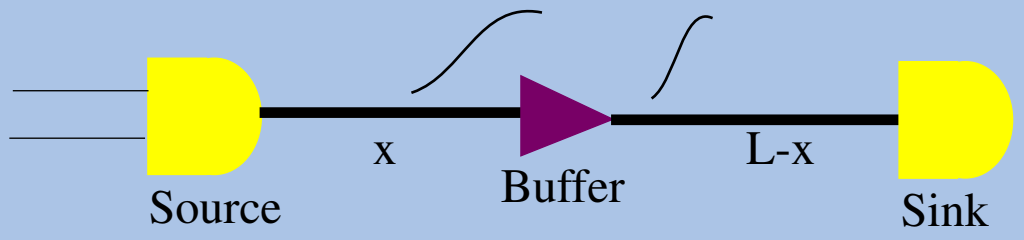
Power Supply System Design



Issue	Power Planes	Power Grids	Routed System
Resistive (IR) Drops	Very low	Low	High
Inductive Drops	Low - best for high performance	High	Very high
Electromigration	Very low	Low, few hot spots	High probability
Floor Plan	May require additional metal level and increases process cost	Very restrictive - but you know it in advance and design around it	Everything looks great until just before design tape-out
Popularity	Will be common for gigahertz designs	Most popular	Only for low-cost sys-tems

Design with Buffer Insertion

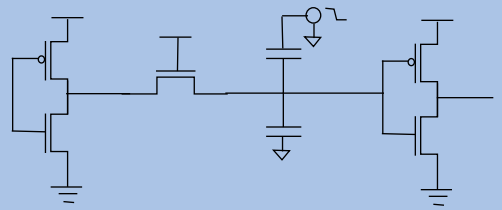
- Interconnect delay quadratic in wire length (RC)
- reduces delay to linear, eliminates long wires (noise)
- e.g. placement of one buffer



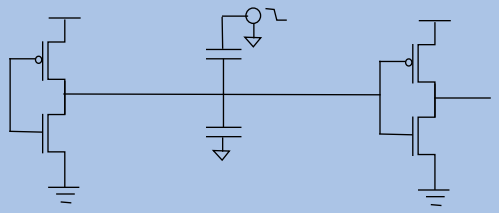
- D_0 = delay without buffer
- D_1 = optimum delay with buffer
- Let $\frac{\partial D_1}{\partial x} = 0$ which gives $x = \frac{L}{2} + \frac{R_b - R_{so}}{2R} + \frac{C_{si} - C_b}{2C}$
- But must be combined with physical design for appropriate placement optimization
- Buffer insertion to reduce noise, inductance,
- Buffer insertion for multi-driver, multi-drop nets

Noise challenges

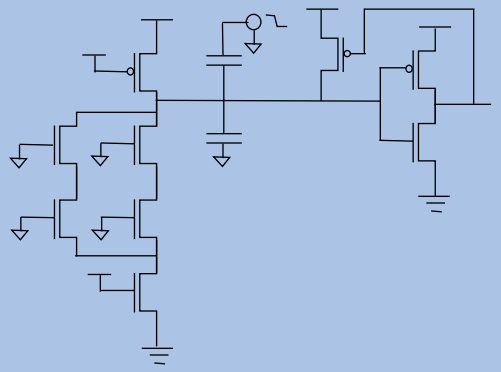
A.

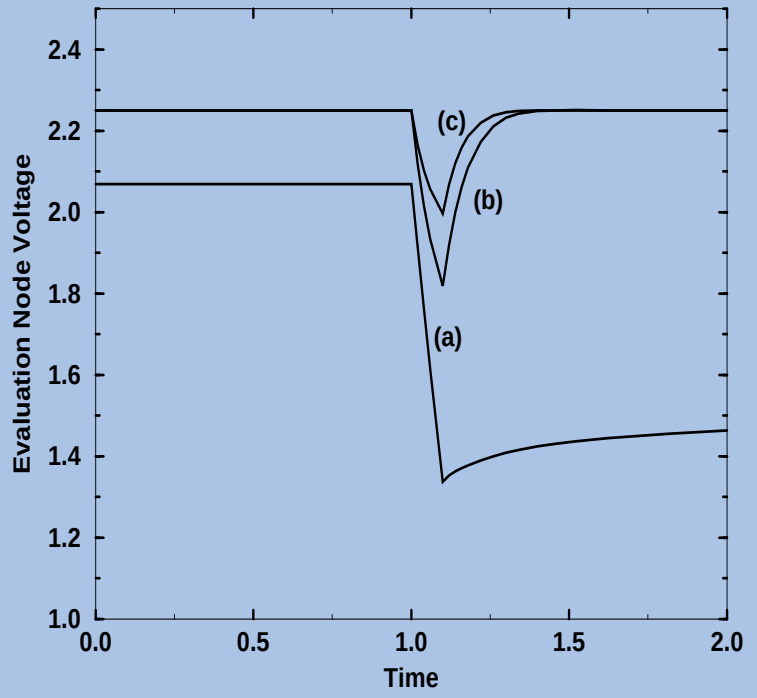


B.

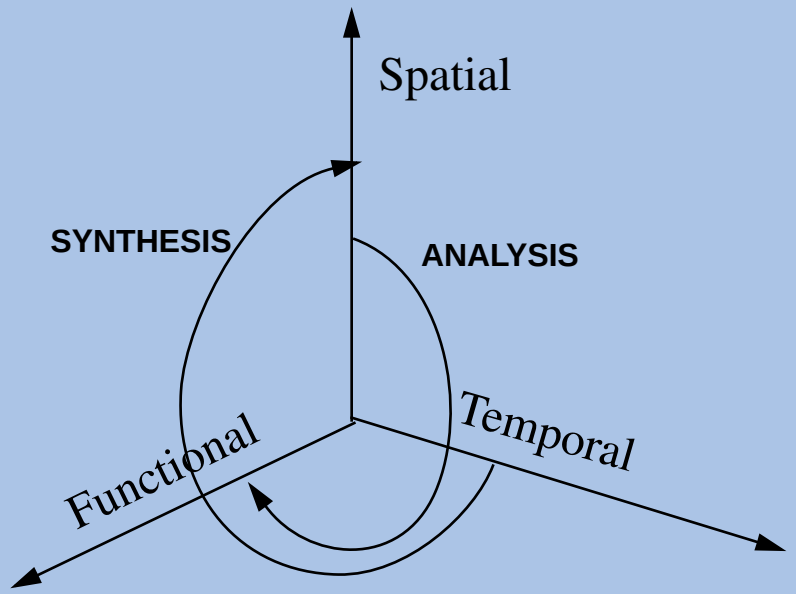


C.



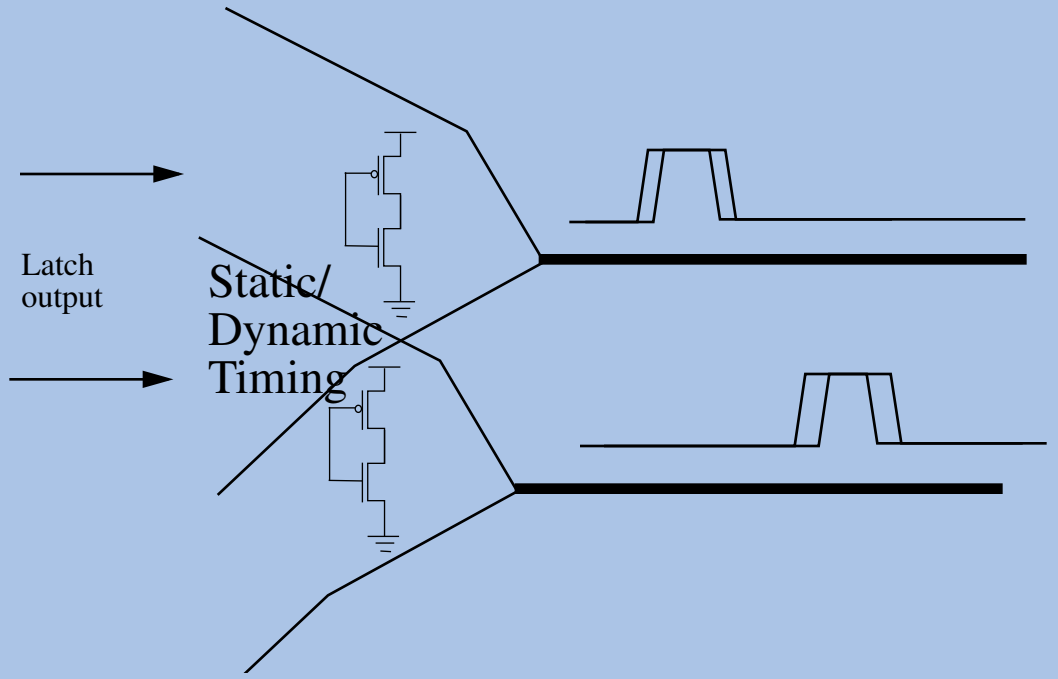


Signal interaction space



Timing orthogonality

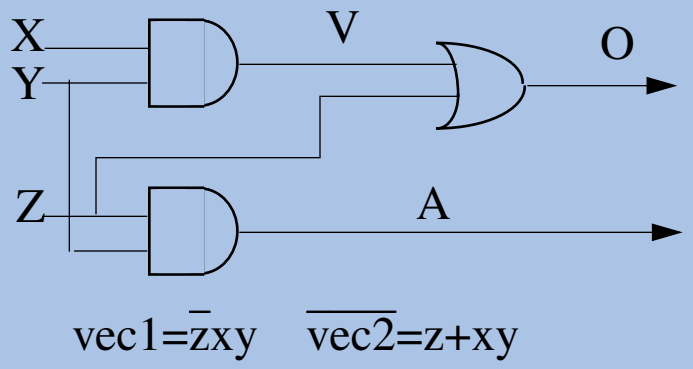
- Timing information obtained from dynamic or static simulation of devices + interconn.
- Design is aimed for signals to be as timing orthogonal as possible
- If orthogonality, no need for expensive extraction/analysis



- In synthesis can route wires close to each other in physical design
- can use wires as isolation
- accurate delays depend on accuracy of timing and on interconnect models (esp. pre-PD)

Logic orthogonality

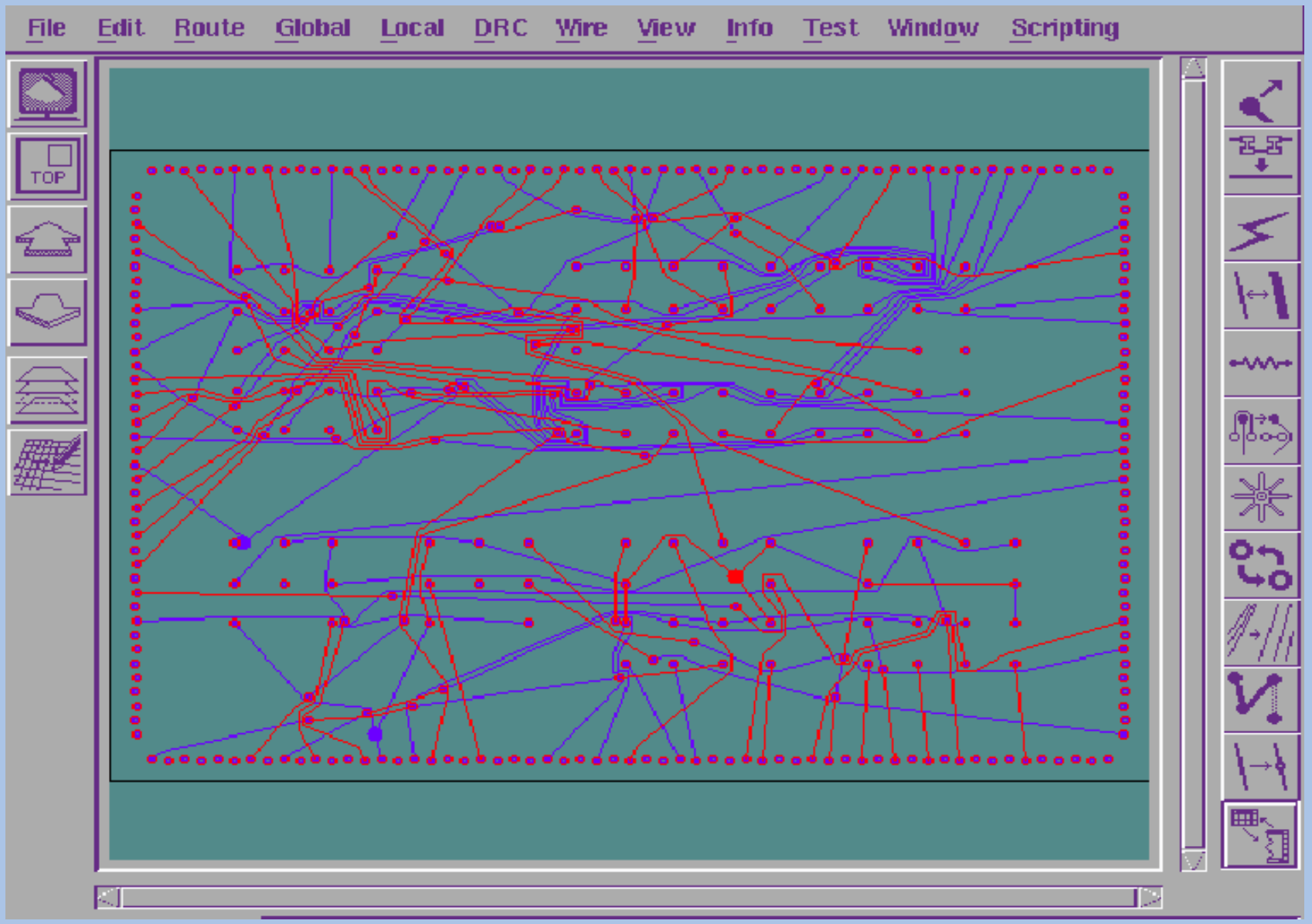
- possible to predict signal interaction by logic functionality alone

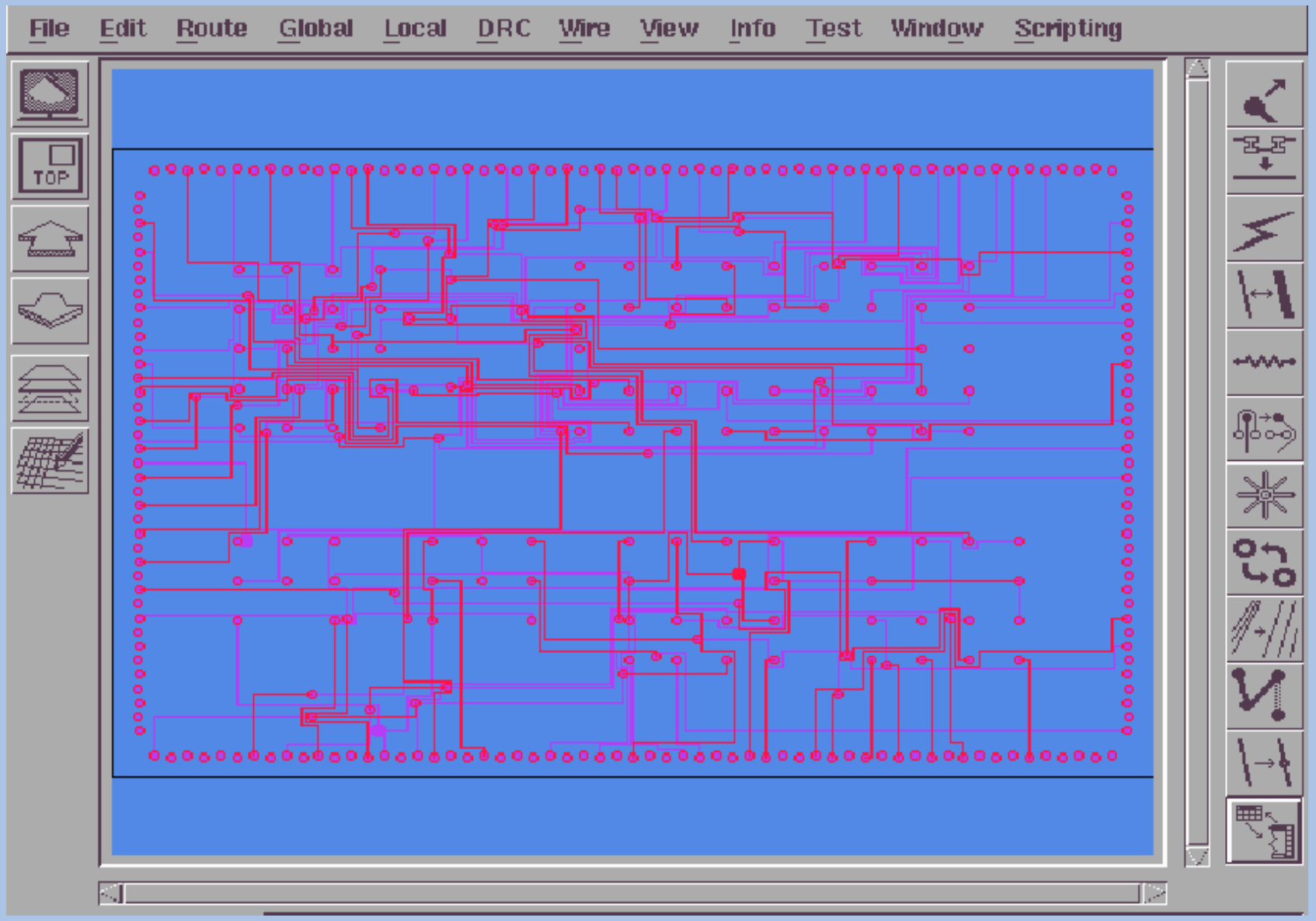


- a pair of input vectors (vec1, vec2) for primary inputs which cause signal A to rise while V falls
- However this interaction is unobservable at the output
- Timing information -> glitches
- Formal mathematical analysis of net coupling matrix which can be used for synthesis or analysis

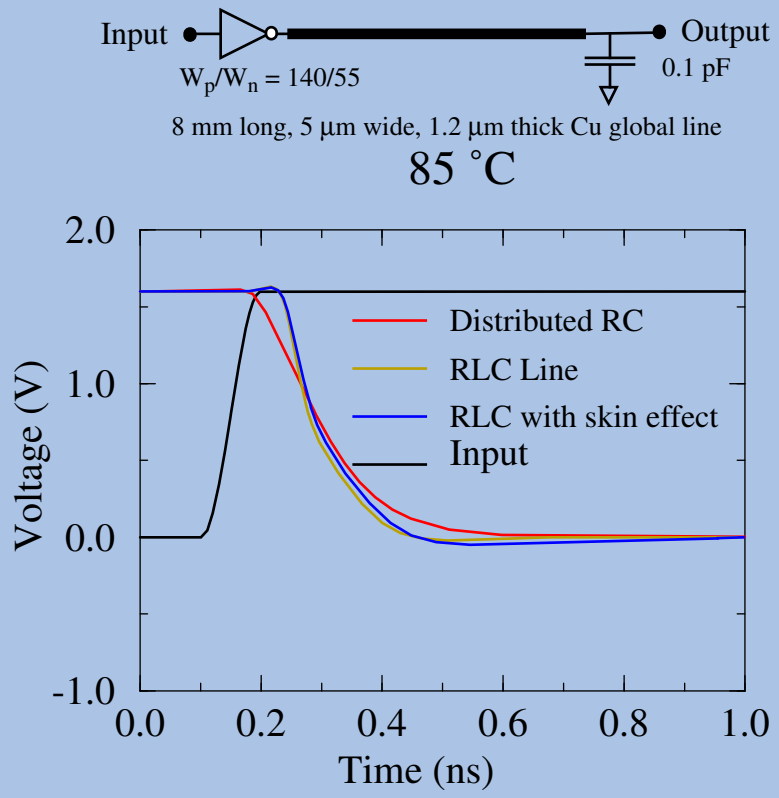
Noise aware routing

- Experimental work going on with router from UC Santa Cruz and IBM
- Global routing (congestion, noise density), local rubber band routing (proximity, length, wiring level)
- Ultimate tool would combine variable accuracy extraction with timing orthogonality, logic orthogonality, device noise sensitivity rules of thumb, and flexible variable width routing algorithms to control resistive loss, decrease noise and decrease RC delay while controlling inductance.
- trade-off in delay (length, resistance, capacitance), noise (proximity, length, devices, shielding), electromigration (thickness, device drivers), routability (density, extra wires due to shielding), inductance (length and width), area usage (forced distance apart, extra width).
- Need different noise metrics at different points in the router


File
Edit
Route
Global
Local
DRC
Wire
View
Info
Test
Window
Scripting
TOP


File
Edit
Route
Global
Local
DRC
Wire
View
Info
Test
Window
Scripting
TOP

Exploiting Inductances



Inductive effects permit significantly higher signal risetimes

- Can increase inter-macro communication and processor-memory data bus bandwidths by at least 50%
- Faster edges also result in slightly (5-10%) smaller interconnection delays and latch delays
- Effects are more pronounced at room temperature
- Smaller rise times also impact cross talk significantly

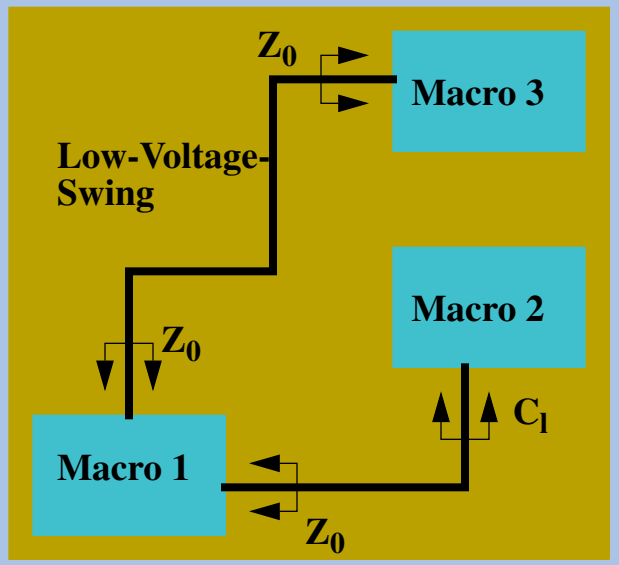
Inductive effects change the delay metrics in physical design

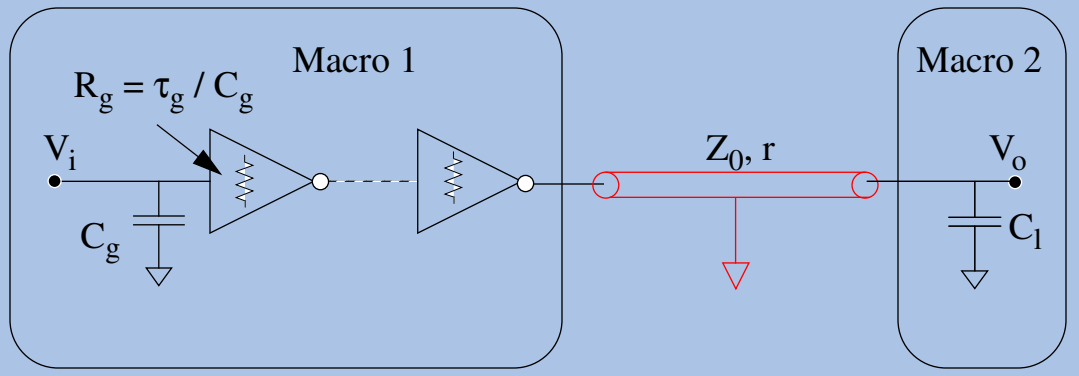
- Time-of-flight effects emphasizes linear delay metrics for interconnection delays as a function of line lengths

Transmission line design

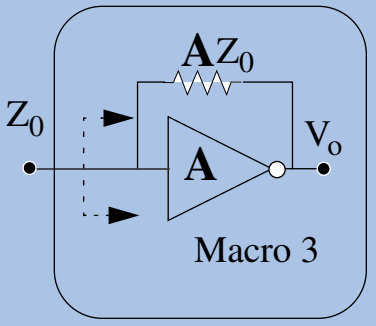
- Provide a controlled ground return for most global lines such that RLC per unit length is predictable
- No need for inductance extraction which normally requires large neighborhood and therefore CPU resource
- Ground return controls coupling to neighboring lines
- Make as lossless as possible to avoid RC delays so only time of flight
- Long experience in microwave design rules can be transformed to VLSI
- Will require new simulation algorithms for large numbers of transmission lines and large number of devices (currently timing simulation for devices only)

Structured Module Integration

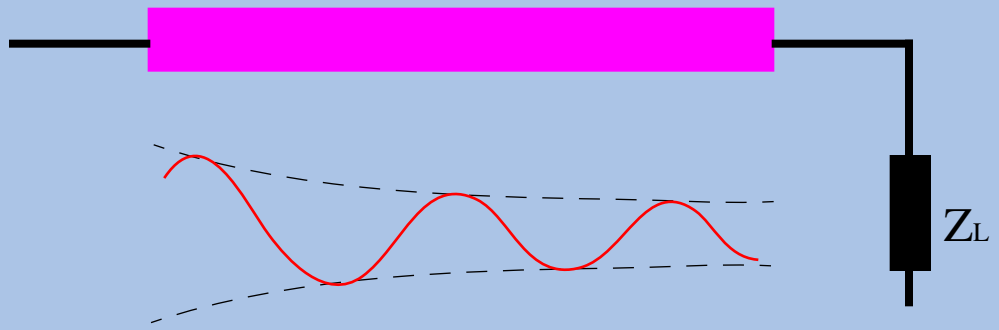




$\tau_B \sim e\tau_g \ln(R_g/Z_0)$ $\tau_L \sim \tau_0(r, Z_0, l)$ $\tau_R \sim Z_0 C_l$

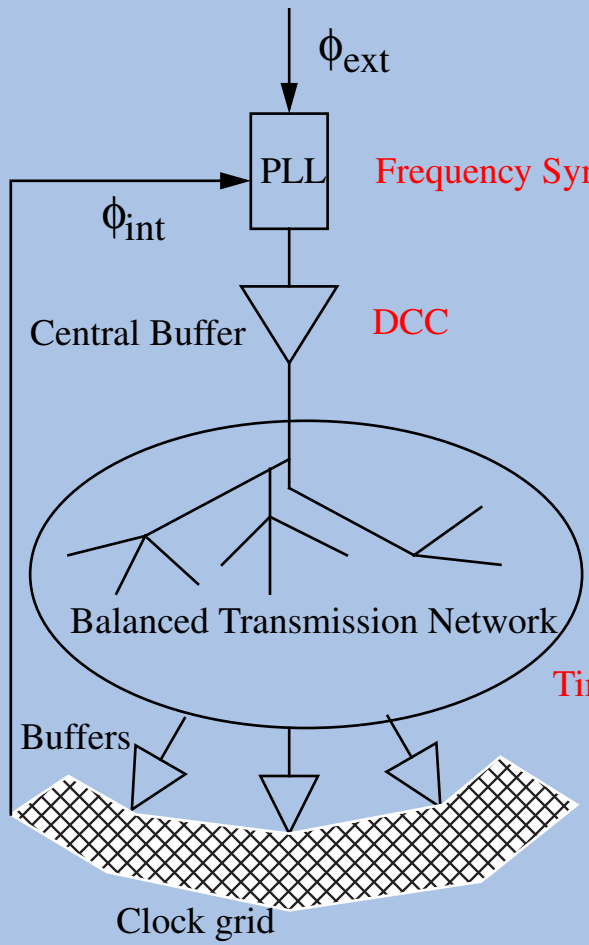


Distortion-free lines



- $\gamma = \sqrt{(R+sL)(G+sC)}$ $Z_0 = \sqrt{\frac{R+sL}{G+sC}}$
- if $\frac{R}{L} = \frac{G}{C} = k$ then $\gamma = \frac{s+k}{c}$ and $Z_0 = \sqrt{\frac{L}{C}}$
- if $Z_L = Z_0$ then $I_0 = I(s)e^{\left(\frac{-sl}{c}\right)}e^{\left(\frac{-kl}{c}\right)}$
- The current is attentuated by a factor that is a function of length only but there is no other distortion

Global Clock Distribution

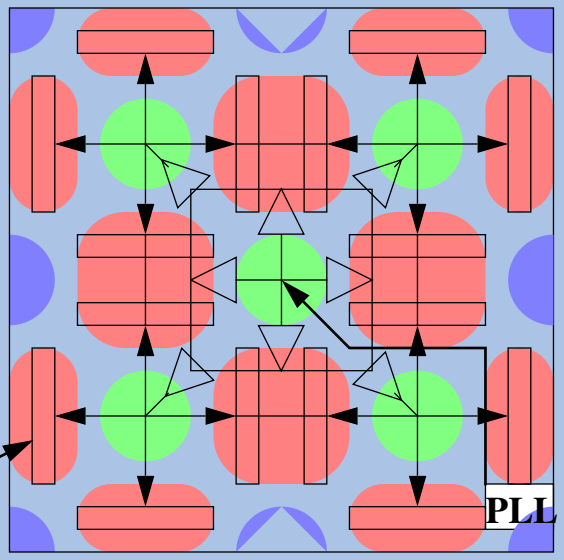


Transmission line networks are essential for reducing global skew

- Time-of-flight delay per unit length ($\tau \approx \sqrt{\varepsilon_r}/c$) where c is the velocity of light in vacuum
- Time-of-flight delays are independent of variations in interconnection geometries due to fabrication processes
- Future designs will emphasize planar clock routing techniques and optimal wire sizing for transmission lines

Temperature Issues in Clock Interconnect

- Power dissipation in clock buffers affect local skews
- Local buffer temperature $\Delta T \propto V_{dd} I_{dsat} R_{th} = V_{dd}^2 (C_l / \tau_r) R_{th}$ depends on local power dissipation, thermal resistance to substrate sink, and the local rise times
- Temperature variations affect threshold voltages (4 mV/˚C) and interconnection resistances (0.4%/˚C)
- Unoptimized designs have temperature variations of 5-10 ˚C
- Distribute buffers to attain uniform die temperature, while minimizing global clock distribution skew



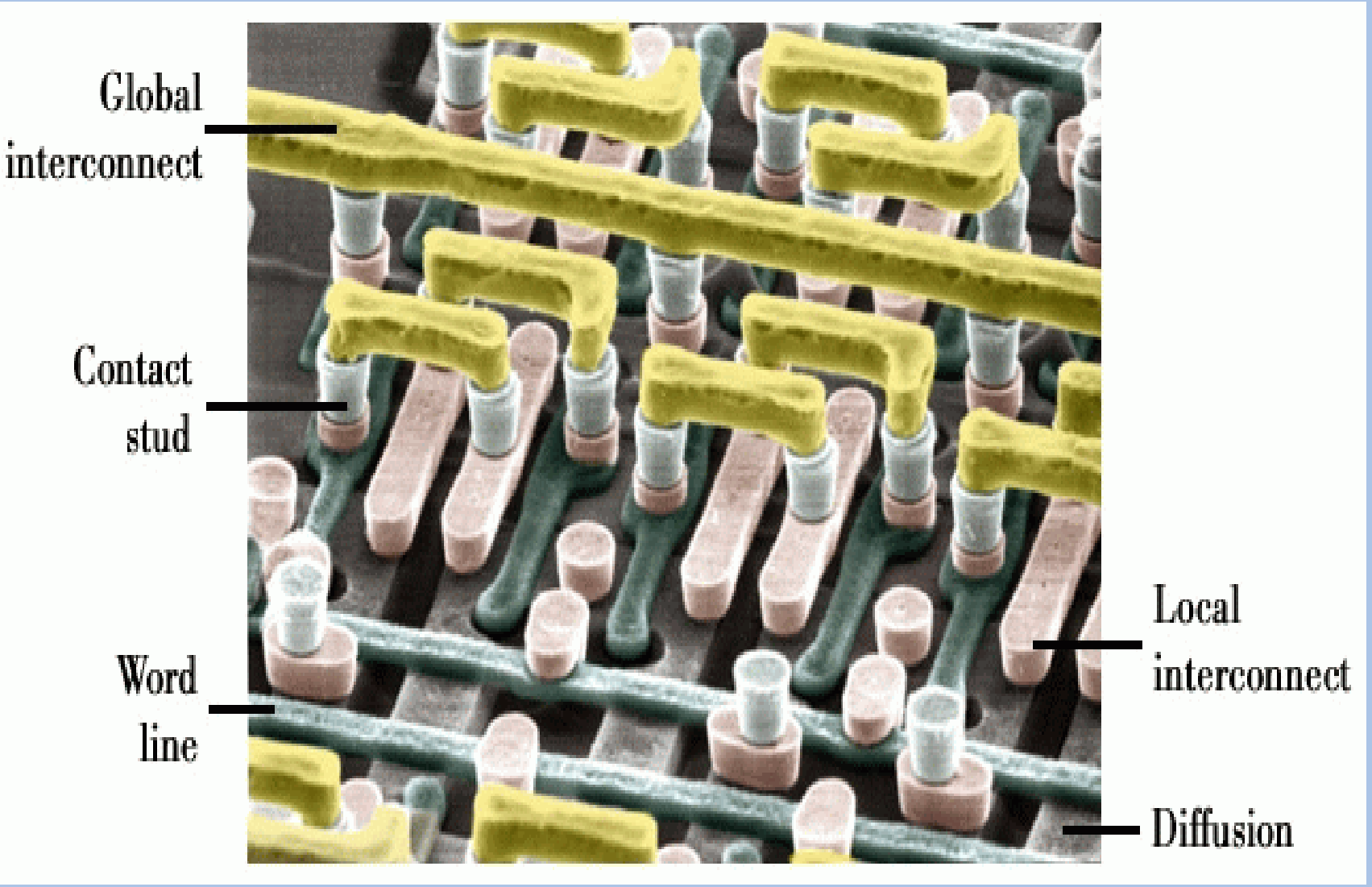
600 MHz Alpha

Interconnect/Dielectric material

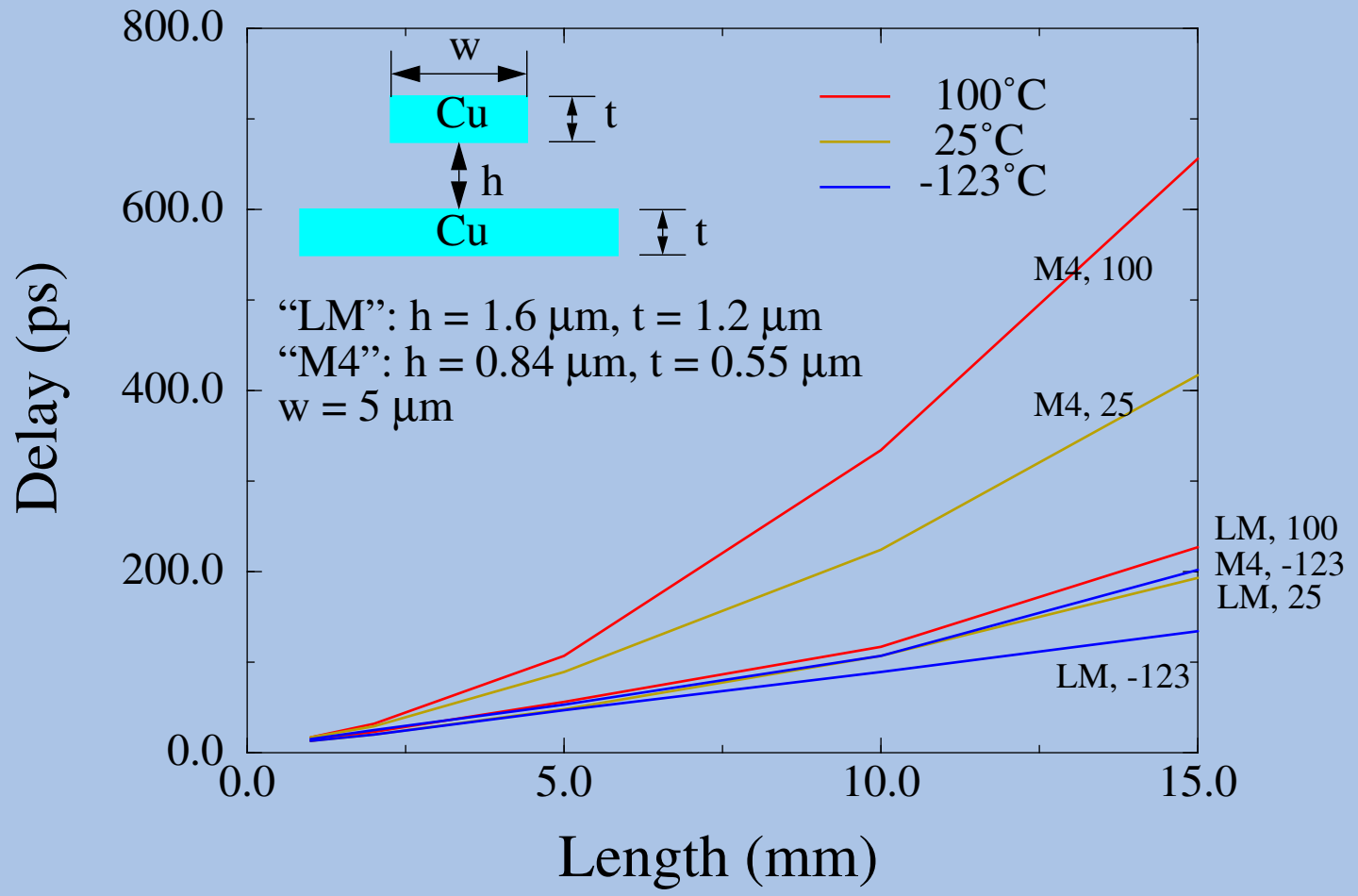
- capacitance a function of geometry dielectric constant $C = \varepsilon_{ox}\frac{W}{Lt_{ox}}$
- resistance a function of geometry and resistivity of metal: $R = \frac{\rho l}{A}$
- Aerogel dielectrics which incorporate microscopic air bubbles (ε_r=3.9->2.0? would half the capacitive effects, speed up RC delay, speed up time of flight propagation by $\sqrt{2}$)
- Aluminum is being replaced by copper (2.417 ohm-m vs. 1.543), the trade-off can be either resistive or capacitive (lines widened)

resistivity (ohm-m) of different metals

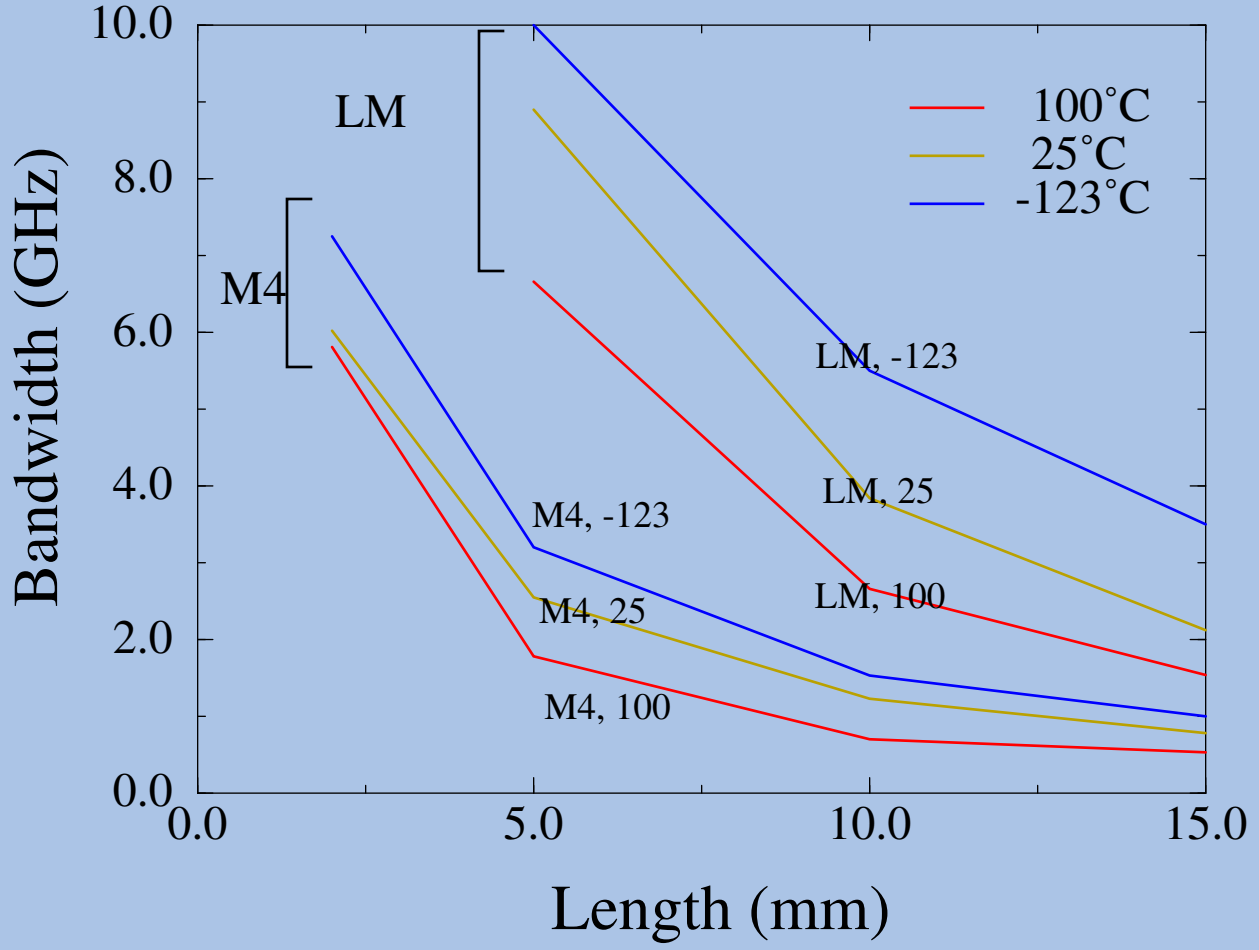
T (Kelvin)	Al	Cu	Ag
100	0.442	0.348	0.418
200	1.587	1.046	1.029
273	2.417	1.543	1.467
293	2.650	1.678	1.587


Global interconnect
Contact stud
Word line
Local interconnect
Diffusion

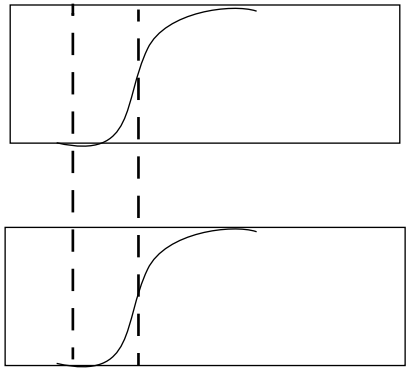
Low temp: interconnection delays



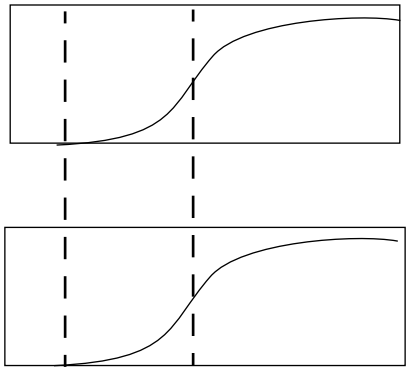
Low temp: interconnection bandwidth



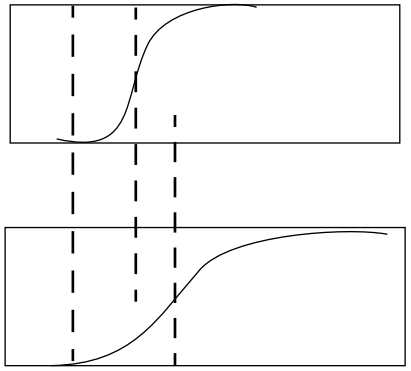
Speed of light effects



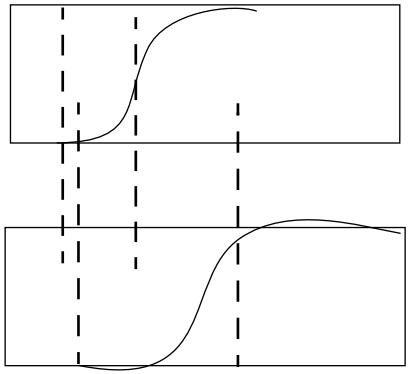
Short (no delay)



C (no delay)

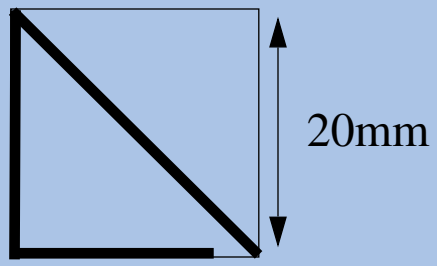


RC (delay)



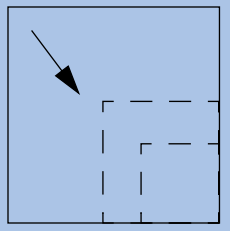
RLC (delay)

Computational area limitations



- Assume large chip size is 20 mm (P2SC->18mm, growing)
- Best case signal to traverse is 28 mm, worst case is 40 mm
- Best case lossless propagation 186 ps, worst case 267 ps
- Implies best case clock rate is 5.3 GHz, worst case is 3.8 GHz
- However, this is without the attenuation delay component and doing no computational work between the latches (RC=140ps/cm, bc=392,wc=560ps -> bc clock=3GHz, wc=1.2GHz! RC+10x5ps. gates, bc 2.6GHz, wc 1.2 GHz)

- the opposite way of looking at it is to suppose a clock speed and calculate the area that will be reached by that clock within a cycle:



- e.g. 10 GHz clock -> interconnect= 2.5 mm, BC=3.4mm, WC=2.5mm
- This clock will not be able to reach even a quarter of the chip!