An Interconnect-Centric Design Flow for Nanometer Technologies

Jason Cong

UCLA Computer Science Department

Email: cong@cs.ucla.edu

Tel: 310-206-2775

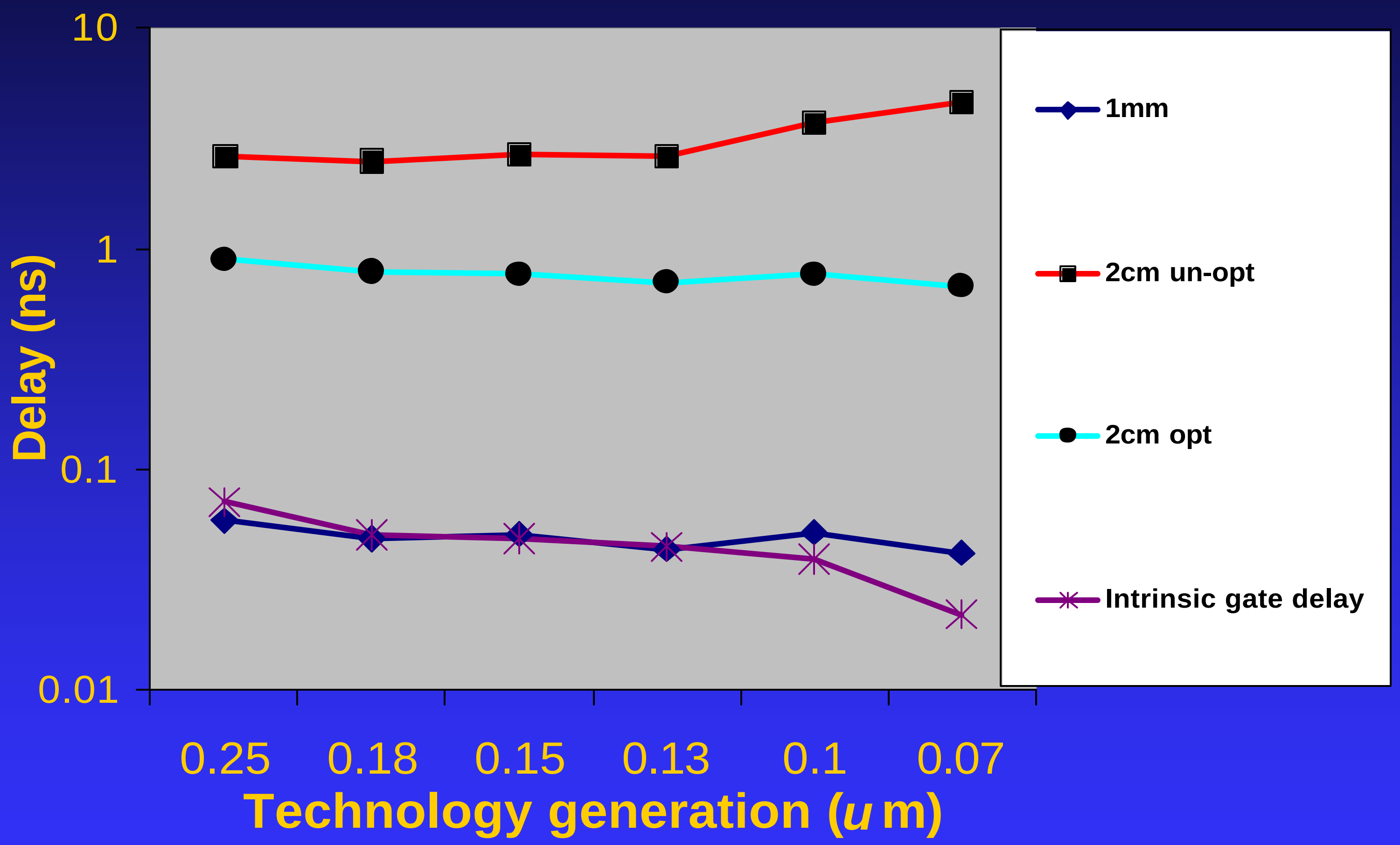
URL: http://cadlab.cs.ucla.edu/~cong

Exponential Device Scaling

- **Moore's Law**
  - **The min. transistor feature size decreases by 0.7X every three years (Electronics Magazine, Vol. 38, April 1965)**
  - **True in the past 30 years!**
- **National Technology Roadmap for Semiconductors (NTRS'97)**

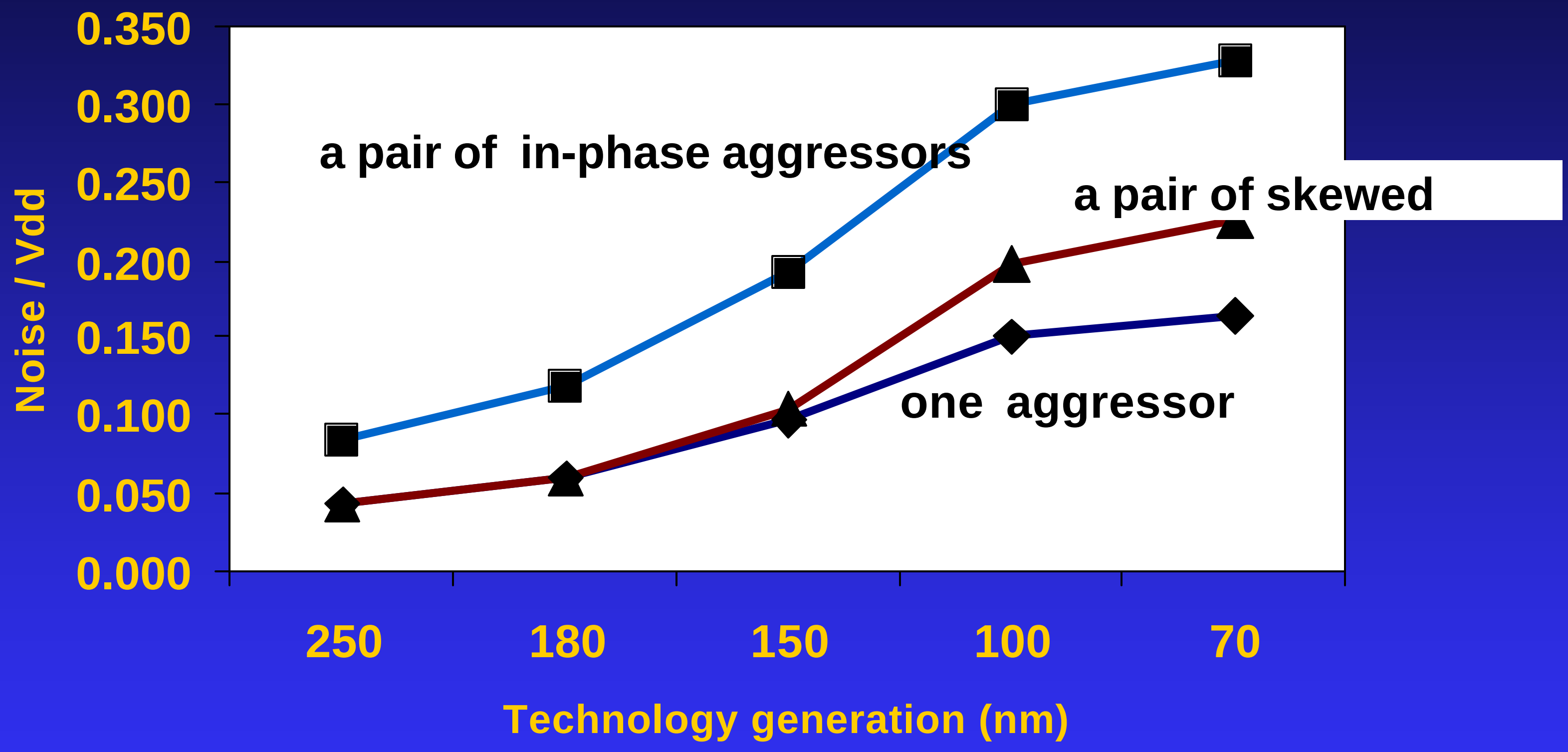
Technology (um)	***0.25***	***0.18***	***0.15***	***0.13***	***0.10***	***0.07***
Year	1997	1999	2001	2003	2006	2009
# transistors	11M	21M	40M	76M	200M	520M
On-Chip Clock (MHz)	750	1200	1400	1600	2000	2500
Area (mm^2)	300	340	385	430	520	620
Wiring Levels	6	6-7	7	7	7-8	8-9

Global/Local Interconnect Delays vs. Gate Delays



Optimization is obtained buffer insertion/sizing and wire sizing

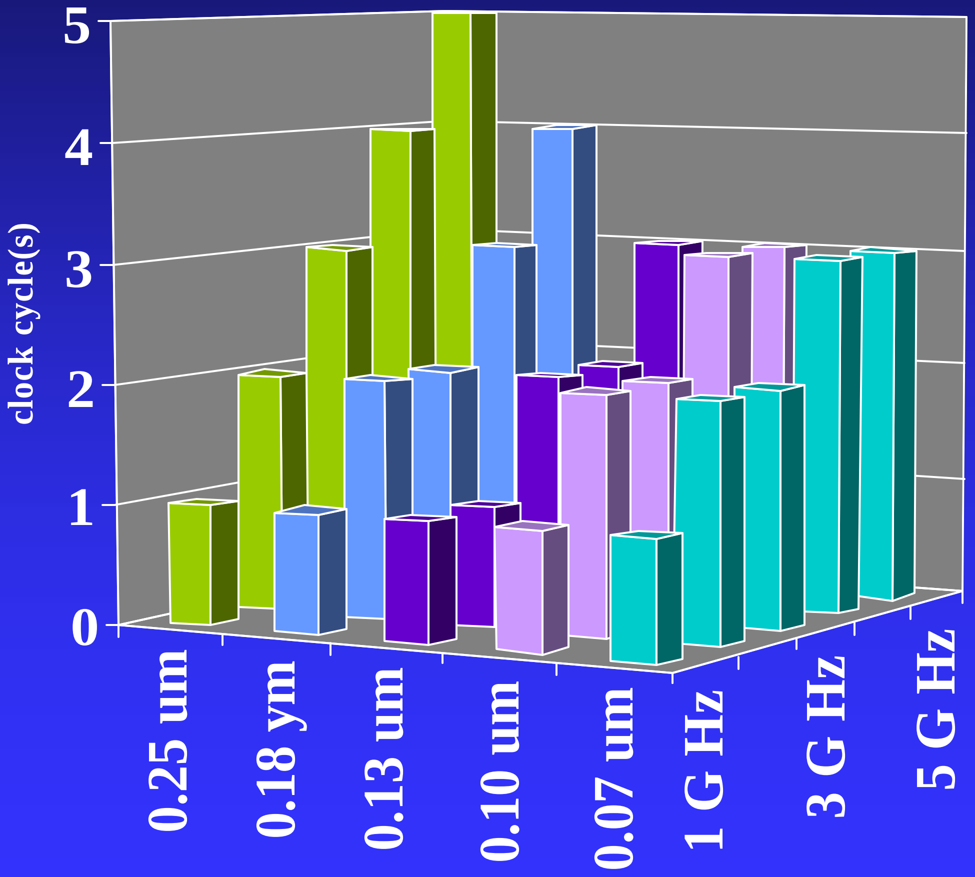
Coupling Noise



Coupling noise from two adjacent aggressors to the middle victim wire of 1mm with 2x min. spacing. Rise time is 10% of project clock period.

- **Coupling noise depends strongly on both spatial and temporal relations!**

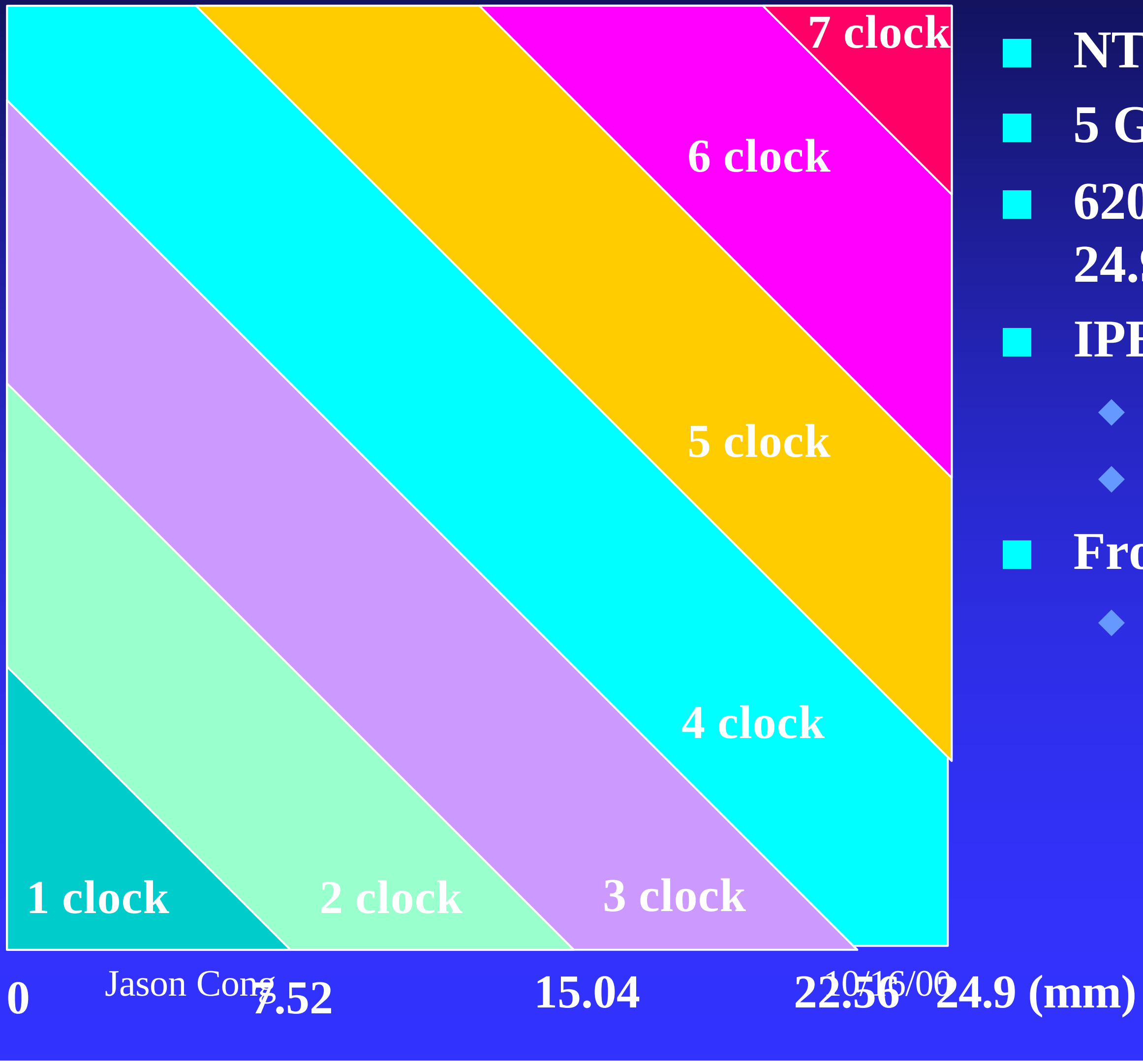
Clock cycles required for traveling 2cm line under BIWS (buffer insertion and wire sizing)



Estimated by IPEM
On NTRS'97 technology

Driver size: 100x min gate
Receiver size: 100x min gate
Buffer size: 100x min gate

How Far Can We Go in Each Clock Cycle



- NTRS’97 0.07um Tech
- 5 G Hz across-chip clock
- 620 mm^2 (24.9mm x 24.9mm)
- IPEM BIWS estimations
    - Buffer size: 100x
    - Driver/receiver size: 100x
- From corner to corner:
    - 7 clock cycles

Two Important Implications

- Interconnects determine the system performance

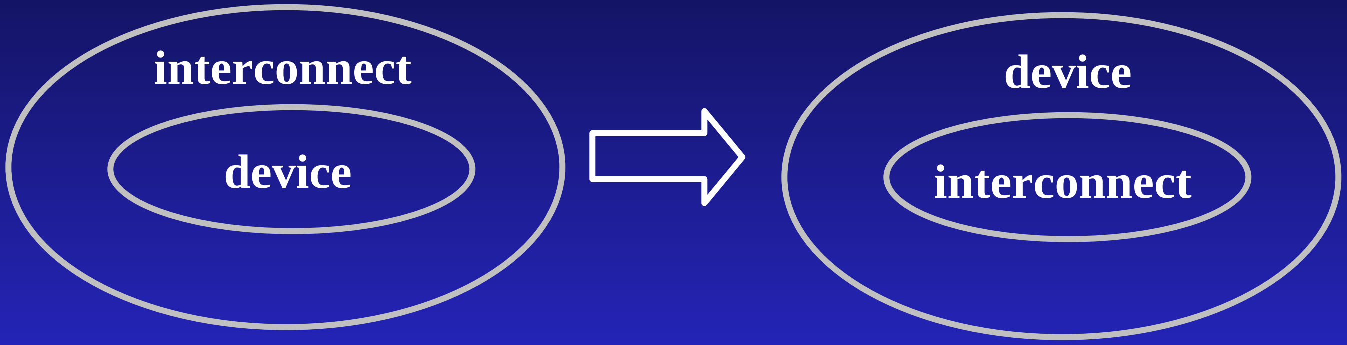
  Interconnect/communication-centric design methodology

- Need multiple clock cycles to cross the global interconnects in giga-hertz designs

  Pipelining/retiming on global interconnects

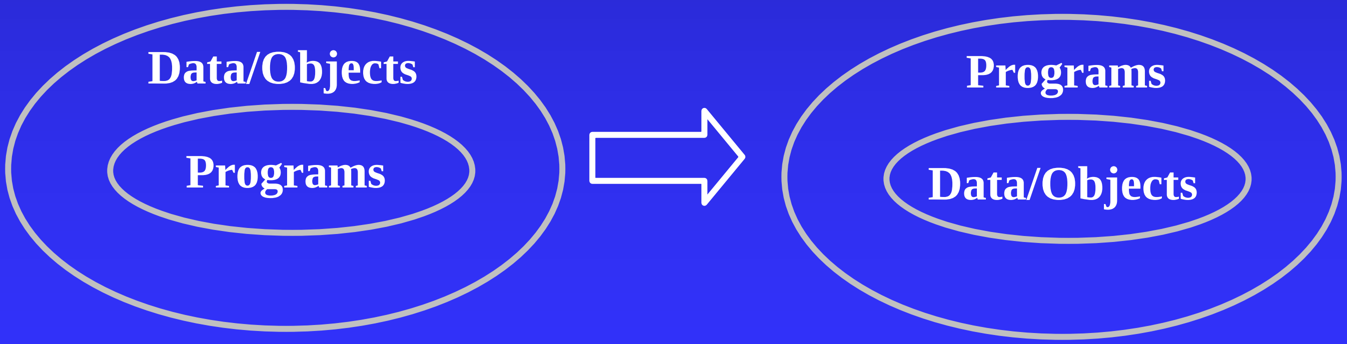
Interconnect-Centric Design Methodology

- Proposed transition

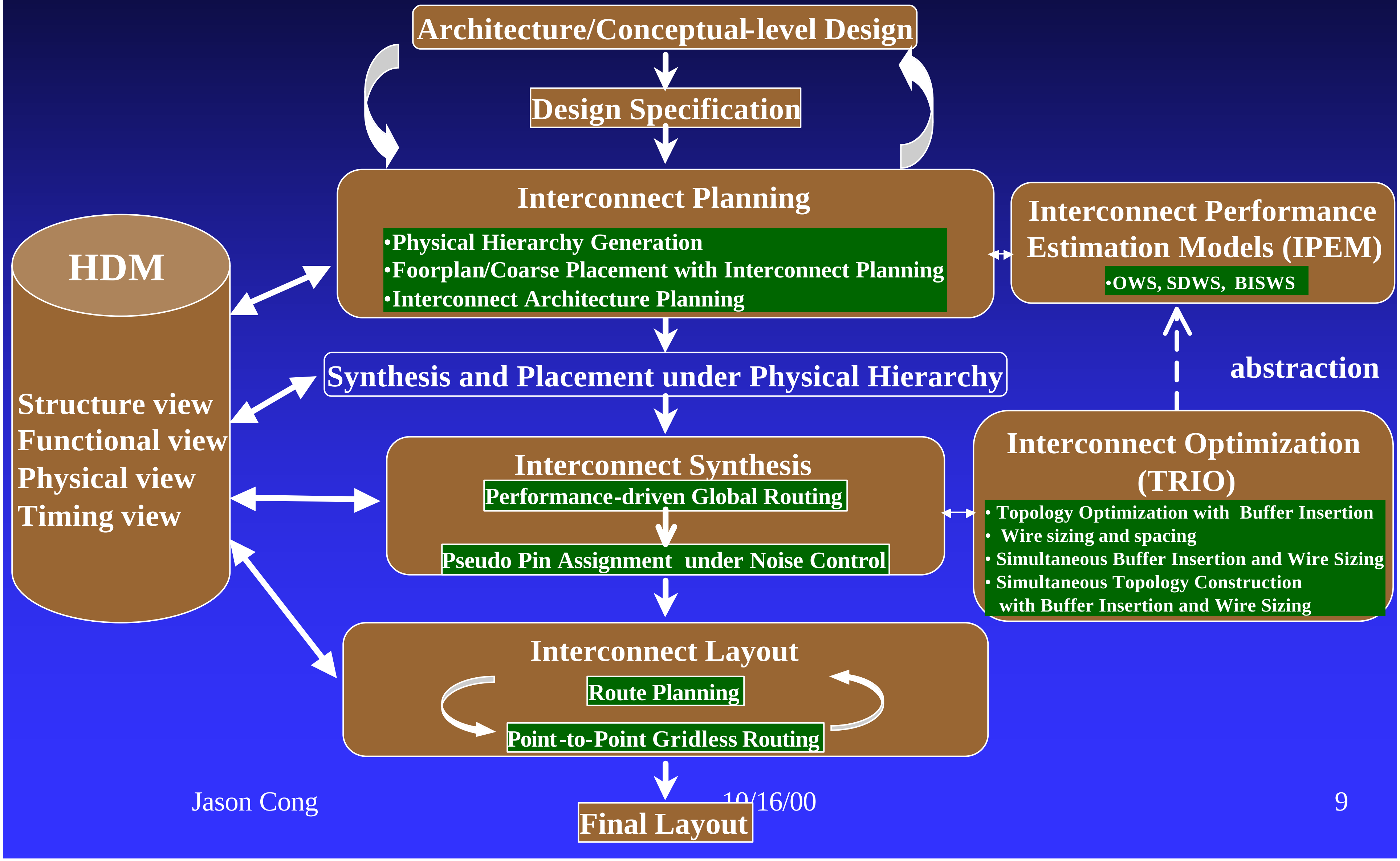


device/function centric interconnect/communication centric

- Analogy



Interconnect-Centric IC Design Flow
Under Development at UCLA



Architecture/Conceptual-level Design

Design Specification

Interconnect Planning

- •Physical Hierarchy Generation
- •Foorplan/Coarse Placement with Interconnect Planning
- •Interconnect Architecture Planning

Interconnect Performance Estimation Models (IPEM)

- •OWS, SDWS, BISWS

abstraction

HDM

Structure view
Functional view
Physical view
Timing view

Synthesis and Placement under Physical Hierarchy

Interconnect Synthesis

Performance-driven Global Routing

Pseudo Pin Assignment under Noise Control

Interconnect Optimization (TRIO)

- • Topology Optimization with Buffer Insertion
- • Wire sizing and spacing
- • Simultaneous Buffer Insertion and Wire Sizing
- • Simultaneous Topology Construction with Buffer Insertion and Wire Sizing

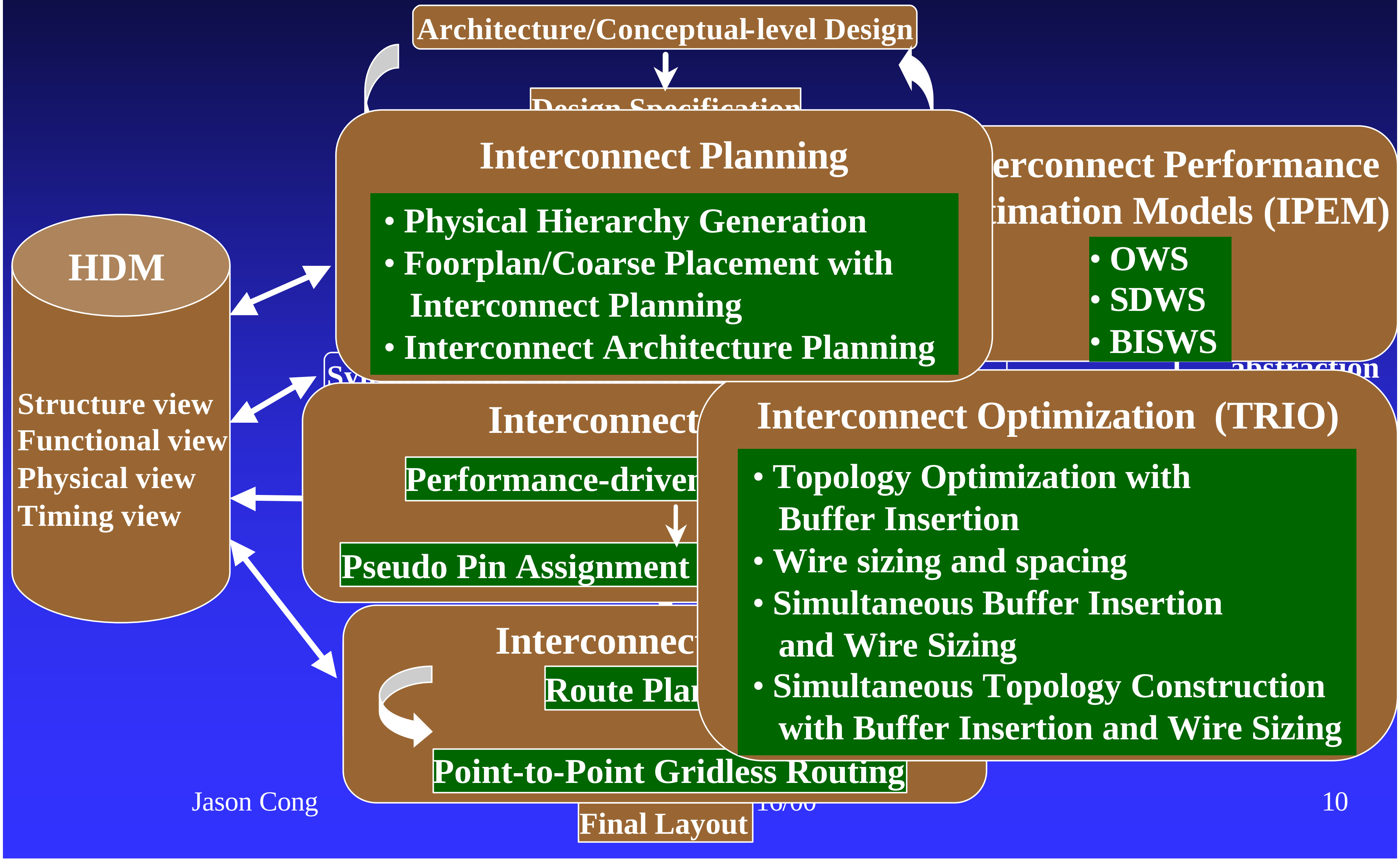
Interconnect Layout

Route Planning

Point-to-Point Gridless Routing

Final Layout

Interconnect-Centric IC Design Flow Under Development at UCLA



Architecture/Conceptual-level Design

Design Specification

HDM

Structure view
Functional view
Physical view
Timing view

Interconnect Planning

- Physical Hierarchy Generation
- Foorplan/Coarse Placement with Interconnect Planning
- Interconnect Architecture Planning

erconnect Performance
imation Models (IPEM)

- OWS
- SDWS
- BISWS

abstraction

Sy

Interconnect

Performance-driven

Pseudo Pin Assignment

Interconnect Optimization (TRIO)

- Topology Optimization with Buffer Insertion
- Wire sizing and spacing
- Simultaneous Buffer Insertion and Wire Sizing
- Simultaneous Topology Construction with Buffer Insertion and Wire Sizing

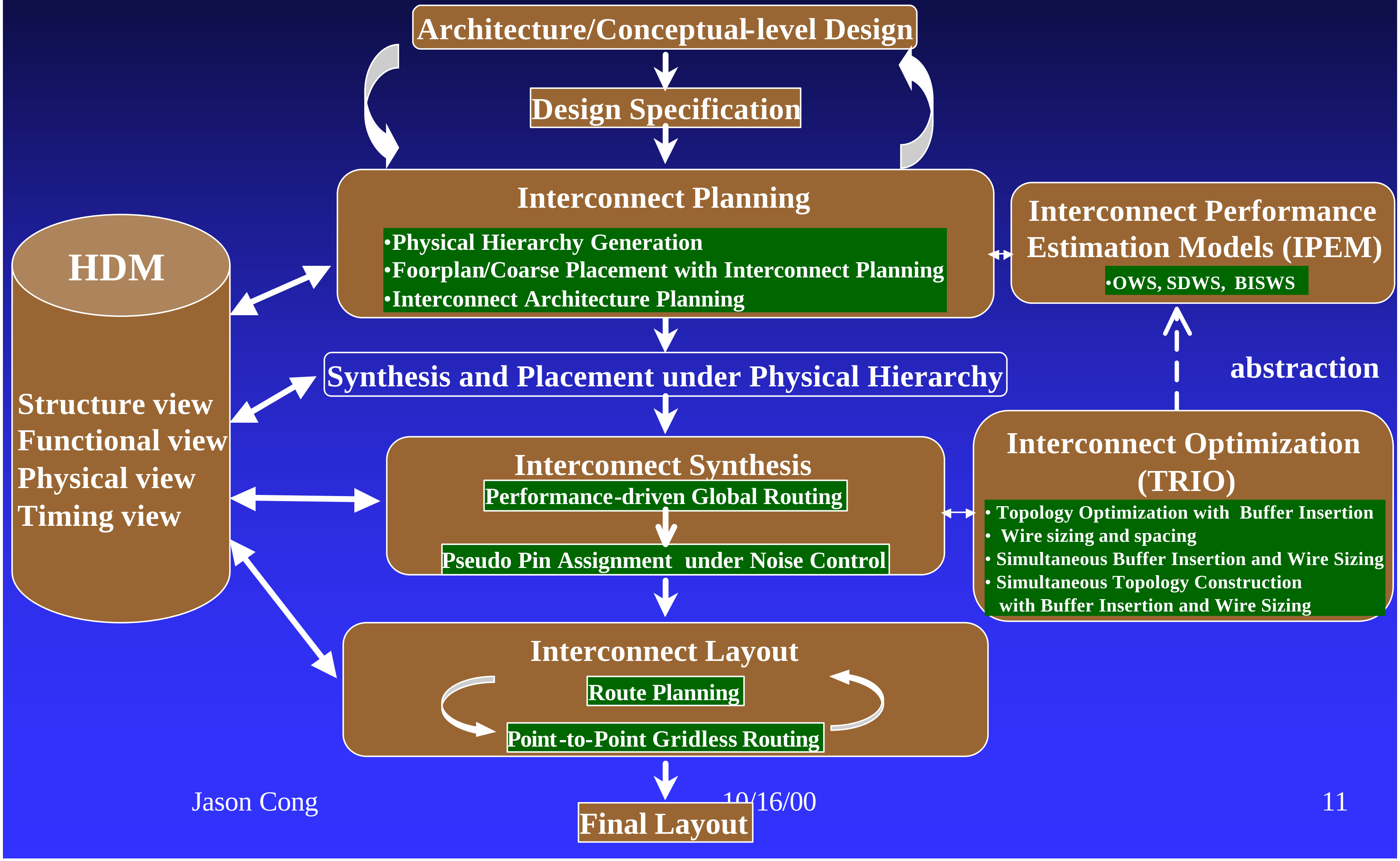
Interconnect

Route Plan

Point-to-Point Gridless Routing

Final Layout

Interconnect-Centric IC Design Flow Under Development at UCLA



Architecture/Conceptual-level Design

Design Specification

Interconnect Planning

- Physical Hierarchy Generation
- Foorplan/Coarse Placement with Interconnect Planning
- Interconnect Architecture Planning

Interconnect Performance Estimation Models (IPEM)

- OWS, SDWS, BISWS

HDM

Structure view
Functional view
Physical view
Timing view

Synthesis and Placement under Physical Hierarchy

abstraction

Interconnect Synthesis

Performance-driven Global Routing

Pseudo Pin Assignment under Noise Control

Interconnect Optimization (TRIO)

- Topology Optimization with Buffer Insertion
- Wire sizing and spacing
- Simultaneous Buffer Insertion and Wire Sizing
- Simultaneous Topology Construction with Buffer Insertion and Wire Sizing

Interconnect Layout

Route Planning

Point-to-Point Gridless Routing

Final Layout

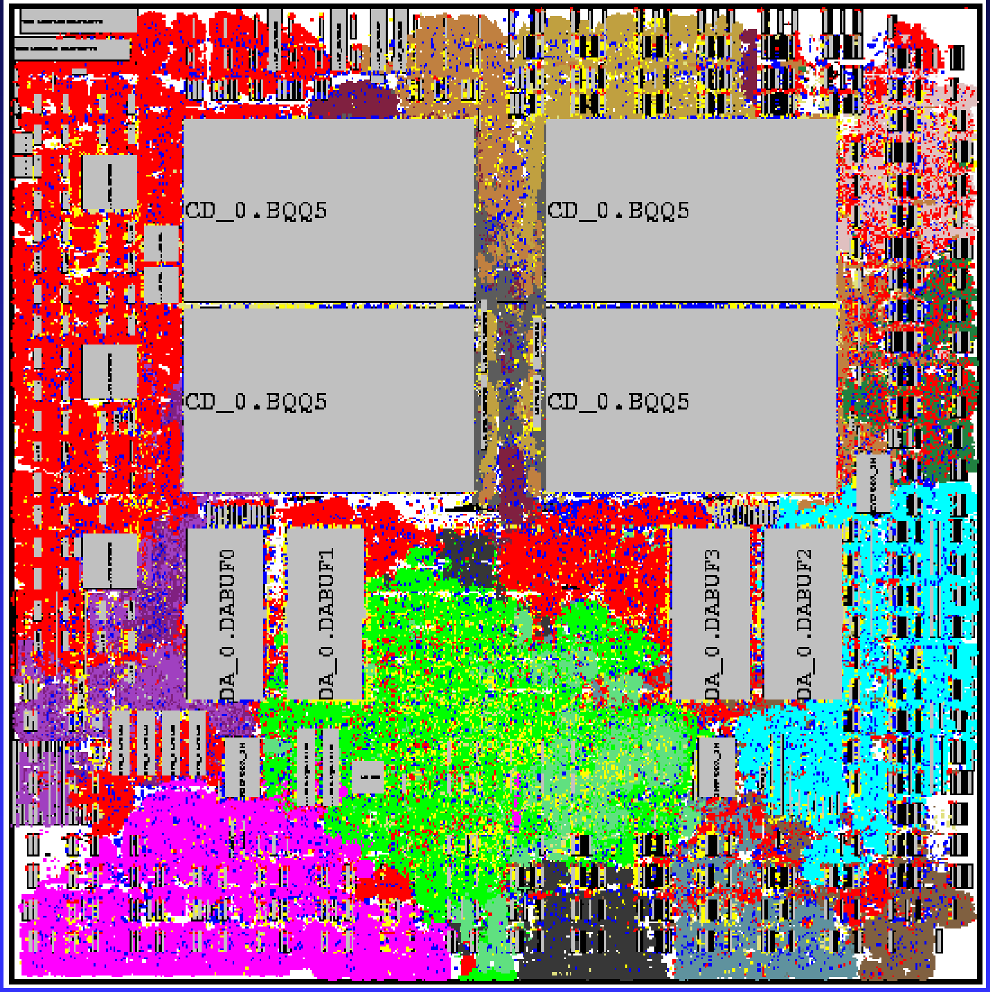
Interconnect Planning

- Physical Hierarchy Generation
- Floorplan/Coarse Placement with Interconnect Planning
- Interconnect Architecture Planning

Physical Hierarchy Generation

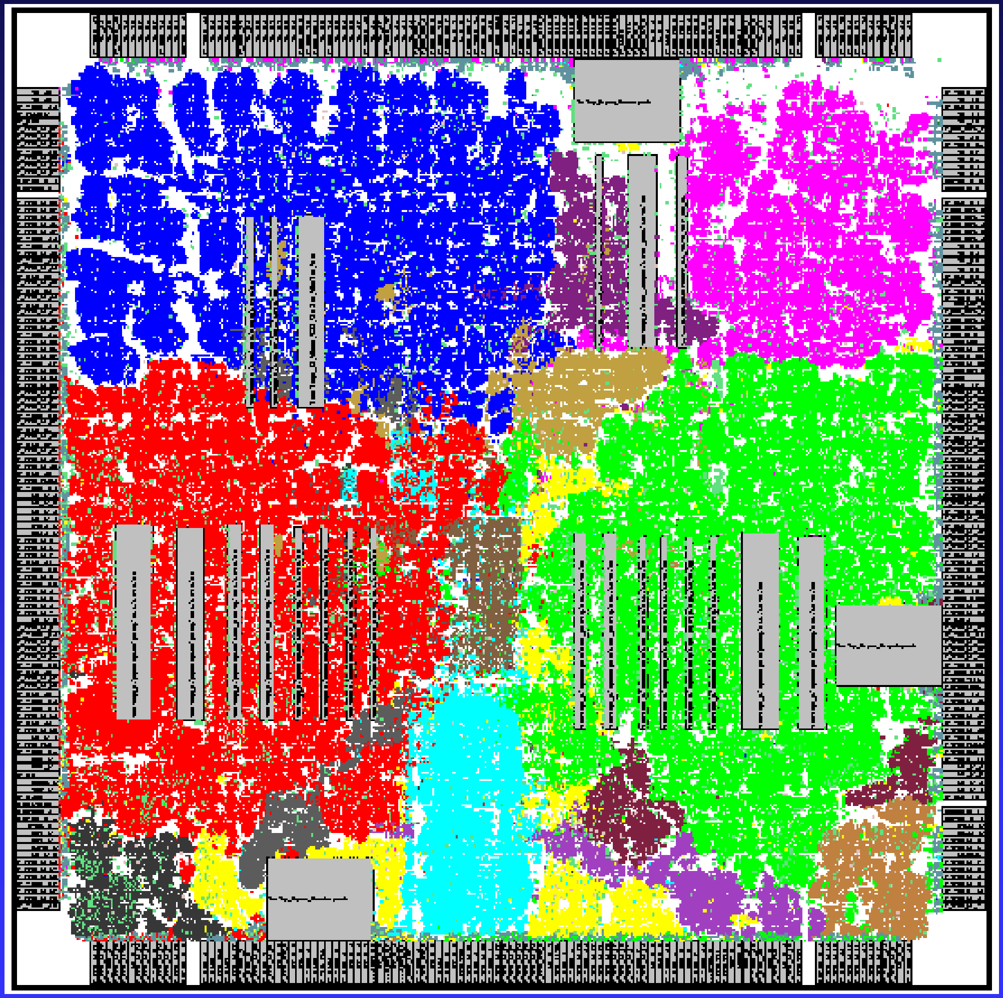
- Designs are hierarchical due to high complexity
- Design specification (in HDL) follows logic hierarchy
- Logic hierarchy may not be suitable to be embedded on a 2D silicon surface, resulting poor interconnect designs
  - RT-level floorplanning is a bad idea!
- Solution: transform logic hierarchy to physical hierarchy

Example of Logic Hierarchy in Final Layout



By courtesy of IBM (Tony Drumm)

Example of Logic Hierarchy in Final Layout



By courtesy of IBM (Tony Drumm)

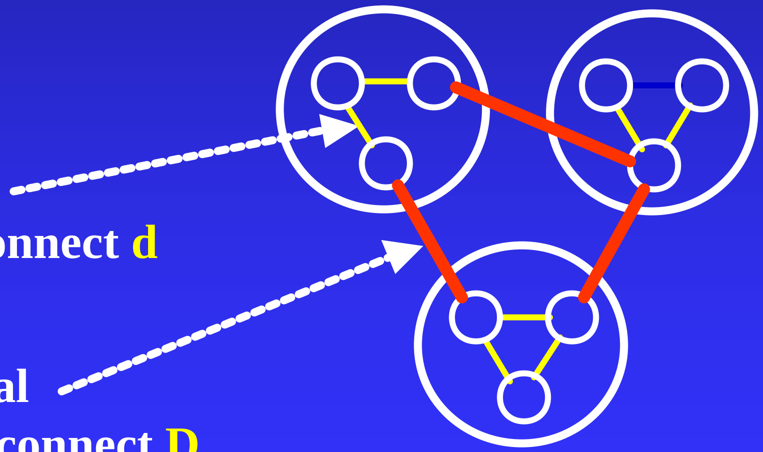
Transform Logic Hierarchy to Physical Hierarchy

- **Simultaneous partitioning, coarse placement, and retiming on the *flat* netlist to generate a good physical hierarchy**
  - **Synthesis will follow**

- **Use multi-level optimization to handle with the complexity**

Role of Partitioning

- Importance of Partitioning:
  - **Conventional view: enables divide-and-conquer**
  - **DSM view: defines global and local interconnects**



Local Interconnect d

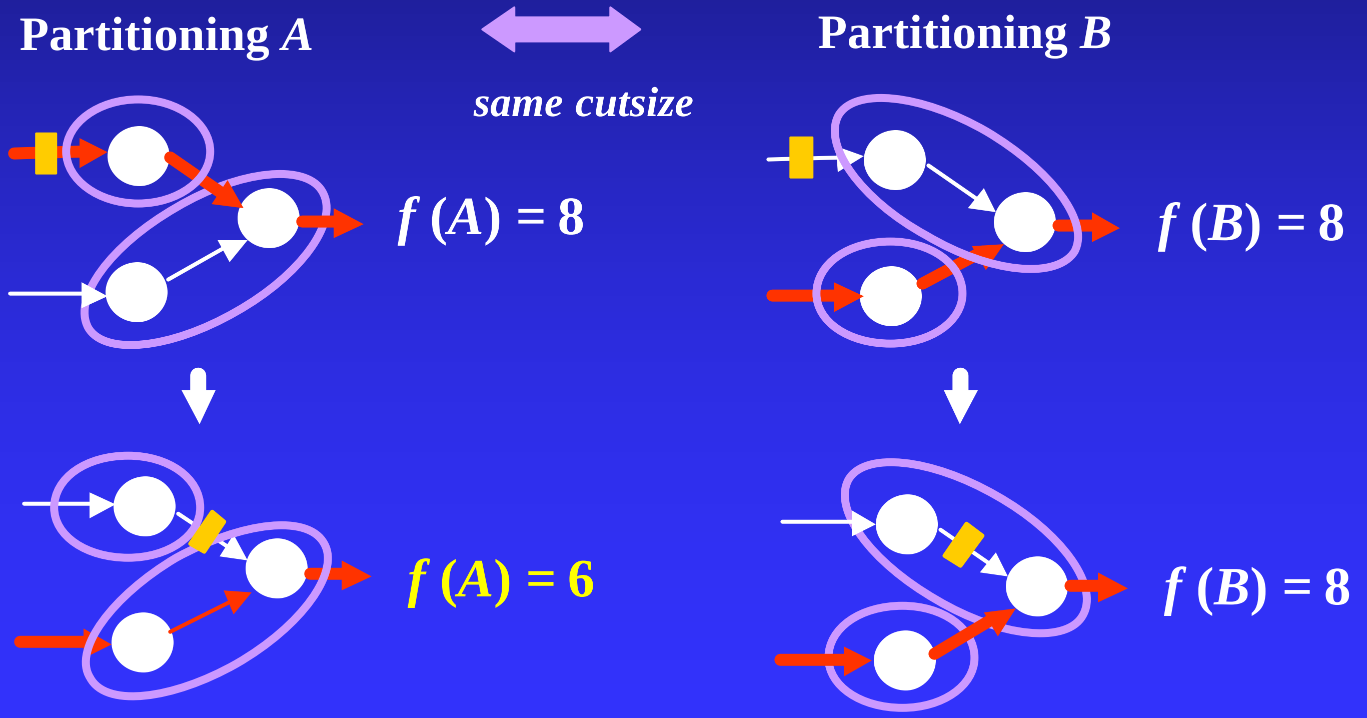
Global Interconnect D

D >> d !!!

Need of Considering Retiming during Partitioning

- Retiming/pipelining on global interconnects

- Multiple clock cycles are needed to cross the chip
- Proper partitioning allows retiming to hide global interconnect delays.



Partitioning *A*

same cutsize

Partitioning *B*

f (*A*) = 8

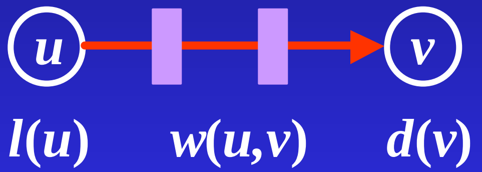
f (*B*) = 8

f (*A*) = 6

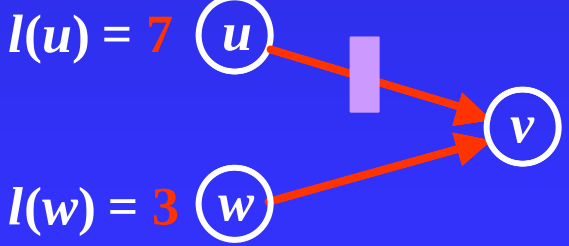
f (*B*) = 8

Sequential Arrival Time (SAT)

- **Definition [Pan et al, TCAD98]**
  - $l(v)$ = max delay from PIs to v after opt. retiming under a given clock period f
  - $l(v) = \max\{l(u) - f \cdot w(u,v) + d(u,v) + d(v)\}$



  - Relation to retiming: $r(v) =$ é$l(v) / f$ ù $- 1$
  - Theorem: P can be retimed to $f + \max\{d(e)\}$ iff $l(\text{POs})$ £ f



$d(v) = 1, d(e) = 2, f = 5$

$l(v) = \max\{7-5\cdot1+2+1, 3+2+1\} = 6$

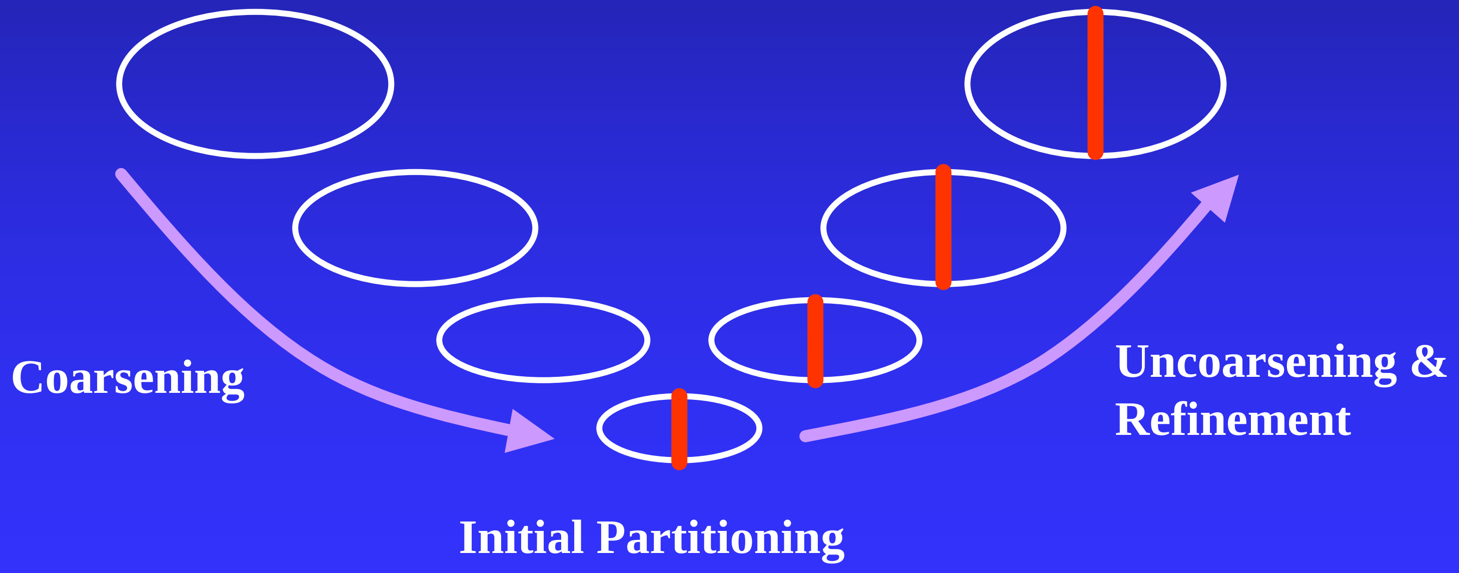
Simultaneous Partitioning/Placement with Retiming

- Minimize SAT during partitioning/placement
- Apply optimal retiming to the resulting solution (best suitable for retiming)
- Partitioning/placement with retiming can be applied recursively to generate physical hierarchy

- Good news: SAT can be computed efficiently (linear time in practice, quadratic time in the worst case)
- Difficulty: Flattened netlist can be very large!
  - Solution: use multi-level method

Multi-level Partitioning

- Iterative coarsening (clustering) to generate a multi-level hierarchy
- Initial partitioning on the coarsest level
- Iterative de-clustering and refinement



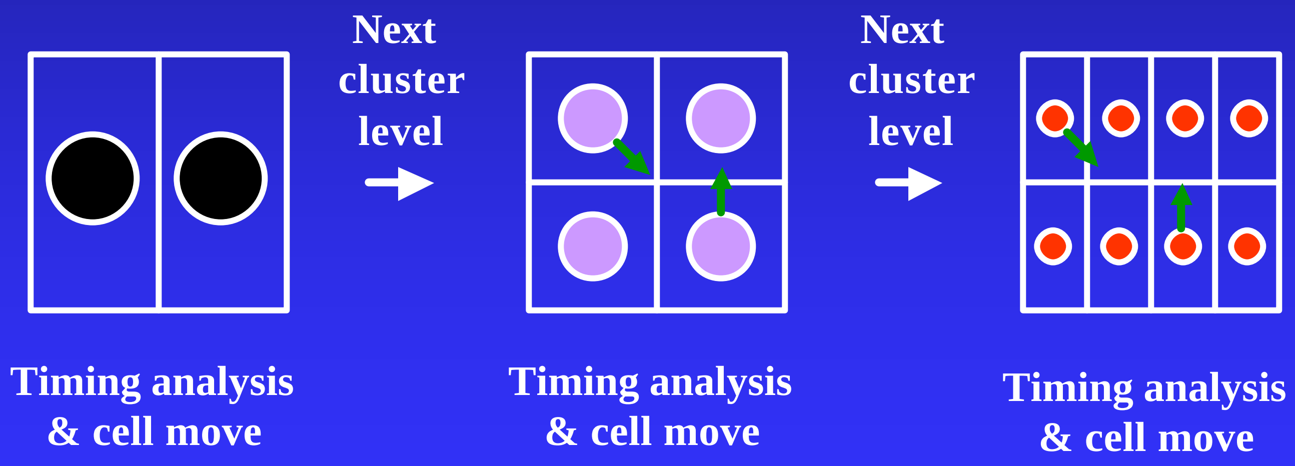
Hierarchical Approach vs Multi-Level Approach

- Hierarchical approach: higher-level design *constrains* lower-level designs
  - Not sufficient information at higher-level
  - Mistake at higher level is impossible or costly to correct
- Multi-level approach: finer-level design *refines* coarse-level design
  - Converge to better solution as more details are considered

Example: Multi-Level Partitioning with Coarse Placement & Retiming

- Bottom-up multi-level clustering
- Top down cell move based multi-level partitioning
- Sequential timing analysis at each level

[Cong and Lim, ICCAD00]



Next cluster level

Next cluster level

Timing analysis & cell move

Timing analysis & cell move

Timing analysis & cell move

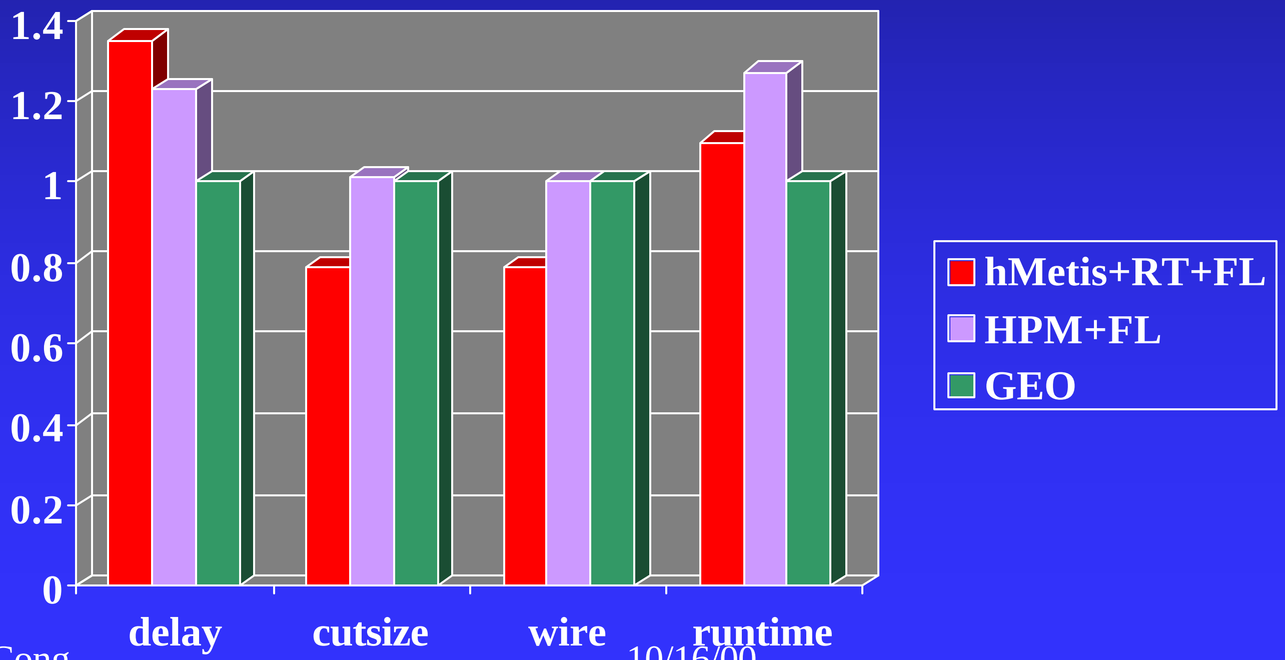
Success of Multi-Level Approach

- First used to solve partial differential equations (multi-grid method)
- Successfully applied to circuit partitioning (hMetis [Karypis et al, 1997])
  - Best partitioner for cut-size minimization
- Successfully applied to physical hierarchy generation (HPM and GEO [Cong et al, DAC’00 & ICCAD’00])
  - 30-40% delay reduction compared to hMetis
- Successfully applied to circuit placement [Chan et al, ICCAD’00]
  - 10x speed-up over GordianL

Experimental Results

- Comparison with existing algorithms
  - hMetis [DAC97] + retiming + slicing floorplan [Algo89]
  - HPM [DAC00] + slicing floorplan [Algo89]
  - GEO: simultaneous partitioning + coarse placement + retiming

Close to 40% delay reduction!



Interconnect Planning

- Physical Hierarchy Generation
- Floorplan/Coarse Placement with Interconnect Planning
  - Example: Buffer Block Planning in Floorplanning
- Interconnect Architecture Planning

Demand of Buffers in Nanometer Designs

- **Need to insert buffers in long global interconnects for performance optimization**

Technology (um)	***0.25***	***0.18***	***0.13***	***0.10***	***0.07***
#buffer per chip	5k	25k	54k	230k	797k

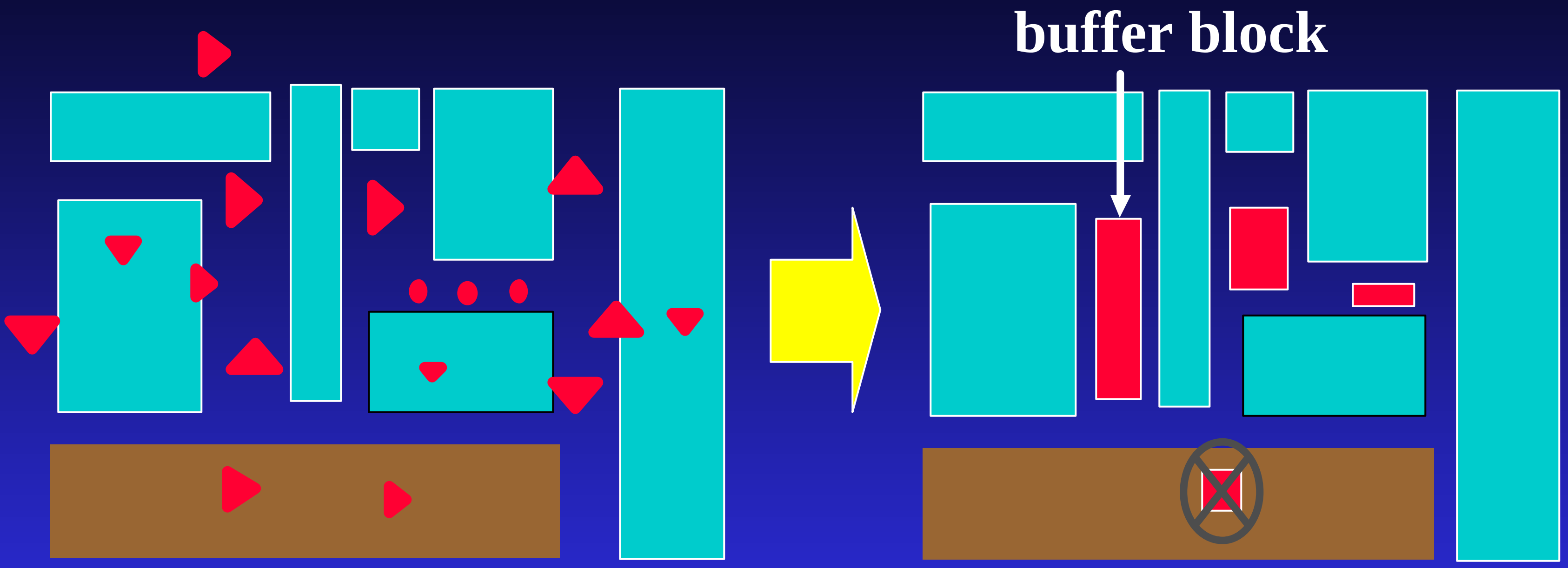
Source: [Cong'97, SRC Work Paper]

http://www.src.org/research/frontier.dgw

(Estimated based on NTRS'97 & [Davis-Meindl'97])

Buffer Block Planning Problem

[Cong-Kong-Pan, ICCAD'99]

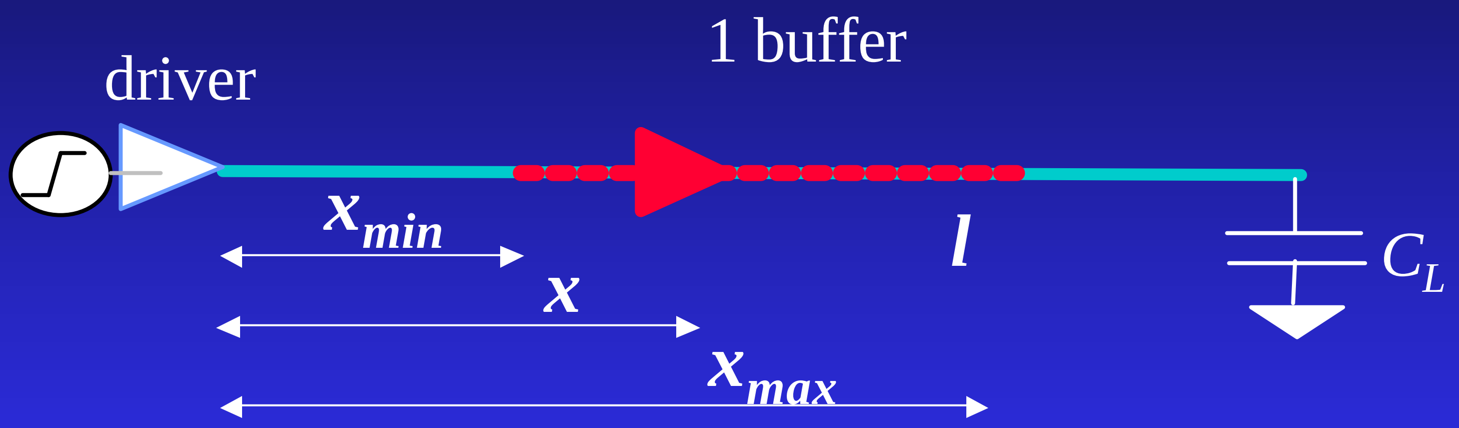


- **Restriction from hard IP blocks**
- **Implications on P/G routing**
- **Impact on floorplan configuration**

=> need to plan ahead for buffers.

Optimal Buffer Location Can Be Relaxed

- **Closed-form formula of feasible region (FR) for inserting one buffer to meet delay constraint**



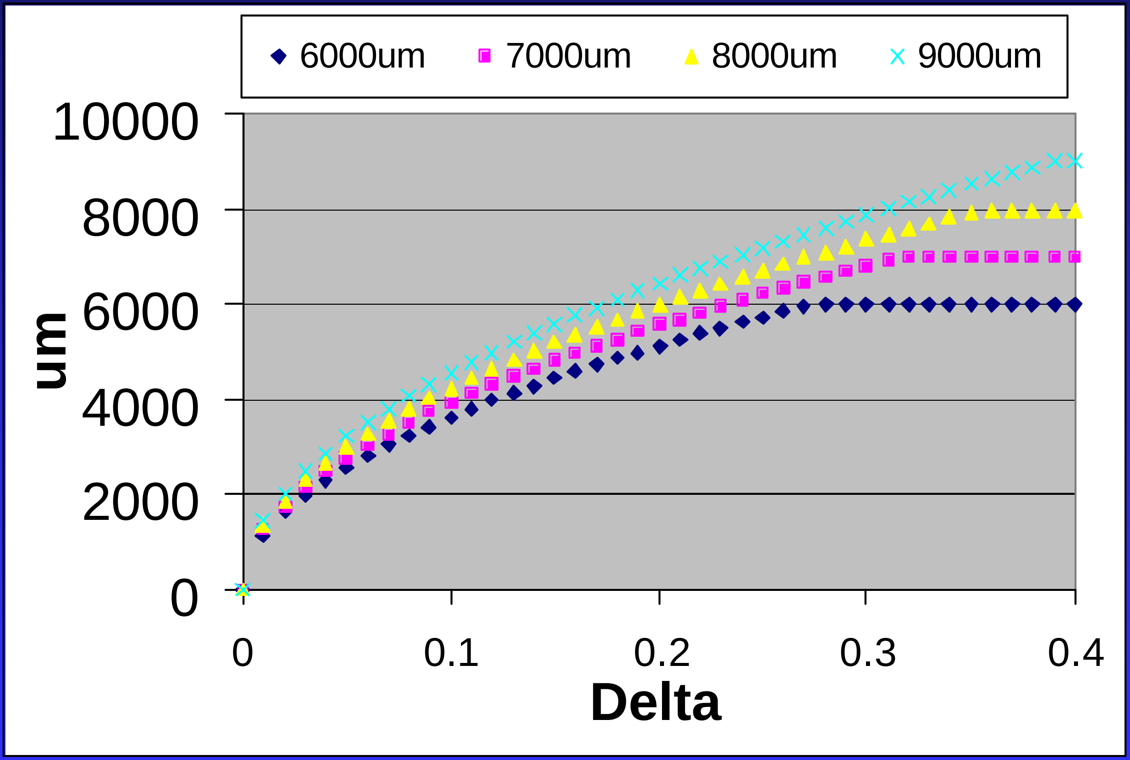
$$x \in [x_{\min}, x_{\max}]$$

$$x_{\min} = MAX\left(0, \frac{K_2 - \sqrt{K_2^2 - 4K_1K_3}}{2K_1}\right)$$

$$x_{\max} = MIN\left(l, \frac{K_2 + \sqrt{K_2^2 - 4K_1K_3}}{2K_1}\right)$$

Feasible Region (FR) Is Very Large

- Even under tight delay constraint, FR for BI can still be very large!



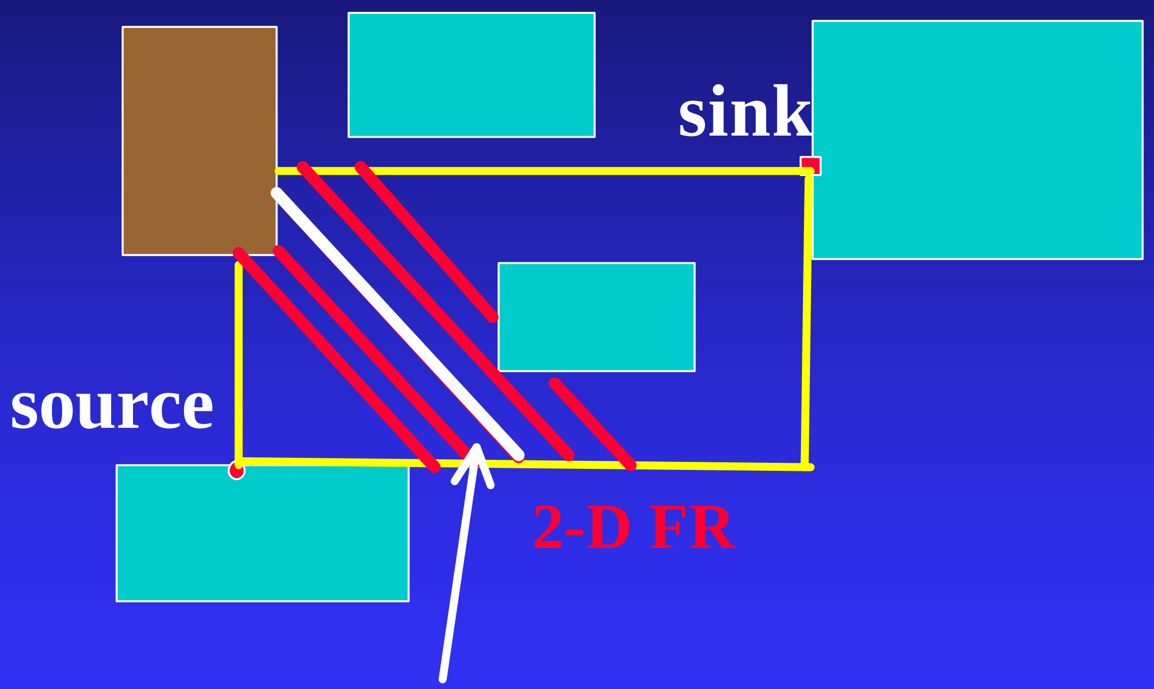
❖ Delay budget is (1+Delta) T_{opt} (the best delay by optimal buffer insertion)

Delta	FR
1%	19%
5%	43%
10%	60%
20%	86%

=> FR provides a lot of flexibility to plan buffer location

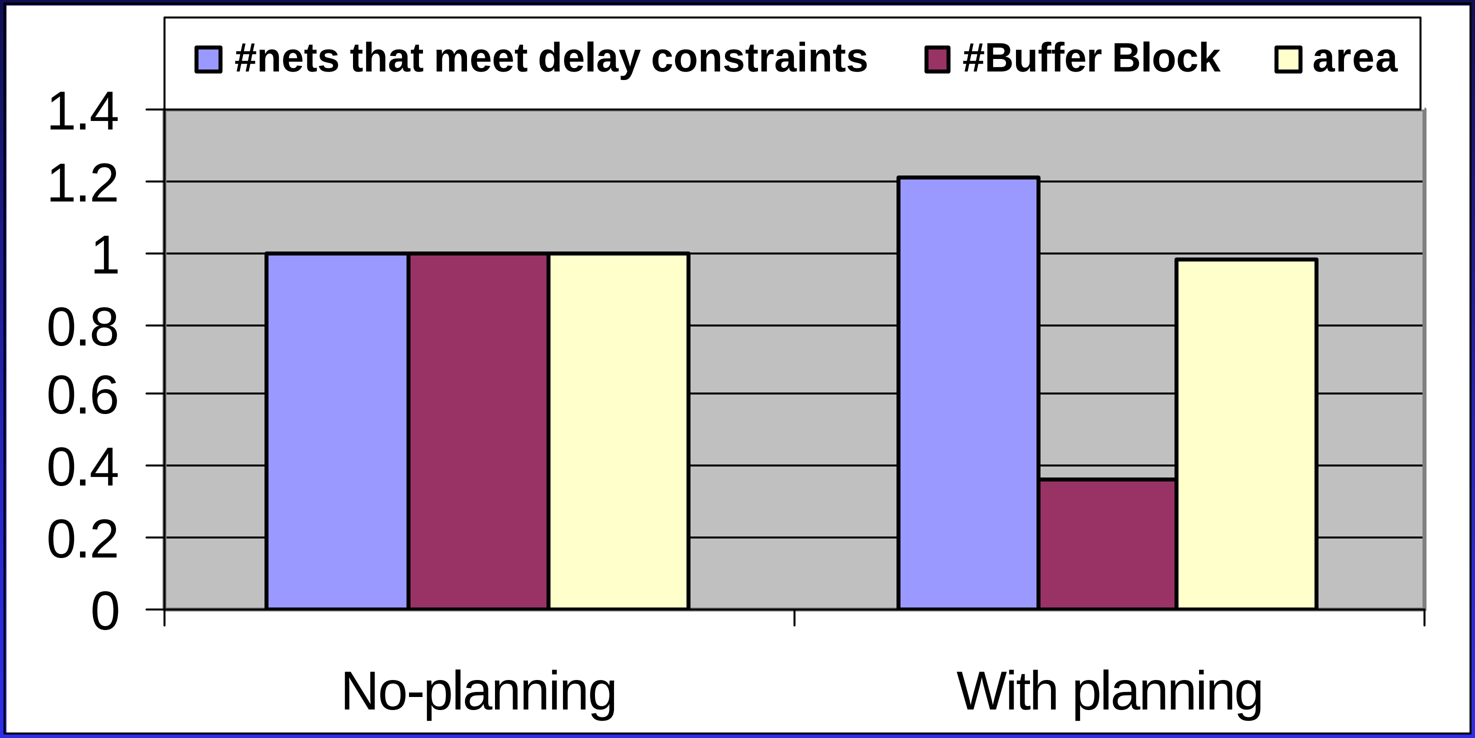
Extension: 2D Feasible Region

- FR extended to 2-dimension with obstacles



Locus of min-delay BI (Restricted lines)

Experimental Results of Buffer Block Planning



Buffer block planning reduces # buffer blocks, better meets timing constraints, and use smaller area

Concluding Remarks

- Interconnects determine system performance
- Interconnect-centric design is needed
  - Interconnect planning
  - Interconnect synthesis
  - Interconnect layout
- Physical hierarchy generation is crucial for interconnect planning
- A good combination of partitioning/placement and retiming can hide global interconnect delays, and lead to good physical hierarchy
- Multi-level method is an effective way to cope with complexity