VLSI architecture,
synthesis & technology

UCLA

Coping with Interconnects

Jason Cong

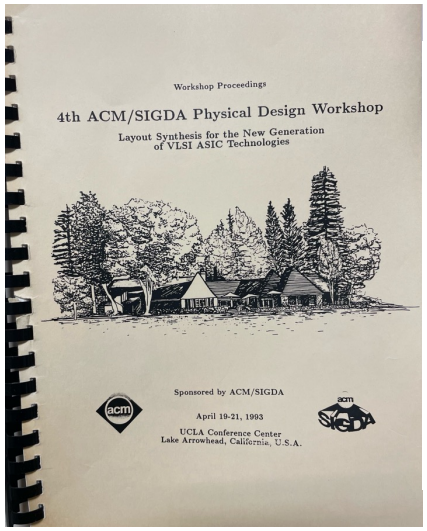
Volgenau Chair for Engineering Excellence, UCLA Computer Science

Director, Center for Domain-Specific Computing (CDSC)

https://vast.cs.ucla.edu/people/faculty/jason-cong

Deep Appreciation of ISPD for this Great Honor

- I have learned a lot from friends and colleagues in the ISPD community
- ISPD has been a wonderful forum where we shared many of our research results
- I am glad that I had the opportunity to contribute to ISPD



WELCOME

Welcome to the 1993 ACM/SIGDA Physical Design Workshop! This is the forth biannual workshop on VLSI physical design automation since the series was started in 1987.

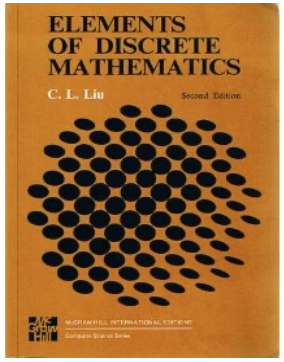
The rapid advances in VLSI ASIC technologies have led to many new challenges in the physical design automation of VLSI systems: The increasing emphasis on system performance requires timing constraints to be considered at every stage of physical design; the constantly decreasing feature size leads to much denser circuits and the interconnection delay becomes the dominating factor in system performance; the widespread use of automatic logic synthesis tools complicates many layout problems; the strong need for shorter design cycles and lower design costs have resulted in the fast development of field-programmable gate-arrays (FPGAs) and field-programmable interconnects (FPICs). The objective of this workshop is to provide a forum to discuss and investigate these emerging problems in physical design automation for the new generation of VLSI ASIC technologies. We hope that you find the workshop program interesting and exciting.

Jason Cong	**Bryan Preas**	**Carl Sechen**	**Mary Jane Irwin**
Workshop Chair	**Workshop Co-Chair**	**Program Chair**	**Publication Chair**
UCLA	Xerox PARC	Univ. of Washington	Penn State

- But let's start from beginning – how did I start with physical design?

From PKU to UIUC

- Started my undergraduate in Peking University
  - Changed the preferred major from math to Computer Science in the last minute
- One of my most favorite courses: Discrete mathematics
  - In particular, enjoy reading Prof. C. L. Liu's textbook
- Started my PhD at UIUC in Jan. 1986
  - The only PhD program I applied (thanks to the recommendation of Prof. Gongben Wang)



An Unexpected Journey into EDA

- Dave started to switch to EDA in mid 1980s
- A great decision -- it was the “Big data” problem at the time
  - Plenty of opportunities for algorithmic innovations
  - Dave received 2011 Phil Kaufman Award for “leading the transformation from ad hoc EDA to algorithmic EDA”
  - The ISPD Lifetime Achievement Award in 2012
- But it was a worrisome to me
  - Not sure if it’s a good fit for me

My First Paper in Physical Design (and EDA)

- On three-Layer channel routing
- Took a novel transformation-based approach
- Used several combinatorial optimizations
  - Two-processor scheduling, shortest path …
- Had a great collaborator – Martin Wong

[ICCAD'1987]

A New Approach to Three- or Four-Layer Channel Routing

JINGSHENG CONG, D. F. WONG, AND C. L. LIU, FELLOW, IEEE

Abstract—We present in this paper a new approach to the three- or four-layer channel routing problem. Since two-layer channel routing has been well studied, there are several two-layer routers which can produce optimal or near optimal solutions for almost all the practical problems. We develop a general technique which transforms a two-layer routing solution systematically into a three-layer routing solution. This solution transformation approach is different from previous approaches for three-layer and multilayer channel routing. Our router performs well in comparison with other three-layer channel routers proposed thus far. In particular, it provides a ten-track optimal solution for the famous Deutsch's difficult example, whereas other well known three-layer channel routers required 11 or more tracks. We extend our approach to four-layer channel routing. Given any two-layer channel routing solution without an unrestricted dogleg that uses w tracks, our router can provably obtain a four-layer routing solution using no more than $\lceil w/2 \rceil$ tracks. We also give a new theoretical upper bound $\lceil d/2 \rceil + 2$ for arbitrary four-layer channel routing problems.

I. Introduction

A KEY PROBLEM in VLSI layout design and implementation is the channel routing problem. The two-layer channel routing problem has been studied exten-

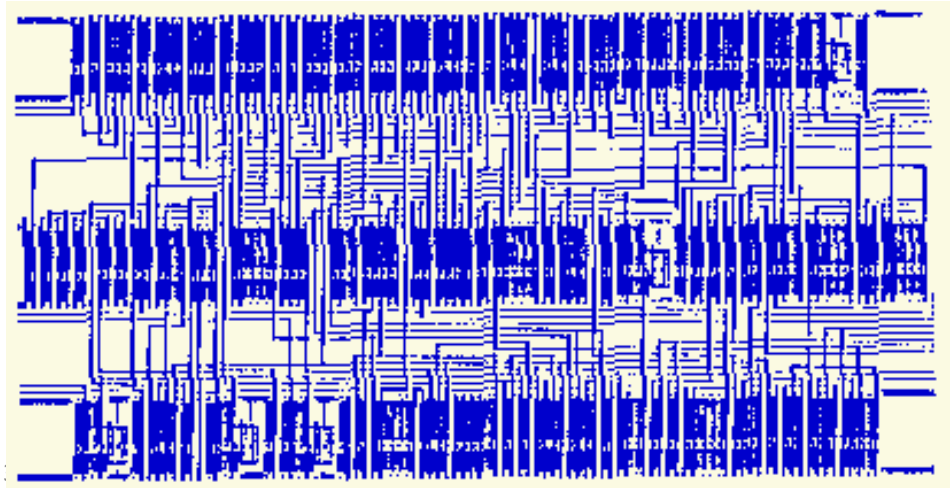
megabit DRAM designed by Taguchi *et al.* uses four routing layers, three layers of polysilicon and one layer of metal. Thus, the design and implementation of channel routing algorithms using a small number of layers (usually three or four layers) are not only practical, but also are becoming more and more important.

The multilayer channel routing problem has been studied in the literature. Chen and Liu [5] presented a three-layer channel router based on the net merging method used by Yoshimura and Kuh [22] for two-layer channel routing. Bruell and Sun [3] designed a "greedy" router for three-layer channel routing and obtained the first 11-track solution for Deutsch's difficult example. Braun *et al.* [2] implemented a multilayer channel router which divides layers into several groups. Each group contains two or three layers and routing for each group is done by the extended two-layer router YACR2 [20]. Enbody and Du [11] developed a multilayer router using leading column heuristics and limited backtracking. As for theoretical results, Hambrusch [15] obtained some near-optimal upper bounds for the case of two terminal nets allowing mixed wiring on the same layer. Brady and Brown [1] proposed

With Martin for his ISPD Lifetime Achievement Award in '2024

My Real Appreciation of EDA

- Summer intern at Xerox PARC with Bryan Preas (1987)
  - Practiced VLSI design beyond graph optimization
- Developed a standard cell global router as part of PARC DATool system
  - Published a paper in ICCAD'88
  - Got a big layout plot (from Tektronix printer)



ISPD'2014 life time achievement award for Bryan

PhD Thesis on “Routing Problems in the Physical Design of VLSI Circuits (1990)

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

THE GRADUATE COLLEGE

JULY 1990

WE HEREBY RECOMMEND THAT THE THESIS BY

JINGSHENG CONG

ENTITLED ROUTING ALGORITHMS IN THE PHYSICAL DESIGN OF VLSI CIRCUITS

BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Director of Thesis Research

Head of Department

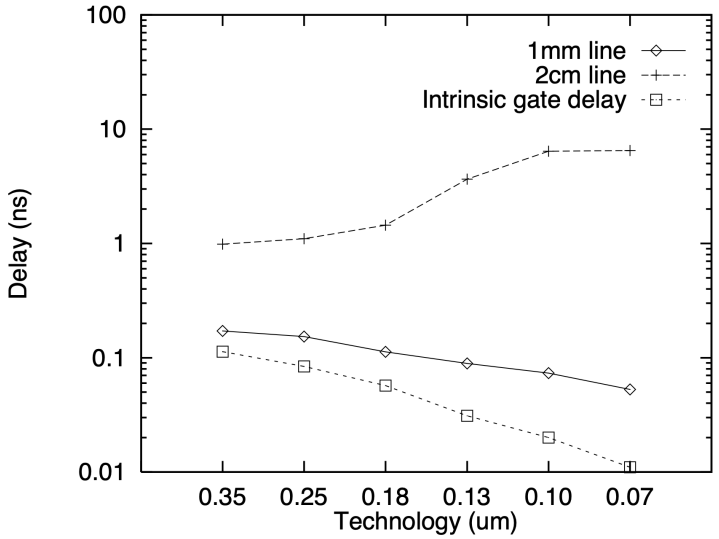
Committee on Final Examination†

Chairperson

† Required for doctor's degree but not for master's.

“Houston, we have a problem” -- Interconnects

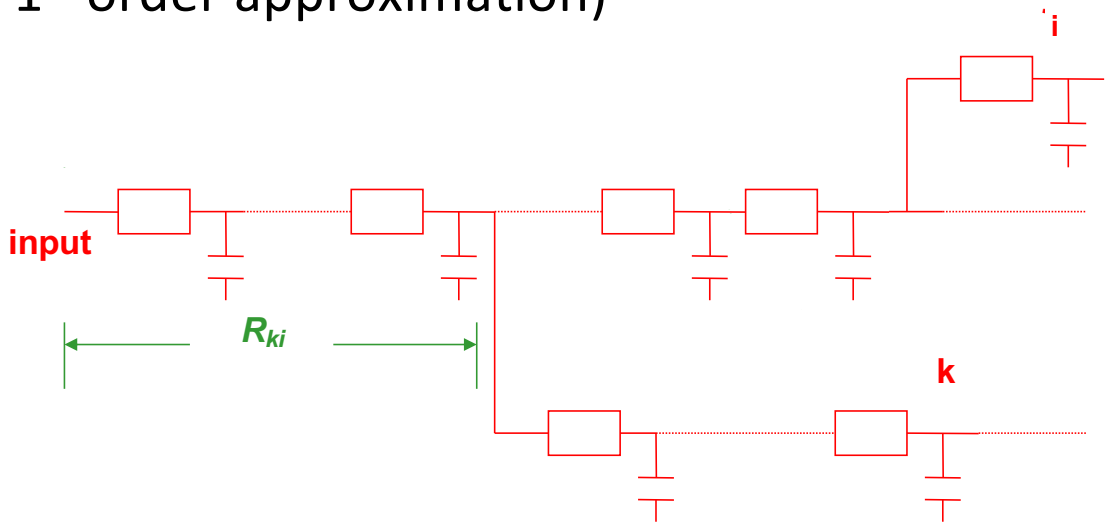
- (Local) Interconnect delays dominate gate delays
- My first NSF proposal
  - “Interconnection Problems for High-Performance VLSI Circuits and Systems”
  - NSF Initiation Award
  - $70K for two years (7/1991 – 6/1993)



After various interconnect optimizations [ICCAD’97]

Understanding the Interconnect Delay

- P_i: path from input to node i
- R_{ki}: resistance of common path $P_i \cap P_k$ from input to $i\&k$
- Elmore delay to node i (the 1st order approximation)
  - $T_{D_i} = \sum_k R_{ki} C_k$



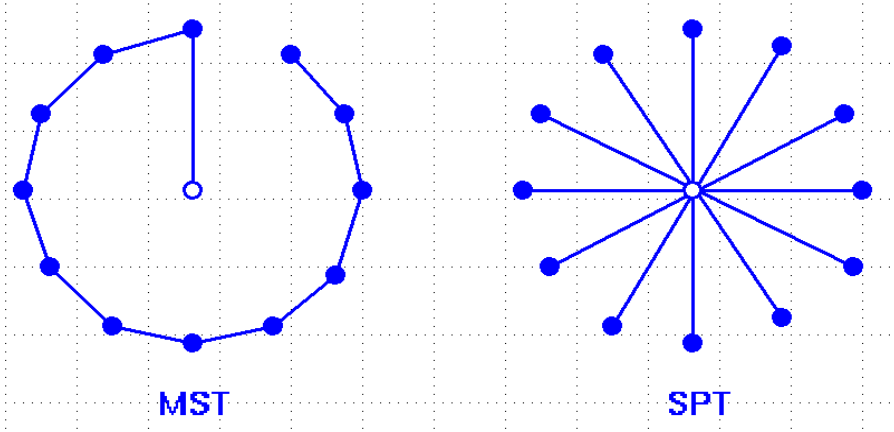
Approach 1: Interconnect Topology and Geometry Optimization

(The decade of 1990s)

Interconnect Topology Optimization

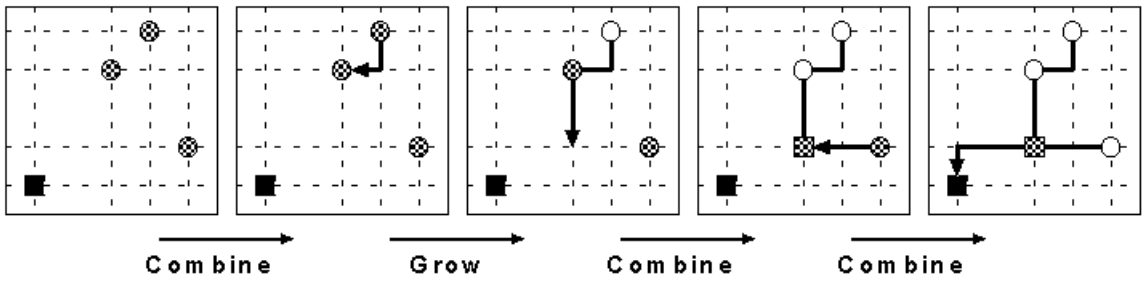
- Conventional Routing Algorithms Are Not Good Enough
  - Minimum spanning tree (MST) may have very long source-sink path
  - Shortest path tree (SPT) may have very large routing cost
- It's possible to bound the radius and total length simultaneously
  - Path length <= $(1+\varepsilon) * R_{SPT}$, and cost <= $(1+(2/\varepsilon)) * cost(MST)$

[Cong-Kahng-Robin-Sarrafzadeh-Wong, T-CAD'92]



In Practice, A-tree Algorithm Works Well [Cong-Leung-Zhou, DAC'93]

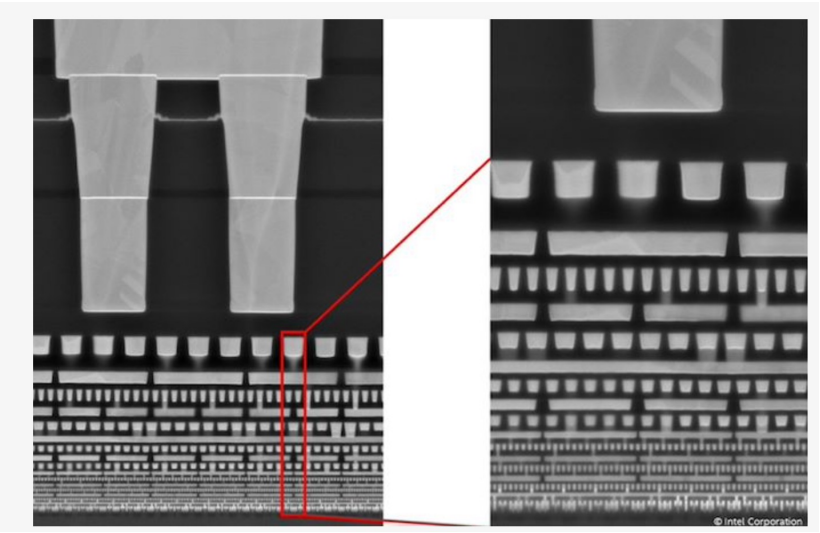
- A-tree: a minimum-cost rectilinear Steiner arborescence (SPT)
- Start with a forest of n single-node A-trees
- Apply a sequence of moves
  - **Grow an existing A-tree, or**
  - **Combine two A-trees into a new one**
- Terminate when only one A-tree is left



An NP-hard problem (shown later [Shi & Su, SIAM JOC'05]).
But the A-tree heuristic was at most 4% from optimal; Can be combined with buffer insertion.

Simultaneous Width and Spacing Optimization

- Recall the Elmore delay
  - $T_{D_i} = \sum_k R_{ki} C_k$
- Reduce path resistance
  - Widen the wires near the source
- Bound the capacitance increase
  - Taper down the wires near the sink
- Care about the coupling cap to the neighbours
- Can solved optimally
  - Discrete sizing [Cong-Leung, ICCAD'93]
  - Continuous sizing [Gao-Wong, T-CAD'99] ...

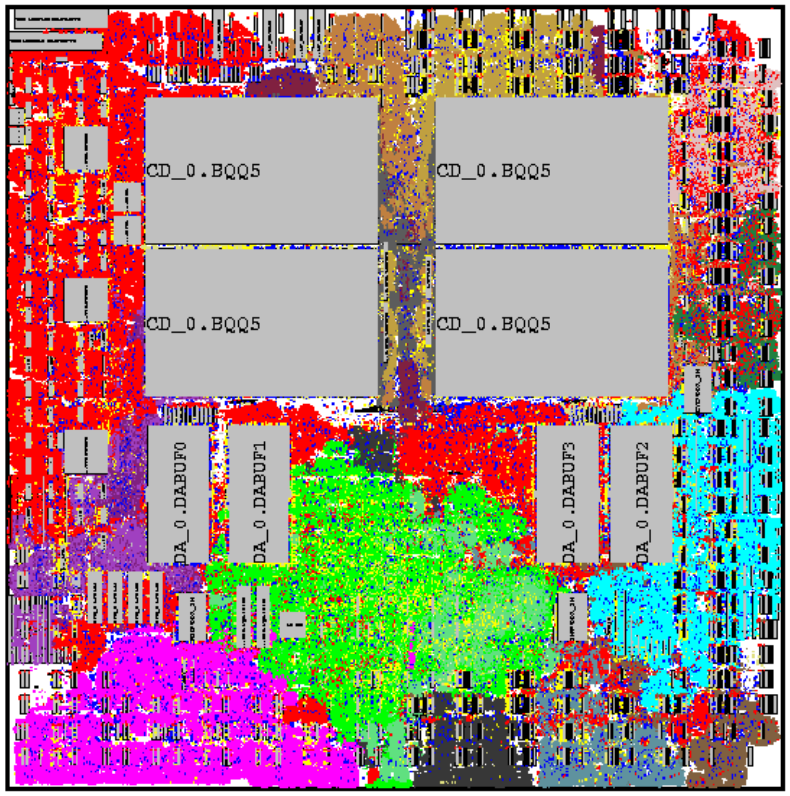


Approach 2: Scalable Placement
(The decade of 2000s)

Routing Is Determined by Placement

- **Also, following logic hierarchy may lead to sub-optimal solutions**
- **Need scalable placement to generate physical hierarchy**

Example of Logic Hierarchy in Final Layout

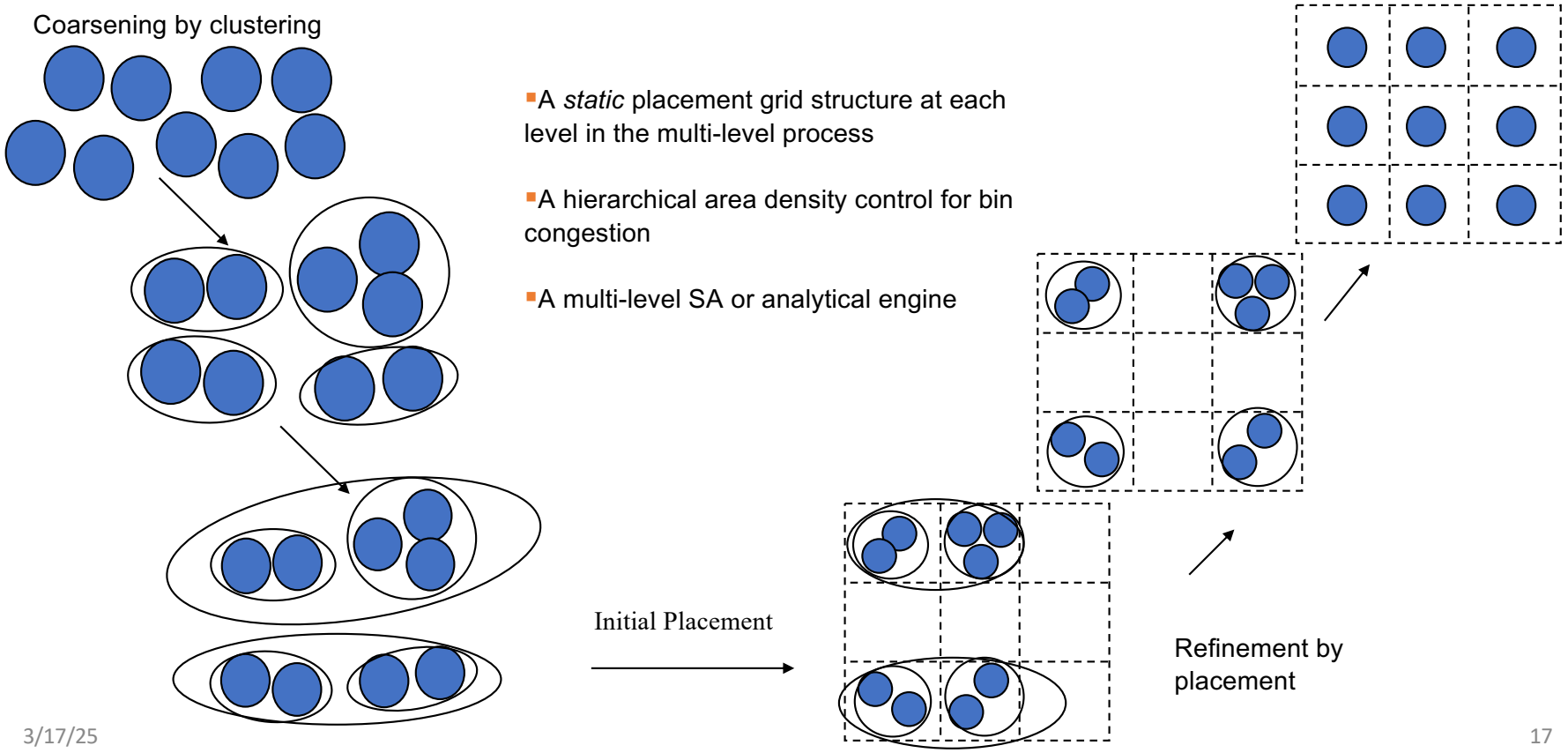


By courtesy of IBM (Tony Drumm)

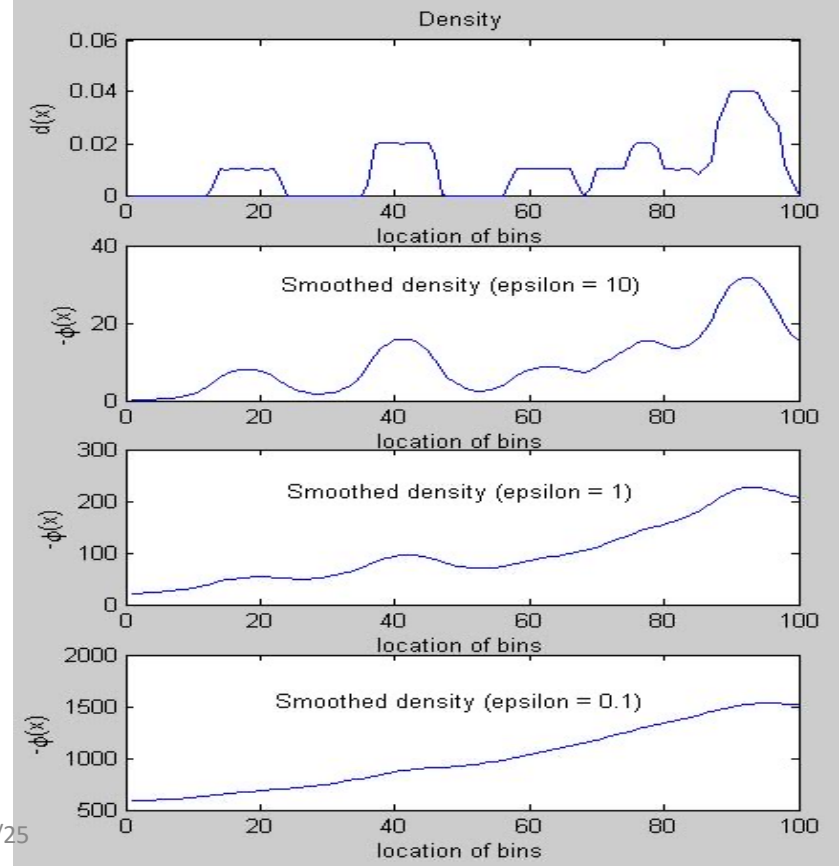
How Good Are Existing Circuit Placement Tools ?

- How do we know? It's NP-hard!
- Placement examples with known optimal (PEKO)
  - Up to 2 million placeable objects [Chang et al., TCAD'04]
  - *Initial WL gap: 1.6x – 2.5x (2003)*
- Multiple EE Times articles coverage, e.g.
  - Placement tools criticized for hampering IC designs [Feb'03]
- Many downloads from our website
  - Cadence, IBM, Intel, Magma, Mentor Graphics, Synopsys, …
  - CMU, MIT, SUNY, UCB, UCSB, UCSD, UIC, UMichgan, UWaterloo, …
- Optimality gap on PEKO was narrowed down to ~20% as of 2007 (from 60% - 150%)
- Improvement on real circuits as well
  - 30+% improvement by mPL placer 2003-06

Our Contribution 1: Multi-Level Placement



Our Contribution 2: Density-Smoothed Constrained Optimization



- Smoothing operator:

$$(\Delta - \varepsilon I)\phi = d$$

- Larger epsilon
  - More local smoothing
  - Slow convergence
- Smaller epsilon
  - More global smoothing
  - Faster convergence

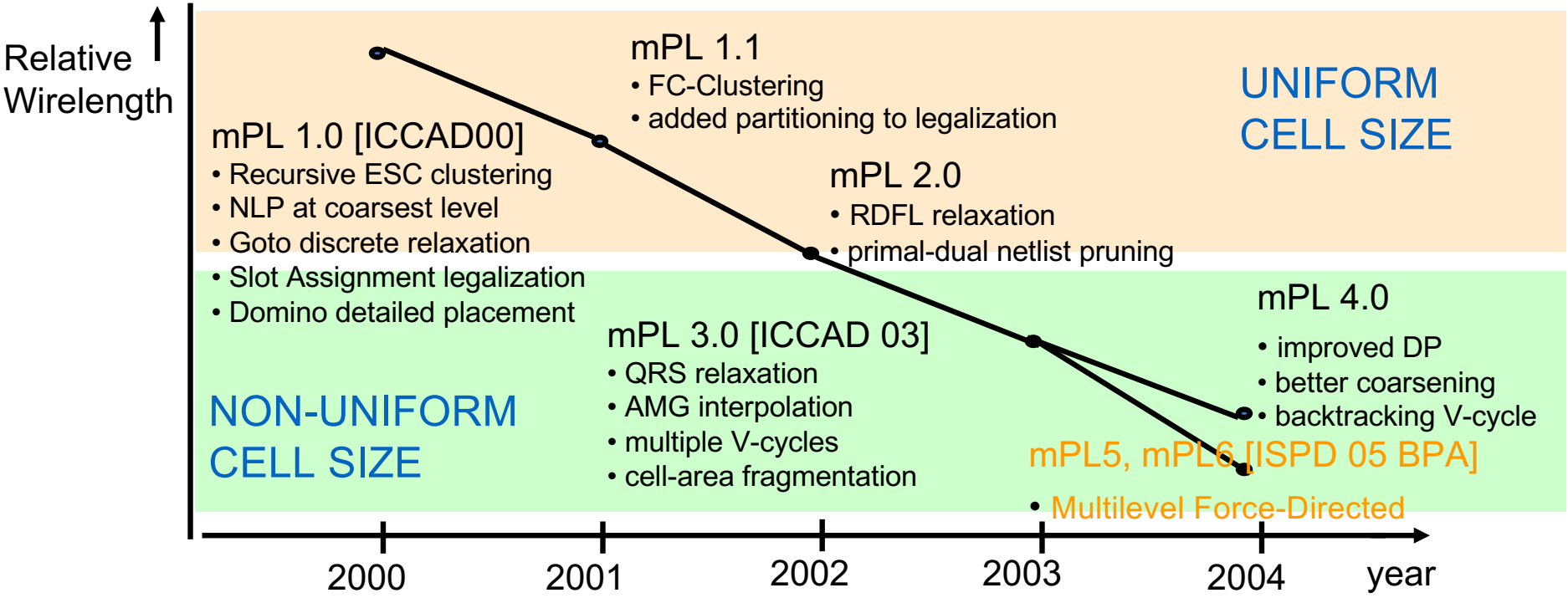
$$\Delta\psi \equiv \frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2}$$

WL Minimization Under the Smoothed Density Constraints [ISPD 2005]

- Minimize wirelength subject to smooth density function:

$$\min \quad W(x)$$
$$s.t. \quad \phi(x) = \kappa,$$
$$\text{where } \phi(x) = \Delta_{\varepsilon}^{-1} d(x), \kappa = \Delta_{\varepsilon}^{-1} c.$$

A Brief Review of Our Multilevel Placer mPL



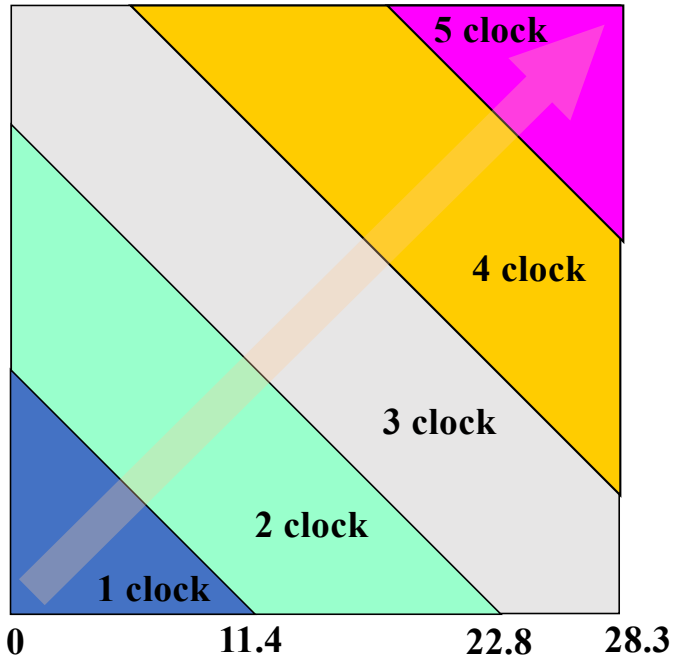
Great recent progress in circuit placement: RePlace [Cheng2018], DreamPlace [Lin2019], DreamPlace3 [Gu2020]

Approach 3: Space-Time Co-optimization for Interconnect Pipelining

(Early 2000s – present)

We Were Losing the Fight for Keeping Single-cycle Cross-chip Communication in Early 2000s …

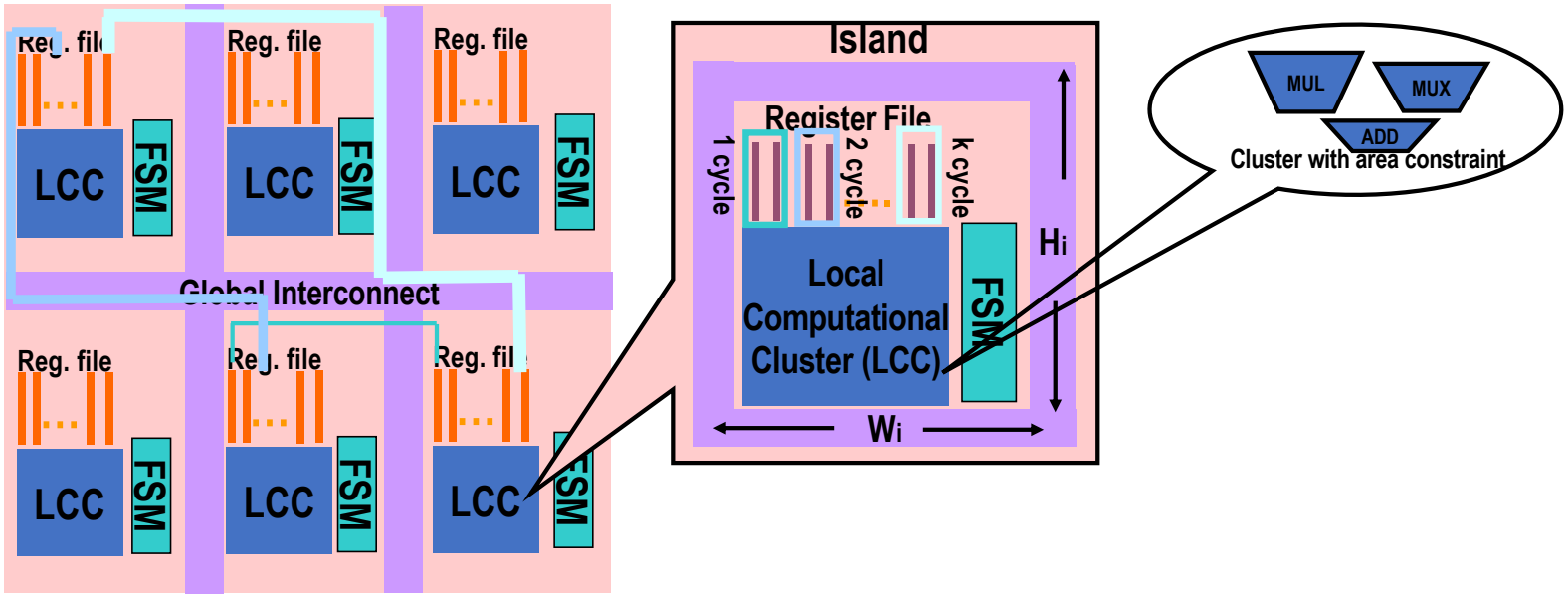
- Single-cycle full chip synchronization is no longer possible



- ITRS'01 70nm Tech
- 5.63 G Hz across-chip clock
- 800 mm^2 (28.3mm x 28.3mm)
- IPEM BIWS estimations
  - Buffer size: 100x
  - Driver/receiver size: 100x
- From corner to corner:
  - at semi-global layer (Tier 3)
  - can travel up to 11.4mm in one cycle
  - need 5 clock cycles

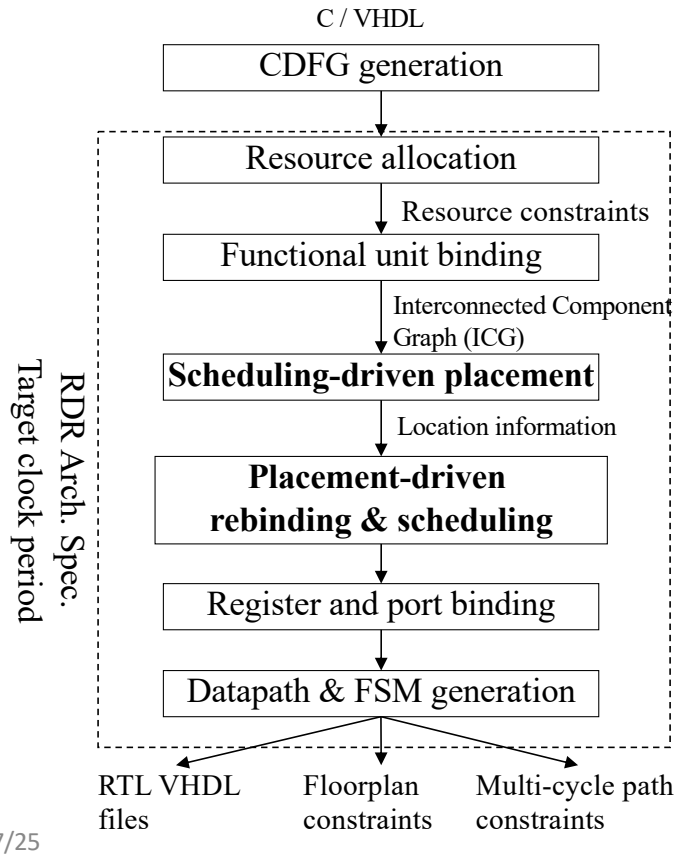
J. Cong, et al, T-CAD 2004

Our Proposal: Regular Distributed Register (RDR) Architecture [T-CAD'2004]



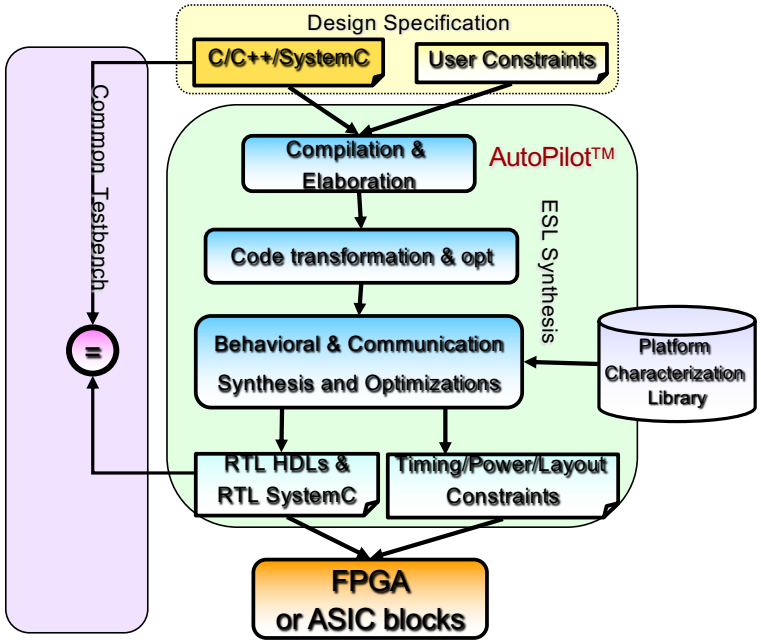
- Use register banks:
  - Registers in each island are partitioned to k banks for 1 cycle, 2 cycle, … k cycle interconnect communication in each island
- Highly regular
- Goal: high-frequency designs

MCAS: Placement-Driven Architectural Synthesis Using RDR Architecture [TCAD'04]



- **Multi-Cycle Communication Architectural Synthesis (MCAS) System**
  - **Scheduling-driven placement**
  - **Placement-driven rescheduling & rebinding**
- **Limitations**
  - **The high-level synthesis (HLS) engine was not robust**
  - **Did not consider interconnect pipelining**
  - **Regularity imposes some overhead**
- **This idea has to wait for another 17 years to mature**

Development of Robust and Scalable HLS Tool



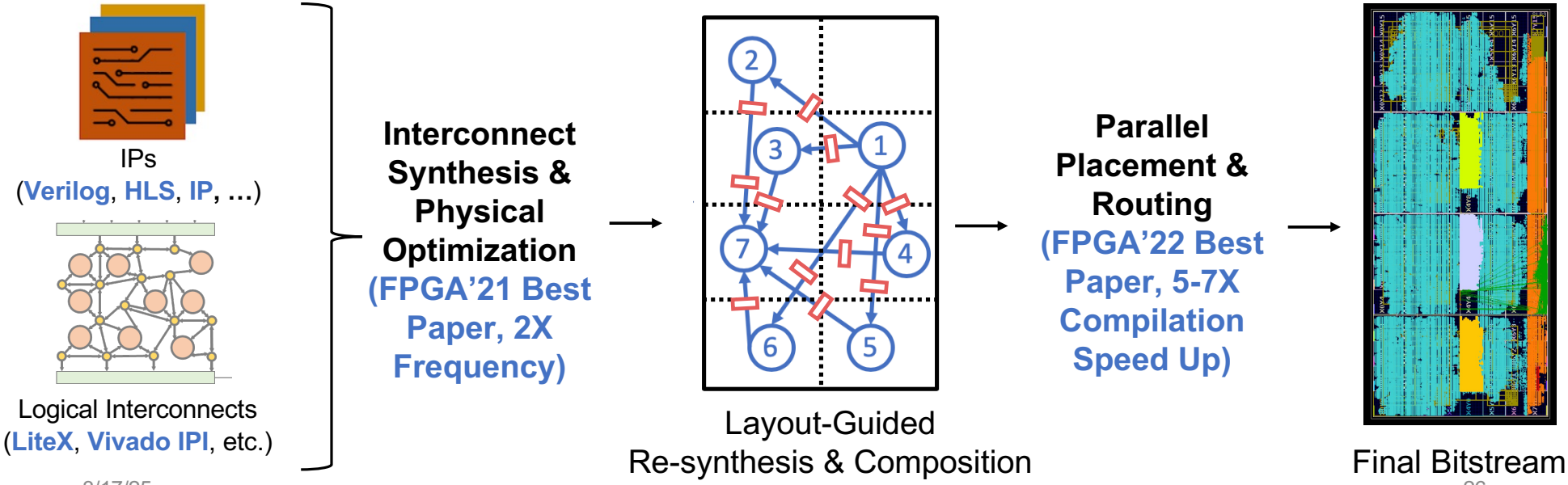
- xPilot (UCLA 2006) -> AutoPilot (AutoESL) -> Vivado HLS (Xilinx 2011-)
  - LLVM based compilation
  - Platform-based C to RTL synthesis
  - Synthesize pure ANSI-C and C++, GCC-compatible compilation flow leveraging LLVM framework
  - Full support of IEEE-754 floating point data types & operations
  - Efficiently handle bit-accurate fixed-point arithmetic
  - SDC-based scheduling
  - Automatic memory partitioning
- …

QoR matches or exceeds manual RTL for many designs

TCAD April 2011 (keynote paper) “High-Level Synthesis for FPGAs: From Prototyping to Deployment”

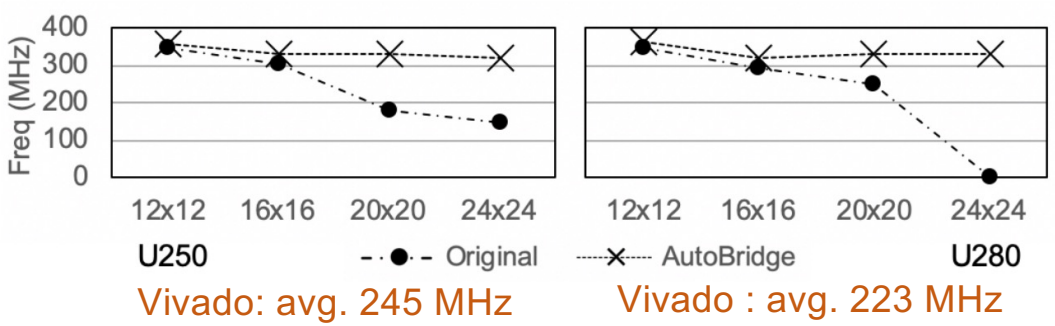
HLS with Automated Interconnect Pipelining

- HLS modules or IP blocks are connected in the dataflow style
- Automated co-optimization of interconnect pipelining, floorplan, and HLS => 2X Fmax
- Parallel placement & routing => 5-7X productivity

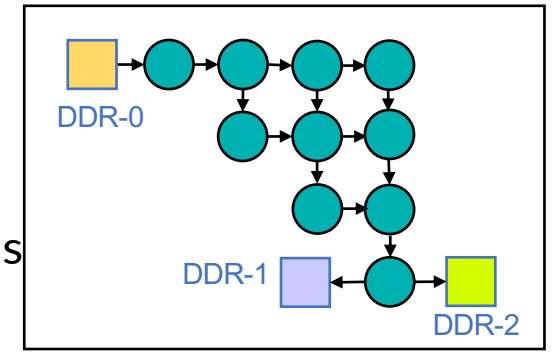


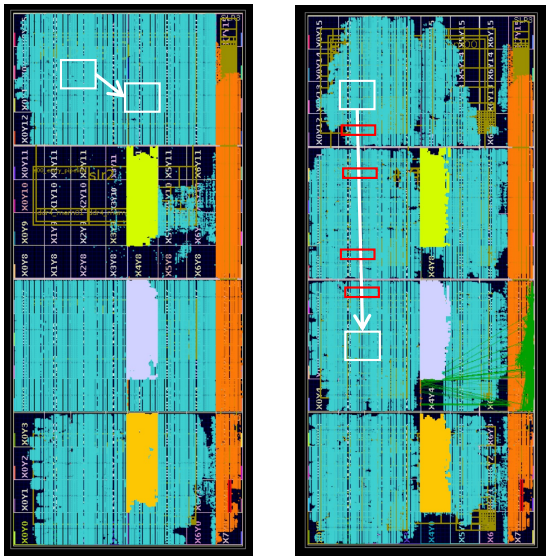
A Case Study

- Systolic array for Gaussian elimination, 8 configurations



- Difference in Resource Utilization
  - LUT: +0.14%
  - FF: +0.04%
  - BRAM: +0.03%
  - DSP: +0.00%





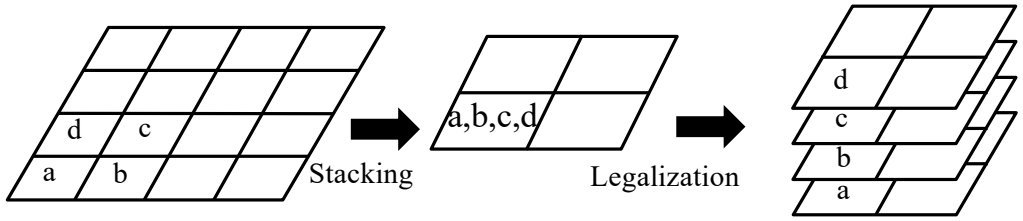
Vivado AutoBridge

Comparison of the 24x24 Design on U250

Approach 4: Novel Technologies for Interconnects

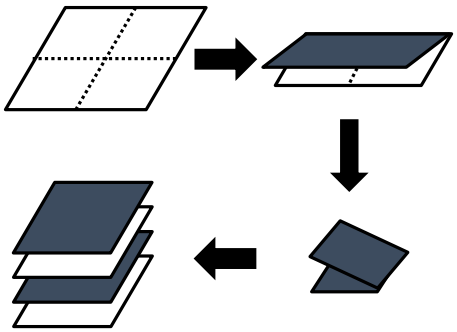
(Early 2000s – present)

Exploration 1: 3D-IC Designs (Early 2000s)

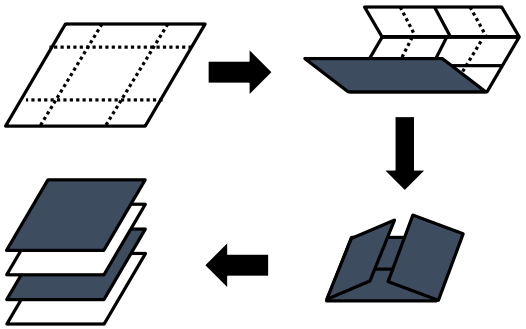


Local stacking transformation

Achieve 2x wirelength reduction Compared with 2D mPL5 with four device layer



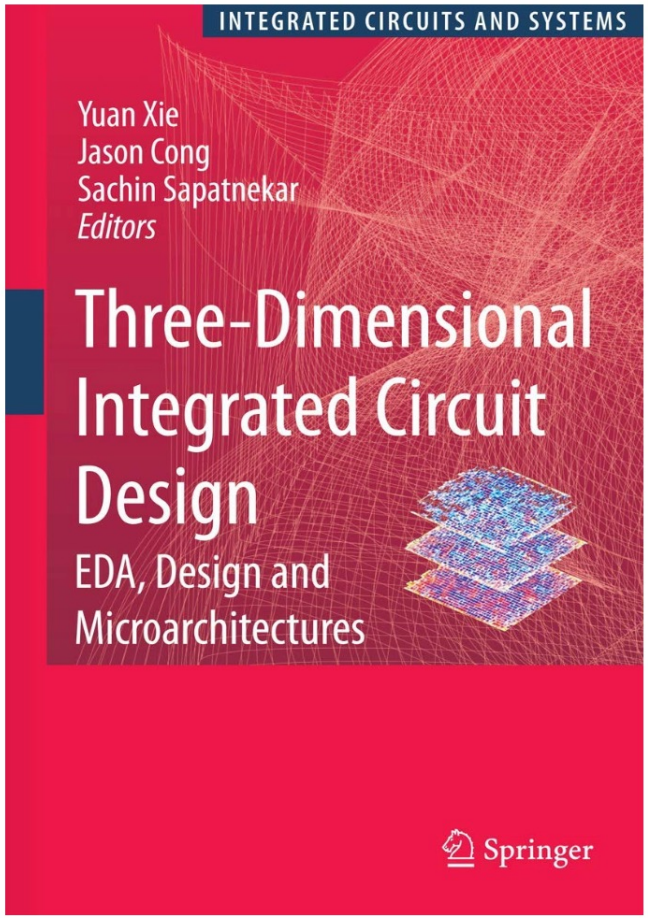
Folding-2 transformation



Folding-4 transformation

J. Cong, G. Luo, J. Wei, and Y. Zhang. Thermal-Aware 3D IC Placement via Transformation. ASP-DAC 2007, Yokohama, Japan, **(The ASP-DAC 2017 Ten-Year Retrospective Most Influential Paper Award).**

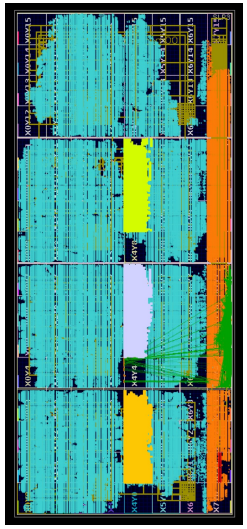
Many Studies on 3D Physical Design and Architecture Exploration



Springer Publishers, 2009.

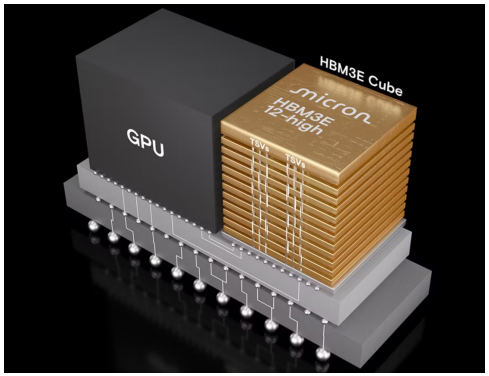
But Industry Adoption of 3D Technologies is Slow … Not until late 2010s (a Decade Later)

- Xilinx Multi-Die FPGA



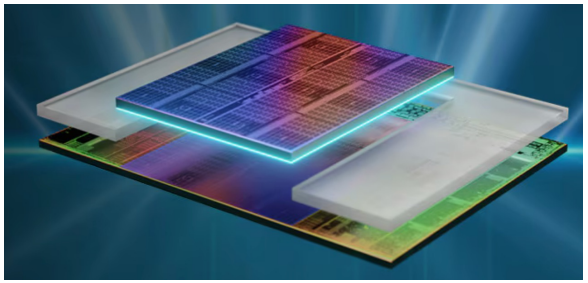
[Guo et al, FPGA'21]

- HBM
  - 3D-stackedsynchronous dynamic random-access memory
  - 12 layers



From MICRON website:
https://www.micron.com/products/memory/hbm/

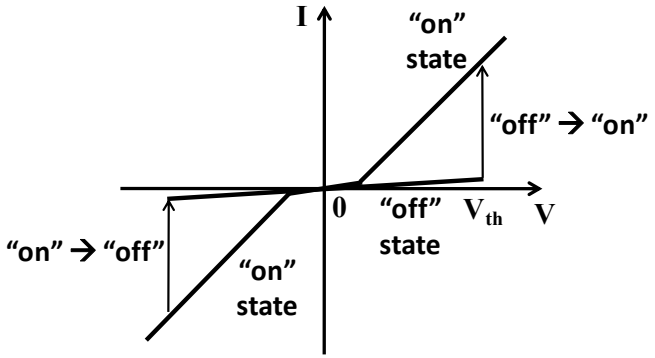
- AMD EPYC 9004 Series Processors with AMD 3D V-Cache™ Technology
  - 3D die stacking

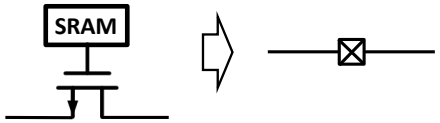


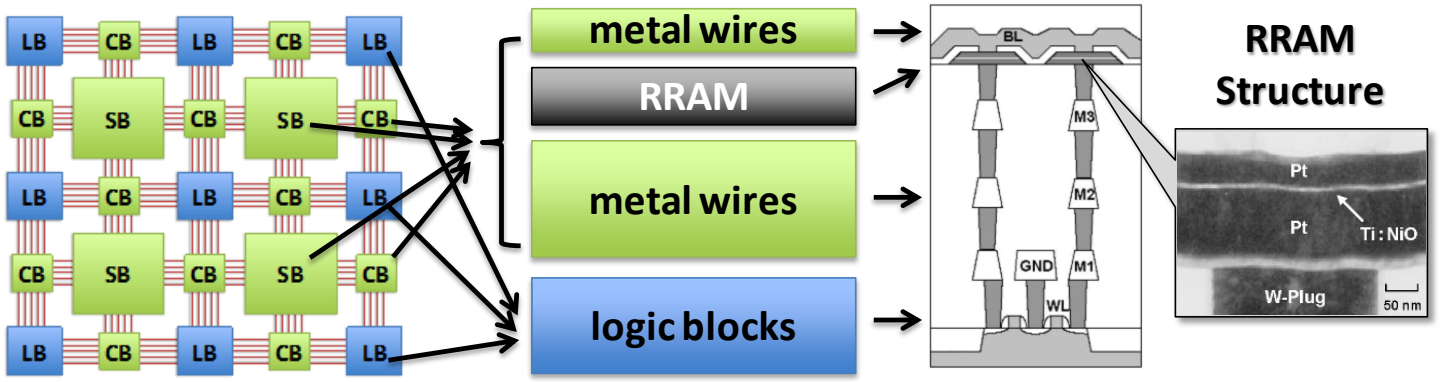
From AMD website:
https://www.amd.com/en/products/processors/technologies/3d-v-cache.html

Our Effort to Promote Use of 3D Technologies: RRAM-Based FPGAs with 3D Stacking [T-VLSI'13]

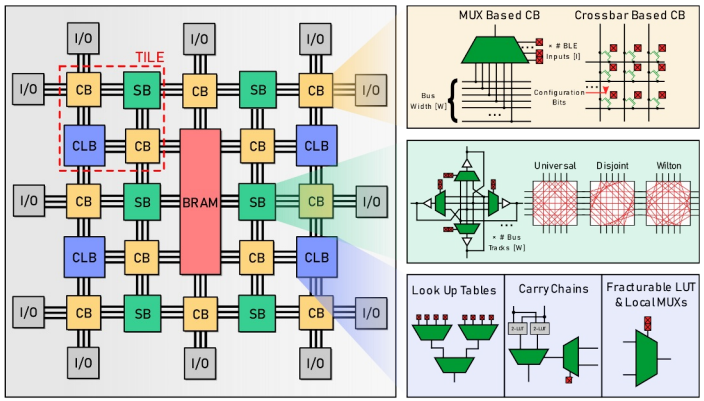
- Use of resistive RAM (RRAM)
  - Functioning as a routing switch
  - May build up programmable interconnects
- Stack RRAM over CMOS
- Built a physical prototype with automated P&R
- >50% savings in area, latency and power compared to pure CMOS design

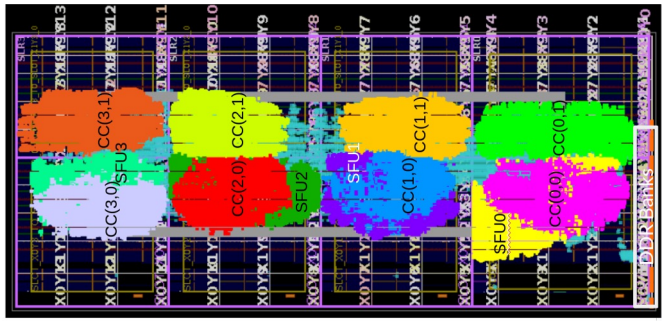






LLM on 3D FPGA with Back-End-of-Line Transistors [DAC'25]

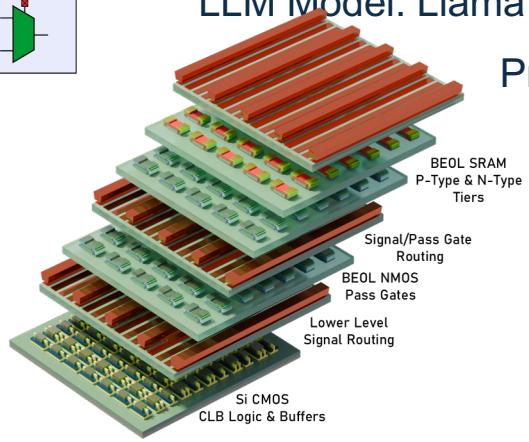




System	Versal VPK180
Area	736.0 mm^2
Delay	4.17 ns
AT^2	12798.9 $mm^2 \times ns^2$

LLM Model: Llama 3.1 8B on 2D FPGA [FCCM'25]

	Si AT^2	M3D AT^2
CONV	15380.8	**3812.2**
FFT	224.7	**51.2**
GEMM	733.23	**187.7**



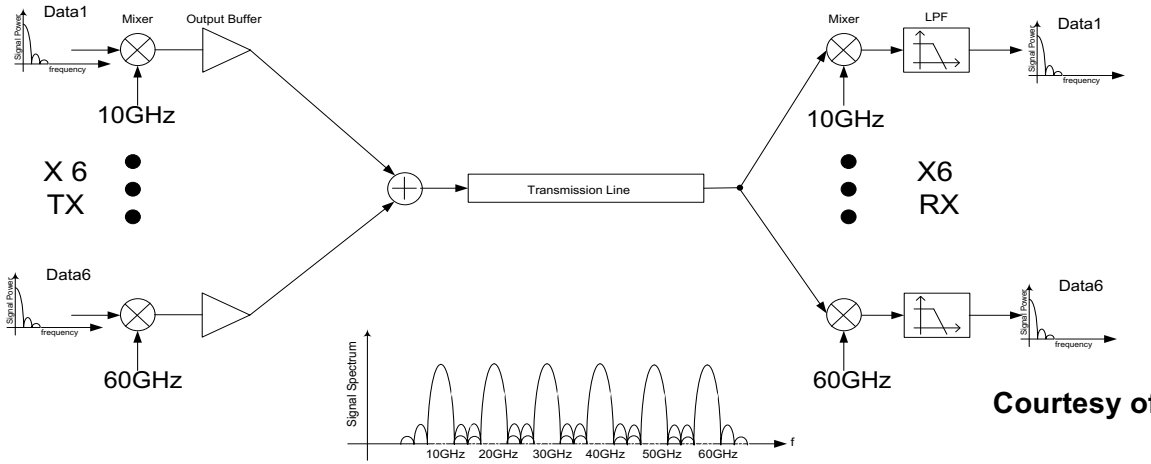
Projected M3D design for LLM

System	M3D FPGA
Area	423.8 mm^2
Delay	3.31 ns
AT^2	4643.2 $mm^2 \times ns^2$

2.76x Projected AT^2 Improvement

PIs: Shimeng Yu, Jason Cong
Students: Faaiq Waqar, Jiahao Zhang, Zifan He

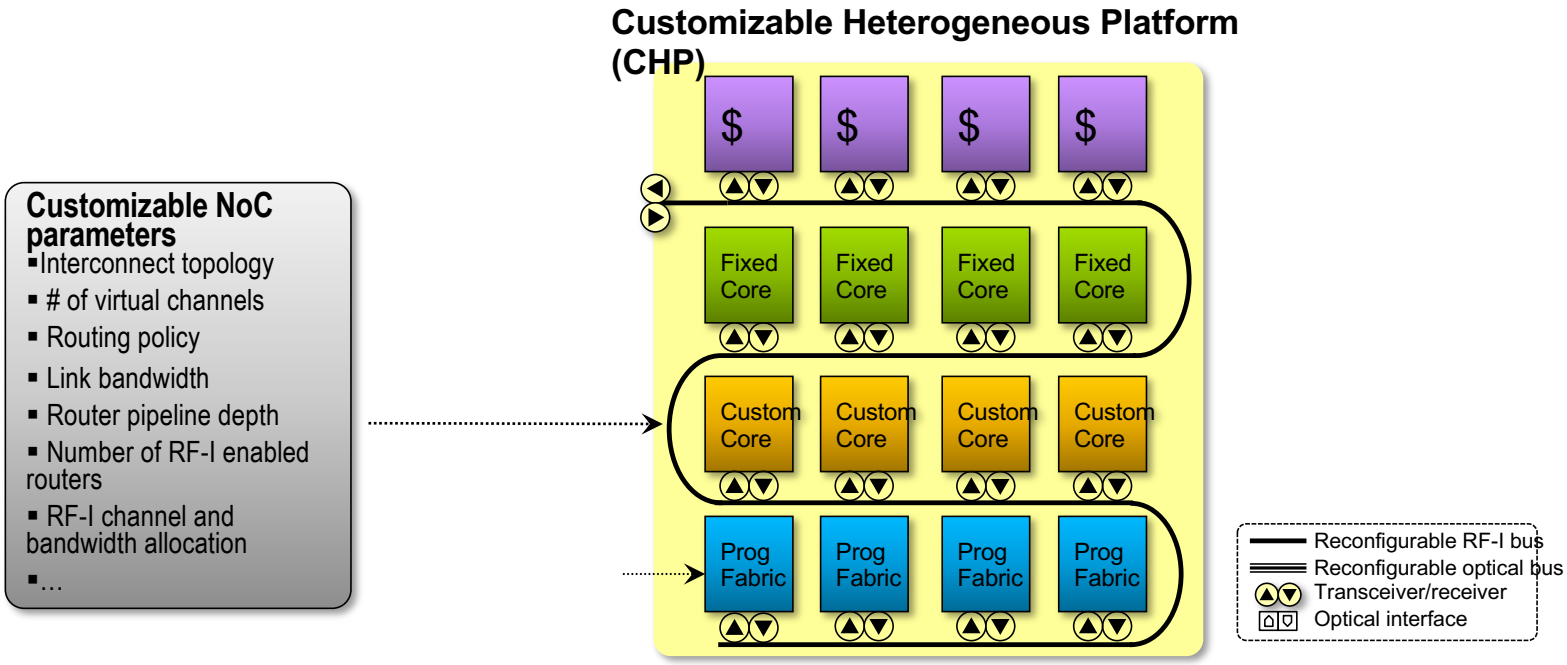
Exploration 2: Use of On-Chip Multiband RF-Interconnects



Courtesy of Frank Chang

- **In TX, each mixer up-converts individual baseband streams into specific frequency band (or channel)**
- **N different data streams (N=6 in figure above) may transmit simultaneously on the shared transmission medium to achieve higher aggregate data rates**
- **In RX, individual signals are down-converted by mixer, and recovered after low-pass filter**

RF-Interconnects in Customizable Heterogeneous Platform (CHP) [Expeditions in Computing Proposal, 2009]

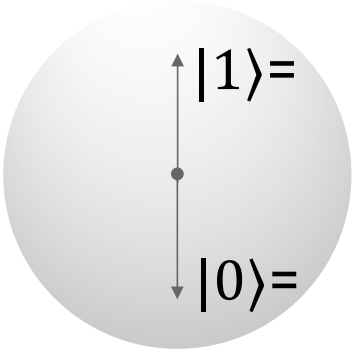


Approach 5: From Interconnect to Entanglement

(Late 2010s – present)

Can "spooky action at a distance" bypass interconnects?

A Simple Example of Quantum Circuits

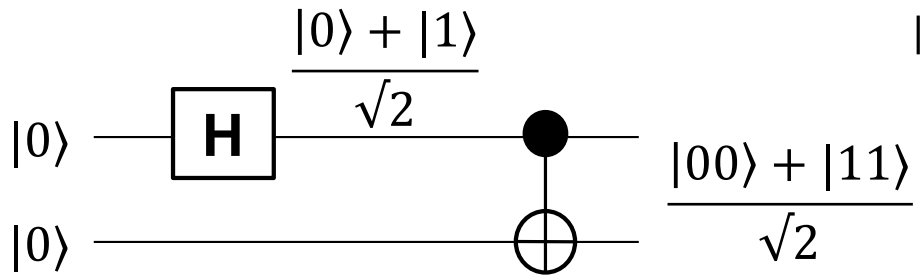


A qubit can be visualized as a point on the Bloch sphere.

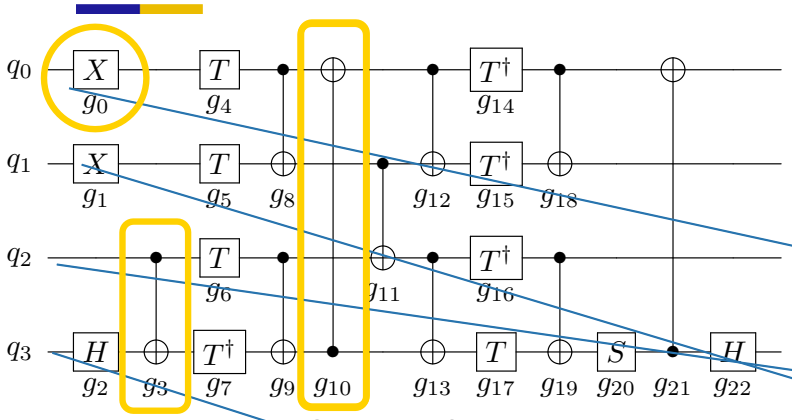
I.e., any unit vector in the Hilbert space spanned by basis vectors |0> and |1>.

- Hadamard gate $H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$
- CNOT:
  - First qubit: control, second qubit: target

$$|00\rangle \rightarrow |00\rangle$$
$$|01\rangle \rightarrow |01\rangle$$
$$|10\rangle \rightarrow |11\rangle$$
$$|11\rangle \rightarrow |10\rangle$$



Quantum Layout Synthesis (QLS)



Input Circuit of Adder

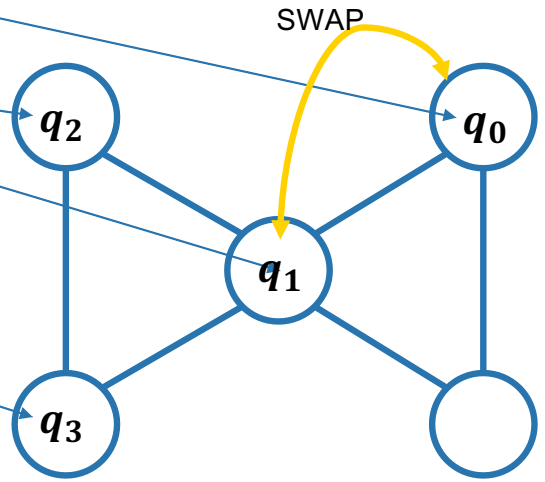
```
# Input quantum program
x q[0];
x q[1];
h q[3];
cx q[2], q[3];
t q[0];
…
```

† means Hermitian conjugate, which is straightforward once we have the original gate implementation.

CX on a pair of adjacent qubits, OK.

CX on a pair of non-adjacent qubits!

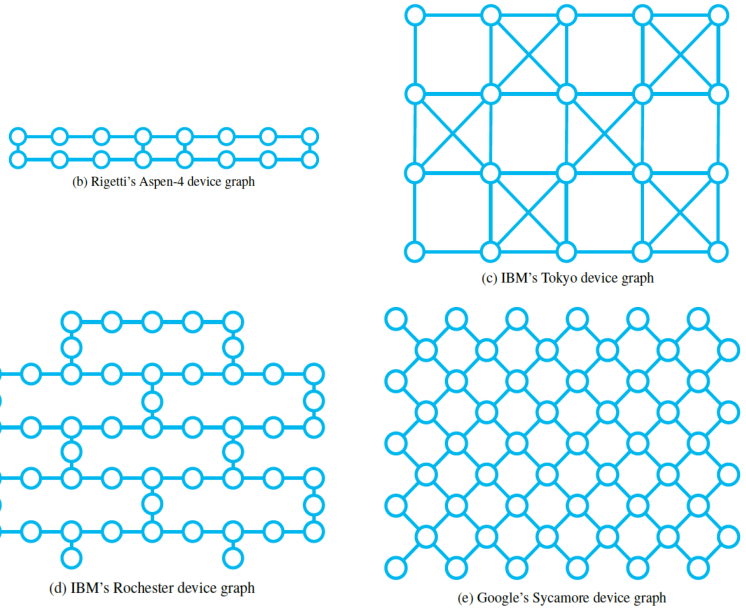
Insert SWAP gate to change the mapping



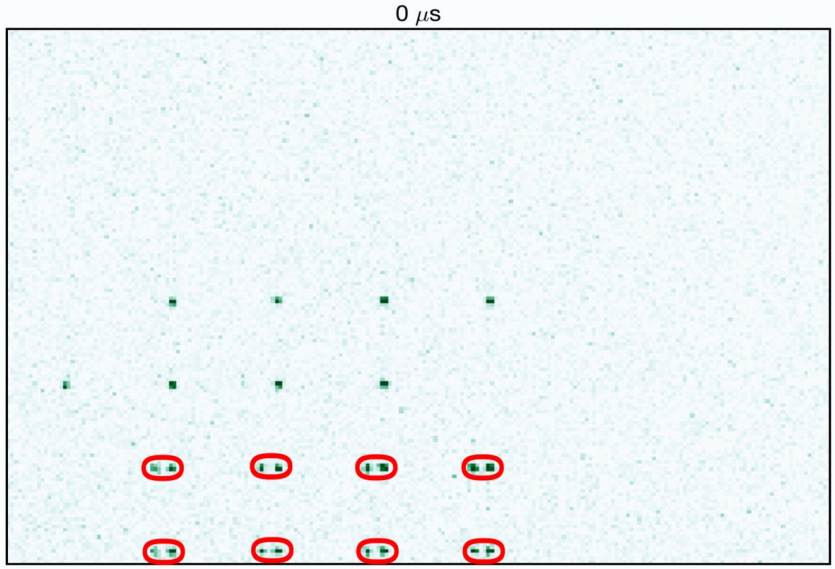
Coupling Graph of IBM QX 2

Need to Support Diverse QC Platforms

- Superconducting devices
  - With different topologies



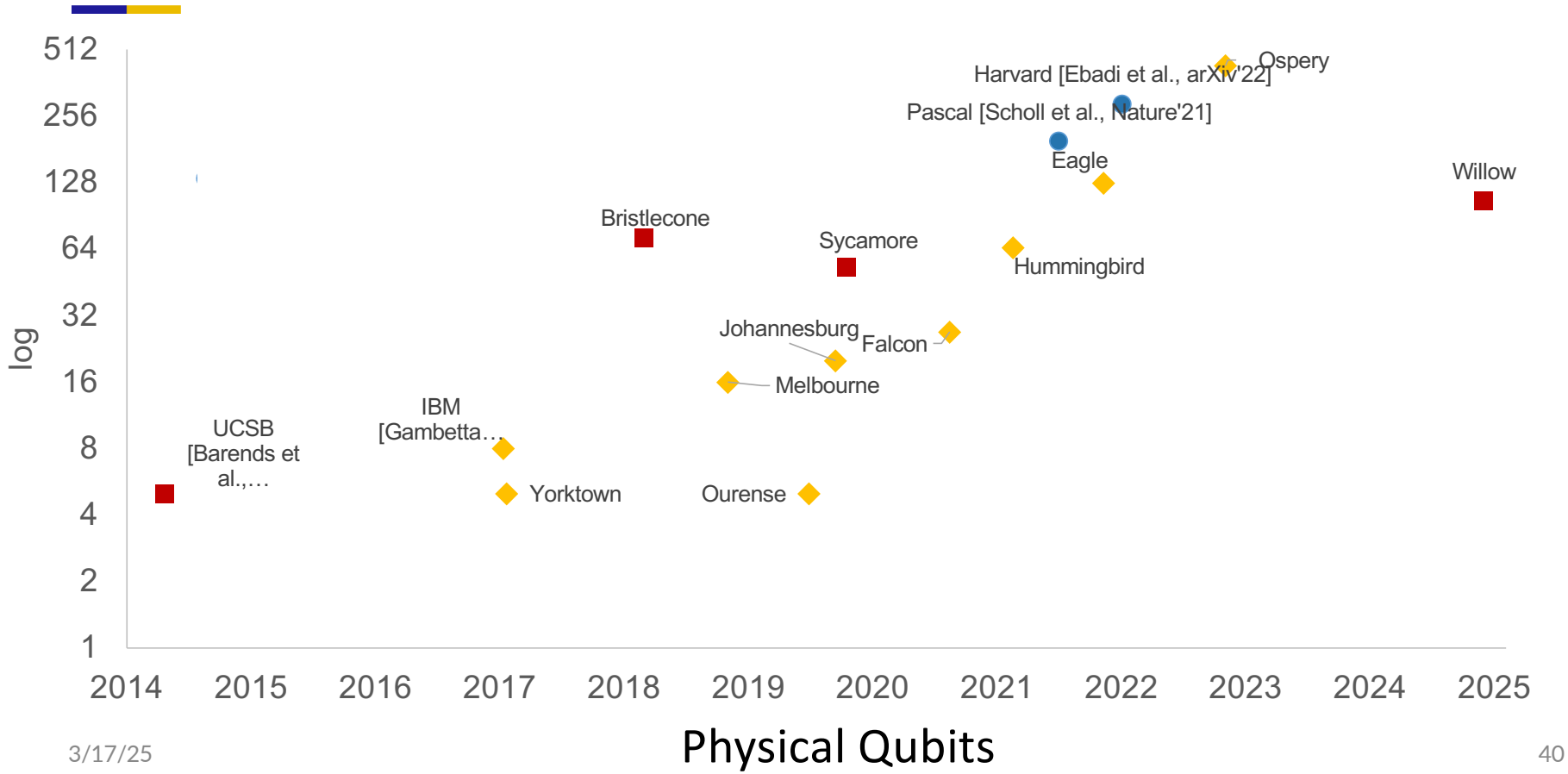
- Neutral Atom devices
  - Qubit movement



[Bluvstein et al., Nature'22]

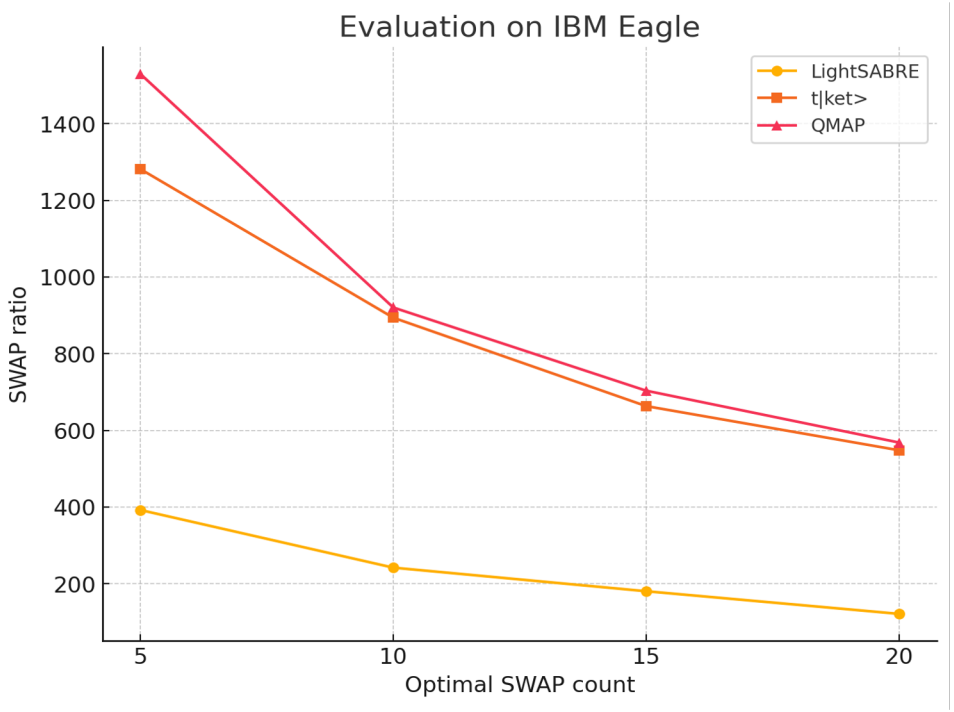
Both types can be supported using the SMT formulation: OLSQ2 [DAC'23] and OLSQ-DRAA [ICCAD'22]

Recent Advancement of Quantum Computing Technologies

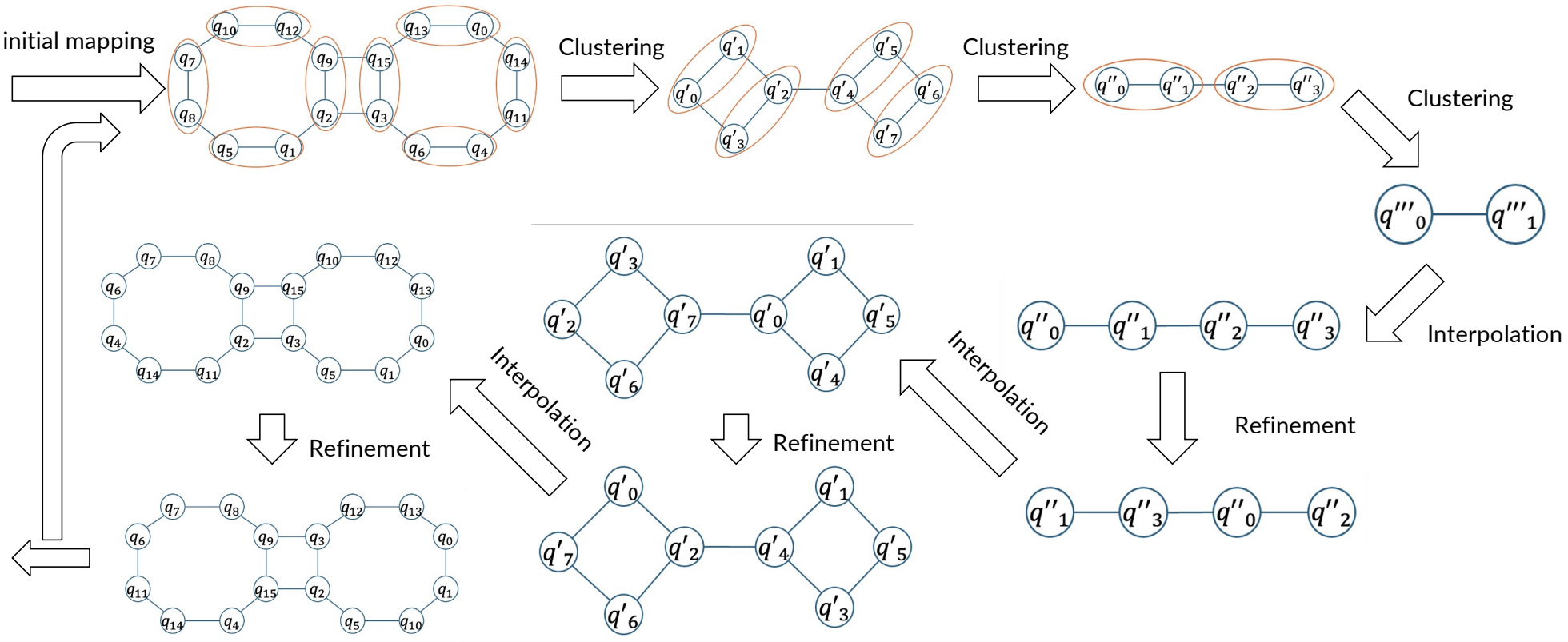


Surprise! Even at the Current Scale, Existing QLS Tools are Far Away from Optimal! [DAC'25]

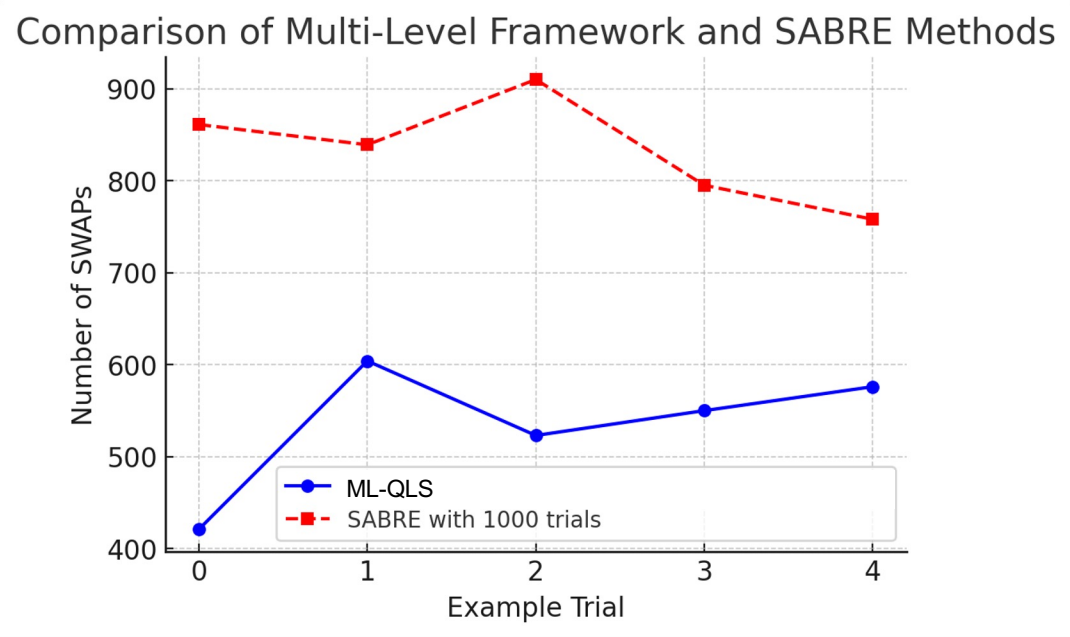
- Evaluated on IBM Eagle architecture (127 qubits)
- Circuit size: 3000 2Q gates
- Average optimality gap: 233.97x
  - SABRE is adopted in IBM Qiskit



Multi-level framework for QLS



Result from Multi-level QLS



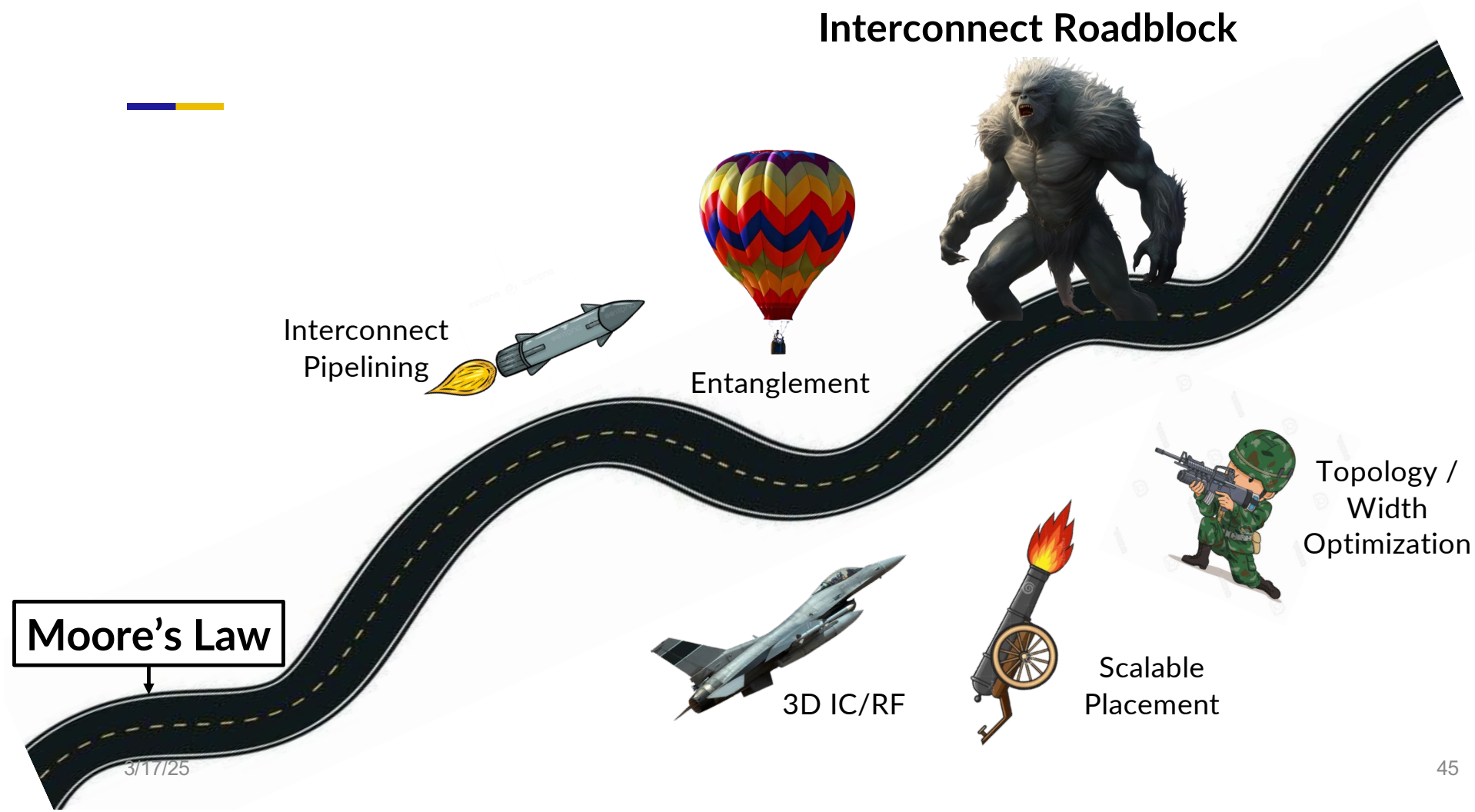
- 36% reduction in optimality gap
- Still a lot of room for improvement

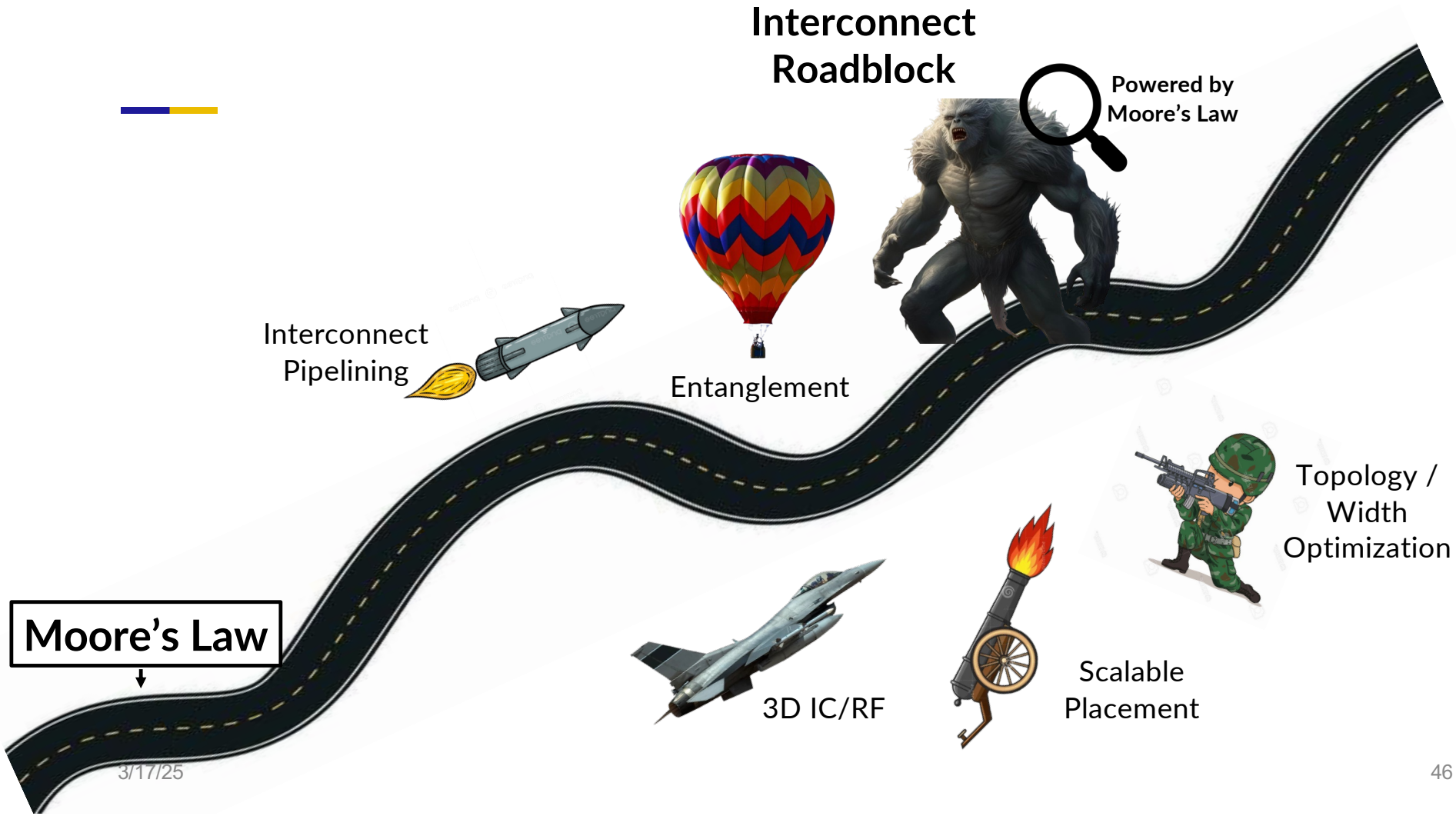
Some Concluding Remarks and Lessons Learned

- Use the right models to guide optimization
  - *Trade-off between accuracy and complexity: e.g. Elmore delay model*
- Device and CAD co-optimization
  - E.g. exploration of 3D architectures
- Space and time co-optimization
  - **"Time and space are modes in which we think, not in which we exist"**

    -- Albert Einstein
- Be patient – some technology adoption takes time (e.g. 3D IC)
- Physical design is valuable to build a digital twin or cousin for concept validation
  - E.g. 3D-stacked RRAM-based FPGAs
- Choose a problem of long-lasting impact

Interconnect Roadblock




Interconnect Roadblock
Powered by Moore's Law
Interconnect Pipelining
Entanglement
Topology / Width Optimization
Moore's Law
3D IC/RF
Scalable Placement

Many Interconnect Challenges Ahead

A serious concern:

- Copper is difficult to further scale
  - First introduced at 0.25um
- Copper mean free path = 40nm
- Resistivity of a 10nm wide Cu line > 4 x Resistivity of bulk material
- Need to search for new materials

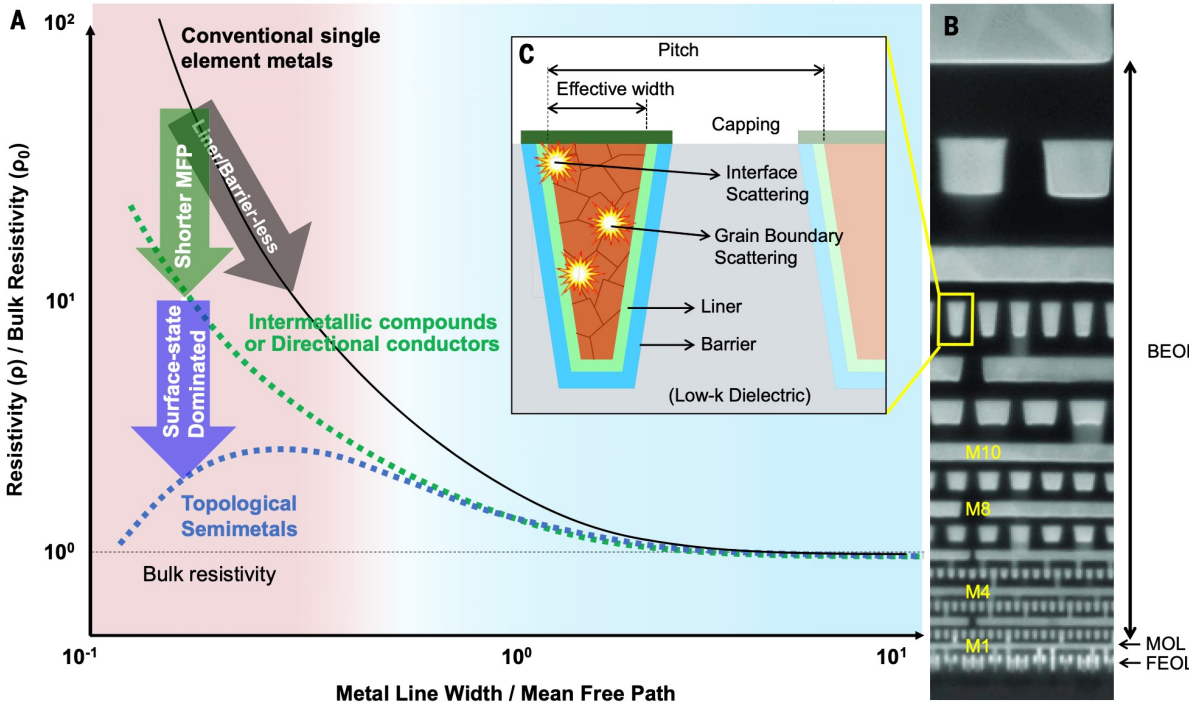


Fig. 1. Origin of increased resistivity. An exponential increase in electron scattering in metal lines is attributed to the abrupt increase in metal resistance.
Source: Kim et al., Science 2024

Acknowledgement – VLSICAD/VAST Lab Members

- Interconnect design and optimization:


Jie Fang


Lei He

Kei-Yong Khoo


Cheng-Kok Koh

Hardy Leung


Patrick Madden


Takumi Okamoto


David Z Pan


Taku Uchino

- Scalable placement:


Chin-Chih Chang


Tianming Kong


Sung-Kyu Lim


Michail Romesis


Toshiyuki Shibuya


Joe Shinnerl


Kenton Sze


Xin Yuan


Min Xie

Acknowledgement – VLSICAD/VAST Lab Members

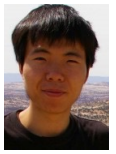
- 3D IC design and architecture:


Wei Jie

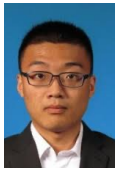

Ashok Jagannathan


Guojie Luo


Yan Zhang


Bingjun Xiao

- Space-time co-optimization with HLS:

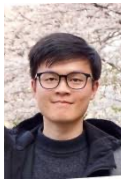

Yuze Chi


Yiping Fan

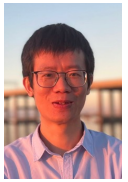

Licheng Guo

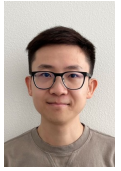

Guoling Han


Jerry Jiang


Jason Lau


Weikang Qiao


Linghao Song


Jie Wang

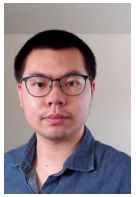

Zhiru Zhang


Peipei Zhou

- Quantum layout synthesis:


Jason Kimko


Wan-Hsuan Lin


Daniel Tan

More Acknowledgement – Many Collaborators

- *Interconnect topology optimization: Chuck Alpert, Andrew B. Kahng, Majid Sarrafazdeh, and C.K. Wong*
- *Scalable placement: Tony Chan, Xiaojian Yang, and Wei Chen*
- *RF-interconnect: Frank Chang and Glenn Reinman*
- *3D FPGA Prototypng: Shimeng Yu*
- *Interconnect pipelining with HLS:, Alireza Kaviani, Moazin Khatti, Chris Lavin, Pongstorn Maidee, Xingyu Tian, Ecenur Ustun, and Yun Zhou,*
- *Quantum layout synthesis: Dolev Bluvstein and Mikhail Lukin*
- ….

Acknowledgements; Funding Agencies & Industry Partners

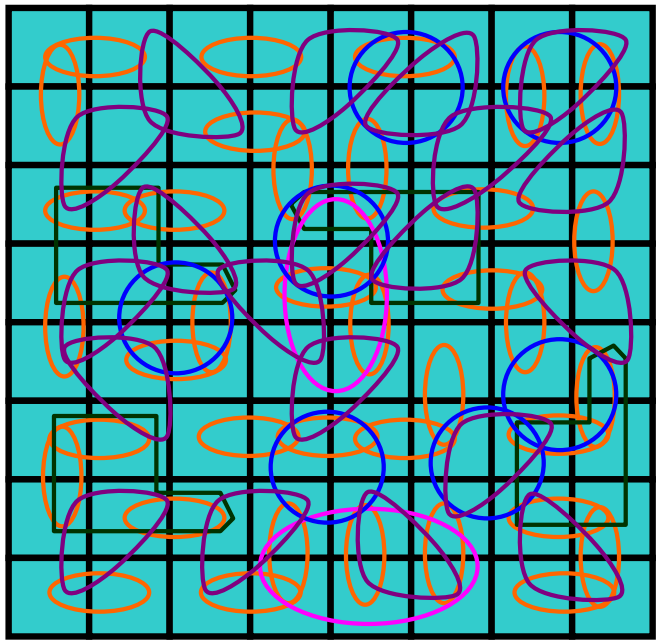
- Funding Agencies :

- MARCO/JUMP

- Many Industry Partners:

Thank You!

Placement Examples with Known Optimal (PEKO) Wirelength [Chang et al., TCAD'04]



- All the modules are of equal size
- For 2-pin nets, connect any two adjacent modules
- Net degree distributions extracted from real industrial benchmarks