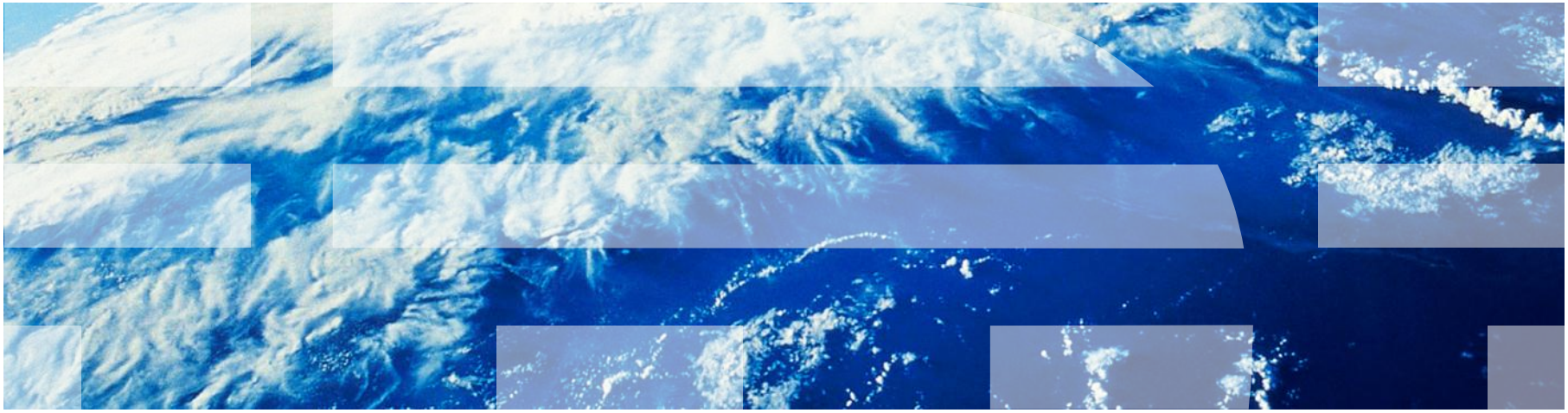
High Performance Microprocessor Design, and Automation:

Challenges and Opportunities

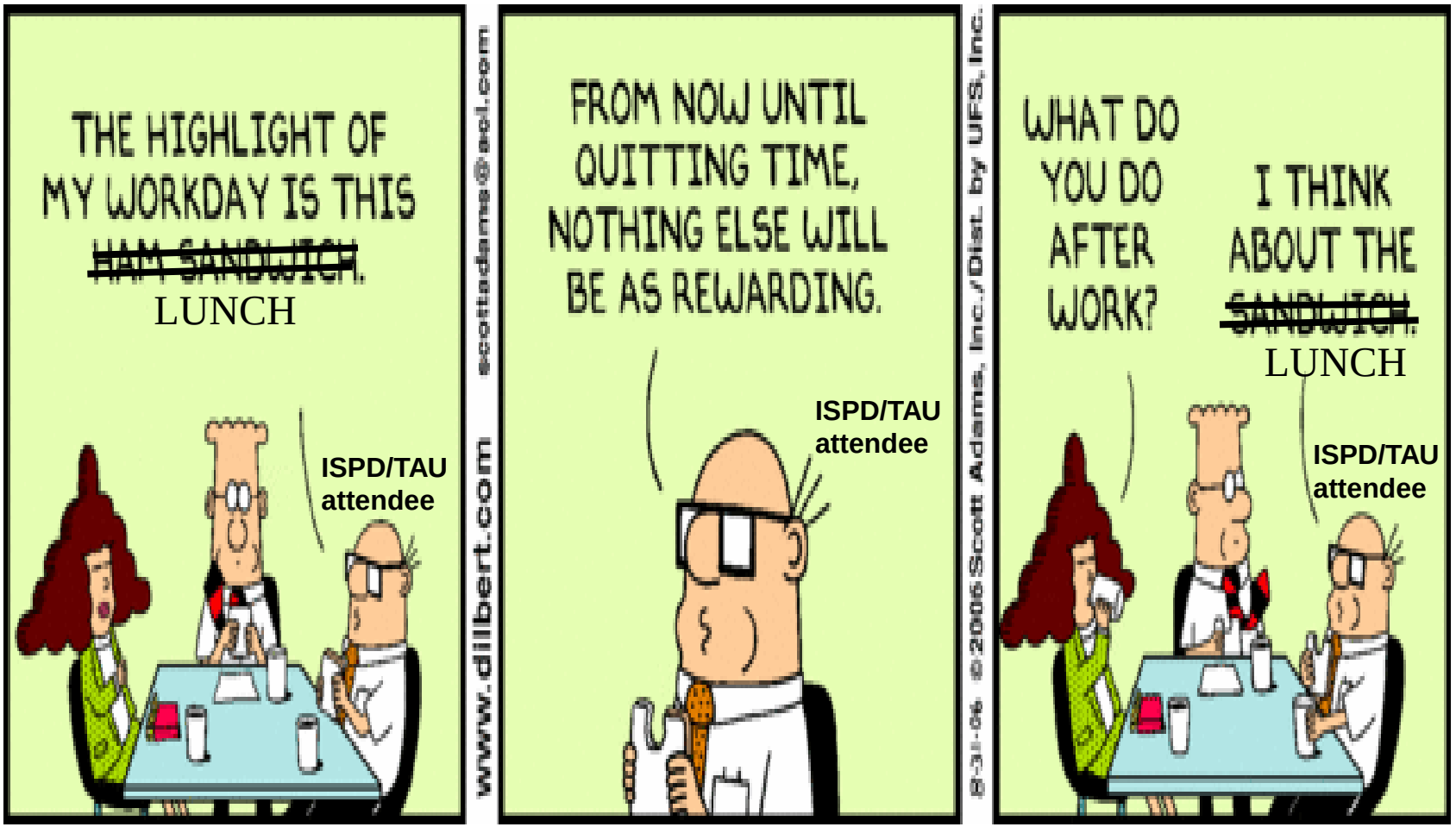
Ruchir Puri

IBM Fellow, VLSI Systems

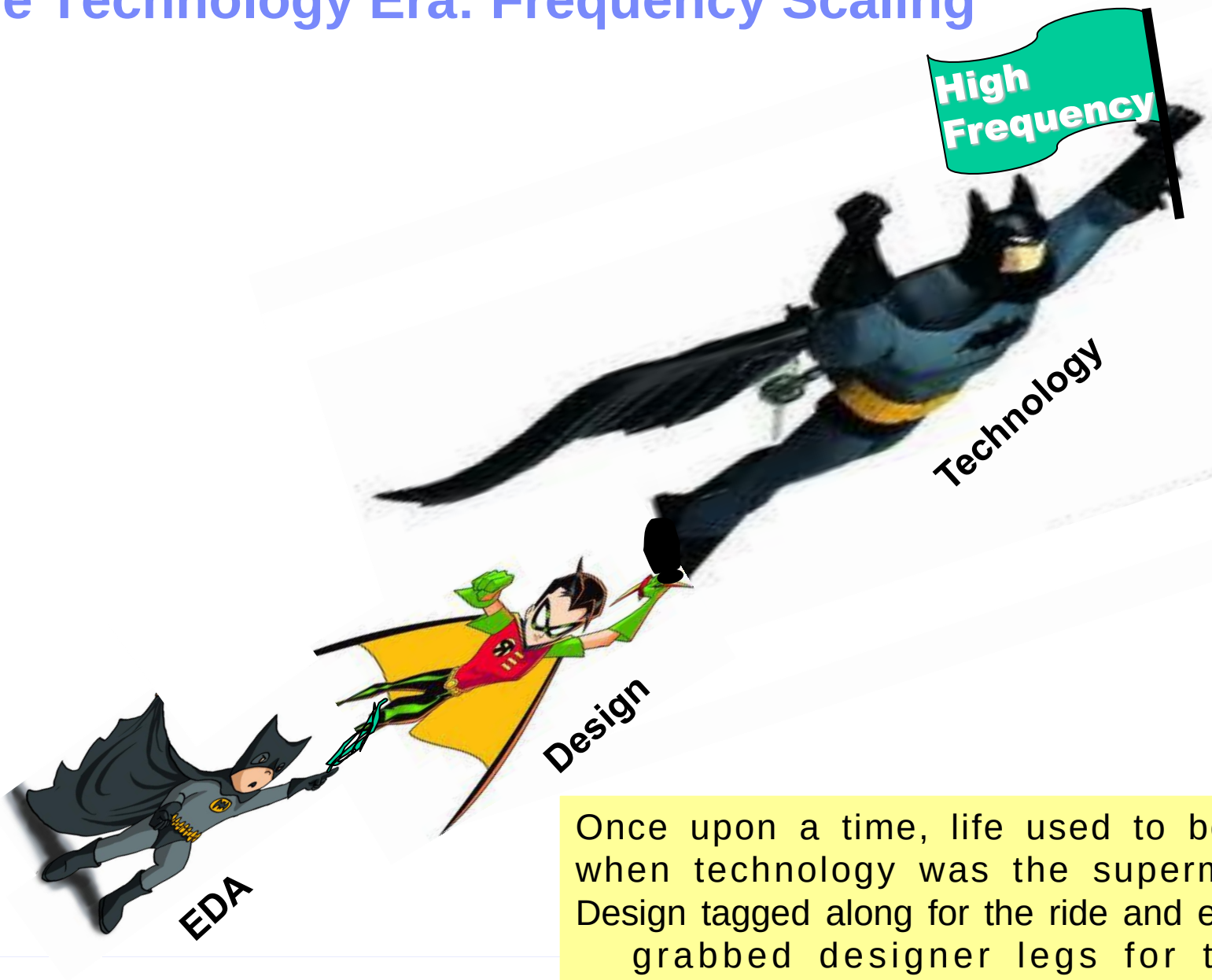
IBM Thomas J Watson Research Center, Yorktown Heights, NY



First, something about lunch and technical work…



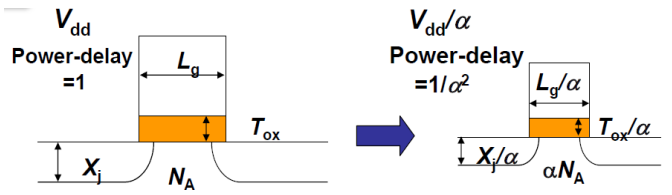
The Technology Era: Frequency Scaling



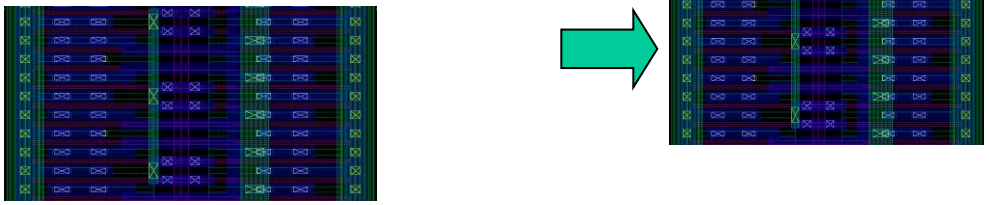
Once upon a time, life used to be Great, when technology was the superman and Design tagged along for the ride and even EDA grabbed designer legs for the fun!

Characteristics of Single Thread Era

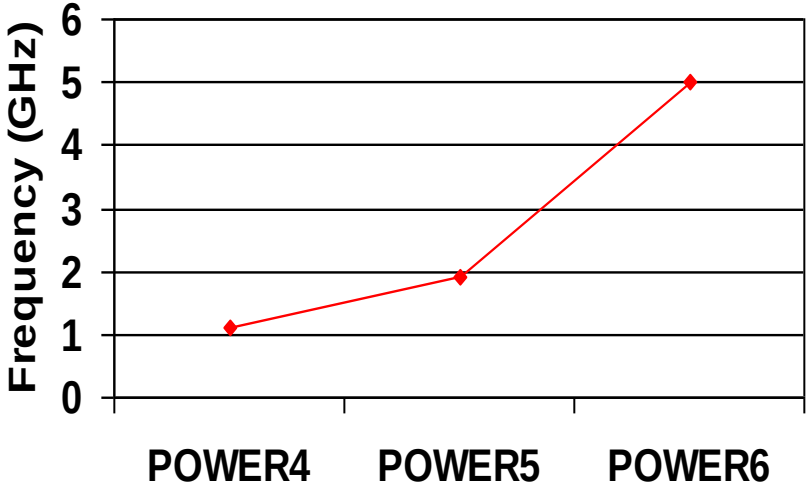
Dennard Scaling



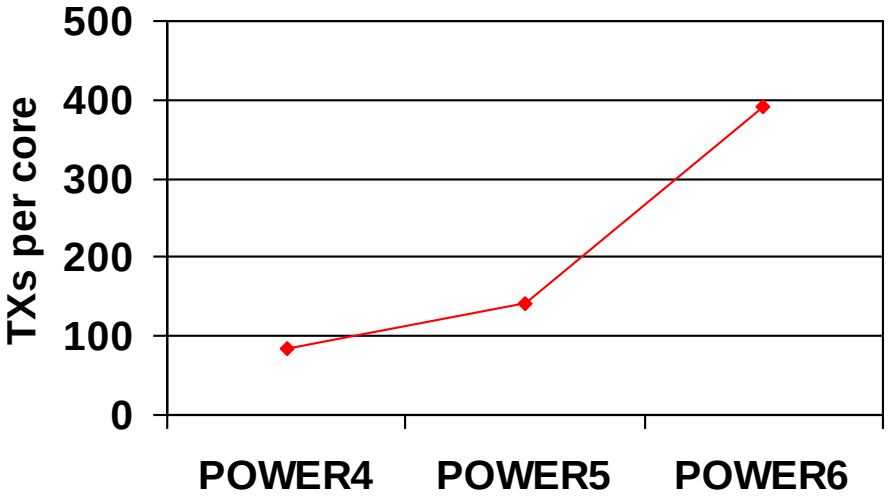
Optical Scaling / Node Migration



Exponential Frequency Growth

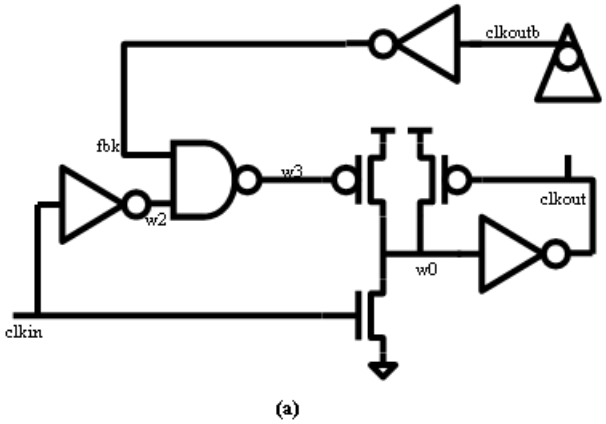


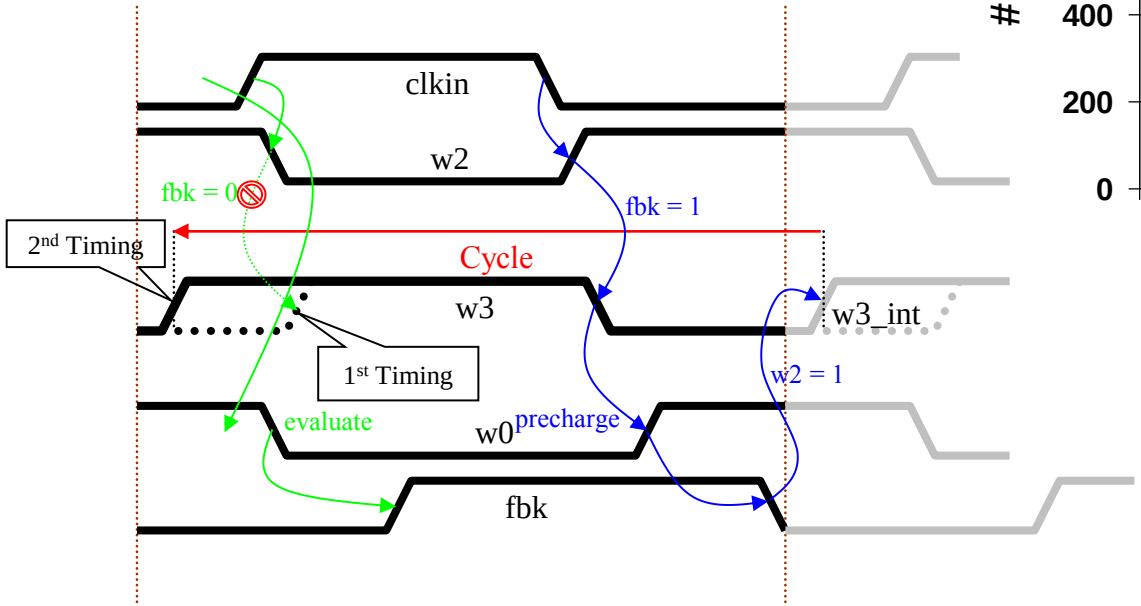
Expanding uArch Complexity



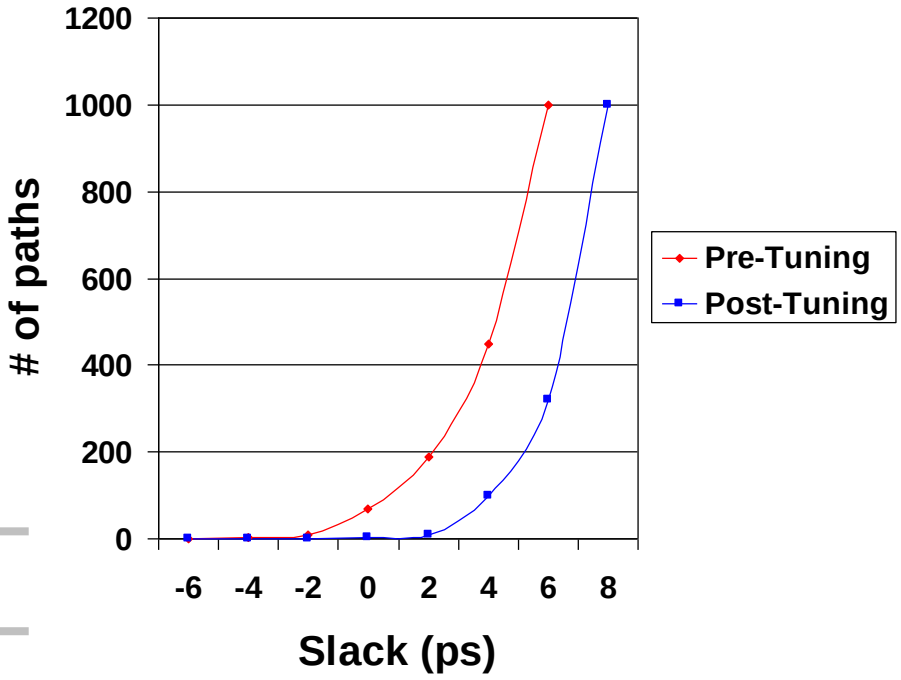
Single Thread Era EDA: Transistor Analysis & Optimization

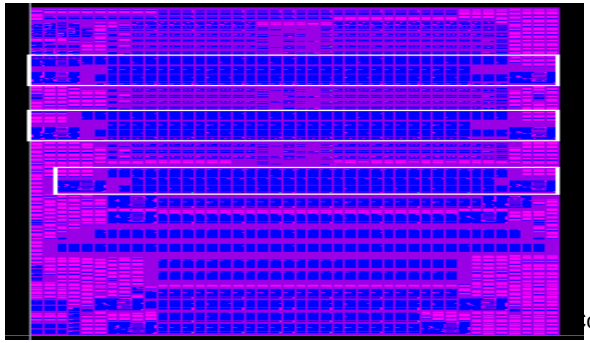
Static timing analysis of complex circuits



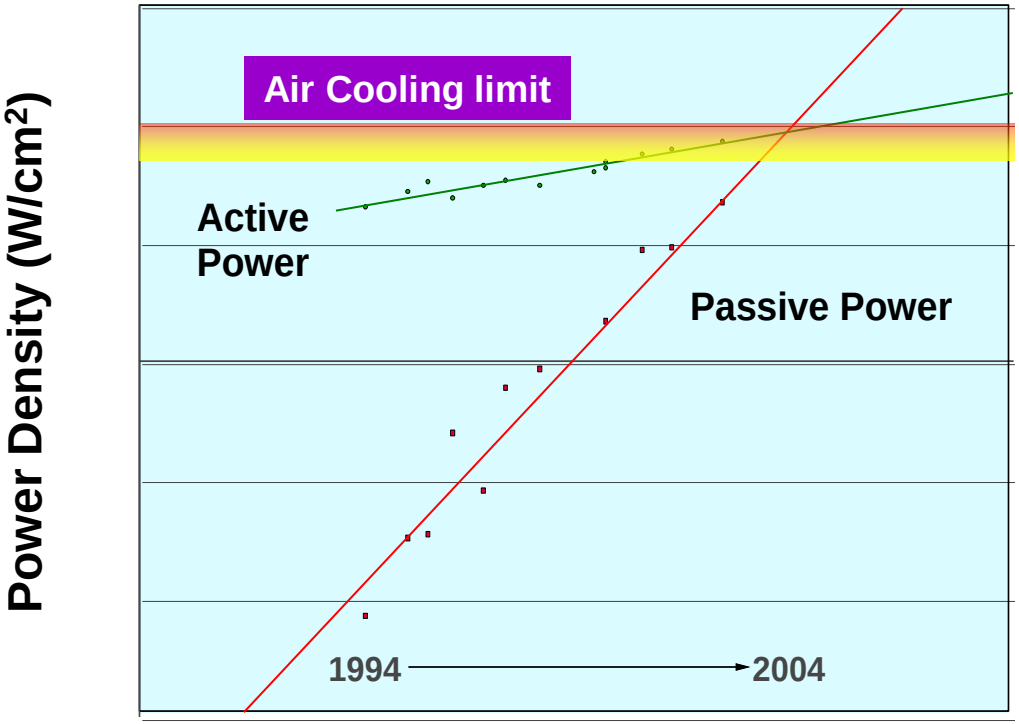


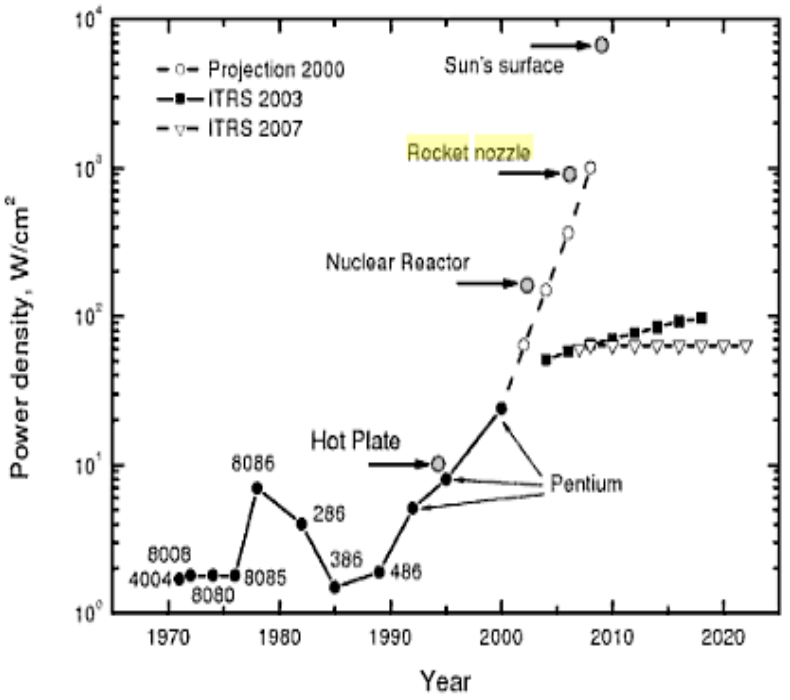
Transistor Level timing optimization





End of Frquency Scaling : The Power Wall

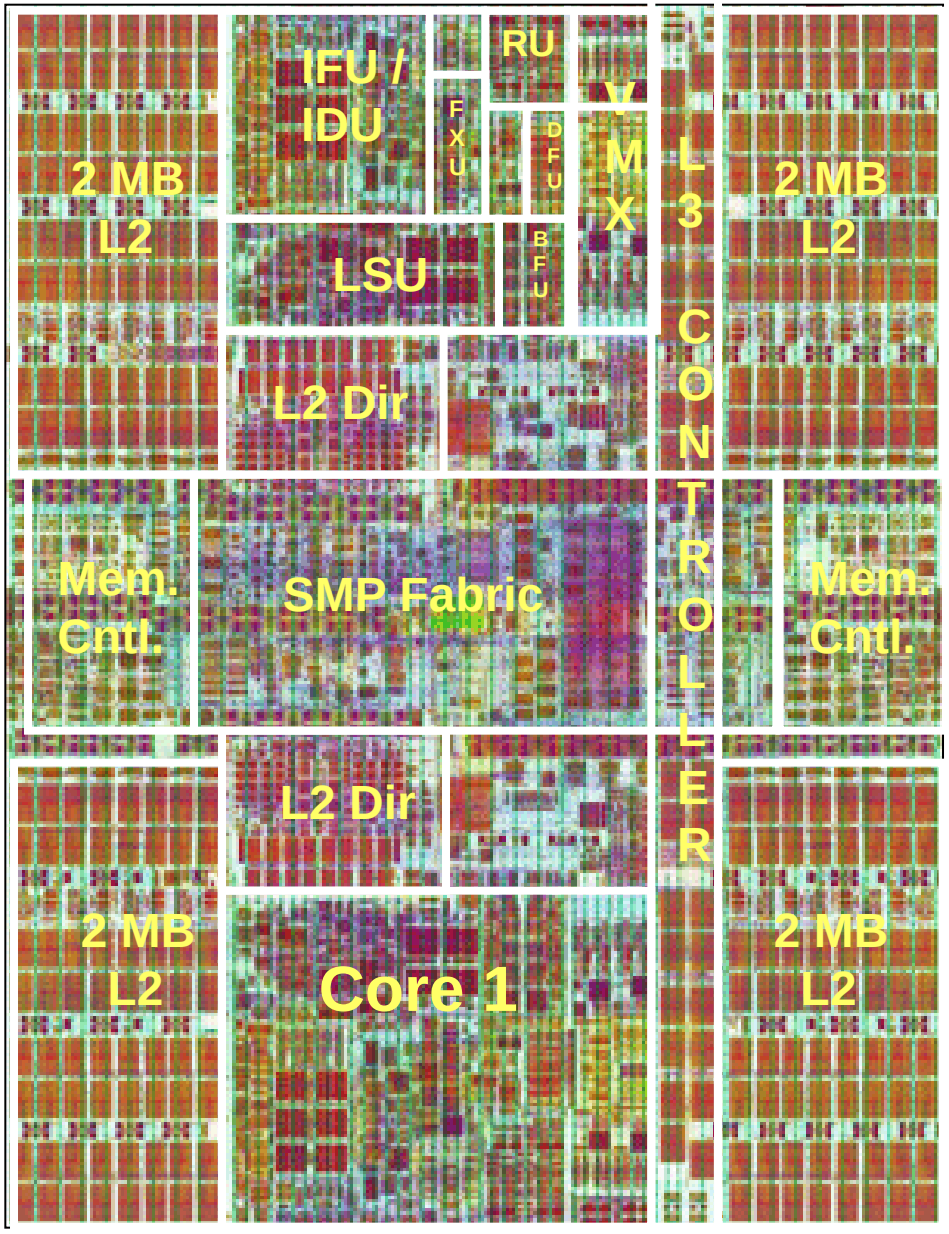




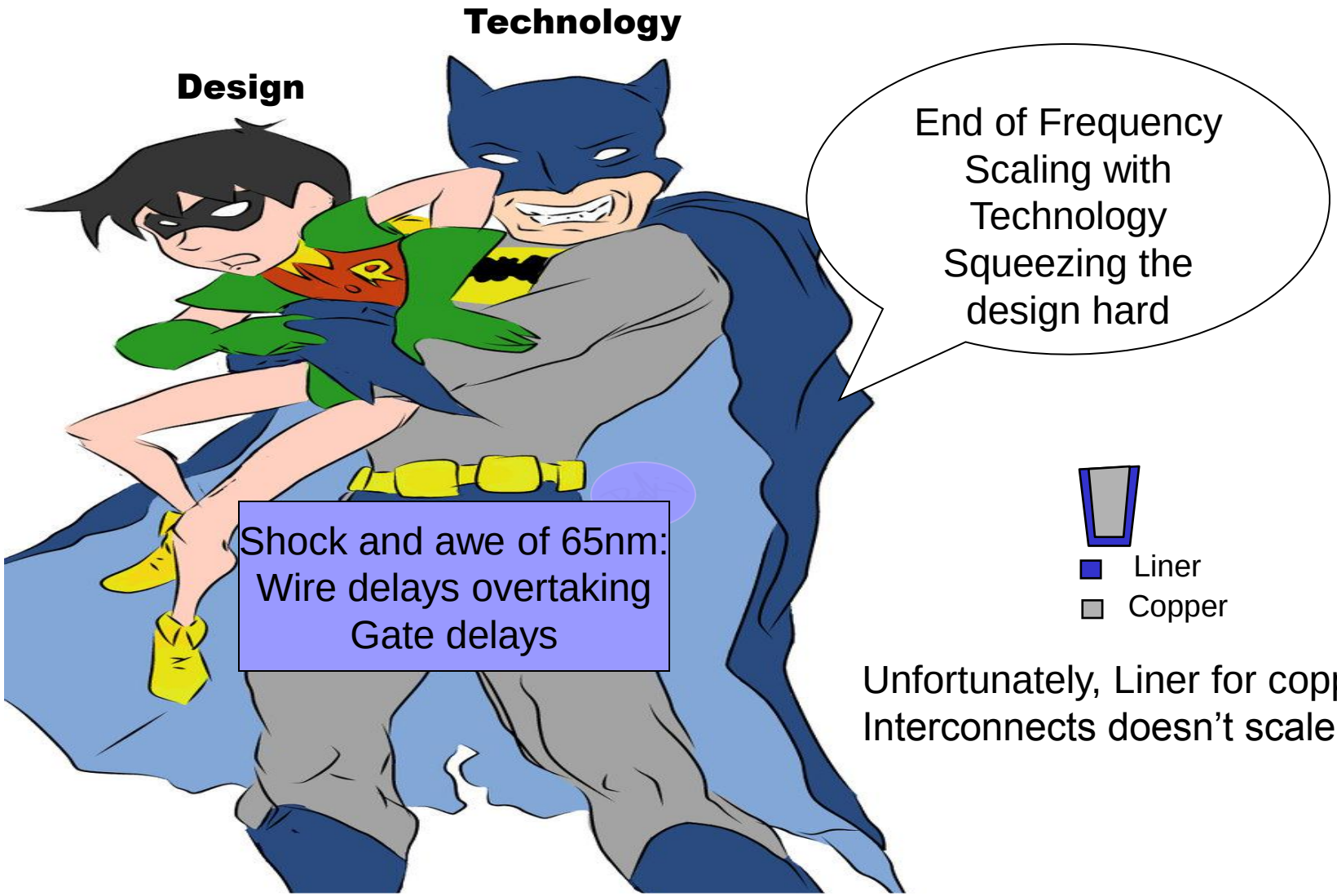
Inability to scale Oxide thickness & lower voltage resulted in a power wall for single thread performance

Frequency Scaling : POWER6 (65nm, 2007)

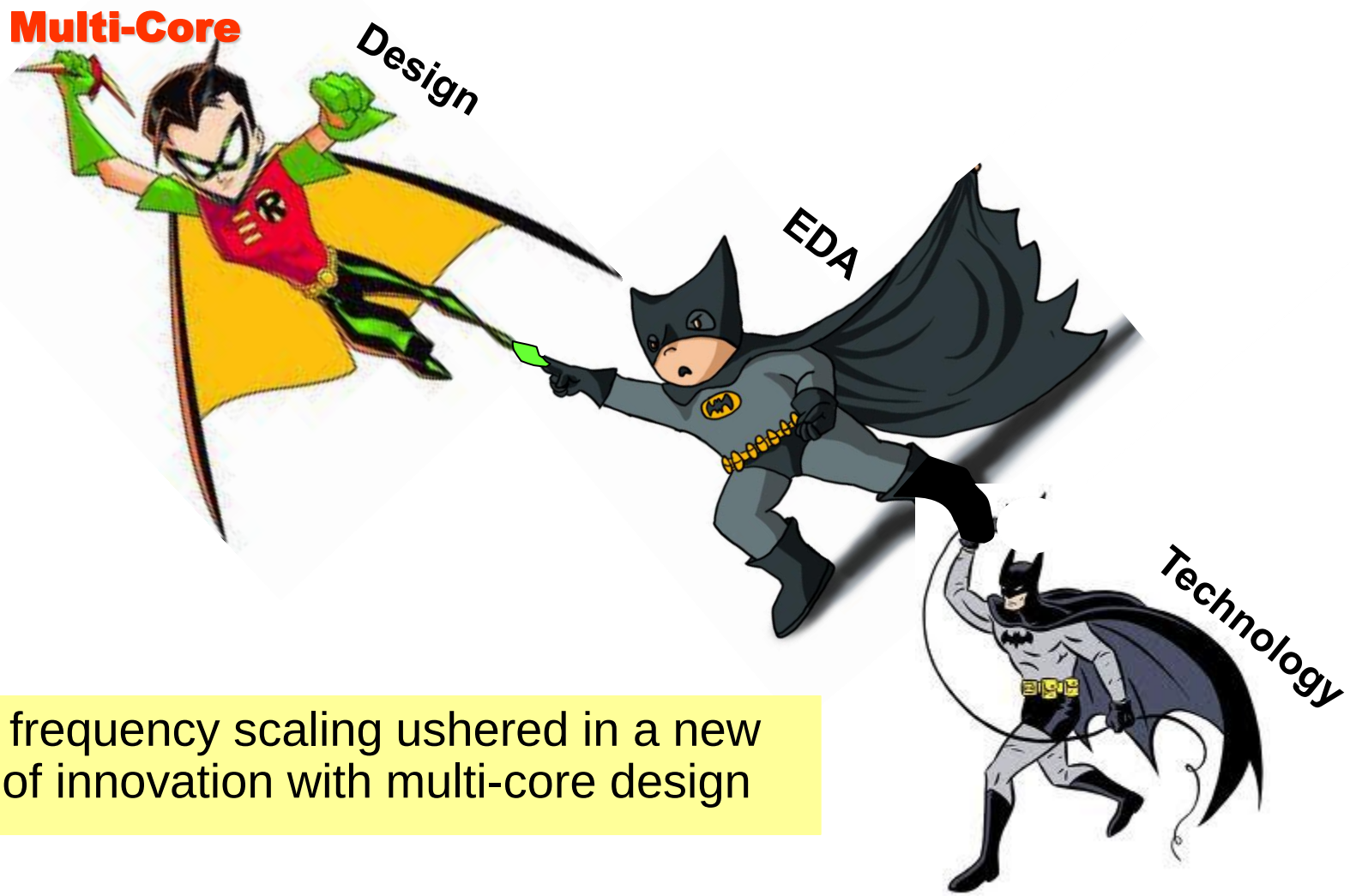
- 5+ GHz operation, >790M transistors, 341mm^2 die
- 65nm SOI with 10 levels of Cu interconnect
- Same pipeline depth & power @ 2x frequency versus POWER5



Technology Tantrums

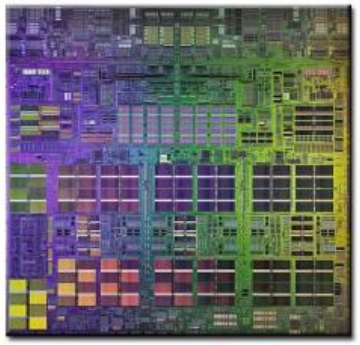


Multi-Core Era

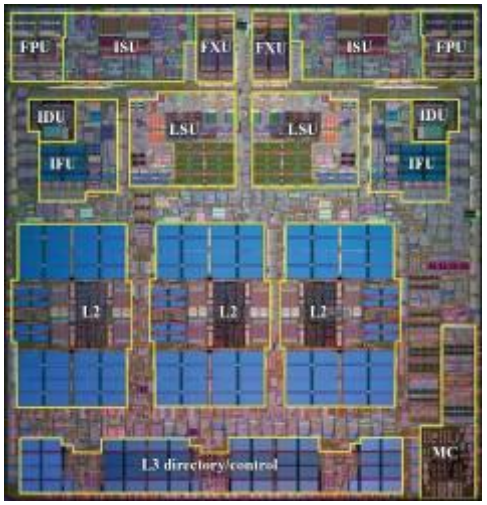


End of frequency scaling ushered in a new era of innovation with multi-core design

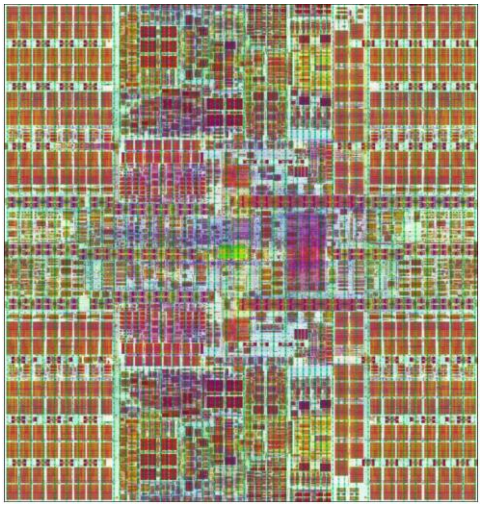
POWER Processors Began the Multi-Core / Multi-Thread Era



Power 4
2001
Introduced First Dual core

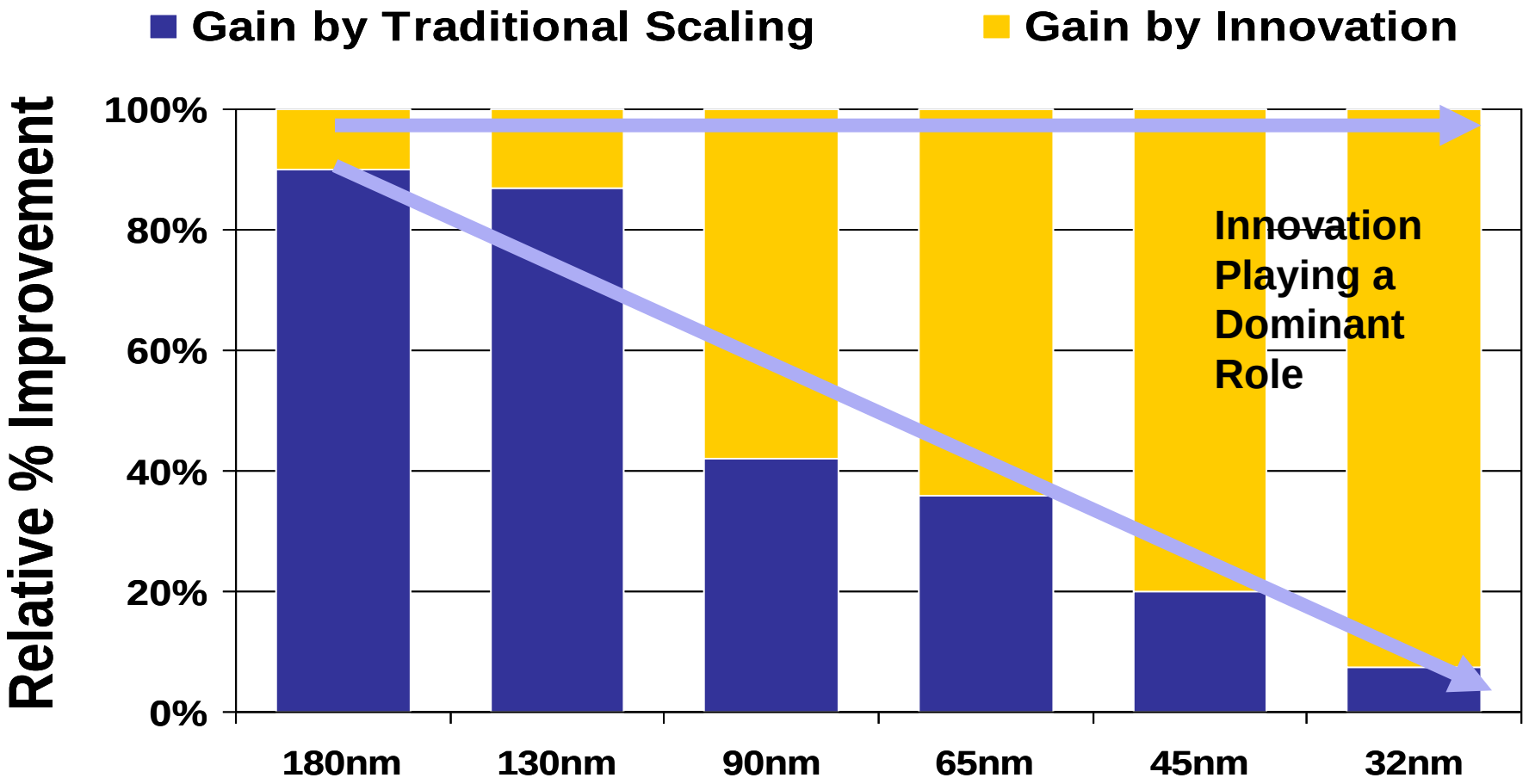


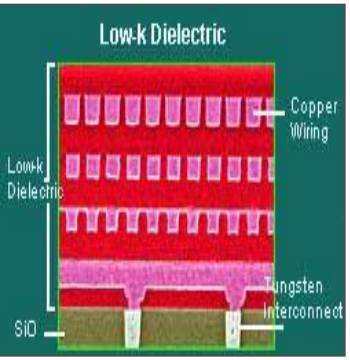
Power 5
2004
Dual Core
Introduces SMT (4 threads)

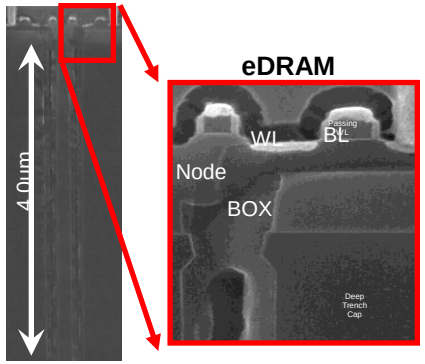


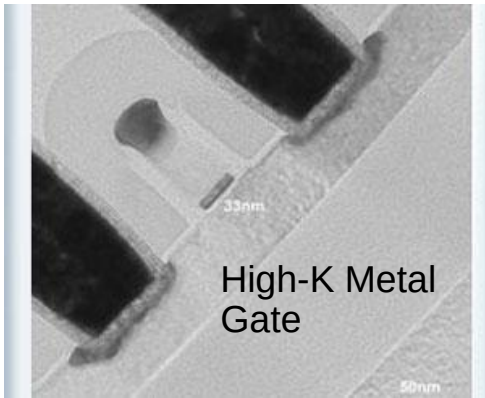
Power 6
2007
Dual Core – 4 threads
Enhances SMT Efficiency

Life starts to become interesting: Technology ride very bumpy



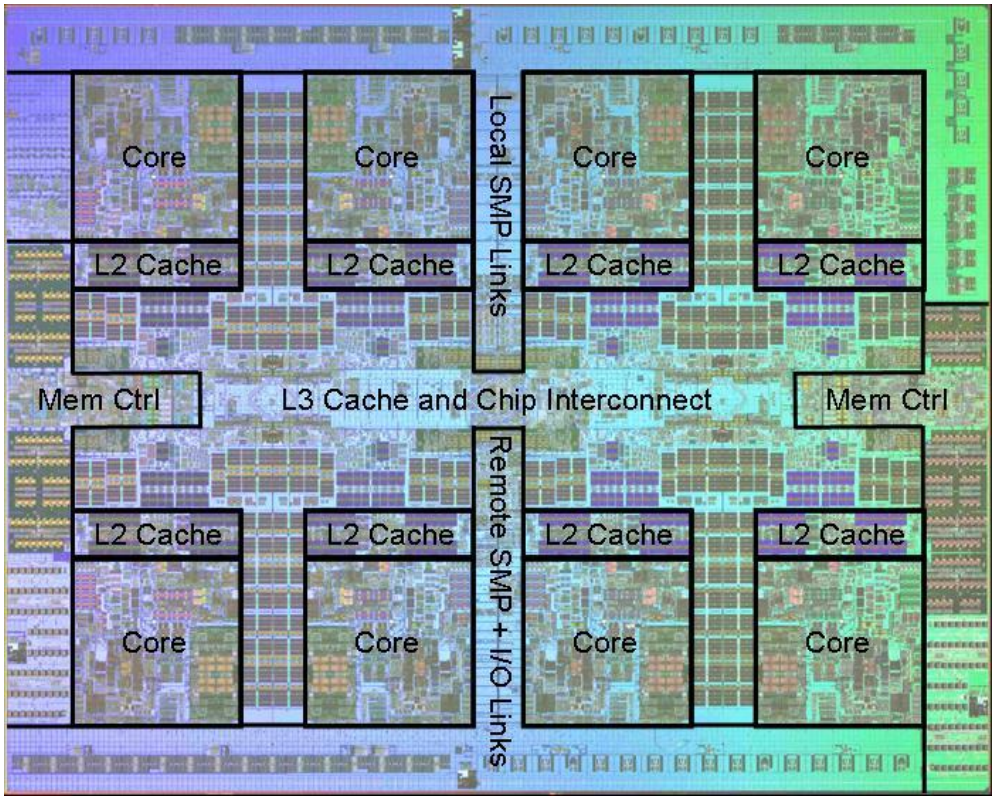


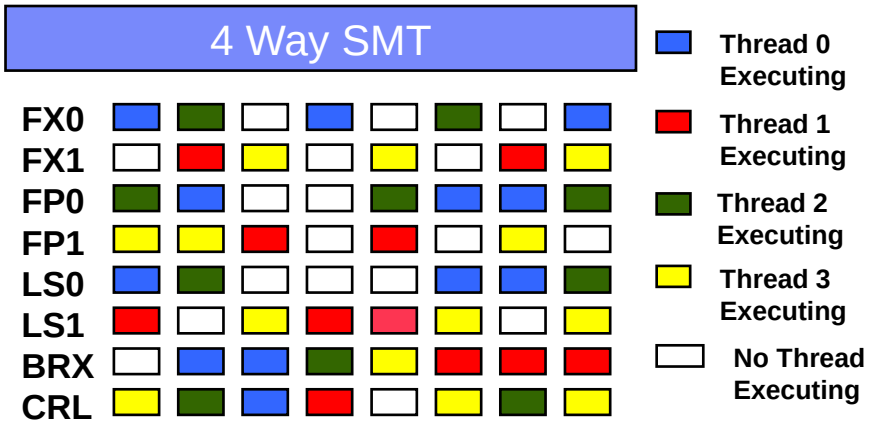




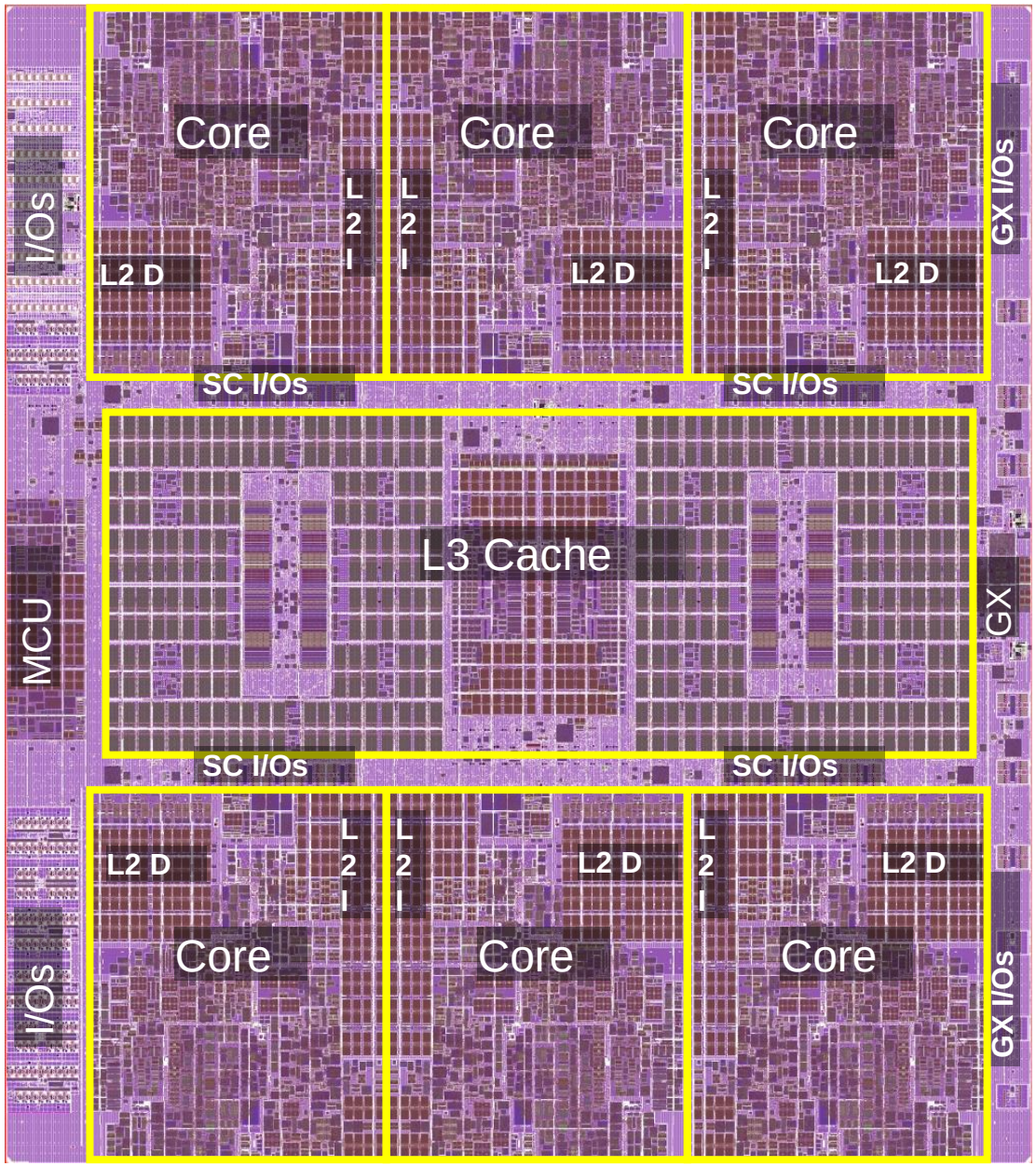
POWER7 Processor Chip

- 567mm^2, 45nm SOI w/ eDRAM
- 1.2B transistors
  - Equivalent function of 2.7B (eDRAM)
- Eight processor cores w/ 4 way SMT
  - 32 threads / chip
- 32MB on chip eDRAM shared L3
- Dual DDR3 Memory Controllers w/ 100GB/s sustained Memory bandwidth
- Scalability up to 32 Sockets
  - 360GB/s SMP bandwidth/chip
  - 20,000 coherent operations in flight





Z EC 12 (32nm, 2012)



32nm high-k CMOS

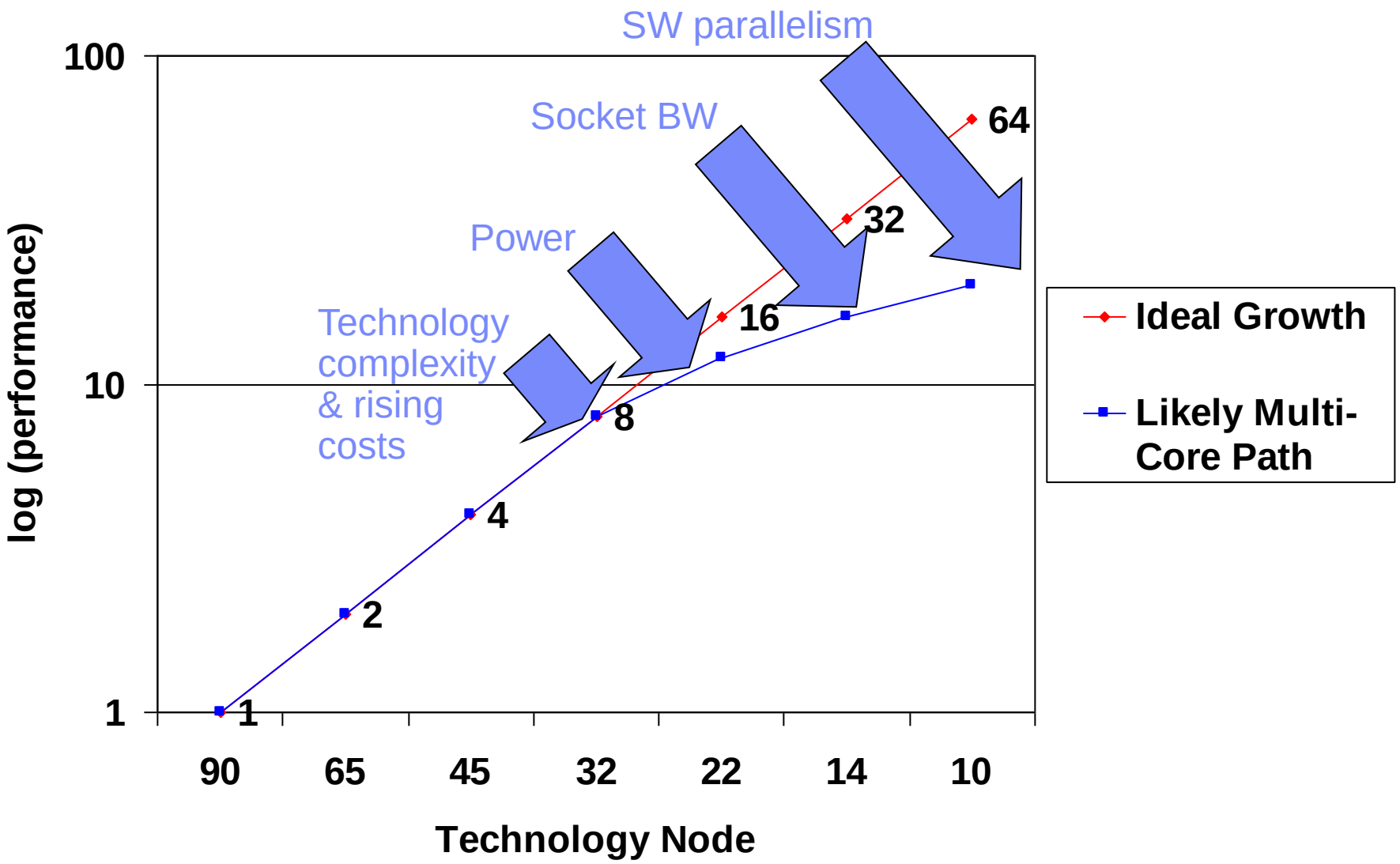
597 mm²

5.5 GHz

2.75B transistors

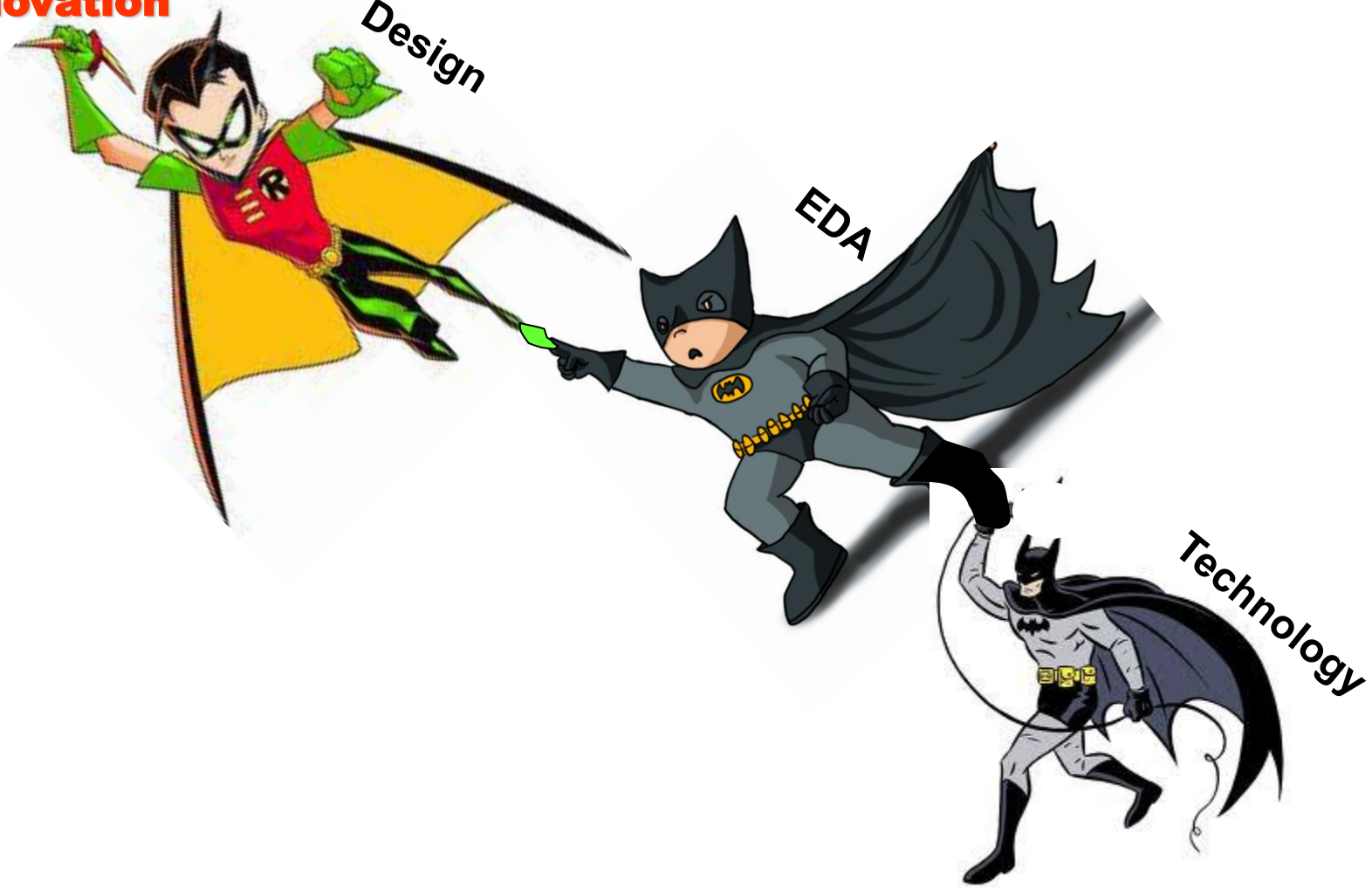
15 levels of metal

Multi-Core Era Limiters

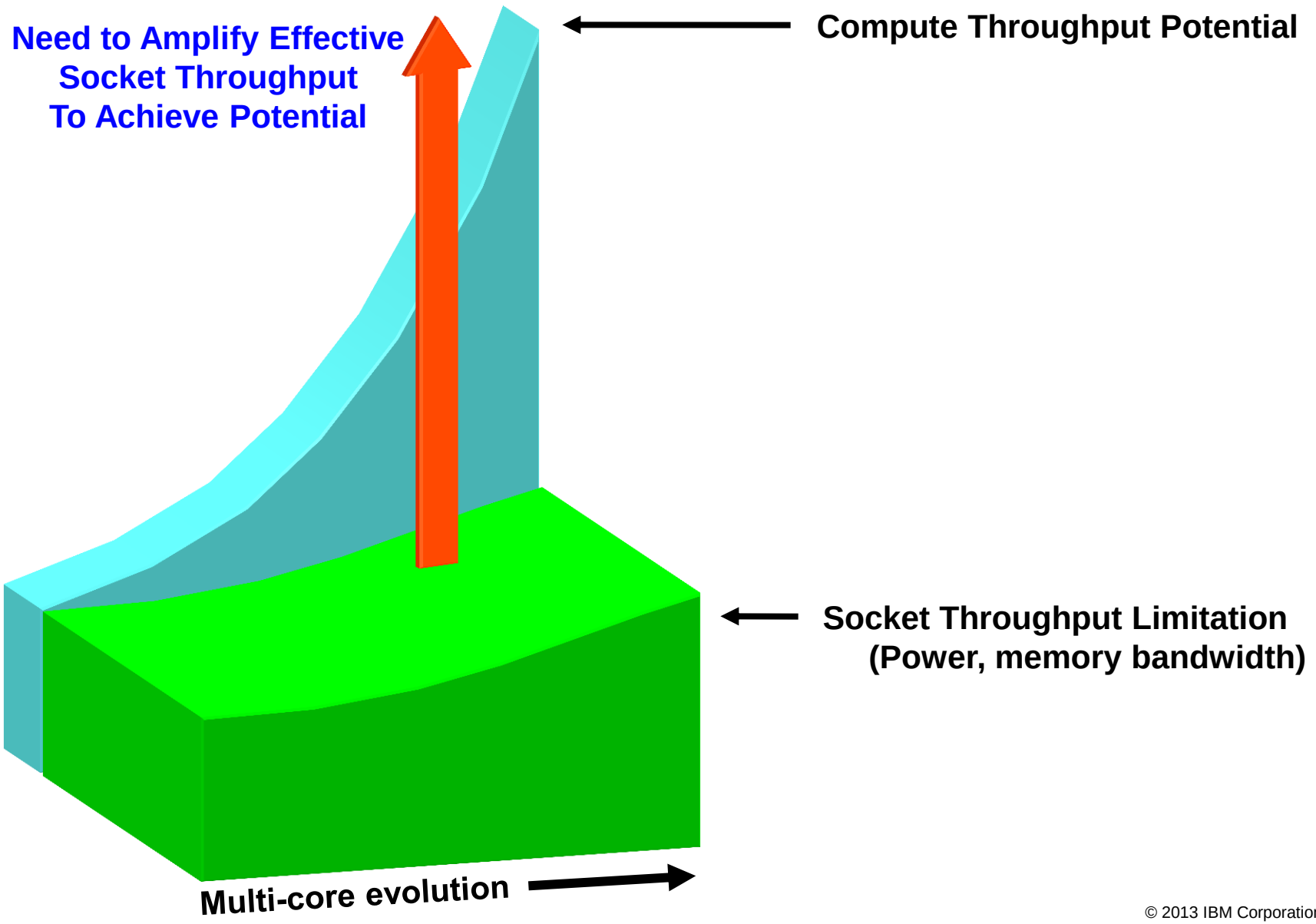


Innovation Drive

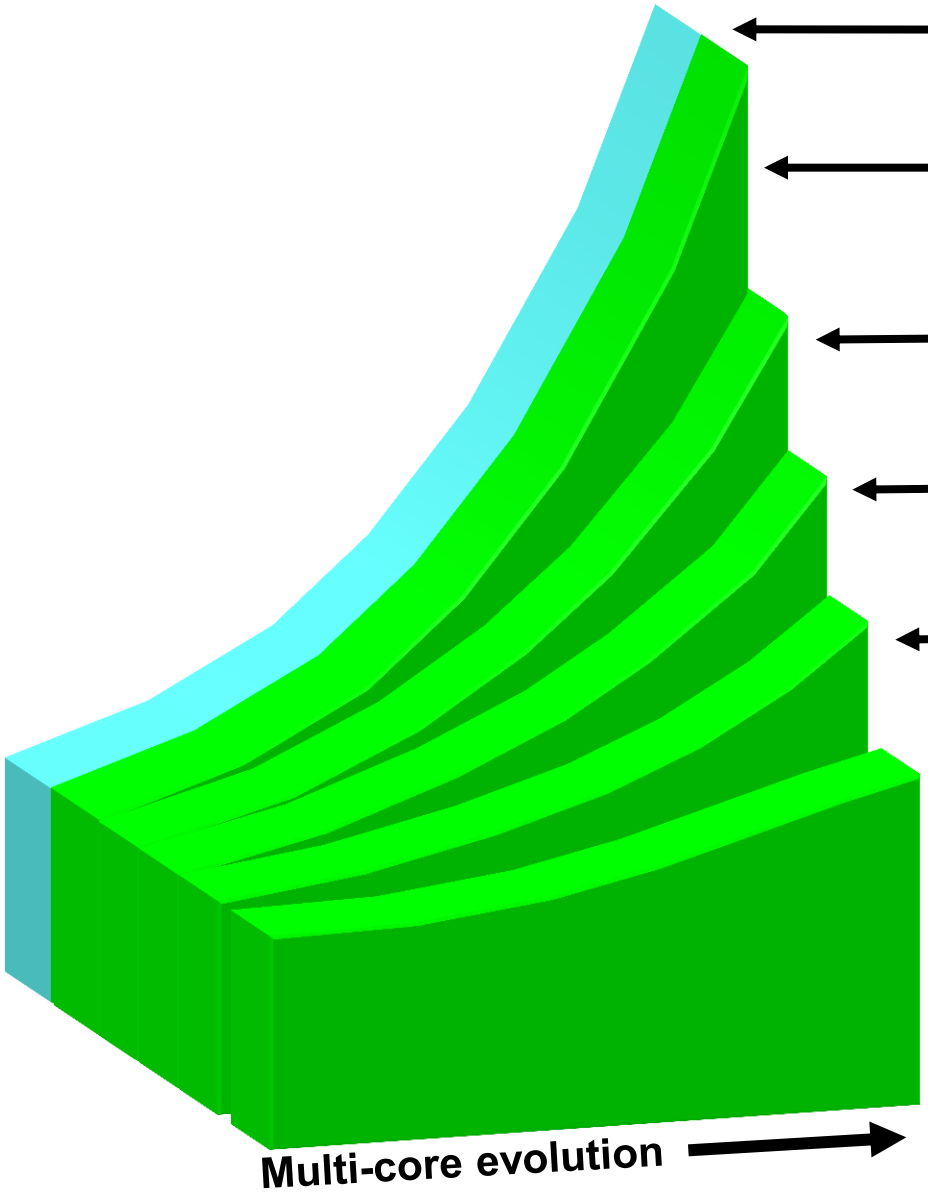
Architecture & Productivity Innovation



Multi-Core Advantage



High performance uP Designs:
Extending Multi-Core Gains (Power processor)



Compute Throughput Potential

Coherence Innovation to minimize socket-to-socket communication

Low-Power Off-Chip Signaling Technology

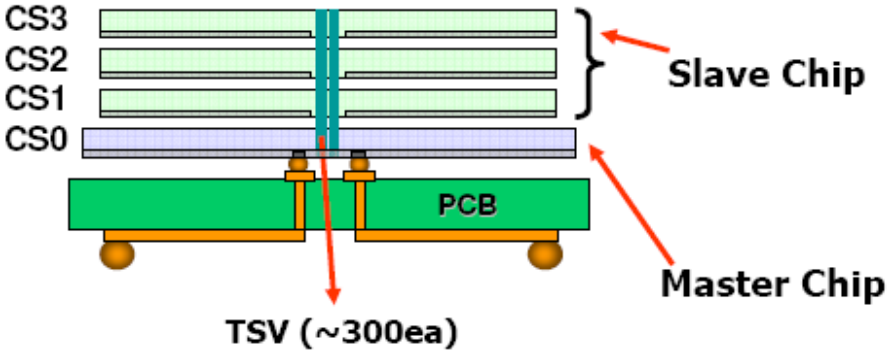
High bandwidth memory buffer

EDRAM = large, low power cache

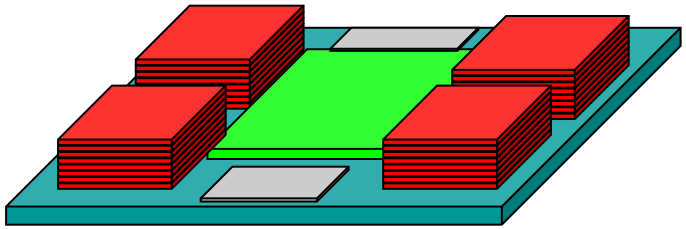
Socket Throughput Limitation (Power, memory bandwidth)

Innovation Drive : System Level Technologies

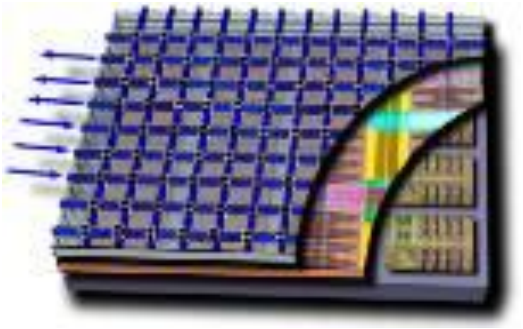
3D Stacking with Through Silicon Vias



Single Processor–Memory Socket



Silicon Photonics

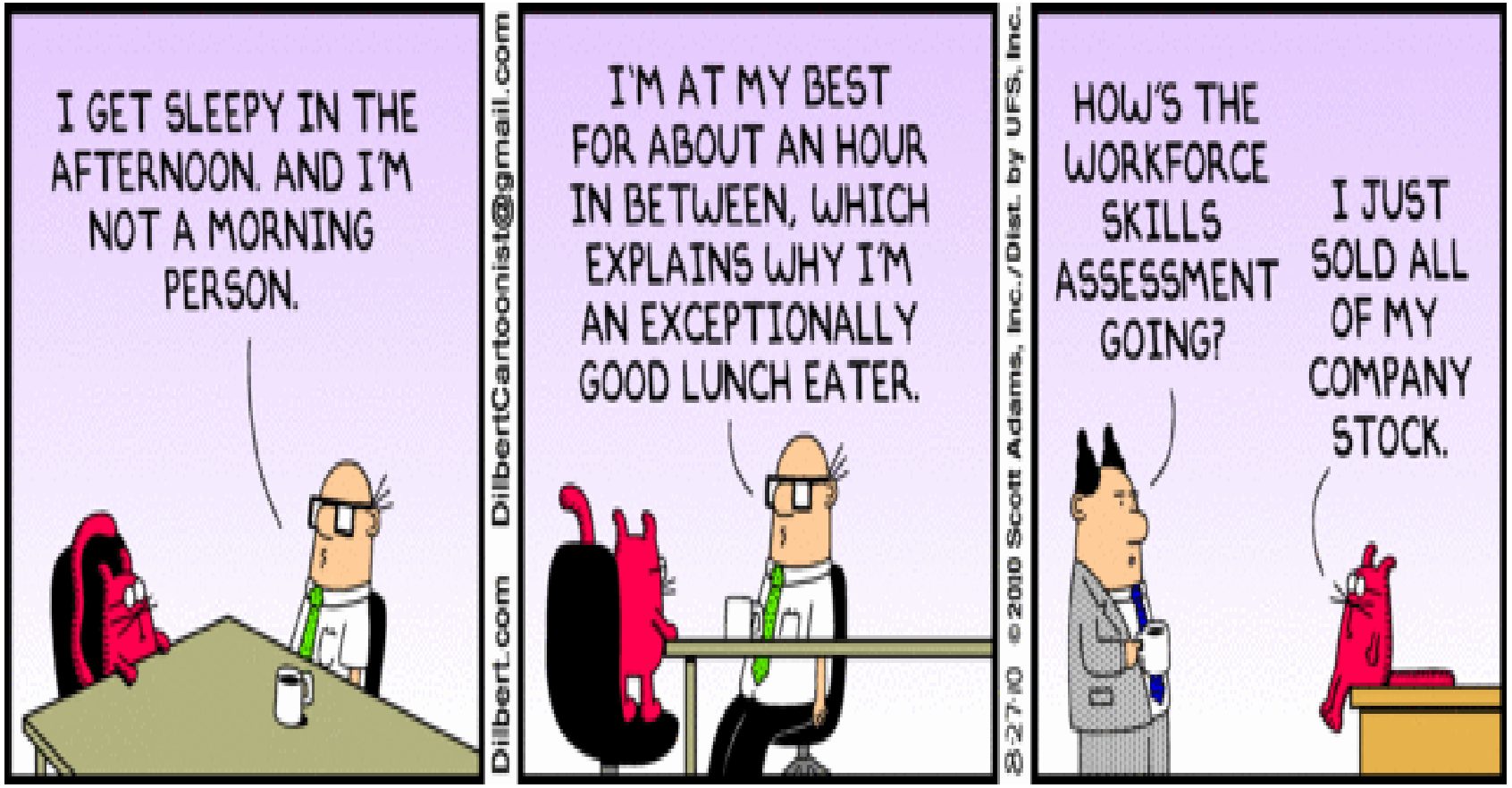


FPGA Accelerators

Flash Memory / SSD

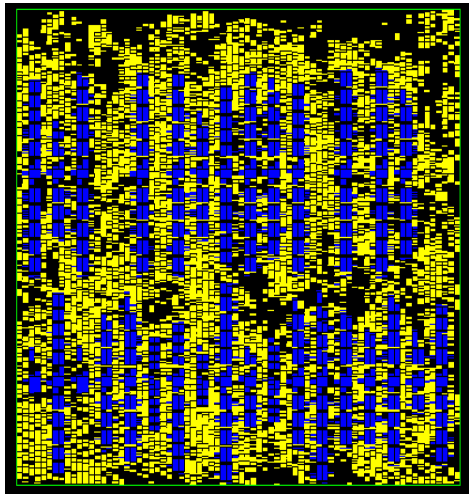
Heterogeneous systems on Chip
Specialized functions
Specialized cores:
- Single thread focused
- Throughput focused

As you start to get sleepy…Dilbert to the rescue..

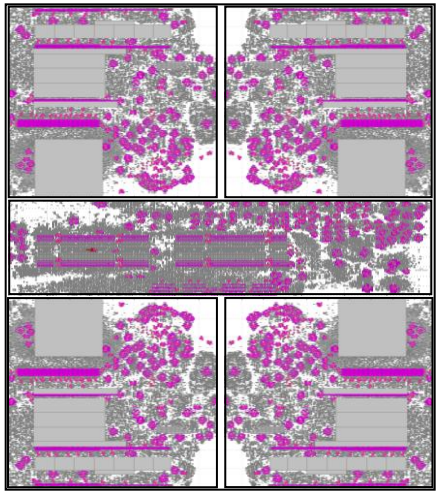


Productivity Innovation: Structured Synthesis and Large Block Synthesis

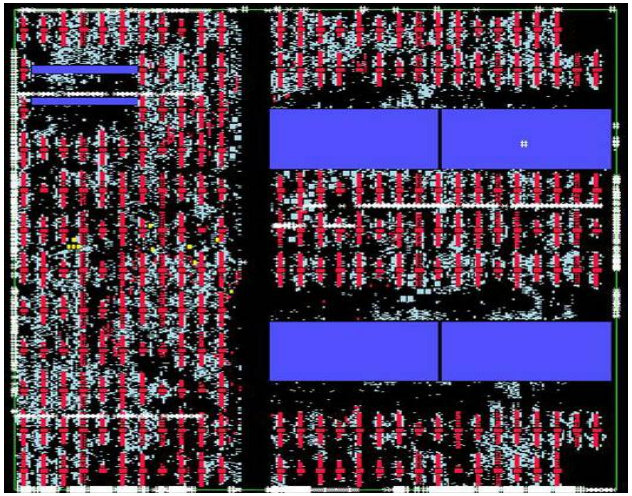
- Customs take large amount of resources and productivity is key
- Merge the domain of customs and Synthesis targeting design productivity and improved quality
  - through merging of custom and synthesis hierarchy with structure in synthesis (not random logic any more)
  - Global Optimization view; Targeted structured data paths and synthesis
  - A methodology with numerous algorithmic and practical innovations spanning from incremental logic design processing, to data paths to structured clocking to custom synthesis merged techniques.



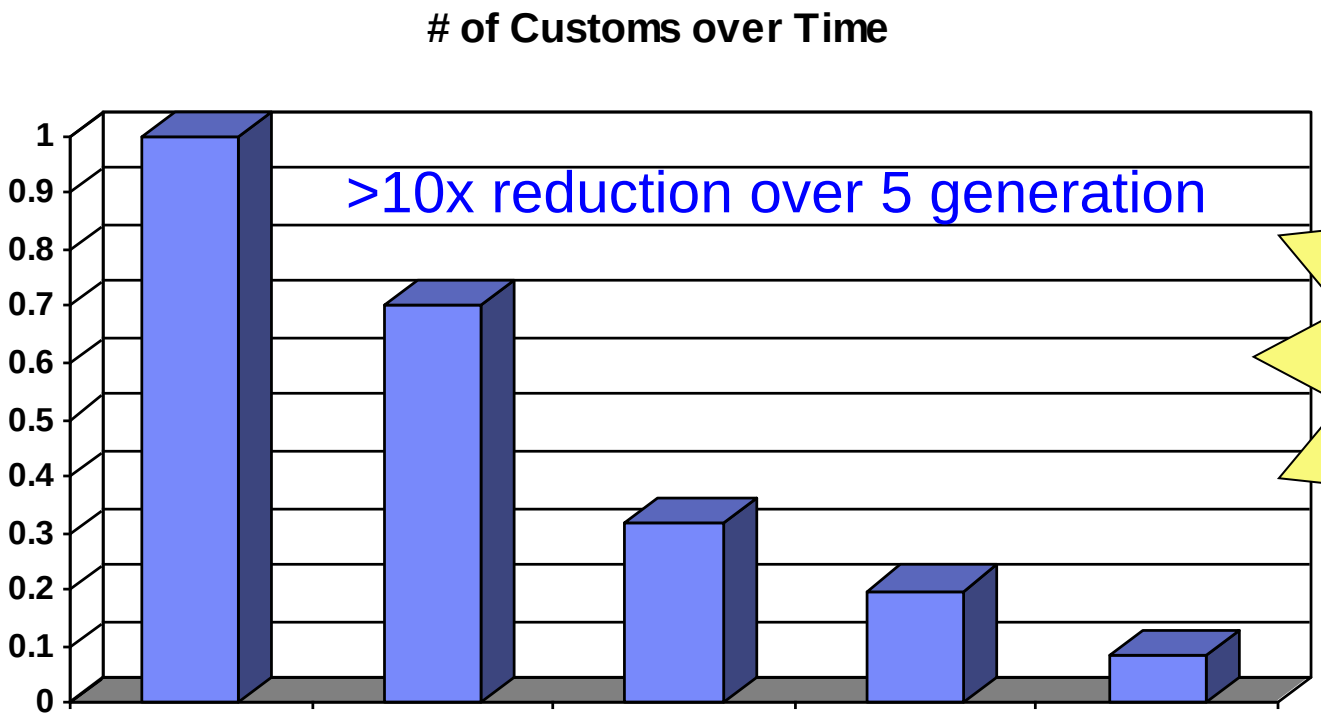
P/Z server Macro



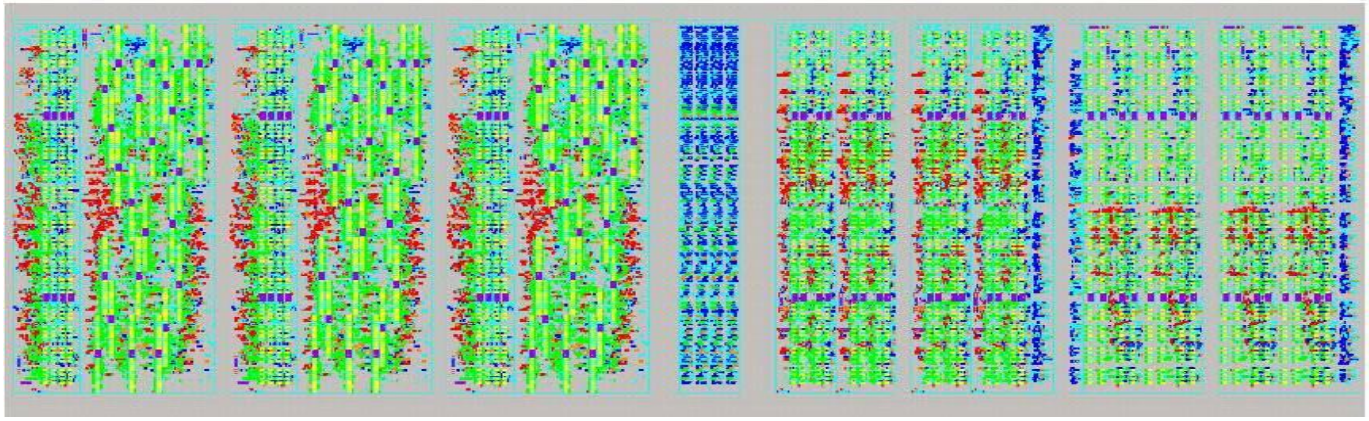
Quad FPU



Productivity Innovation : Reduce Custom Design (Structured Synthesis)



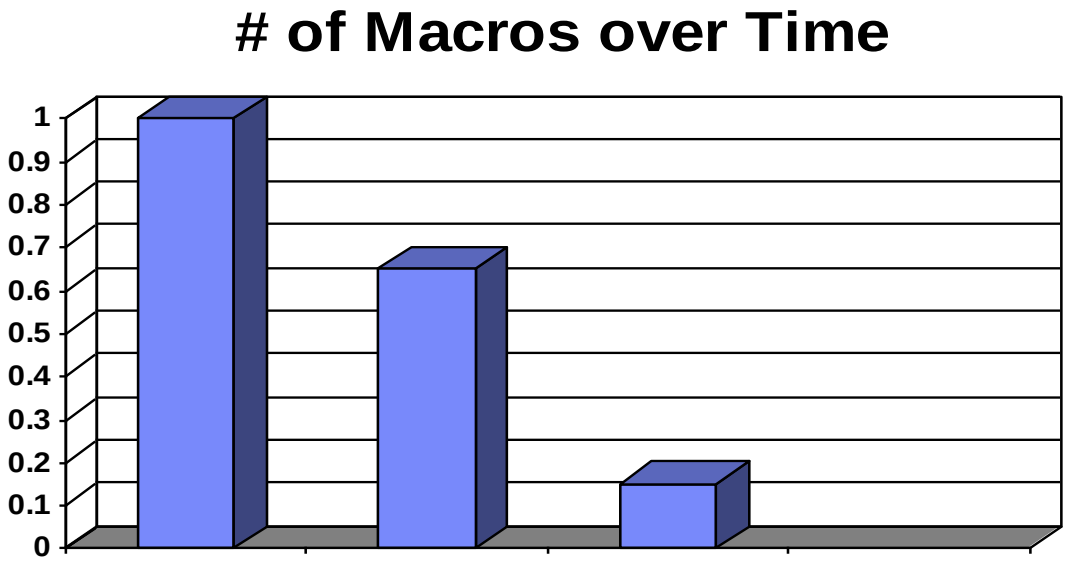
Milestone:
Digital Logic in 22nm server class Microprocessors 99% synthesized and signed-off by Gate Level signoff

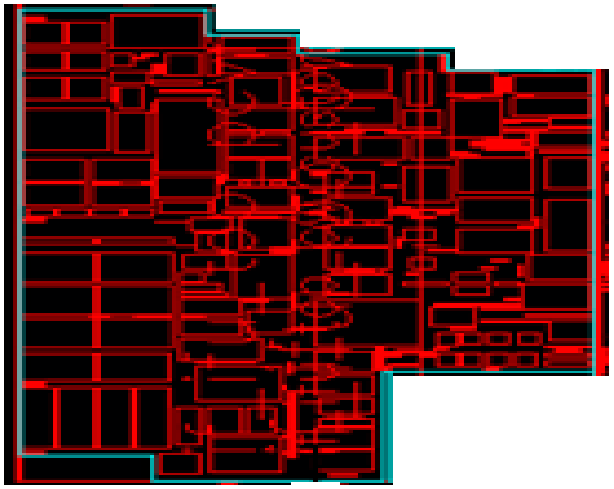


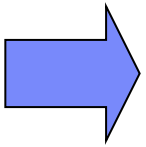
Synthesis results w/ custom-like data flow alignment.

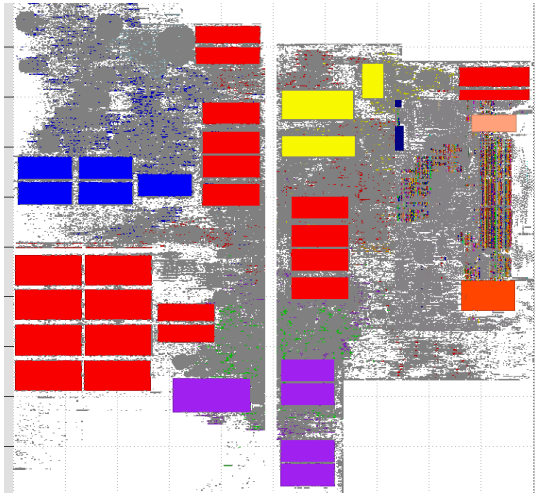
*ISPD 2013 Best Paper Award: “Network Flow Based Datapath Bit Slicing” H.Xiang et al.

Productivity Innovation: Reduced # of Design Partitions (Large Block Synthesis)





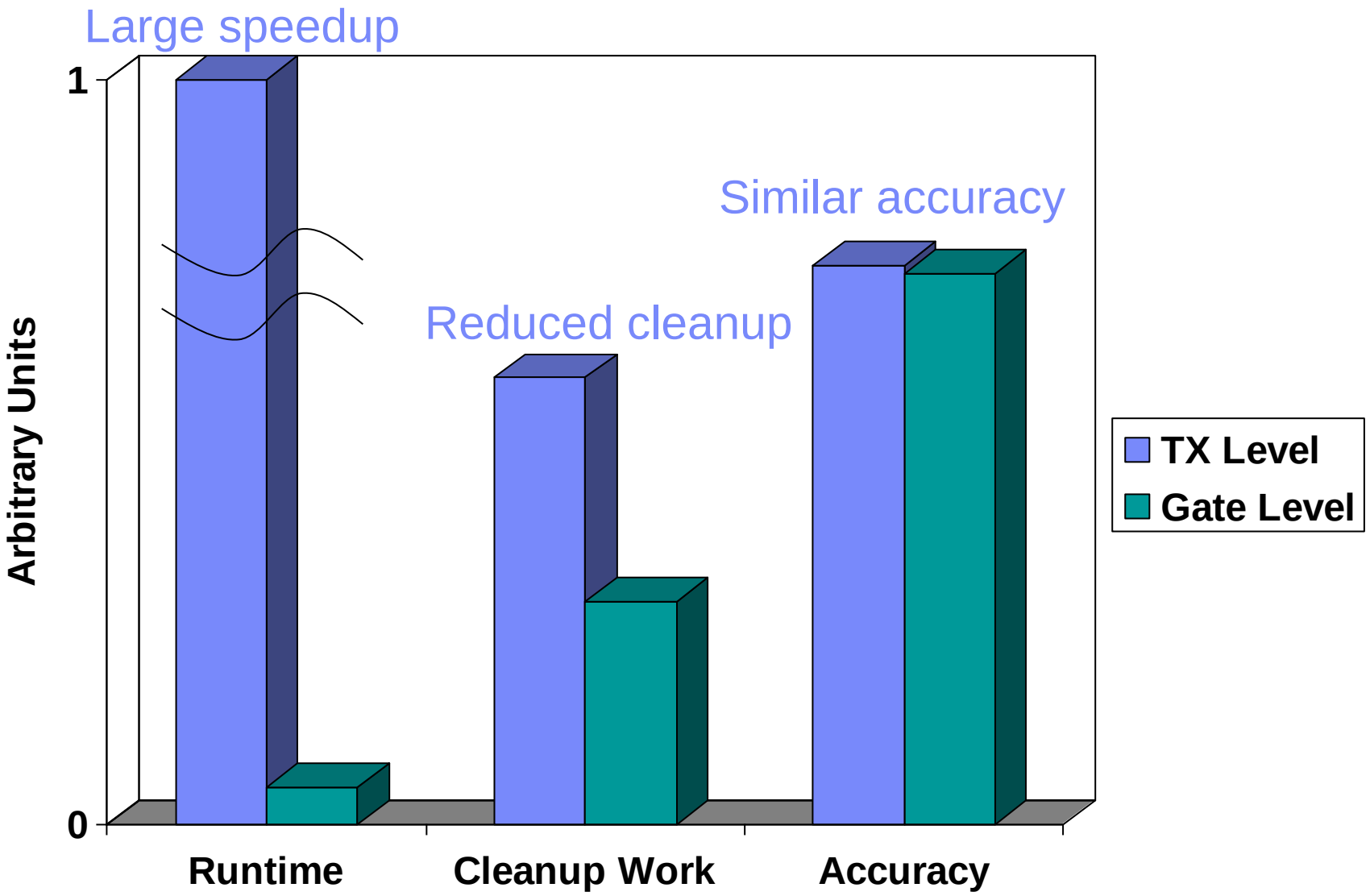




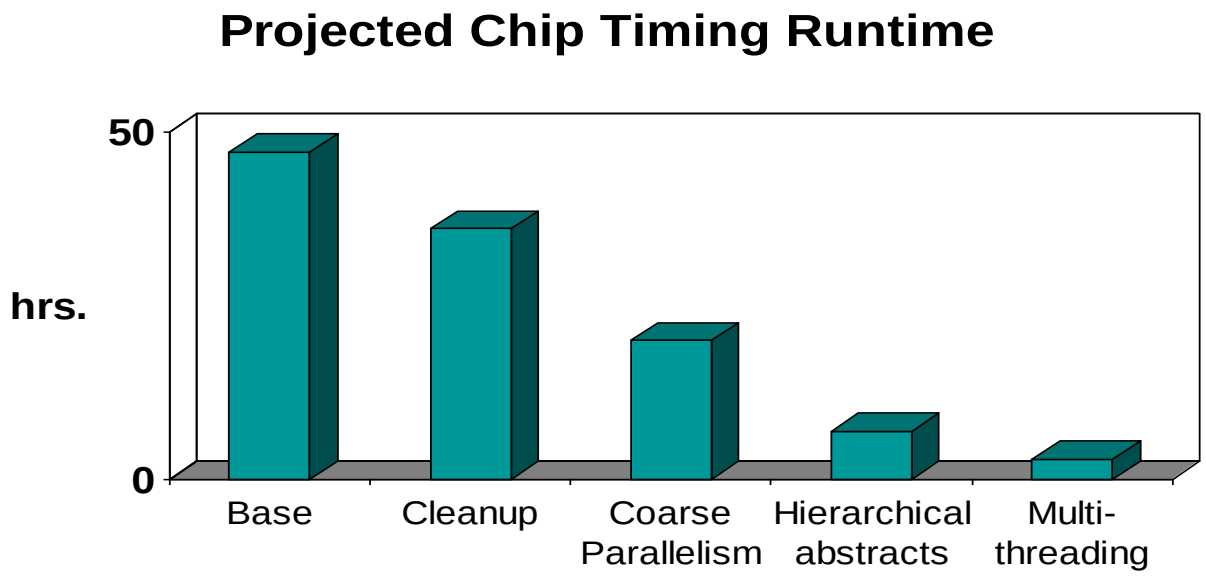
60 logic macros, 25 customs, 14 unique arrays/RFs

1 macro, 0 customs, 9 unique arrays / RFs
Reduced area & power; equal cycle time

Productivity & TAT Innovation: Gate Level Analysis & Signoff

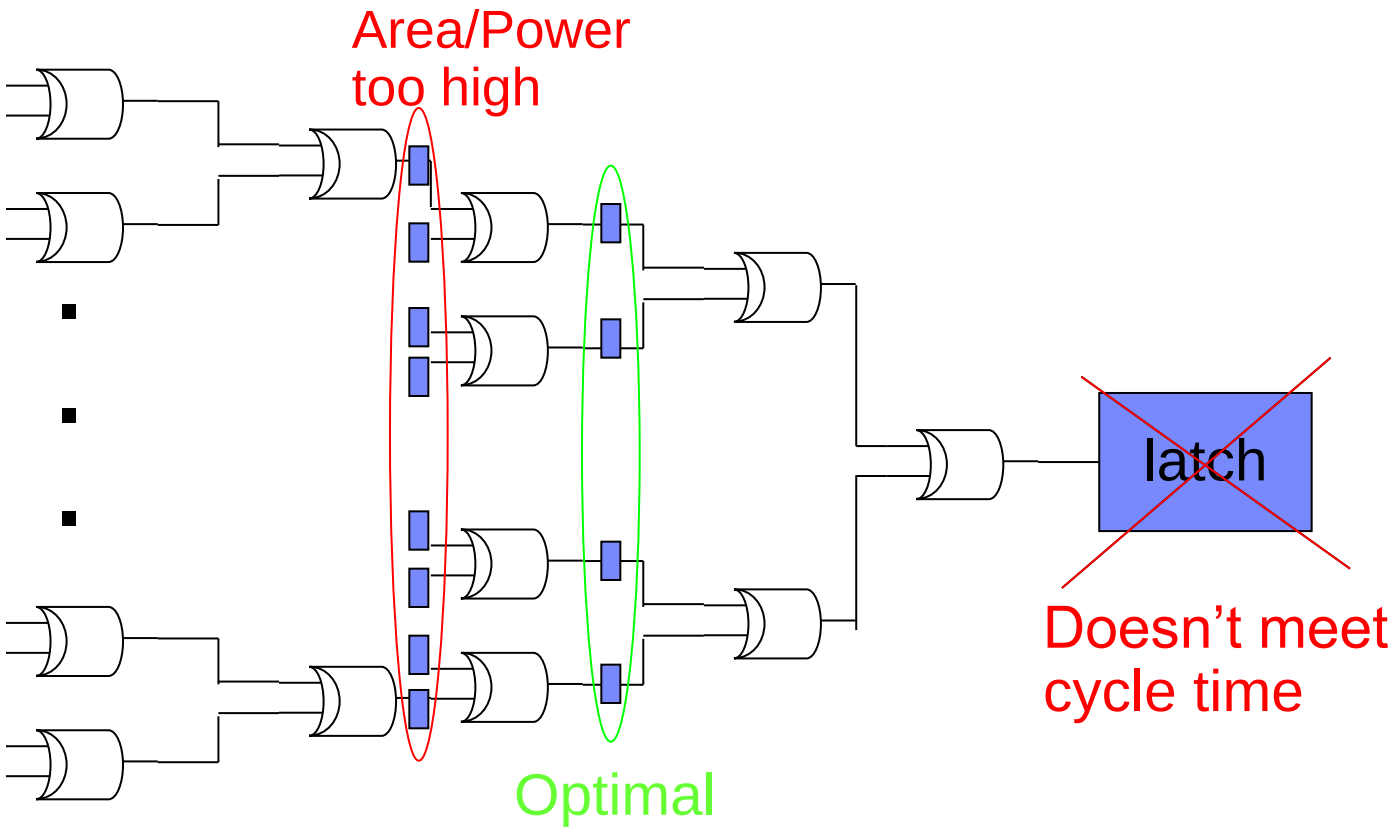


Productivity & TAT Innovation: Hierarchical Abstraction & Multi-Threading



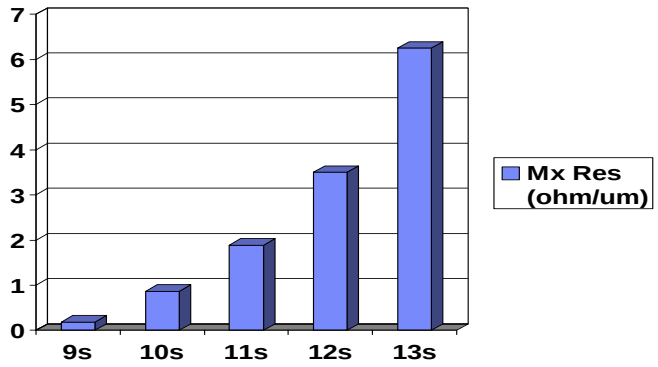
- Fast global analysis tools allow designers to iterate more often resulting in improved final designs.
- Hierarchical abstraction & multi-threading are the most promising ways to minimize TAT.
  - –Applies to all disciplines (timing, verification, etc)

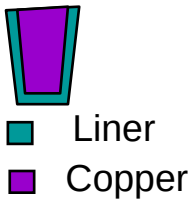
Productivity Innovation Challenges: Retiming

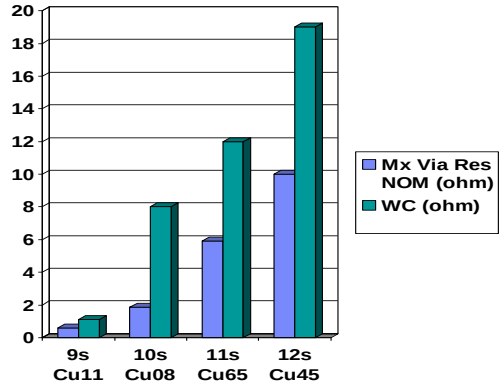


- Significant fraction of logic designer effort spent in optimizing cycle boundaries
- Retiming enables physical synthesis to optimally place latches in logic cones to balance timing/area/power
- Invention is required to seamlessly handle divergence between functional RTL (Verilog/VHDL) and physical implementation throughout methodology.

Challenges Back-end: Scaling and Interconnects







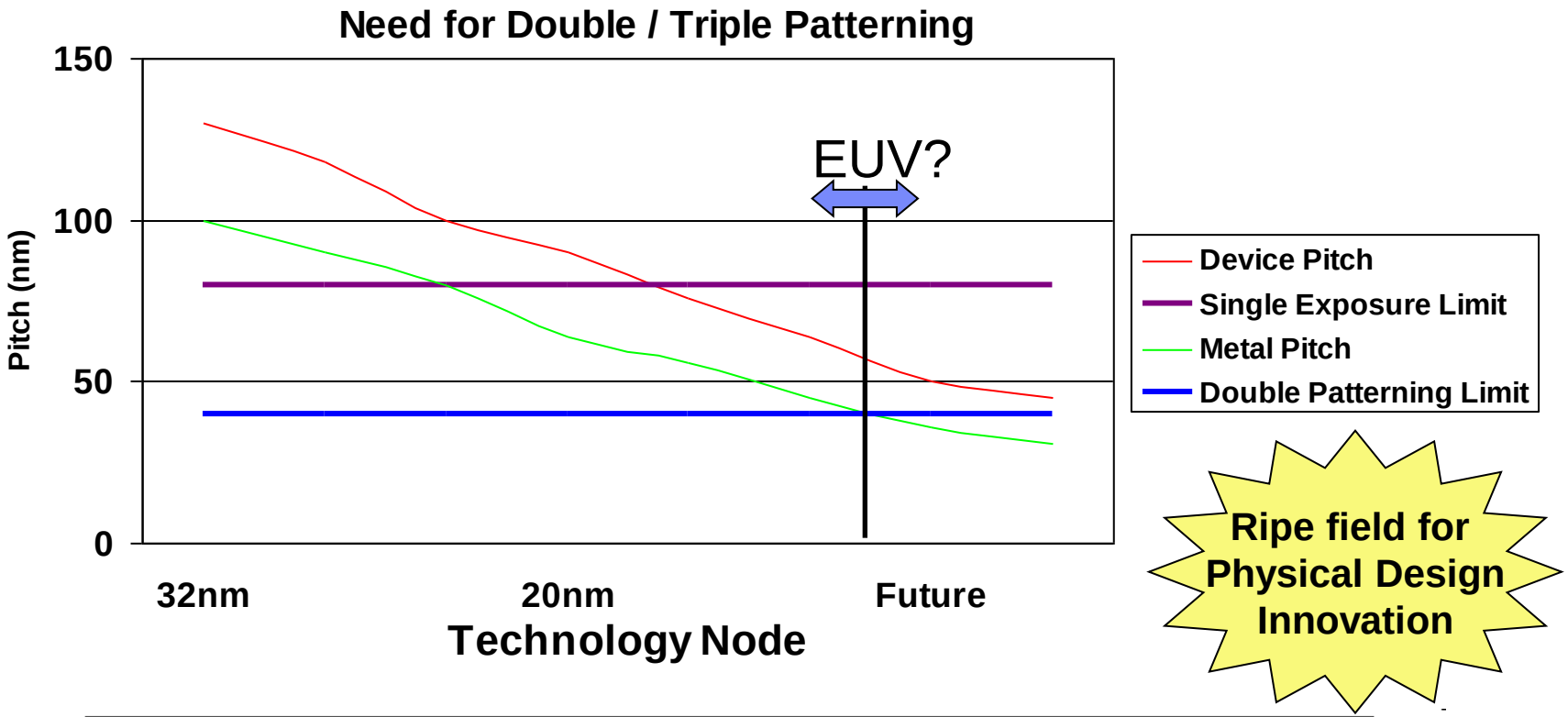
High-performance designs will not be able to tolerate such large RC increases
- Push for more wiring interconnect layers (coarse-pitch)
- Will still need some number of fine-pitch layers for short run local connections

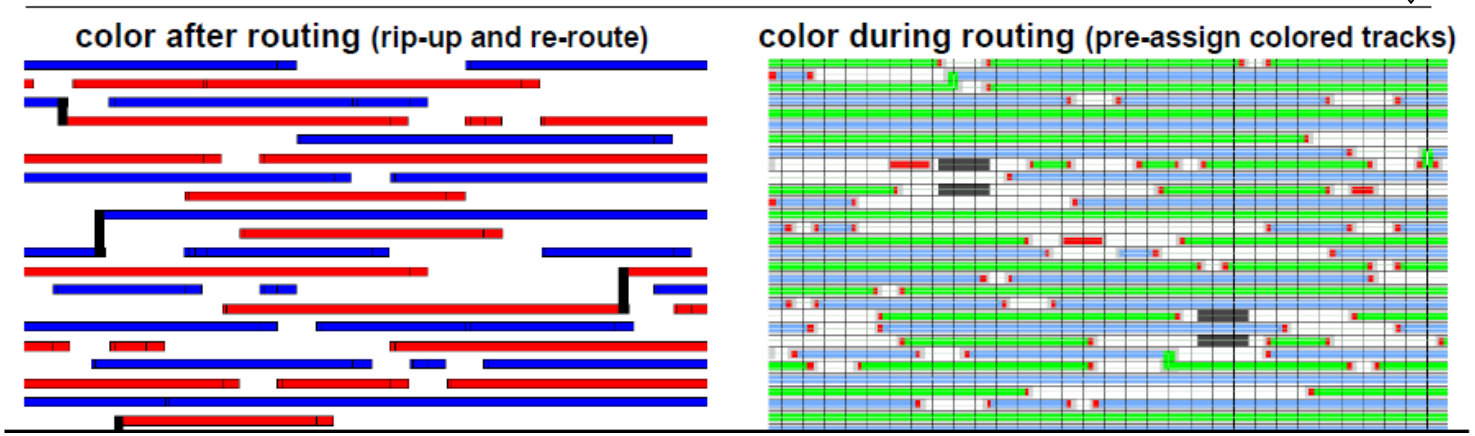
Improved DA tools (routers) needed*
- Optimize wire plane usage to limit technology complexity
- Negotiate through special design rules for the finest levels
- Via optimization, especially at driver end
- Tricky performance vs wireability tradeoffs
- Many wires will need “special” treatment
- Increase width, push higher, add buffers, etc.

Continued Pressure on Wiring. Innovation needed

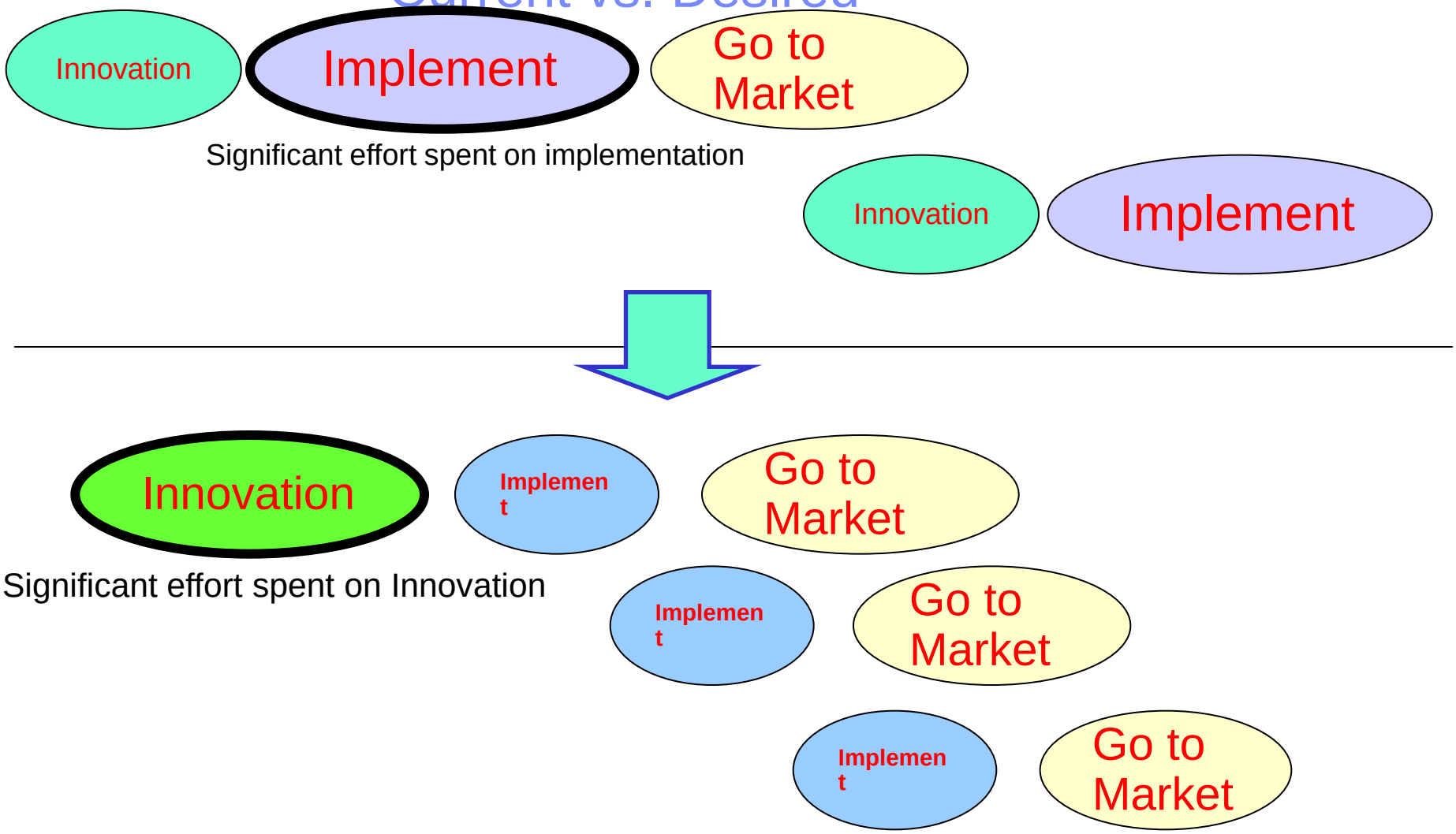
*J.Warnock’s talk earlier yesterday at ISPD

Innovation at Technology, Design Interface: Double/Triple Patterning

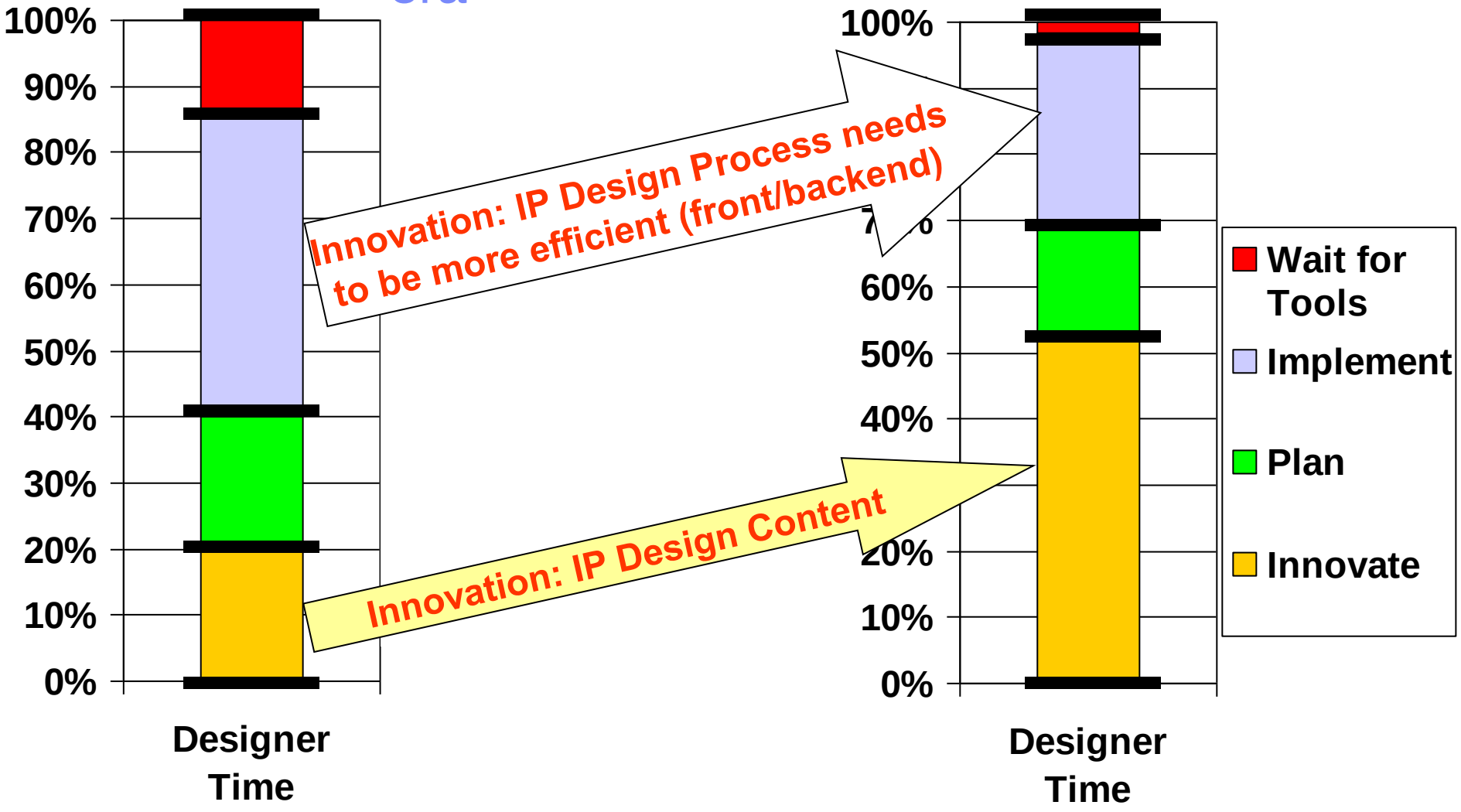




How We Stage Designs to Market: Current vs. Desired

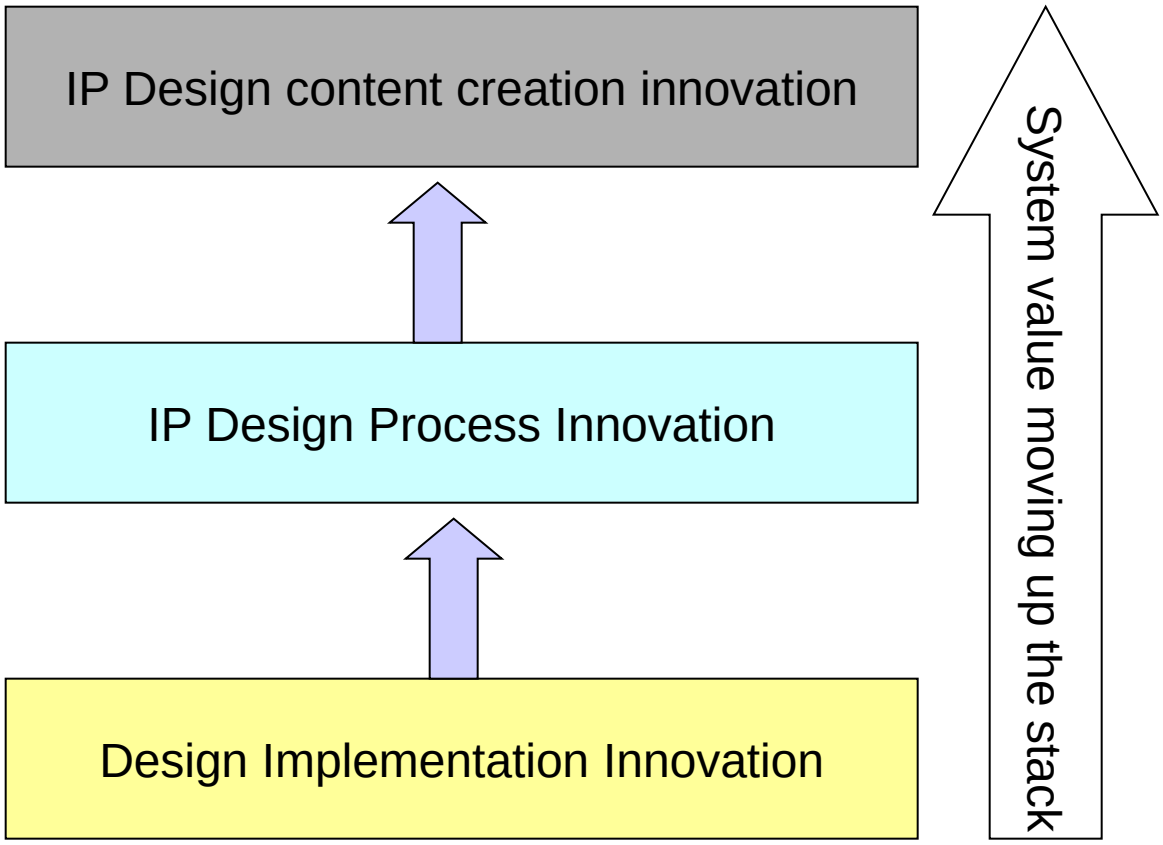


Innovation: The sweet spot in this new era

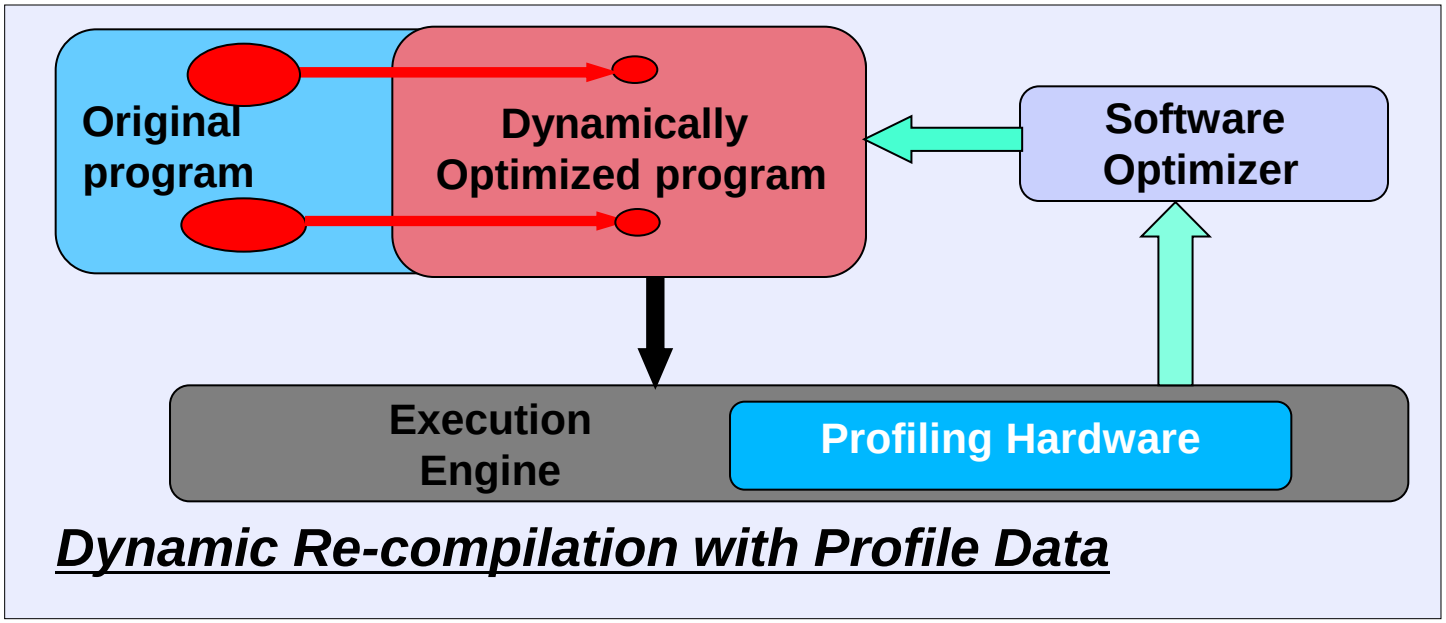


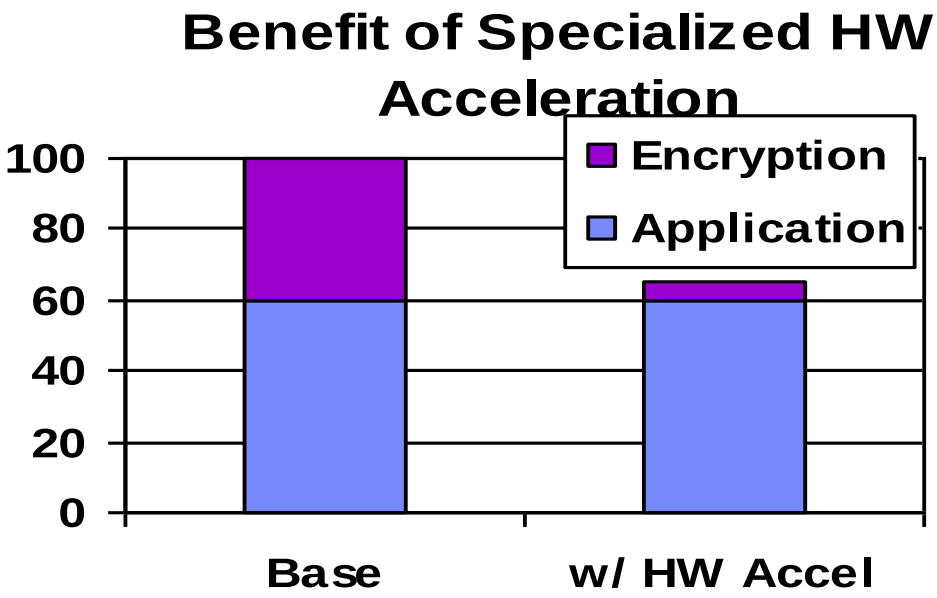
New Frontier: Innovating up the Value Stack

:



Innovation Drive : Hardware-Software Co-Optimization





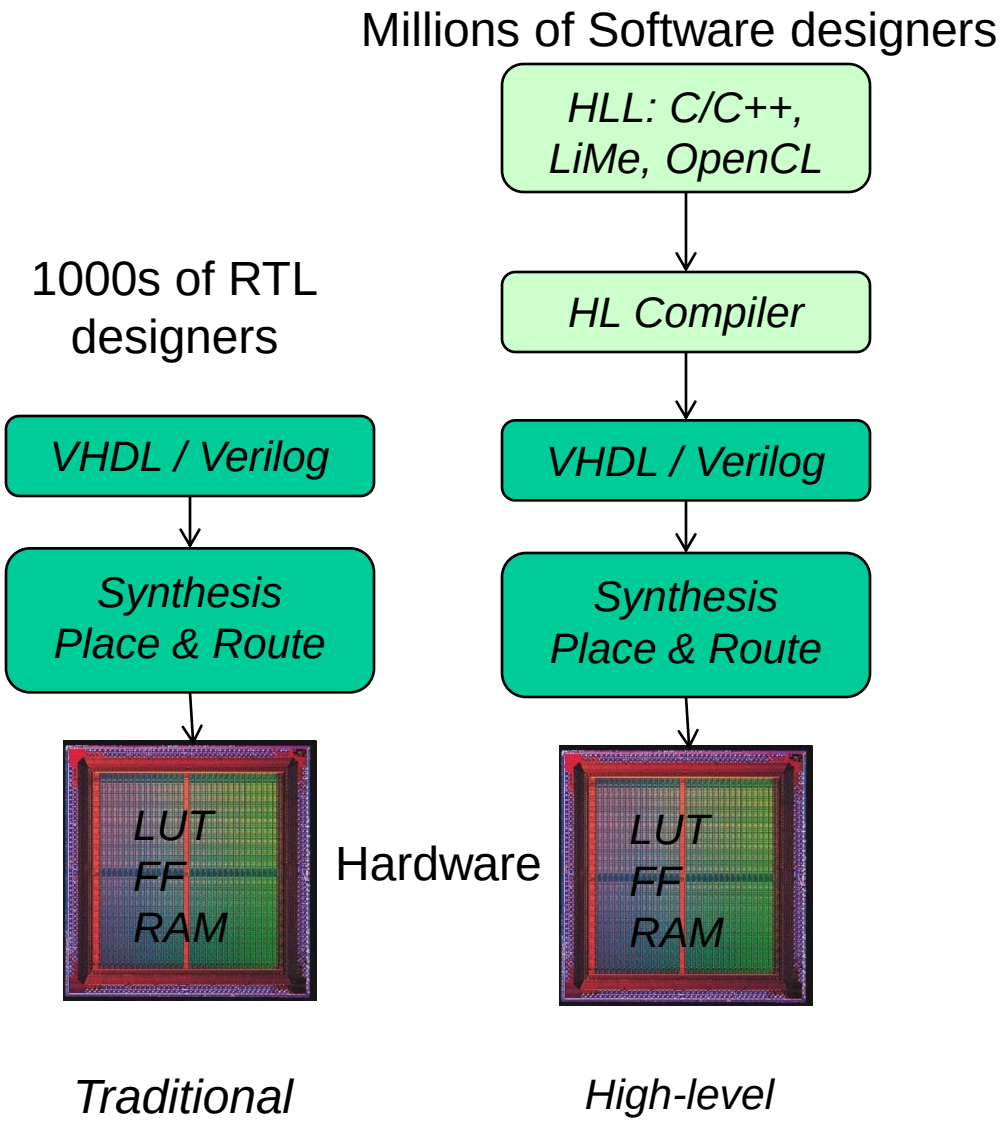
FPGA Accelerators

Innovation on Acceleration in PD/timing domain?

Heterogeneous design: Specialized HW acceleration viable in many areas

- Graphics
- Compression
- Cryptography…

Hardware Programming



- Hardware mistakes cannot be patched

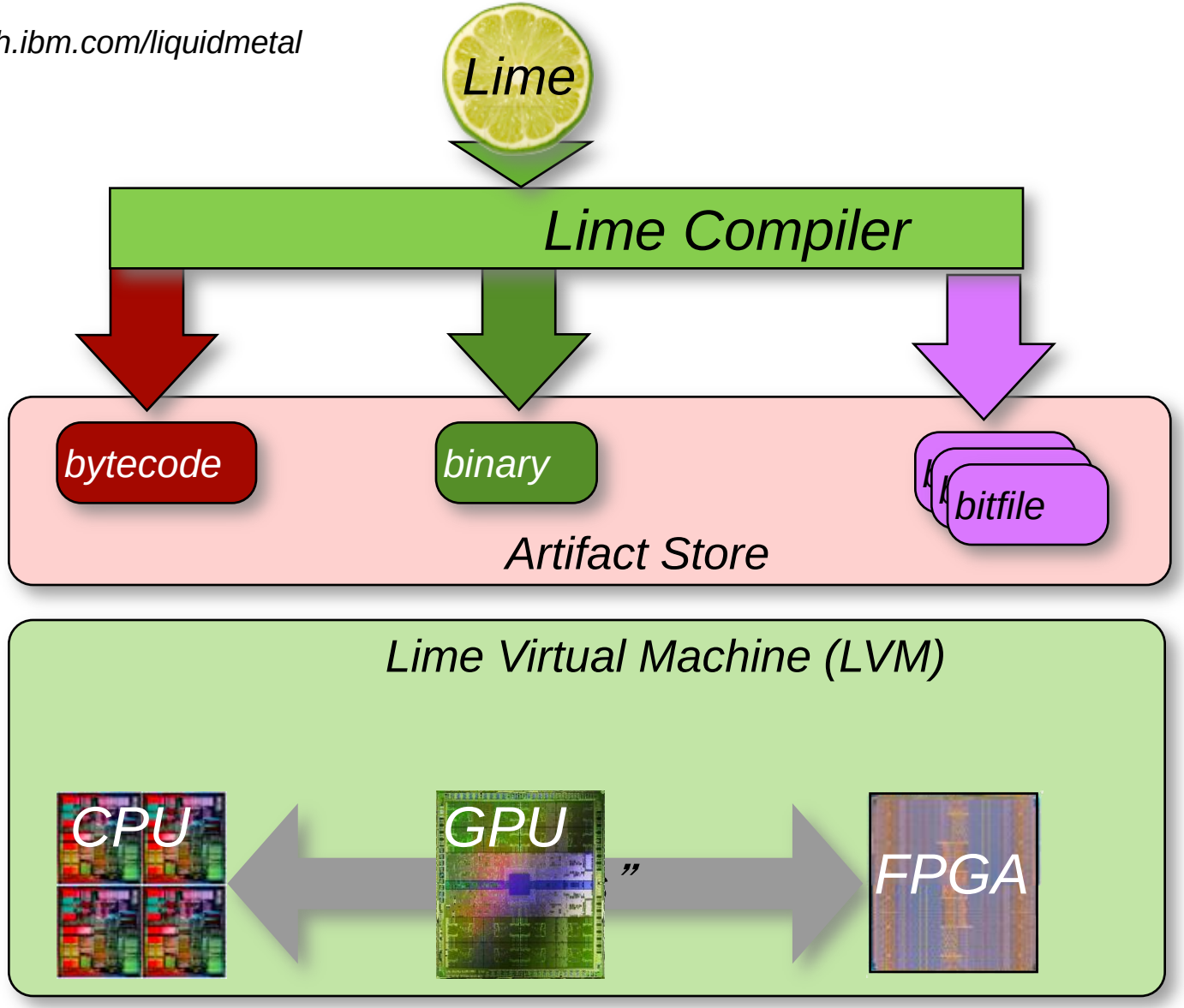
- Significant barriers to be overcome between RTL, physical design domain and high level languages.

- Correlation between higher level spec. and lower level design.

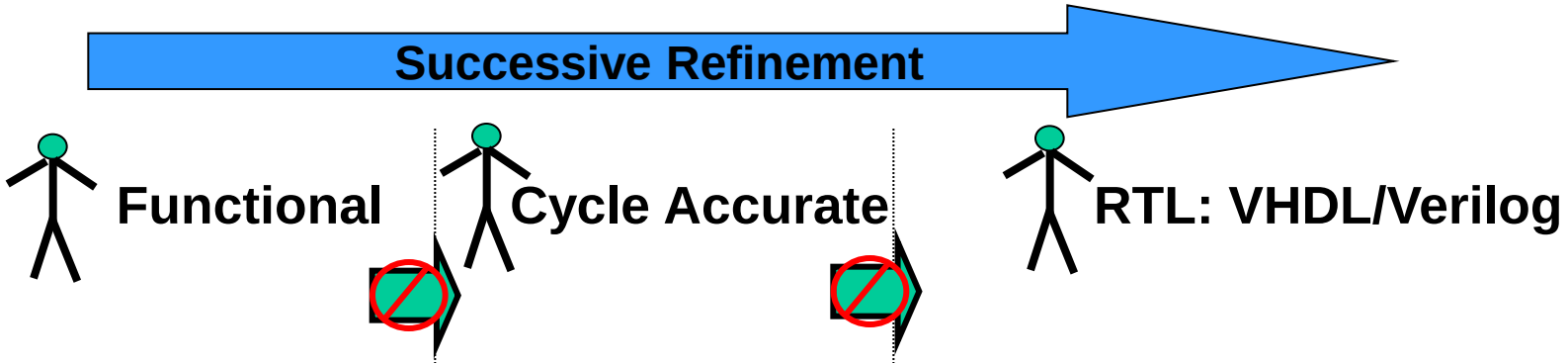
There is more religion in logic design than number priests in religion

Liquid Metal (A high level harware compiler)

http://www.research.ibm.com/liquidmetal



Architectural Synthesis



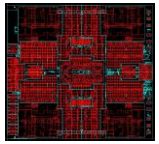
C/C++ Model

Metric: Cache Miss rate etc.

C/C++ Model

Metrics: Performance Models, CPI etc

Back End PD Implementation and Analysis

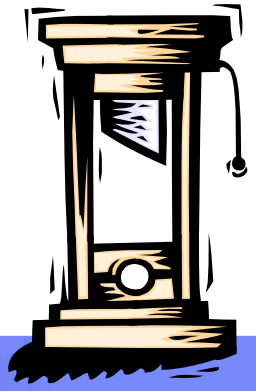


Metrics: Electrical, Timing, Area, Noise etc.

Capturing the Vibrant spirit of DA Researcher

Once upon a time there lived three men: a doctor, a chemist, and a DA Researcher. For some reason all three offended the king and were sentenced to die on the same day.

The day of the execution arrived, and the doctor was led up to the guillotine.



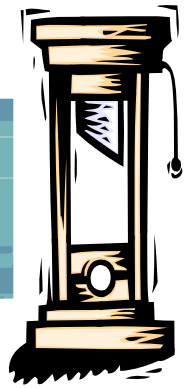
Capturing the Vibrant spirit of DA Researcher

As he strapped the doctor to the guillotine, the executioner asked, "Head up or head down?"

"Head up," said the doctor.

"Blindfold or no blindfold?"

"No blindfold."



So the executioner raised the axe, z-z-z-z-ing! Down came the blade--and stopped barely an inch above the doctor's neck. Well, the law stated that if an execution didn't succeed the first time the prisoner had to be released, so the doctor was set free.

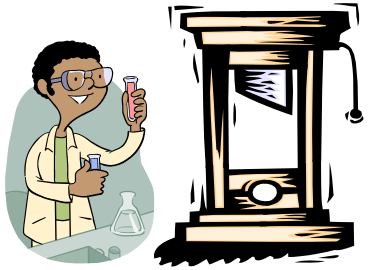
Capturing the Vibrant spirit of DA Researcher

Then the chemist was led up to the guillotine.
"Head up or head down?" said the executioner.

"Head up."

"Blindfold or no blindfold?"

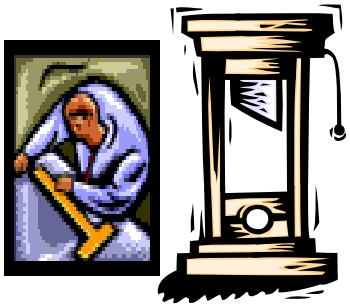
"No blindfold."



So the executioner raised his axe, z-z-z-z-ing! Down came the blade--and stopped an inch above the chemist's neck. Well, the law stated that if the execution didn't succeed the first time the prisoner had to be released, so the chemist was set free.

Capturing the Vibrant spirit of DA Researcher

Finally the DA Researcher was led up to the guillotine.



"Head up or head down?"

"Head up."

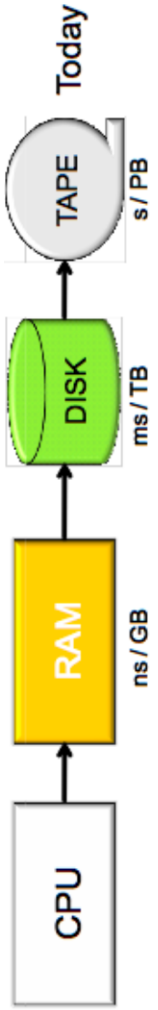
"Blindfold or no blindfold?"

"No blindfold."

So the executioner raised his axe, but before he could cut the rope, ever curious Chandu yelled out:

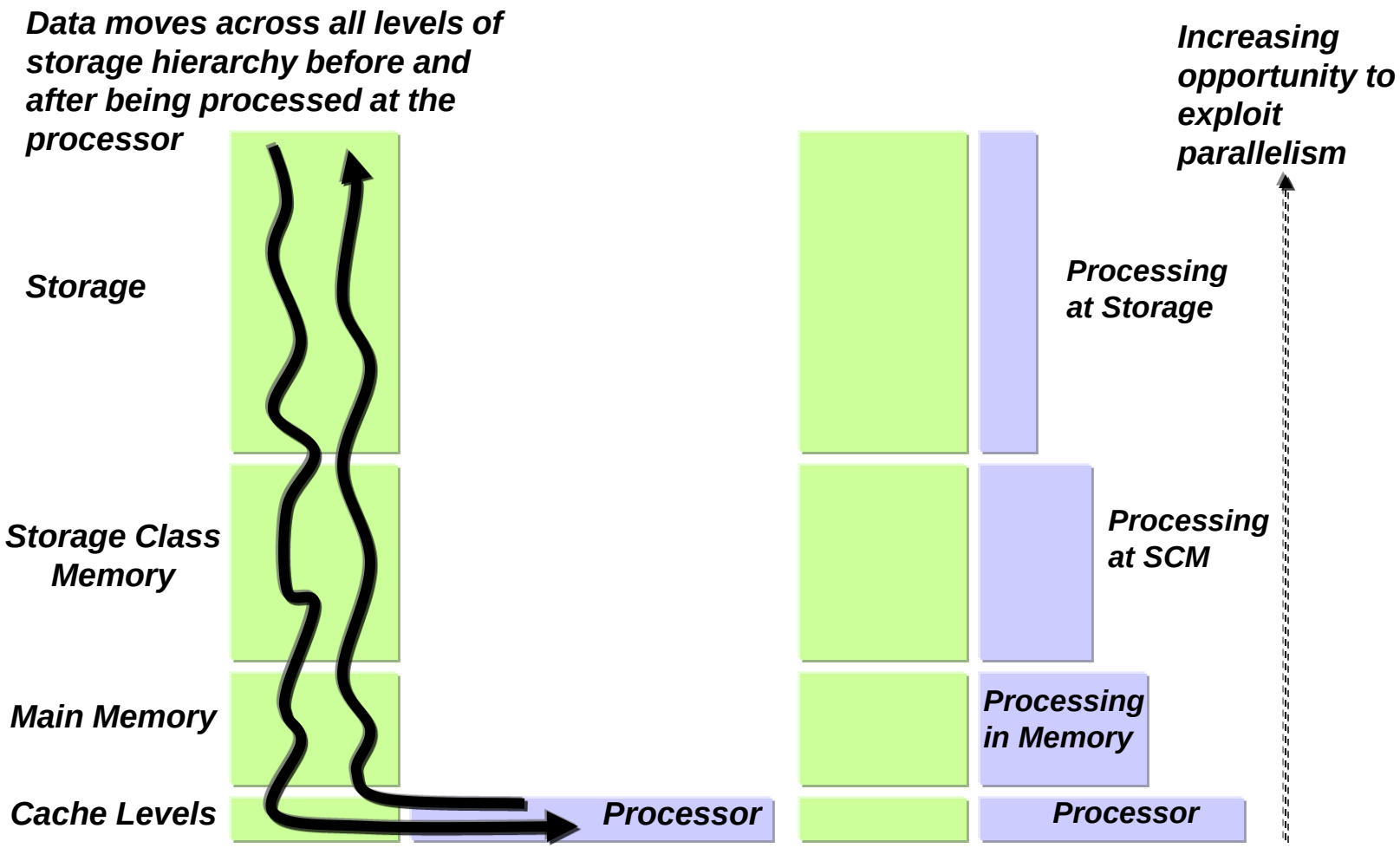
"WAIT! I see what the problem is!".

Role of Memory



- Von Neumann Bottleneck gating system performance (memory bandwidth)
  - Data resides many hierarchy levels away from the computation.
  - Get 100 pages of data 4-5 hierarchy levels down into the cache, use 1 page, discard the rest 99.
    - Latency bottleneck and energy waste.
- Solution increasingly being adopted:
  - Computation right where data is:
    - Netezza for database filtering is one such example
    - Hybrid Memory Cube (Micron); AMC (IBM/Micron)
    - For functions which are critical and need very active access to data, perform next to data.

Systems view of Data processing

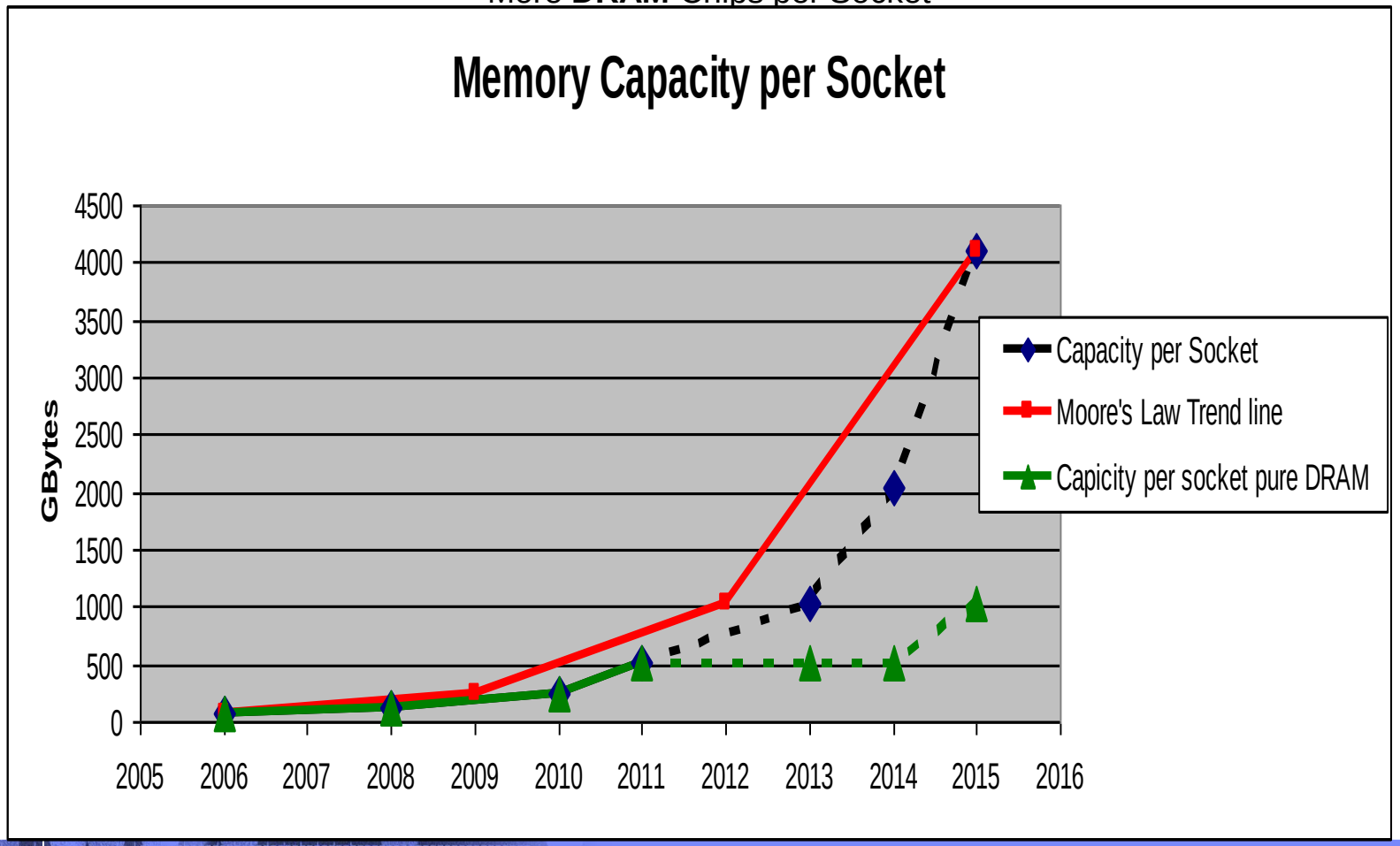


Traditional Computing → Hierarchical Data Processing

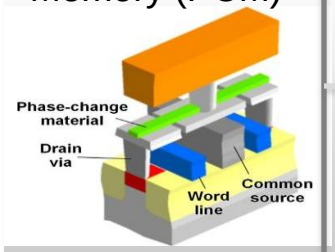
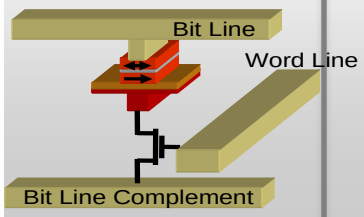
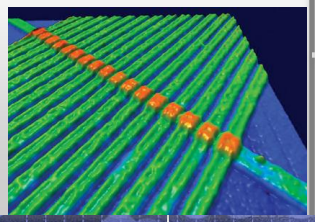
Scope of memories is expanding

- Memory is becoming pervasive throughout, and bits are becoming plentier and cheaper every day (< 50c /GB for flash)
  - Memory technology continues to evolve even if scaling slows down.

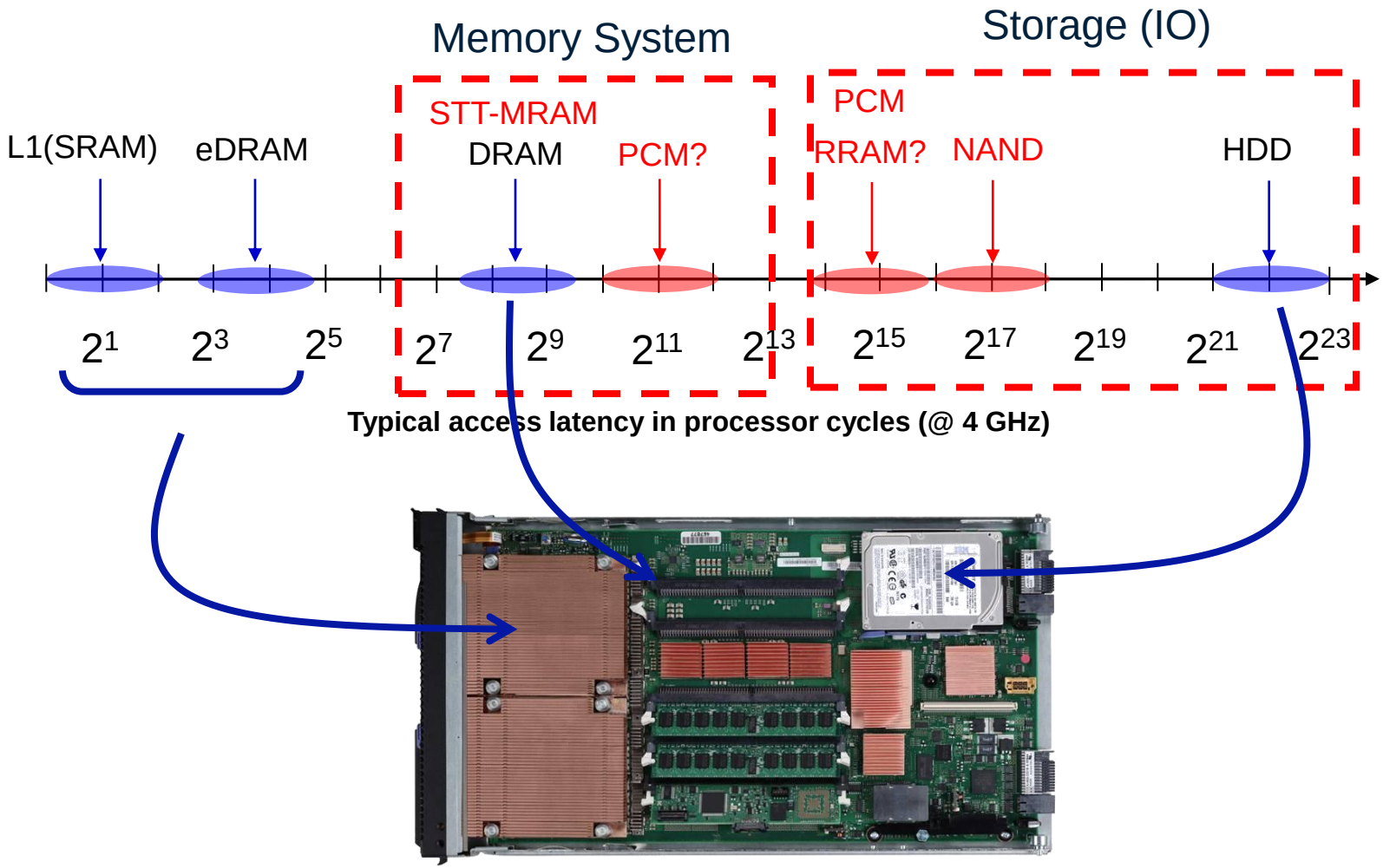
More Memory Capacity per Socket though
More **DRAM** Chips per Socket



Storage Class Memory – Market View

	Projected Date of Enterprise Availability	Vendors Researching	Initial Characteristics: Density	Initial Characteristics: Cost	Initial Characteristics: Latency	Risk Comments
Phase Change Memory (PCM) 	Low Density 2015+	Micron Hynix	= DRAM	Parity in 2015	3x DRAM Reads 10x DRAM Writes	Most advanced technology Productization and Aggressive dev. on hold until market is identified
	High Density 2018+	Samsung	0.25x NAND	4x NAND		Current very low density NOR Flash Offerings exist for a small market
Spin Torque Transfer MRAM (STT-MRAM) 	2016/2017+	Micron- Partnership with IBM but also pursuing other DRAM replacement tech Hynix-allied w/Toshiba Samsung- Acquired start-up *Grandis*	= DRAM	Target DRAM Parity	= DRAM in both Reads & writes	Early in technology dev. Promising scalability Manufacturability unexplored
Resistive RAM (RRAM) 	Low Density 2016+	Micron Hynix-partnered with HP	2x NAND	Target NAND Parity	> DRAM, < NAND (~us)	Early in technology dev. Promising scalability & performance Manufacturability unexplored
	High Density 2018+	Samsung Sandisk/Toshiba				

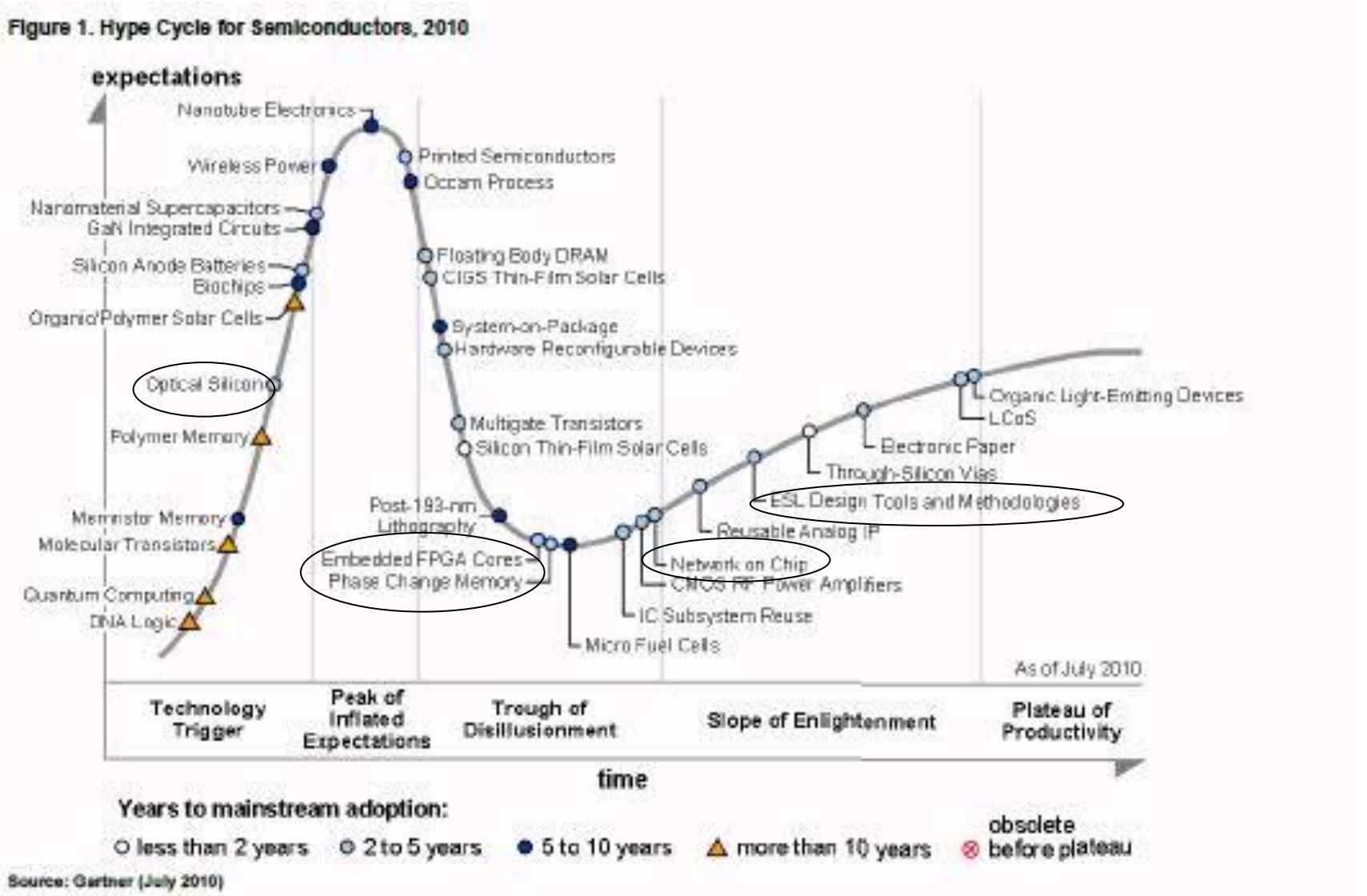
Scope of memories: The whole spectrum



Computing with Memory

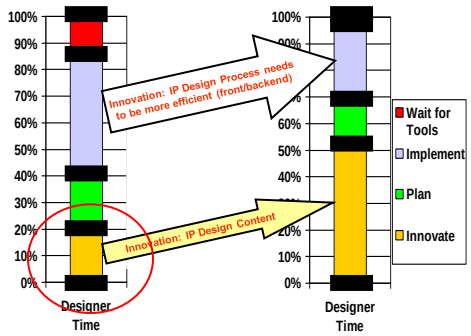
- Potential to solve a systems bottleneck of memory bandwidth in a flexible way for data intensive applications/tasks
  - Reconfigurable on demand (compilation step)
  - Robust (ECC protected)
  - Verification advantage
  - Fast and enables implementation of explicit parallelism
    - Utilizes bit level parallelism.
- One must think about computational memory from a systems and software perspective (How and where will computational memory fit in and what functions can it implement).. Active Ongoing research…

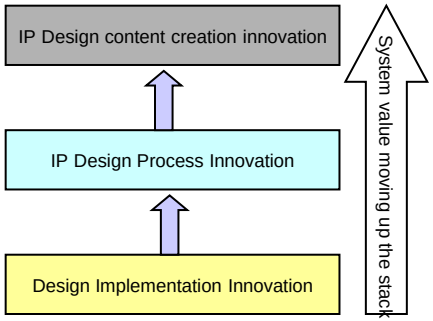
Life Cycle: *Technology Trigger, Hype, Disillusionment, Enlightenment, and Commoditization*



Summary

- Information technology landscape is changing dramatically
  - Value is in innovating across the entire stack and increasingly higher up in the stack
  - Key problems remain to be solved in technology, design and automation as technology continues to scale
  - Significant emerging opportunities in new ways to solve system bottlenecks at every levels: Logic, Architecture, Memory.
  - In last several years, life became very challenging but also very interesting as the ride has gotten a lot choppier…
  - With challenges and opportunities abound, Winners and Losers will be decided by organizations that grab these challenge and innovate their way out of the current dilemmas…





Memoirs of a DA Engineer

- A lawyer is flying in a hot air balloon over lake tahoe and realizes he is lost.

- He reduces height and spots a man down below. He lowers the balloon further and shouts, "Excuse me, can you tell me where I am?"

Memoirs of a DA engineer

- The man below said, "Yes, you're in a hot air balloon, hovering 30 feet above this field."

  "You must be a DA engineer," said the lawyer.

  "I am," replied the man. "How did you know?"

- "Well," said the lawyer, "everything you have told me is technically correct, but it's of absolutely no use to anyone."

  The man below said, "You must be a lawyer."

  "I am," replied lawyer, "but how did you know?"

- **"Well," said the man, "you don't know where you are, or**
- **where you're going, but you expect me to be able to help.**

- **You're in the same position you were before we met, but now it's my fault."**