New Dimensions in Performance

Harnessing 3D Integration Technologies

Kerry Bernstein

IBM T.J. Watson Research Center

Yorktown Heights, NY

9 September, 2008 **Fort Collins, CO**

"Escher Envy" courtesy of David Bryant

3D Press Release

EE Times:

3D interconnect firms form development, marketing alliance

Peter Clarke
(02/28/2005 5:57 AM EST)
URL: http://www.eetimes.com/showArticle.jhtml

LONDON — Two companies developing differe
Semiconductor Corp. and Ziptronix Inc., have s

The companies said they wanted their alliance t
developers, and manufacturers.

ElectronicNews
Your World in Real Time
February 14, 2006

Future Certain for 3D Chip Scale Packaging

Online Staff – Electronic News, 5/28/2004

There is no doubt that 3D chip scale packaging technology will play an

Ziptronix Targets Vertical Scalability

Ziptronix has unveiled what it is calling the world's first successful direct bon technology to enable 3D chips

The method can replace through-die vias used in previous 3D chip integration

ElectronicsWeekly.com

ISSCC: IMEC outlines 3D chip progress

Tuesday 17 February 2004

EW Daily News 17/02/2004 10:26:27 - ISSCC: IMEC outlines 3D chip progress
ISSCC: IMEC outlines 3D chip progress David Manners
Belgian microelectronics research organisation IMEC has raised the questio
of 3D chips at the International Solid State Circuits Conference (ISSCC) in San Francisco.

Wall Street Journal: IBM Touts Breakthrough in 3-D Chips

By WILLIAM M. BULKELEY

April 12, 2007; Page B3

International Business Machines Corp. said it achieved a breakthrough in developing a three-dimensional semiconductor chip that can be stacked on top of another electronic device in a vertical configuration long sought by engineers to reduce size and power use.

Agenda

1) Workloads
2) The Memory Wall, Bandwidth, and Latency
3) The Technologies of 3D Integration
4) “The future ain’t what it used to be.”
5) Summary
6) Human Scaling

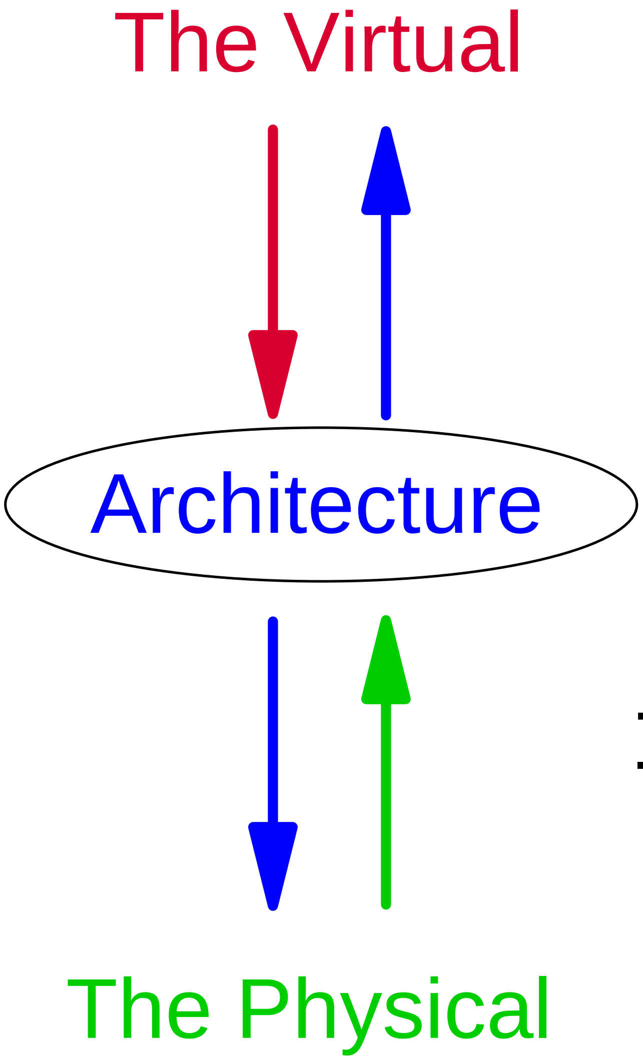
Computer Workloads & Thru-put

Workloads – What do Computers do?

- **Scientific** **(i.e. Lawrence Livermore National Labs)**
  - Highly regular, predictable patters allows streaming data from cache to processor
  - Performance is directly proportional to bus bandwidth
  - High utilization, full data bus at all times
- **Commercial** **(i.e. Starbucks)**
  - Unpredictable irregular patterns
  - Miss rate follows Poisson process (random)
  - Requires low bus utilization to avoid clogs in the event of a burst of misses (usually 30% bus utilization)
- **Both application spaces need BW, but for different reasons**

The Software Stack:

- System
  - Hypervisor
  - Operating System
- Applications Layer
- Program
- Compiler
- Machine Language



The Virtual

Architecture

The Physical

Architecture:

A fully-specified unambiguous contract

The Hardware Stack:

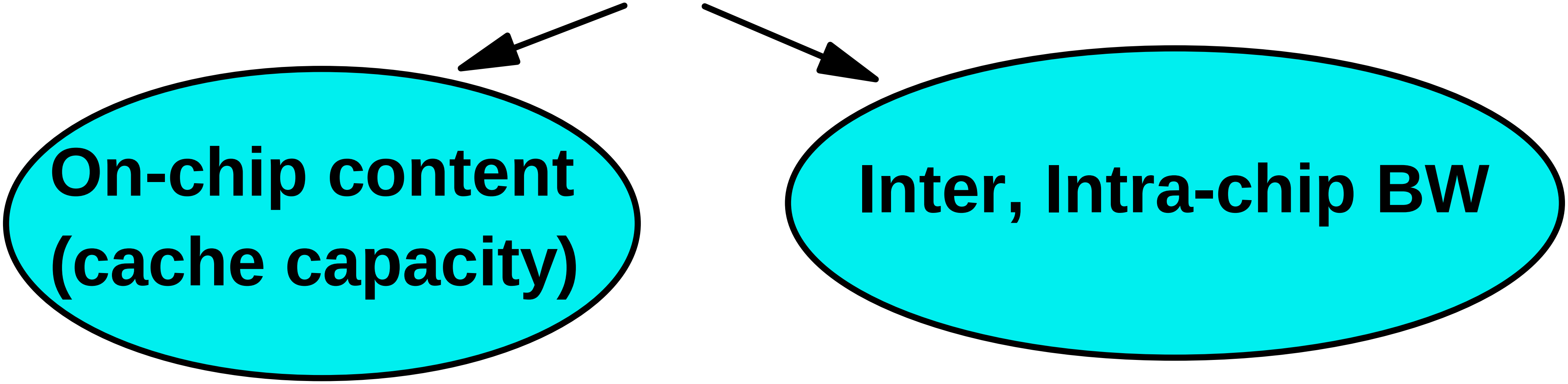
- Logical Level Description
- Machine Organization
- Schematic Representation
- Circuit Design
- Physical Design
- Device Level (transistors)
- Atomic Level

Processor Cores and Memory Subsystems

The New Units of Design

(Systems/Thread) x (Threads/Core) x (Cores / Die)

Puts pressure on Memory Subsystem, Communication



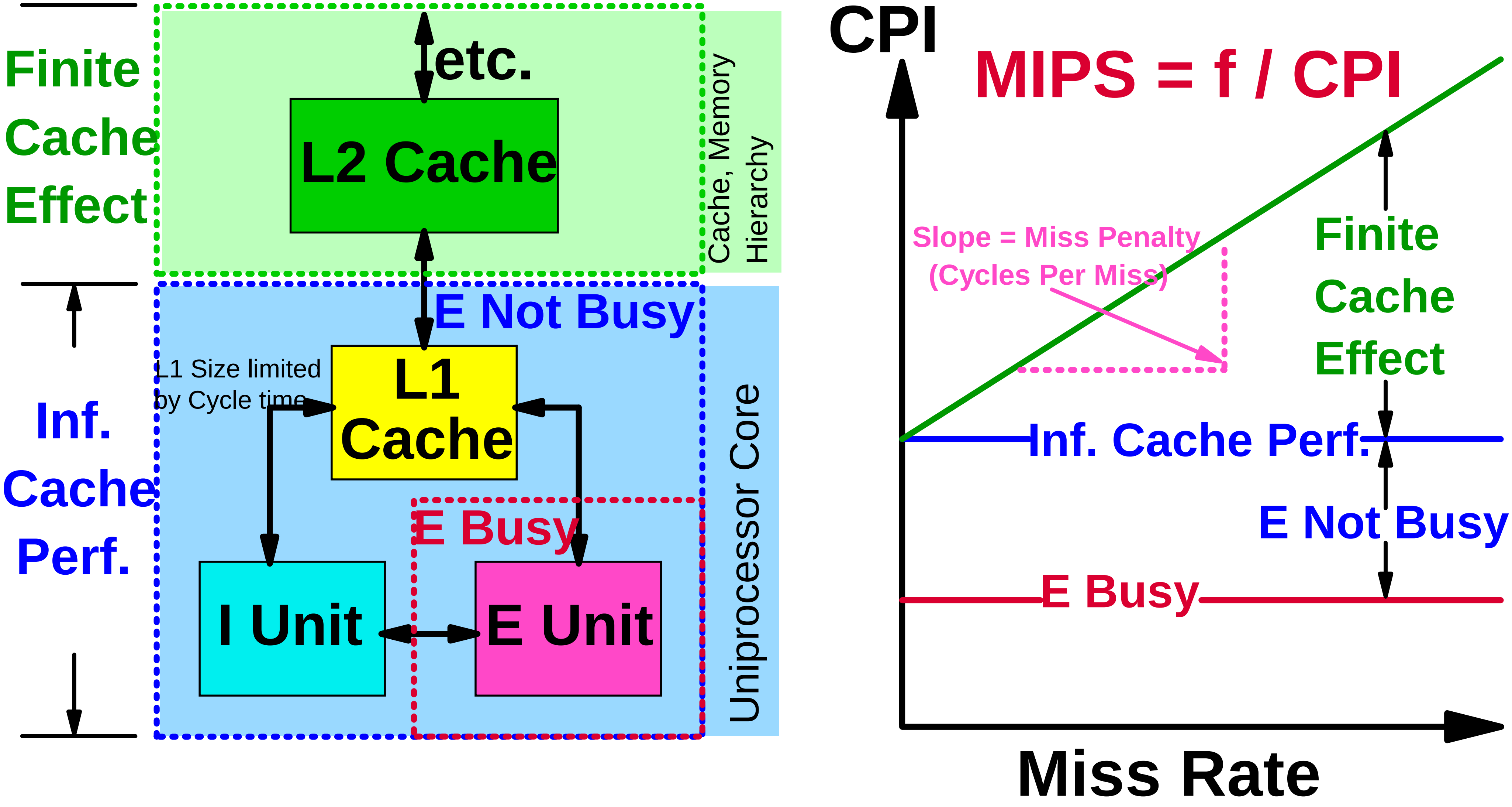
On-chip content (cache capacity)

Inter, Intra-chip BW

Integration Focus moves from the device and circuit to core

Components of Processor Performance

From ISCA '06
Keynote address by
Phil Emma, IBM

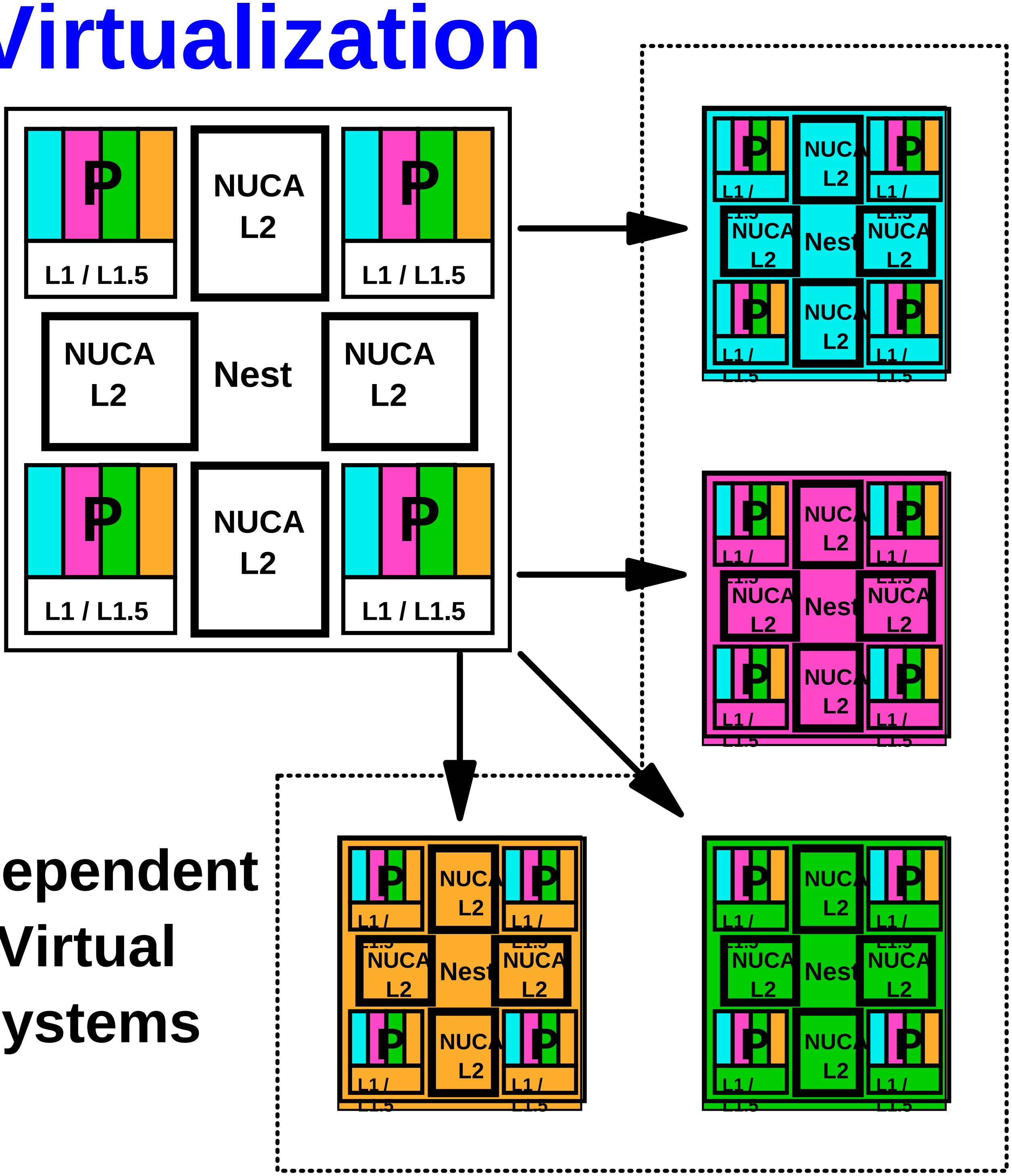


Delay is sequentially determined by a) ideal processor, b) access to local cache, and c) refill of cache

Virtualization

From ISCA '06
Keynote address by
Phil Emma, IBM

Real
System

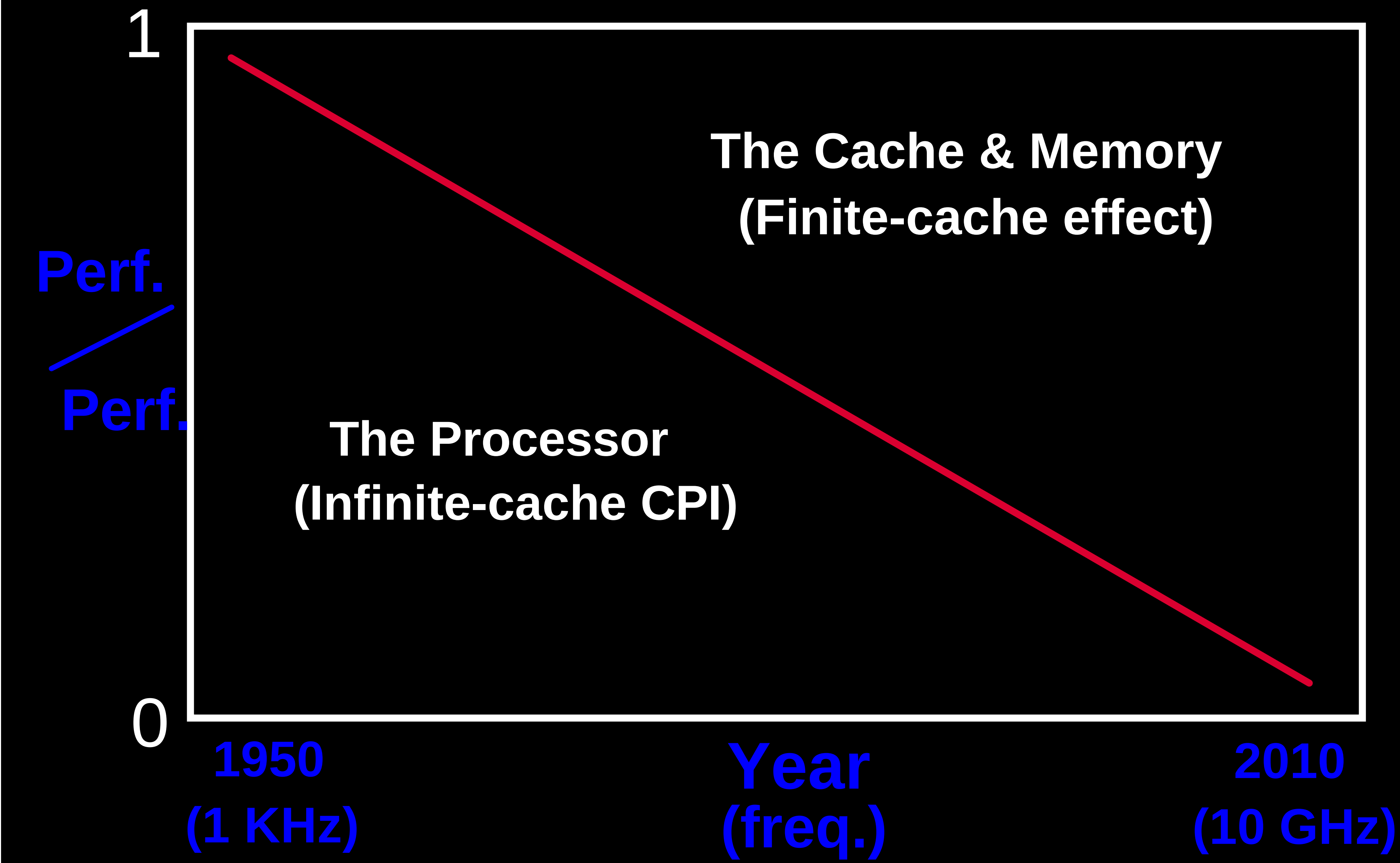


Independent
Virtual
Systems

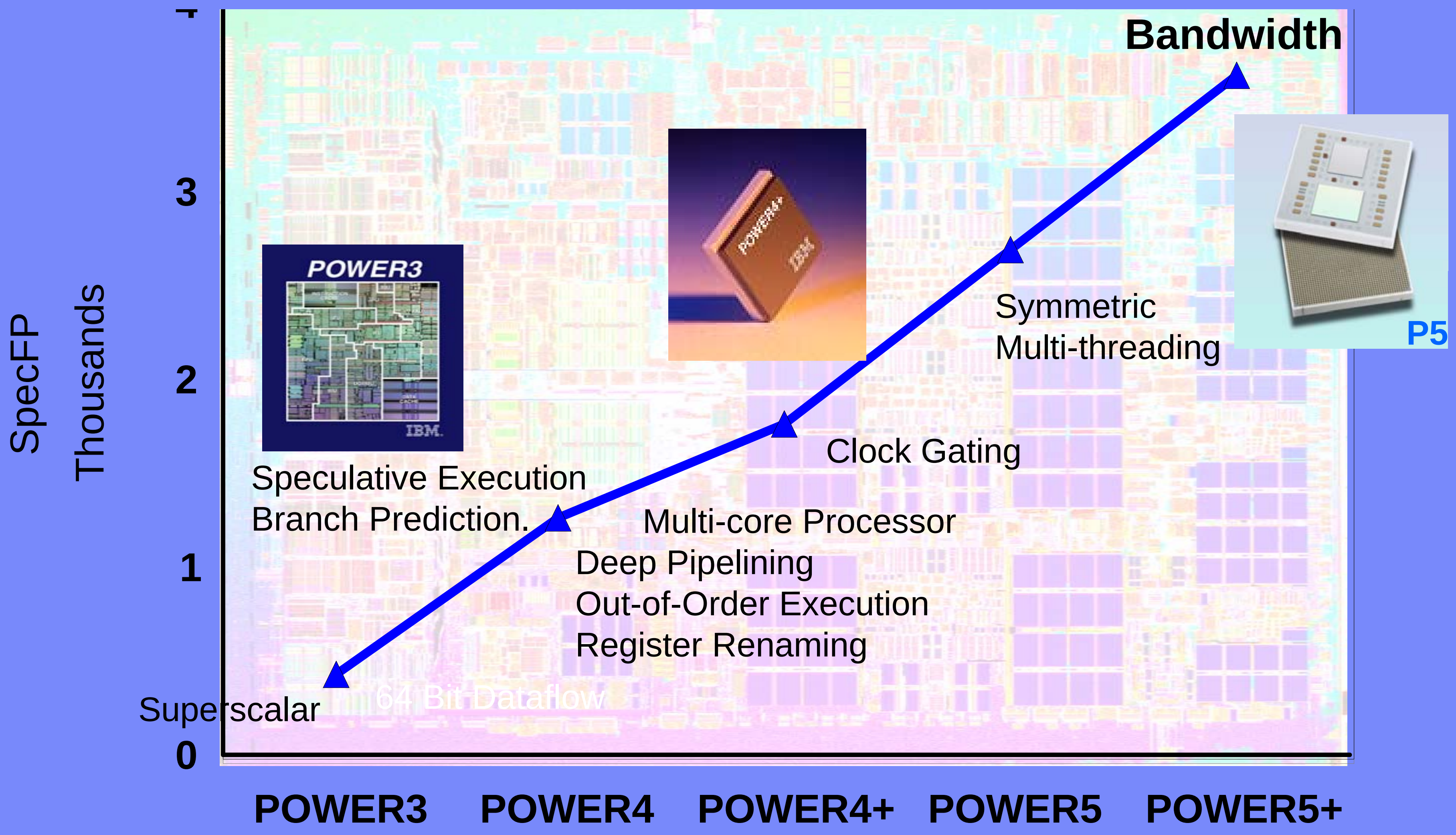
"Looks like"
4 independent
systems, each
with 16 cores!

What Dominates System Performance?

From ISCA '06
Keynote address by
Phil Emma, IBM



POWER Series Architectural Perf Contributions

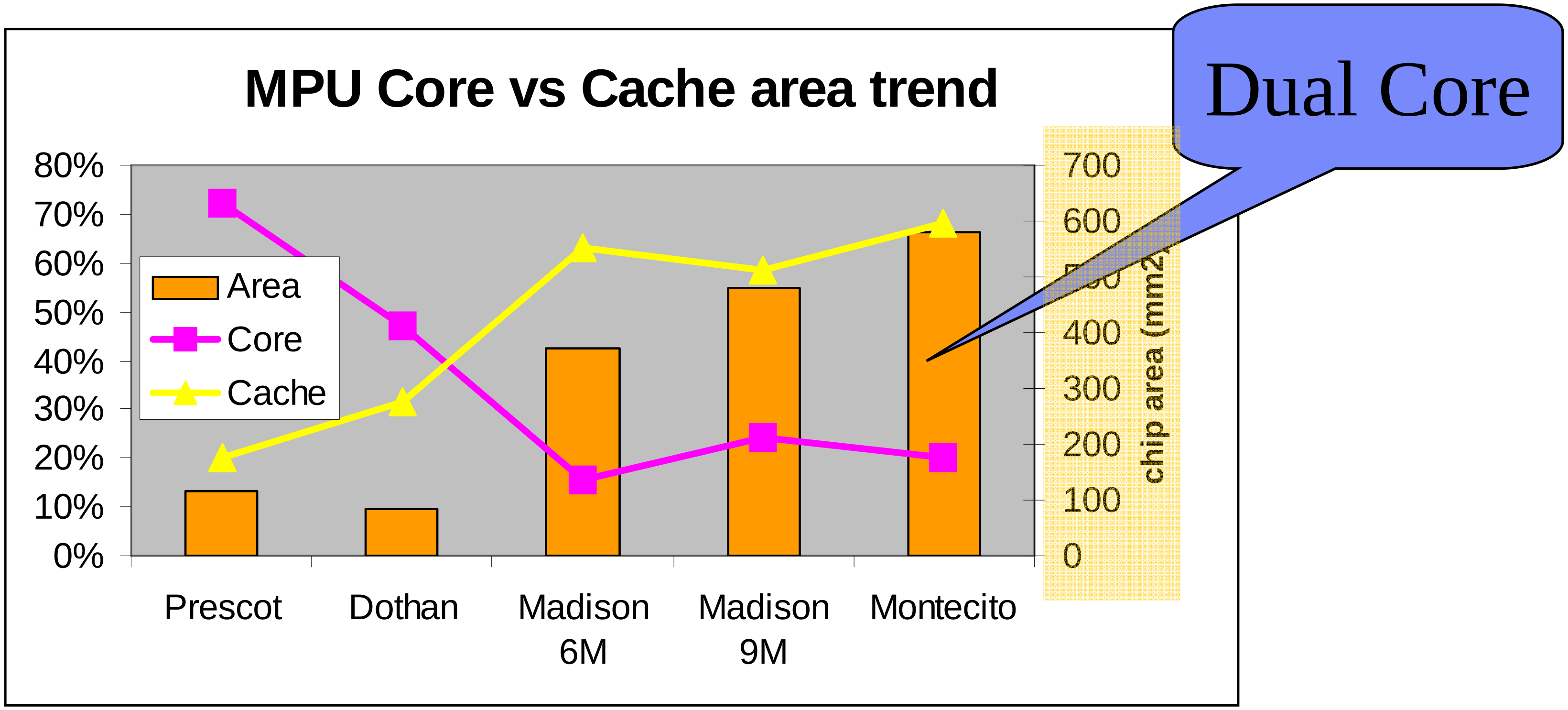


Processor Perf — Data Delivery

Transaction Rate Dependence

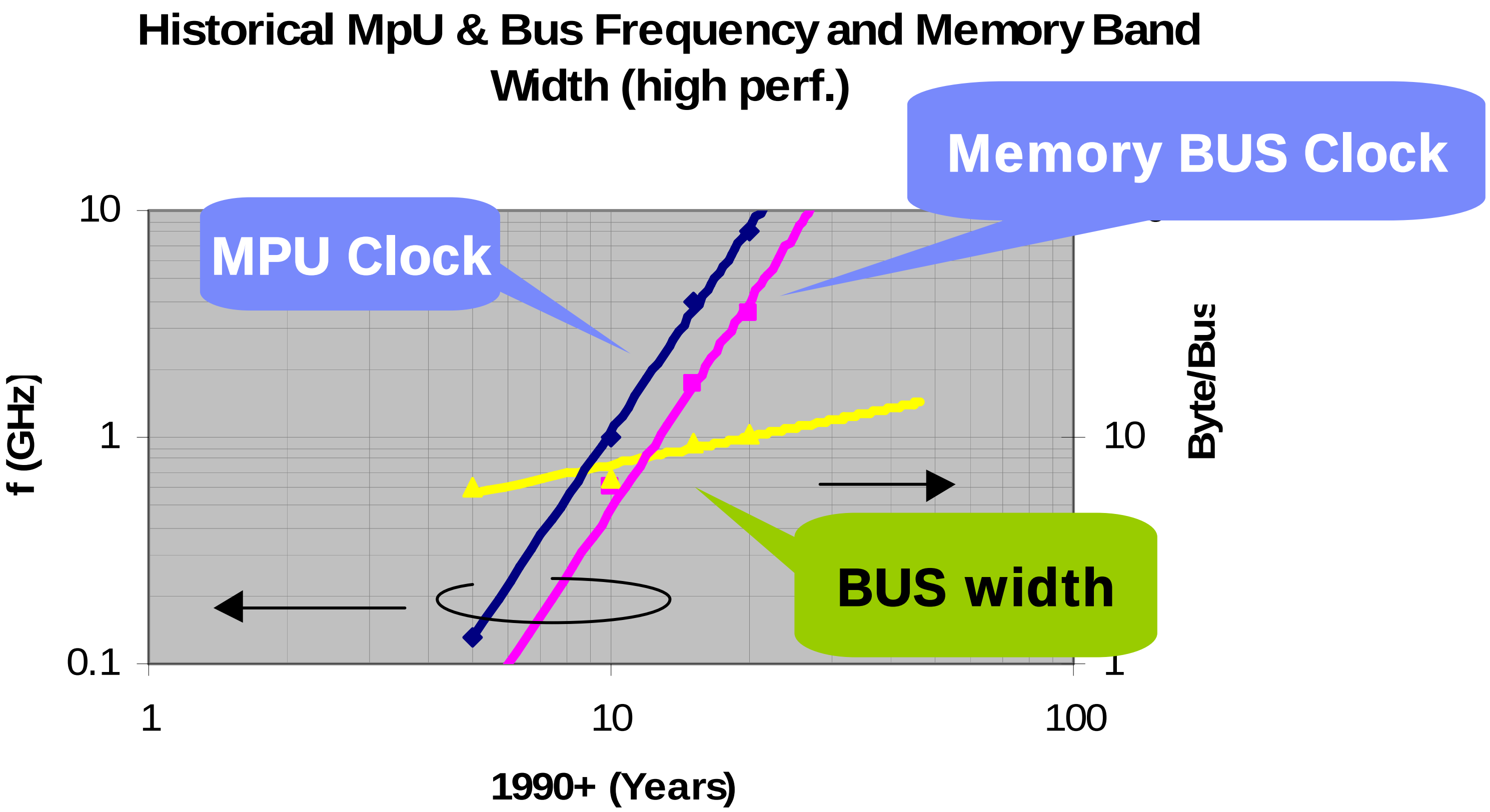
Increasing Cache Size Drives Chip Size

Floorplans source: P. DeMone, “Sizing Up the Super Heavyweights,” *Real World Technologies* Report, 9/17/2004



- Growing data sets will increasingly stress cache size
- Multi-core floor planning and SRAM concerns will halt cache size growth to maintain manageable chip size

Frequency Drives Datarate

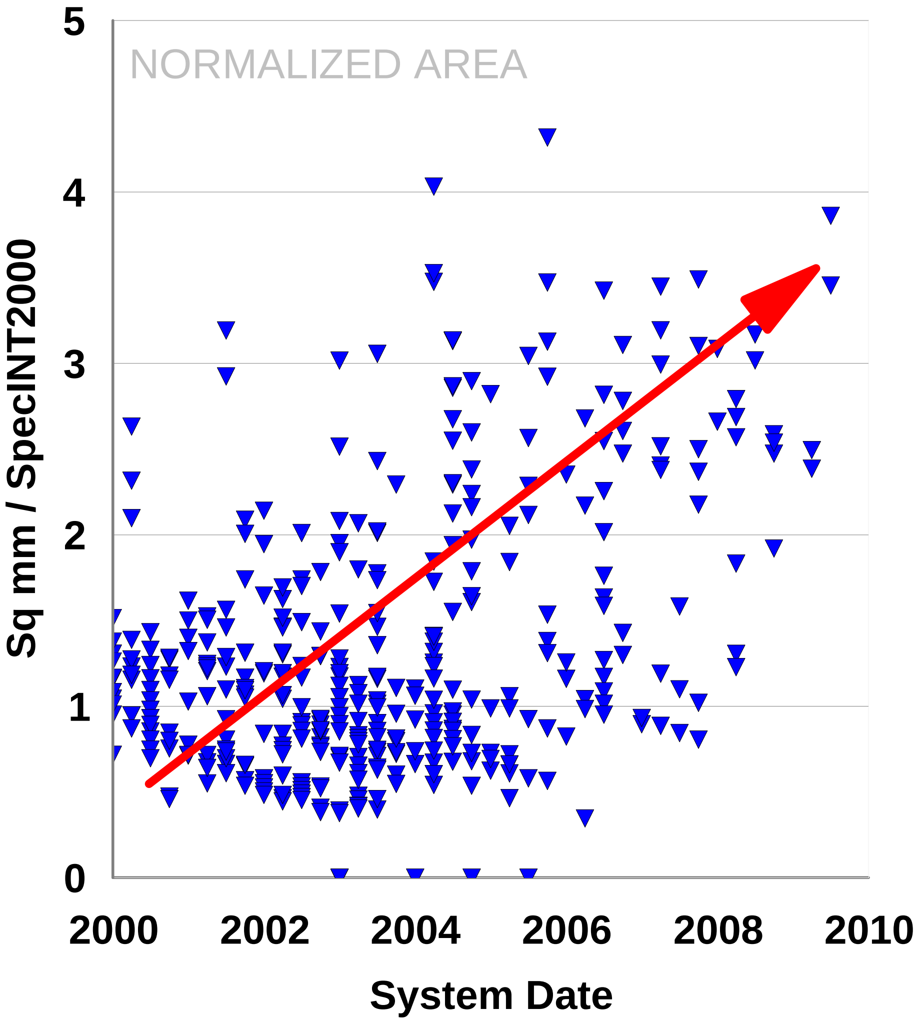


- Data bus frequency follows MPU frequency at a ratio 1:2 roughly doubling every 18 to 24 month
- Data bus band width shows only a moderate increase
  - Data bus transfer rate is basically scaled by bus frequency
- **When clock growth slows, BUS data rate growth will slow too!**

Architecture Net

- **Growing the number of cores/chip increases demand for bandwidth**
- **Transaction retirement rate dependence on data delivery is *increasing***
- **Transaction retirement rate dependence on λP performance is *decreasing***

Die Area Increase



1) Architecture overhead increasing area of die

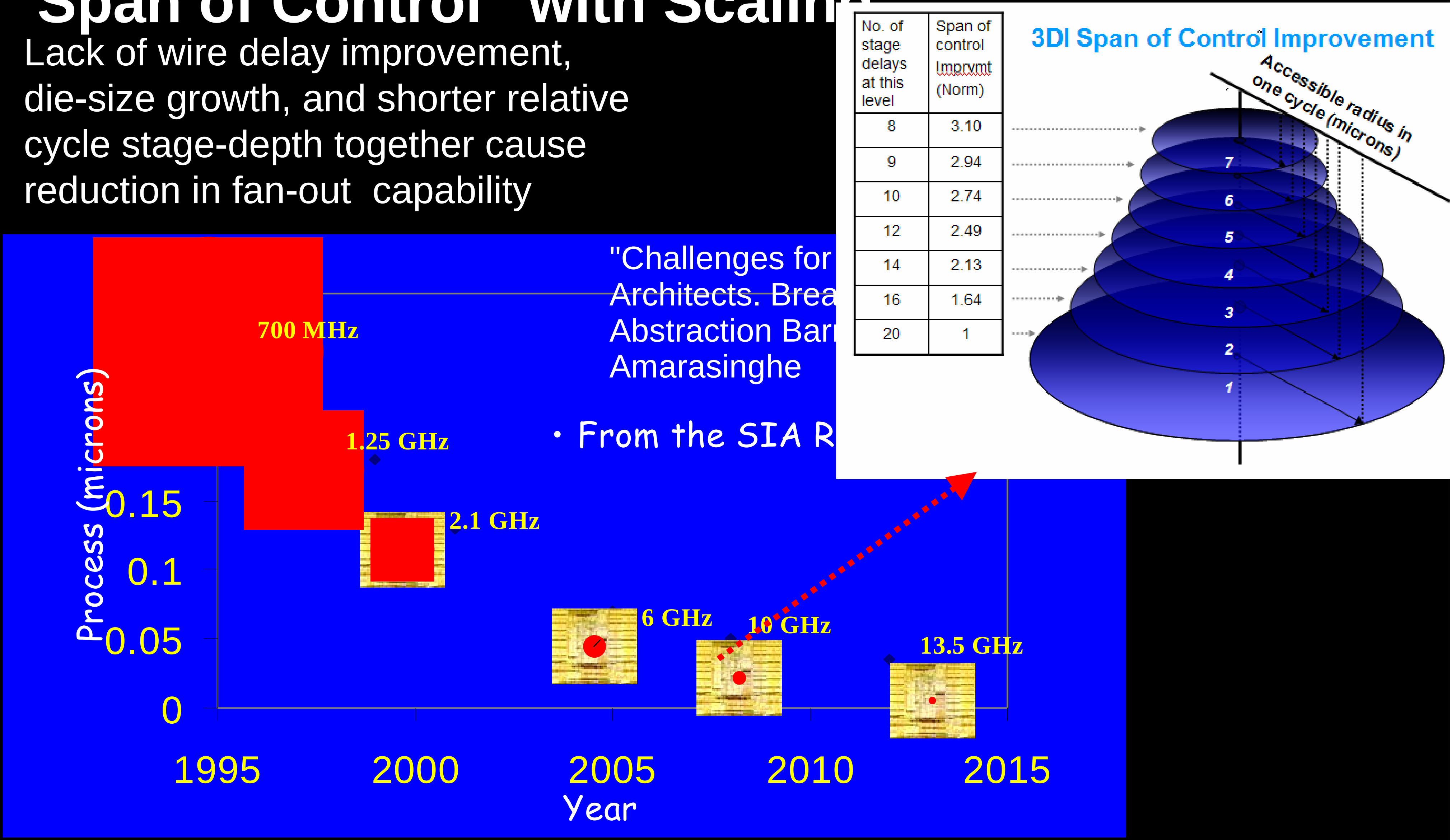
2) Accessible portion of chip over normalized cycle time is decreasing generation over generation

3) Deeper Pipes are decreasing delay per cycle

Performance is expensive when left to architects!!

“Span of Control” with Scaling

Lack of wire delay improvement, die-size growth, and shorter relative cycle stage-depth together cause reduction in fan-out capability



No. of stage delays at this level	Span of control Imprvmt (Norm)
8	3.10
9	2.94
10	2.74
12	2.49
14	2.13
16	1.64
20	1

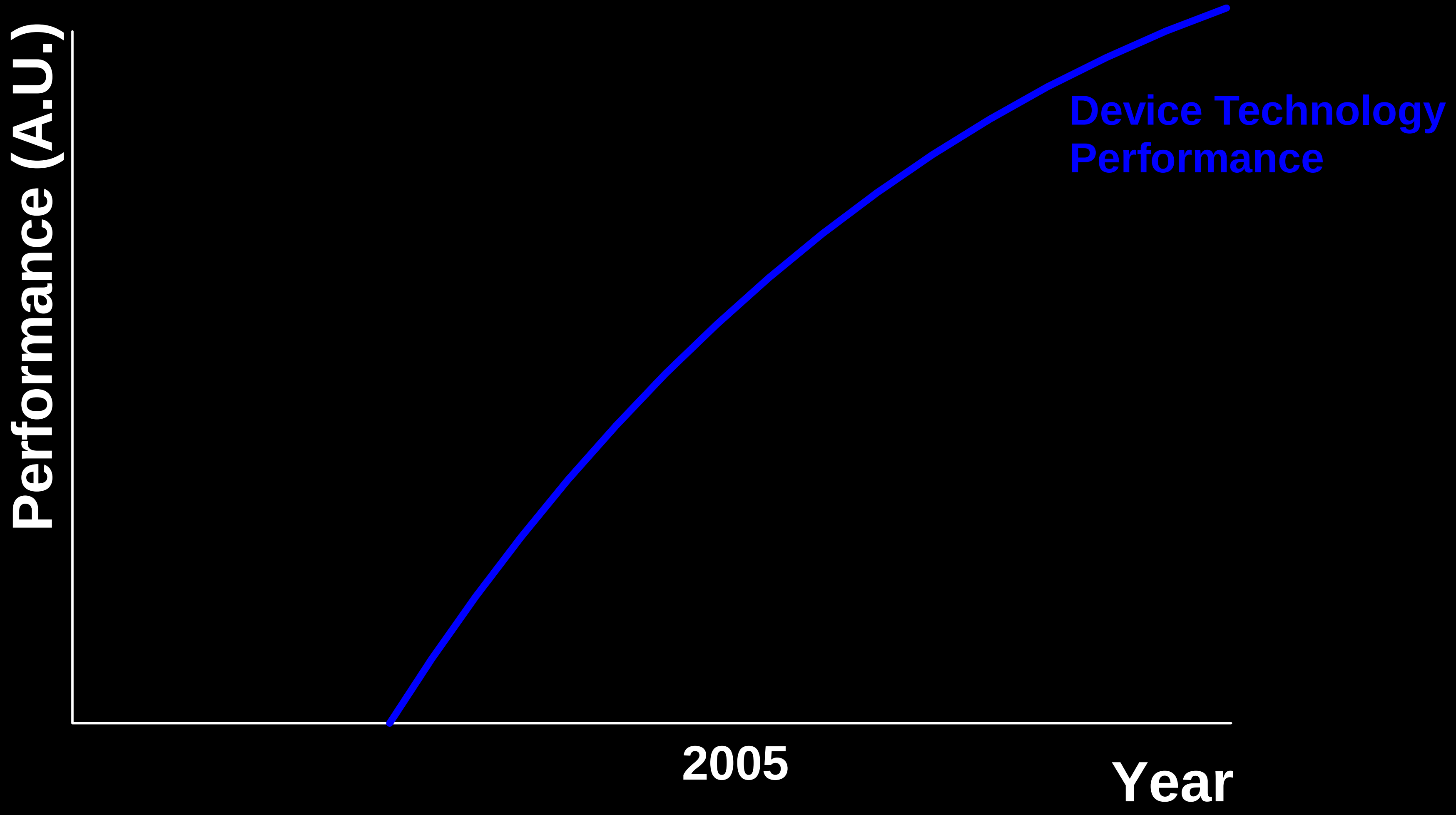
"Challenges for Architects. Brea Abstraction Barr Amarasinghe

- From the SIA R

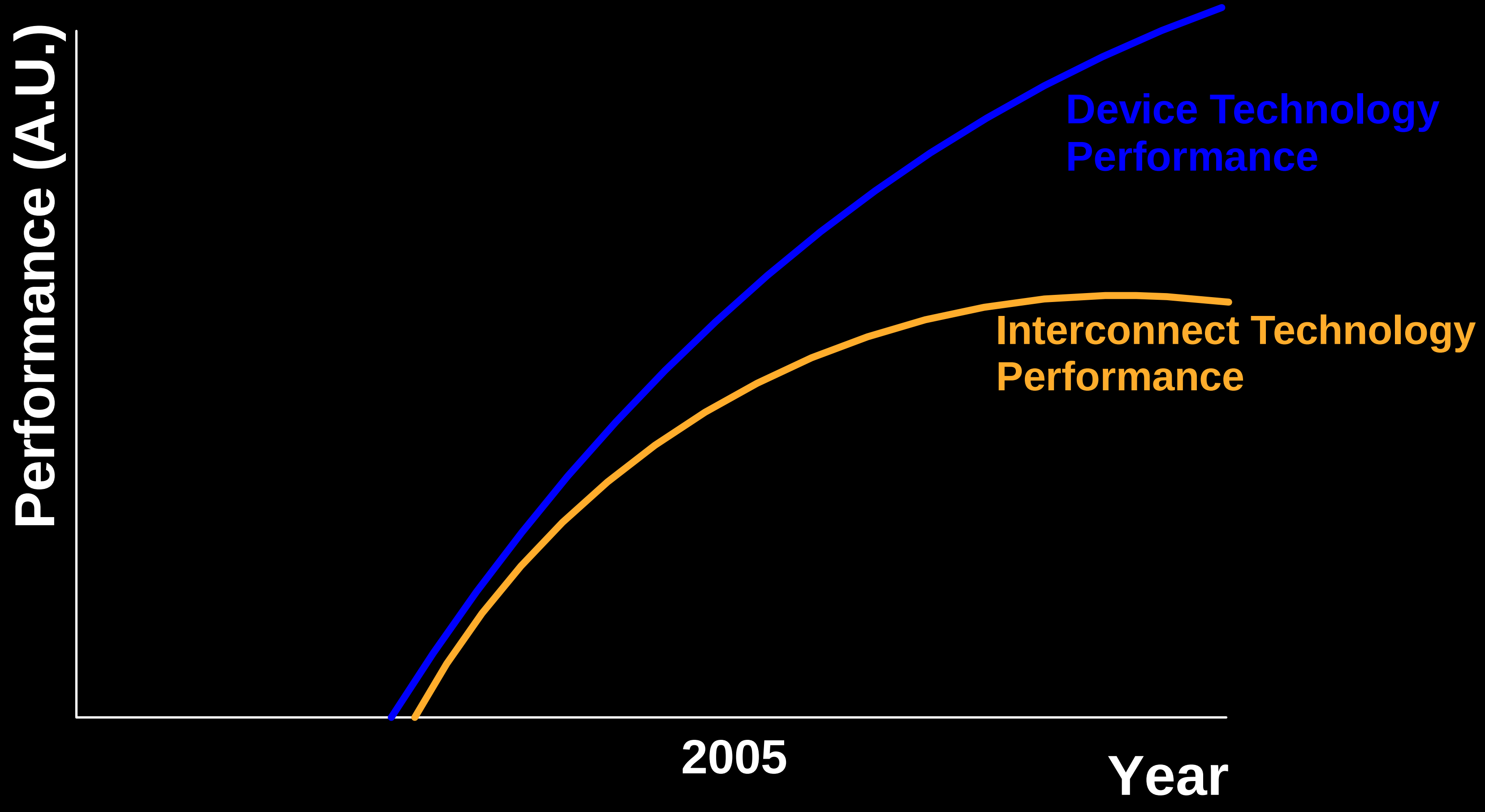
- Perfect Storm: (a)Wire non-scaling; (b) die size growth; (c)Shorter FO4 stages
- Power Cost of Cross-Chip Latency increase

First, a look at a “coincidence”……

Device performance (i.e. I/C_G) continues to improve, however at a decreasing rate…….

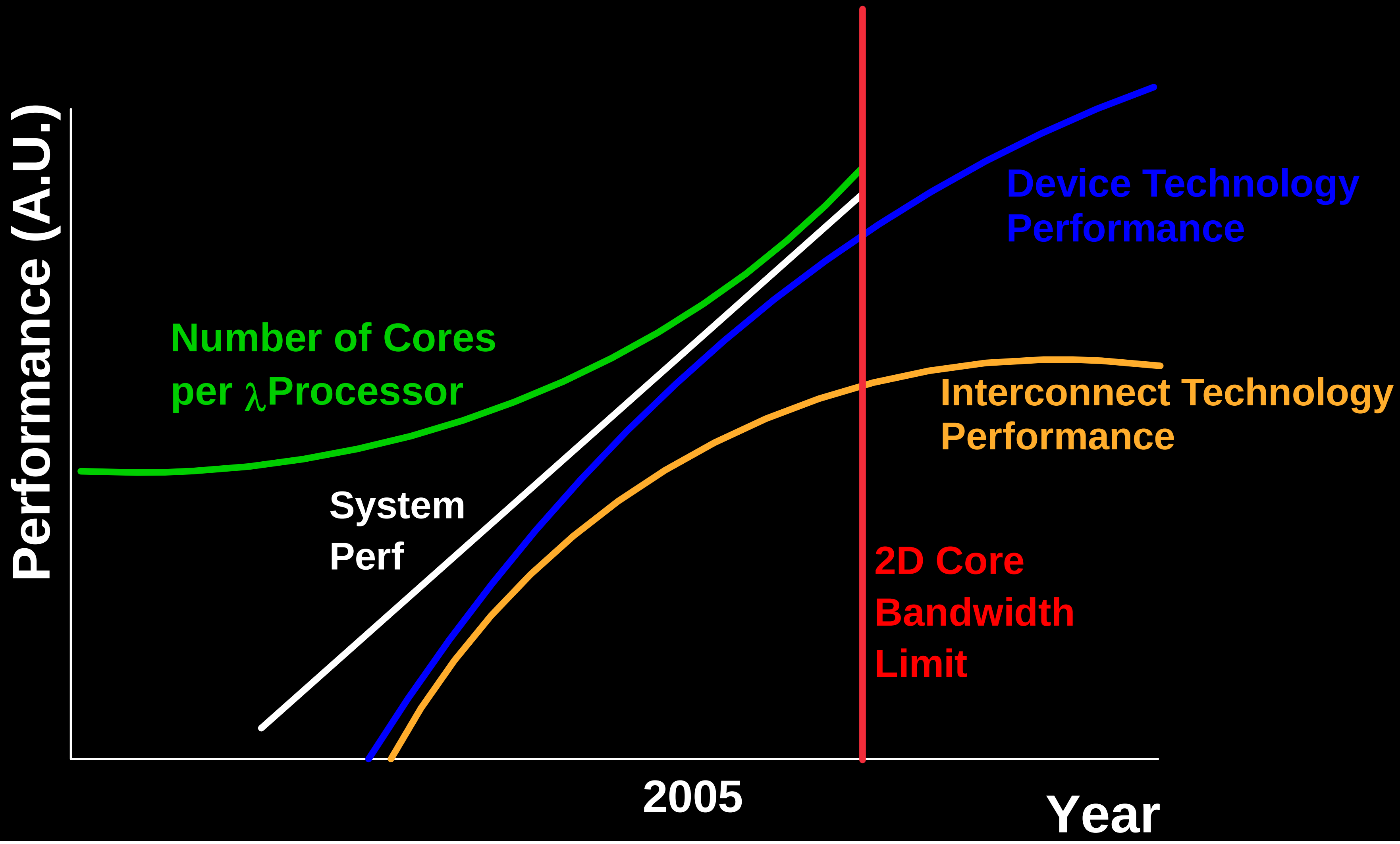


Despite constant infusion of new materials and processes however, interconnect technology performance has at best remained flat.



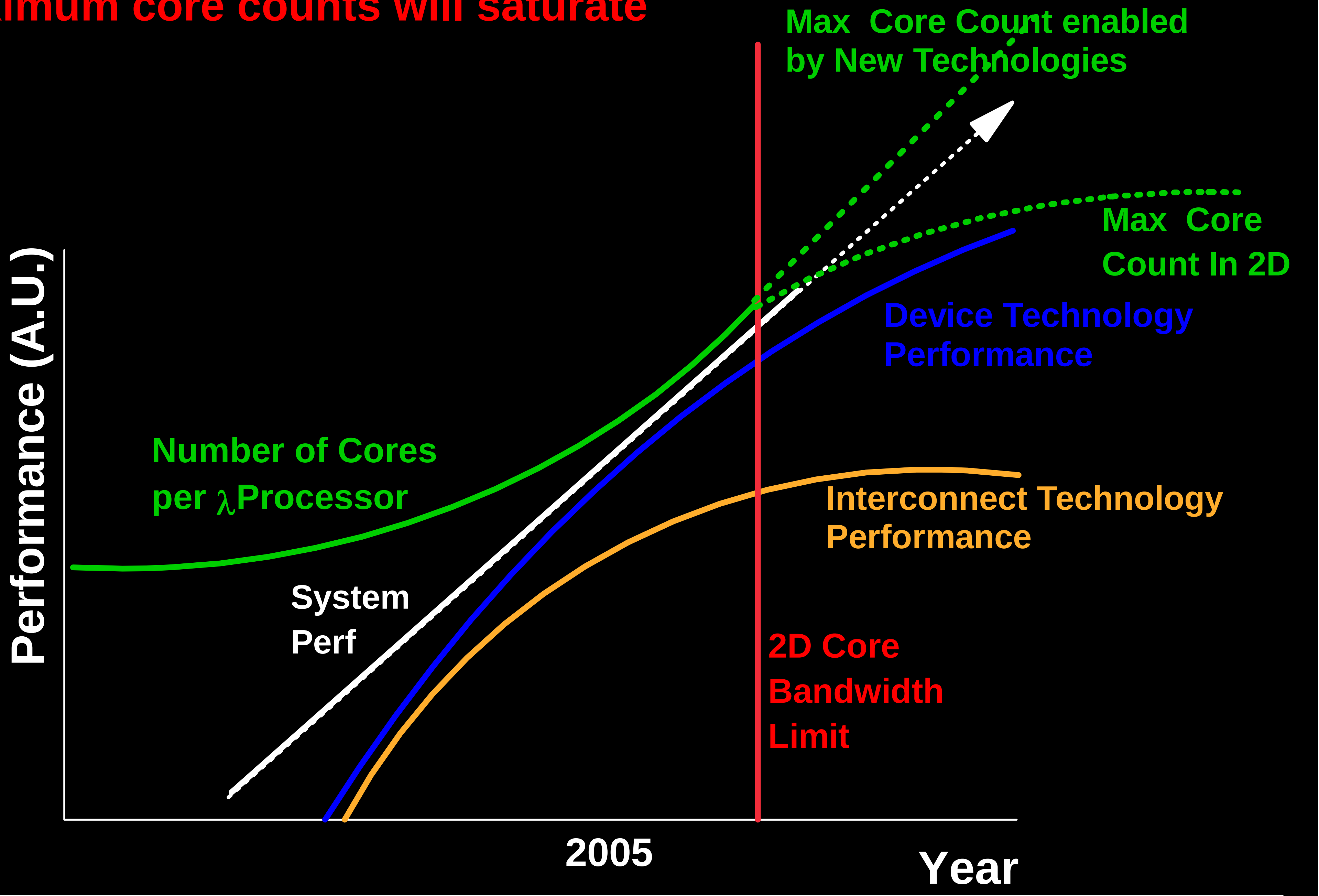
Scaling has increased the divergence between FEOL and BEOL contributions to performance improvement

As effective distances on chip increased due to interconnect, cores/chip has begun to climb. The bandwidth needed to feed these cores will ultimately **limit** number of cores



System Performance improvement is sustained more by the number of cores rather than by the performance of each core

Without more bandwidth at low latencies, maximum core counts will saturate



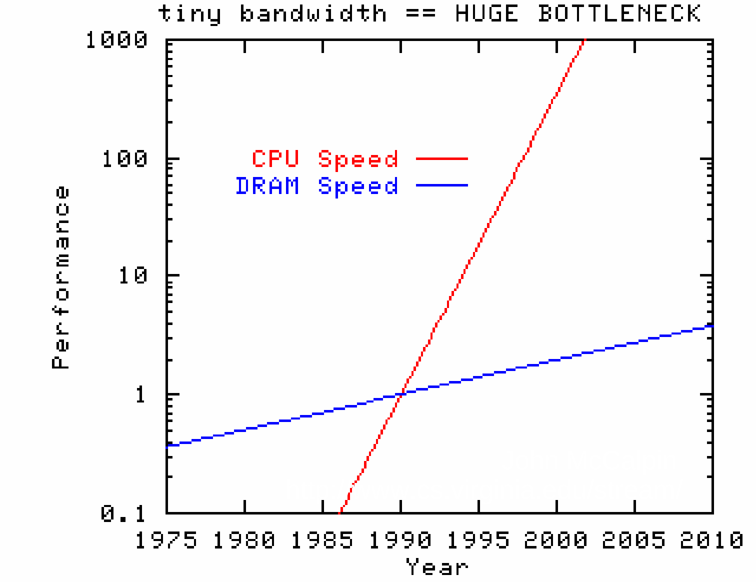
3D extends transfer of performance from the device to the core level

The Memory Wall, Bandwidth, and Latency

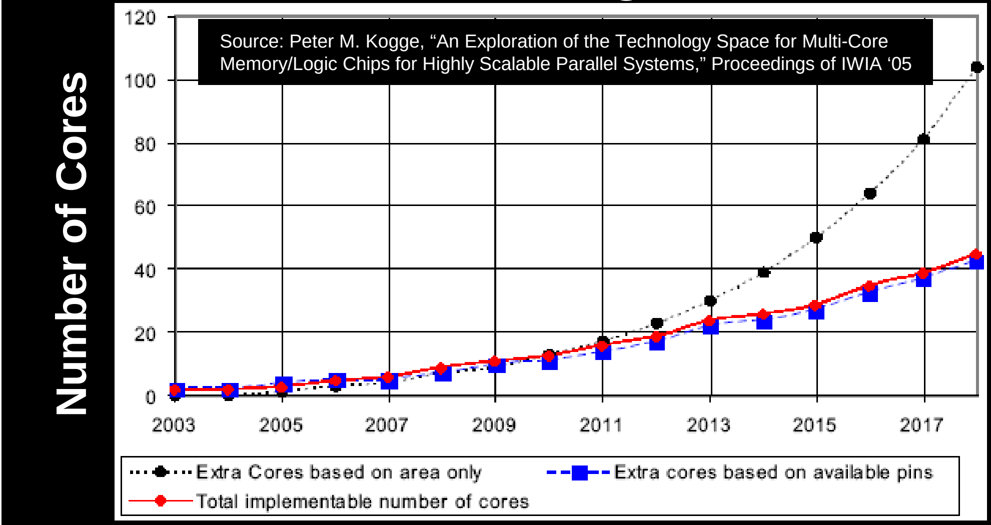
Getting over the Memory Wall

Microprocessor Architectures Fundamental Bus Limits

Latency Challenge



Bandwidth Challenge

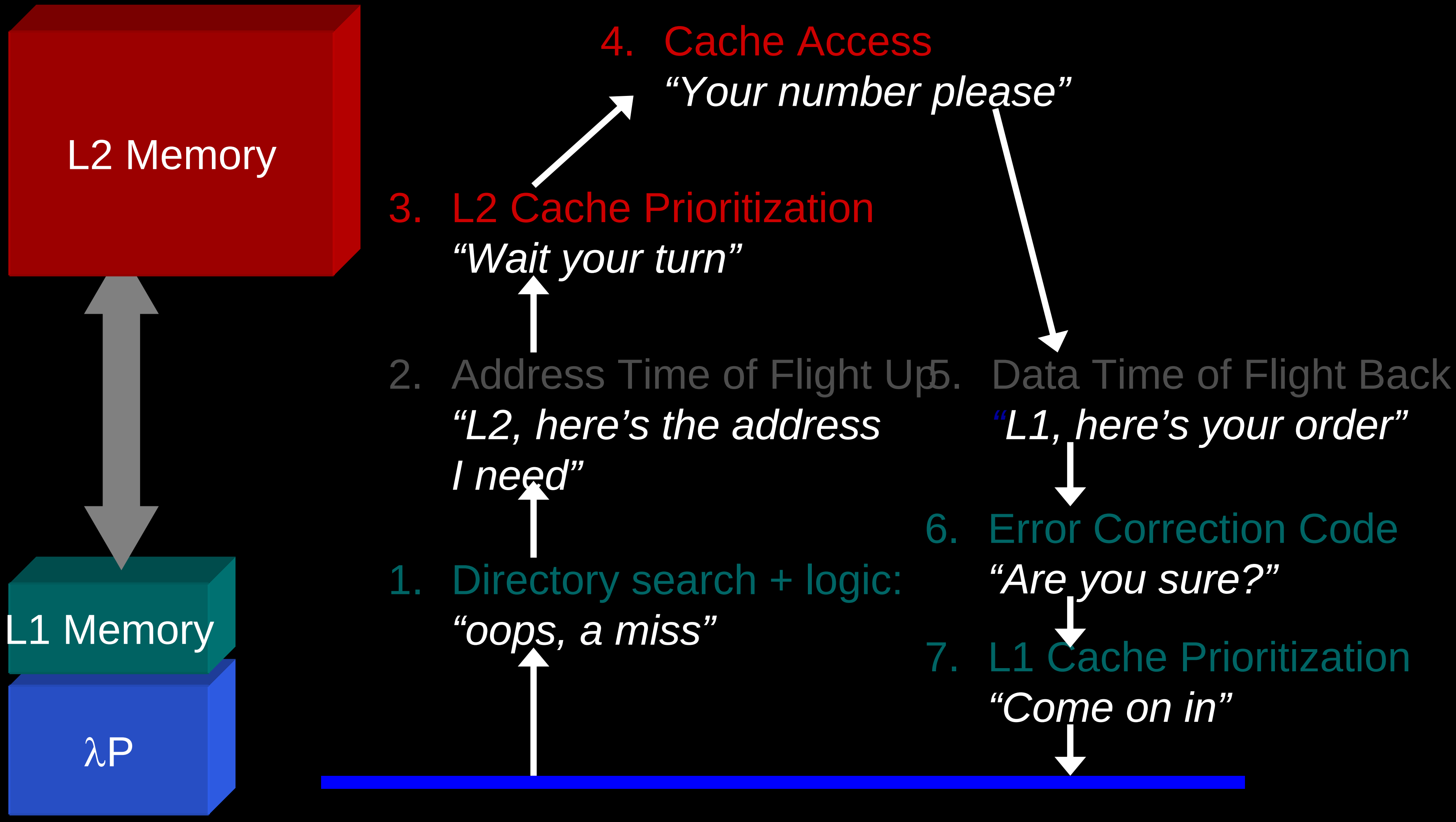


Source: Peter M. Kogge, “An Exploration of the Technology Space for Multi-Core Memory/Logic Chips for Highly Scalable Parallel Systems,” Proceedings of IWIA ‘05

- **Processor speed has increased much quicker than memory access**
- **Result: λP’s data appetite has grown quicker than ability to feed it.**
  - What needs higher BW?
    - •Multi-cores with limited cache
    - •Multi-threading
    - •Virtualization
- **Increasing “cores per chip” addresses memory latency. Core count Limit after 2010 will be from pins used to provide memory bandwidth**
  - The “Memory Wall” is back with a vengeance

Architecture

Cache Miss Penalty Calculation



4. Cache Access
   "Your number please"

3. L2 Cache Prioritization
   "Wait your turn"

2. Address Time of Flight Up
   "L2, here's the address I need"

1. Directory search + logic:
   "oops, a miss"

5. Data Time of Flight Back
   "L1, here's your order"

6. Error Correction Code
   "Are you sure?"

7. L1 Cache Prioritization
   "Come on in"

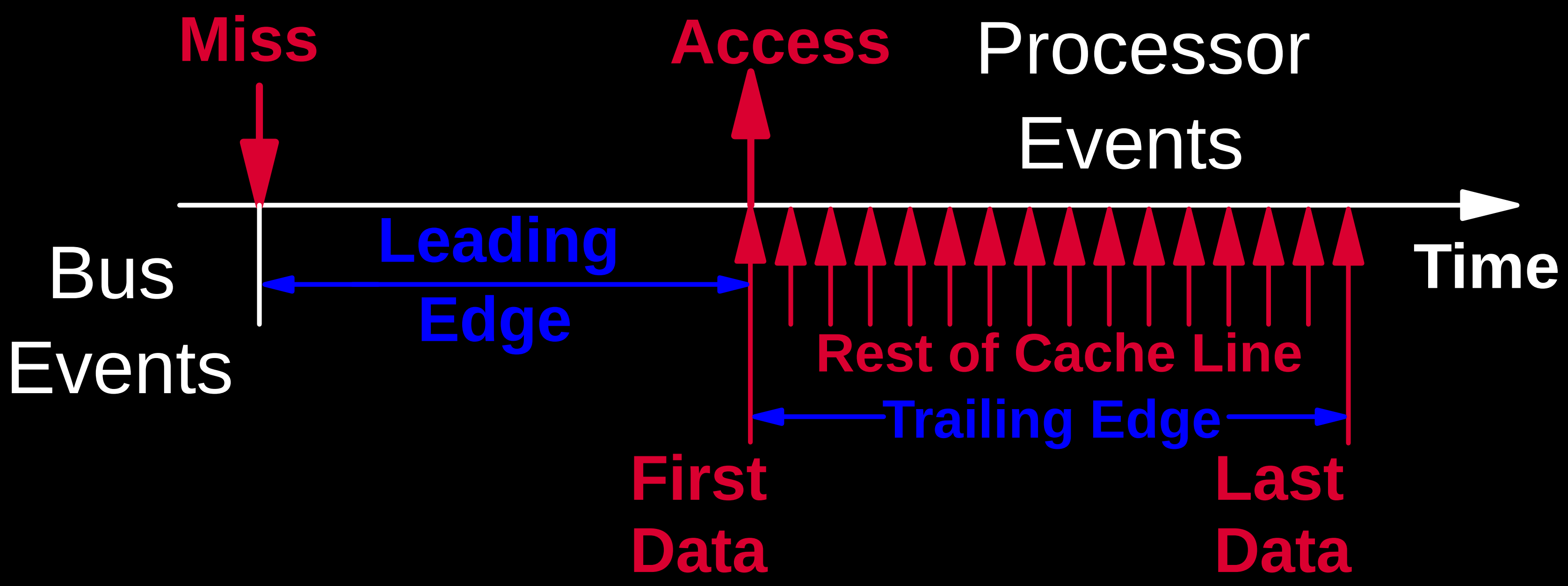
Memory Latency is the delay encountered completing the loop above

Bernstein

From ISCA '06
Keynote address by
Phil Emma, IBM

What Is Bandwidth Used For?

In a computer, it is mostly for handling cache misses:1



Miss Penalty = Leading Edge + Effects(Trailing Edge)

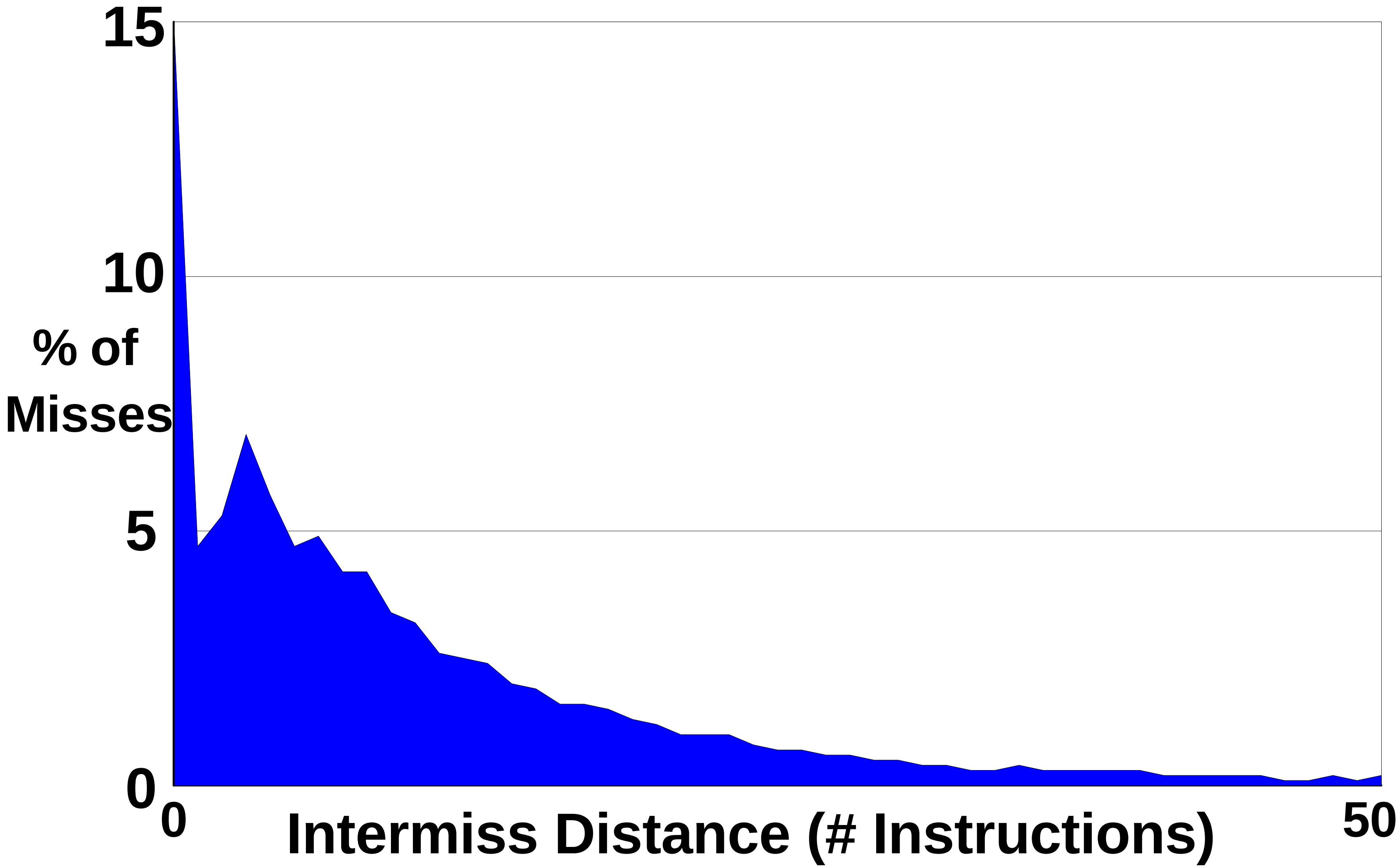
Where

Trailing Edge Effect = (Line Size / Bus Width) x $(F_{(\lambda P)} / F_{(Bus)})$

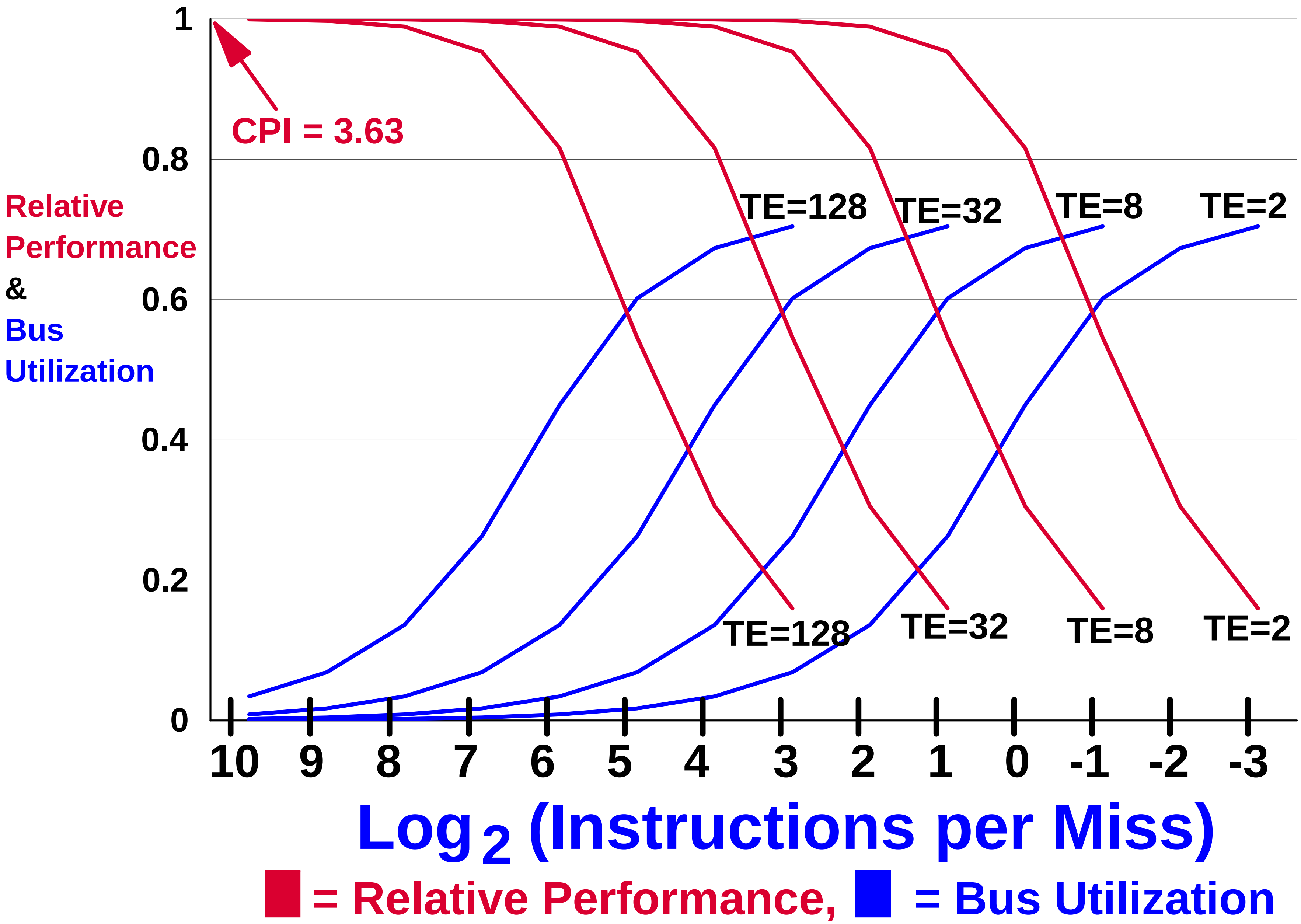
Bus Utilization = (Trailing Edge / Intermiss Distance)

From ISCA ’06
Keynote address by
Phil Emma, IBM

Intermiss Distance Density



Queueing Effects vs. Log Miss Rate



Server Trends are hard on Bandwidth

- **Frequency is no longer increasing**
  - Logic speed scaled faster than memory bus
  - (Processor clocks / Bus clock) consumes bandwidth
- **More speculation multipliers prefetch attempts**
  - Wrong guesses increase miss traffic
- **Reducing line length is limited by directory as cache grows**
  - But doubling line size doubles bus occupancy
- **Cores / die increasing each generation**
  - Multiplies off-chip bus transactions by N / 2*Sqrt(2)
- **More threads per core, and increase in virtualization**
  - Multiplies off-chip bus transactions by N
- **Total number of Processors / SMP increasing**
  - Aggravates queuing throughout the system

3D - Bandwidth and Latency

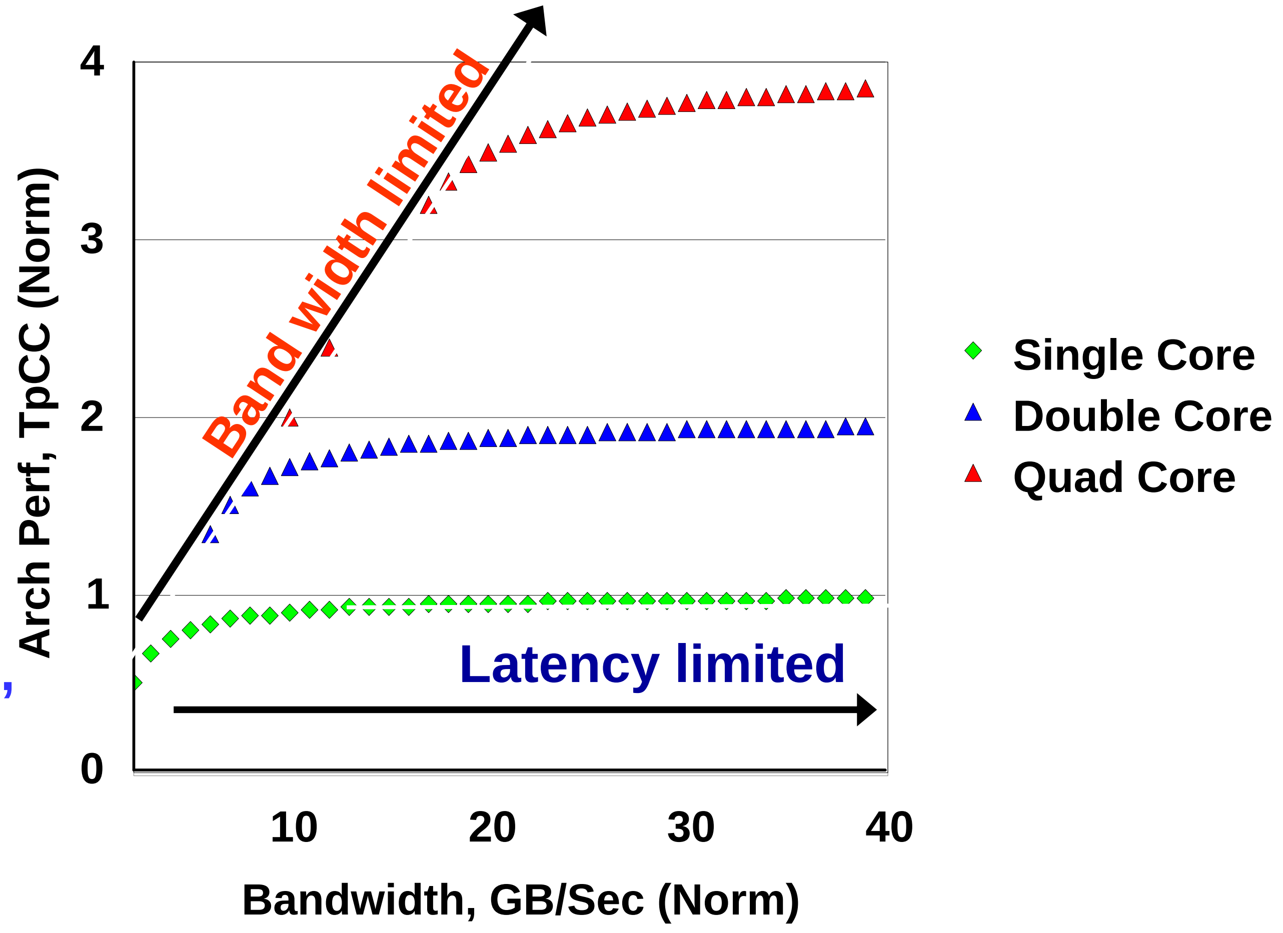
Processor load trade-off between I/O Bandwidth, Bus Latency.

- For generic workloads, uni-processor perf saturates bandwidth benefit, becomes latency-limited.

- As core counts increase, I/O Bandwidth becomes increasingly important

Bandwidth and Latency Boundaries

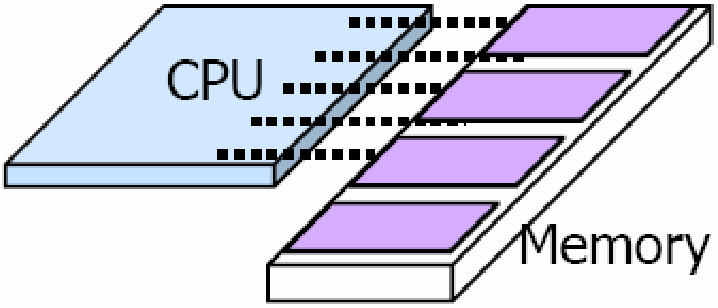
General Purpose Processor Loads



3D opportunity for improving High Perf Compute thru-put in sustaining a higher number of cores per chip

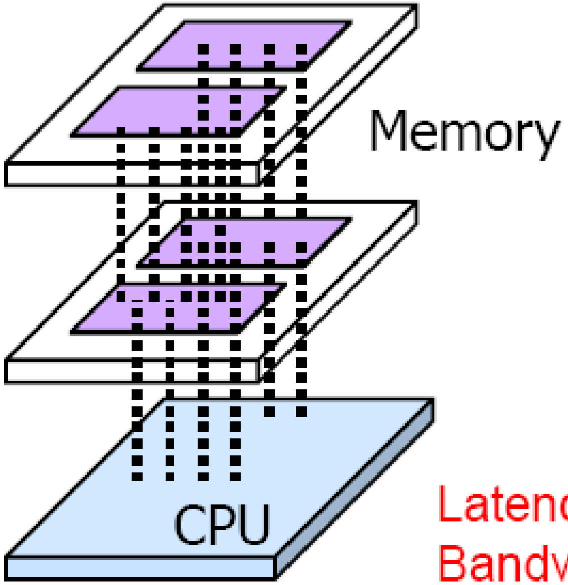
3D Solution

Hierarchical Memory Access



2-D: Connections on the periphery

- Long global connections
- CPU to off-chip main memory with latency and misses



3-D: Connections across the area

- Connections short + vertical
- Suitable for high-bandwidth and vector operations
- No pin cost, large block access of data

Latency: Important for random access (servers, e.g.), single core

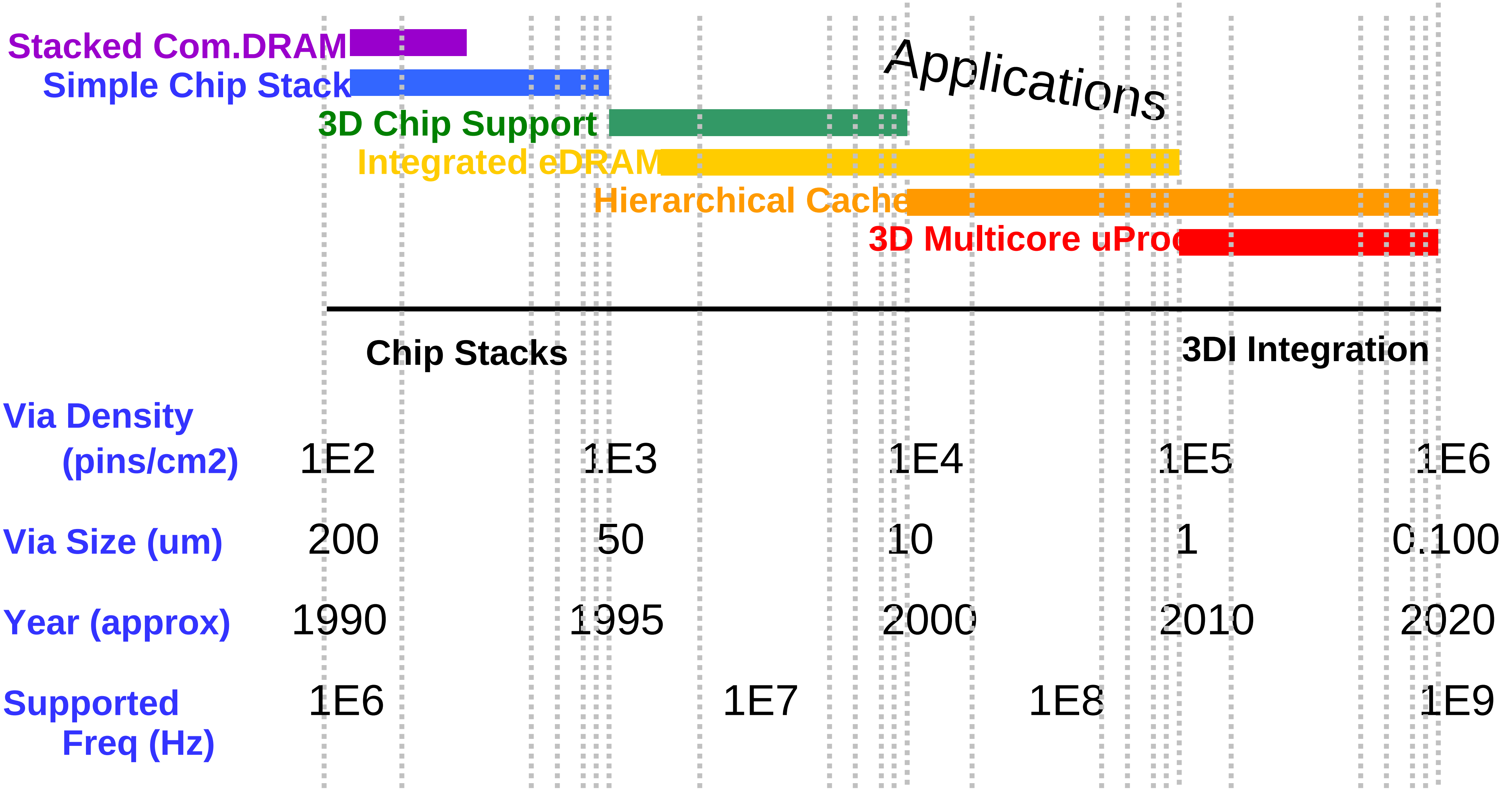
Bandwidth: Multiple cores, multi-threads, graphics

S. Tiwari; “Potential, Characteristics, and Issues of 3D SOI; 3D SOI Opportunities” Short Course, 2005 International SOI Conference

The Technologies of 3D Integration

(and their challenges)

The 3D Integration Technology Spectrum



Applications

Stacked Com.DRAM

Simple Chip Stack

3D Chip Support

Integrated eDRAM

Hierarchical Cache

3D Multicore uProc

Chip Stacks

3DI Integration

Via Density (pins/cm2)	1E2	1E3	1E4	1E5	1E6
Via Size (um)	200	50	10	1	0.100
Year (approx)	1990	1995	2000	2010	2020

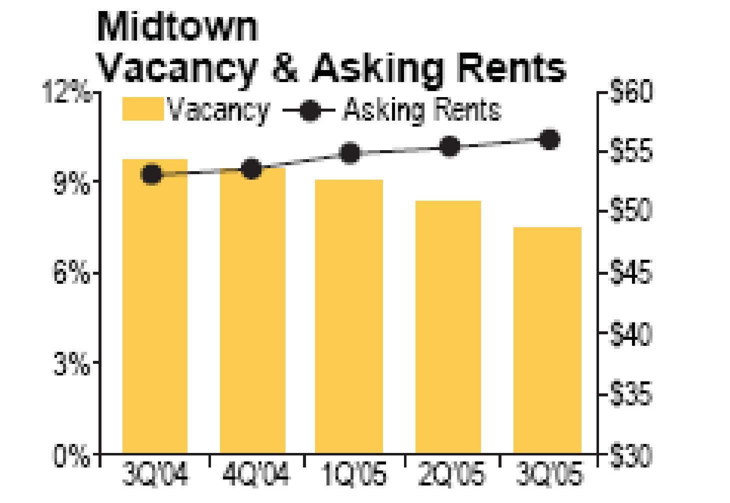
Supported Freq (Hz) 1E6 1E7 1E8 1E9

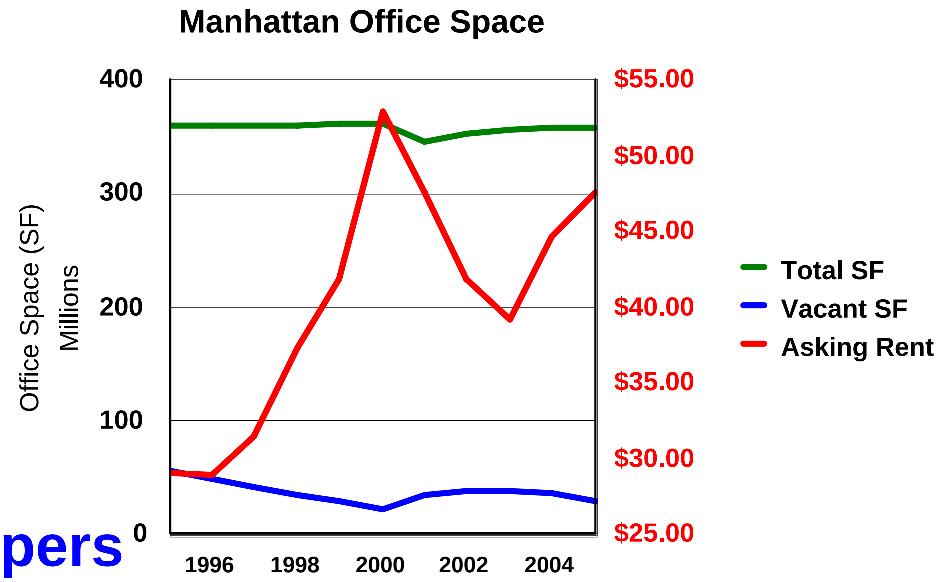
Precedent for 3D Integration:

When Real Estate Becomes Pricey

NYC Office Inventory, Rent, and Skyscrapers

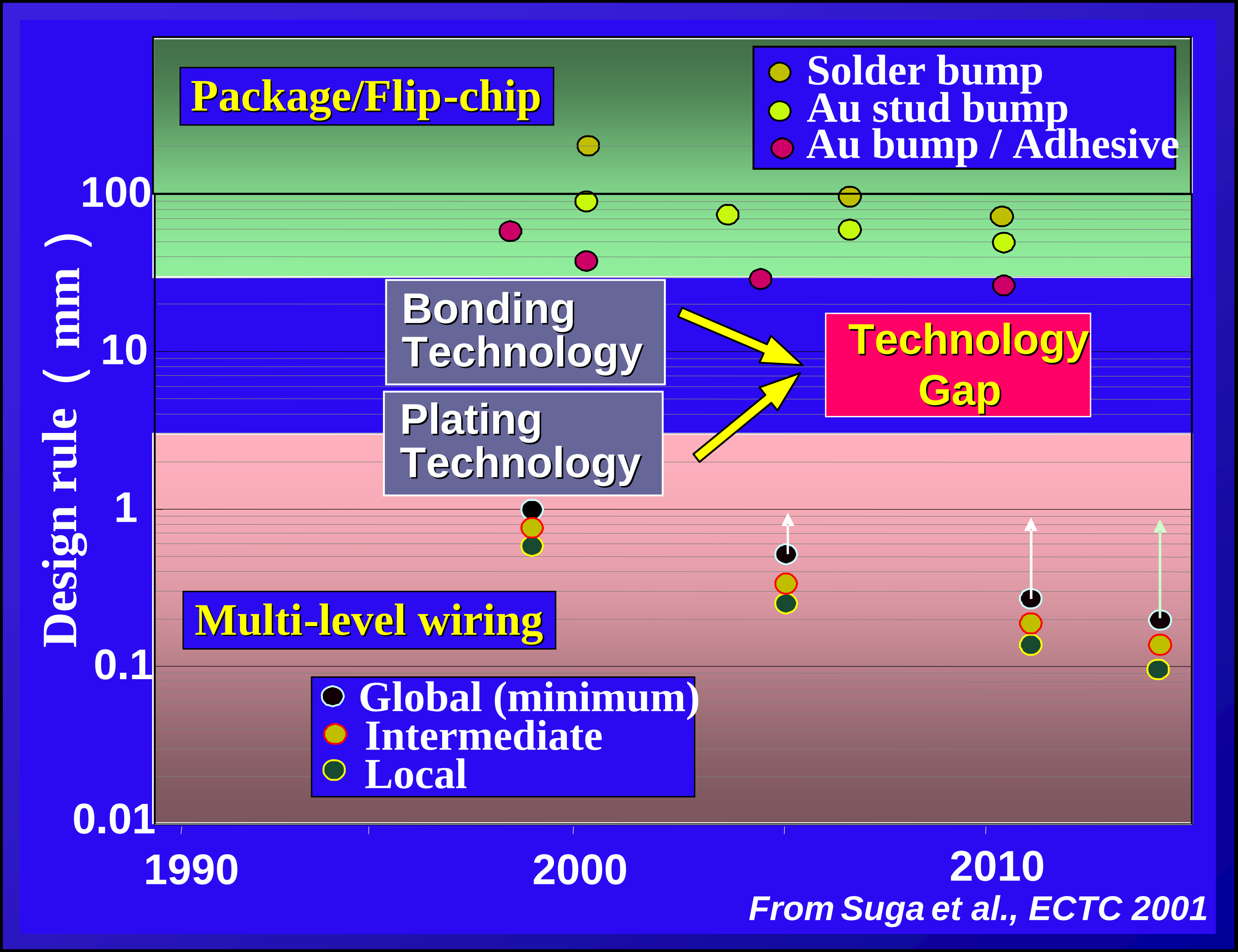
Vertical Integration isn’t new!





Data courtesy of Richard Persichetti
Grubb & Ellis, New York, NY

Chip-Package Technology Gap

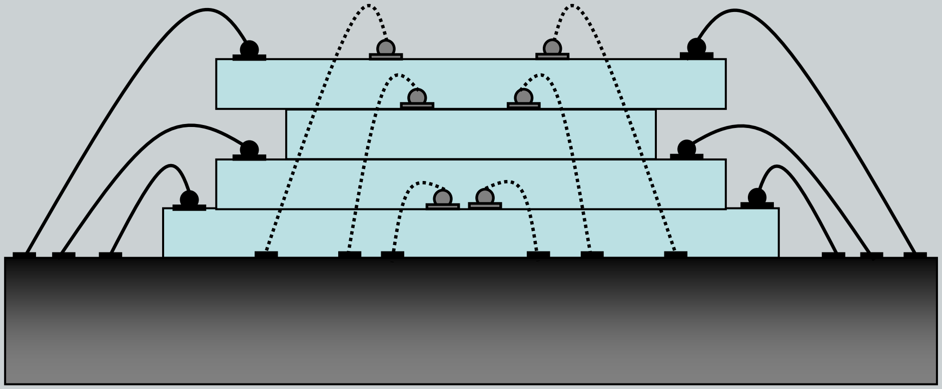


From Suga et al., ECTC 2001

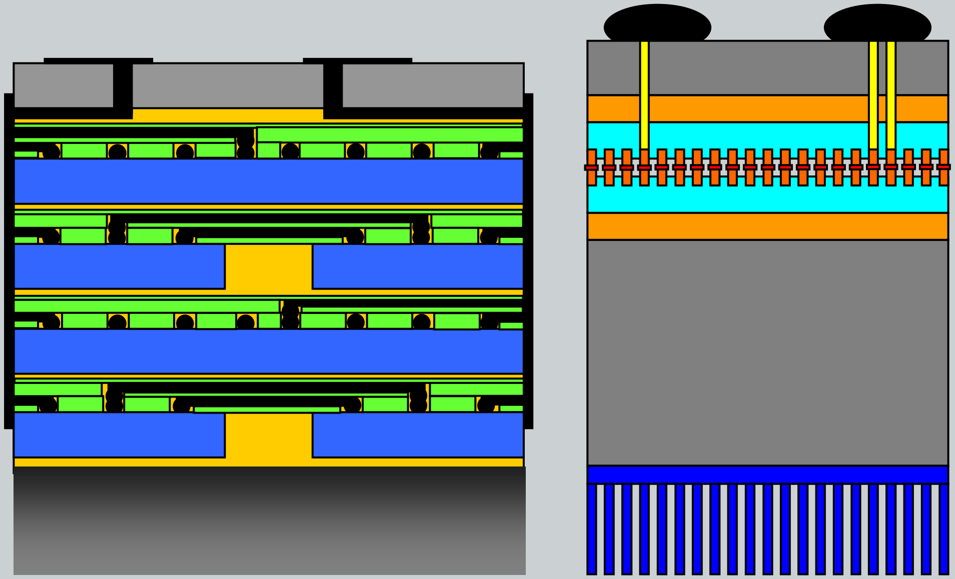
- Technology gap in the design rule between on-chip wiring and packaging interconnects

2. Present Vertical Interconnect Schemes

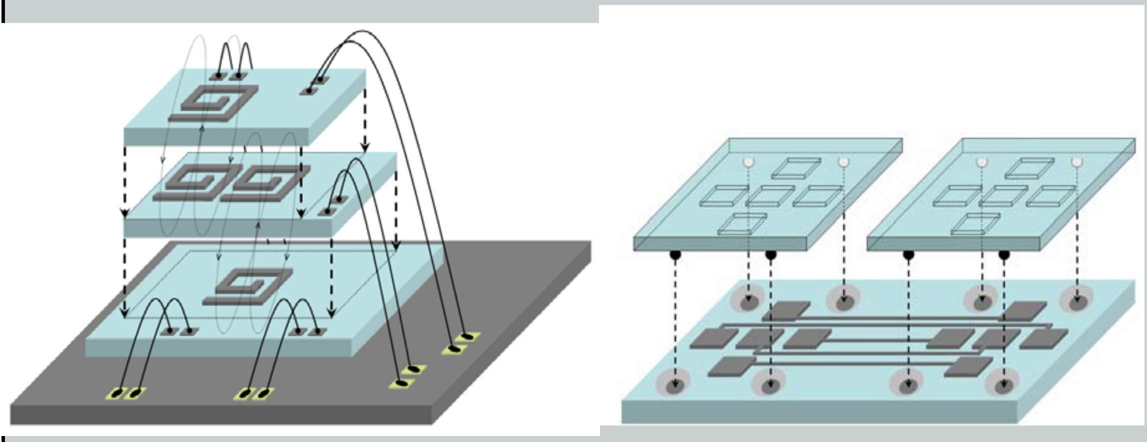
Images used by permission, W.R. Davis, North Carolina State Univ,



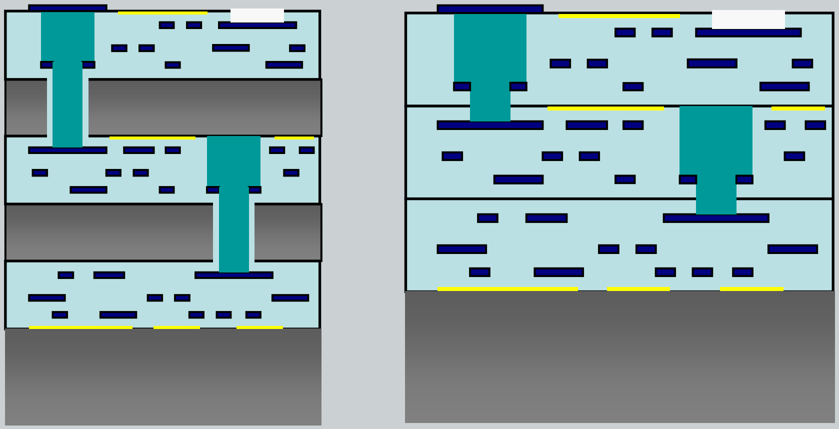
Wire Bonding



Microbump



Coupled Virtual Connections



(a) Bulk

(b) SOI

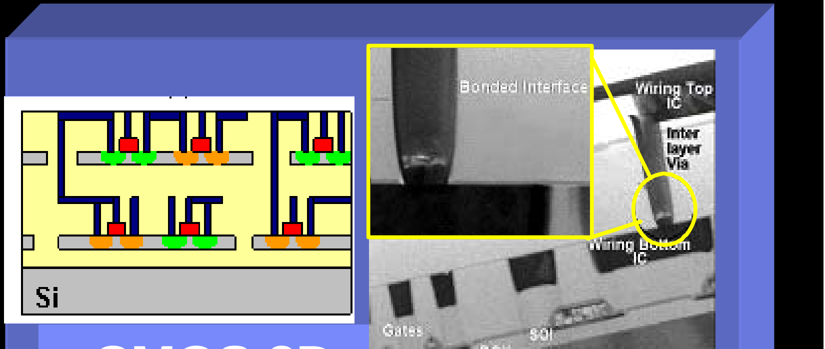
Through-Via

Evolution of 3D Integration

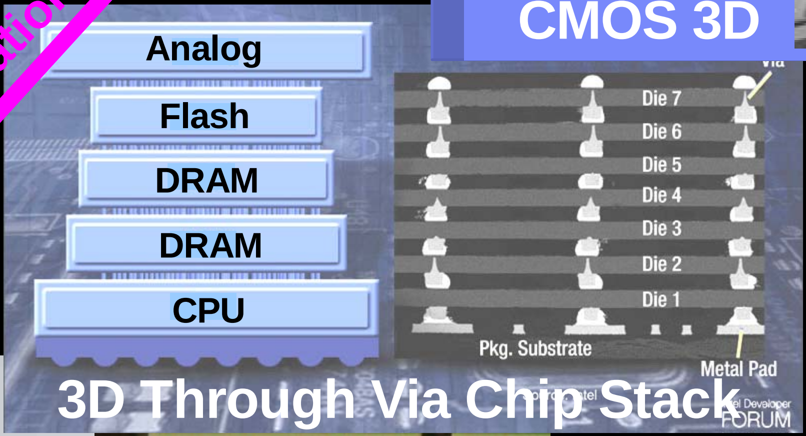
Technology Investment in the Z-Dimension

- 3D Technologies continue the sequence of interconnect advances
- Return balance to device scaling
- Enable new capabilities not available in 2D

Increasing 3D Integration

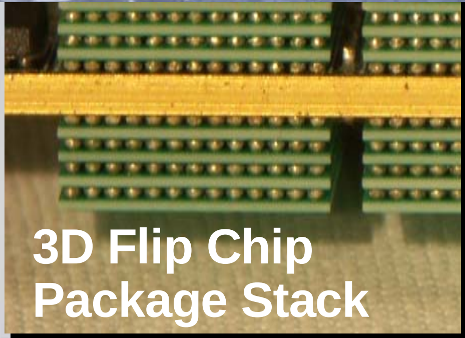


CMOS 3D

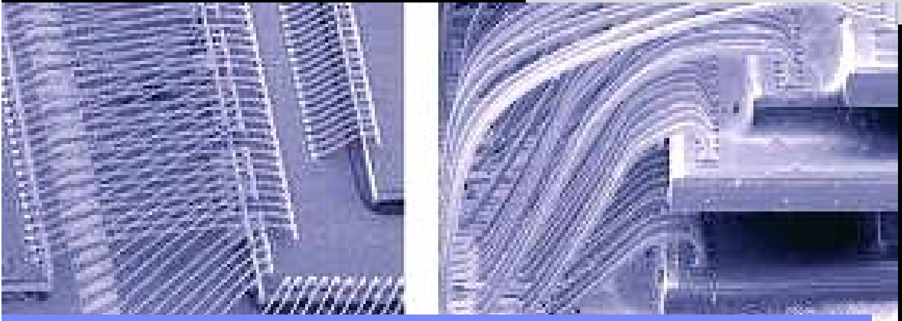


3D Through Via Chip Stack

3D Pkg Chip Stack



3D Flip Chip Package Stack



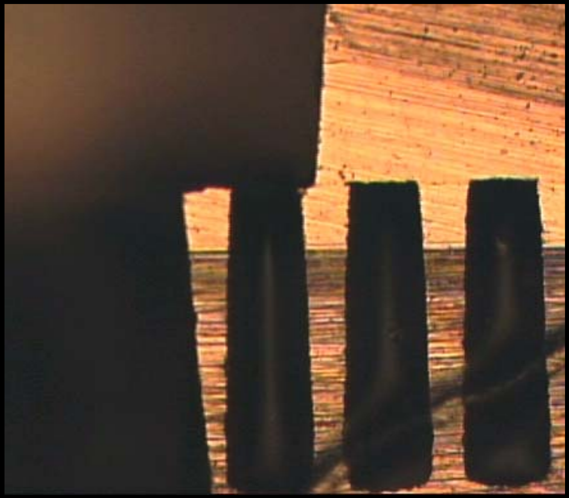
Wire bonded chip stacked 3D

- 3D Packaging R&D now pervasive in industry, academia
- Through-via technology emerging as predominant path
- 3D has *always* been large volume, but now integrating higher technologies

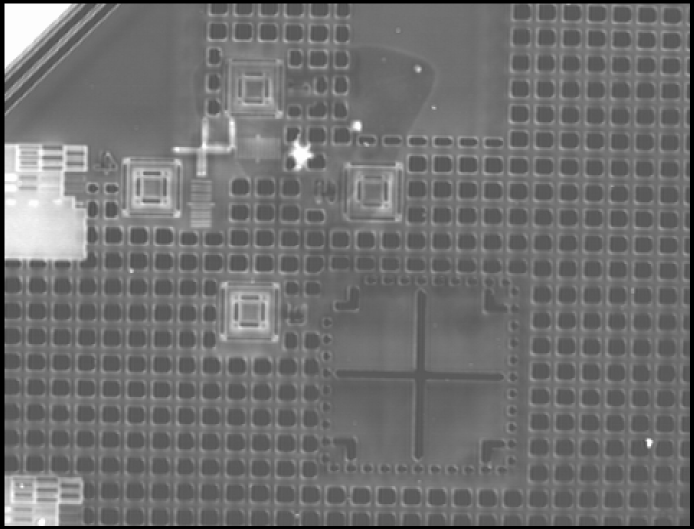
Key 3DI Processes

Images courtesy of Anna Topol,
IBM T.J. Watson Research Center

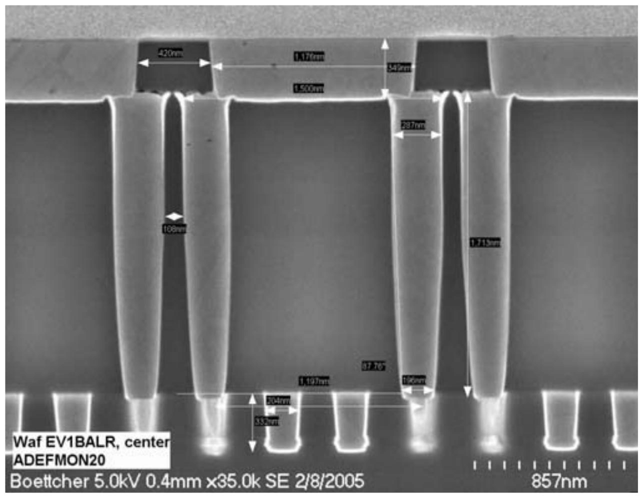
Bonding



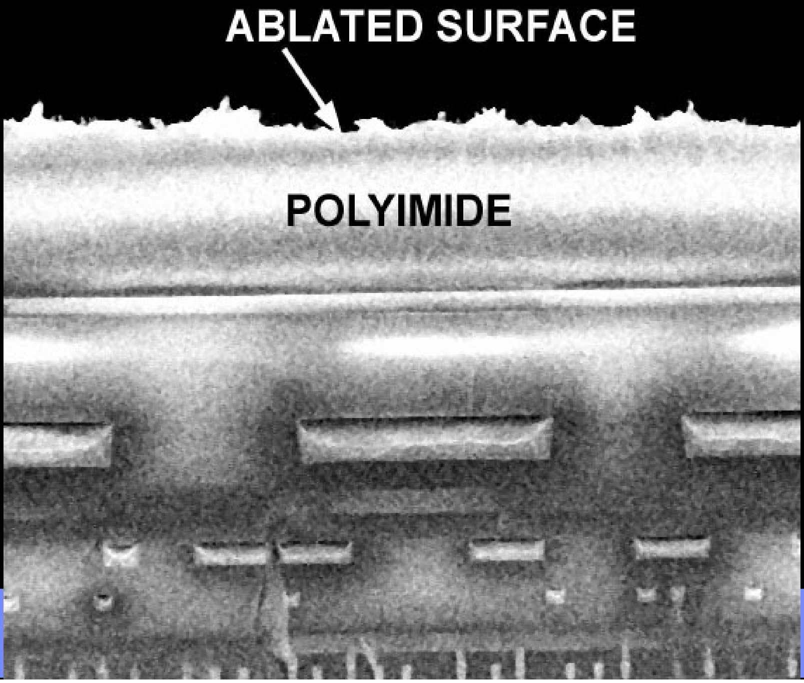
Transfer/Alignment



Electrical Contacting

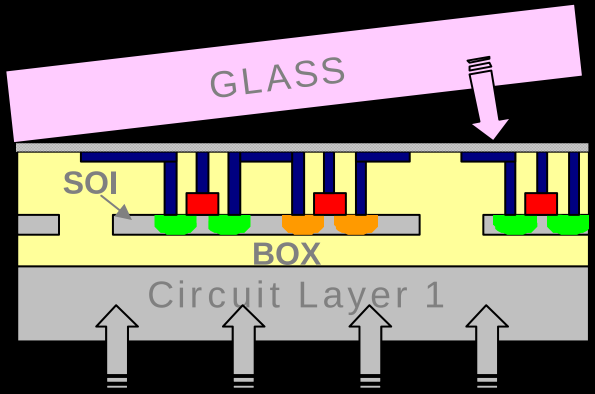


Release Process

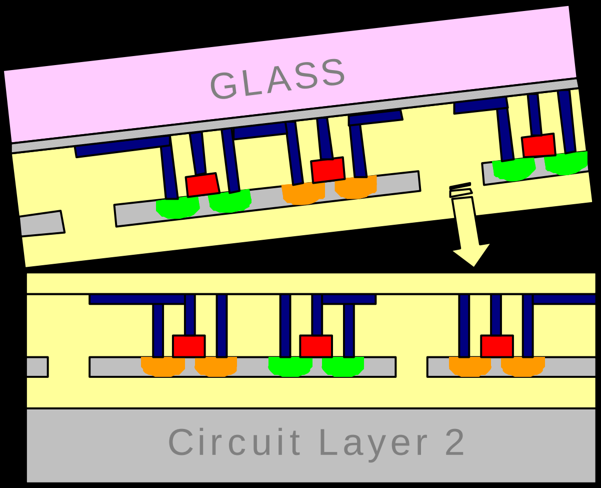


IBM 3D Process:
SOI-Based 3DI Layer Transfer

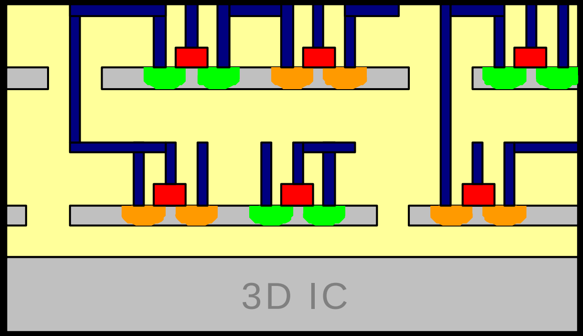
- Device layers stacked using wafer bonding
- Each layer fabricated by conventional processes
- Layers fabricated and tested simultaneously



- Attach circuit to glass handle wafer
- Remove original substrate



- Align & bond top circuit to bottom circuit

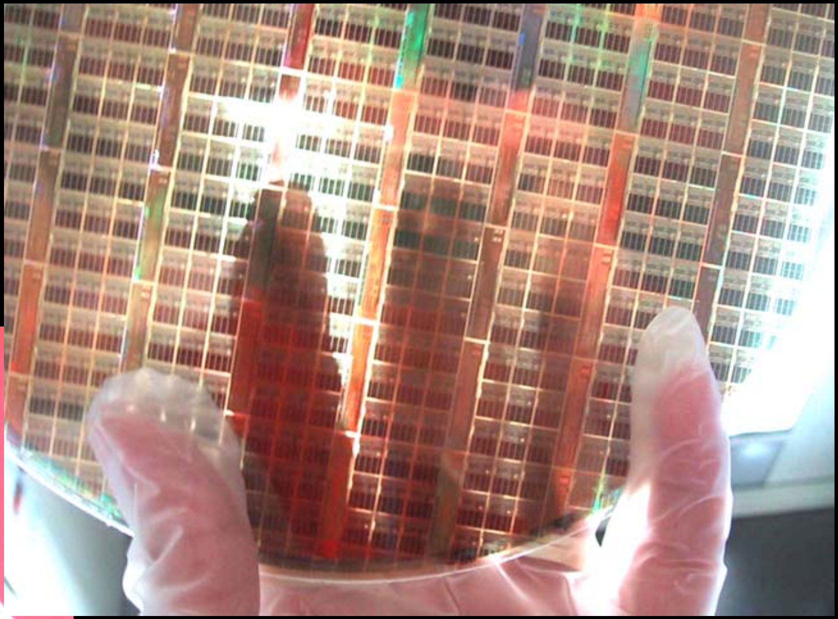


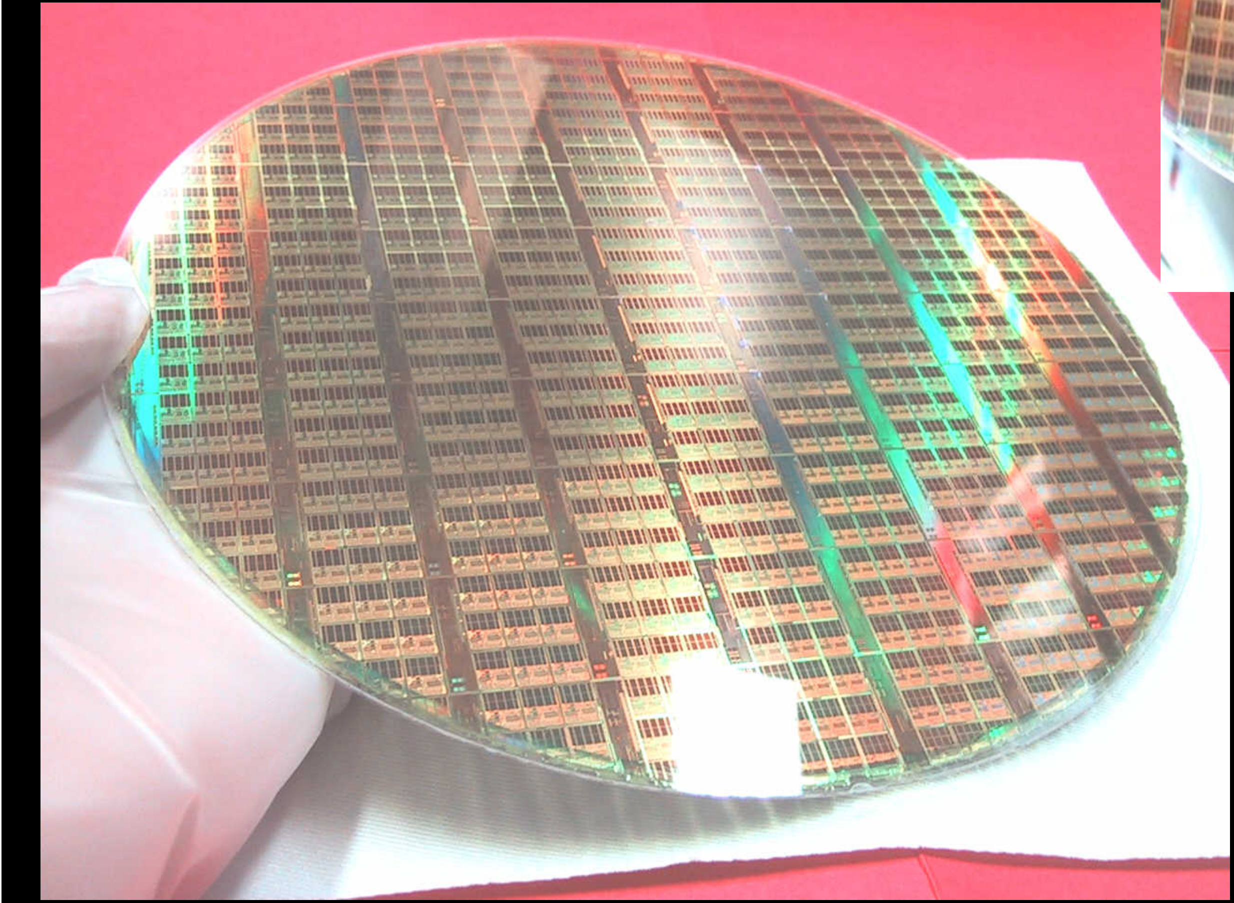
- Remove handle wafer & adhesives
- Form vertical interconnects

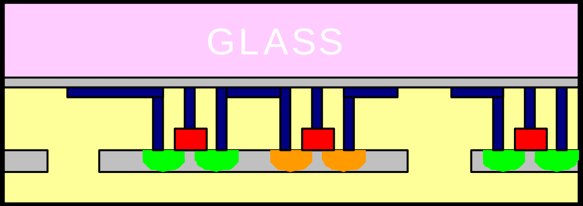
Wafer Transfer / Thinning

K. Guarini, IEDM, 2002.

Transparent Circuit
200 mm Wafer
130nm SOI Technology

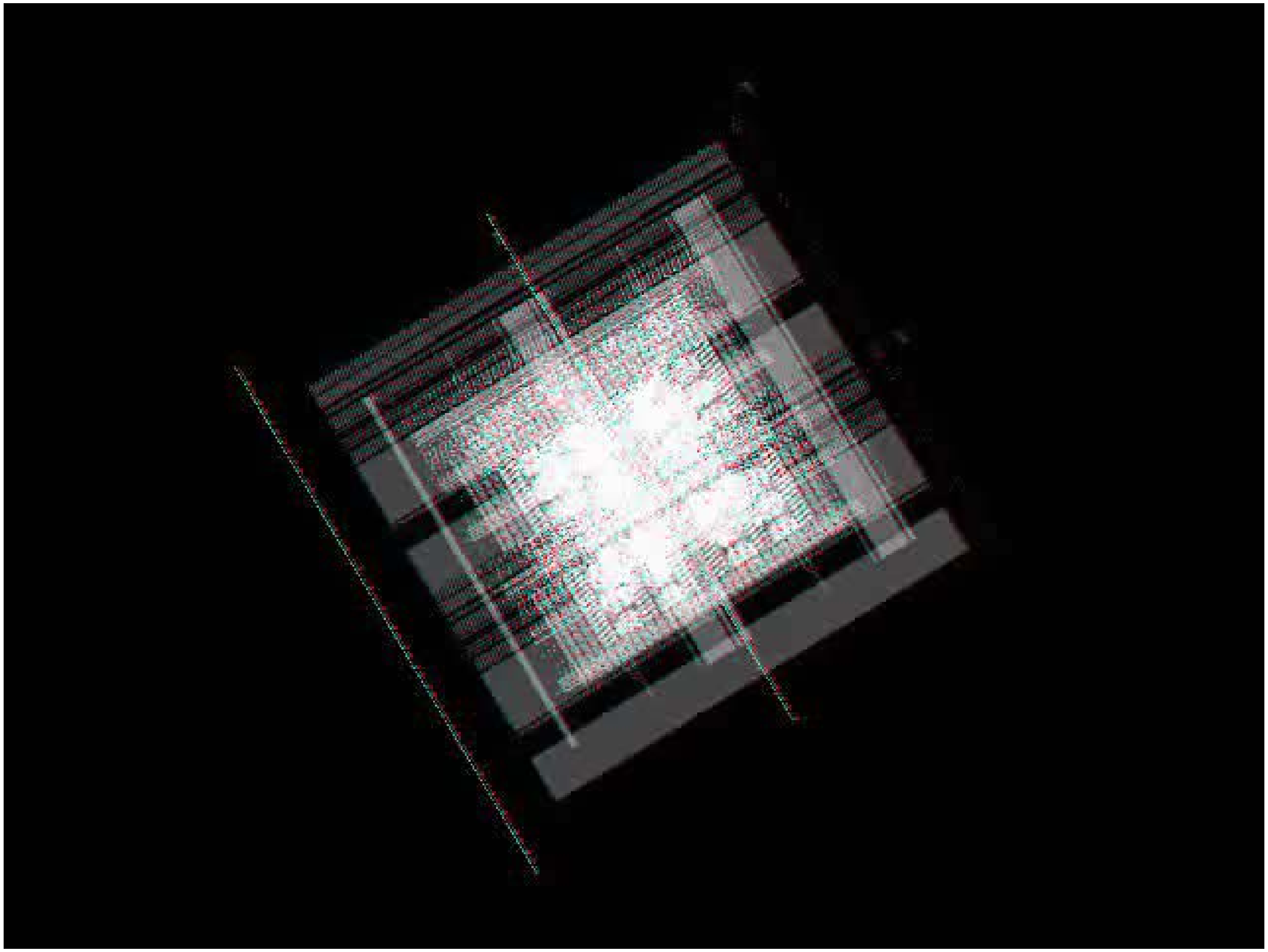






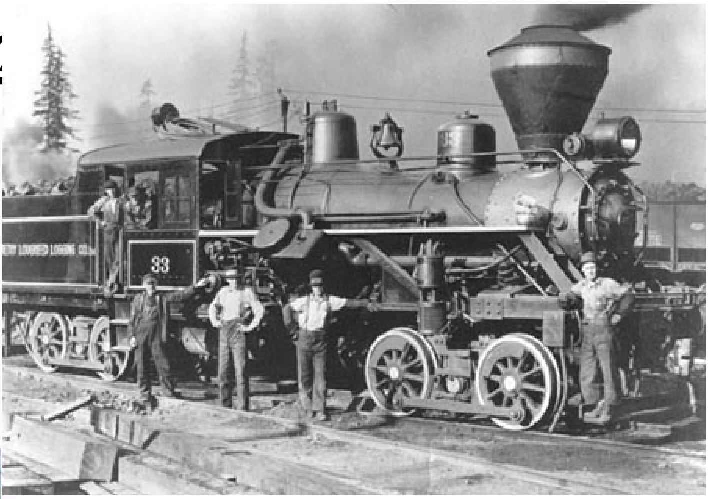
- SOI device layer + back-end metallization transferred onto glass
- Defect-free lamination over 200 mm wafers

3D Fly-Thru Movies of IBM Assembly



3D Challenges
Heat Dissipation and Natural Selection

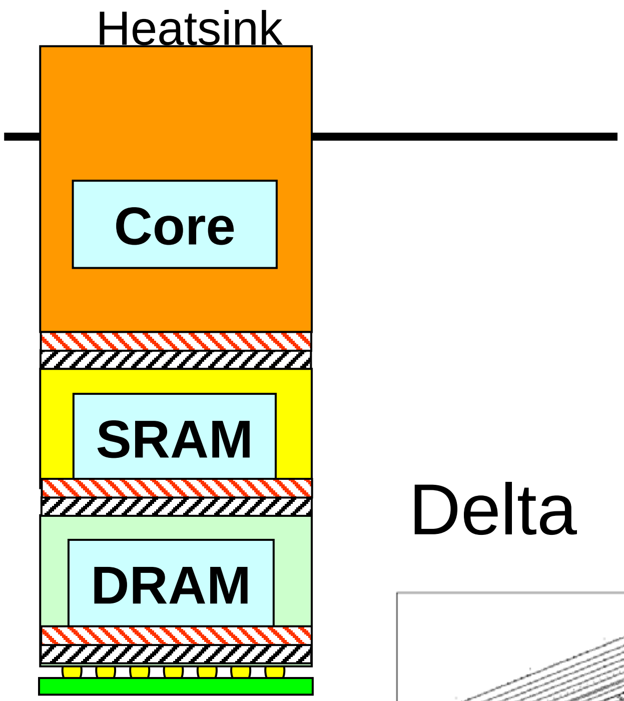
Why is area vs volume such a big deal?

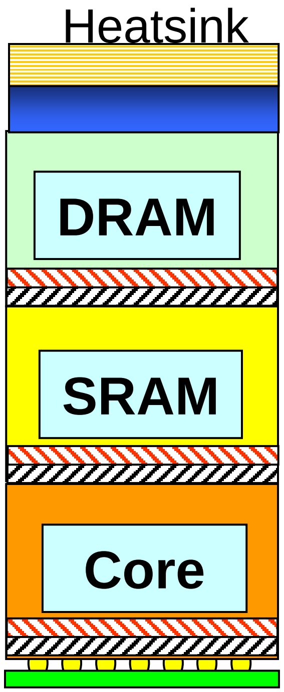


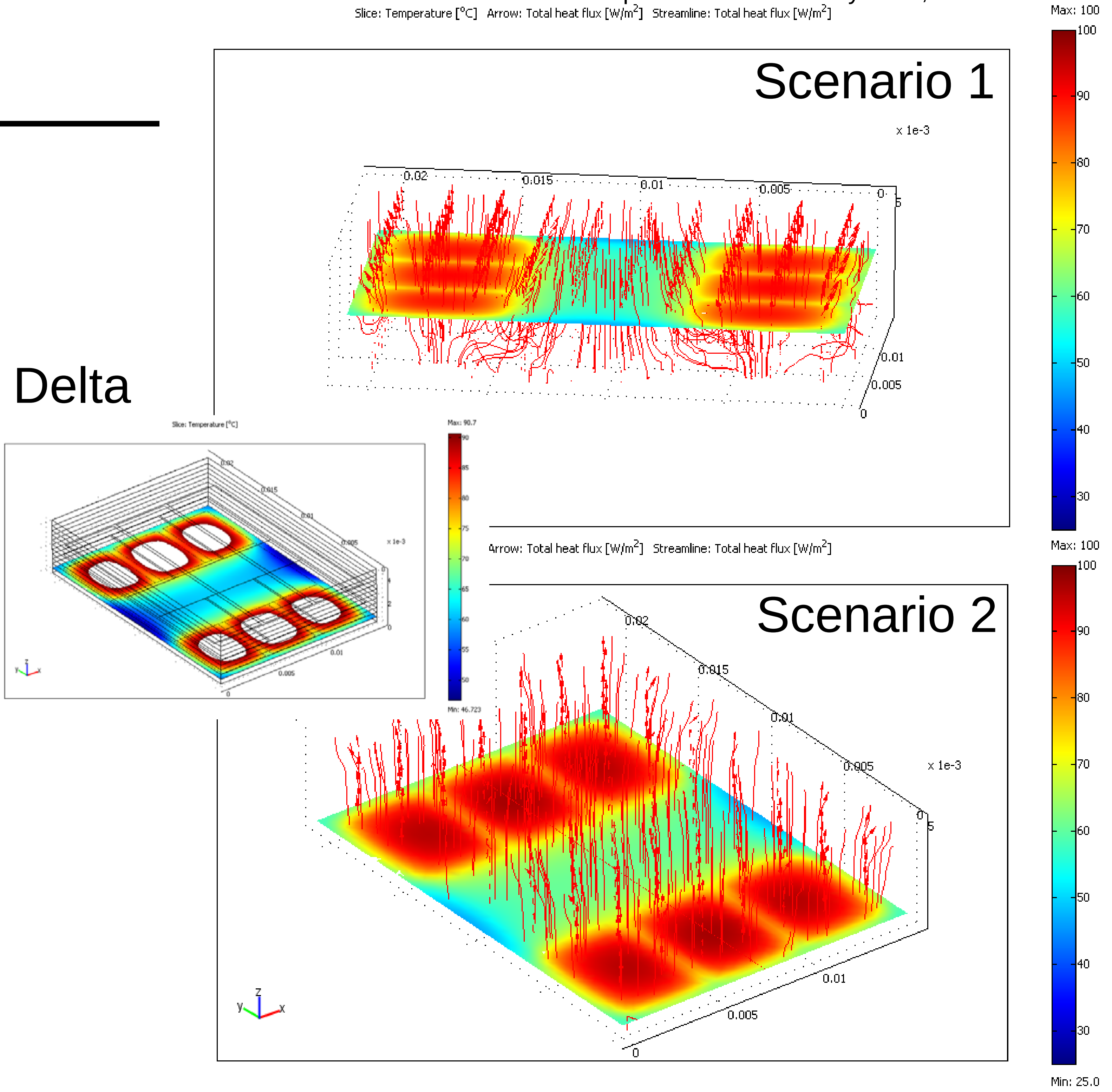
Power/Energy Issues

- **It now takes more energy to move data than to generate it, even just across chip**
  - Compute: 50pJ / FLOP / bit
  - Read: 10 pJ / operand from Reg….but
    1 nJ / operand from cache
- **Worst power nets on chip are data, instruction nets: go from mm(2D) to λm(3D)**

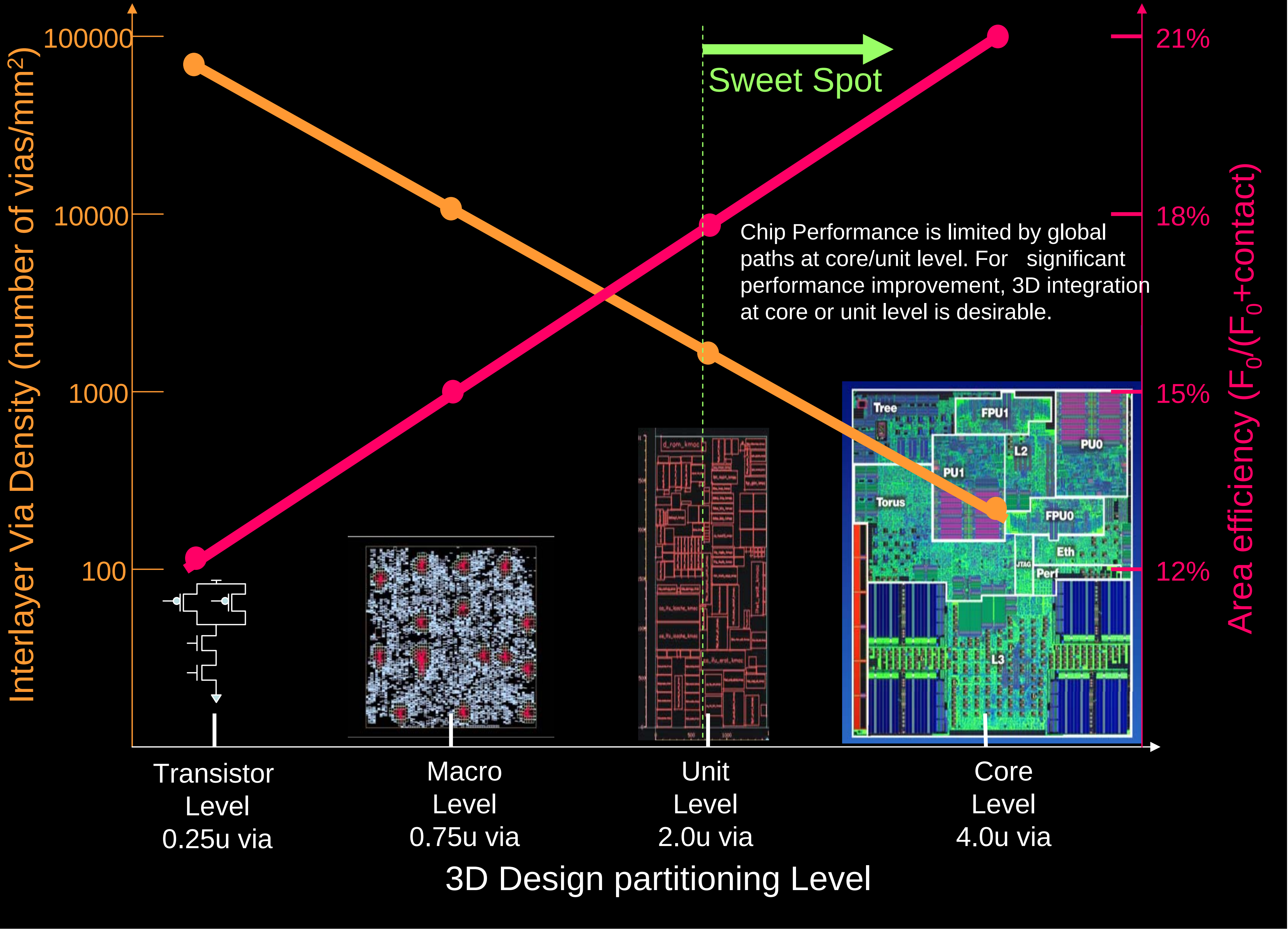
Compliments of Sri Sri-Jayantha, IBM Research







C. EDA and 3D Integration Trends



Chip Performance is limited by global paths at core/unit level. For significant performance improvement, 3D integration at core or unit level is desirable.

7. Summary

- λP architecture tricks to avoid atomistic, QM scaling boundaries overwhelm present interconnects
- Integration into Z-plane again postpones interconnect-related limitations to extending classic scaling.
- Transaction retirement rate dependence on data delivery is *increasing*: dependence on λP performance and CMOS device speed is *decreasing*
- 3D Integration improves storage density & access to that storage
- 3D Integration will enable previously unattainable capabilities characterized by realtime access to massive amounts of storage.

Human Scaling

Tomorrow's microprocessors will be improved with capabilities developed using today's machines

Tomorrow's engineers will design microprocessors with insights they learn from today’s engineers and professors.

Engineers/professors today insure a bright tomorrow by transferring **ideas** as well as **technologies** to the next generation.