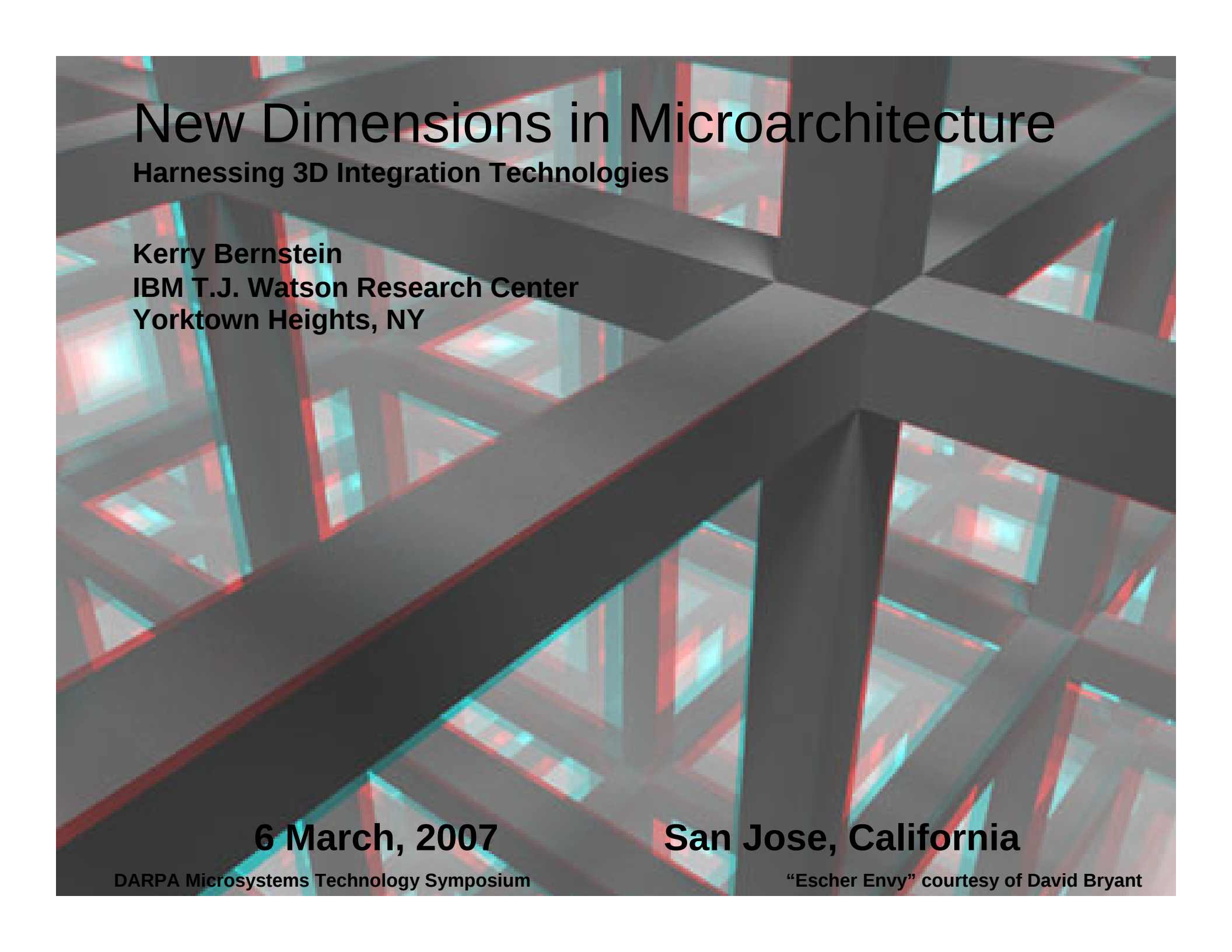




New Dimensions in Microarchitecture

Harnessing 3D Integration Technologies

Kerry Bernstein

IBM T.J. Watson Research Center

Yorktown Heights, NY

6 March, 2007

San Jose, California

DARPA Microsystems Technology Symposium

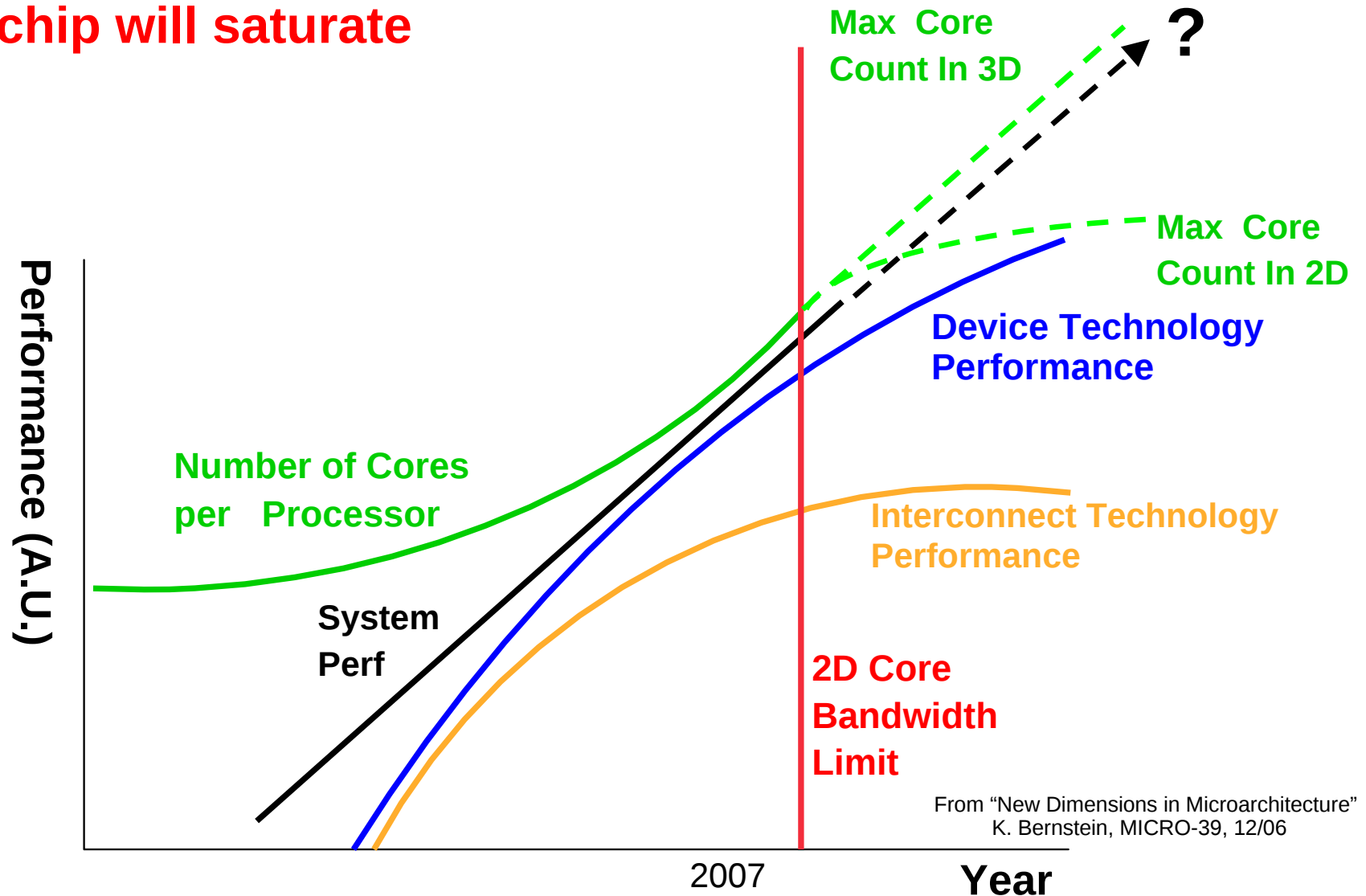
“Escher Envy” courtesy of David Bryant

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE 06 MAR 2007		2. REPORT TYPE N/A		3. DATES COVERED -	
4. TITLE AND SUBTITLE New Dimensions in Microarchitecture Harnessing 3D Integration Technologies				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) IBM Corporation				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release, distribution unlimited					
13. SUPPLEMENTARY NOTES DARPA Microsystems Technology Symposium held in San Jose, California on March 5-7, 2007. Presentations, The original document contains color images.					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 12	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

Server Trends

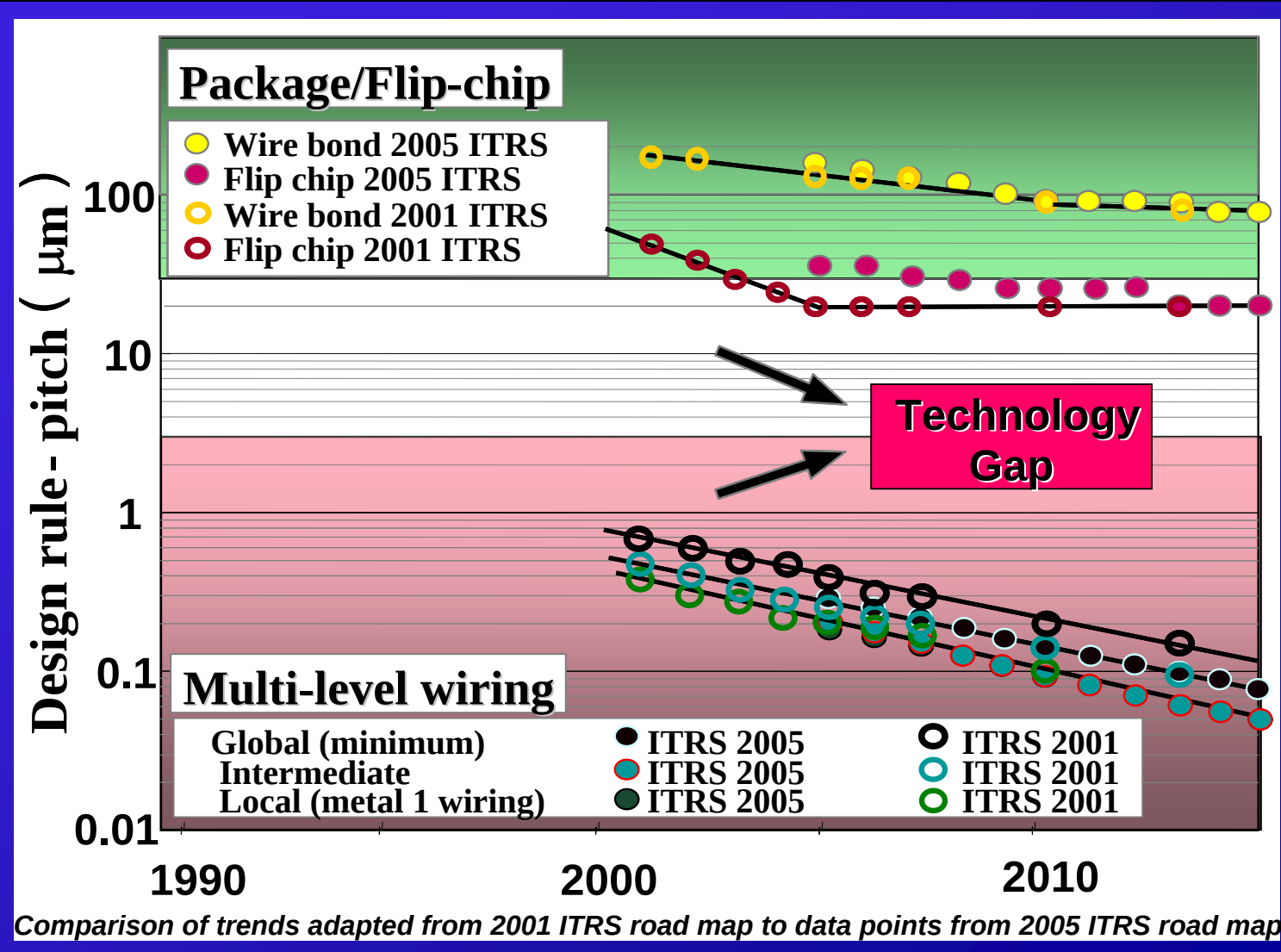
- Frequency no longer increasing
 - Logic speed scaled faster than memory bus
 - (Processor clocks / Bus clock) consumes bandwidth
- More speculation; attempts to prefetch
 - Wrong guesses increase miss traffic
- Shortening linesize limited by directory as cache size grows
 - But doubling linesize doubles 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
- Processors / SMP increasing
 - Aggravates queueing throughout the system

Without more bandwidth at low latencies, core counts on chip will saturate



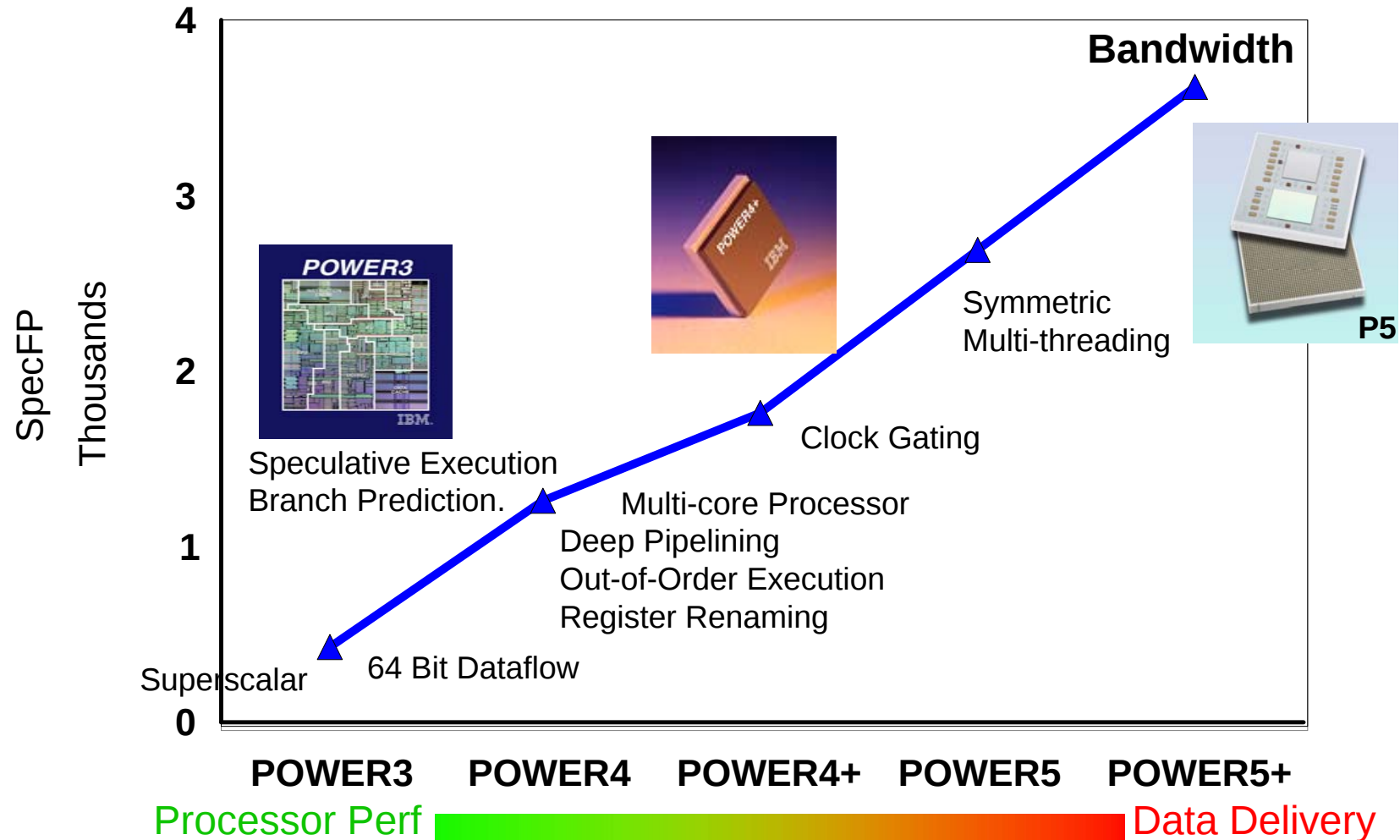
3D extends transfer of performance from the device to the core level

Chip-Package Technology Gap



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

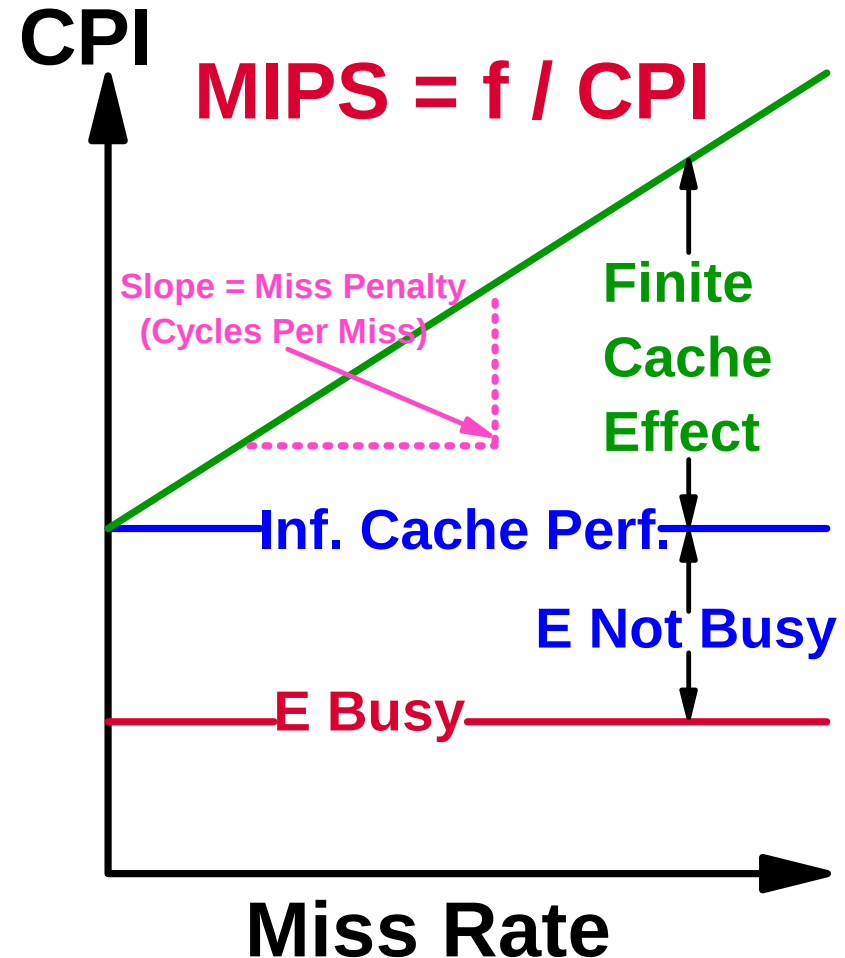
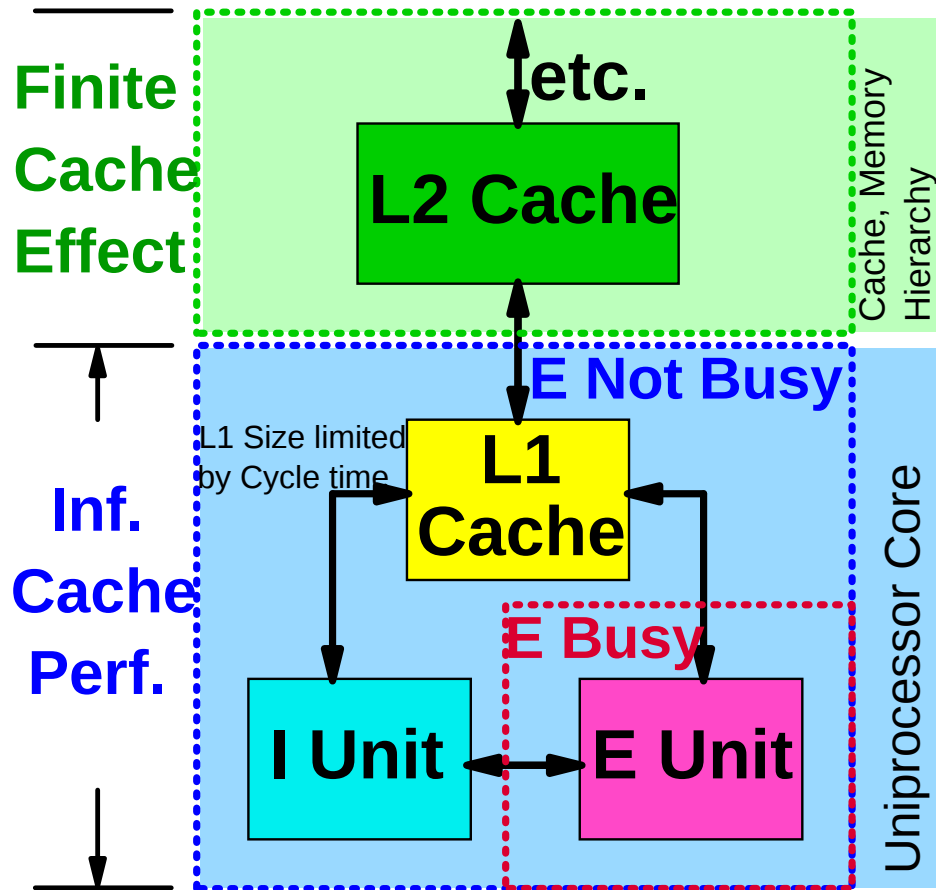
POWER Series Architectural Perf Contributions



Transaction Rate Dependence

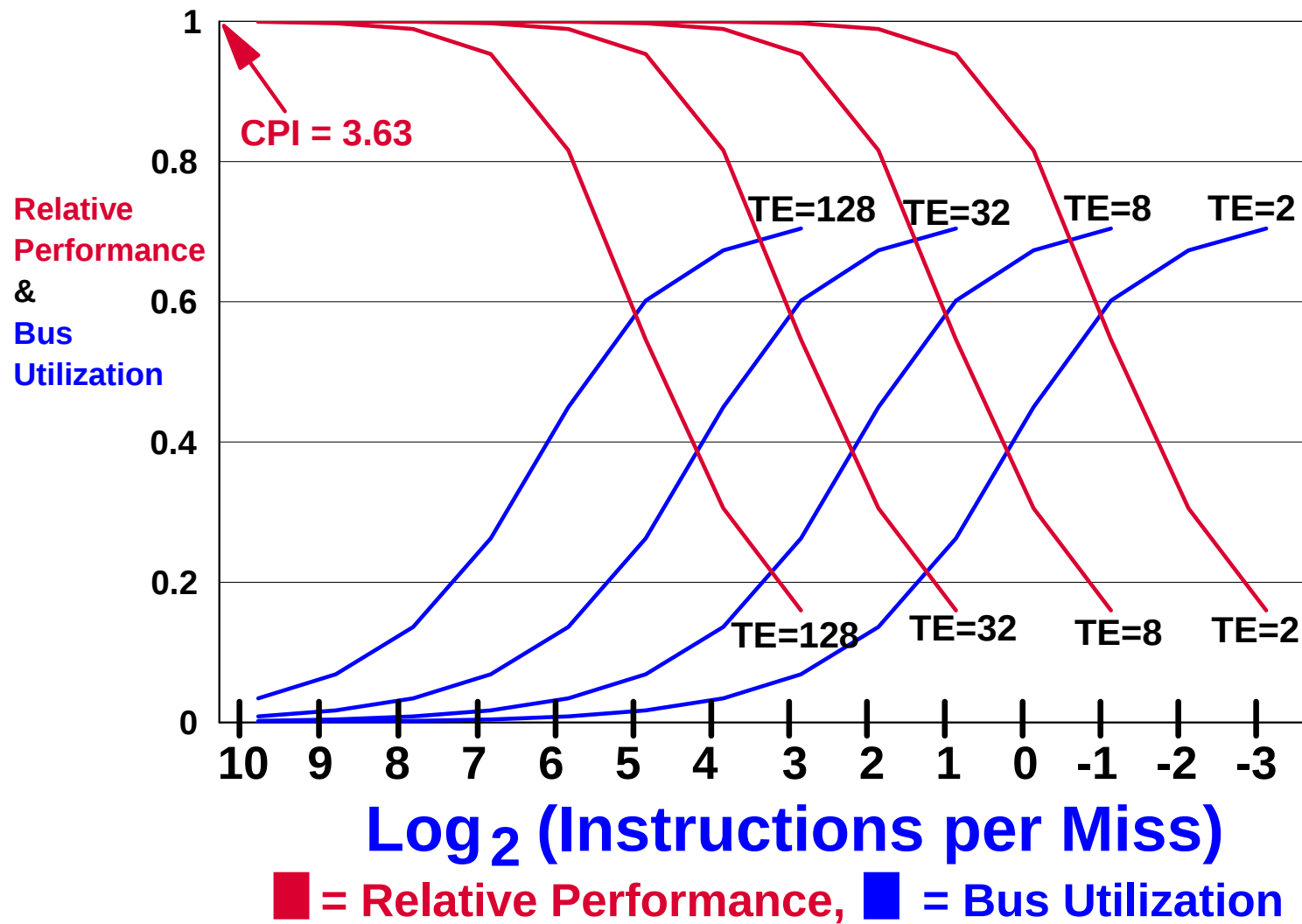
Components of Processor Performance

From ISCA '06
Keynote address by
Phil Emma, IBM



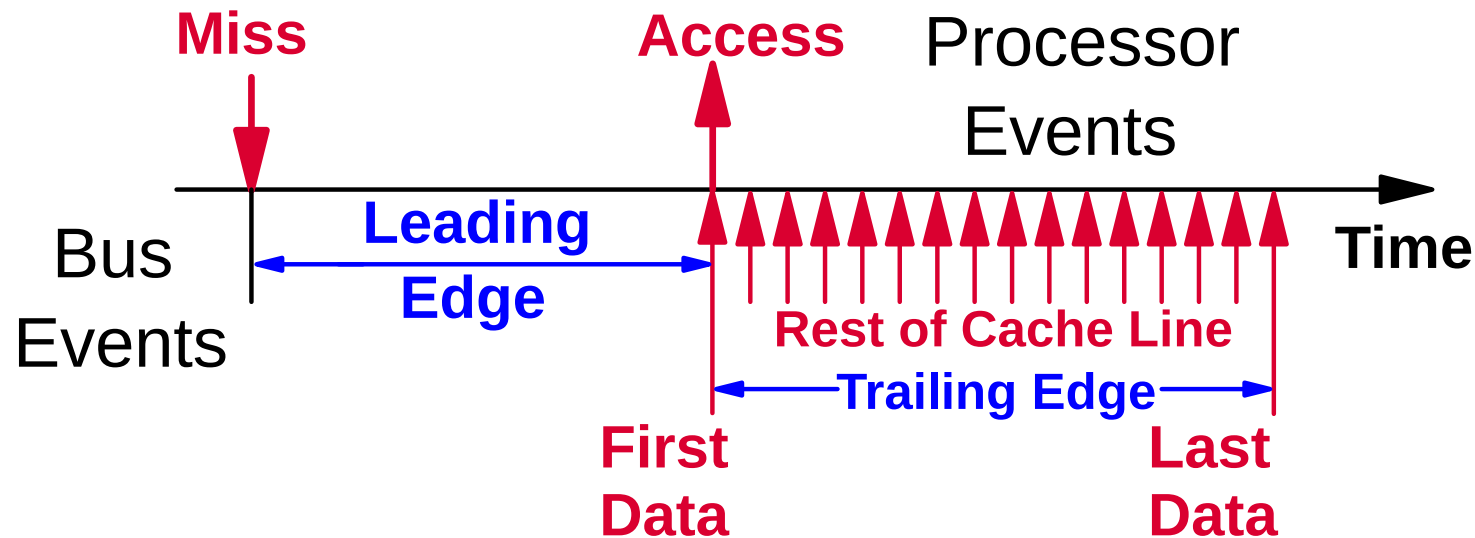
Delay is sequentially determined by a) ideal processor,
b) access to local cache, and c) refill of cache

Queueing Effects vs. Log Miss Rate



What Is Bandwidth Used For?

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

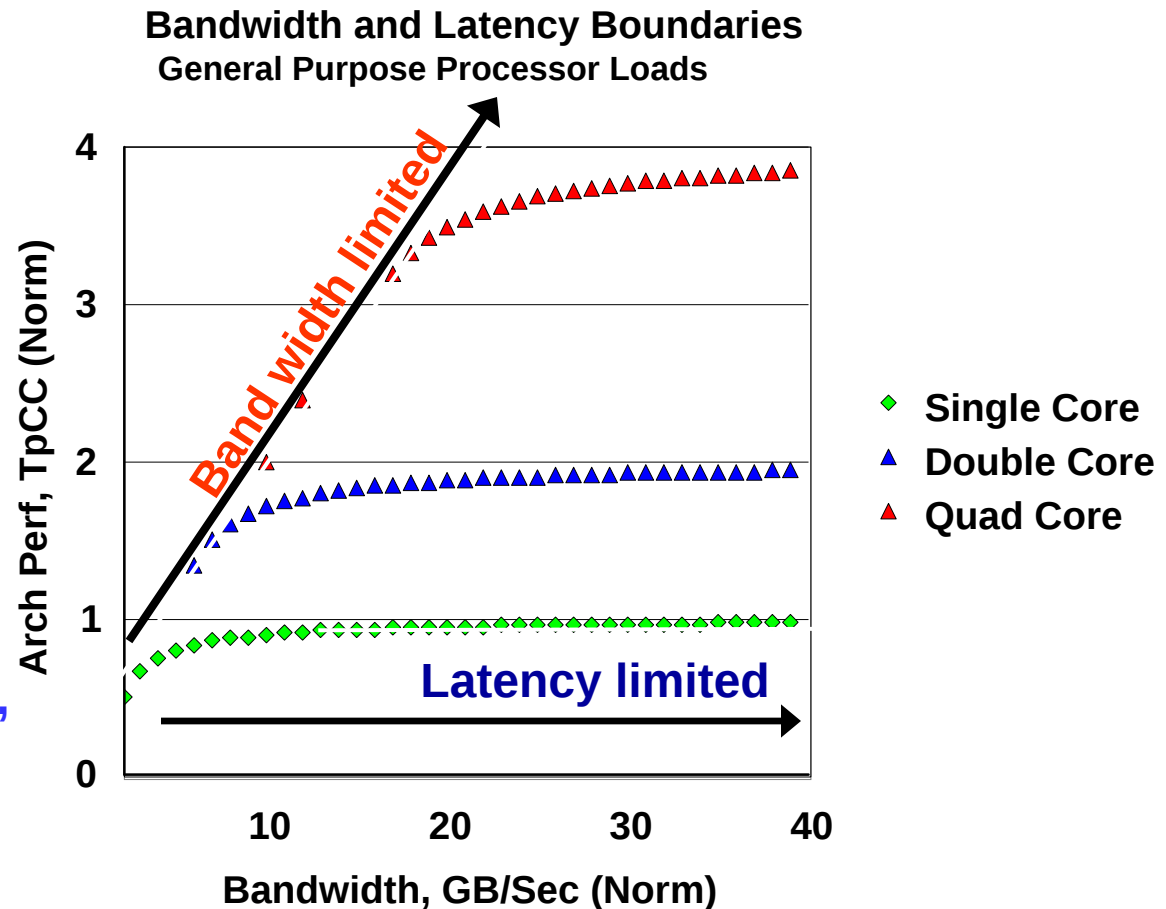


Miss Penalty = Leading Edge + Effects(Trailing Edge)

3D - Bandwidth and Latency

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

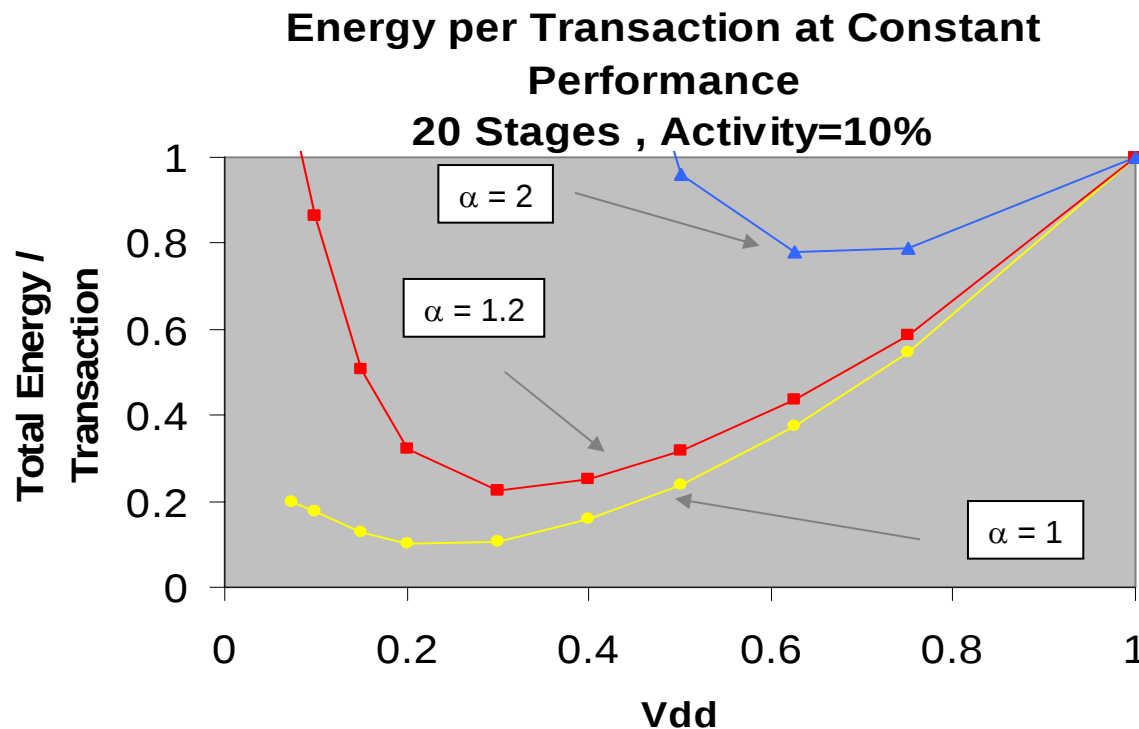
- For generic workloads, uni-processor perf saturates bandwidth benefit, becomes latency-limited.
- As core counts increase, I/O Bandwidth becomes increasingly important



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

Low Vdd Technology and Parallelism

- Energy optimum for fixed performance as function of V_{dd} , V_T and effectiveness of parallelism
 - α determines the device (circuit) overhead to maintain constant performance through parallelism
 - $\alpha = 1$ **no overhead**: half the speed double the devices
 - $\alpha > 1$ **increasing overhead**: passive power becomes dominant



$$P \sim P_0 \frac{d_0}{d} \left(\frac{N}{N_0} \right)^{\frac{1}{\alpha}}$$

N =number of ckts,

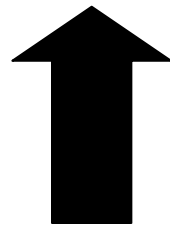
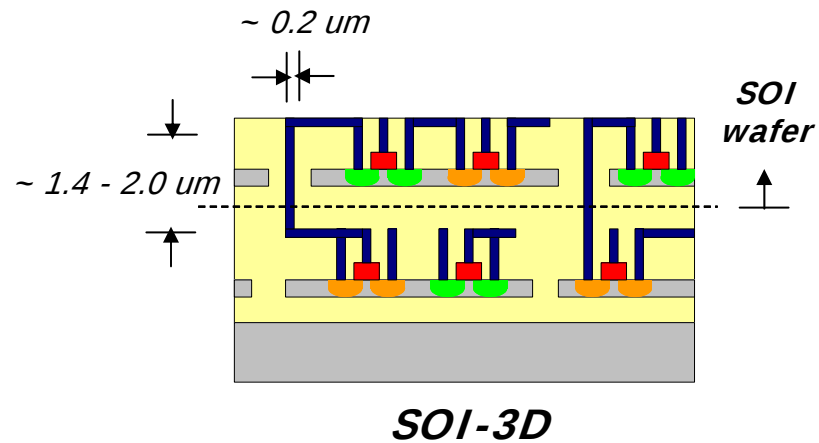
d =ckt delay

N_0 = number of ckts at 1V

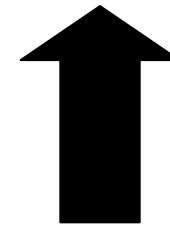
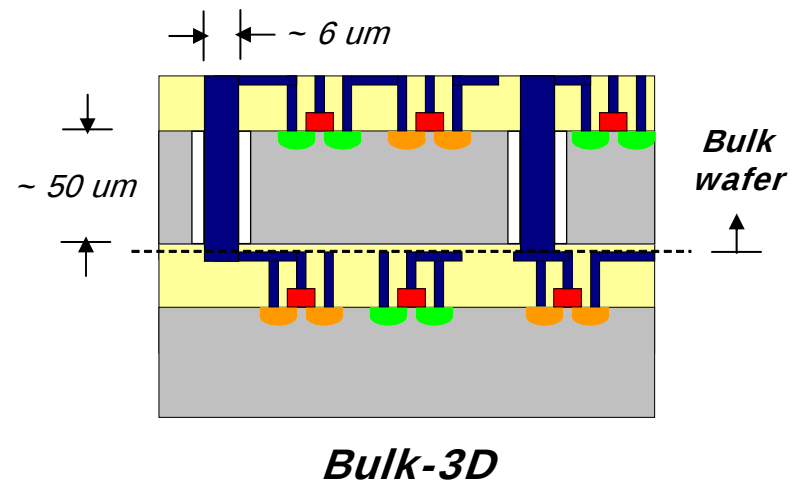
d_0 = ckt delay at 1V

From "3D Intergration" Special Topic SessionI
 W. Haensch, ISSCC '07, 2/07

Two Classes of 3DI Processes at IBM



SOI top layer
**Advantage: Smallest 3D
vias**



Bulk top layer
**Advantage: Broader foundry
compatibility**

Summary

- λ P architecture tricks to avoid atomistic, QM scaling boundaries overwhelm present interconnect technologies
- Integration into Z-plane again postpones interconnect-related limitations to extending classic scaling.
- No aspect of architecture or technology remains 2D, so why even view chips as being monolithic anymore?
- Transaction retirement rate dependence on data delivery is *increasing*: dependence on λ P performance and CMOS device speed is *decreasing*
- 3D Integration improves storage density and access to that storage
- The last remaining opportunity in CMOS to save power is in delivery of data rather than in its generation.