“Breaking the Memory Wall for Multi-core Processors with 3D Hyper-integration and Hetero-integration”

SRC/NSF/ITRS
Forum on Emerging nano-CMOS Architectures -
Virtual Immersion Architectures: VIA 2020,
July 10-11, 2008

John F. McDonald
Center for Integrated Electronics
Rensselaer Polytechnic Institute
Troy, NY 12180
mcdonald@unix.cie.rpi.edu

Rensselaer

A New Motivation for Virtual Immersion

In search of a policy to stop oil addiction

By ROD DREHER

You've heard it said the world is flat — that today, all commerce is global. It's time to rethink that in light of the global energy crisis. The world is being re-rounded, its horizons shrinking. Localism is the new globalism.

Cheap, abundant and accessible fossil fuels allowed us to create a world in which we are relatively unconstrained by geography. That era is passing into history, and it is not likely this process can be reversed.

The reason: simply not enough oil being extracted quickly or inexpensively enough to meet global demand — nor, in all likelihood, will there be again. This is called peak oil.

Economic analysts say Americans have never before spent a greater part of their income on energy costs. The sooner we come to terms with this reality, the sooner we can begin taking serious steps to adapt.

By this fall, chances are John McCain and Barack Obama will be talking more about energy than any other issue. They'll have to. That would be a final change from now. Peak oil is a far more urgent crisis than

Read & React

▶ "The American way of life is not up for negotiation," so says Vice President Dick Cheney, at least, in regards to our reliance on oil. Is it time to reconsider such thinking? Tell us what you think. In 250 words or less, by noon Tuesday. Here are three ways to do so:

■ Go to http://blogs.timesunion.com/readandreact/

■ Email comments to: luxinmentory@timesunion.com. Include your name, community and a daytime phone number where you can be contacted for confirmation.

■ Send a letter to us at: Letters to the Editor, Times Union, Box 15000, Albany, NY 12212. Or deliver it to our main office at 645 Albany Shaker Road in Colonie between 8 a.m. and 5 p.m. Include your name, community and a daytime phone number where you can be contacted for confirmation.

and social effects are numerous on the candidates' agendas. Every petroleum-dependent aspect of our economy, including the far-flung distribution systems for consumer goods and the daily commute, will be difficult to sustain. The only question is how

social activity will, of necessity, be local.

A comprehensive domestic energy policy should be geared toward helping regions, cities and neighborhoods depend as little as possible on petroleum. That could mean:

■ Dramatically changing zoning restrictions to permit small retailing in residential areas, making it possible for people to walk or bike to do their shopping. Refuse to approve new housing developments unless they are designed for pedestrian accessibility to retail areas.

■ Through regulation and tax code changes, encouraging the development of local farming, so population centers can better afford to feed themselves. Similarly, discouraging the use of arable land for development.

■ Government investing in expanding broadband infrastructure in order to make high-speed Internet access more accessible and affordable. A recent study by the Information Technology and Innovation Foundation found ranked the United States 15th out of 30 industrialized countries in terms of broadband performance. Offering tax incentives to companies that use the Internet to decentralize their workforce to homes and neighborhood clusters.

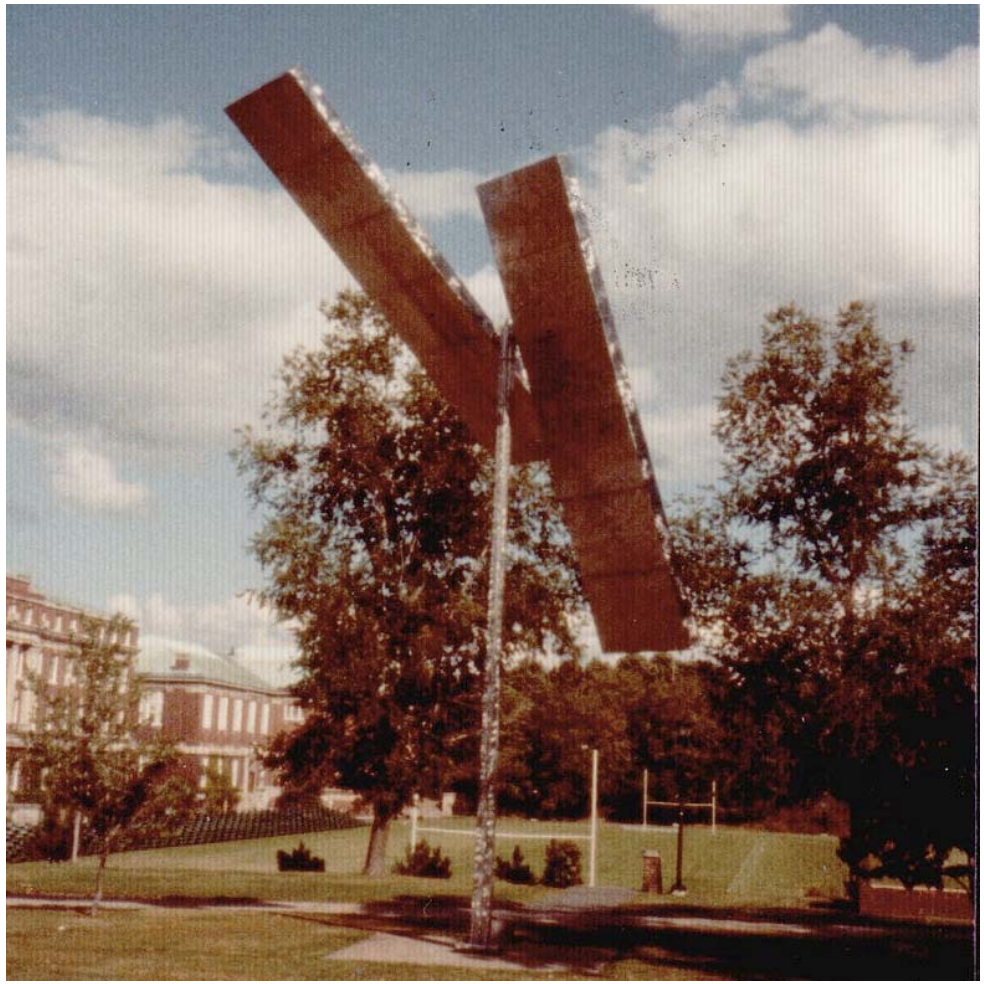
Beyond localism, a far-thinking federal energy policy would consider expanding the national rail system as an increasingly cost-effective alternative to air travel. More broadly, federal and state governments also could accelerate energy smart consumer

Excerpt from the Albany Times Union Newspaper -

"Offer tax incentives for companies to use the internet to decentralize the workforce into VIRTUAL office clusters closer to home…to hold virtual meetings with minimal travel…."

Curtis Priem(§)1982 RPI Undergraduate Project -

Model the Aerodynamics of the RPI Wind Driven Chrinitoid and Display Motion Graphically (George Rickey - RPI Art Professor, ca. 1974)



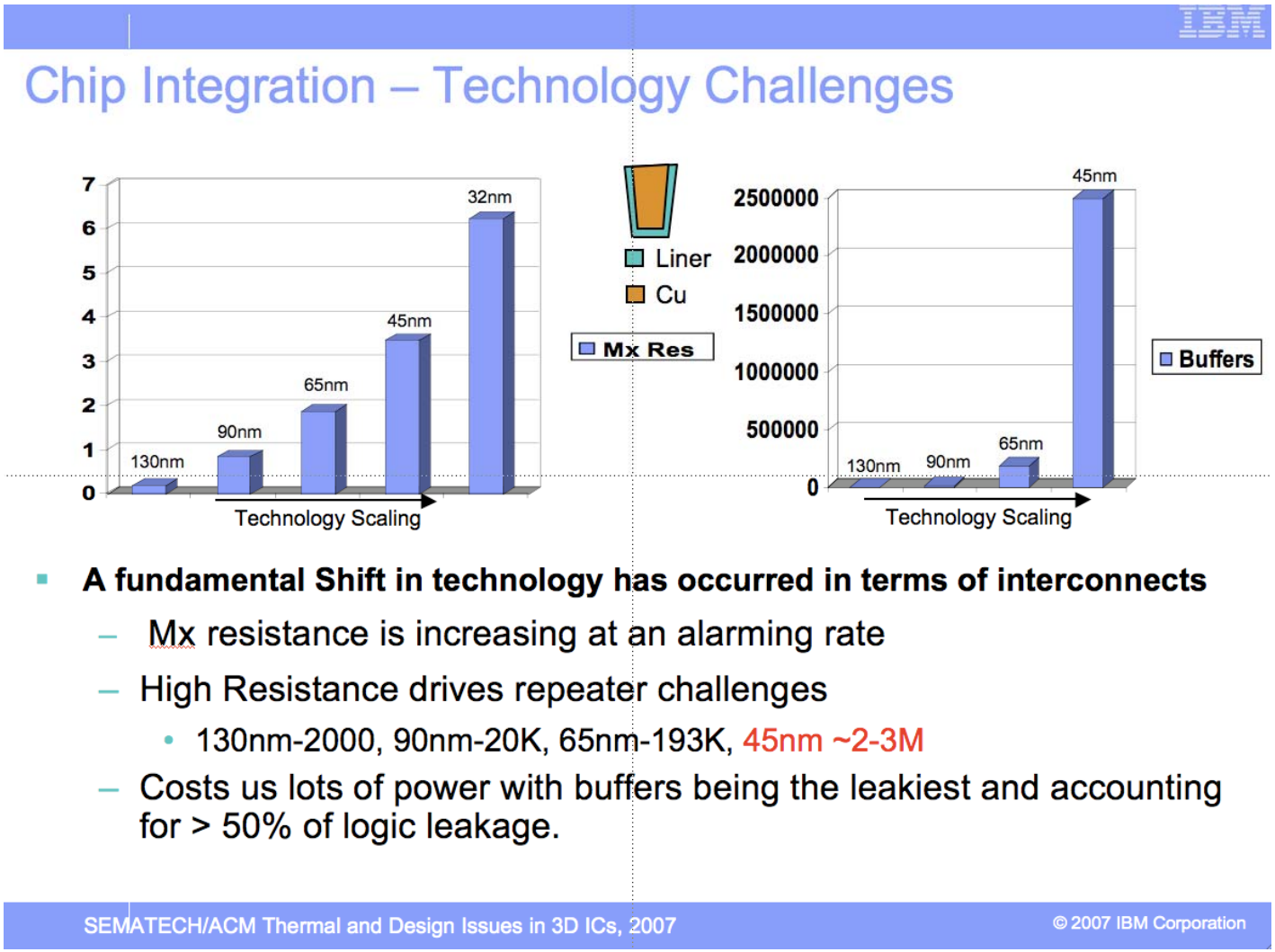
Illustrated two Parts of Virtual Reality: Computation And Display Rendering

§ Co-founded NVIDIA with Jen-Hsun Huang and Chris Malachowsky and was its Chief Technical Officer from 1993 to 2003.

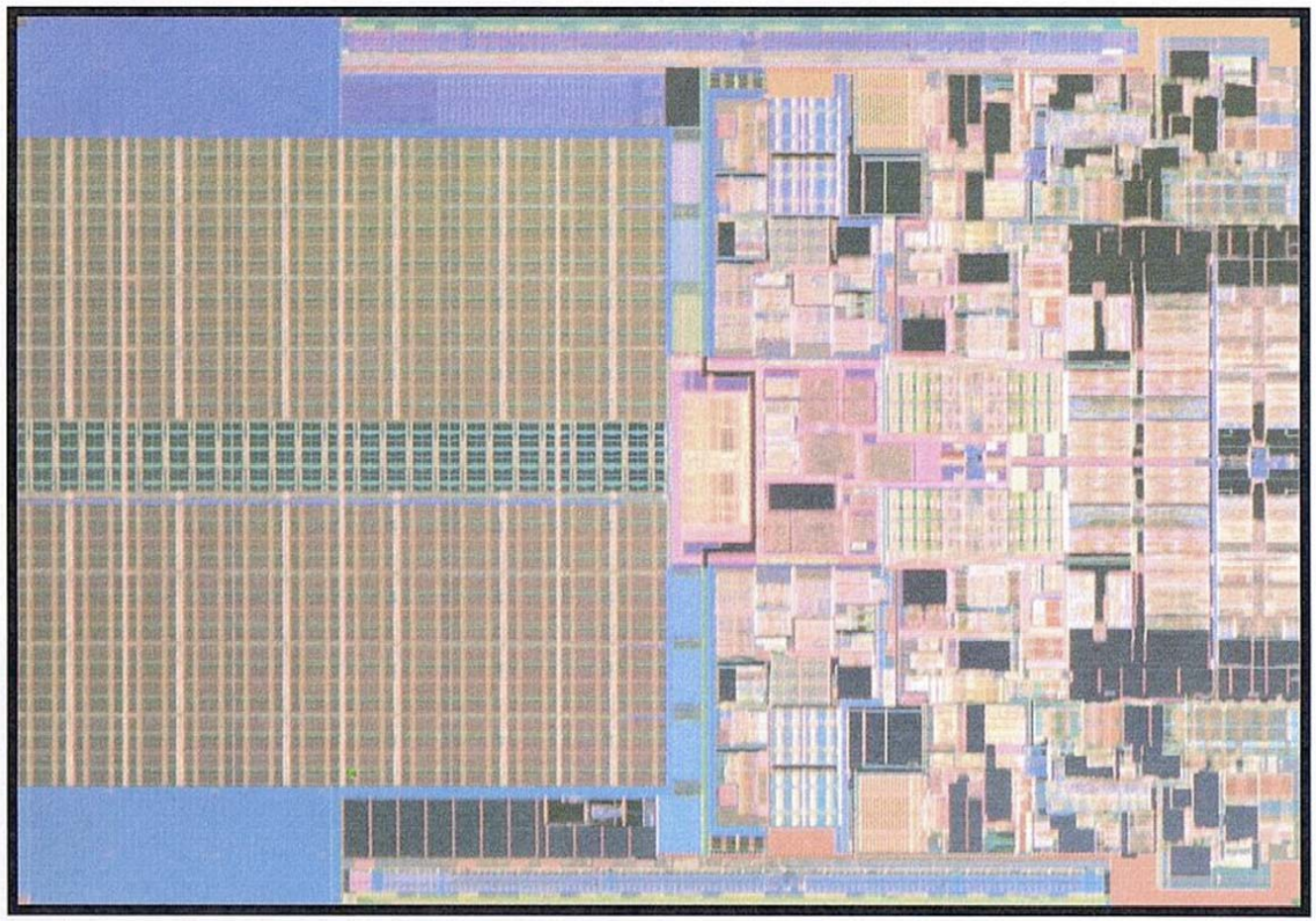
Outline

- Multicores are coming!
- Are they any good?
- Are they any good for Virtual Immersion?
- What can go wrong?
- Can 3D Chip-stack Memory on Processor help?
- Is 3D Real and Ready for Prime Time
- Will it Melt?
- Conclusions

More Motivation:
Faster Clock Repeater Crisis (Ruchir Puri @ IBM) -
Wires Don't Scale Well -
Number of Repeaters is Exploding as a Power of 10 per 33% Shrink

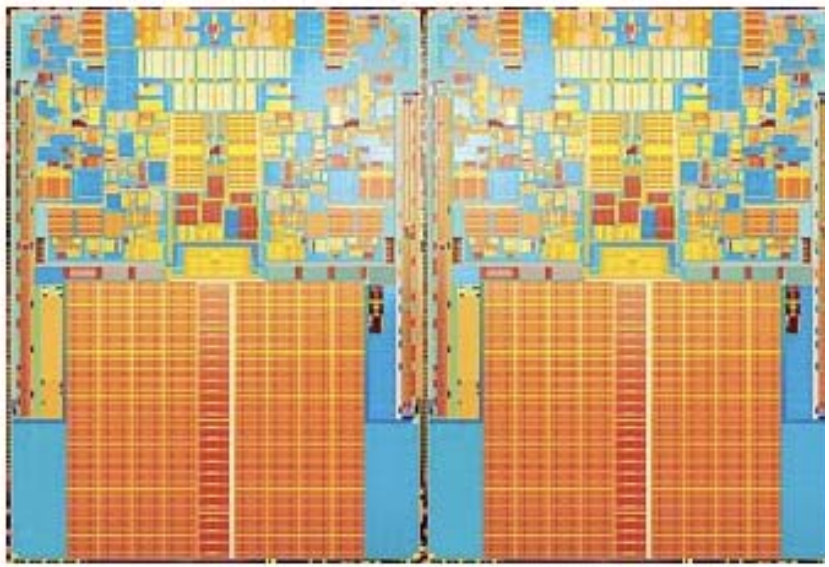


Result: Multicores

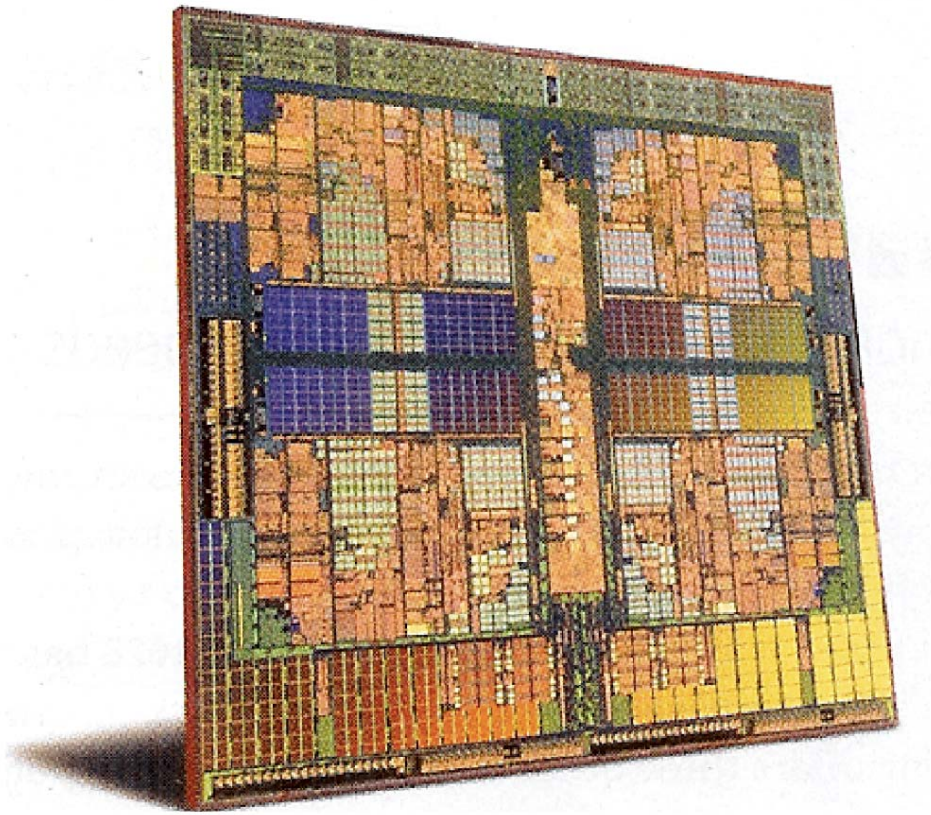


INTEL Penryn dual core 65nm

45nm Quad Core Generation

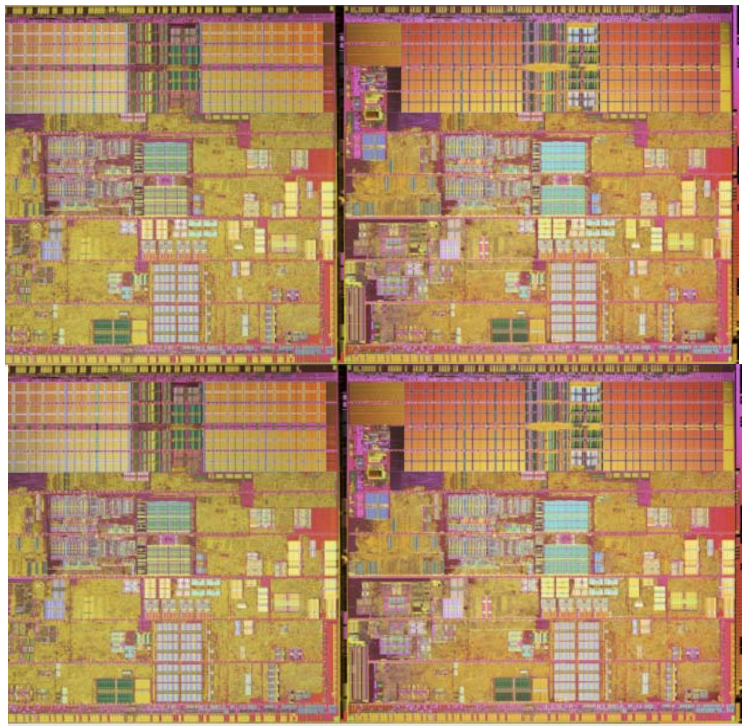


Intel 45nm quad-core processor die

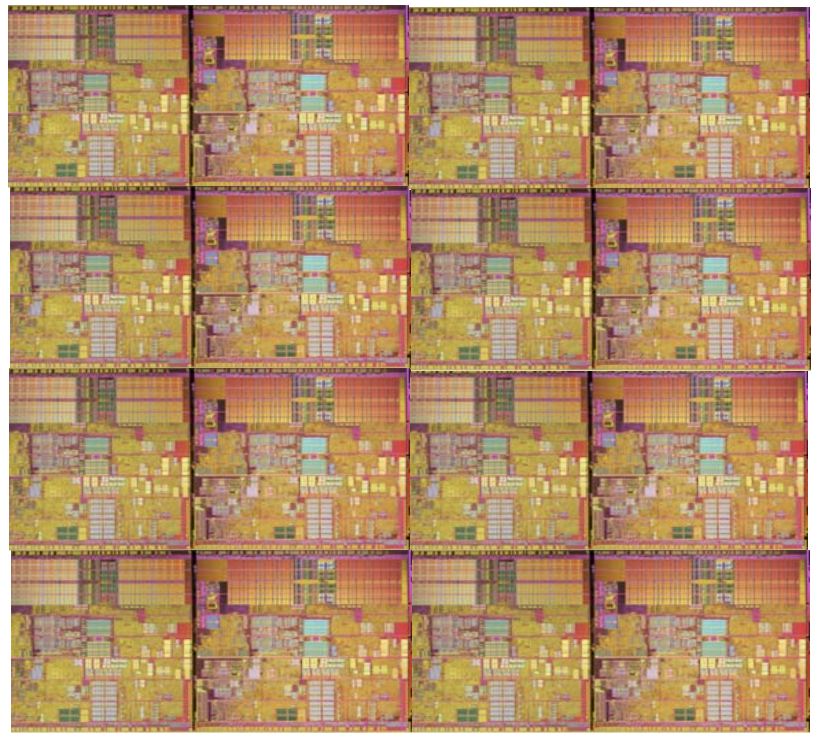


AMD X4 9850 Quad

Future of CMOS - More Multi-cores? Some say 1000!

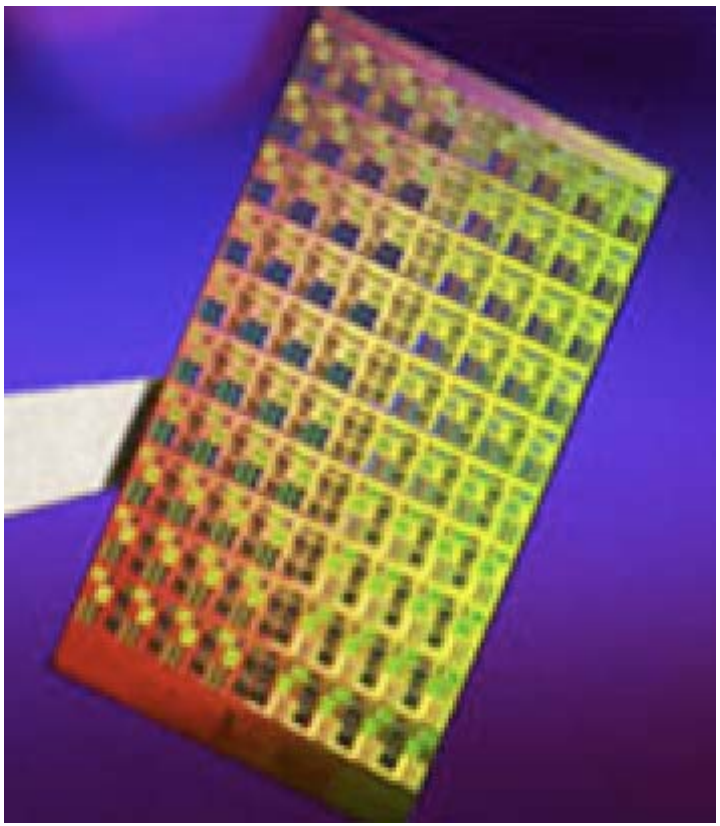


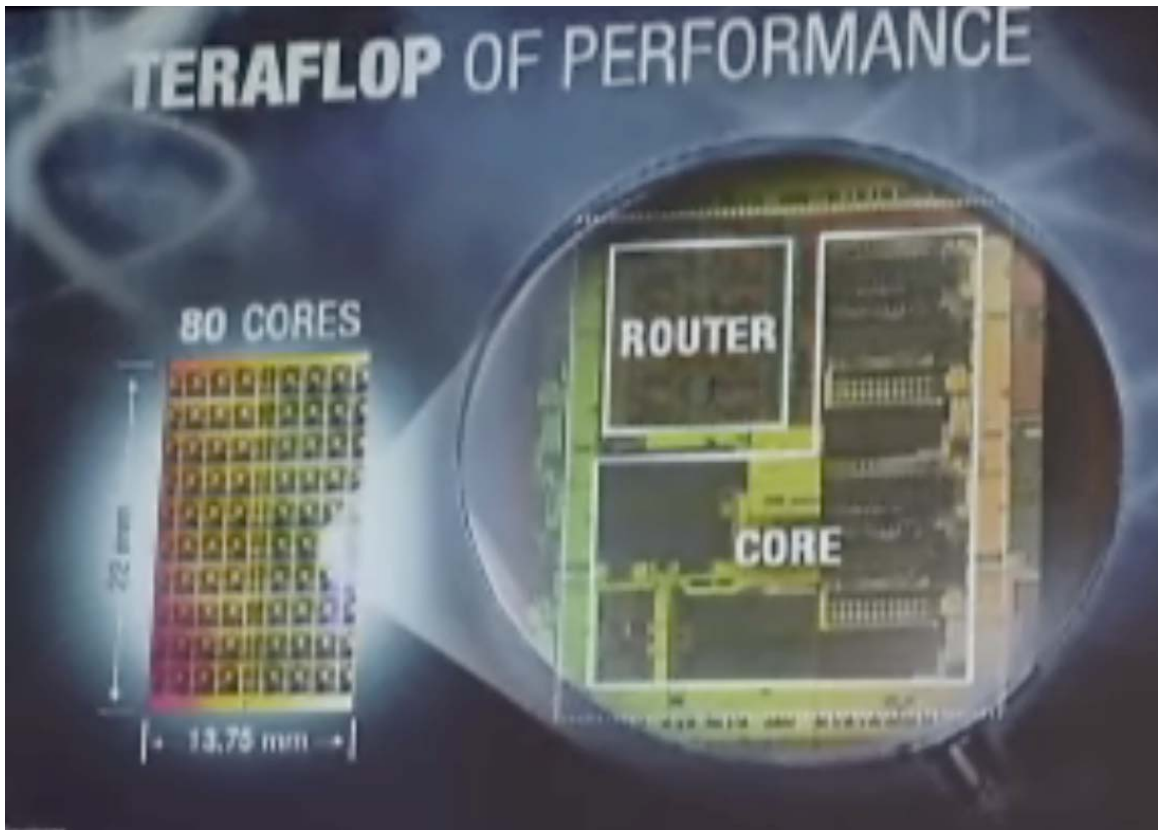
45nm: n = 4 cores



22nm: n = 16 cores

Intel 3.1GHz 80 core TERA - more (but smaller) cores on BIG dies (22mm x 13.75mm)





NVIDIA 8800 GTS 512 - 128 tinier streaming cores at 65nm @ only 1.6 GHz

Like it or NOT multicores are coming!

Are they any good?

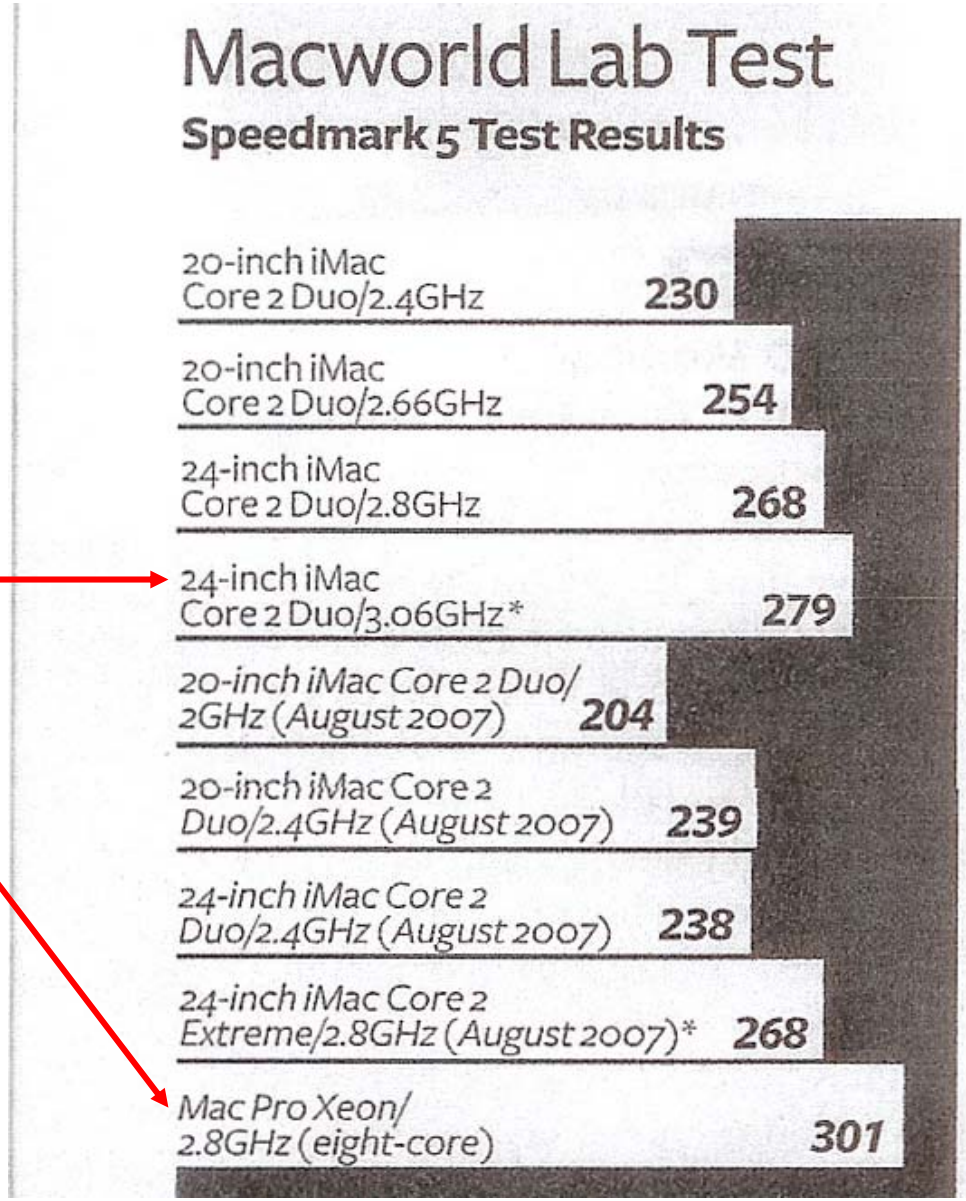
Multi-core Graphics sustains parallelism and is Spectacular!

	Peak pixel fill rate (Gpixels/s)	Peak bilinear texel filtering rate (Gtexels/s)	Peak bilinear FP16 texel filtering rate (Gtexels/s)	Peak memory bandwidth (GB/s)	Peak shader arithmetic (GFLOPS)
GeForce 8800 GT	9.6	33.6	16.8	57.6	504
GeForce 8800 GTS	10.0	12.0	12.0	64.0	346
GeForce 8800 GTS 512	**10.4**	**41.6**	**20.8**	**62.1**	**624**
GeForce 8800 GTX	13.8	18.4	18.4	86.4	518
GeForce 8800 Ultra	14.7	19.6	19.6	103.7	576
Radeon HD 2900 XT	11.9	11.9	11.9	105.6	475
Radeon HD 3850	10.7	10.7	10.7	53.1	429
Radeon HD 3870	12.4	12.4	12.4	72.0	496

How Good Are Multicores Generally?

(Sometimes, it's hard to tell. Here's one recent study - last month)

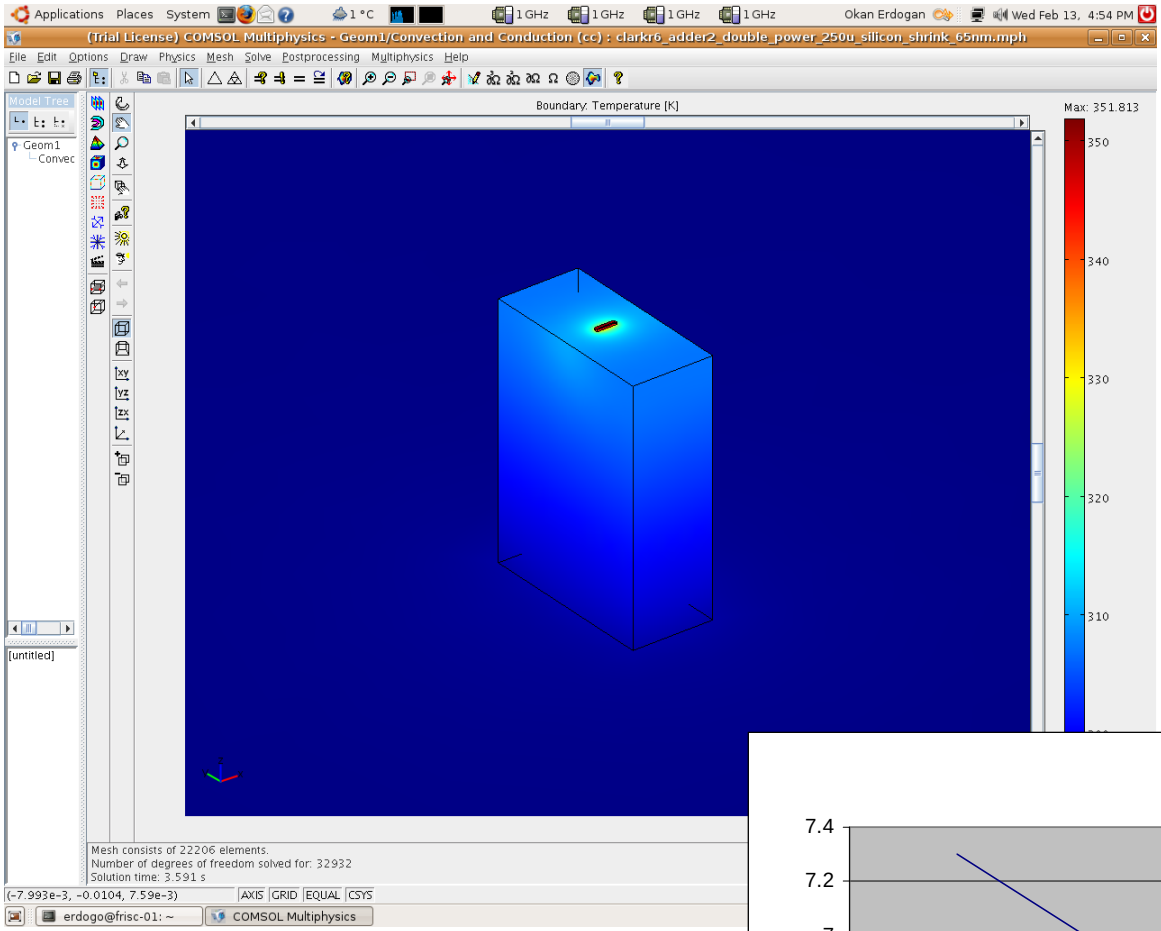
Only 8% Difference for Twice the Number of Cores?



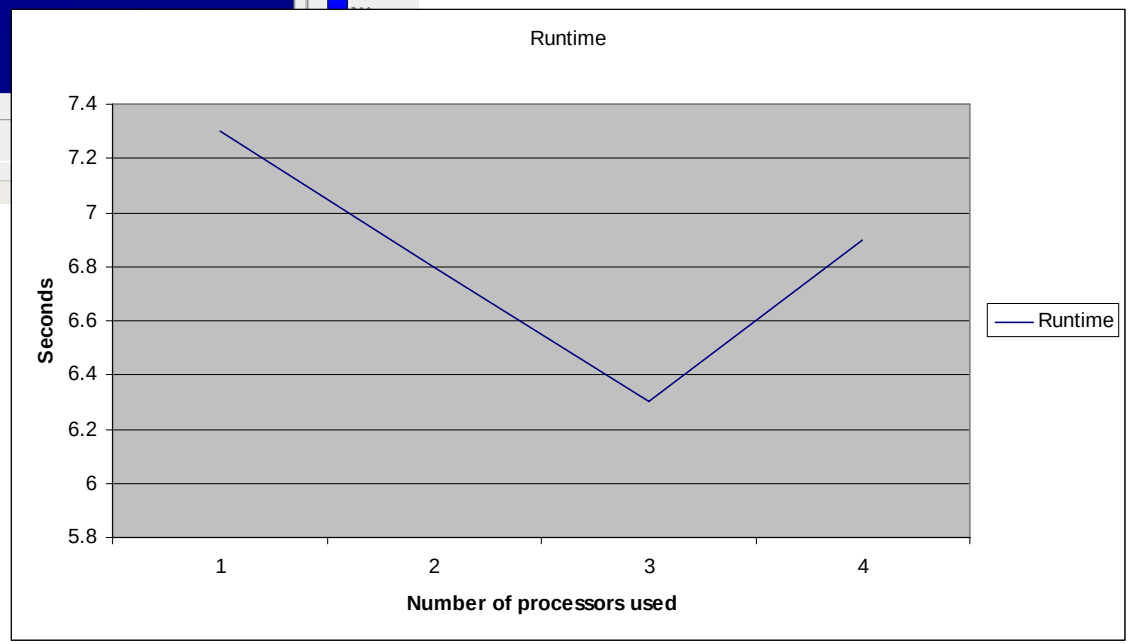
As a Professor I Would Grade this as C work.

But what does the User Community Think?

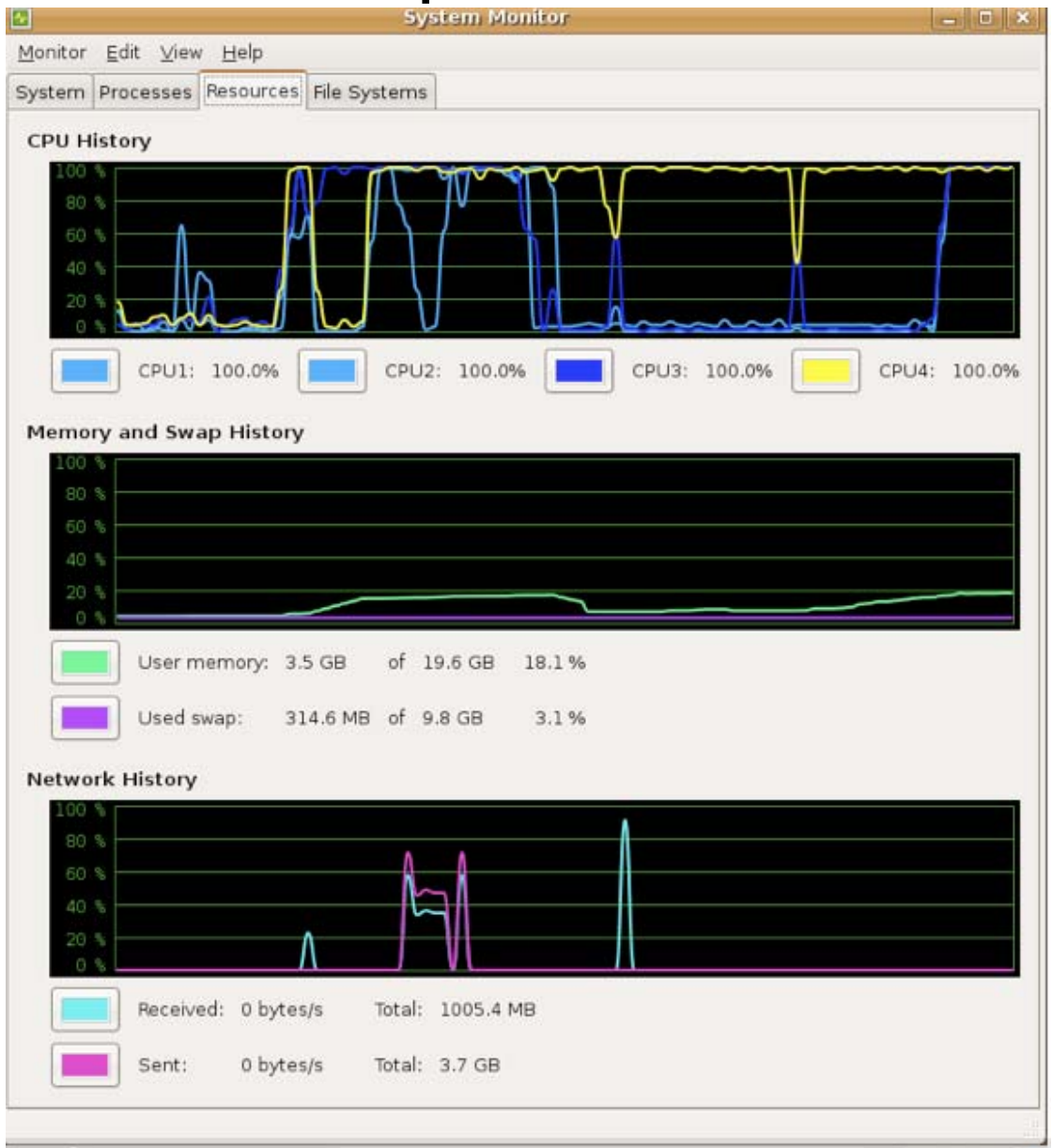
COMSOL thermal analysis (FEM)



Runtime Comparison on Quad Opteron64 Sun 40 (2.88GHz, 20GB) vs. # of Cores Enabled.



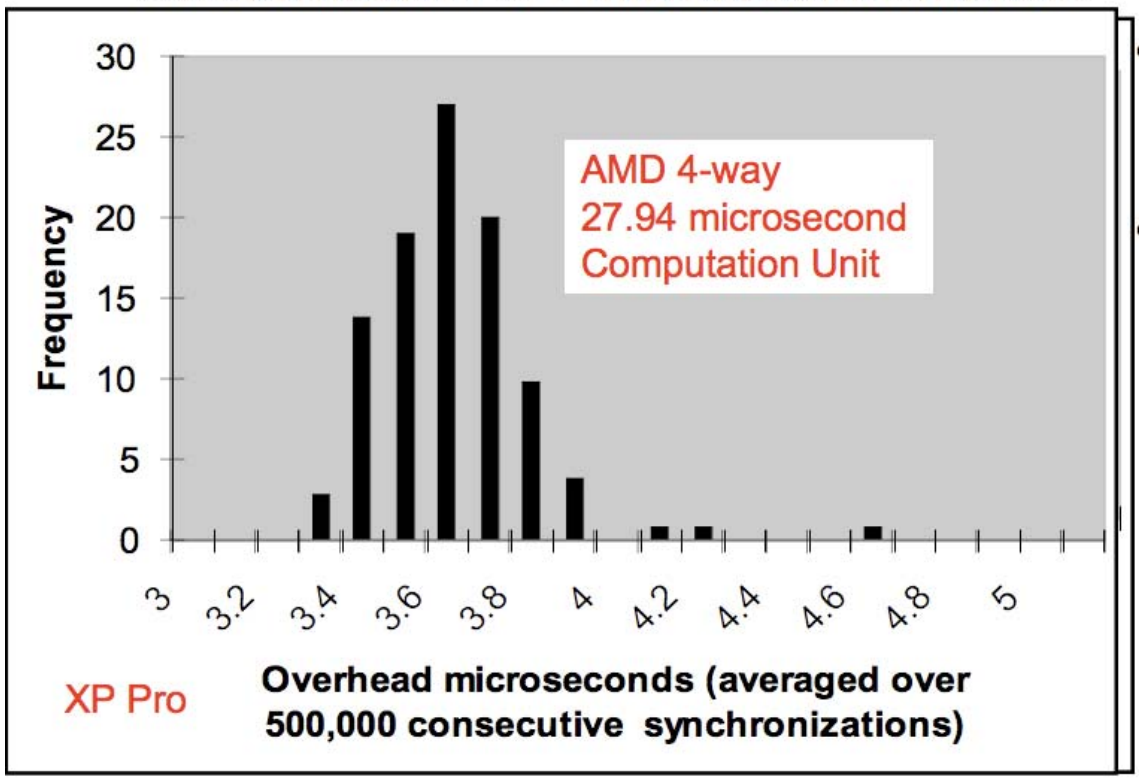
COMSOL CPU Utilization of 4 cores vs. time (Problem: Usable parallelism varies a lot.)



Software Headaches -Synchronization

8-way Parallel Pipeline on two 4-core Xeon

- Histogram of 100 runs -- each run has 500,000 synchronizations following a thread execution that takes 33.92 microseconds
  - So overhead of 6.1 microseconds modest



- Message size is just one integer
- Choose computation unit that is appropriate for a few microsecond stage overhead

What's the best we could EVER hope for?

Amdahl's 1967 Figure of Merit (FOM) for Parallelism

$$PerformanceFOM = \frac{S + P}{S + \frac{P}{n}}$$

$$S = fraction - serial - code$$

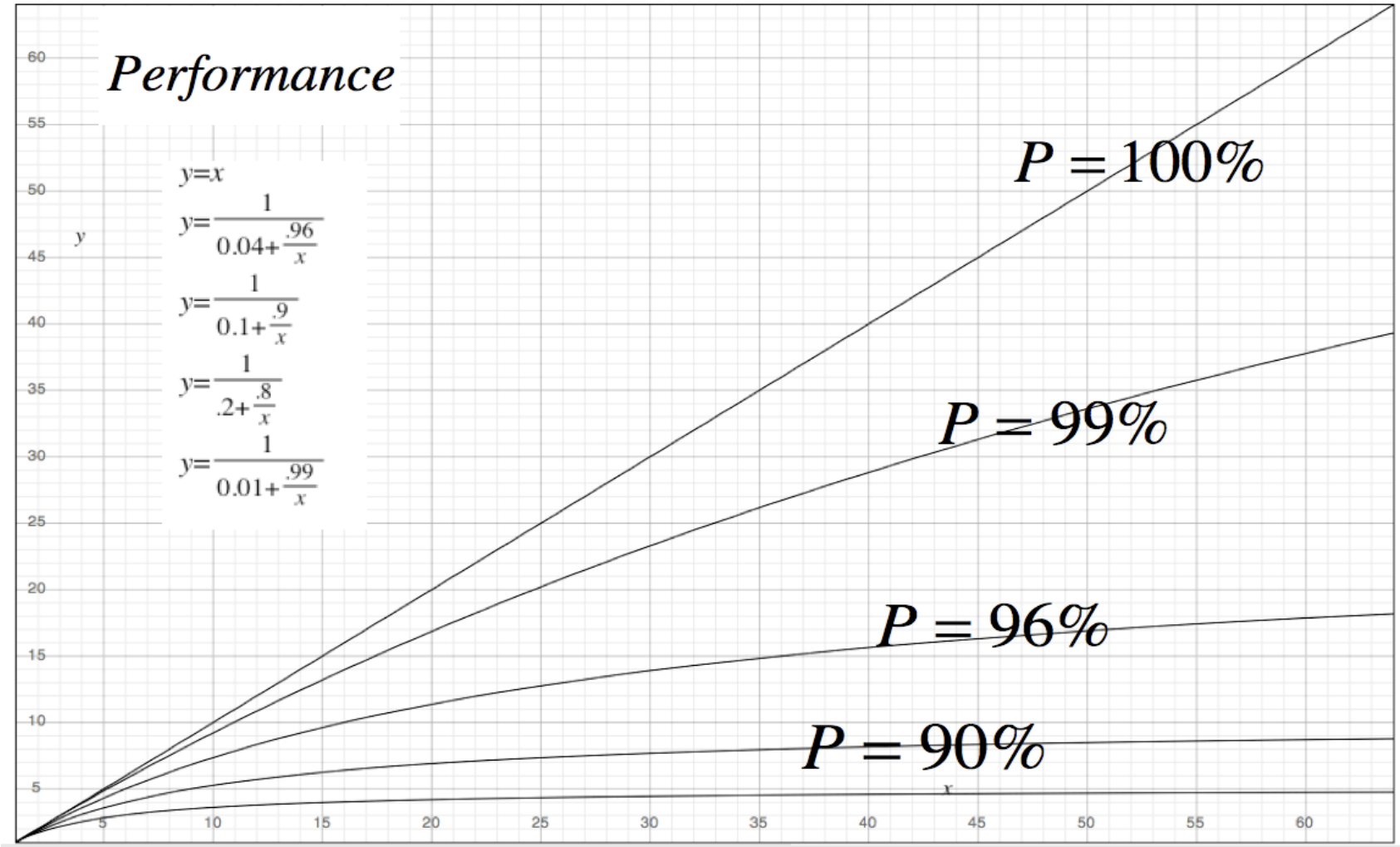
$$P = fraction - parallel - code$$

$$S + P = 1$$

$$\lim_{n \to \infty}[FOM] = 1 + P / S; \lim_{S \to 0}[FOM] = n$$

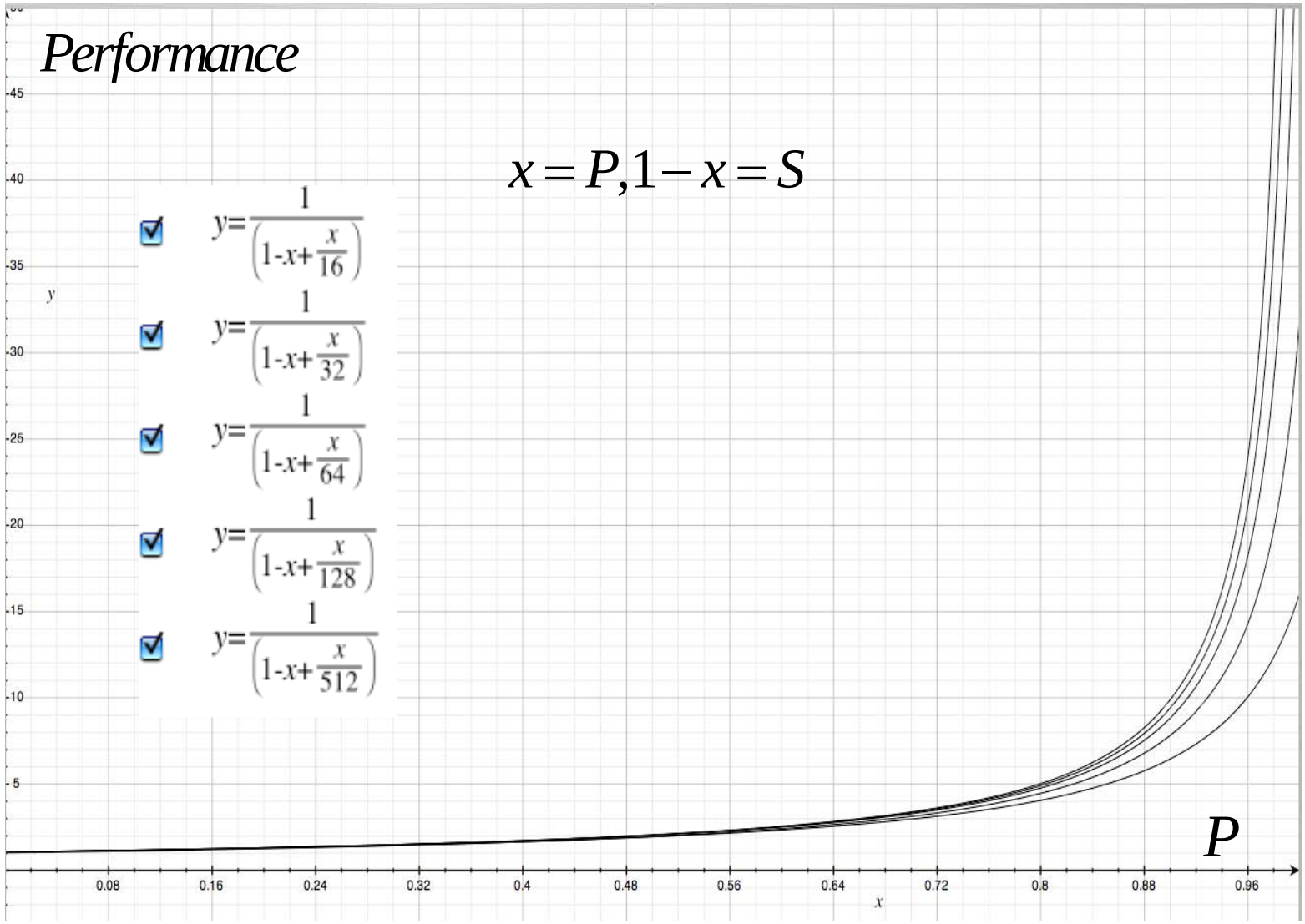
Pretty nifty if S is near ZERO. How close does it have to be?

Usual Way to Plot Amdahl's FOM vs. n (parameterized by P)



$x = n$

Alternate Graphical View: Calculated Performance (parameterized by n for n=16,32,64,128,512 Parallel Processors) vs. P



So….Unless S is less than 4% there is no Big Bang for the Buck!

- FORTUNATELY THERE ARE SUCH ALGORITHMS, BUT THEY CAN BE DEFEATED IF ONE IS NOT CAREFUL.
- THE GOOD NEWS: ONE OF THEM IS GRAPHICS!
- BUT VIRTUAL IMMERSION IS MORE THAN GRAPHICS. SO THE BAD NEWS IS WE MAY STILL HAVE A PROBLEM.
- LATENCY IS ONE ENEMY, DELAY IS ANOTHER, AND THERE ARE OTHERS.

WHAT CAN GO WRONG?

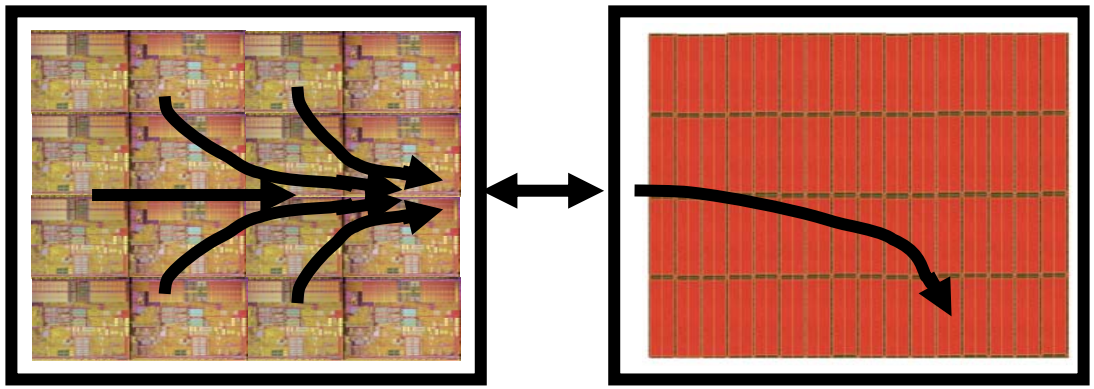
Task Synchronization, Memory Access, Threading Overhead and Message Routing
cause loss of cycles, L(n).

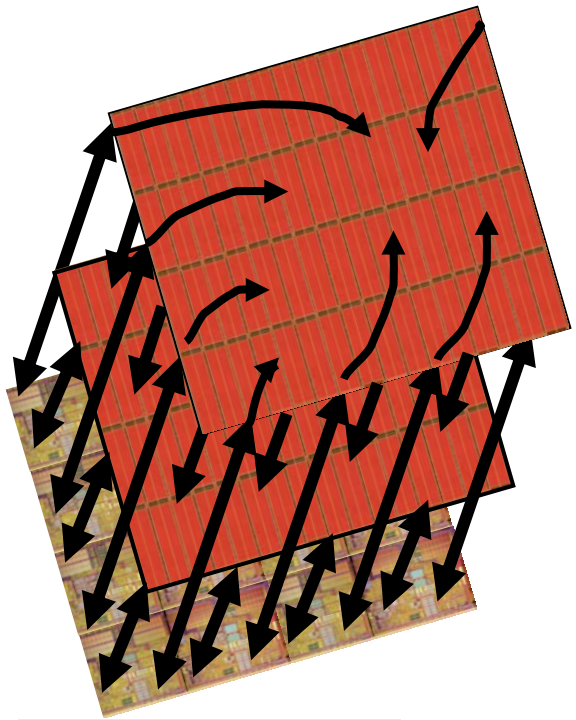
$$FOM_2 = \frac{S+P}{S+\frac{(P+L)}{n}} = \frac{S+P}{(S+B)+\frac{P}{n}}; B = L/n$$

This shows that the Lost cycles, L(n) , can *masquerade* as pseudo sequential code in this figure of merit if B(n) [the average lost cycles per processor] is constant vs. n, but if L(n) is constant we get good results as B(n) then goes down with increasing n. Reality is somewhere in between.

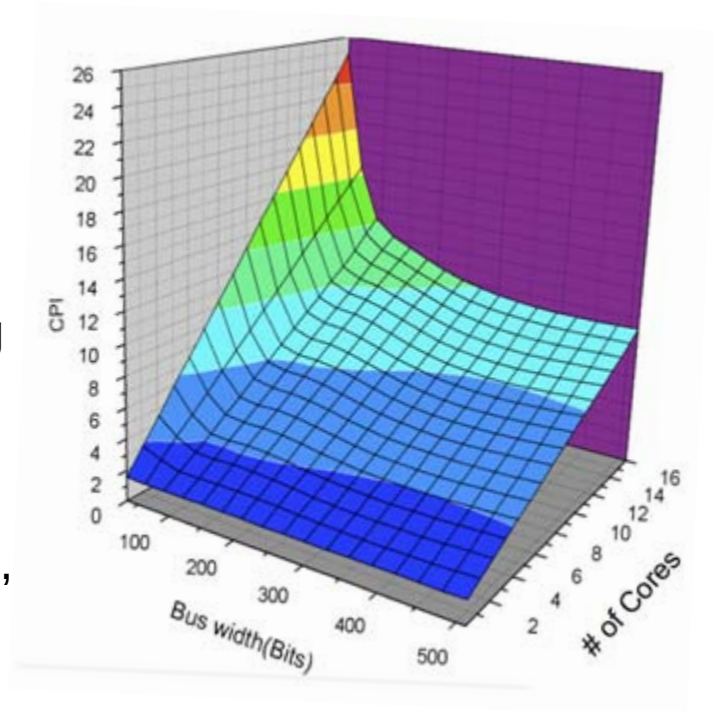
Multi-core Memory Processor Throttling Effects with Conventional Packaging

Multiple Cores may all try to access external memory at the same time and through the same narrow package portal -- Resulting in Blockage of all but One Access





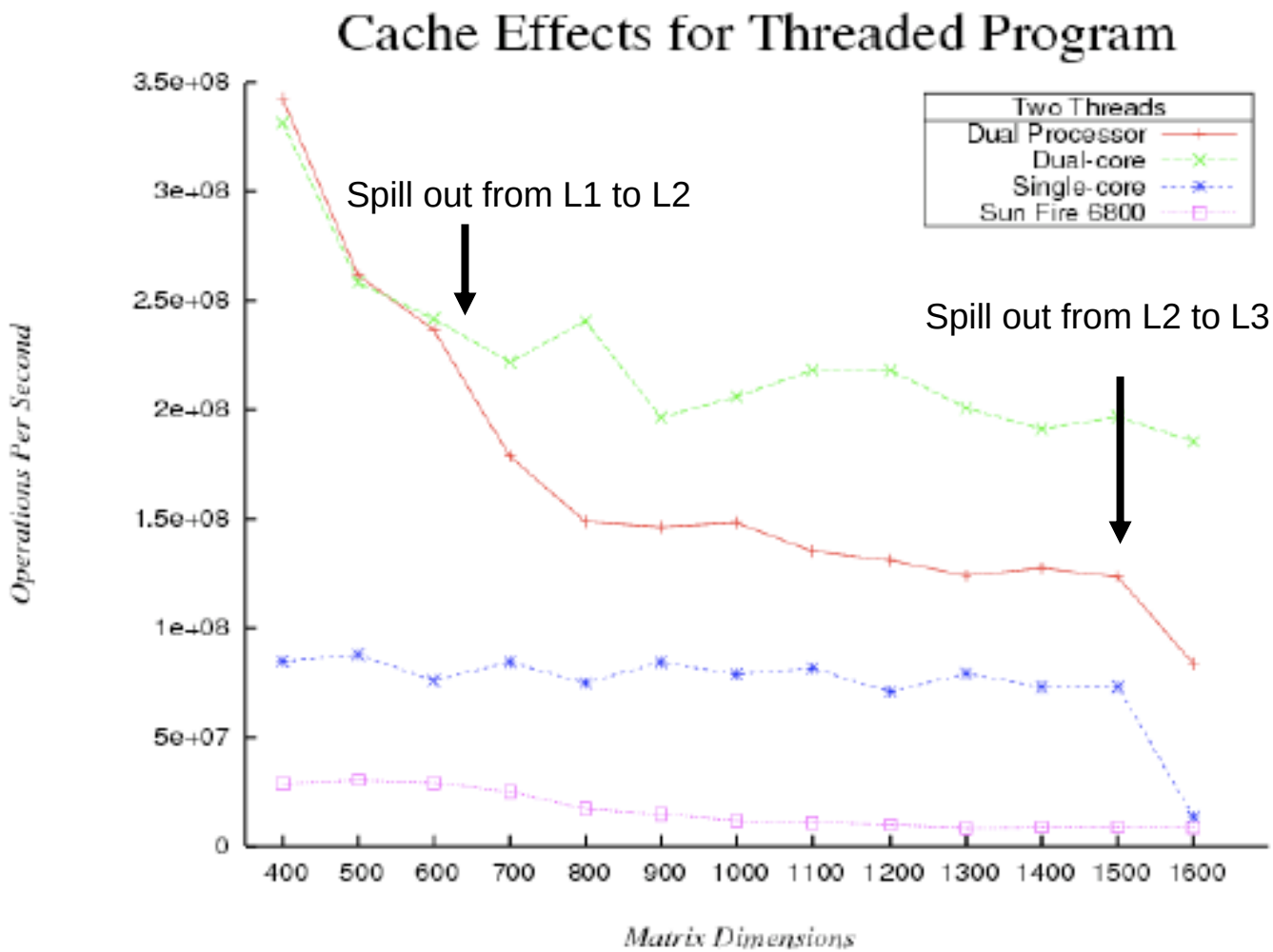
Ultra-wide TSV Bit Path Mitigation of Multi-core Memory Wall CPI Using 3D Memory Processor Stacks, Sophisticated Memory Management, Heterogeneous Memory, and Heat Spreaders

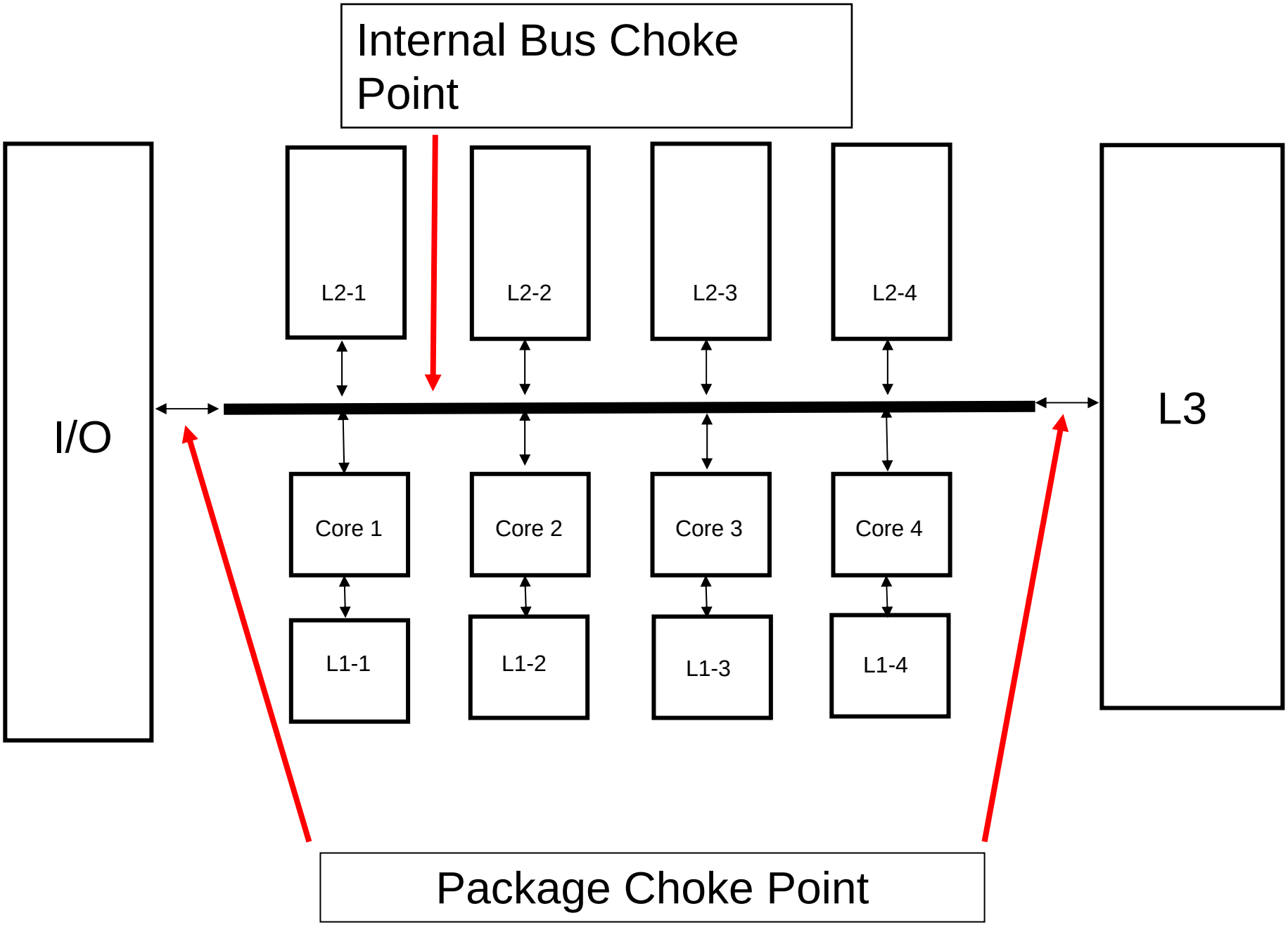


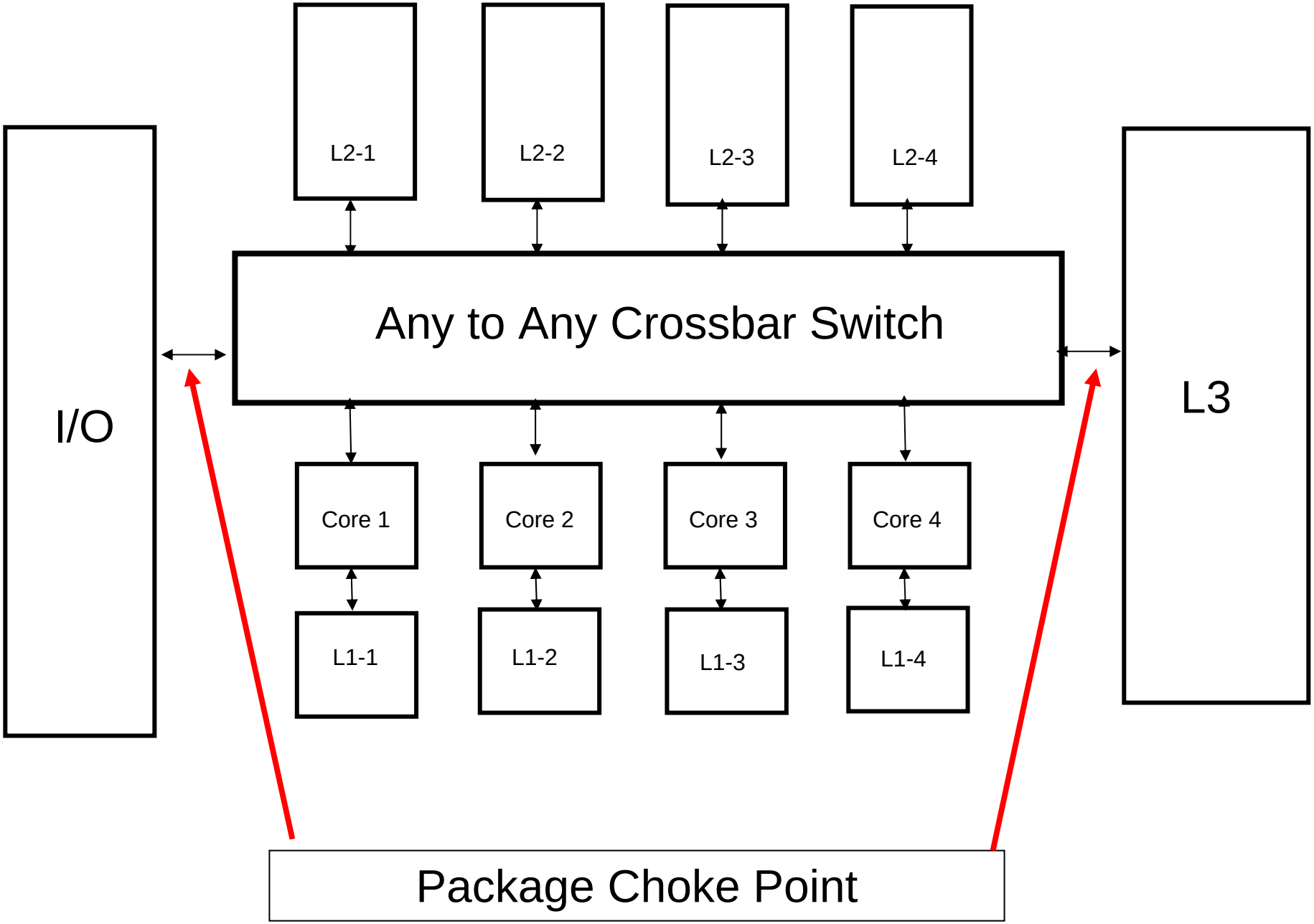
Semiconductor Research Corporation

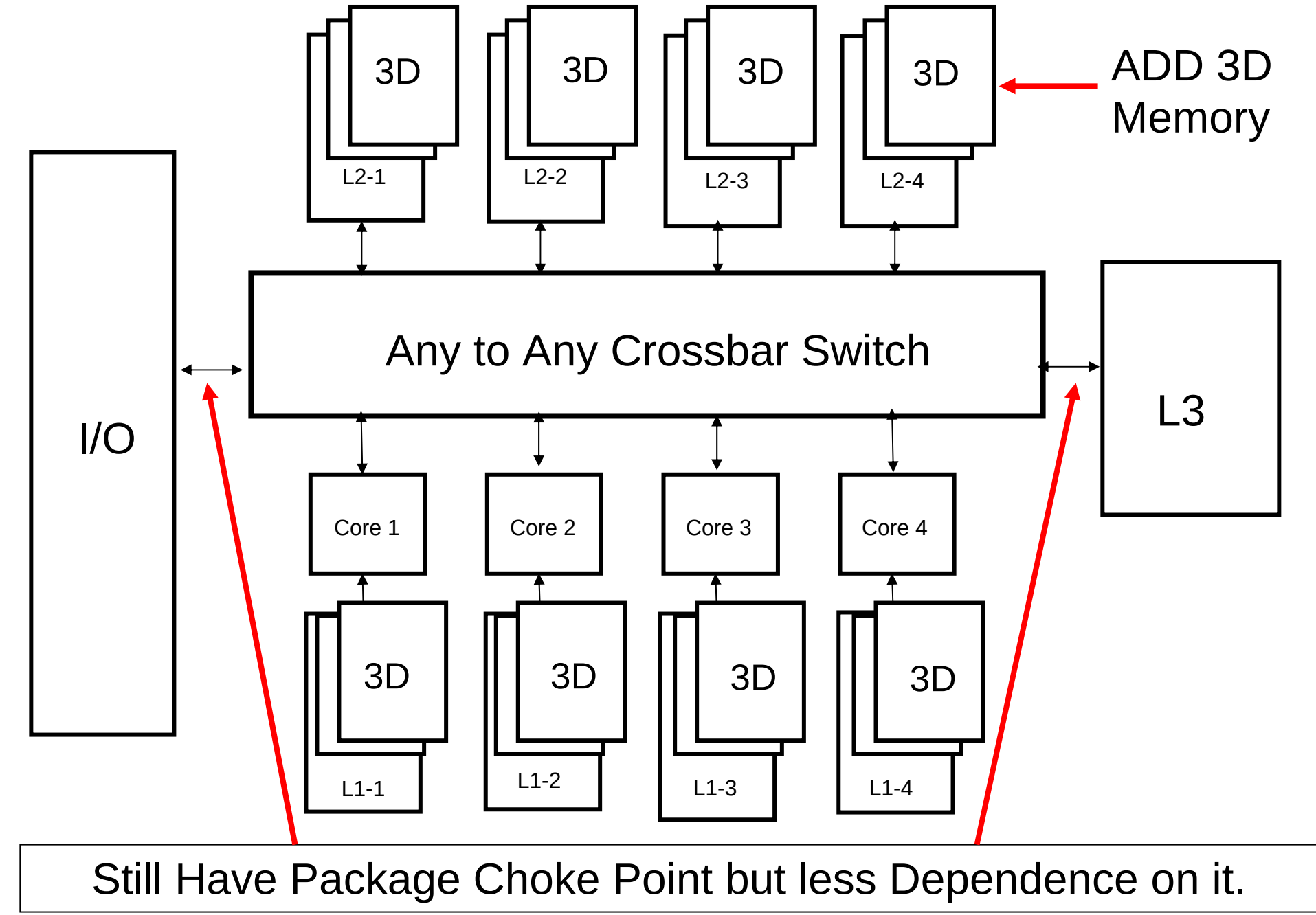
Rensselaer

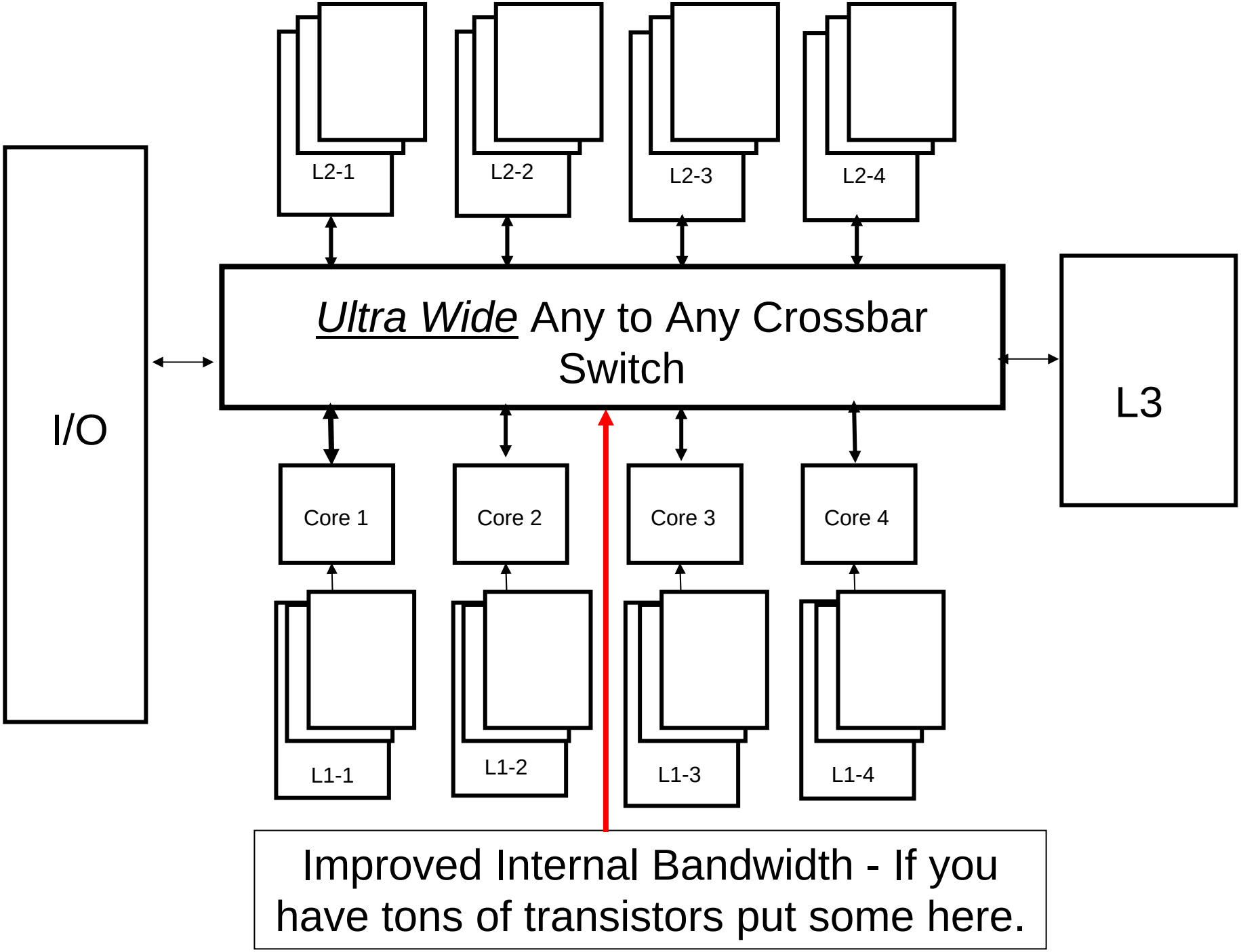
Excerpt from 2007 Paper by Stephanie Dinkins (2007) - on “Embarrassingly Parallel” Matrix Multiplication

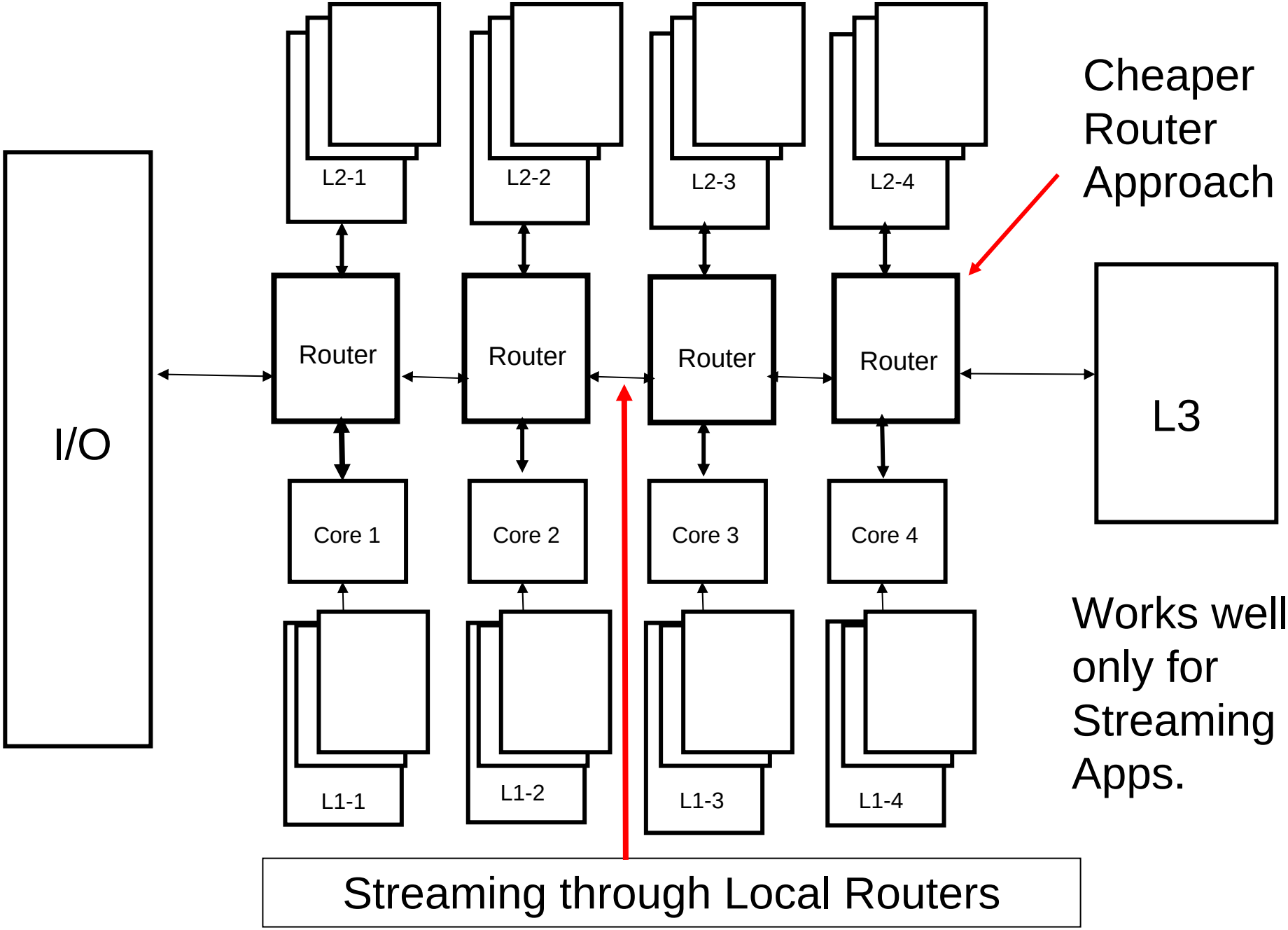



Internal Bus Choke Point
I/O
L2-1
L2-2
L2-3
L2-4
L3
Core 1
Core 2
Core 3
Core 4
L1-1
L1-2
L1-3
L1-4
Package Choke Point

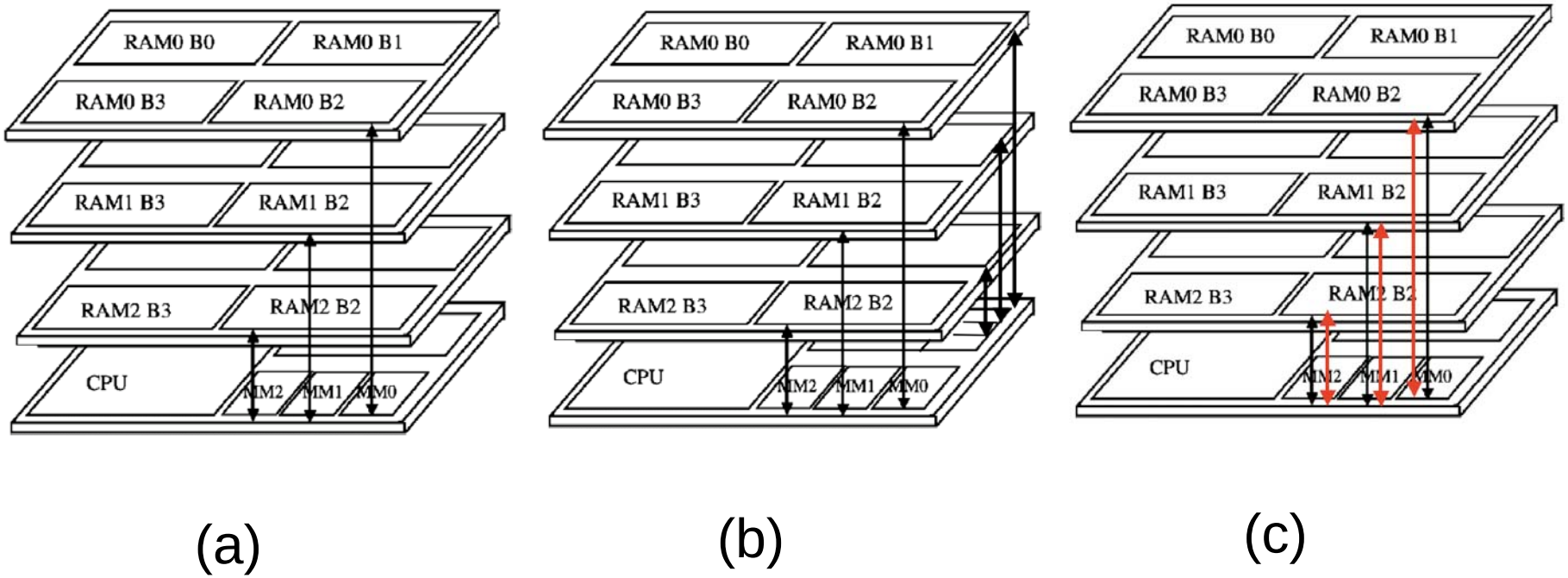

L2-1
L2-2
L2-3
L2-4
I/O
Any to Any Crossbar Switch
L3
Core 1
Core 2
Core 3
Core 4
L1-1
L1-2
L1-3
L1-4
Package Choke Point


3D
3D
3D
3D
ADD 3D Memory
L2-1
L2-2
L2-3
L2-4
I/O
Any to Any Crossbar Switch
L3
Core 1
Core 2
Core 3
Core 4
3D
3D
3D
3D
L1-1
L1-2
L1-3
L1-4
Still Have Package Choke Point but less Dependence on it.


L2-1
L2-2
L2-3
L2-4
I/O
Ultra Wide Any to Any Crossbar Switch
L3
Core 1
Core 2
Core 3
Core 4
L1-1
L1-2
L1-3
L1-4
Improved Internal Bandwidth - If you have tons of transistors put some here.


Cheaper Router Approach
I/O
L2-1
L2-2
L2-3
L2-4
Router
Router
Router
Router
L3
Core 1
Core 2
Core 3
Core 4
L1-1
L1-2
L1-3
L1-4
Works well only for Streaming Apps.
Streaming through Local Routers

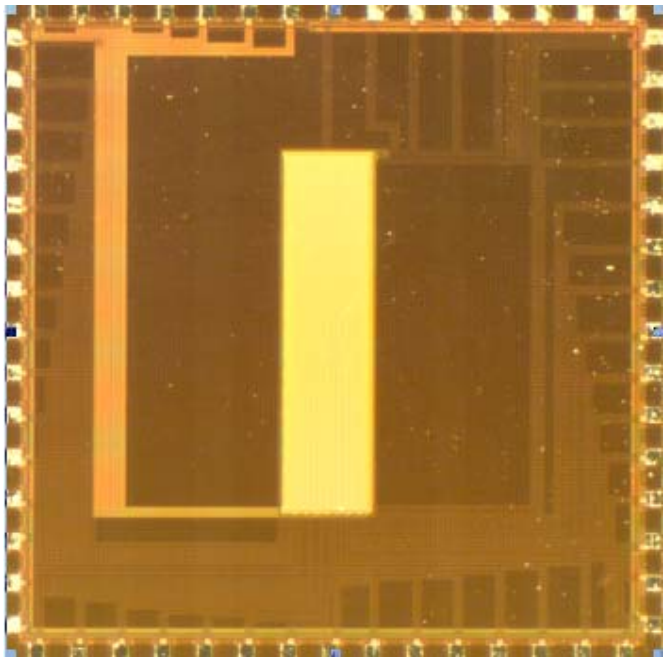
Three Ways 3D can help with Memory Latency



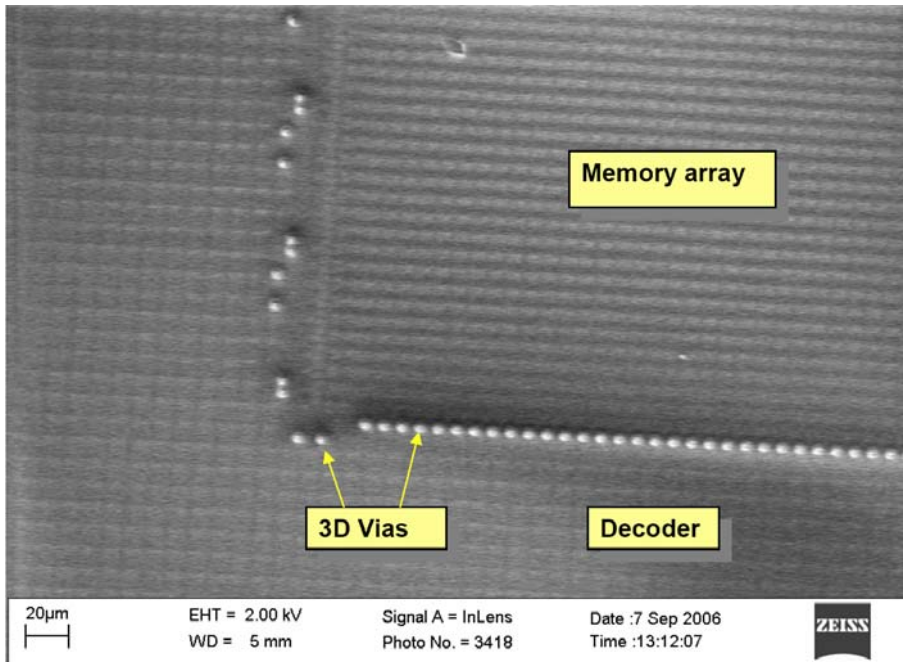
(a) (b) (c)

RPI 3D SRAM Fabricated using Lincoln Labs 1st 3D MPW - 2005.

(a)

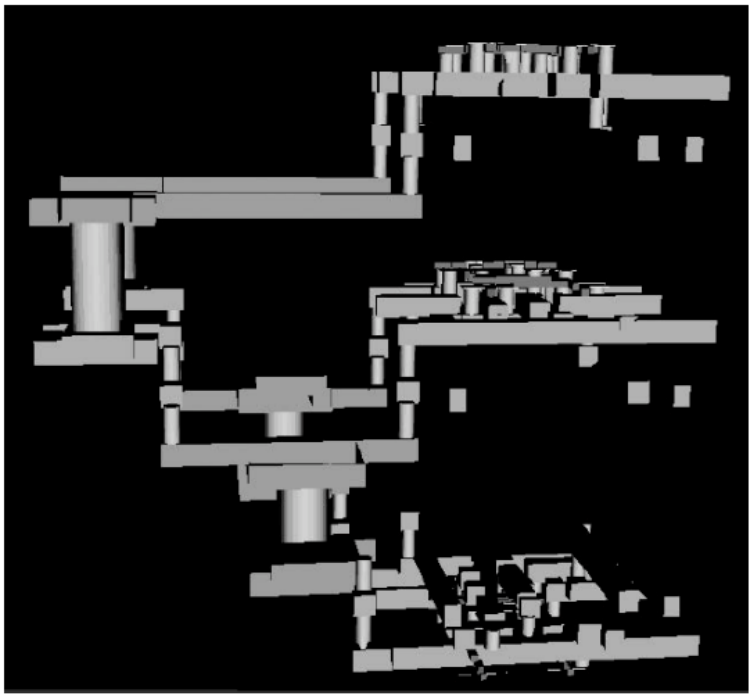


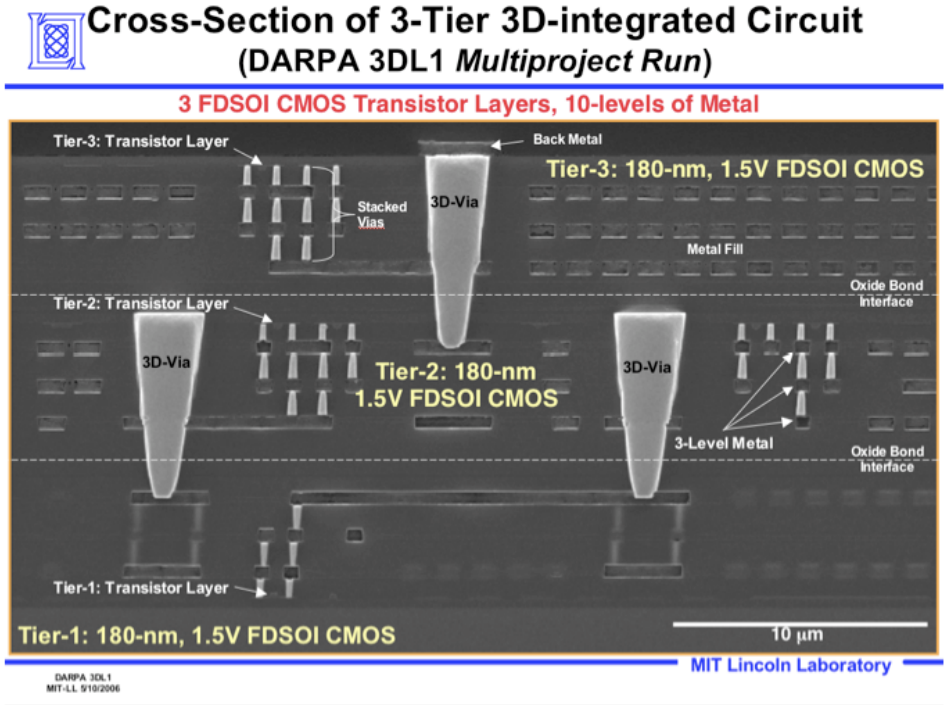
(b)



Color optical micrograph (a) and SEM top view (b). On the top layer some features can be seen with the SEM, but underneath these lie the obscured remaining circuits. Optically the top layers are thin enough to see some wiring under the unused areas, but only if there are no overlying circuits.

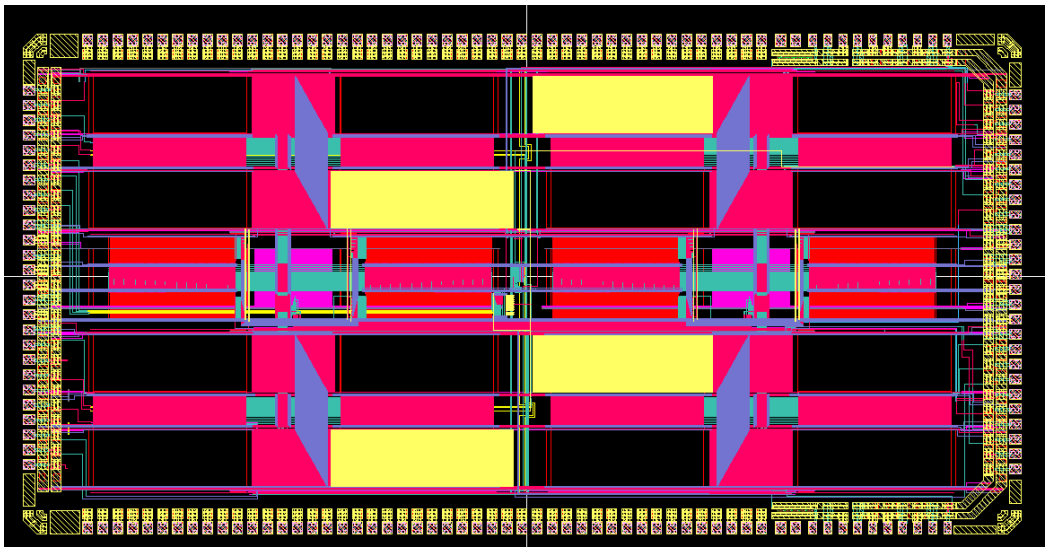
Lincoln Labs 3D Wafer to Wafer Oxide-Oxide Chemical Bonding Process

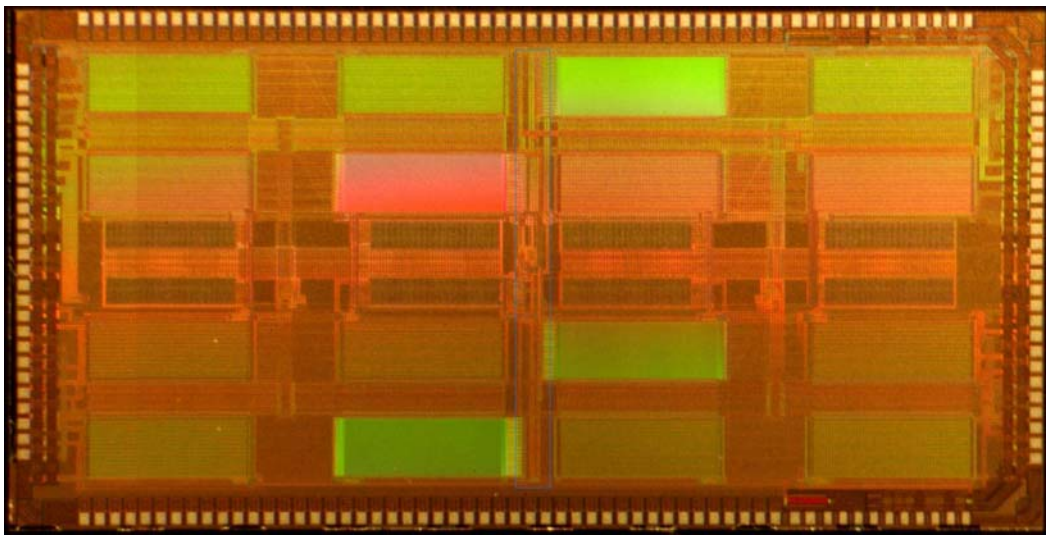




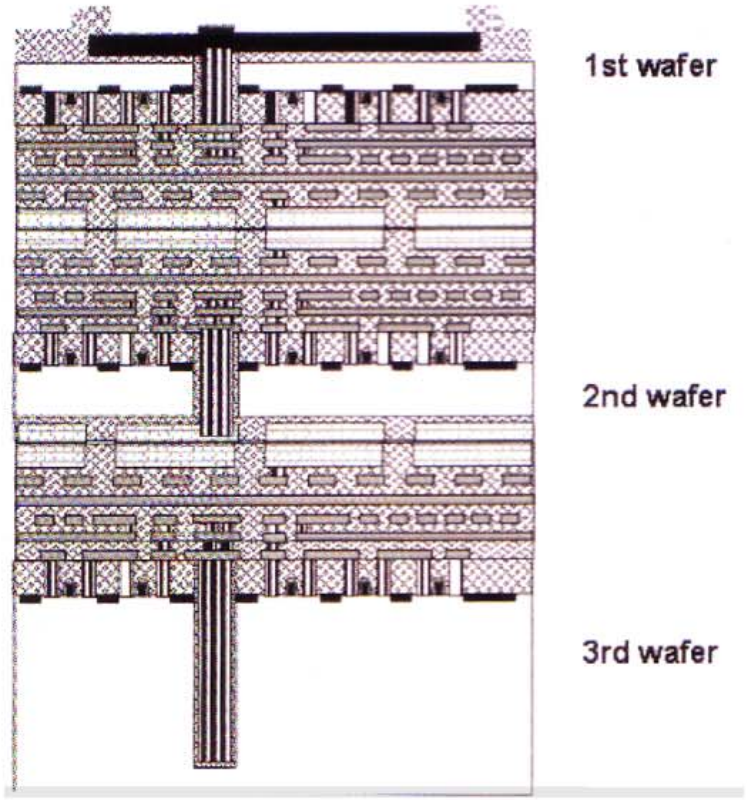
Oxide-Oxide OR Chemical Bonding is so strong that attempts to delaminate will destroy the silicon in the stack.

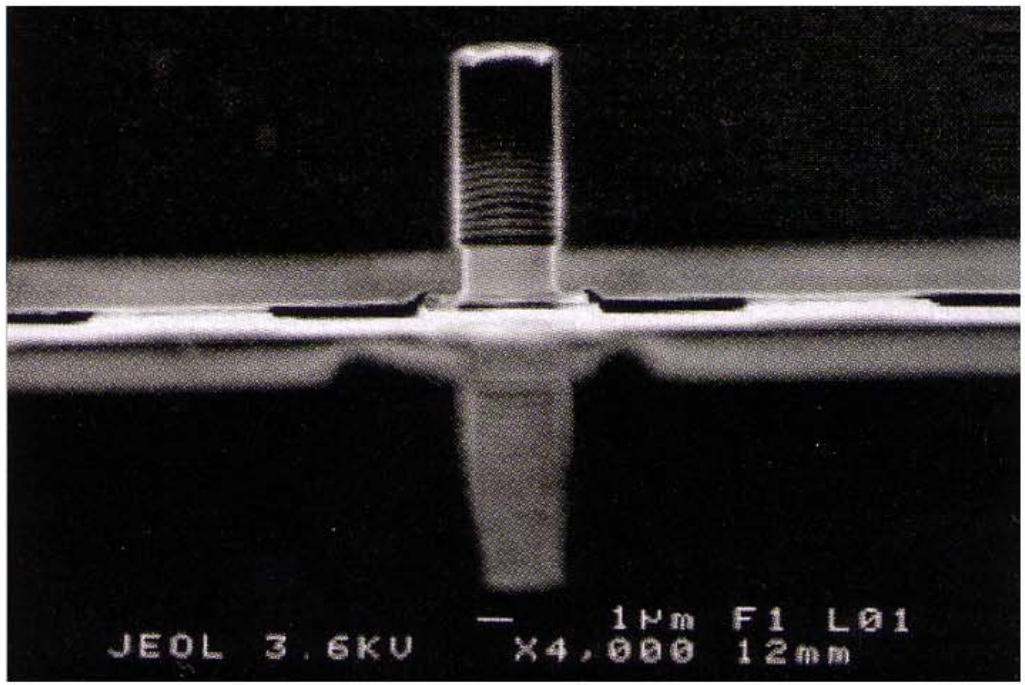
RPI 3D L2 Cache: Chip layout and photomicrograph
Lincoln Labs MPW2 (memory actually works!)





Tezzaron 3D Wafer to Wafer Atomic Cu-Cu Bonding



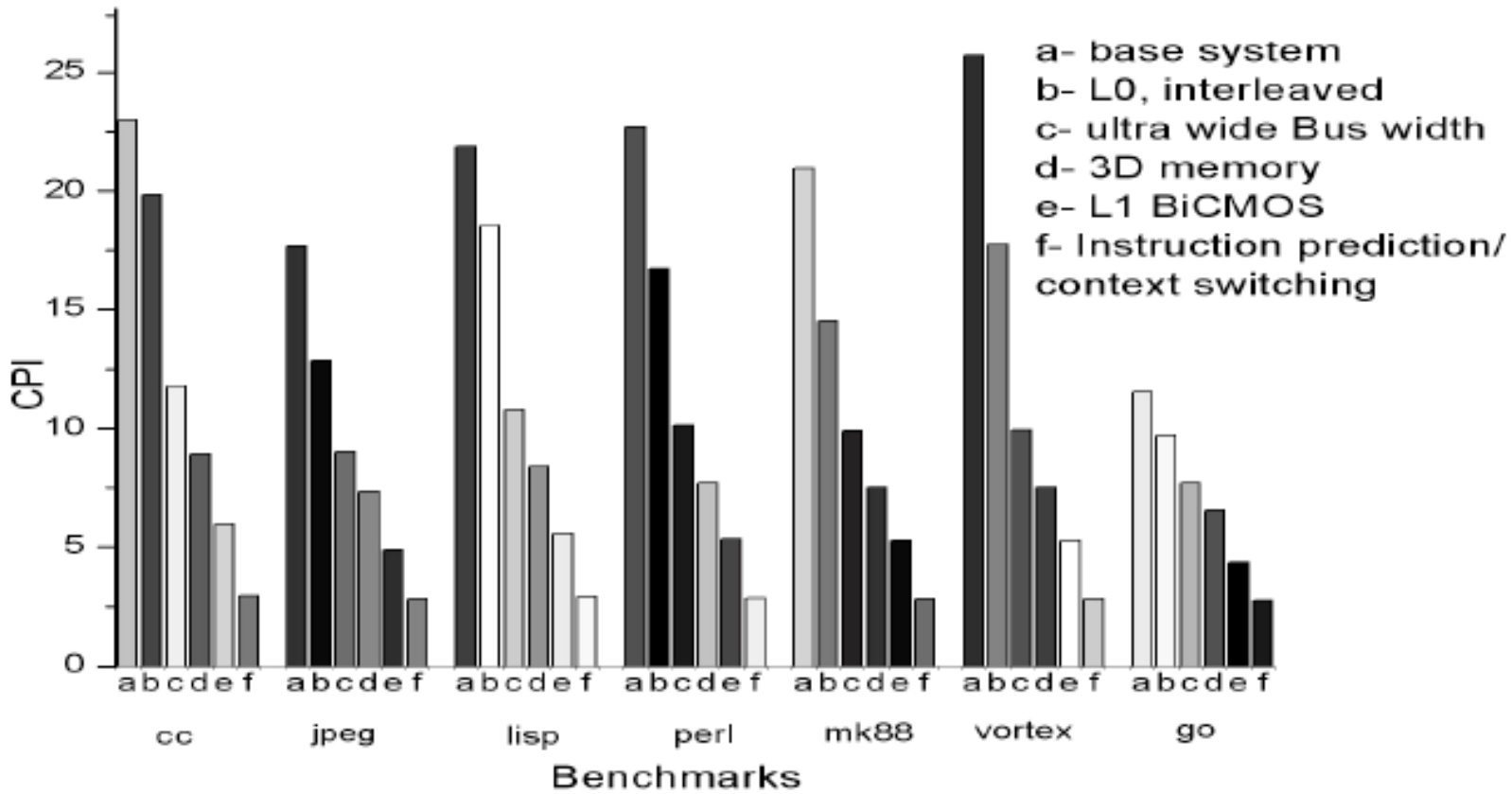


Cu-Cu bonding offers another pathway to 3D stacking

BENEFITS OF ULTRA WIDE VERTICAL 3D VIAS, AND SHORTER WIRE TO AND INSIDE MEMORY.

BUT - MANY STRATEGIES ARE NEEDED,

Hence 3D for Heterointegration may be needed.



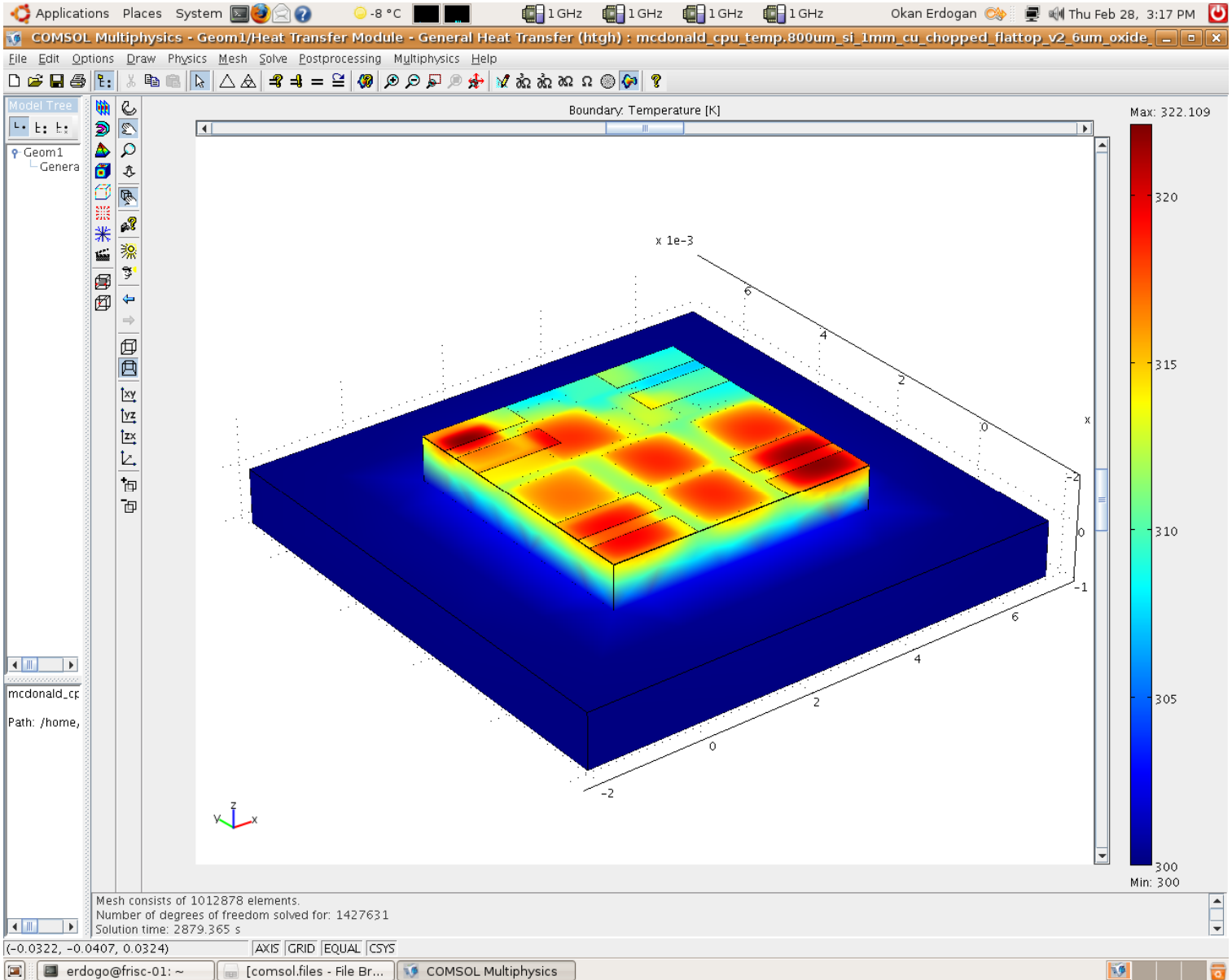
The Terahertz Processor Guy

Memory bandwidth is the biggest challenge tomorrow's CPUs face, said Jerry Bautista, the man responsible for managing technology directions at Intel's microprocessor research lab. Techniques such as 3D chip stacks and new kinds of memory hierarchies hold the most promise for busting through the bottleneck, said Bautista who helped supervise the 80-core Terahertz Processor Intel discussed at ISSCC last year.

Thermal Issue: Will 3D MELT?

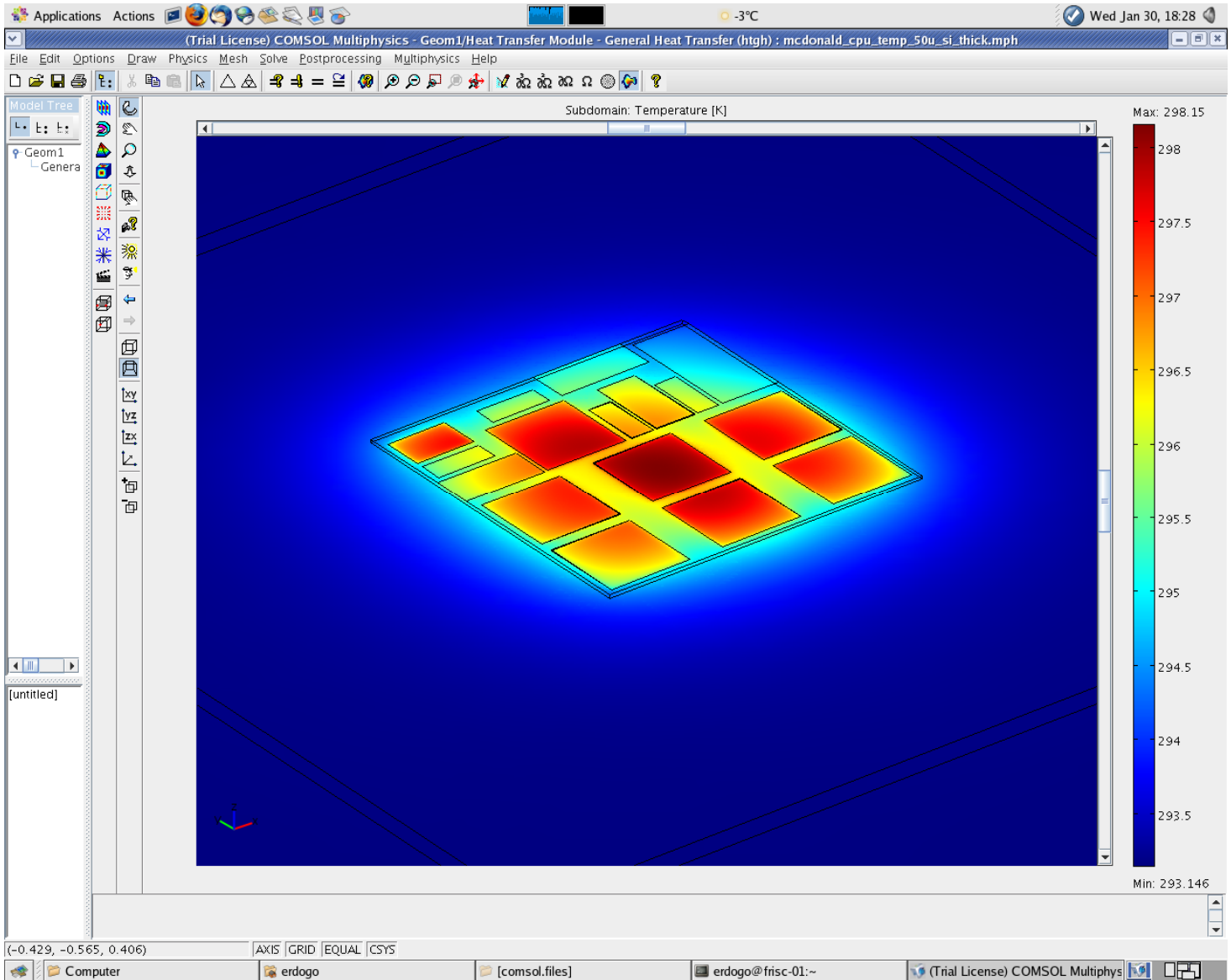
Single Tier of 3D memory over Processor --

show's thermal plume of underlying processor hot spots penetration through memory layer



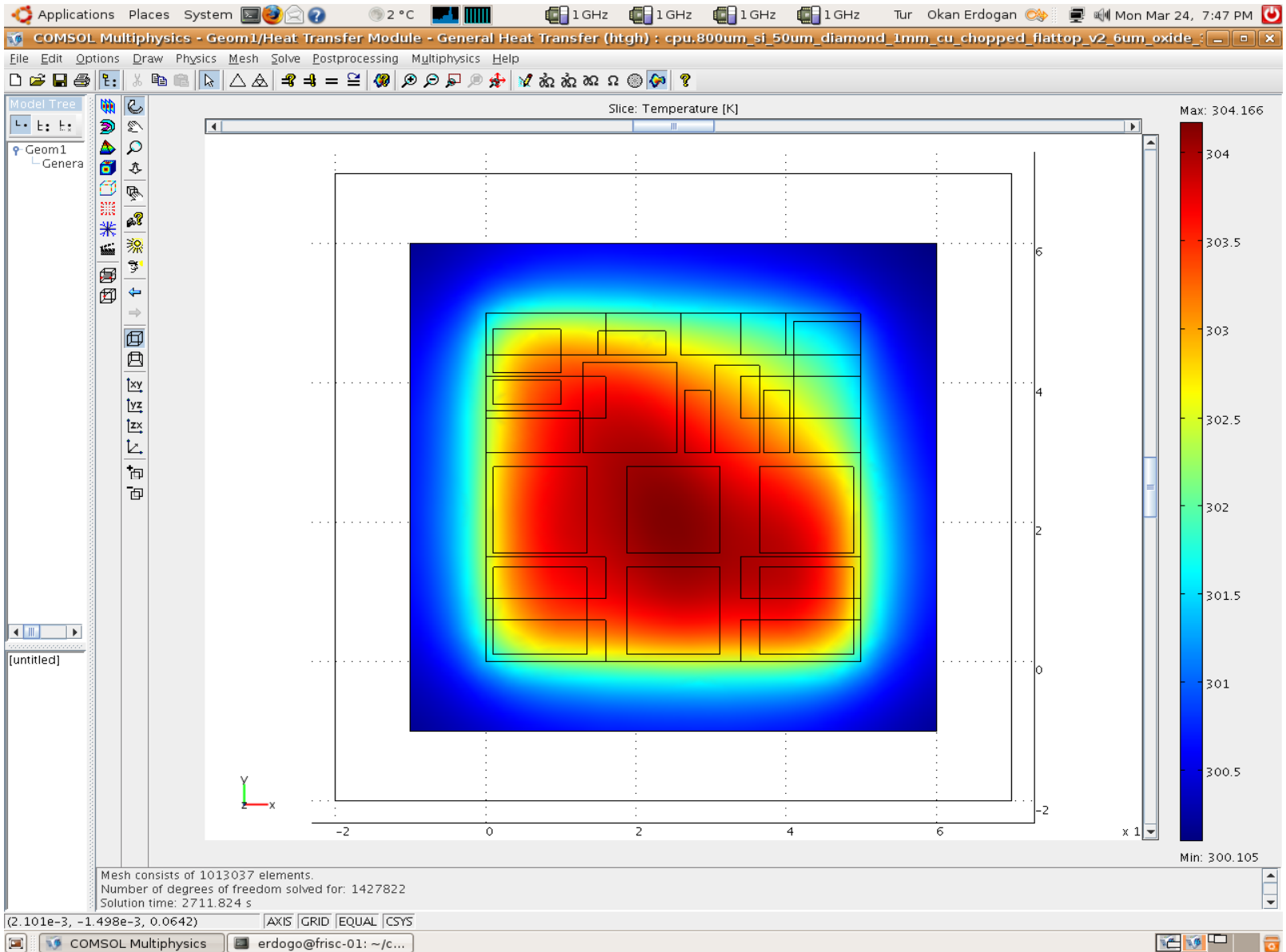
Si Thinned to 50 microns on Diamond of 150 microns on Cu of 1mm (Processor Only)

298ºK, Only 3ºK of non-uniformity (Top View)



View at diamond-Cu boundary

for 150 µm Diamond layer under CPU with One Tier of 3D Memory



Conclusions:

- Multi-core architectures can facilitate the Graphics part of Virtual Immersion.
- Non Graphics, Non Streaming portions of Virtual Immersion may or may not parallelize well, depends on the inherent parallelism involved. Even if there is, Latency is one of the opposing forces.
- System architects have a tradeoff to make allocating transistors for 2D for memory or cores. 3D opens new vistas.
- 3D chip stacking provides a way to have the best of three worlds, many cores and a lot of memory, with high bus width.
- Other forms of latency are important (bus structure, and IO)
- Heat spreading and thermal management will be important.