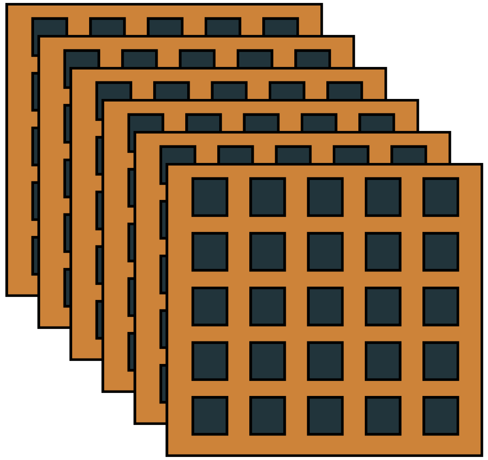
Multicore:
Why is it happening now?
eller
Hur Mår Moore's Lag?

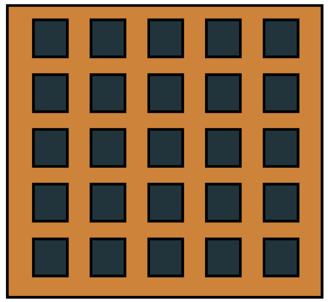
Erik Hagersten
Uppsala Universitet

Darling, I shrunk the computer

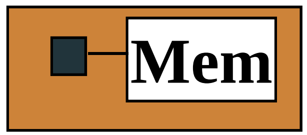
Mainframes



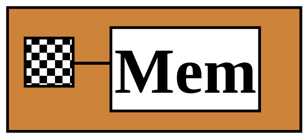
Super Minis:



Microprocessor:



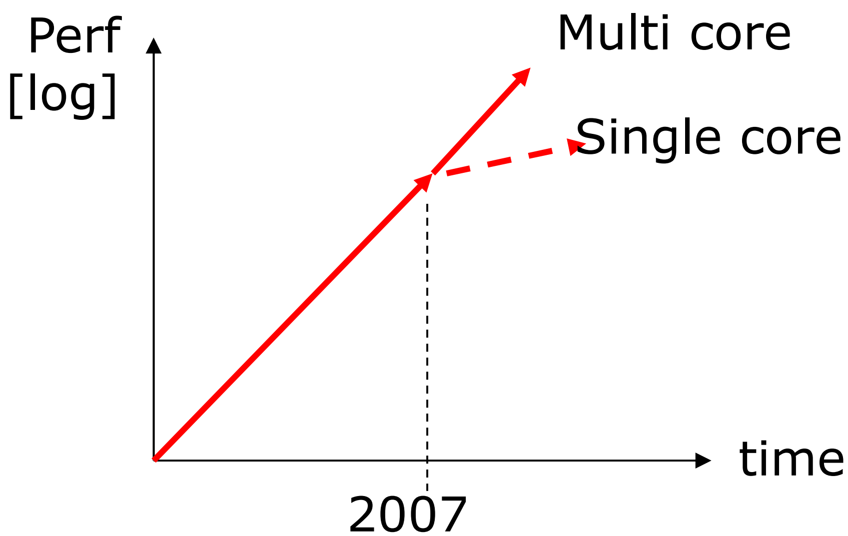
Multicore: Many CPUs on a chip!



Outline

- Why multicore now?
- Performance bottlenecks in MCs
- Commercial offerings
- Reflection for the future

Everybody is doing it!
But, why now?

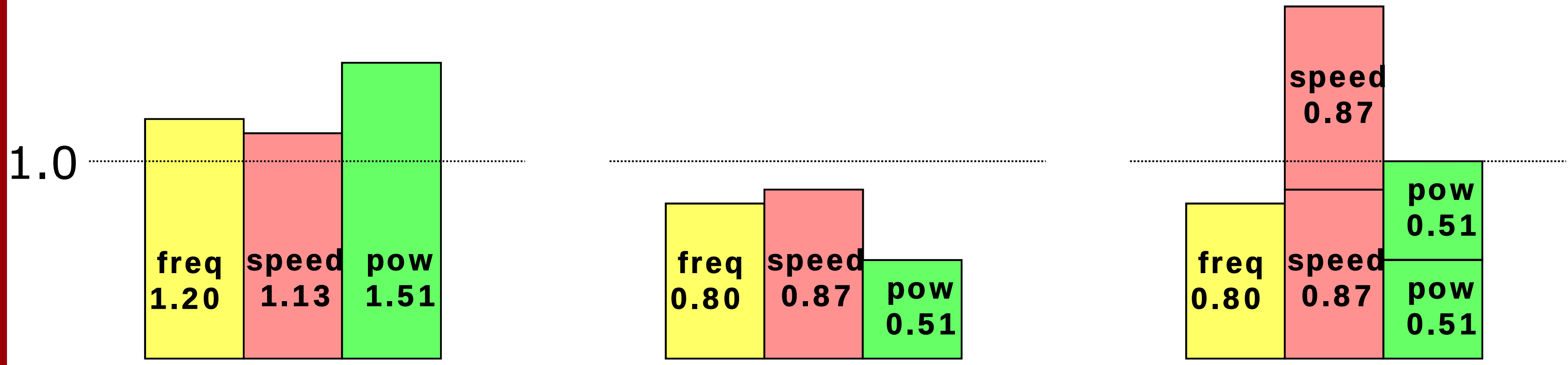


1. Not enough ILP to get payoff from using more transistors
2. Signal propagation delay » transistor delay
3. Power consumption $P_{dyn} \sim C \bullet f \bullet V^2$

UPPSALA UNIVERSITET

Example: Freq. Scaling 2006

$$P_{dyn} = C * f * V^2 \approx \text{area} * \text{freq} * \text{voltage}^2$$



20% higher freq. and higher voltage

20% lower freq. and lower voltage

20% lower freq. Two cores

DVFS: Dynamic Voltage/Frequency Scaling

Lower voltage and lower frequency

Today V does not vary as much with f

Dynamic energy is proportional to:

- ❑ Frequency ^ 2
- ❑ Voltage ^ 2
- ❑ Capacitance ^2

DVFS saves energy by:

- ❑ Lover voltage and higher frequency
- ❑ Lover voltage and lover frequency
- ❑ Higher voltage and higher frequency
- ❑ Higher voltage and lover frequency

UPPSALA UNIVERSITET

Dennard Scaling

$$P_{dyn} = C * f * V^2 \approx \text{area} * \text{freq.} * \text{voltage}^2$$

Roughly constant between generations

Can potentially increase with smaller transistors

Robert Dennard (IBM): Lower voltage when you shrink the transistor. This trend has stopped ☹

➔ Can not increase the frequency! Other options? TLP!

UPPSALA
UNIVERSITET

Darling, I shrunk the computer

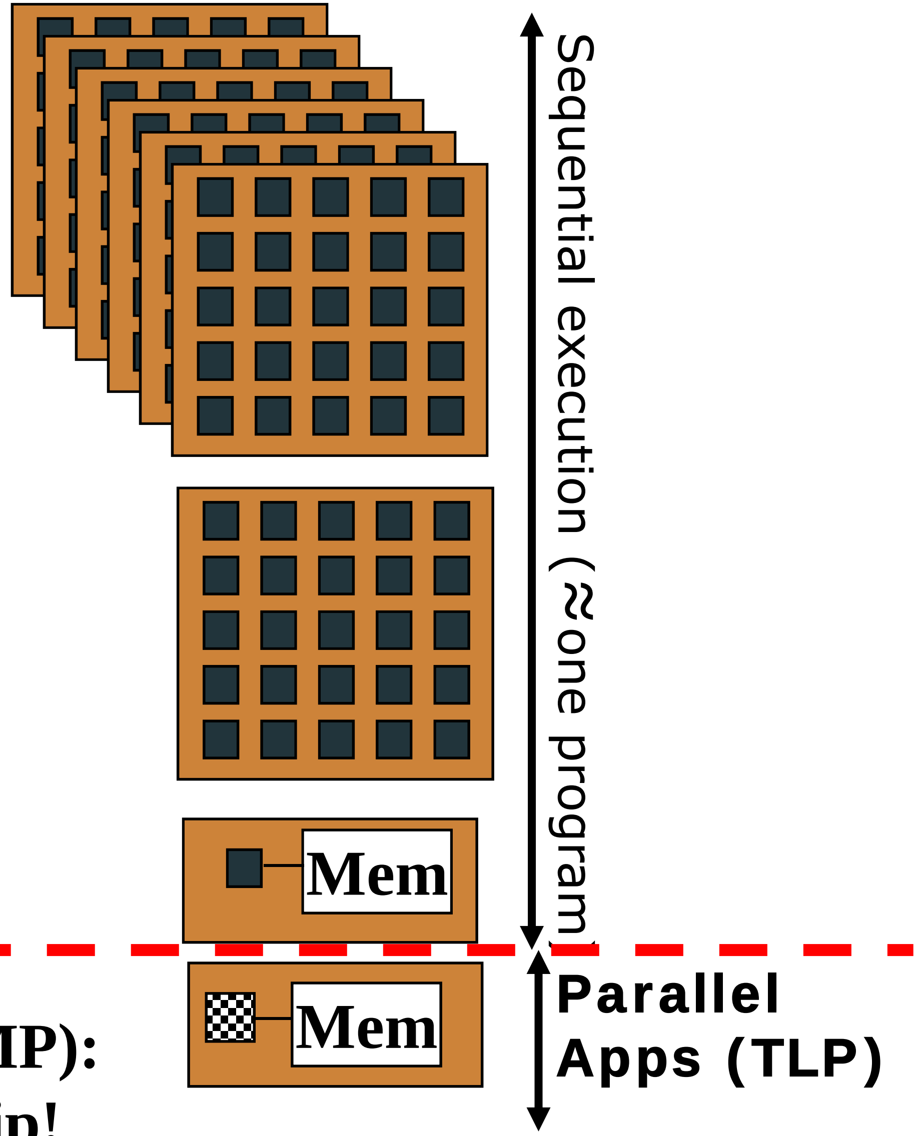
Mainframes

Super Minis:

Microprocessor:

Paradigm Shift

Chip Multiprocessor (CMP):
A multiprocessor on a chip!

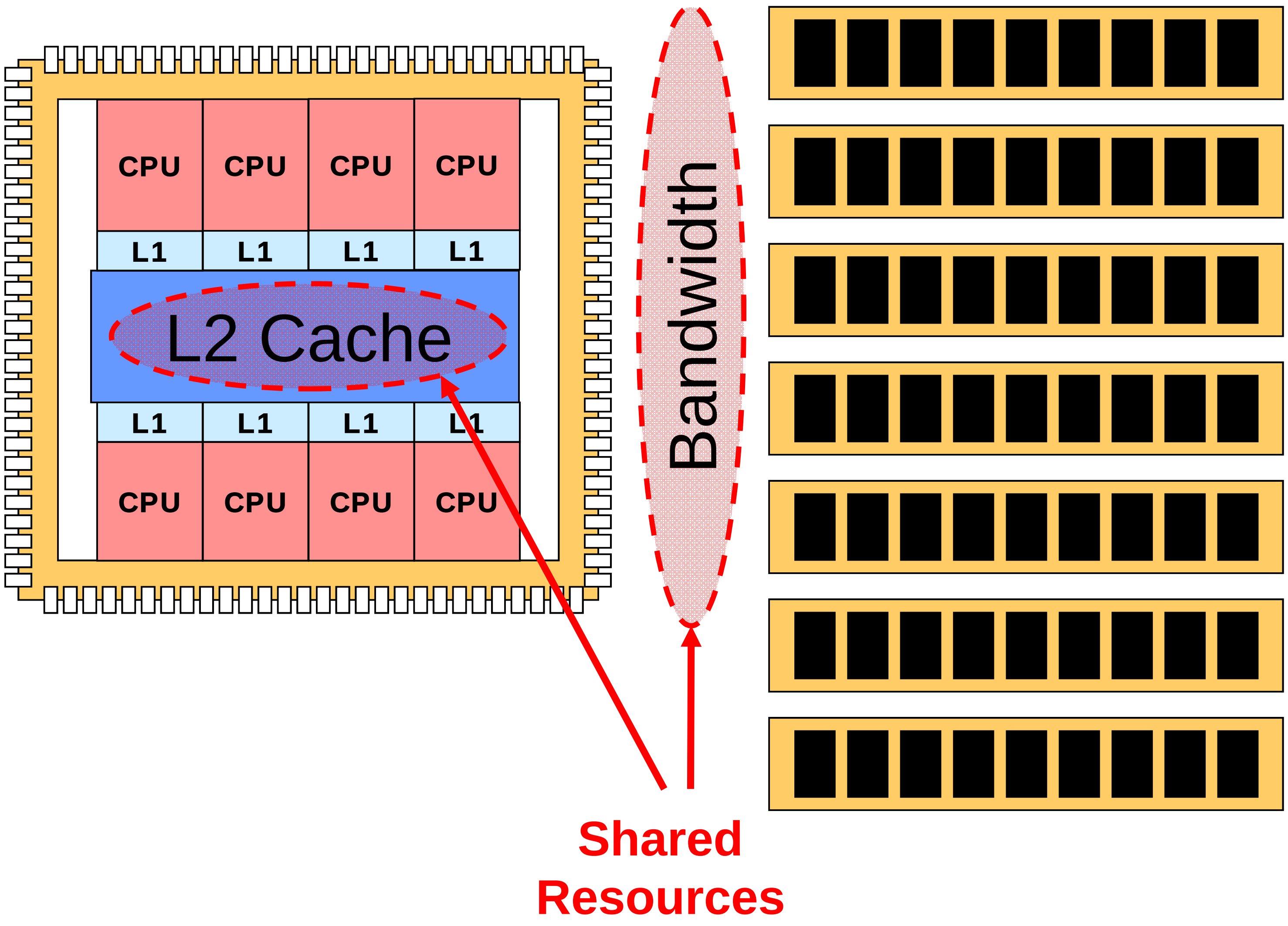


MC Bottlenecks

Erik Hagersten
Uppsala University

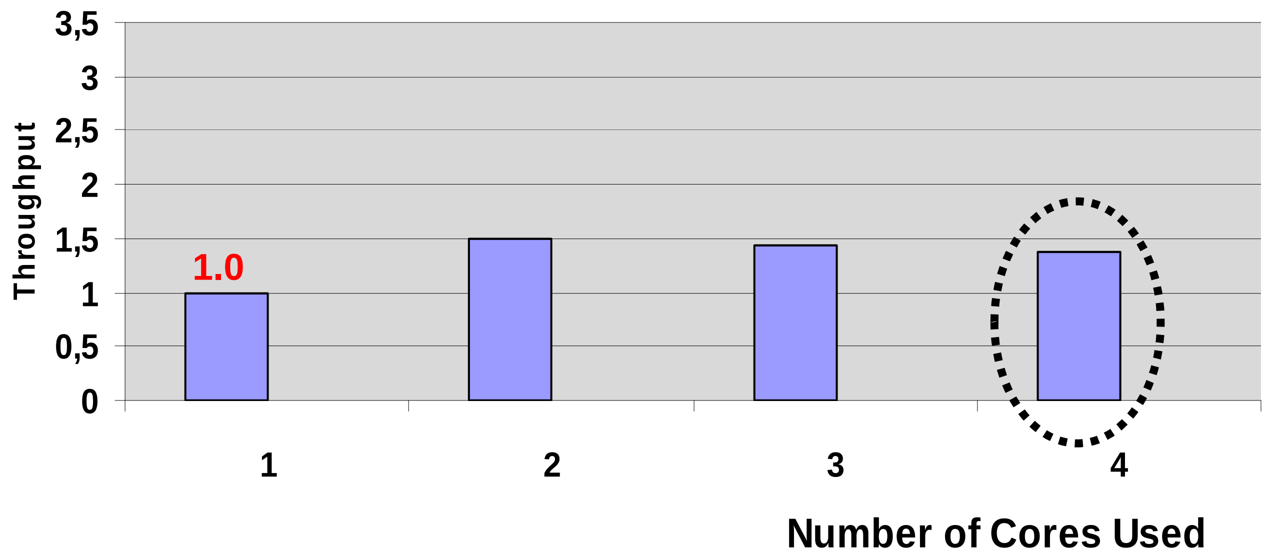
Shared Bottlenecks

(the MCs on these slides are generalized)



Example: Poor Throughput Scaling!

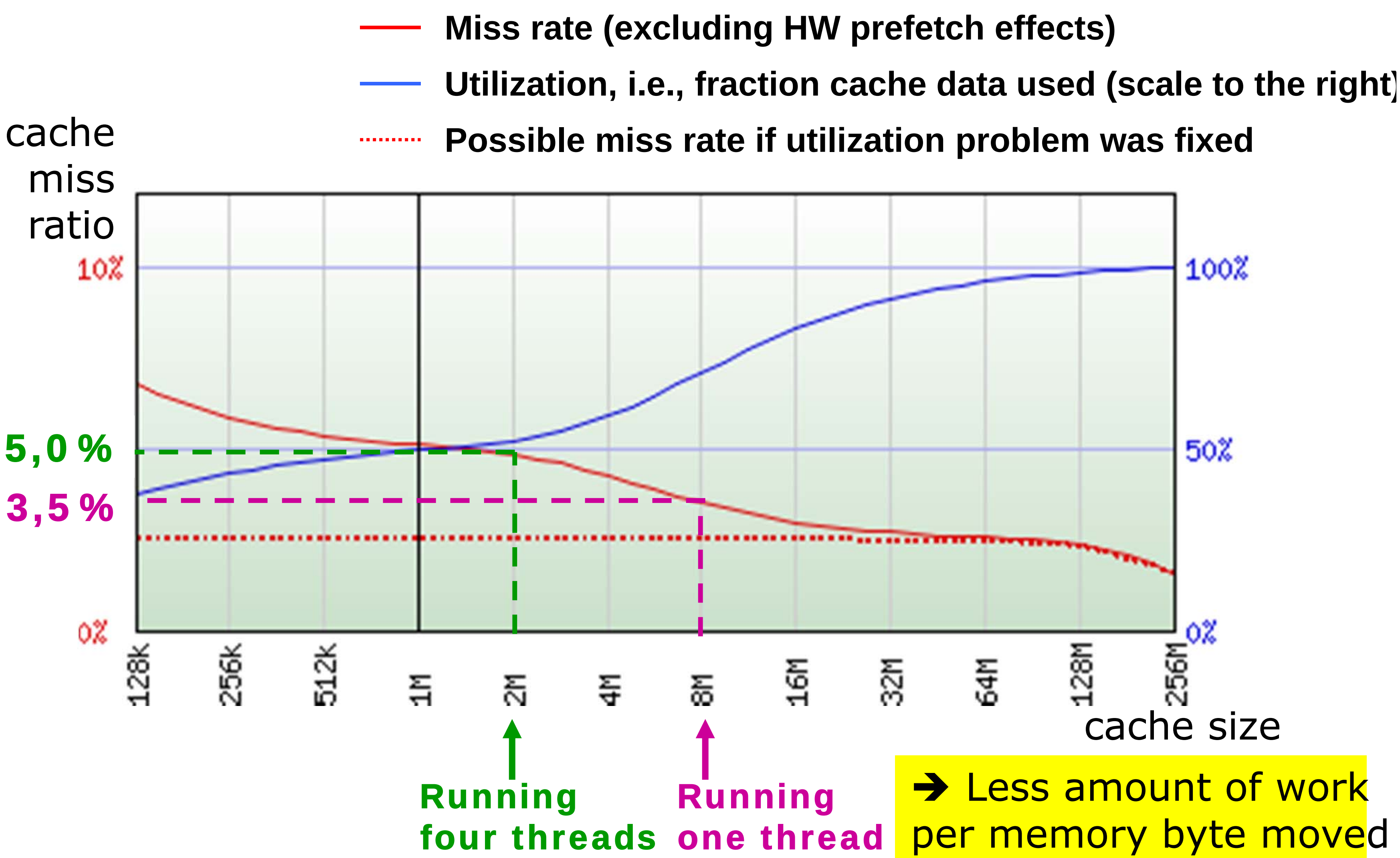
Example: 470.LBM
"Lattice Boltzmann Method" to simulate incompressible fluids in 3D



Throughput (as defined by SPEC):
Amount of work performed per time unit when <u>several instances</u> of the application is executed simultaneously.
<u>Our TP study</u>: compare TP improvement when you go from 1 core to 4 cores

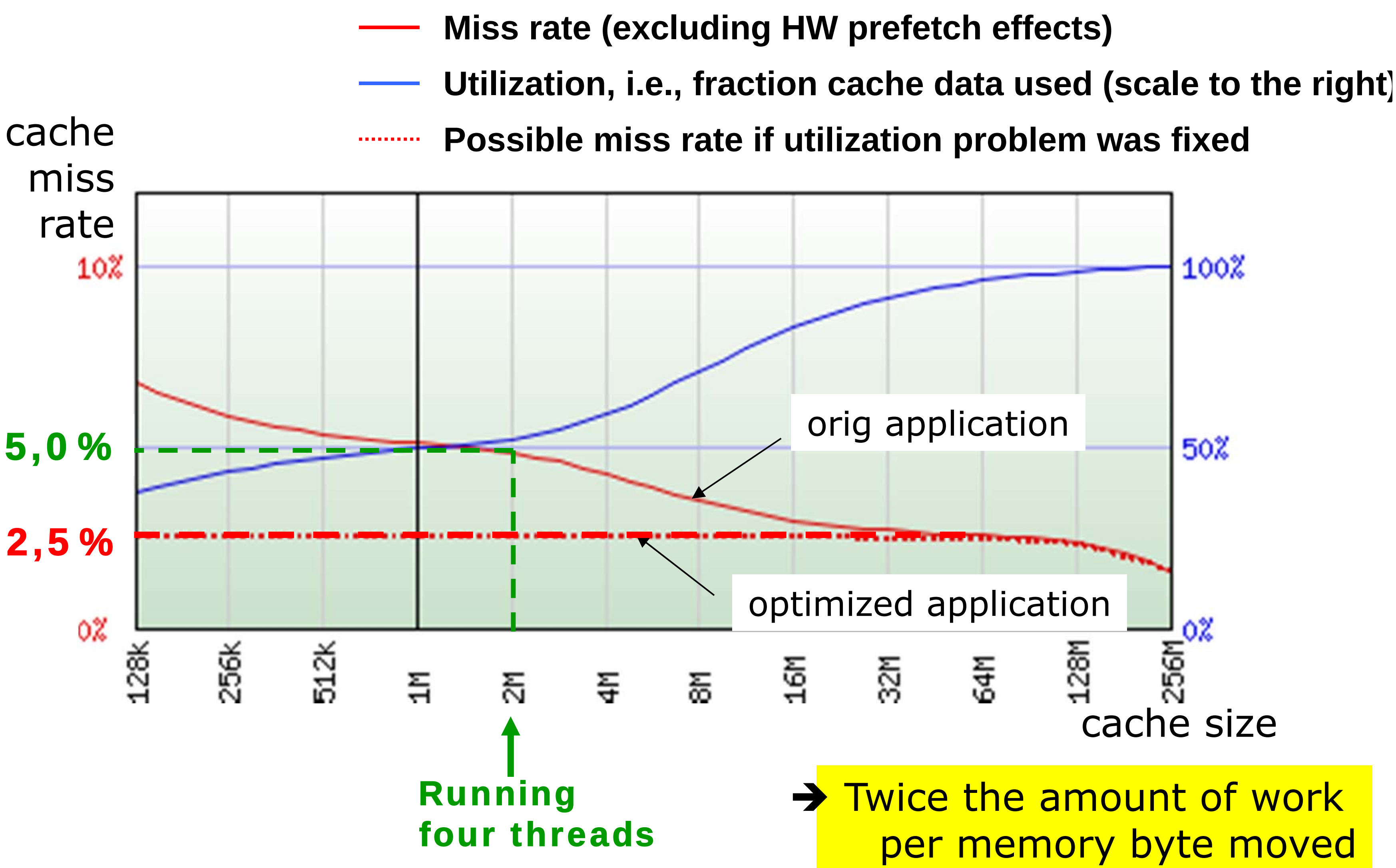
UPPSALA UNIVERSITET

Nerd Curve: 470.LBM



➔ Less amount of work per memory byte moved @ four threads

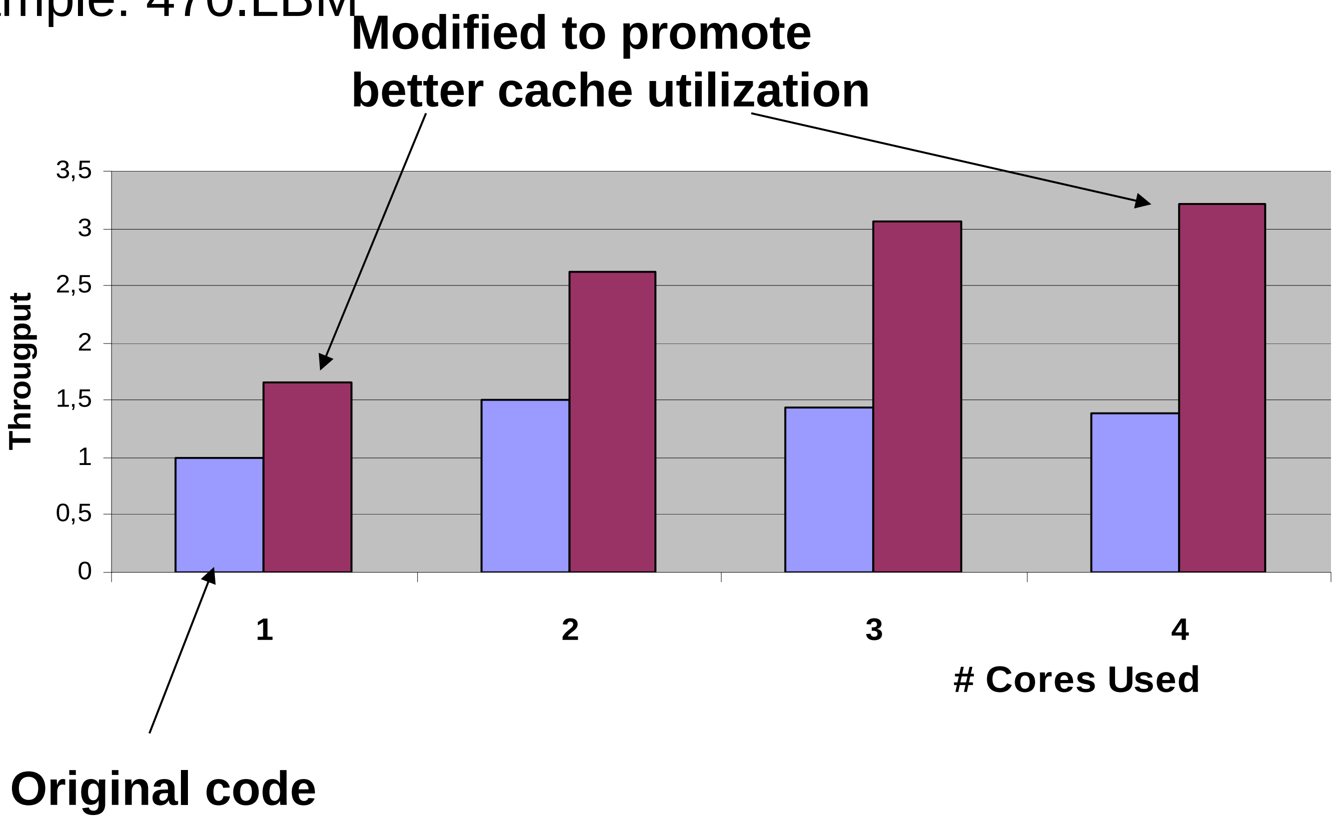
Nerd Curve (again)



➔ Twice the amount of work per memory byte moved

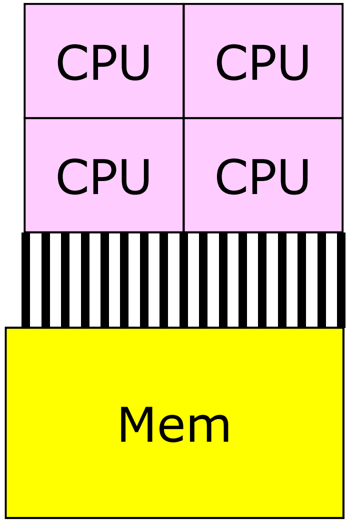
➔ Better Memory Usage!

Example: 470.LBM

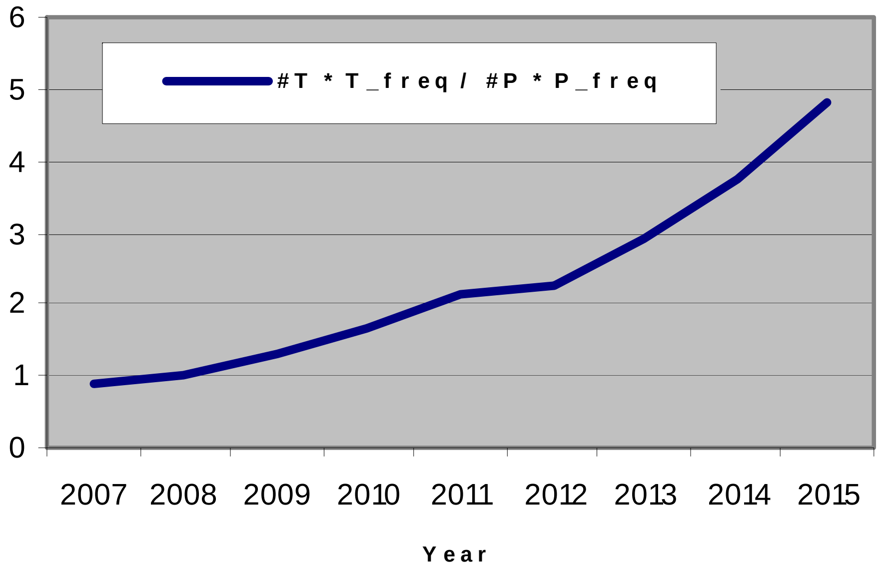


BW in the Future?

#Cores ~ #Transistors



Computation vs Bandwidth



Semiconductors (ITRS)

ultiprocessors

HPCWire.com 2007
Up Against the Memory Wall
”Nevermind the cores. Just hand over the cache”

HPCWire 2007:
More Than 16 Cores May Well Be Pointless
[by Sandia Labs]

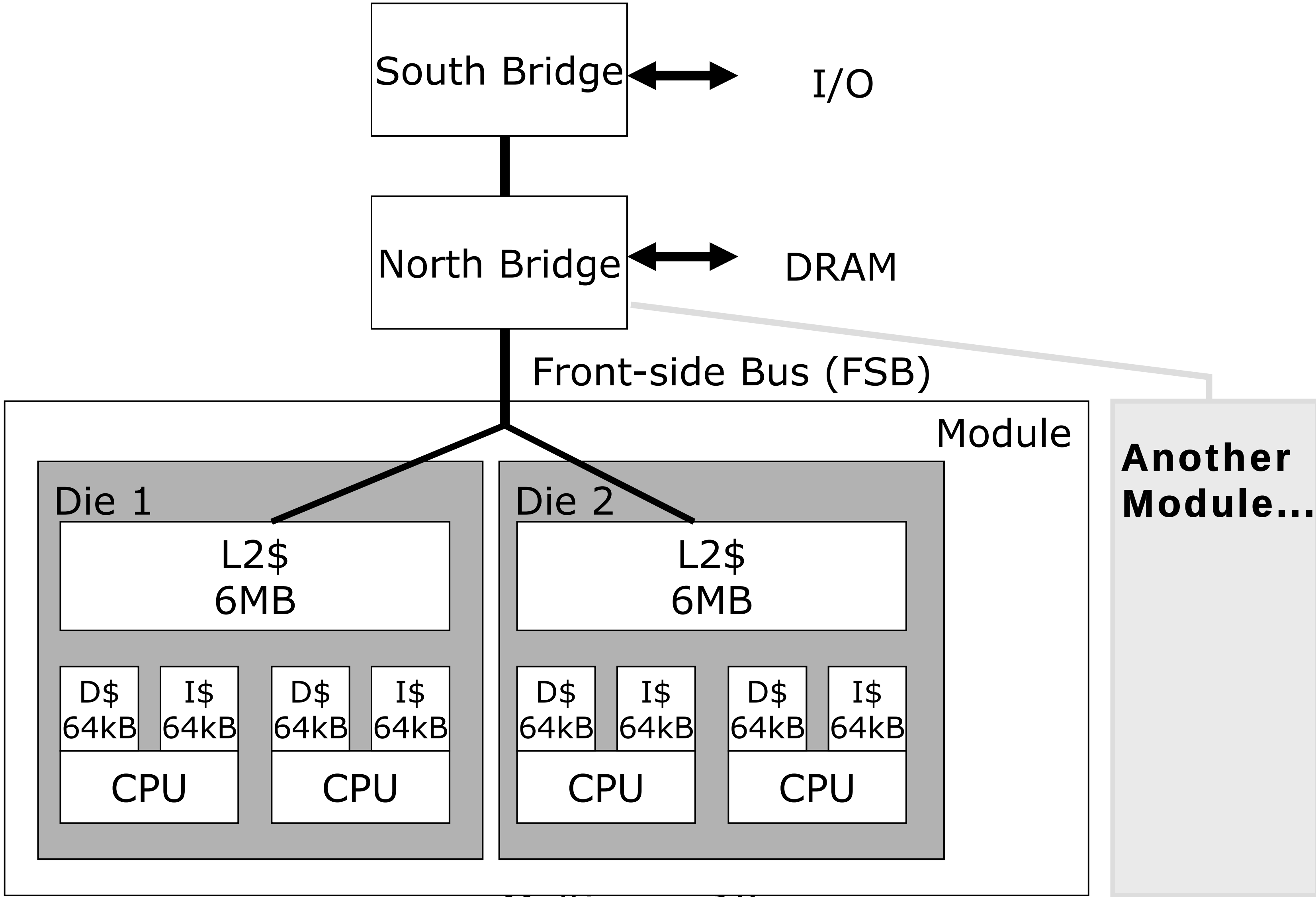
Multicore has the following bottlenecks shared between cores

- ❑ Memory bandwidth
- ❑ Register files
- ❑ Cache capacity
- ❑ CPU frequency

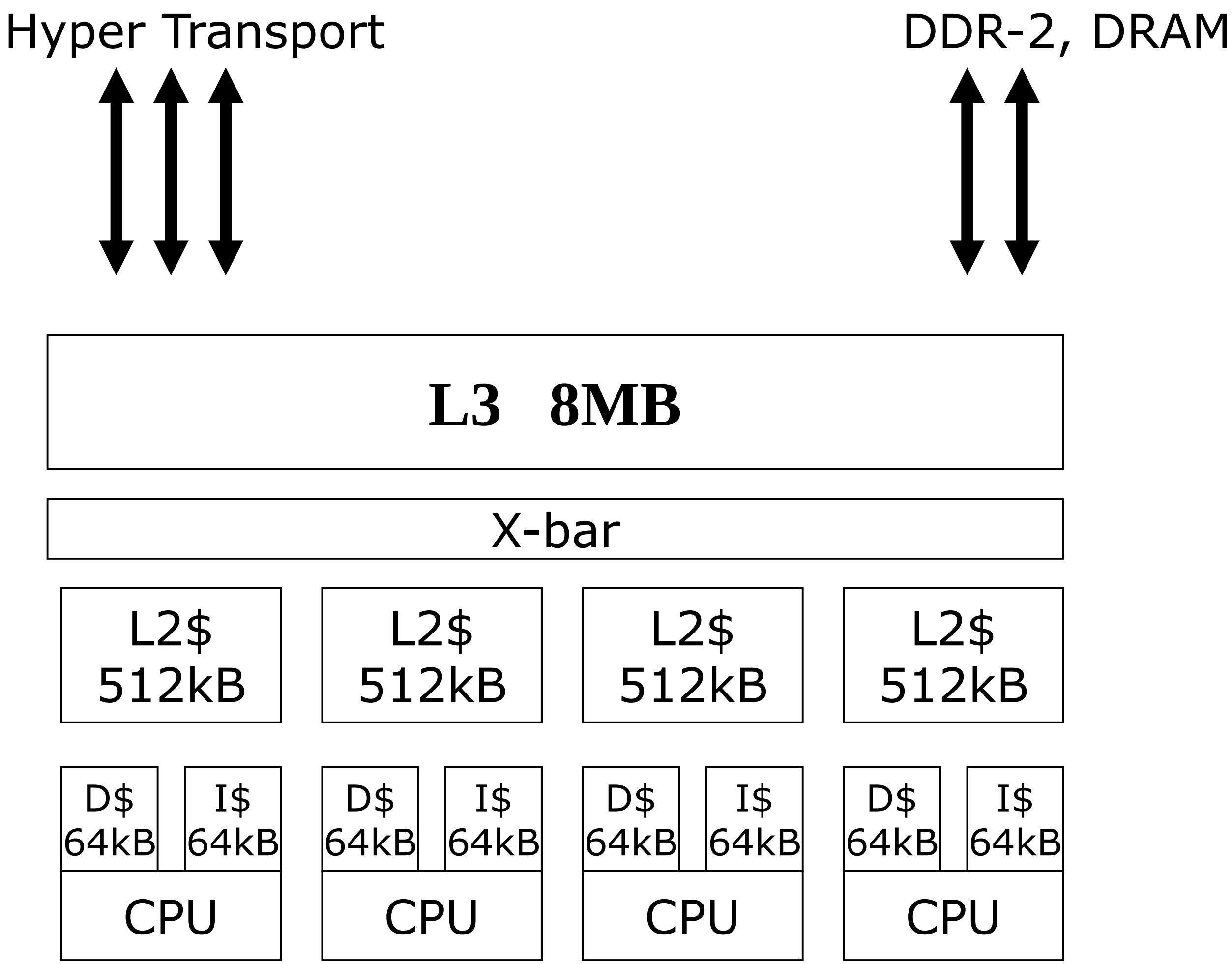
Commercial snapshot
eller
“Kärnornas krig”

Erik Hagersten
Uppsala Universitet

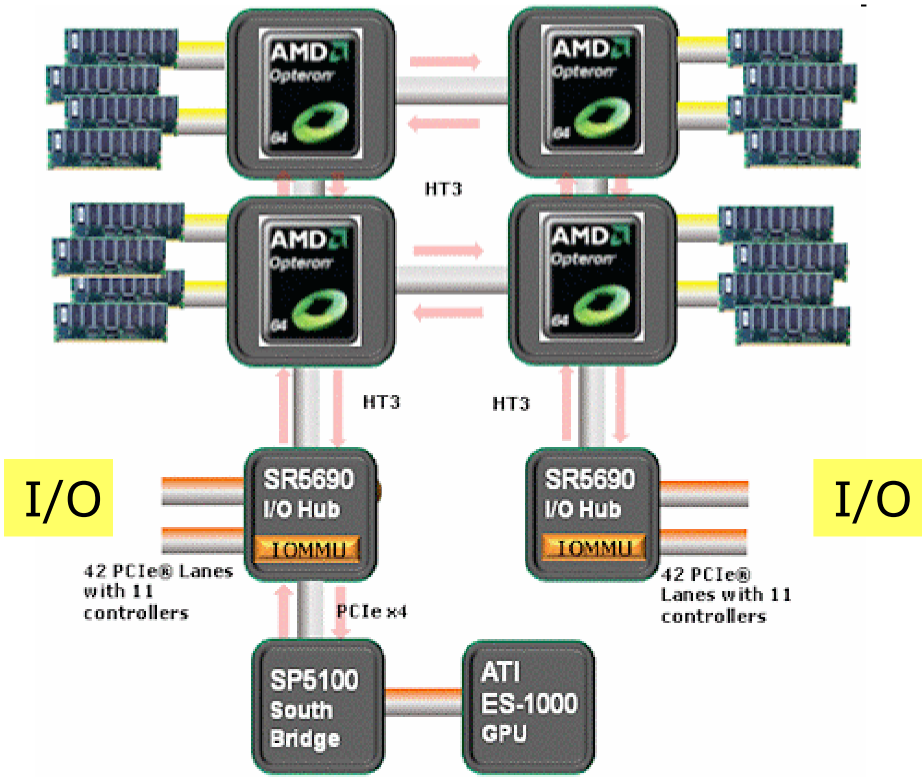
Intel Core2 Quad, 2006



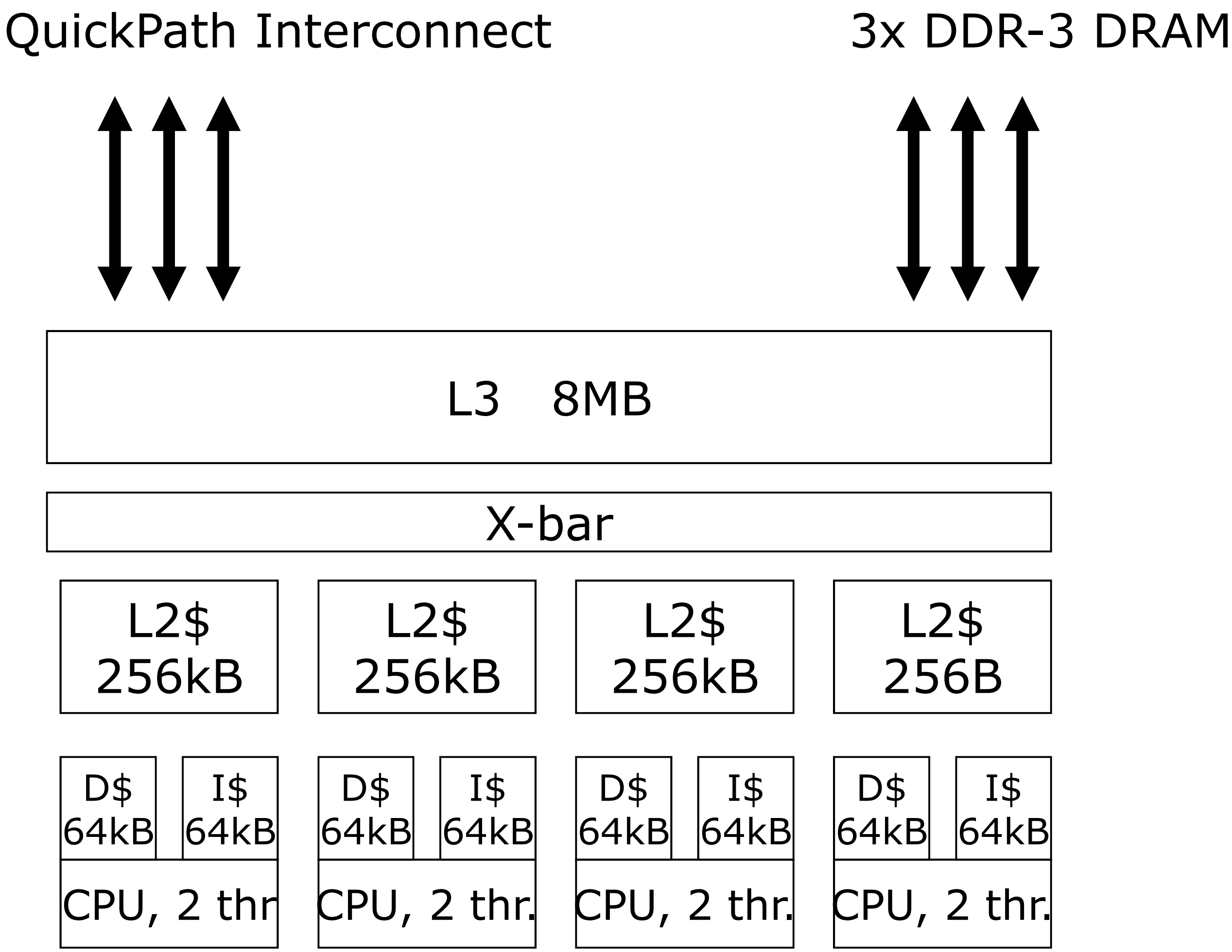
AMD Shanghai, 2007



AMD MC System Architecture

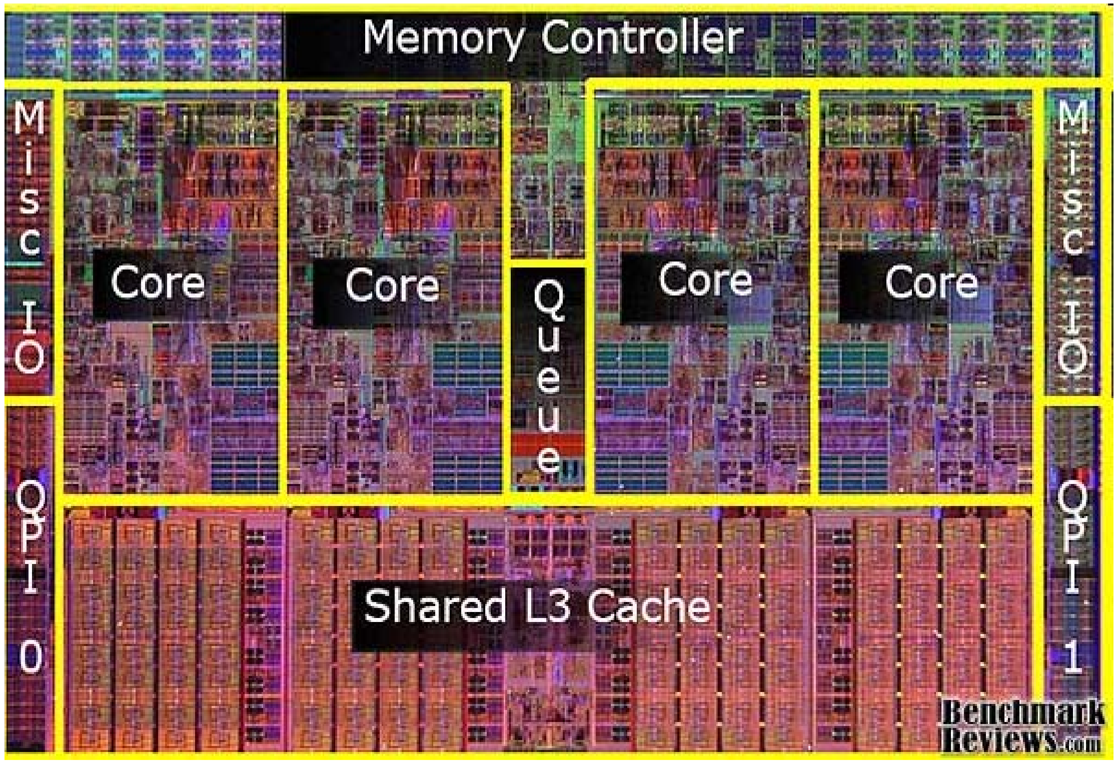


Intel: Nehalem, Core i7
Q1 2009 (4 cores)

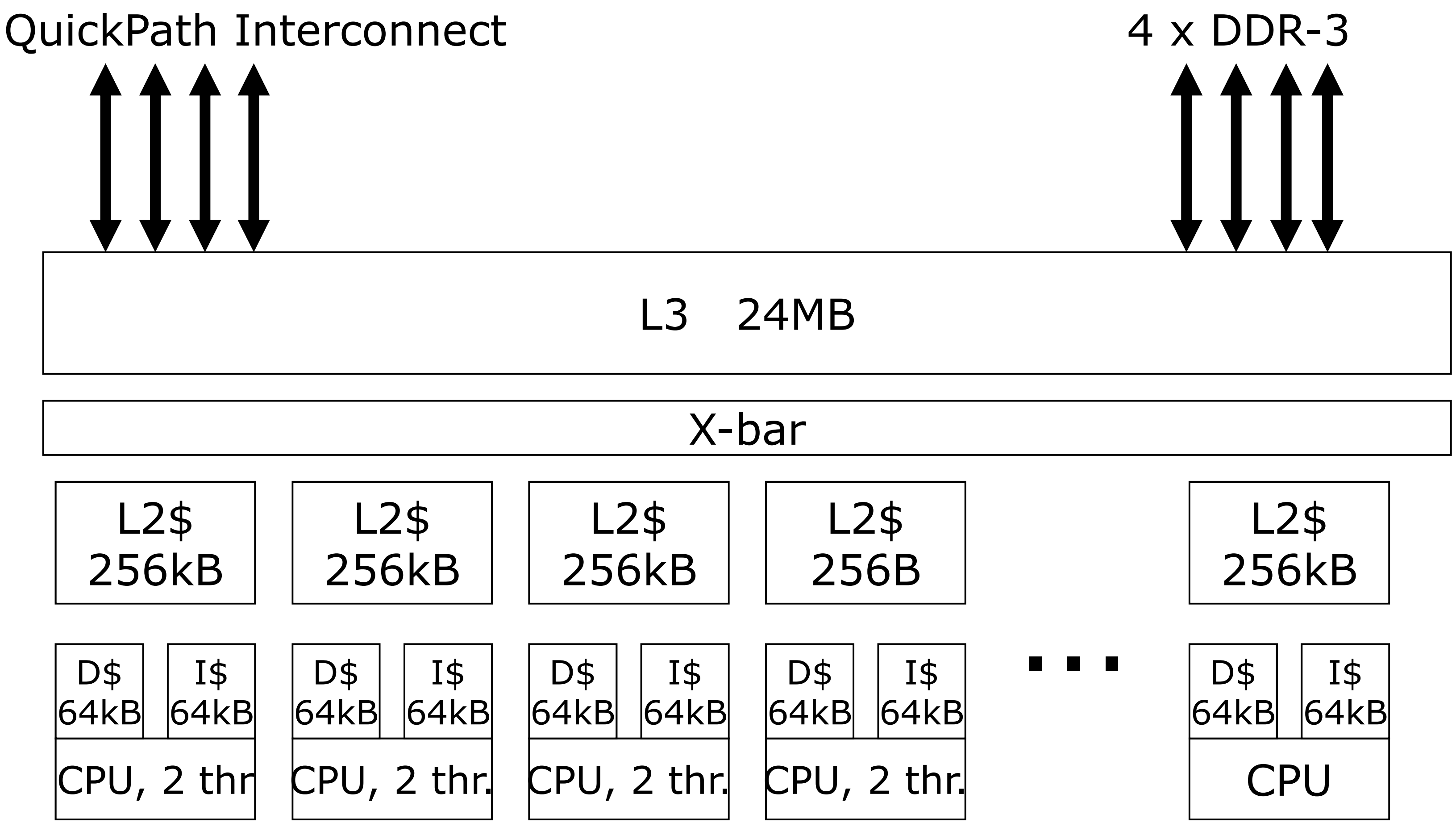


Up to 4 cores x 2 threads

Nehalem "Core i7"



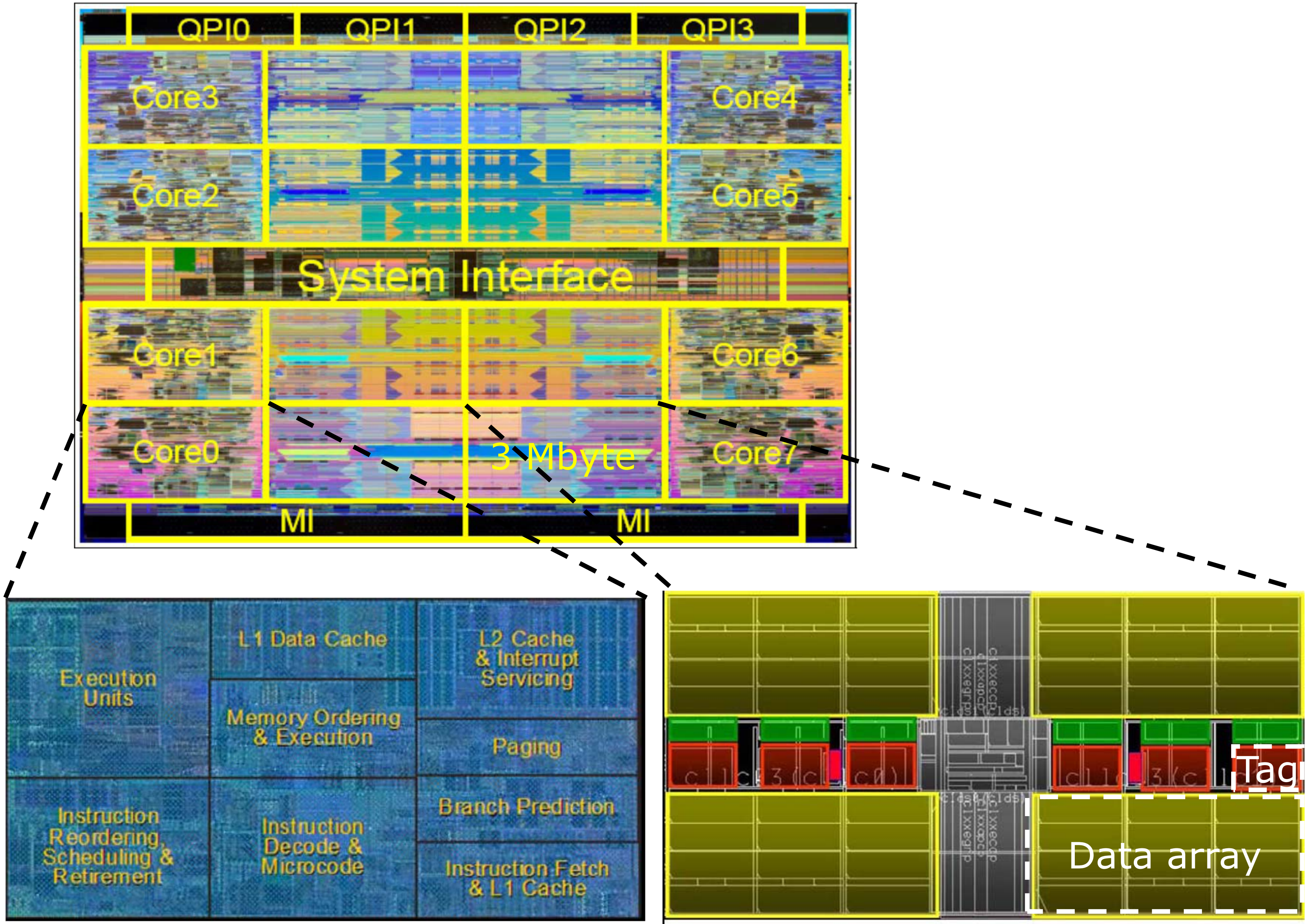
Intel: ”Nehalem-Ex” (i7)



8 cores x 2 threads

UPPSALA UNIVERSITET

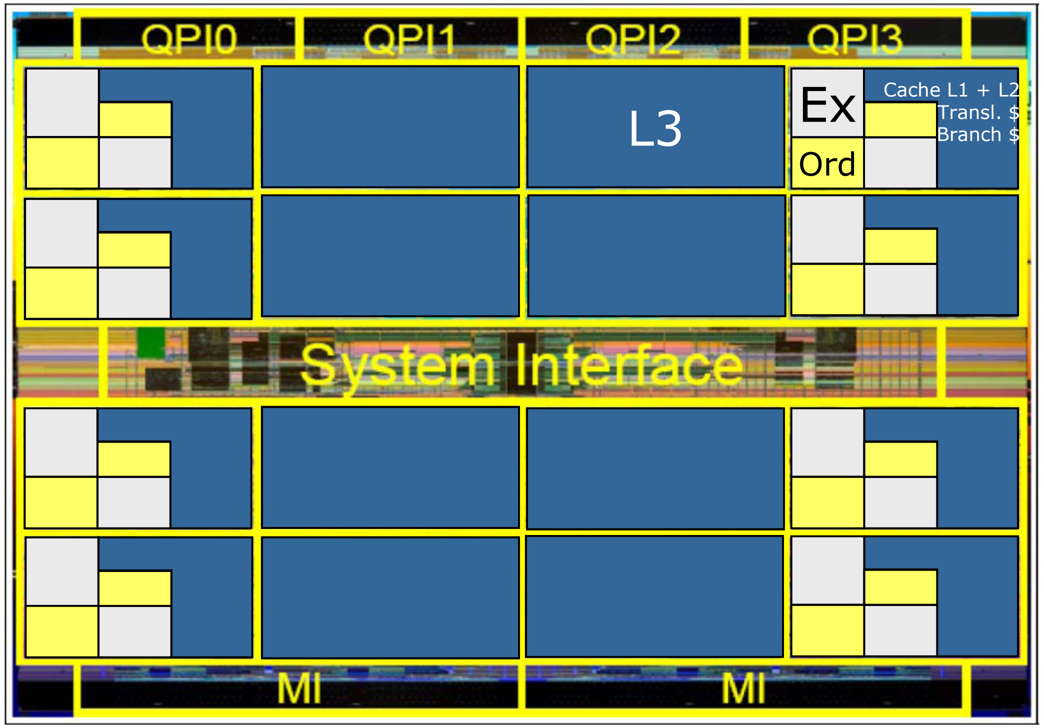
How is the silicon used (i7-Ex)?

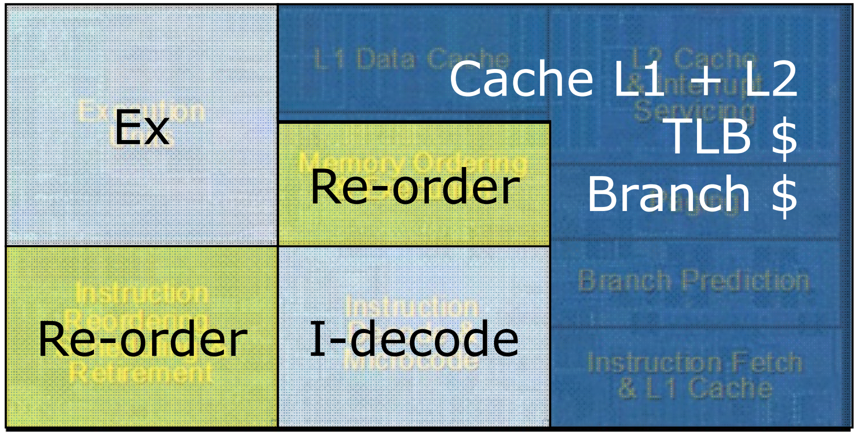


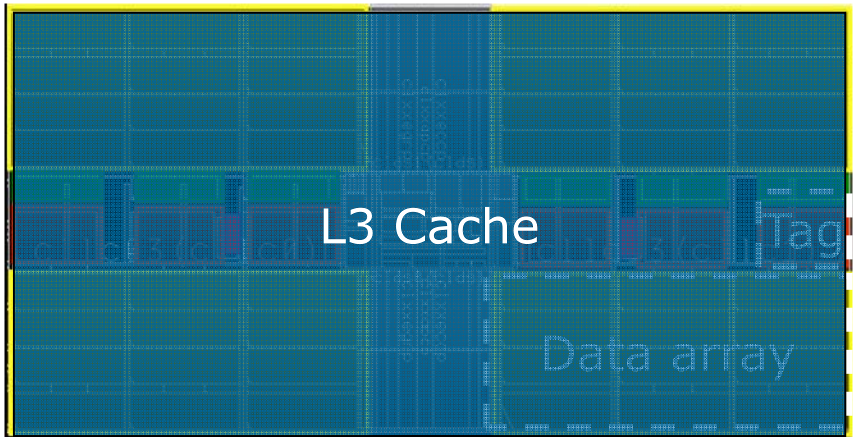
Source: JSSC Jan 2010, Rusu et. al

UPPSALA UNIVERSITET

How is the silicon used?

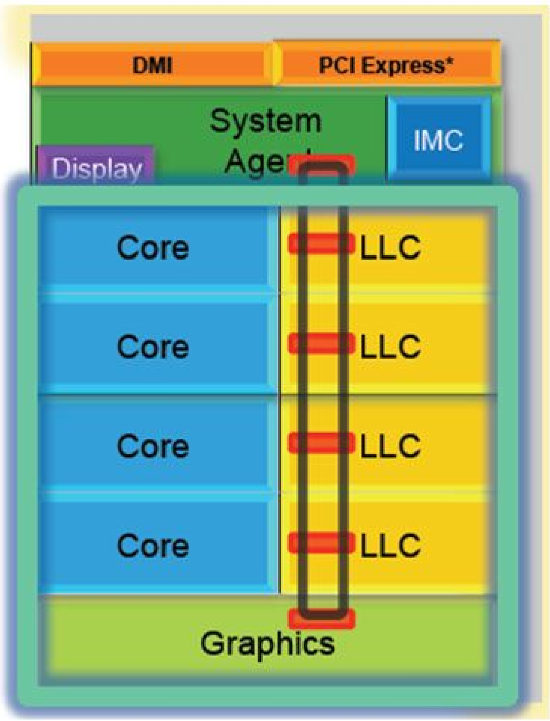




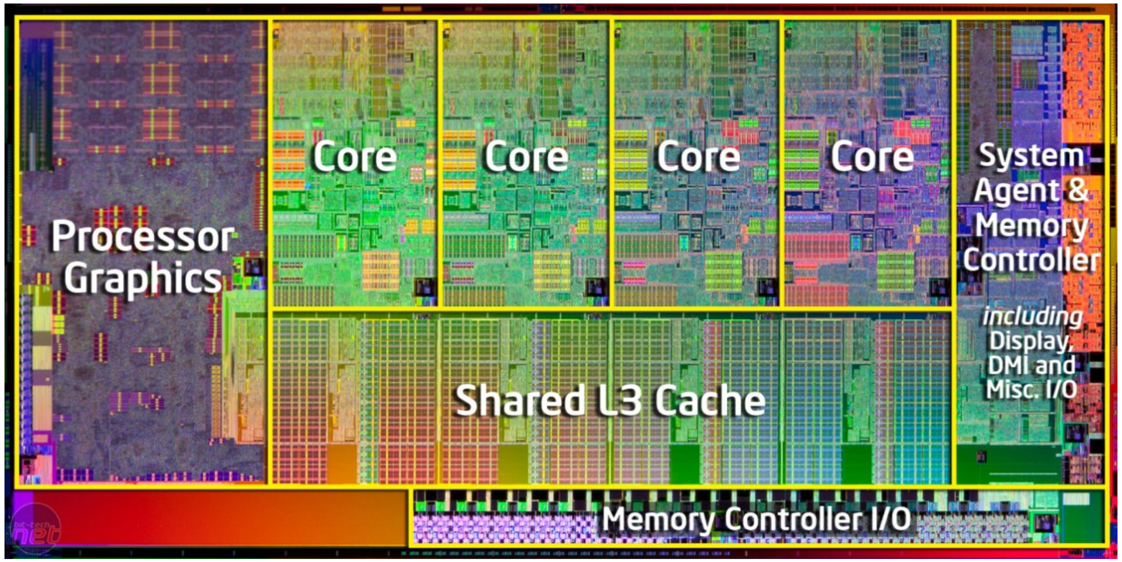


Source: JSSC Jan 2010, Rusu et. al

More recent:
Sandy Bridge, 32 nm & Ivy Bridge, 22nm



- Integrated graphics!
- Ring bus for coherent LLC
- Up to 12 cores (24 threads)
- ~3GHz-ish



UPPSALA UNIVERSITET

General Purpose: Is there a demand for more cores?

- Intel released their quad core at Supercomputing 2006.
- Six years later we are only(!!) at 12 cores and 24 threads.
- Until last year almost nobody har 4-core laptops
- Core complexity and caches seem to be favoured over number of cores

Yes! Certain market segments and special applications need cores...

Speicalizing for Throughput/Commercial Computing

Erik Hagersten

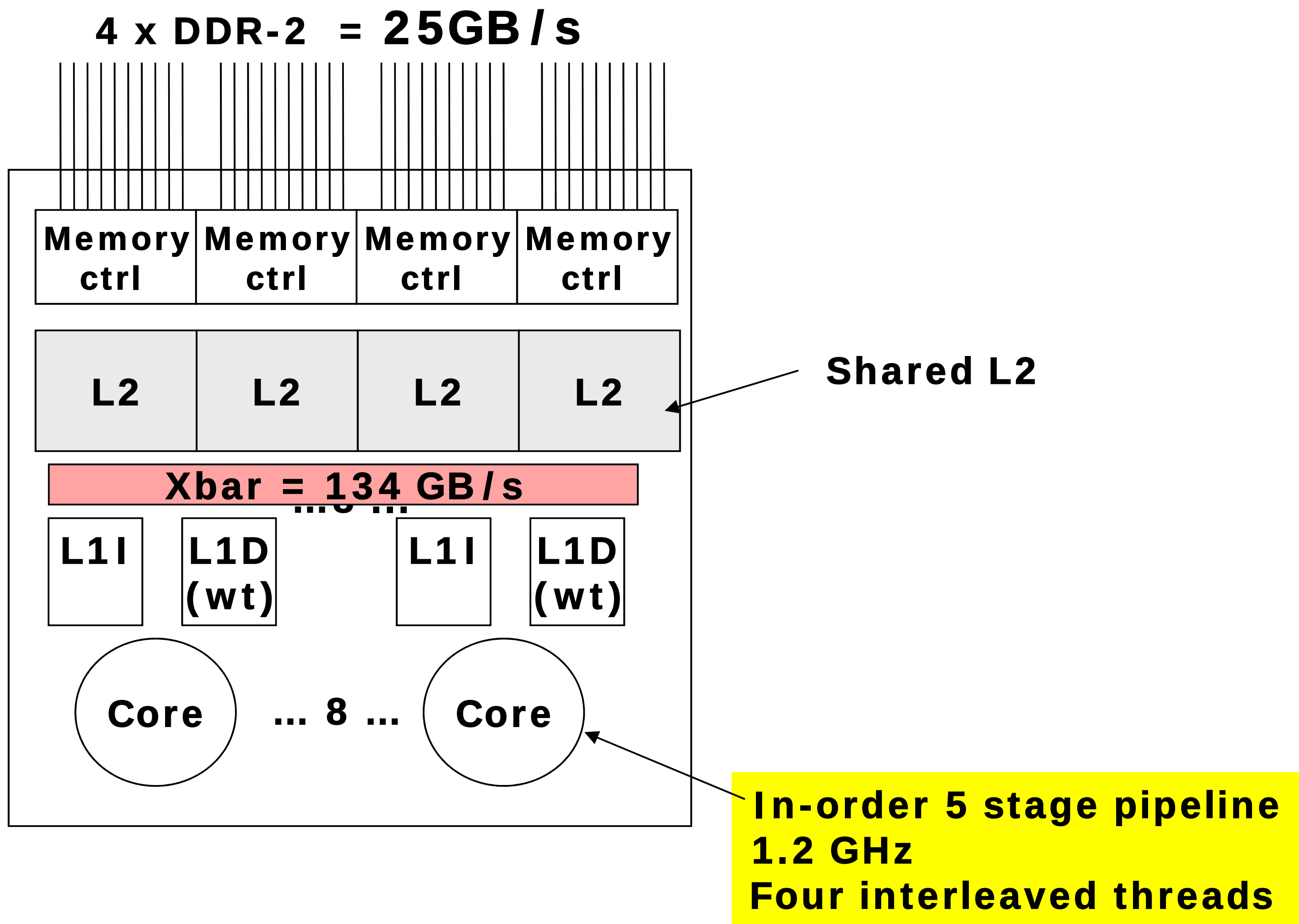
Uppsala University

Hot Chips Conference 2012

” 16-core SPARC T5 CMT Processor with glueless 1-hop scaling to 8-sockets”

” Knights Corner, Intel’s first Many Integrated Core (MIC) Architecture Product”

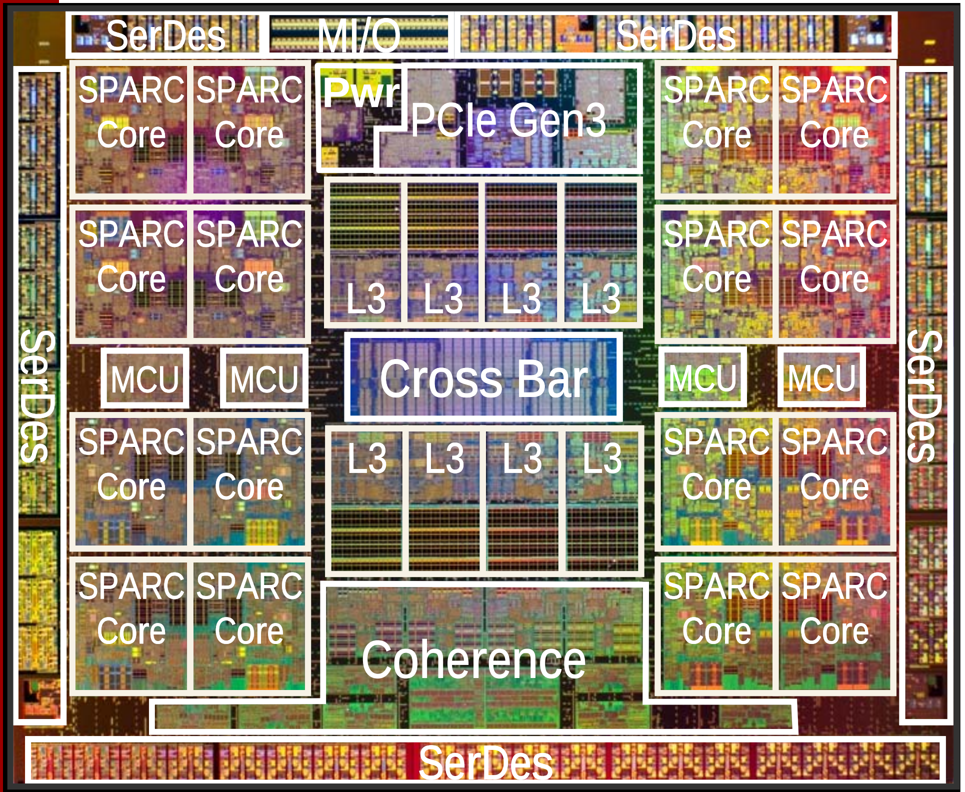
Sun Niagara T1, 2005
Targeting transactions and DB



2012: SPARC Niagra T5

- 16 cores
- 8 threads per core (128 threads/chip)
- 2x superscalar out-of-order @3.6GHz
- 8MB L3 cache (64kB/thread)
- Glueless 8 multisockets (1-hop)
- Coherent global NUMA memory
- Lots of HW accelerations

T5 Processor Overview



- 16 S3 cores @ 3.6GHz
- 8MB shared L3 Cache
- 8 DDR3 BL8 Schedulers providing 80 GB/s BW
- 8-way 1-hop glue-less scalability
- Integrated 2x8 PCIe Gen 3
- Advanced Power Management with DVFS

Oracle marketing slide: Special purpose HW

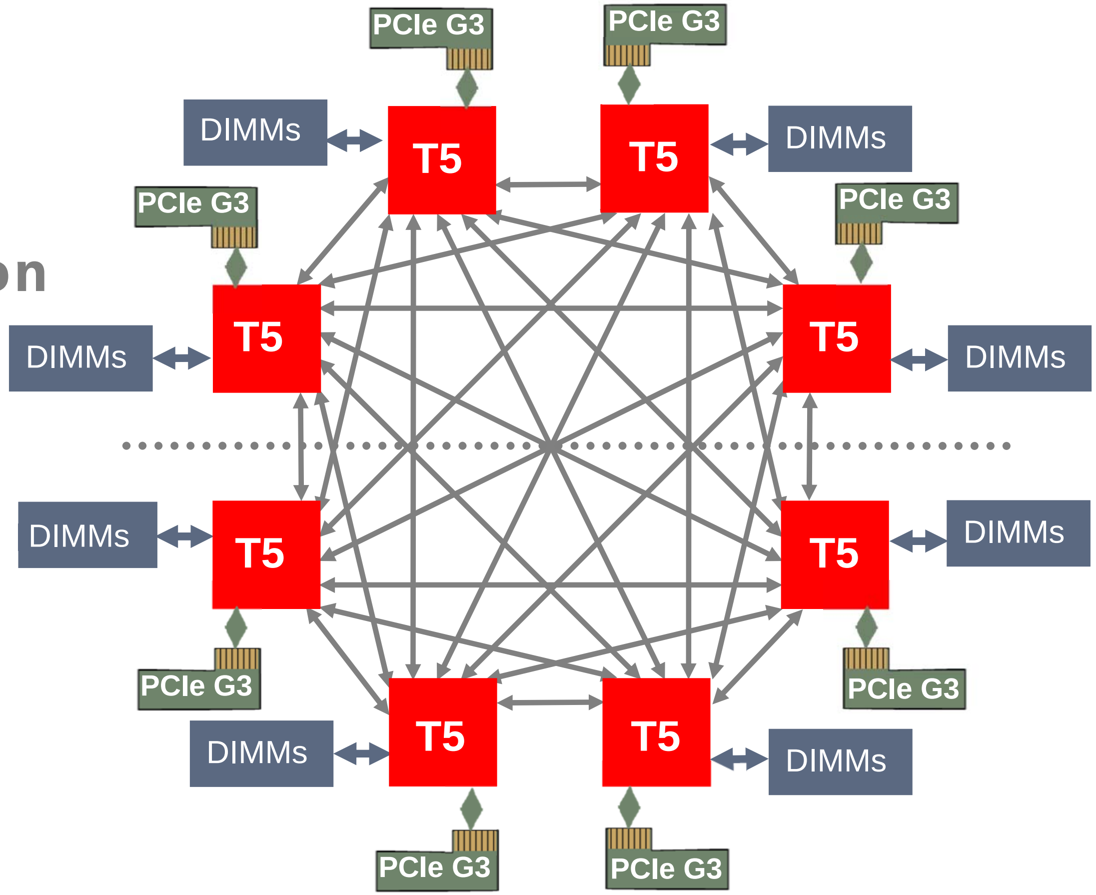
On-Chip Accelerators	SPARC T5	IBM Power7, HP	Intel Westmere/ Sandybridge
Asymmetric /Public Key Encryption	RSA, DH, DSA, ECC	none	RSA, ECC
Symmetric Key / Bulk Encryption	AES, DES, 3DES, Camellia, Kasumi	none	AES
Message Digest / Hash Functions	CRC32c, MD5, Sha-1, SHA-224, SHA-256, SHA-384, SHA-512	none	none
Random Number Generation	Supported	none	Supported

Oracle marketing slide: Speeds and feeds, 8 sockets

DDR3-1066
1+TB/sec

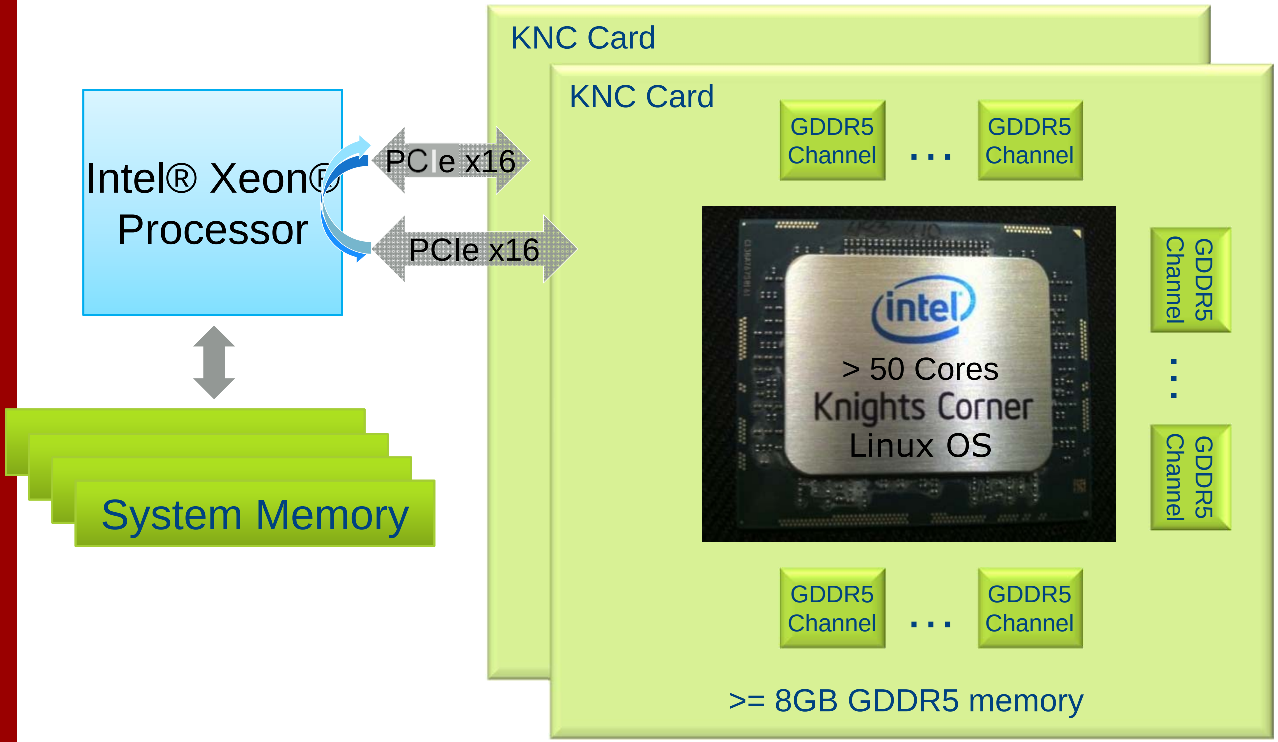
Coherency Bisection Bandwidth 840 GB/sec

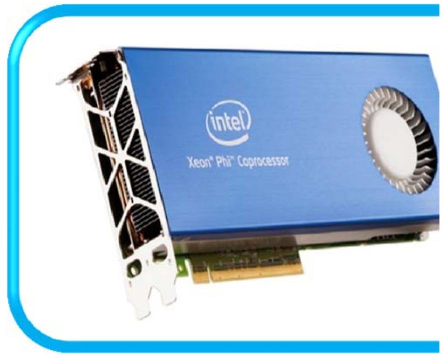
PCI Gen3 Bandwidth 256 GB/sec



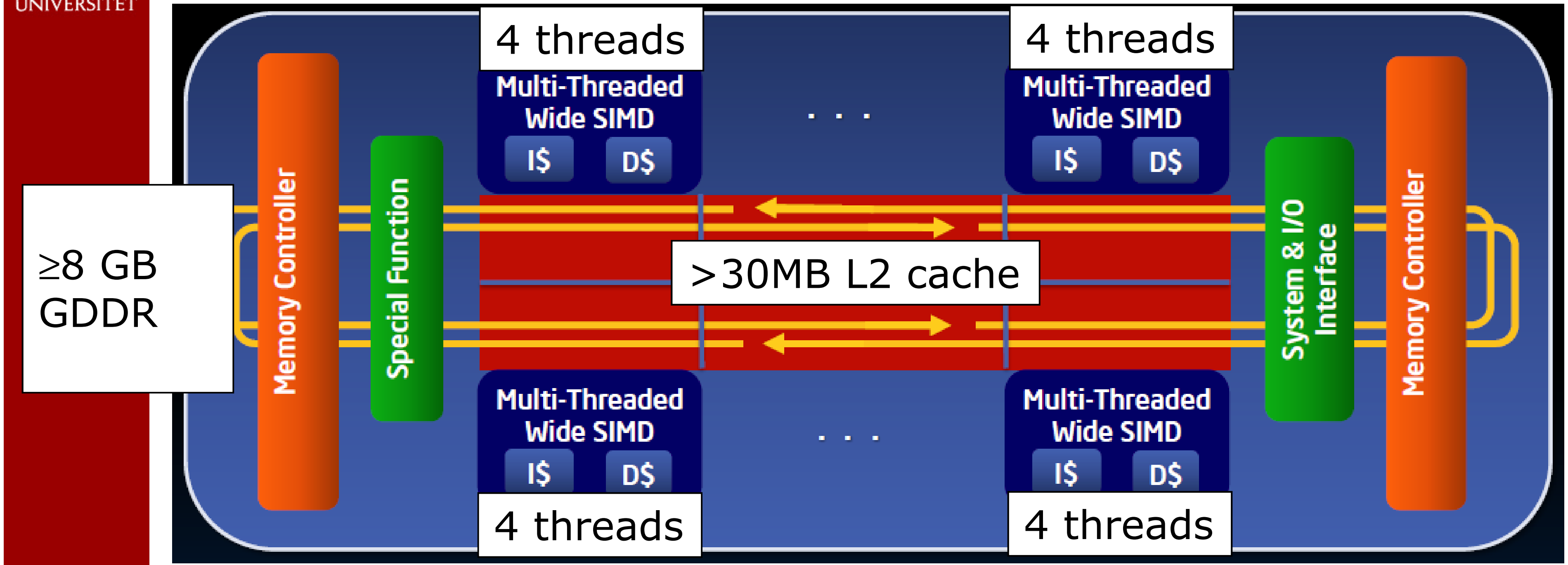
Non OoO Multicores

MIC = Multiple Integrated Cores
Xeon Phi (Knights Corner)
first commercial MIC implementation



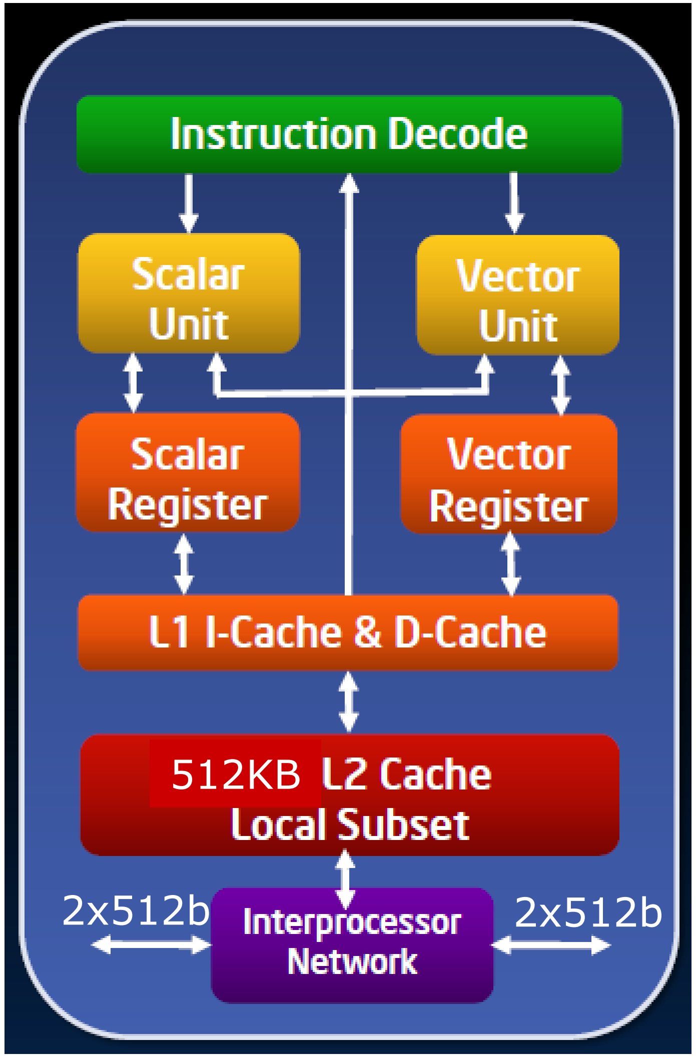


Intel's MIC HPC architecture



- Knights Corner (Intel Xeon Phi), 22nm
- >60 cores, X86 instructions ("enhanced"), 4 Threads
- Coherent caches. Runs Linux. "Just recompile and run"
- > 1TFLOPS (DP).
- 2% of the transistors devoted to x86 instructions

The MIC core ”tile”



- 2x superscalar in-order
- 4 Threads SMT/core
- 512b SIMD instr. (16x SP)
- Scatter/gather operations
- 32k L1 I and D cache
- 512kB coherent L2
- 16 HW prefetch streams
- 2x512bit ringbus
- …

UPPSALA UNIVERSITET

MIC Tool Stack

- It is ”just an x86 Linux box”
- Use your favorit language
- Intel compiler, tuning tool etc.
- 3rd party tools (for example mine:-)
- A huge alpha program since a year back
- Already on the Top500 list (#150)
- Low power/FLOP
- Many strategic sales already
- Question: can we afford 200+ threads/chip

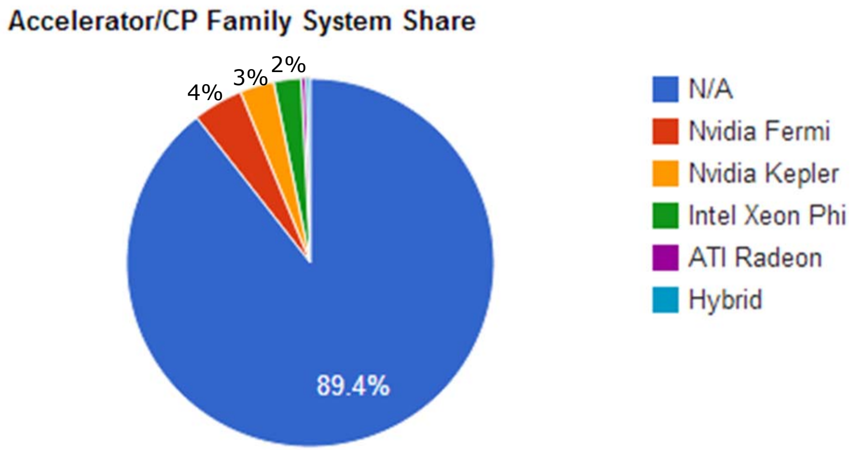
MIC / Xeon Phi

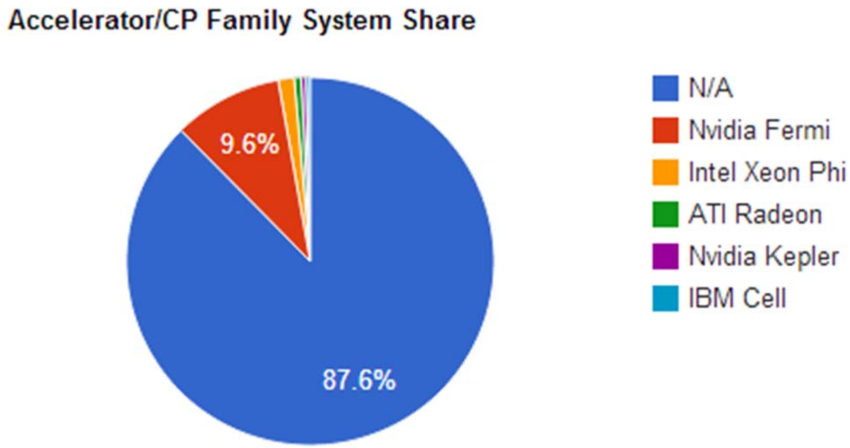
- Intel: It is ”just an x86 Linux box”
- Intel: Use your favorit language. Just re-compile and run.

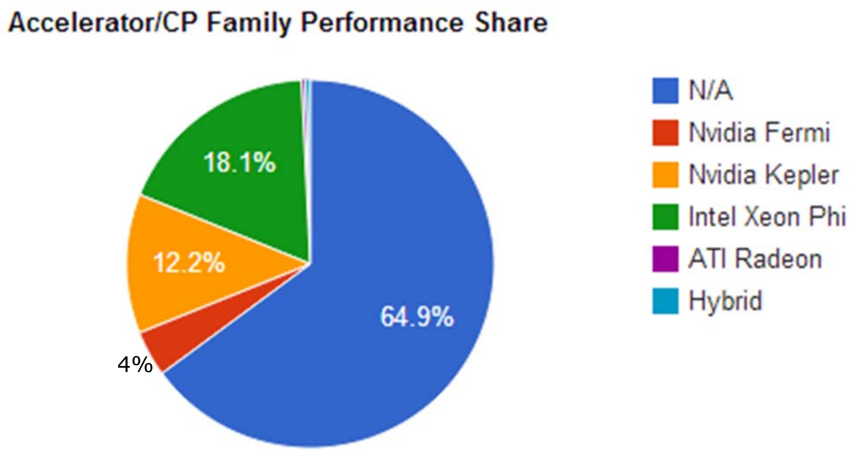
- Intel: compiler, tuning tool etc.
- Impressive on the Top500 list (e.g., #1, 15% of FLOPS)

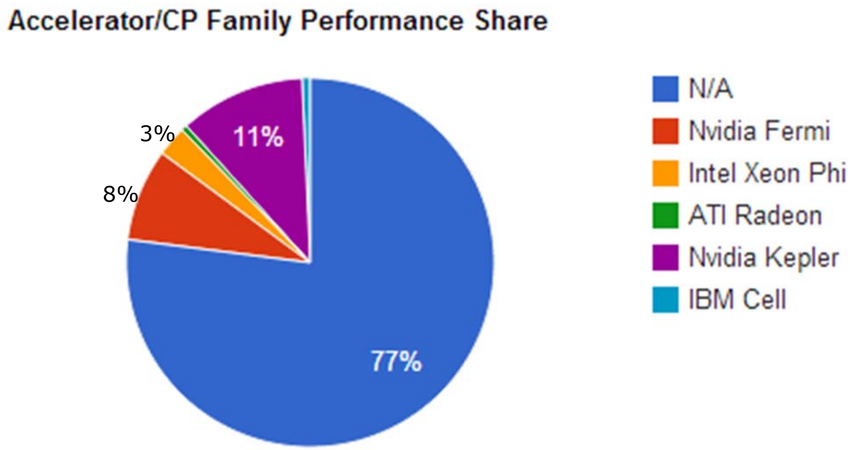
- Q: Can we spend 200+ threads to make one chip happy?
- Q: How much more than just recompilation is needed?
- Q: Lower performance but better productivity compared with GPUs. What will the market say about that?

Accelerators in Top500





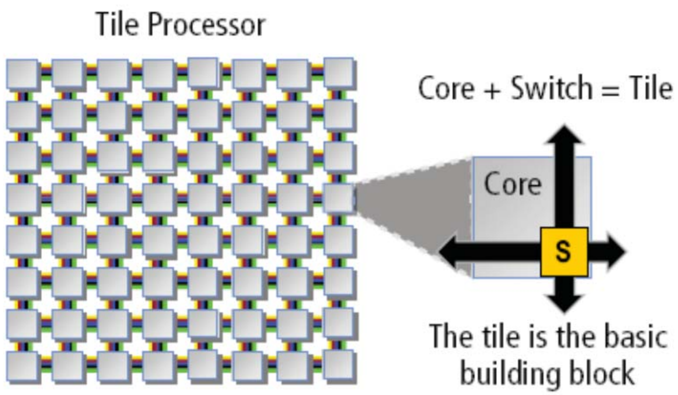




[Nov 18 2013]

[Nov 2012]

TILERA Architecture Targeting Embedded



- 64 cores connected in a mesh
- Local L1 + L2 caches
- Shared distributed L3 cache
- Linux + ANSI C
- New Libraries
- New IDE
- Stream computing
- …

Multicore trends

UPPSALA UNIVERSITET

Multicores: A different beast

Dr Dobb's 2005: The Free Lunch Is Over

A Fundamental Turn Toward Concurrency in Software

By Herb Sutter

"The biggest change in software development since the object-oriented revolution is knocking at the door, and its name is Concurrency...

Dr Dobb's 2012: Welcome to the Jungle

By Herb Sutter

"Mainstream computers [...] are being permanently transformed into heterogeneous supercomputer clusters. Henceforth, a single compute-intensive application will need to harness different kinds of cores, in immense numbers, to get its job done.".

Trends 1 (2)

Varies with application areas:

- Number of cores,
- Cache capacity,
- Bandwidth,
- Threads/core,
- Cache/thread varies.

- Important SW functions moved to HW

→ HW diversity is larger than ever

→ Intel has never released so many different chips per year

Trends 2

- Active power management (DvFS)
- Non-uniformity
- Heterogeneity
  - tiny/fat cores in the same system
  - different ISA
  - specialized HW
- Dark silicon is around the corner (!!)
  - Cannot have all transistors turned on
  - Reconfigurable architectures

Wrapping up about multicores

Erik Hagersten
Uppsala Universitet

What matters for multicore performance?

- Are we buying…
  - CPU frequency?
  - Number of cores?
  - MIPS and FLOPS?
  - Memory bandwidth?
  - Cache capacity?
  - Memory capacity?
  - Performance/Watt?

UPPSALA UNIVERSITET

MC Questions for the Future

- How to get parallelism?
- How to get performance/data locality?
- How to debug?
- A case for new funky languages?
- A case for automatic parallelization?
- Are we buying:
  - compute power,
  - memory capacity, or
  - memory bandwidth?
- Will 1000 cores be mainstream in 5 years?
- Will the CPU market diverge into desktop/capacity/capability/special-purpose CPUs again?
- **A non-question: will it happen?**