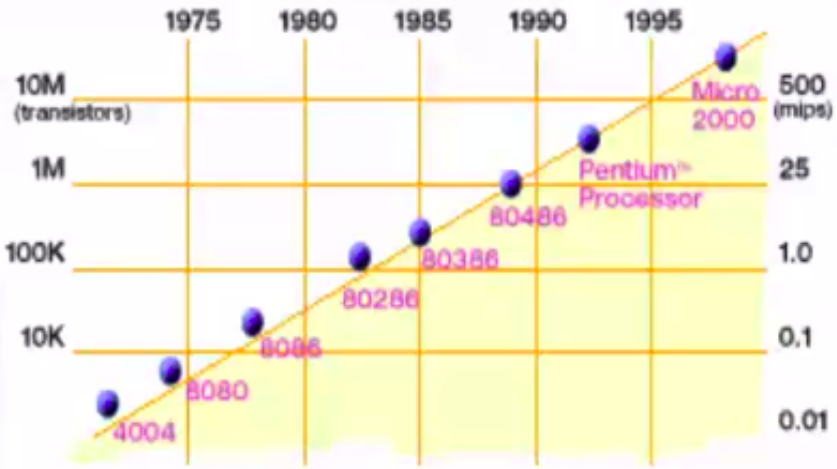
Multiprocessor Revolution

Hemangee K. Kapoor

Department of CSE, IIT Guwahati

Moore's law is Alive and Well

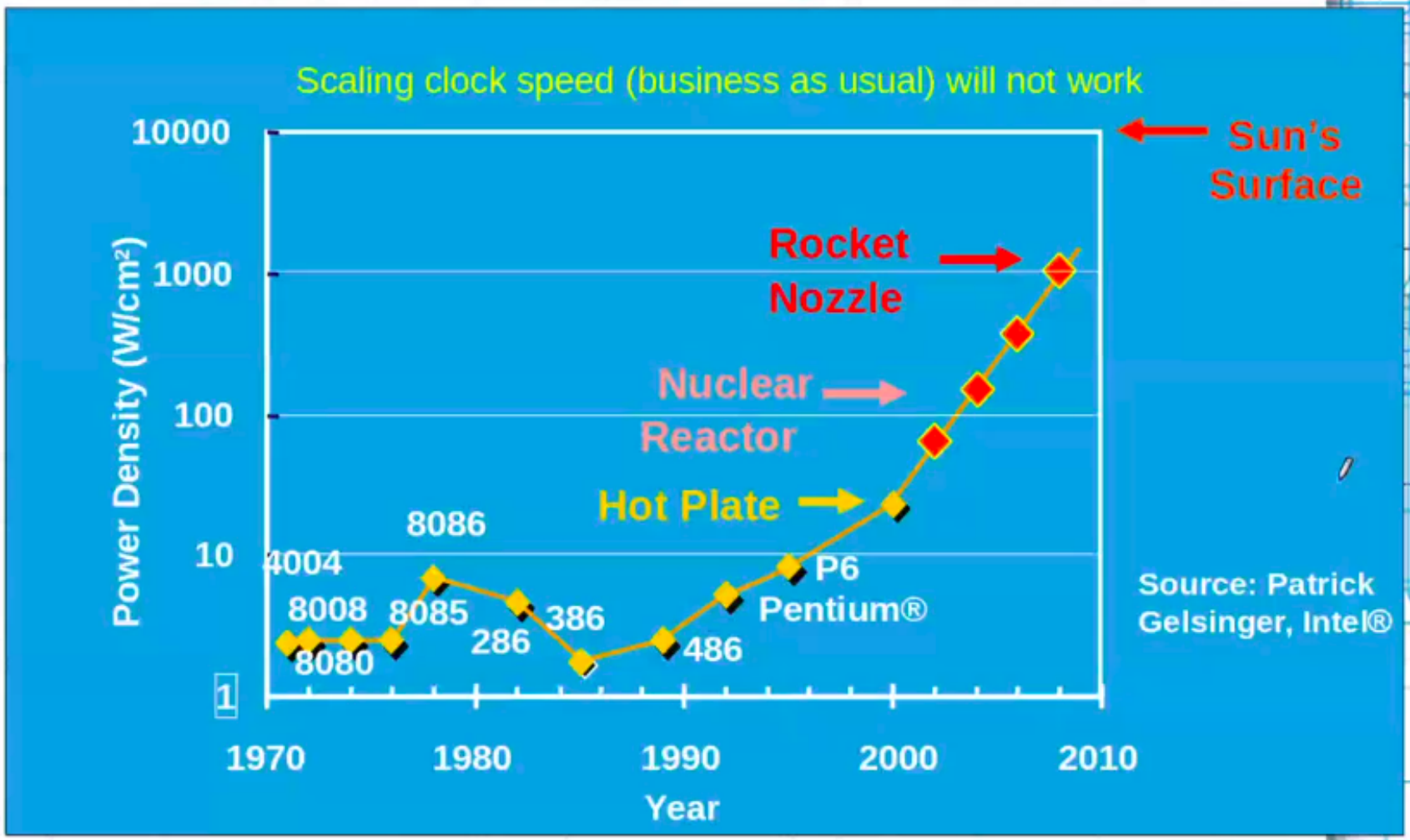


2X transistors/Chip Every 1.5 years
Called "Moore's Law"

Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

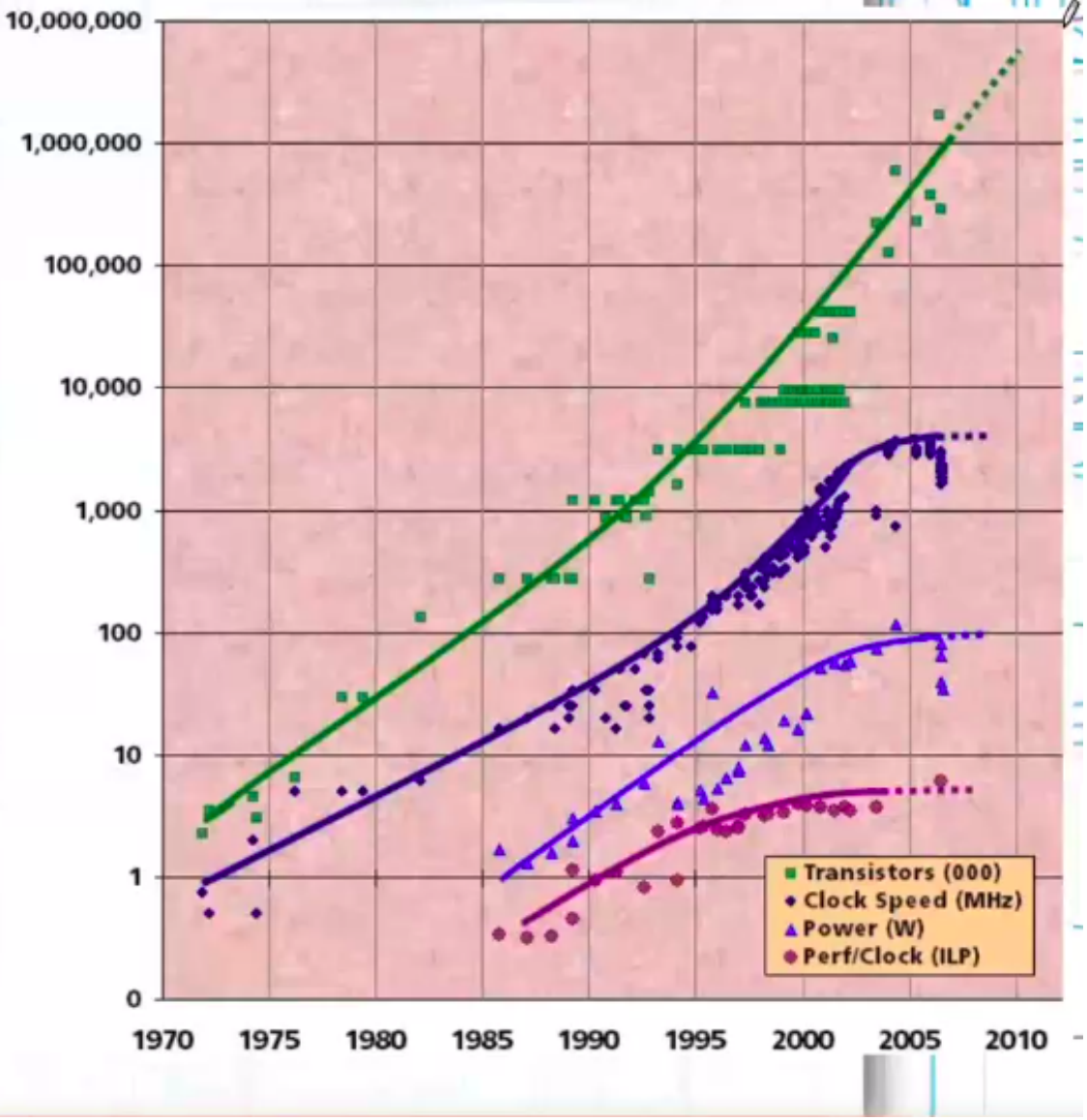
Microprocessors have become smaller, denser, and more powerful.

Clock scaling hits Power Density Wall



Revolution is happening

- Chip density is continuing to increase approx. 2 times every 2 years
  - Clock speed is not
- There is little or no hidden parallelism (ILP) to be found
- Parallelism must be exposed to and managed by software
- Number of processor cores may double instead

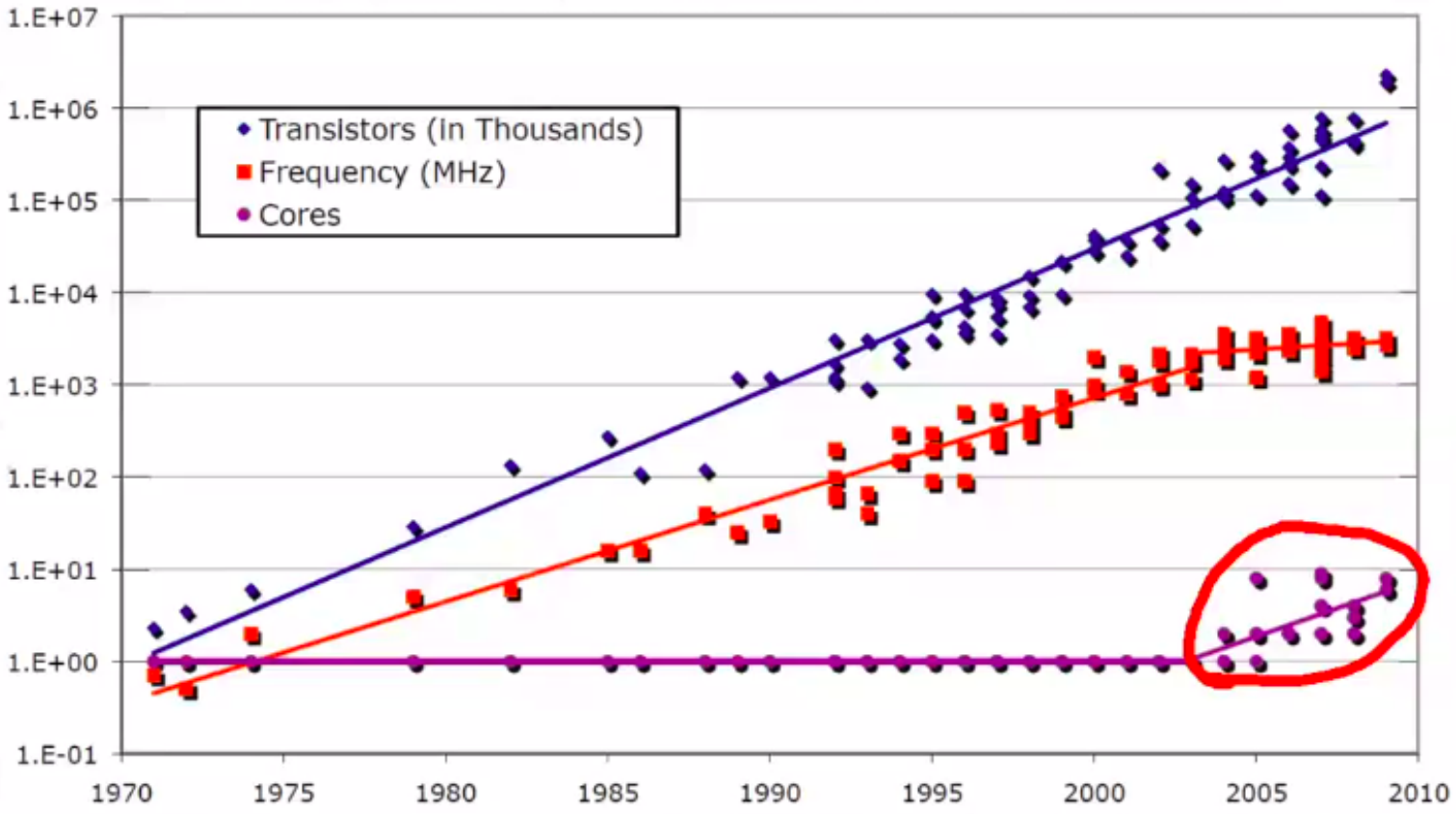


Source: Intel, Microsoft (Sutter) and Stanford (Olukotun, Hammond)

Processor designers forced to go “multicore”

- Heat density: faster clock means hotter chips
  - more cores with lower clock rates burn less power
- Declining benefits of “hidden” Instruction Level Parallelism (ILP)
  - Last generation of single core chips probably over-engineered
  - Lots of logic/power to find ILP parallelism, but it wasn’t in the applications
- Yield problems
  - Parallelism can also be used for redundancy
  - IBM Cell processor has 8 small cores; a blade system with all 8 sells for $20K, whereas a PS3 is about $600 and only uses 7

Moore's law

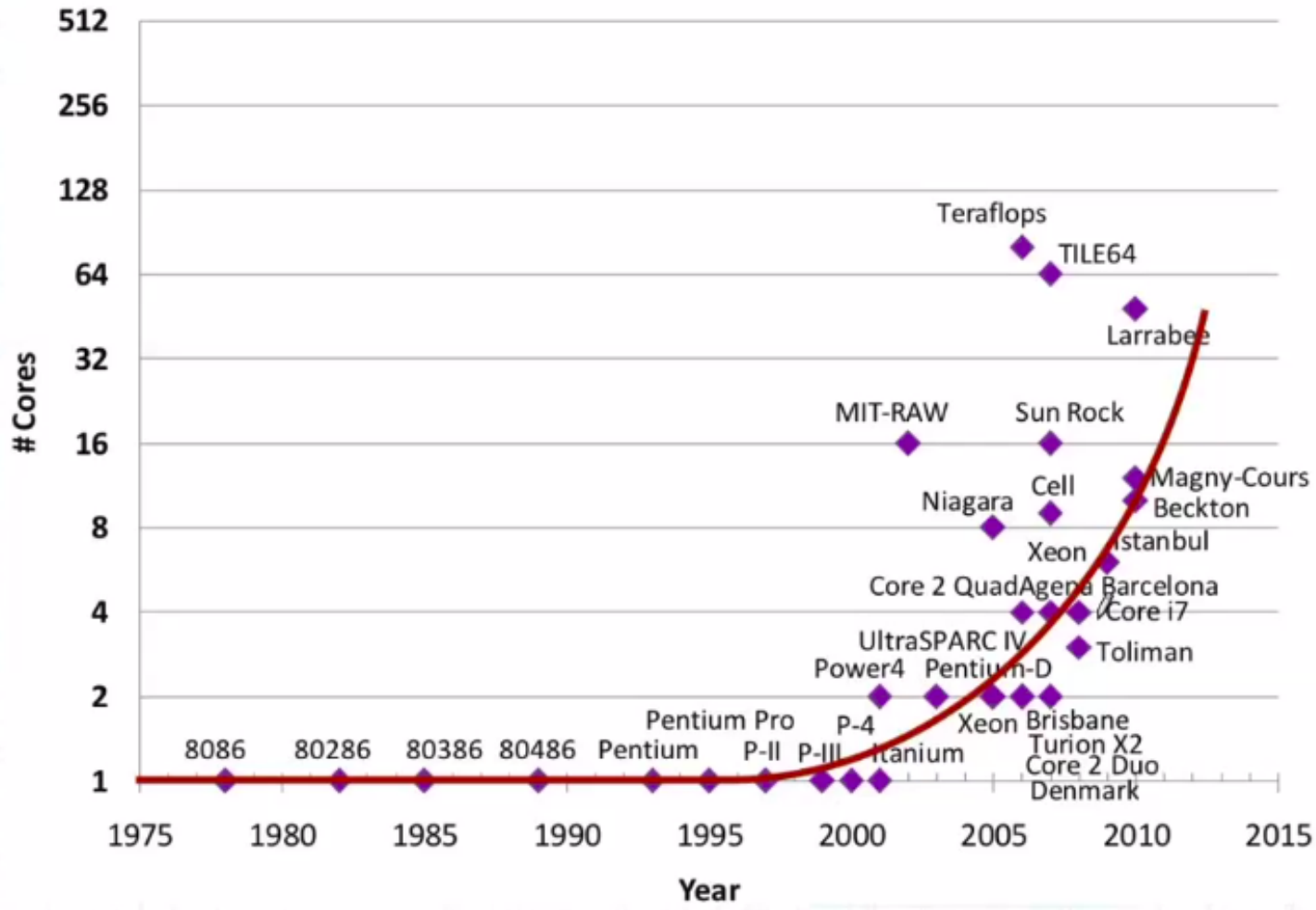


- "New" Moore's Law: 2x cores with every generation
- On-chip cache grows commensurately to supply all cores with data

Moore's law reinterpreted

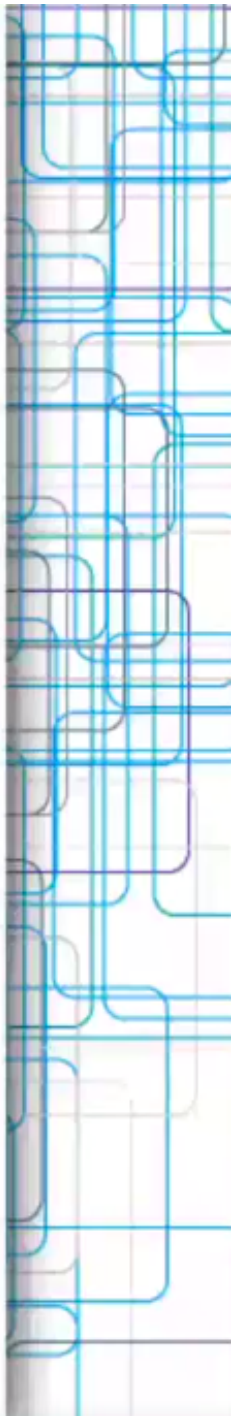
- Number of cores per chip will double every two years
- Clock speed will not increase (possibly decrease)
- Need to deal with systems with millions of concurrent threads
- Need to deal with inter-chip parallelism as well as intra-chip parallelism

New Moore's law



Number of cores/chip: Manycore

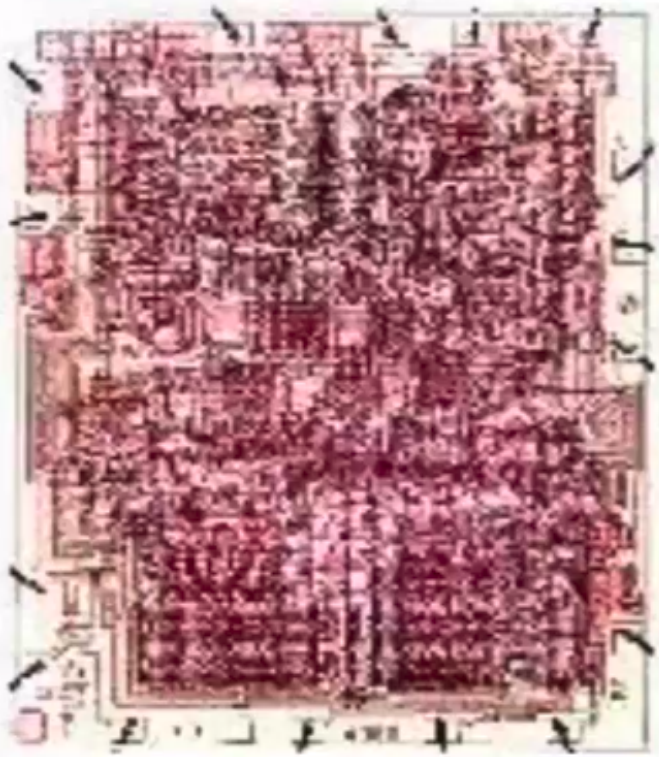
- "Multicore" 2X cores per generation: 2, 4, 8, …
- "Manycore" 100s of cores
- Multicore architectures & Programming Models good for 2 to 32 cores won't evolve to Manycore systems of 100's of processors
  ⇒ Desperately need HW/SW models that work for Manycore or will run out of steam (as ILP ran out of steam)
- We need revolution, not evolution



Multi-core examples

From “old” unicore

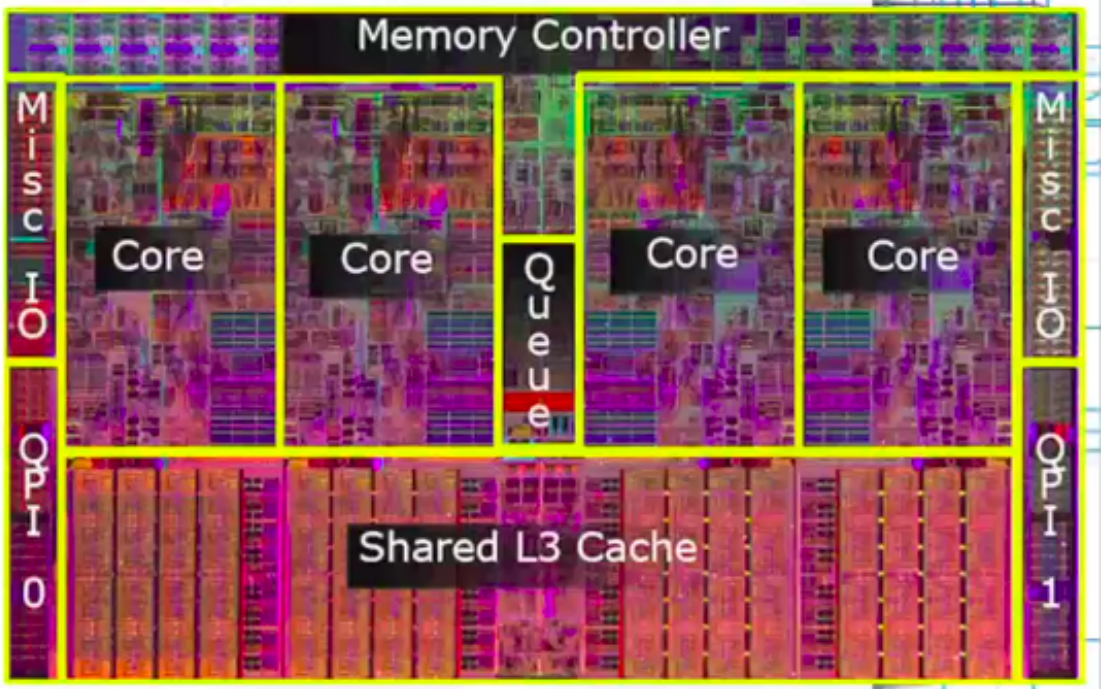
Intel 4004 (1971): 4-bit processor,
2312 transistors, ~100 KIPS,
10 micron PMOS, 11 mm^2 chip



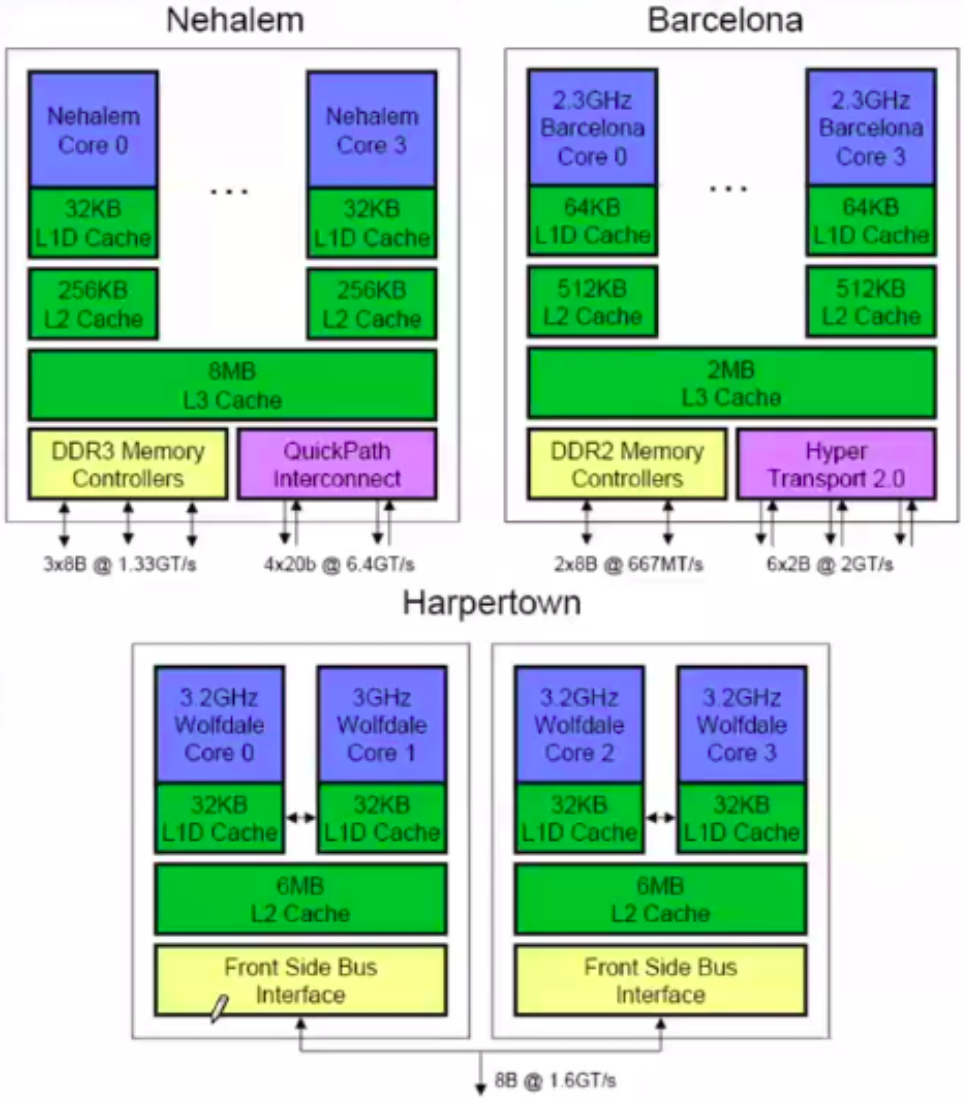
To Intel core i7

Nehalem: Quad core, 8 threaded, 64-bit,
4 issue superscalar, OoO, 16-stage pipeline,
48 bit virtual, 40 bit physical addressing

- 4 cores
- 731 million transistors, 263 mm^2 area, 45nm technology
- L3 is the last level cache, 8MB, 16-way set-assoc
- Each core has: 32KB, 8-way set-assoc, L1 (I and D), 256KB, 8-way set-assoc, L2
- Point-to-point interconnect called QuickPath

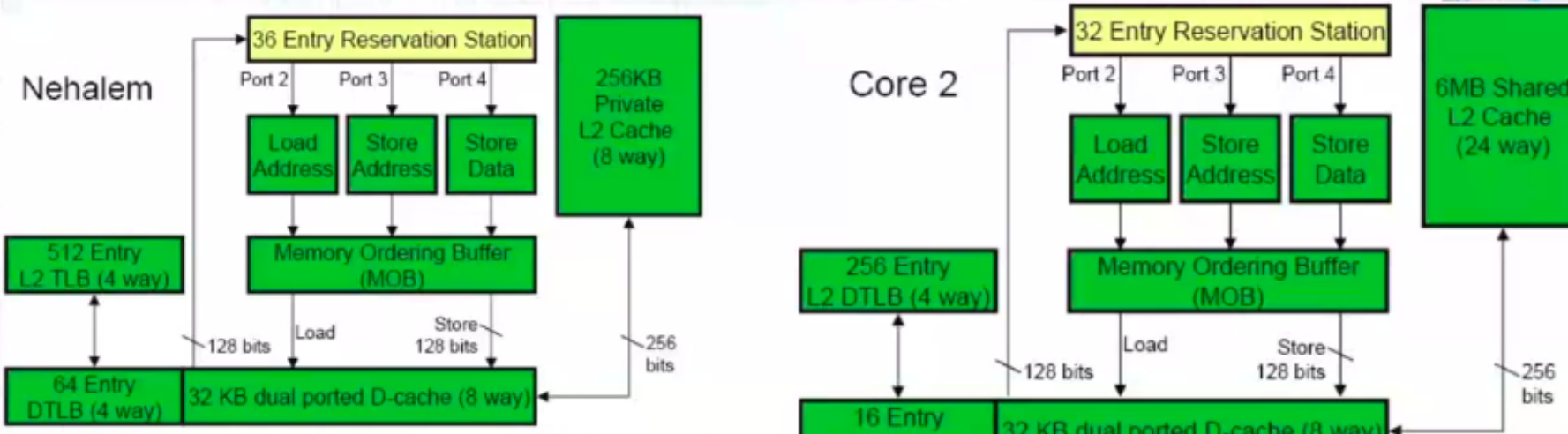


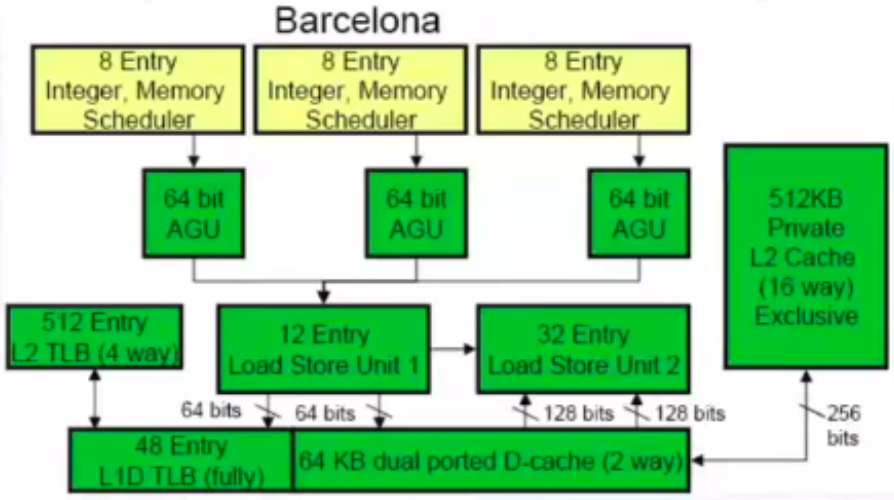
System Architecture comparison



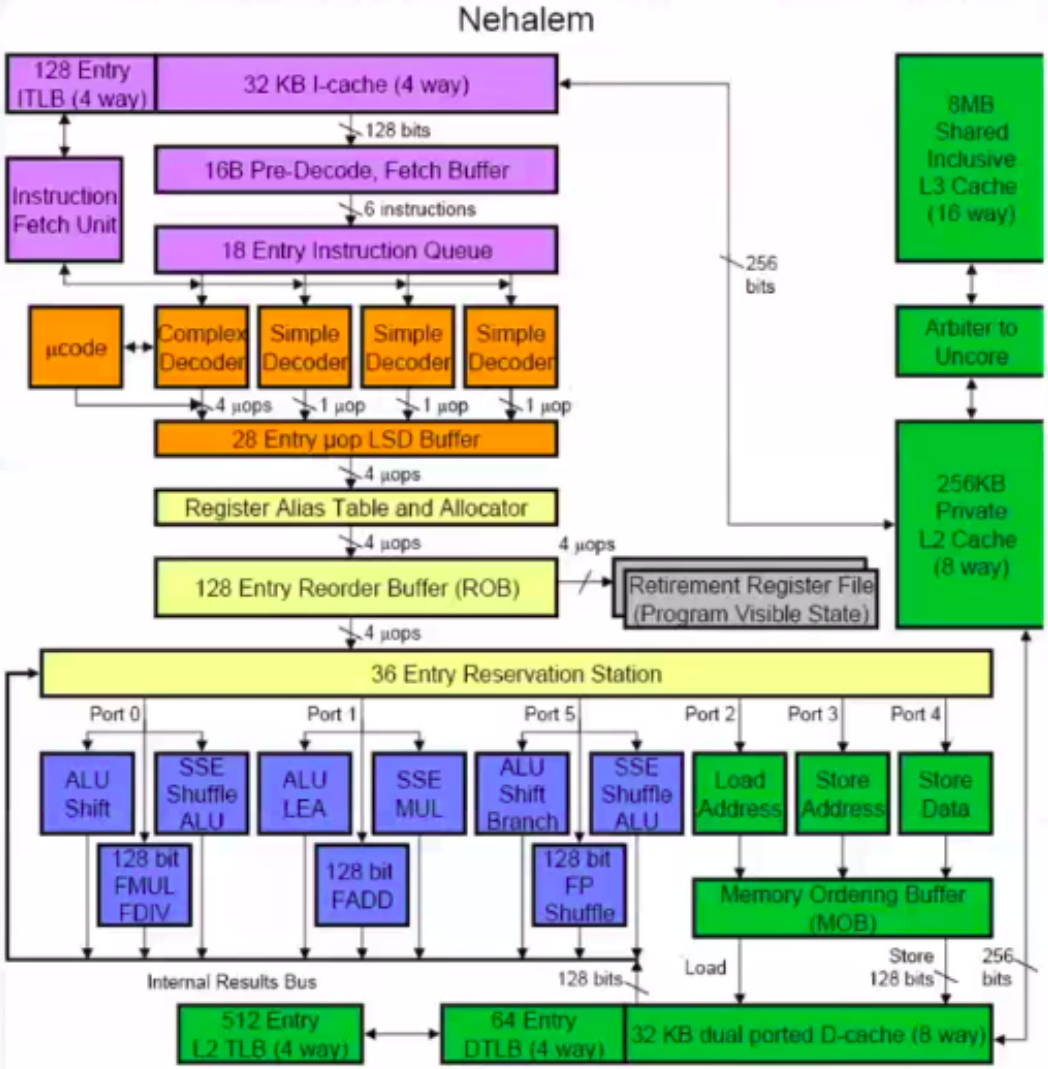
Nehalem, Harpentown: Intel
Barcelona: AMD

Memory subsystem comparison



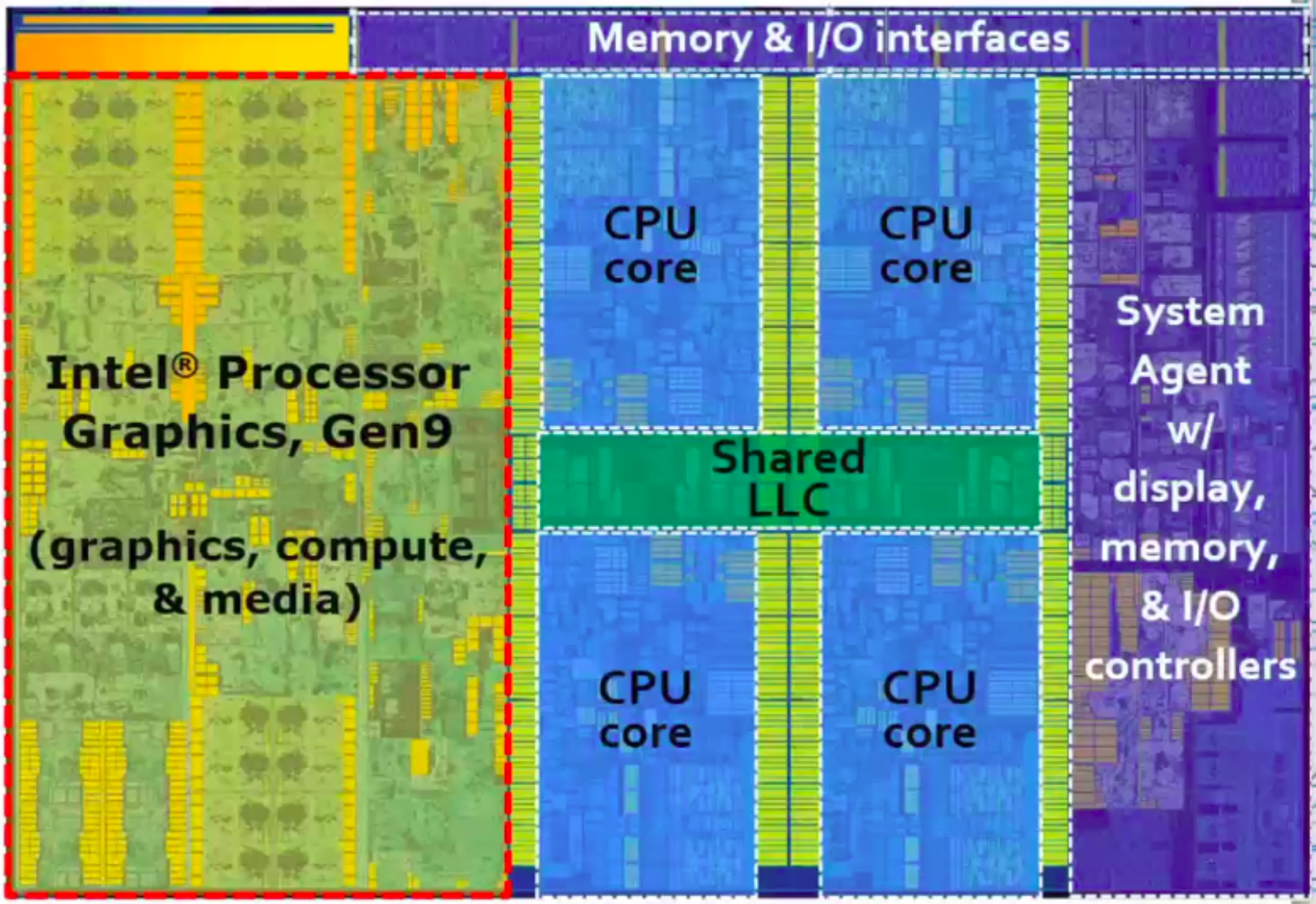


Nehalem architecture

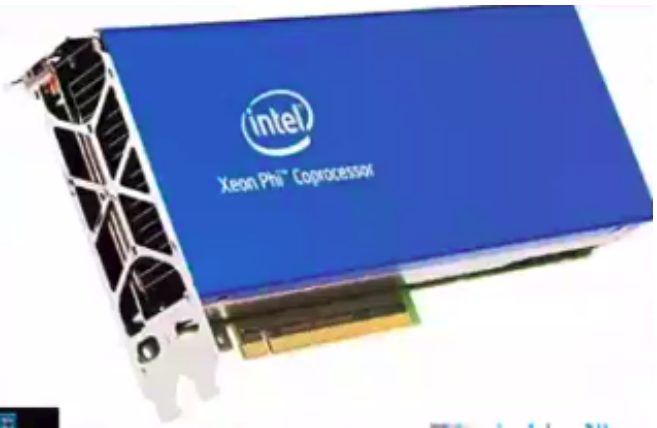


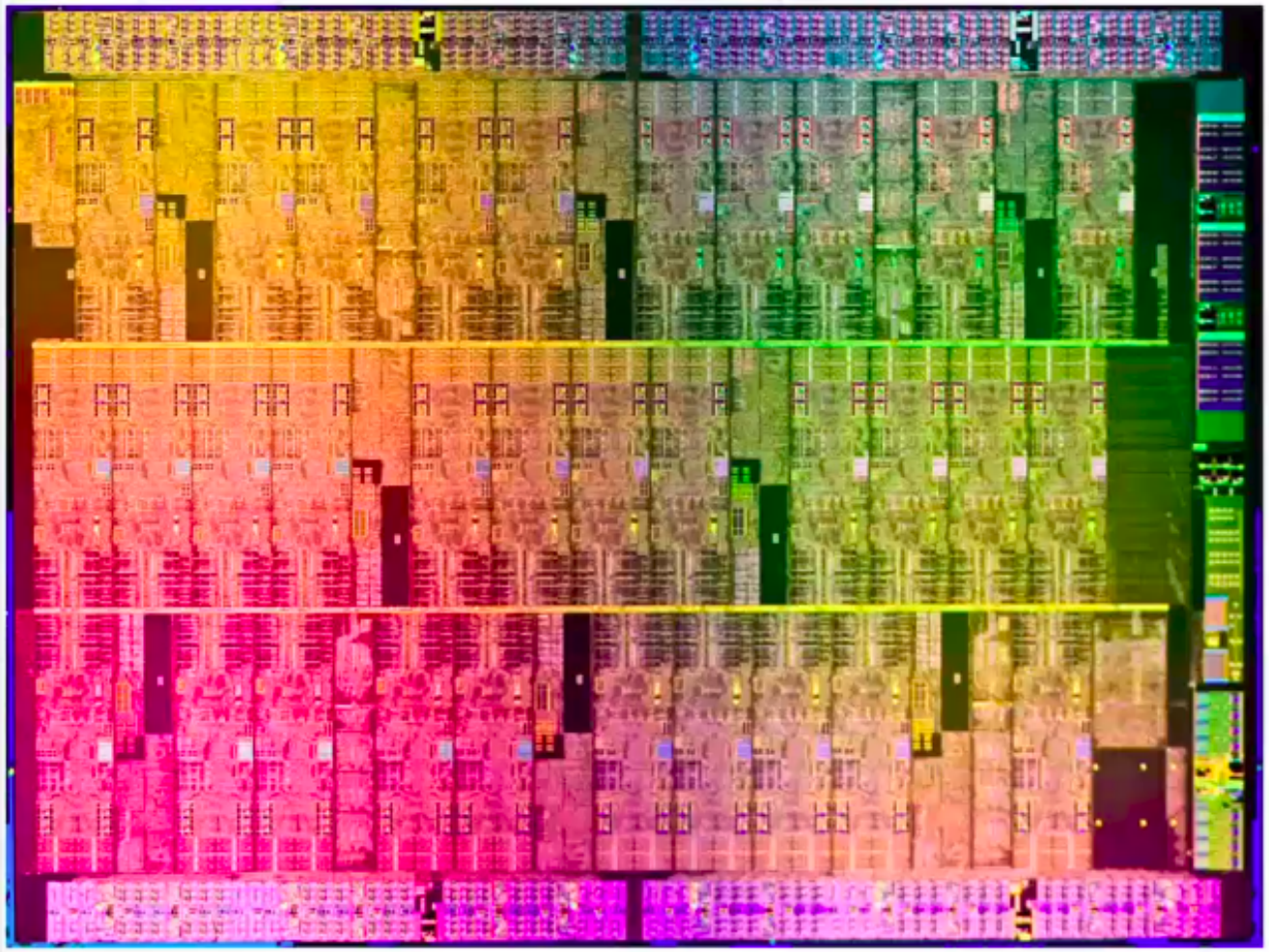
64-bit CPU, Register size = 64-bits
Virtual memory=2^{64}
Data bus=64 bits
64-bits can can be loaded/stored to memory

Intel Skylake – core-i7 (6th gen)



Intel Xeon-Phi





61 simple x-86 cores, 1.3 GHz

Accelerator for supercomputing applications

TESLA V100

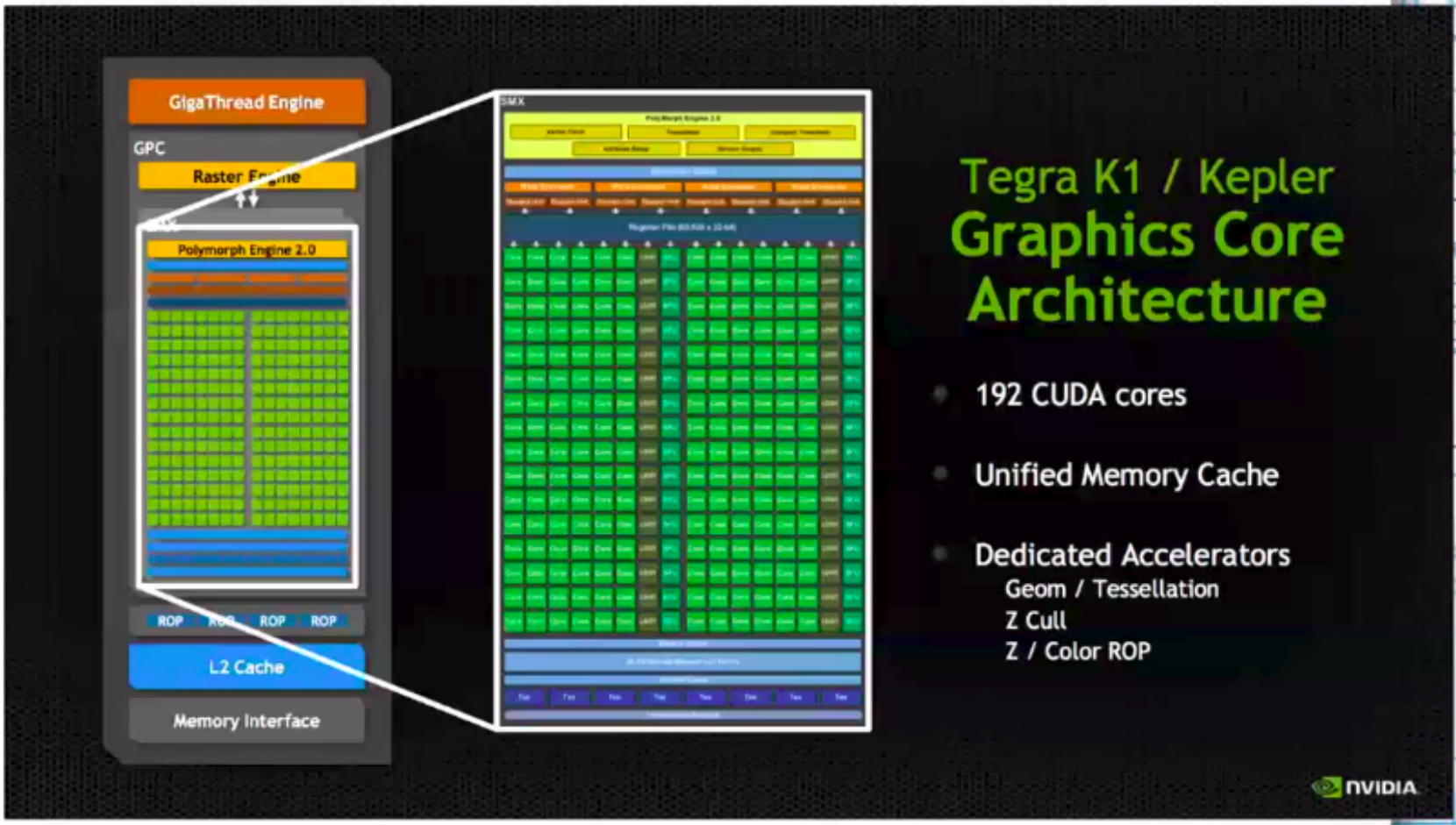
21B transistors
815 mm^2

80 SM
5120 CUDA Cores
640 Tensor Cores

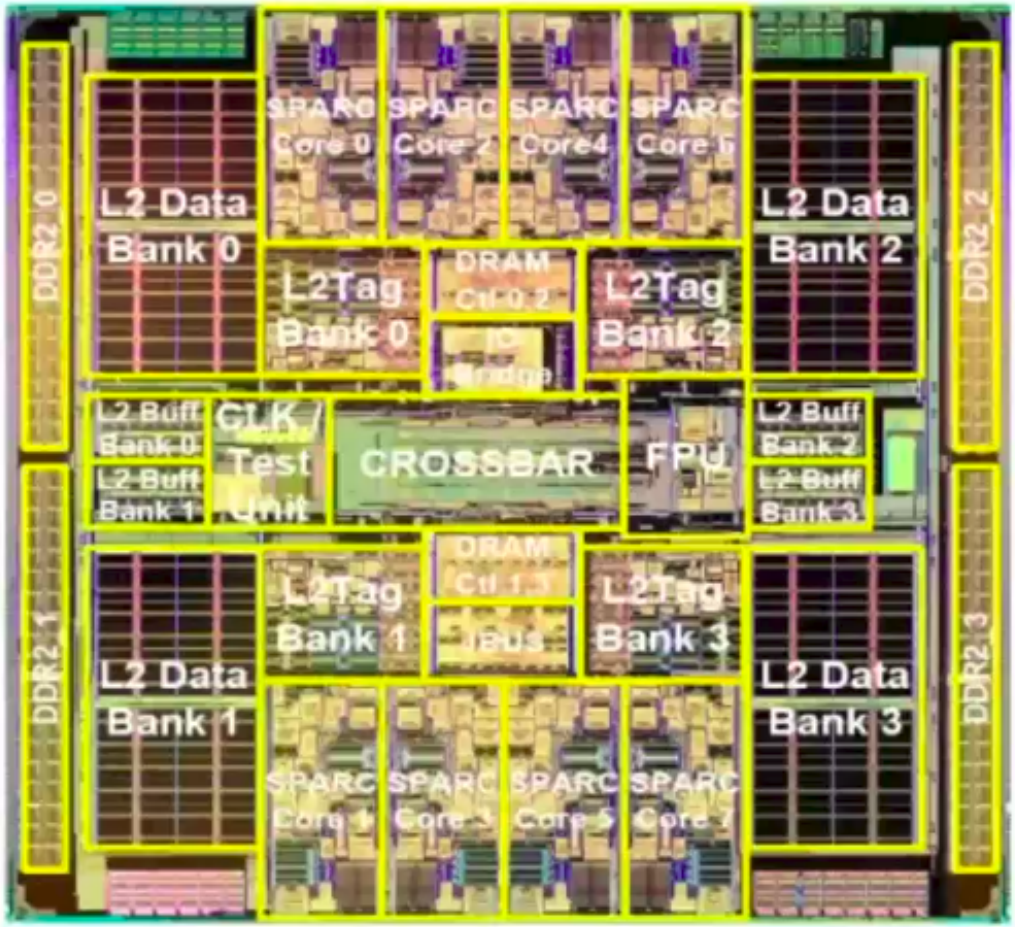
16 GB HBM2
900 GB/s HBM2
300 GB/s NVLink

*full GV100 chip contains 84 SMs

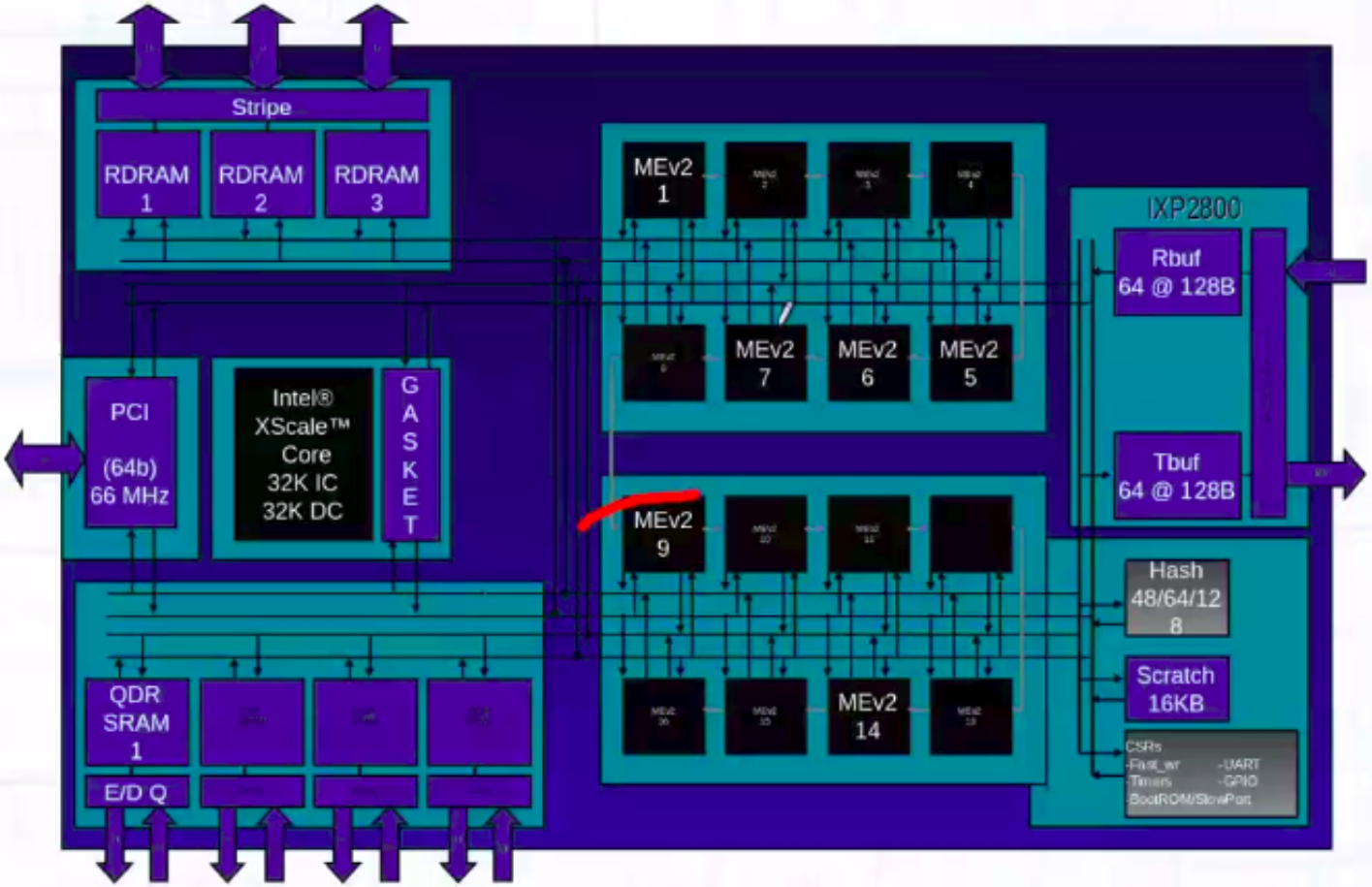
Mobile parallel processor



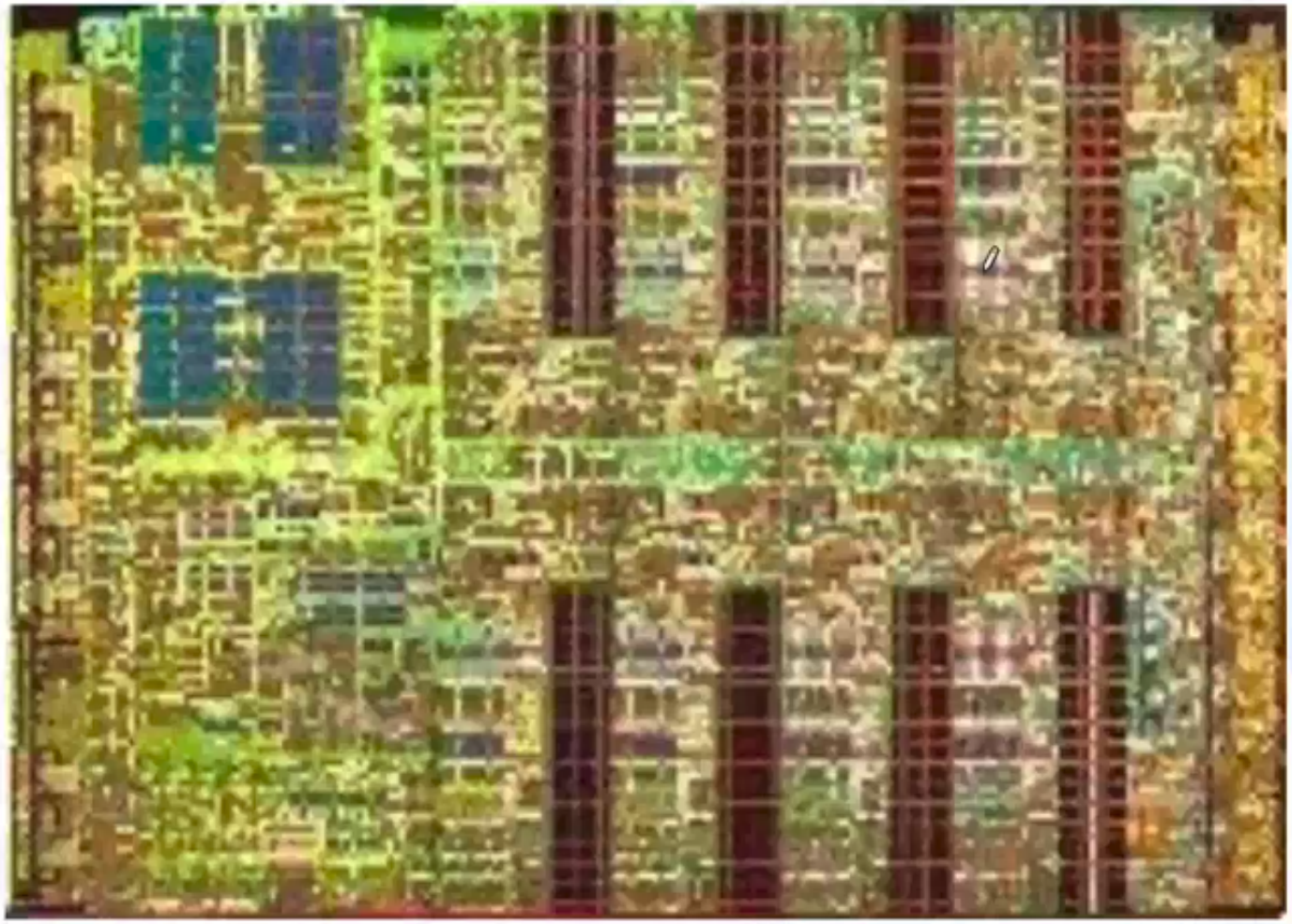
Sun Niagara
8 GPP cores (32 threads)



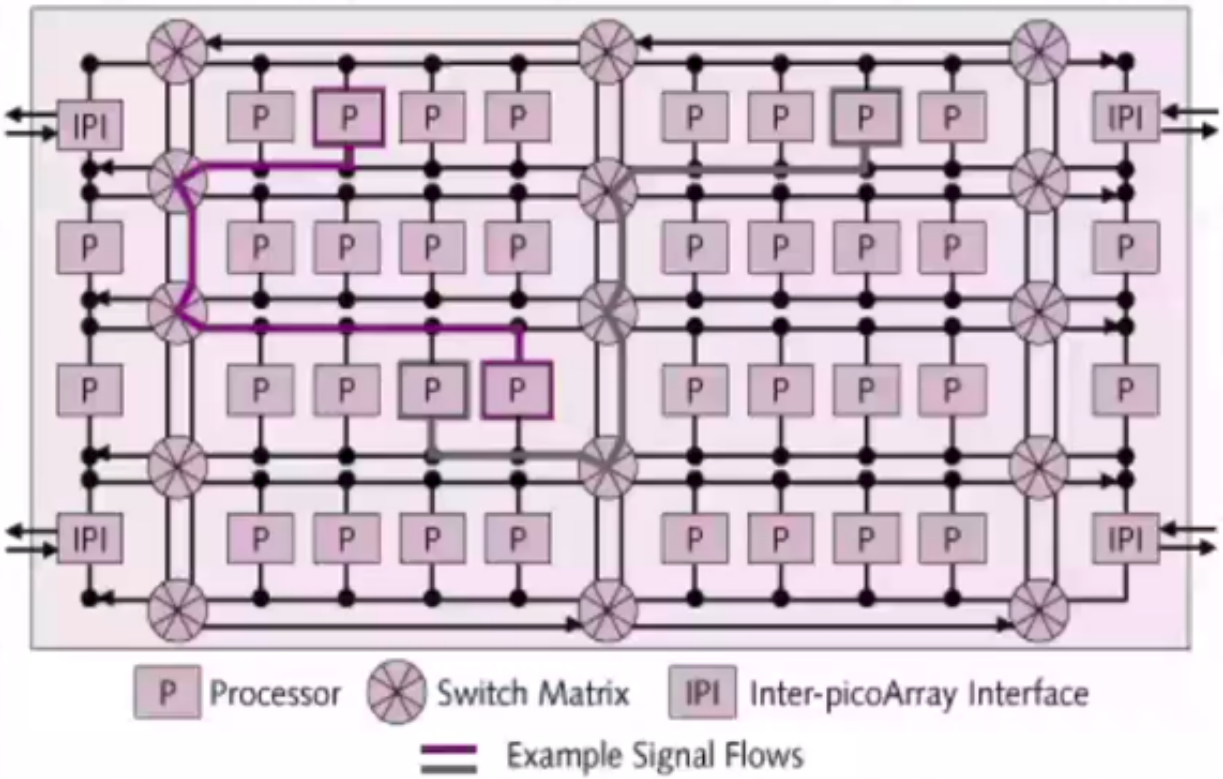
Intel Network Processor
1 GPP Core + 16 ASPs (128 threads)



IBM Cell
1 GPP (2 threads) + 8 ASPs



Picochip DSP
1 GPP core + 248 ASPs



Cisco CRS-1
188 Tensilica GPPs