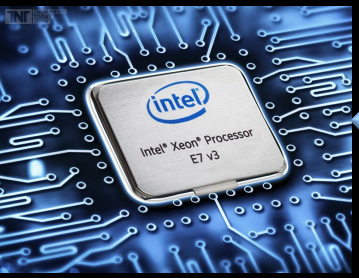
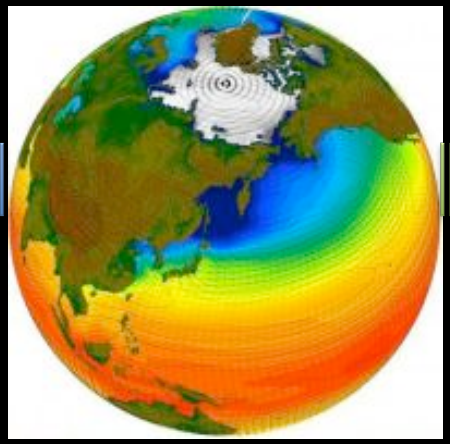
Parallel Computing Trends

Dr. Richard D. Loft
Director, Technology Development Division
Computational and Information Systems Laboratory
National Center for Atmospheric Research
loft@ucar.edu



?



?

Micro-processors

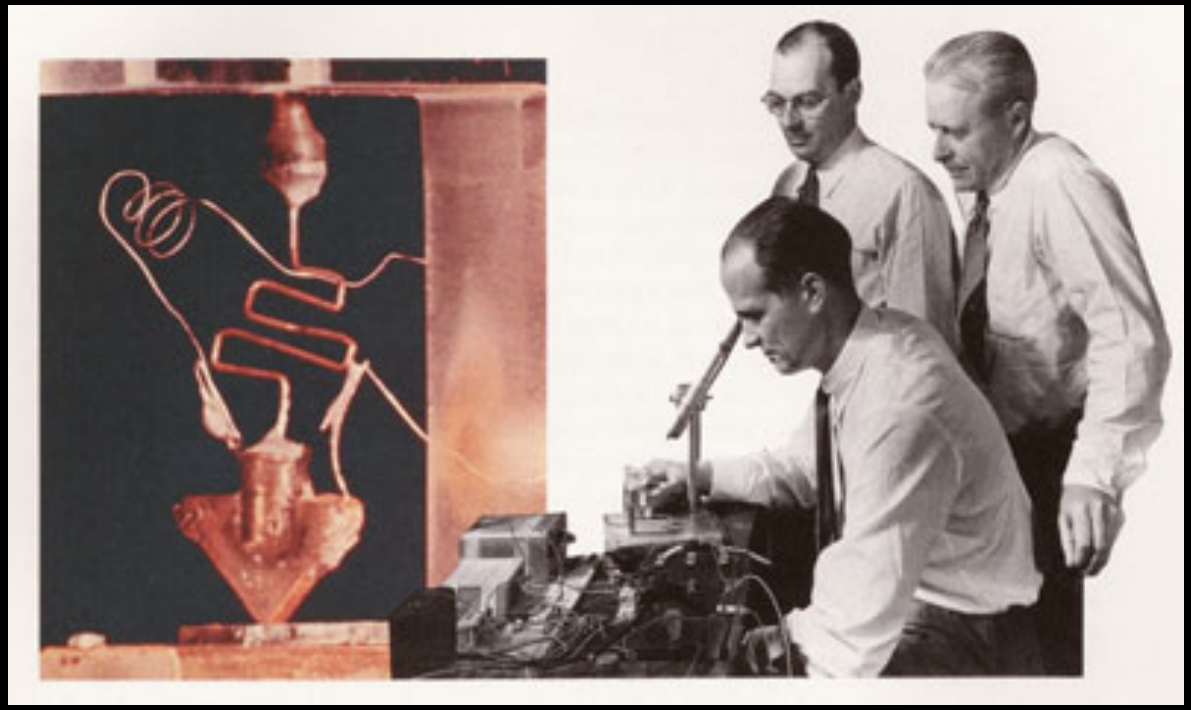
Earth System Models

Many-core
Accelerators

NCAR

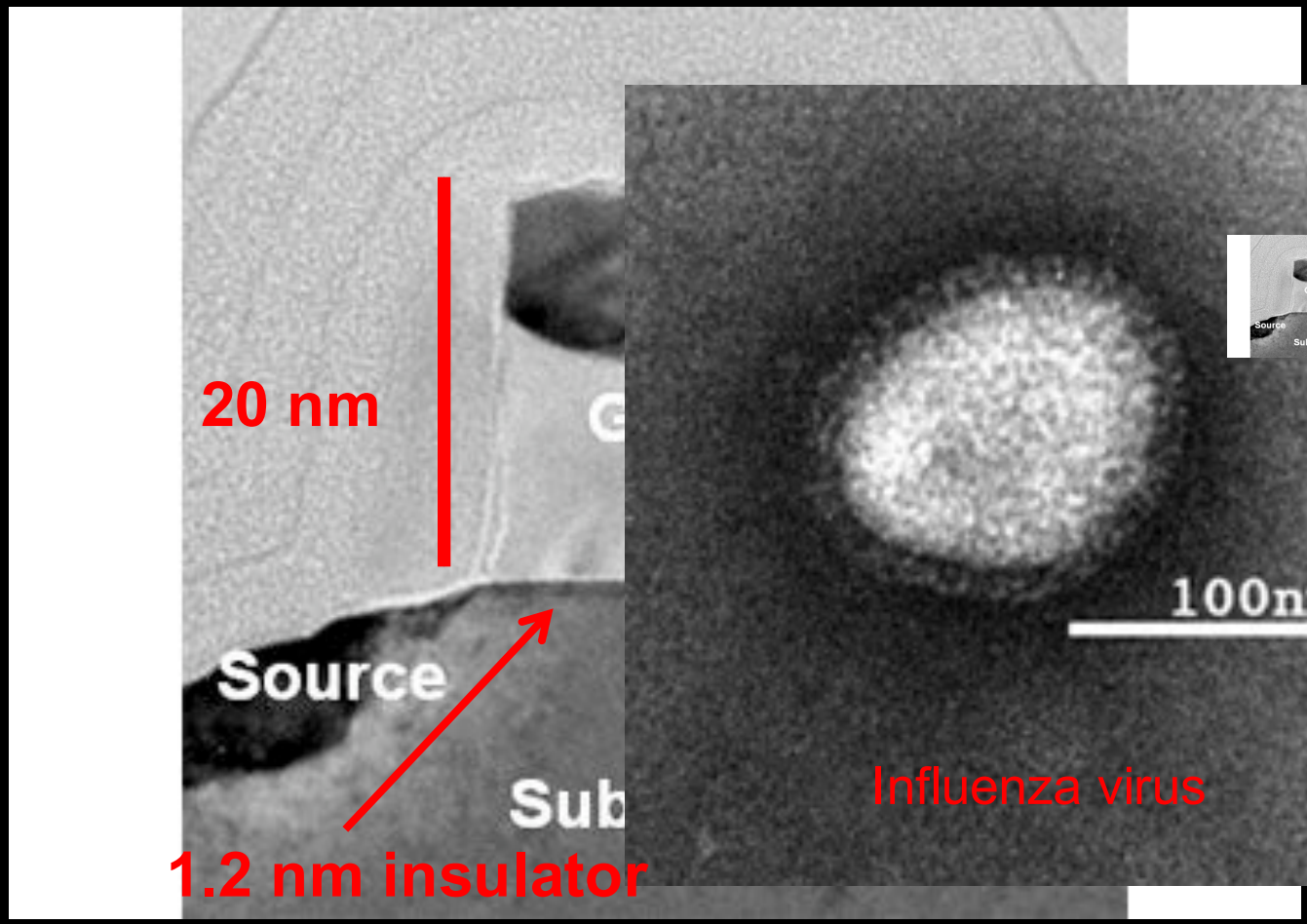
June 15, 2016

Transistor in 1947: ~ 5 cm across



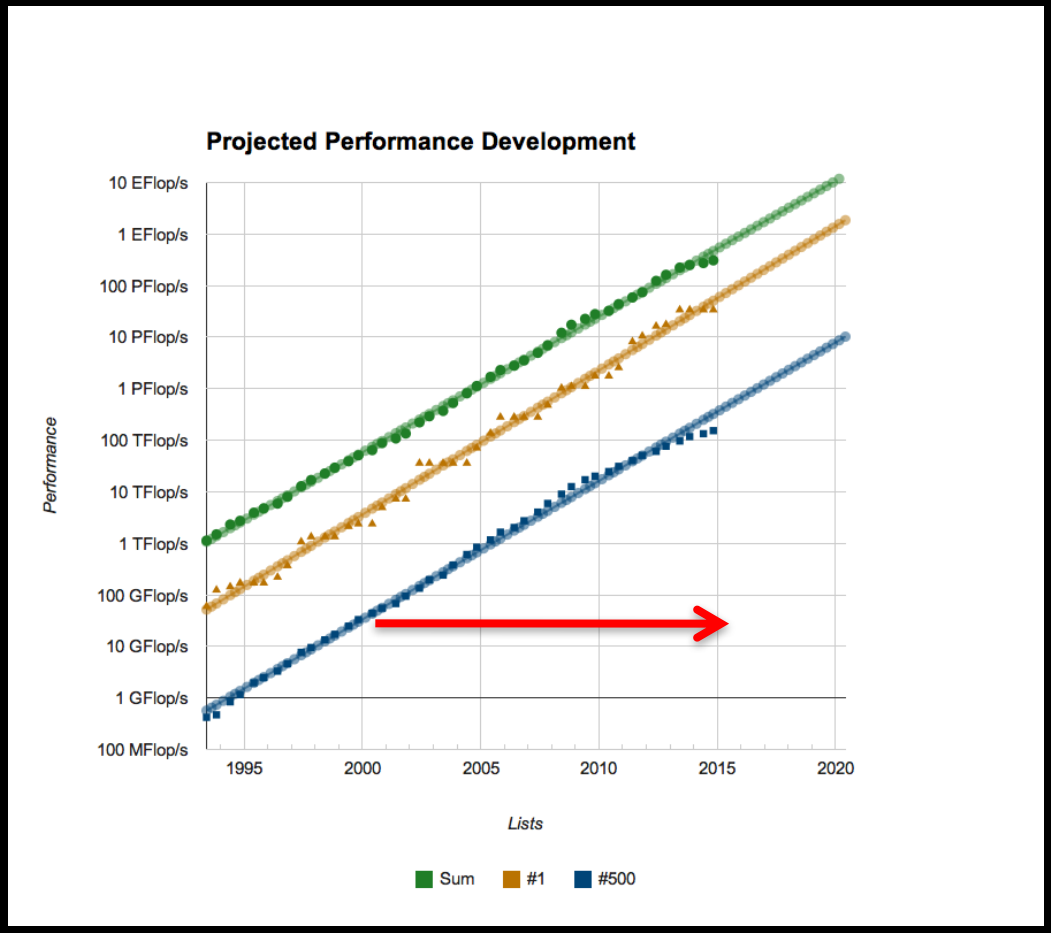
William Shockley (1910-1989), John Bardeen (1908-1991),
Walter Brattain (1902-1987)

The modern transistor: ~200 atoms across



So small that quantum tunneling is a problem!

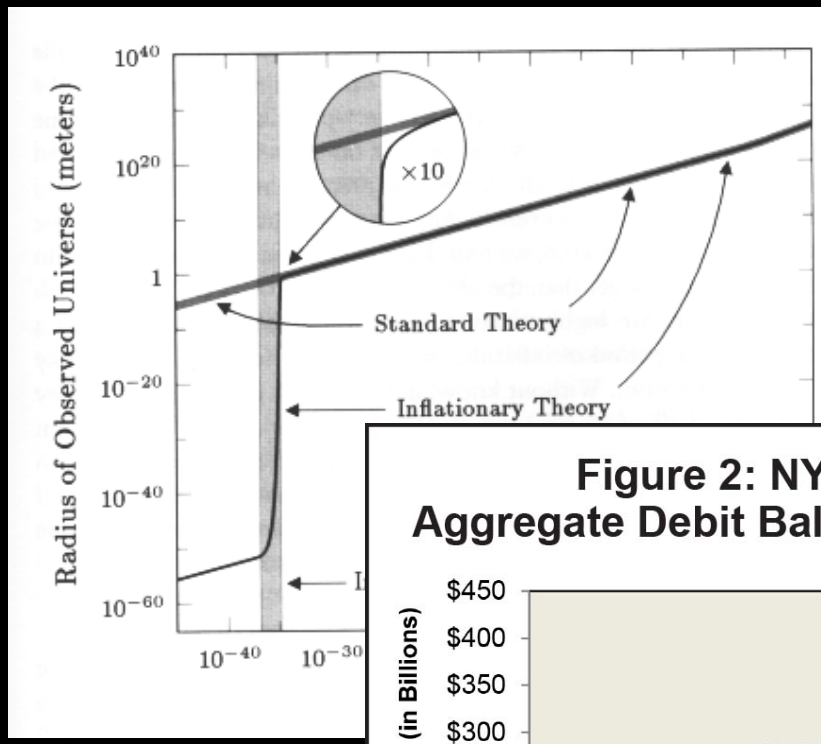
The exponential increase in the speed of computers. Looks relentless.

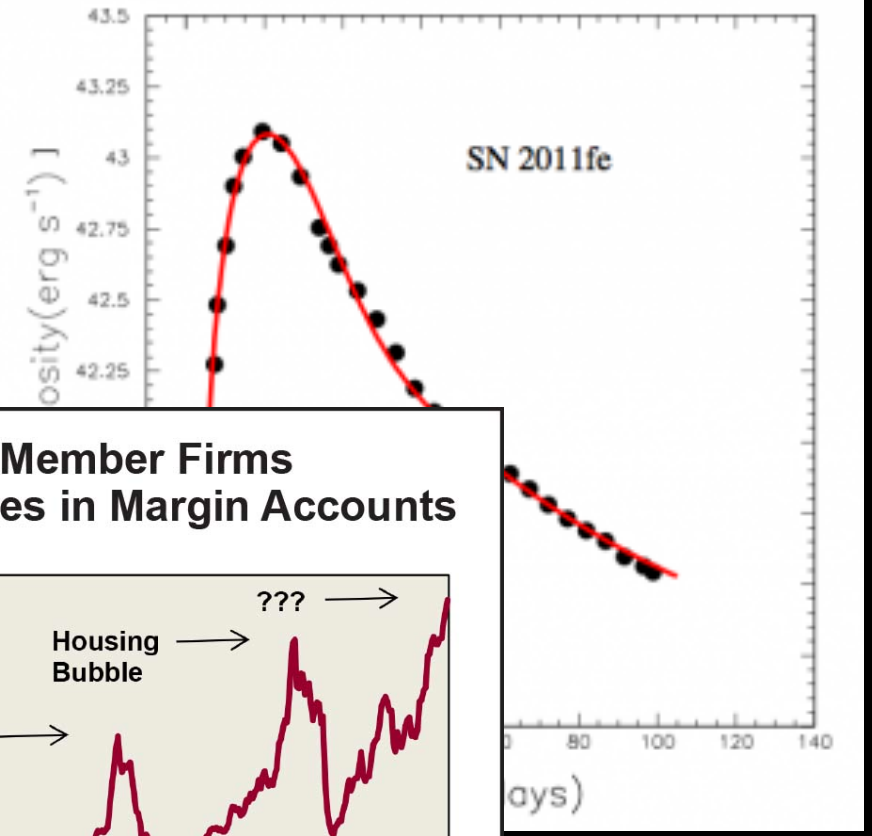


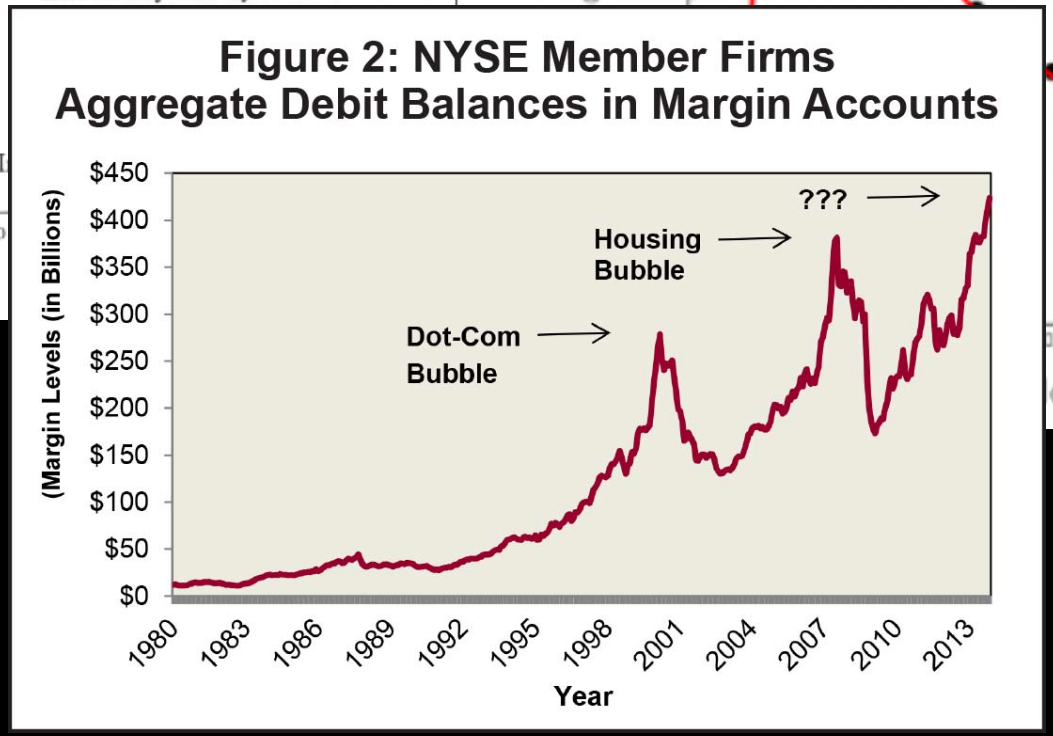
… This laptop would have been on the top 500 list in 2000

Courtesy of top500.org

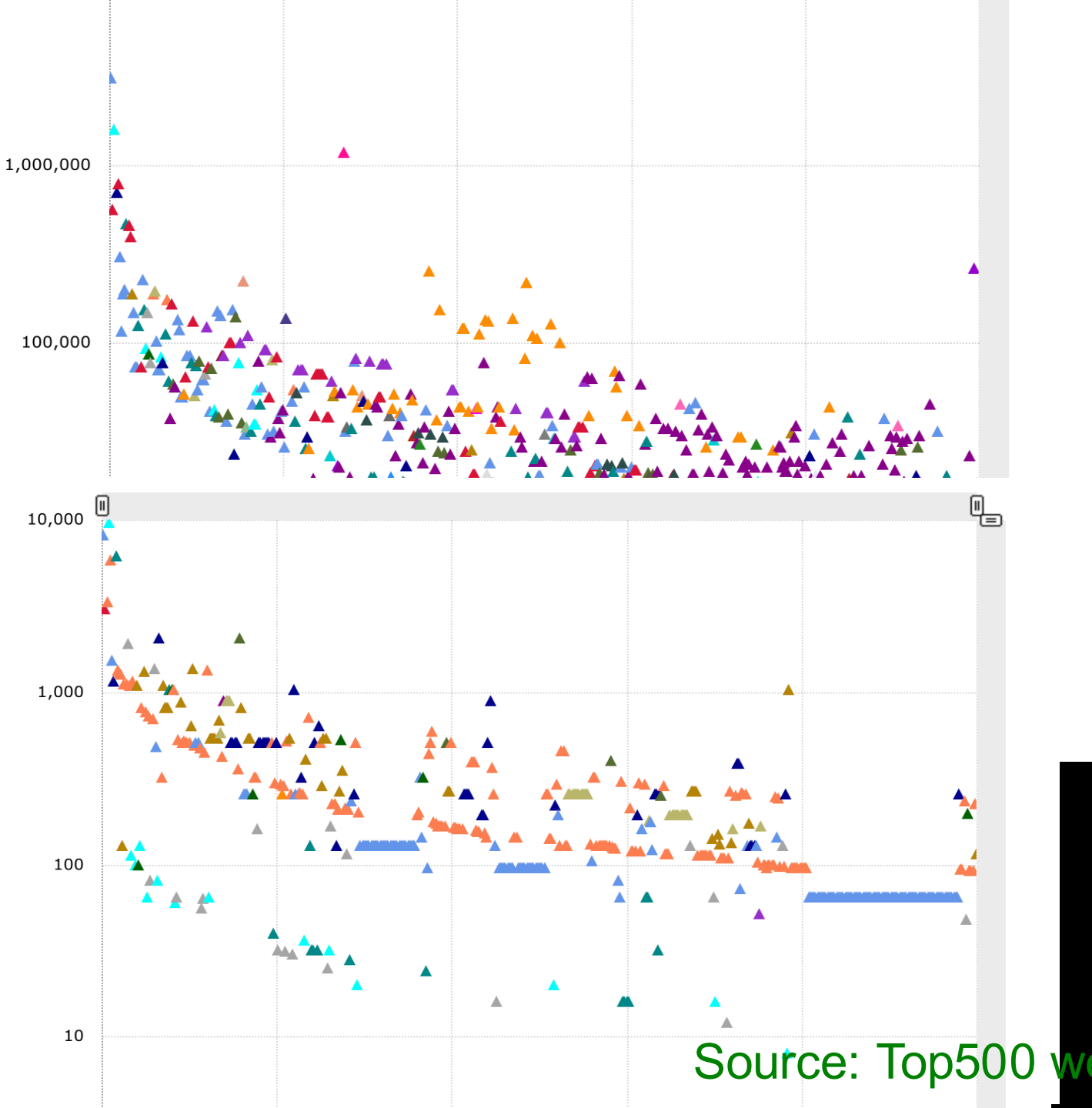
No exponential lasts forever!







To ng



Source: Top500 website

Path Forward is through Parallelism

The largest supercomputers now have O(10^5 -10^6) compute cores...

Mira: 786,432 cores

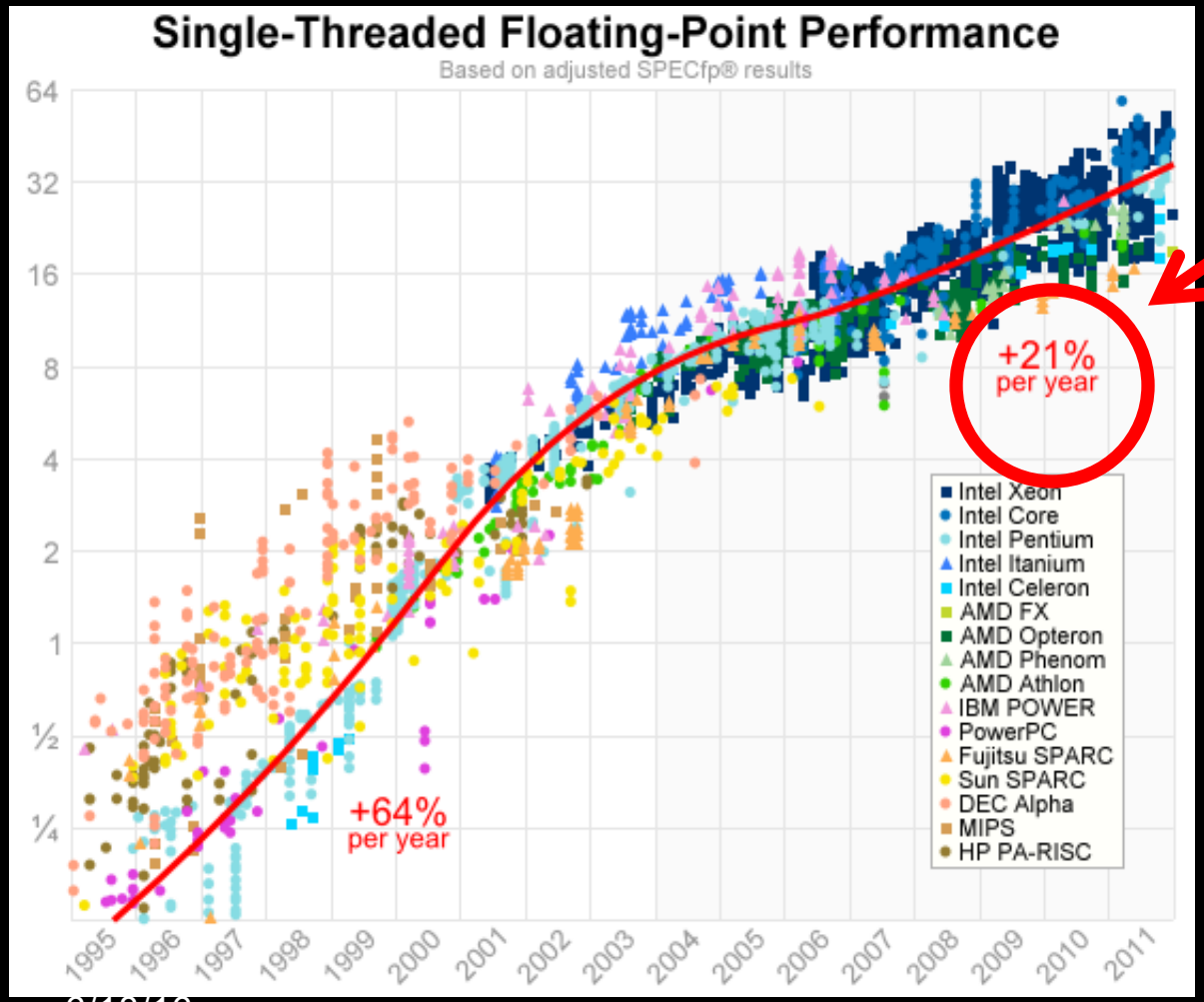
Blue Waters: 362,240 cores

Processor Trends

What happened?

- Around 2000 chip designers realized:
  - Power wall
    - Clock speed couldn't continue to rise due to
    - power density
  - Memory wall
    - Memory system speed couldn't keep up with processor speed
  - ILP wall
    - Benefit of adding processor "smarts" reached point of diminishing returns
- Solution:
  - Break up monolithic single processor into multiple computing regions: cores

Clock speeds stopped improving and single thread improvements slowed



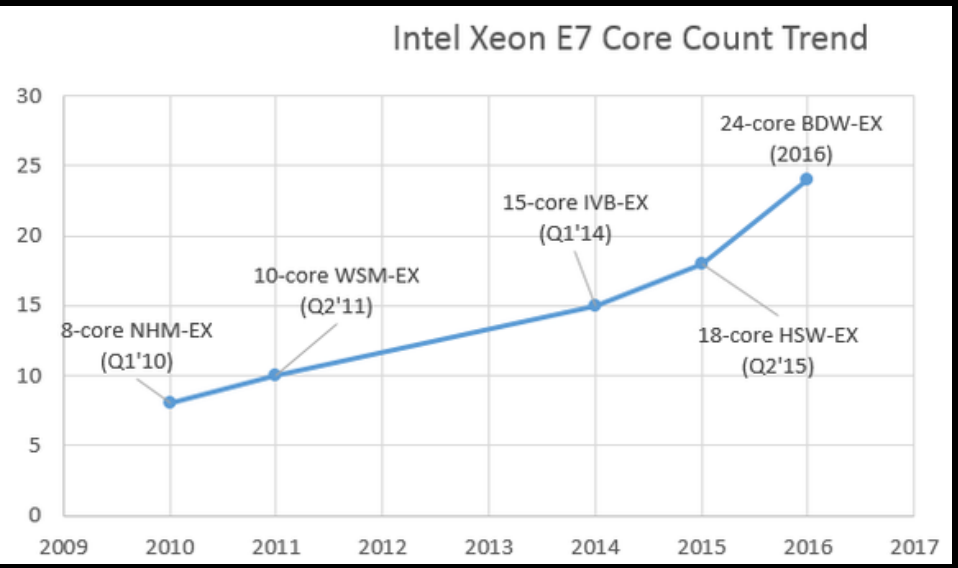
This translates to 5% Model resolution improvement per year

Bias corrected SpecFP results Preshing.com

Original Source- Sutter, H. "The Concurrency Revolution" Dr. Dobb's Journal, 30(3), March 2005.

Many-core: evolutionary approach

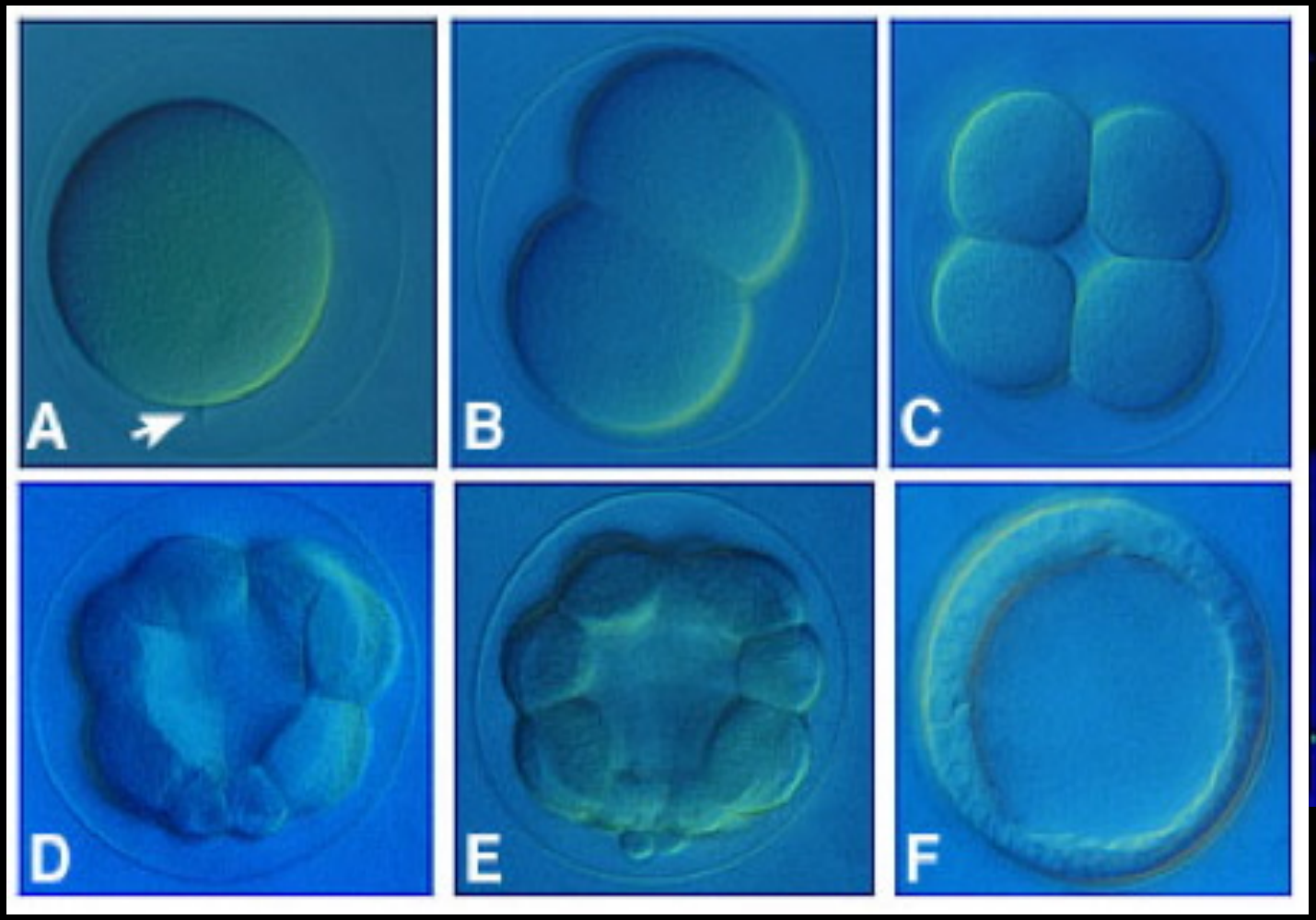
- Hold clock speed ~constant
- Hold core complexity ~constant
- Gradually add more cores on the die as transistors shrink
- Gradually add SIMD features (vector units)



Source: Ashraf Eassa, The Motley Fool

SIMD =
Single Instruction Multiple Data

The Processor Zygote: Speed-up through increasing core parallelism on processor

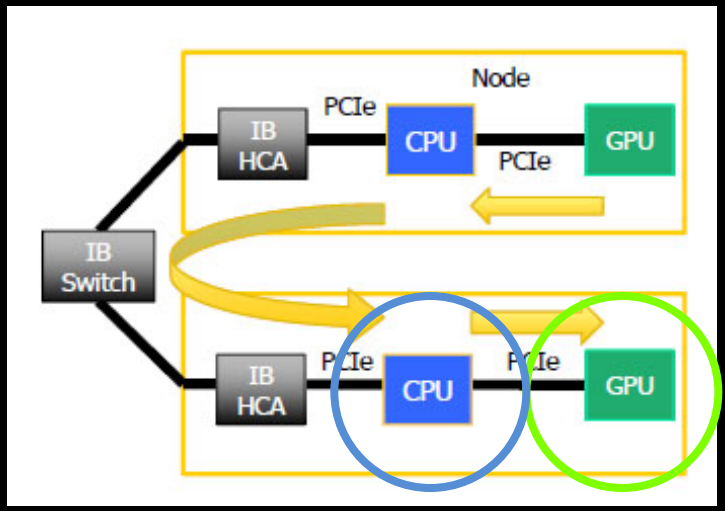


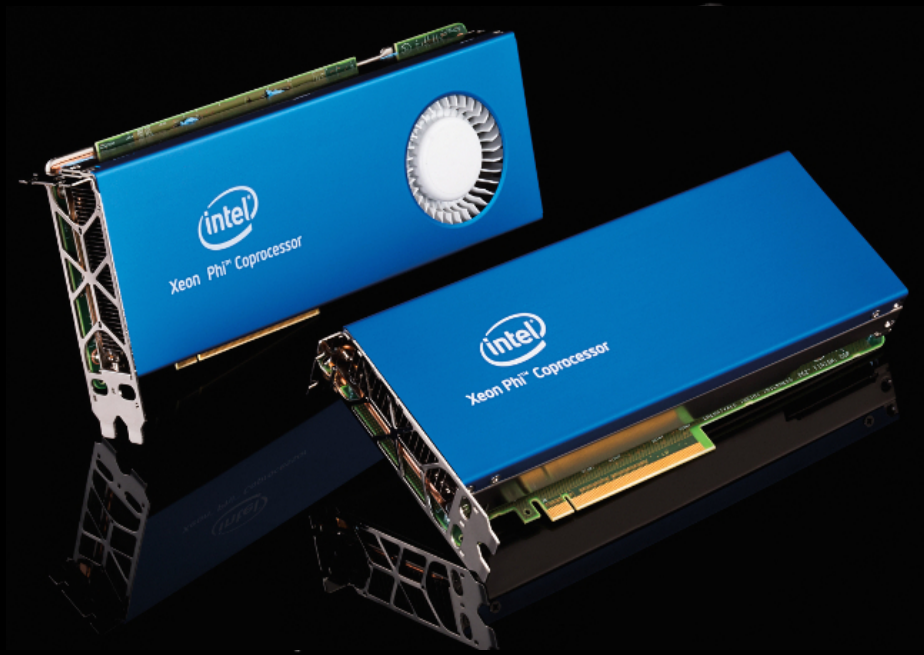
What are processors developing into?

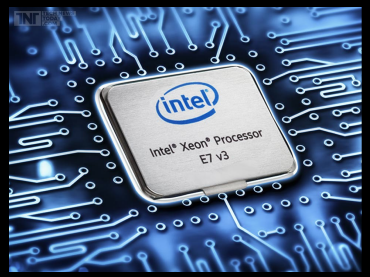
Many-core: revolution

- Solution 2: single core -> many-core
  - Drop clock speed (i.e. by 2-3x)
  - Drop core complexity
  - Stamp out a lot more simplified cores on chip
  - Add even more cores as transistor shrinks
  - Add much more “SIMD-ness” to architecture

Added complexity: coprocessors and heterogeneous computing





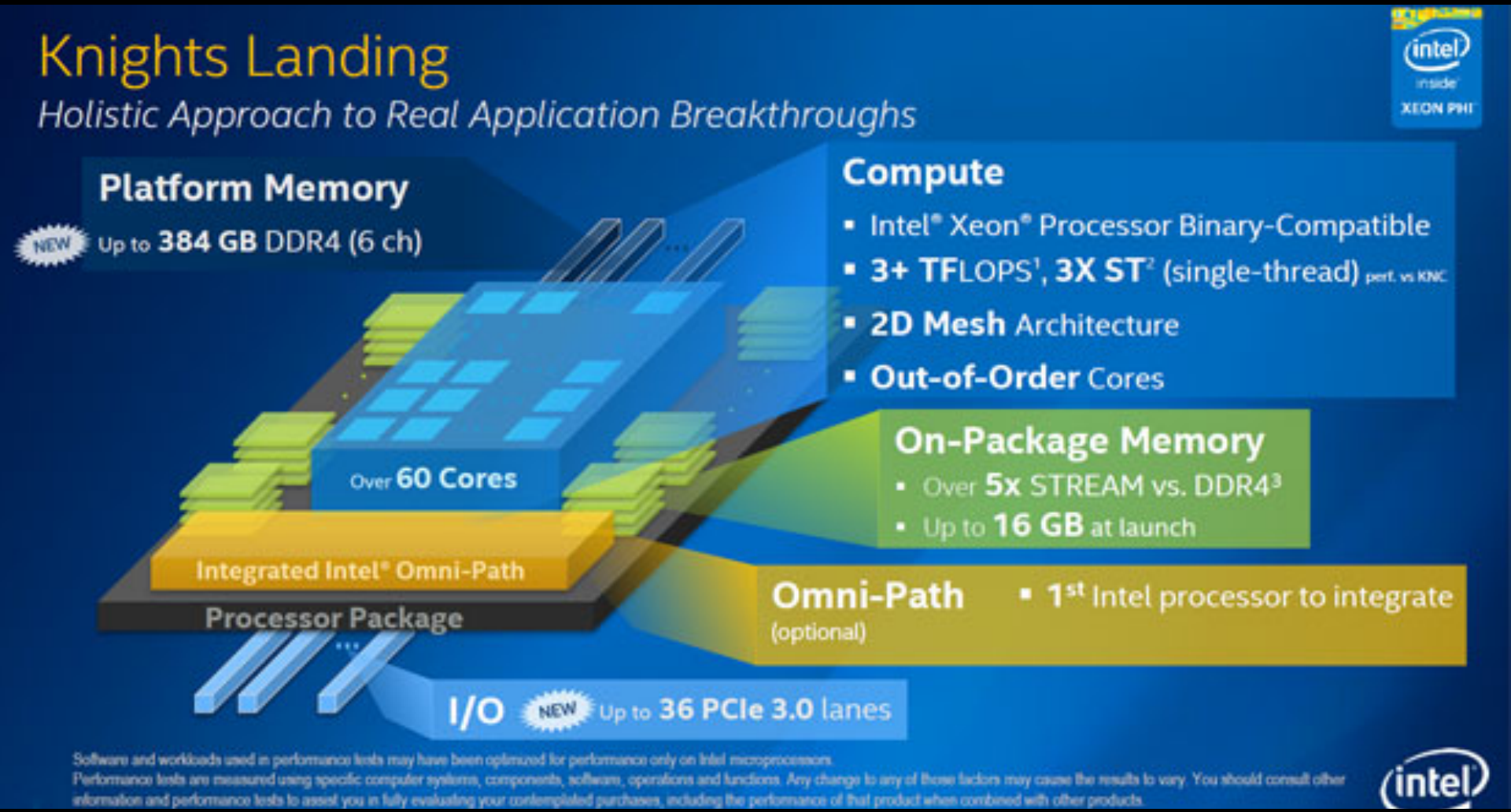


Traditiona[illegible]his model:

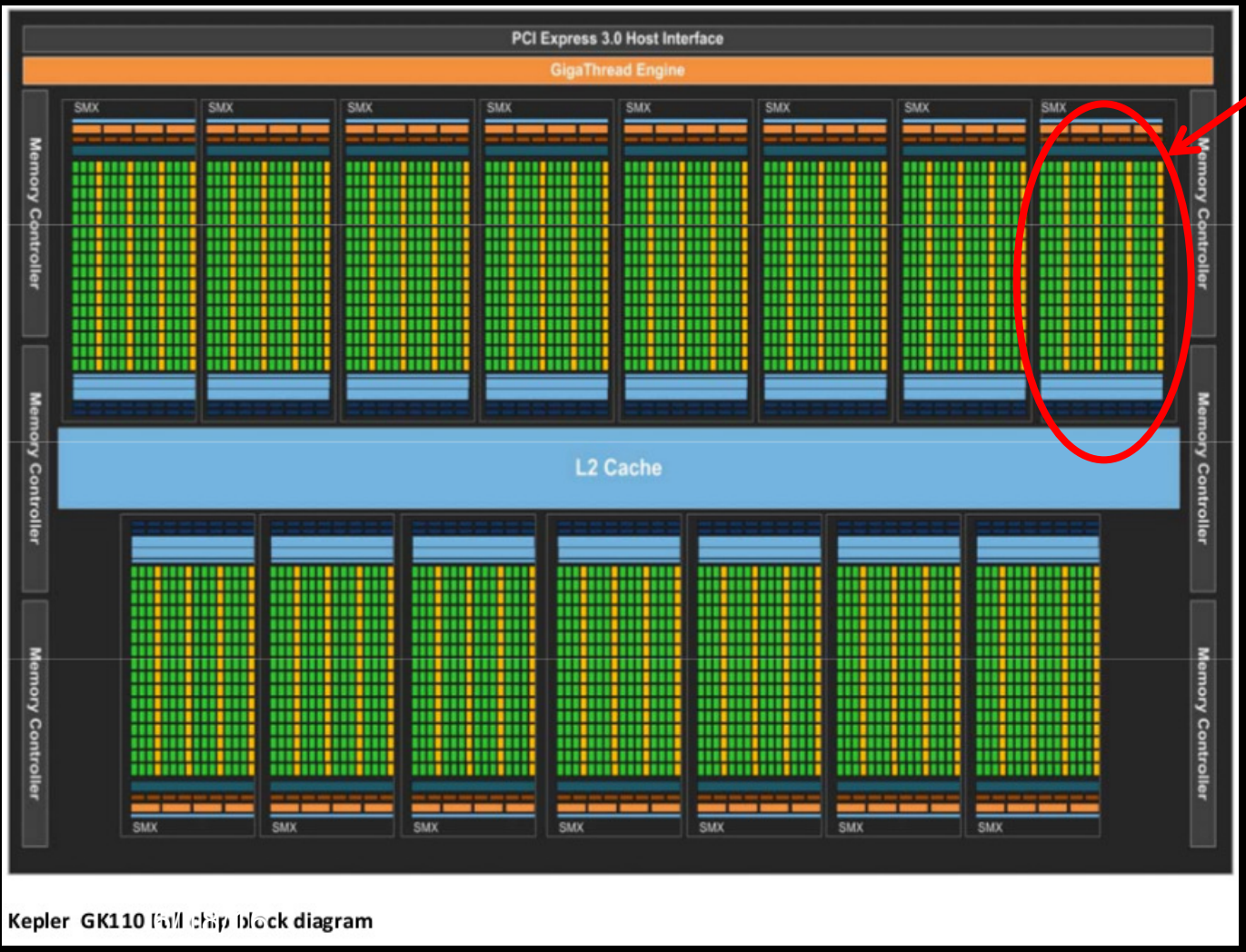
- Two mem[illegible]ssue (host,device)
- CPU medi[illegible]
- Limited [illegible]
- Compile[illegible]gua[illegible]ssues (e.g. CUDA)

Slowly going away over time

Intel Xeon Phi Knights Landing: 72 cores; 512 vectors; 3 TFLOPS



Many-core architecture example: NVIDIA Kepler GPU (K20X) circa 2012



Kepler GK110 Full chip block diagram

SM = Streaming Multiprocessor

192 SIMD CUDA processors per SM

X

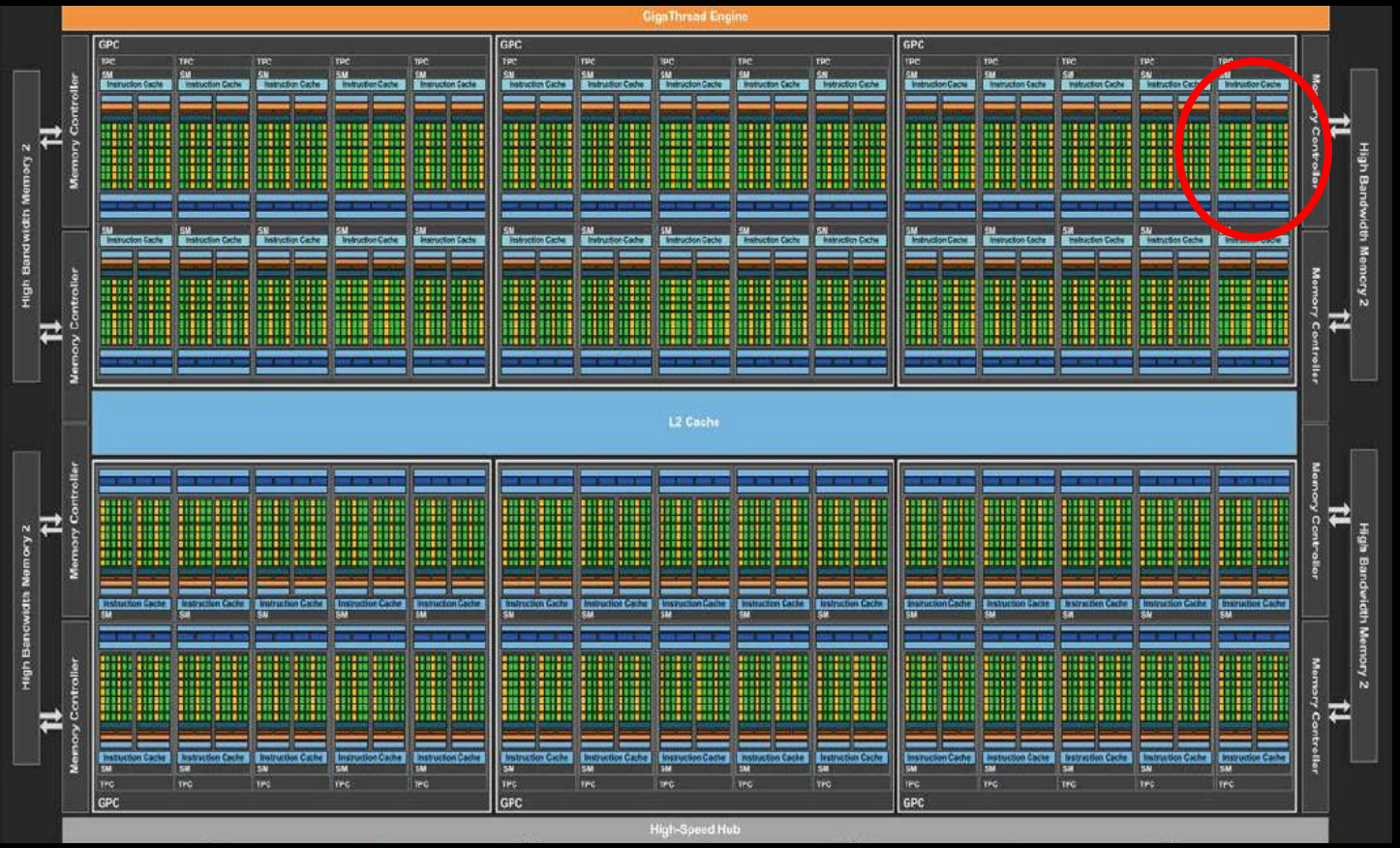
15 SMX pipes

= 2688 cores

K20X memory bandwidth:

= 250 GB/s

Many-core architecture example: NVIDIA Pascal GPU circa 2016



64 SIMD CUDA cores per SM

X

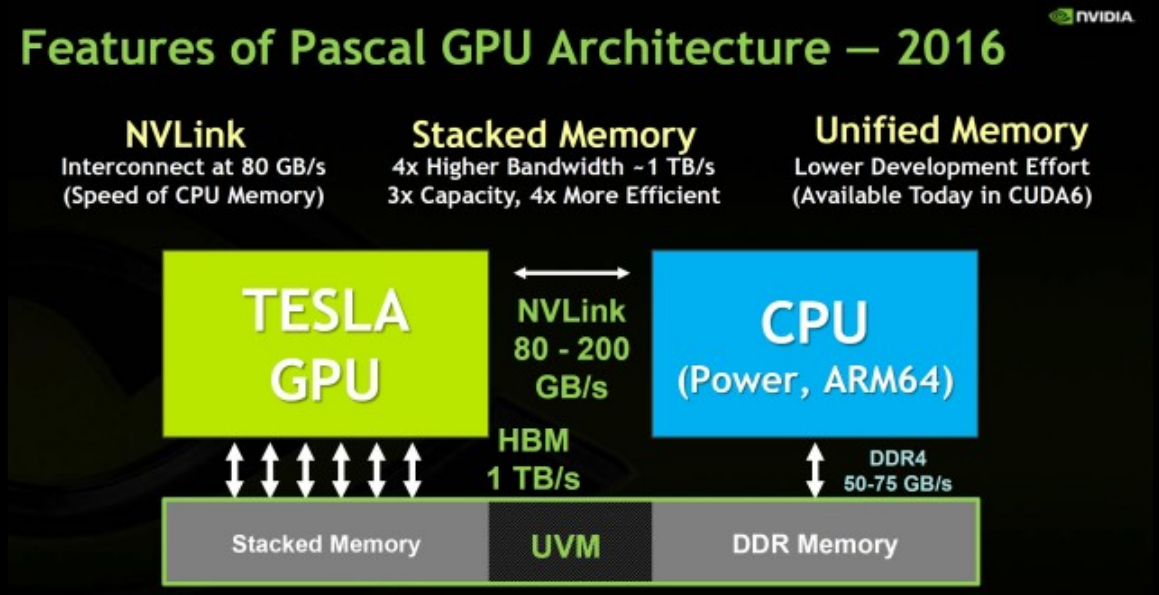
56 SMs pipes

= 3584 cores

stacked memory bandwidth:
= 1000 GB/s

8-GPU DGX-1 architecture with 8 NVIDIA Pascal GPUs and NVLink

- 3584 x 8 = 28,672, 32-bit CUDA cores
- 4 dp teraflops x 8 = 32 dp teraflops/node
- One dp petaflops peak in 32 nodes



Pascal: double, single, ***half*** precision?

- Interestingly, pascal supports fp16 for neural net calculations (>170 TF in a box)
- Question: is fp16 useful for *anything* in weather and climate?
- In a larger context this is an attempt to address the incredible inefficiency of conventional computers relative to biological neural networks (brains!)

Computers vs Brains

>35,000 kg
1,900,000 watts

1.4 kg (3 lbs)
20 watts

Why are supercomputers so bad and being “smart”?

- Supercomputers are designed to be:
  - Very good at mathematical calculations.
  - Remember everything (data integrity)
  - Predictable (reproducibility)
- Comparatively, brains are:
  - Are terrible at math
  - Forget stuff
  - Are generally unpredictable

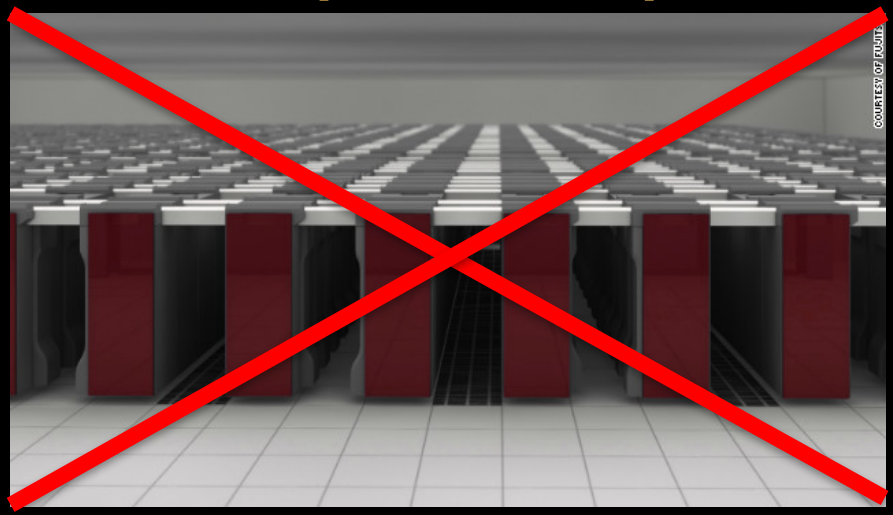
How “smart” is a supercomputer?

House Cat

760 million neurons
10 trillion synapses

OR

K computer in Japan



Simulated 1.7 billion neurons with 10 trillion synapses
But 2400 times slower!

trillion = 1,000,000,000,000 (a million million)

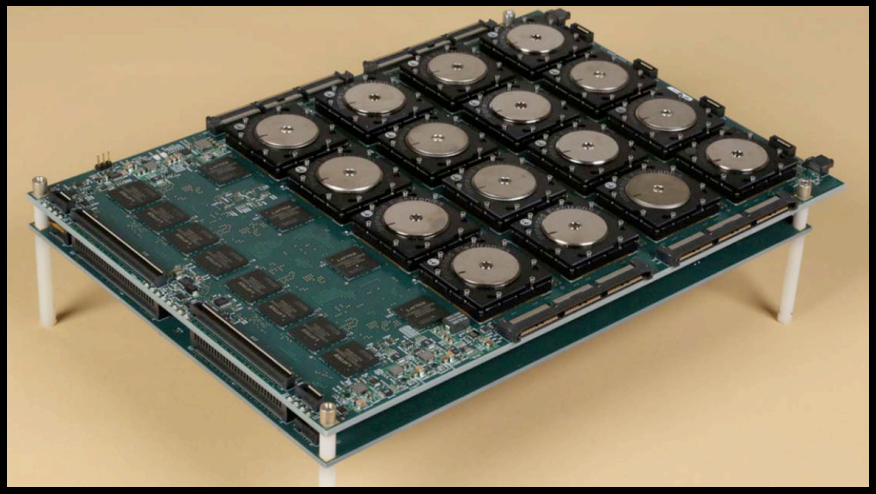
Neuromorphic computers: work like the brain!

Honey Bee

1 million neurons,
1 billion synapses
2 mW

=

IBM TrueNorth Processor (2015)

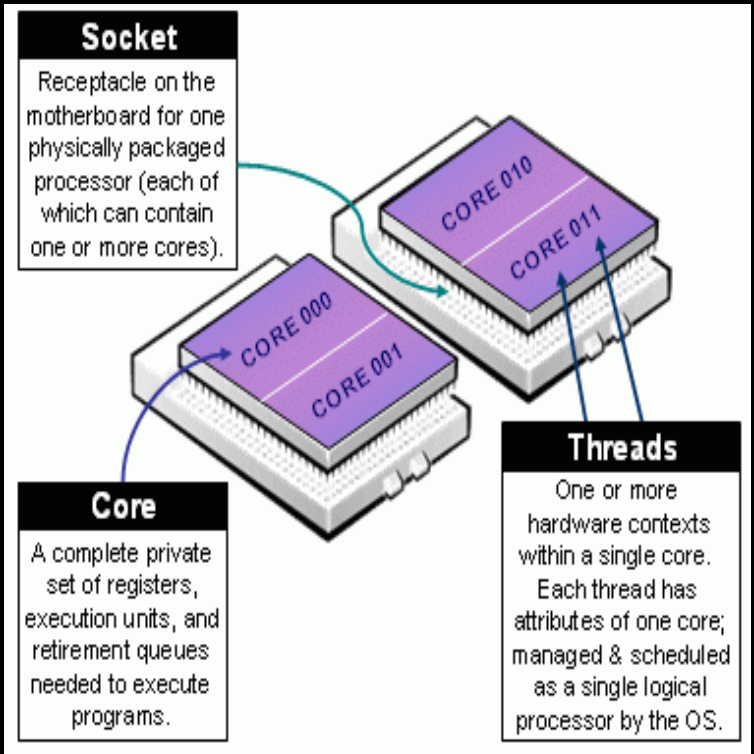


1 million neurons,
268 million synapses
@ 80 mW TrueNorth is
10,000 x more efficient
Than regular computers!

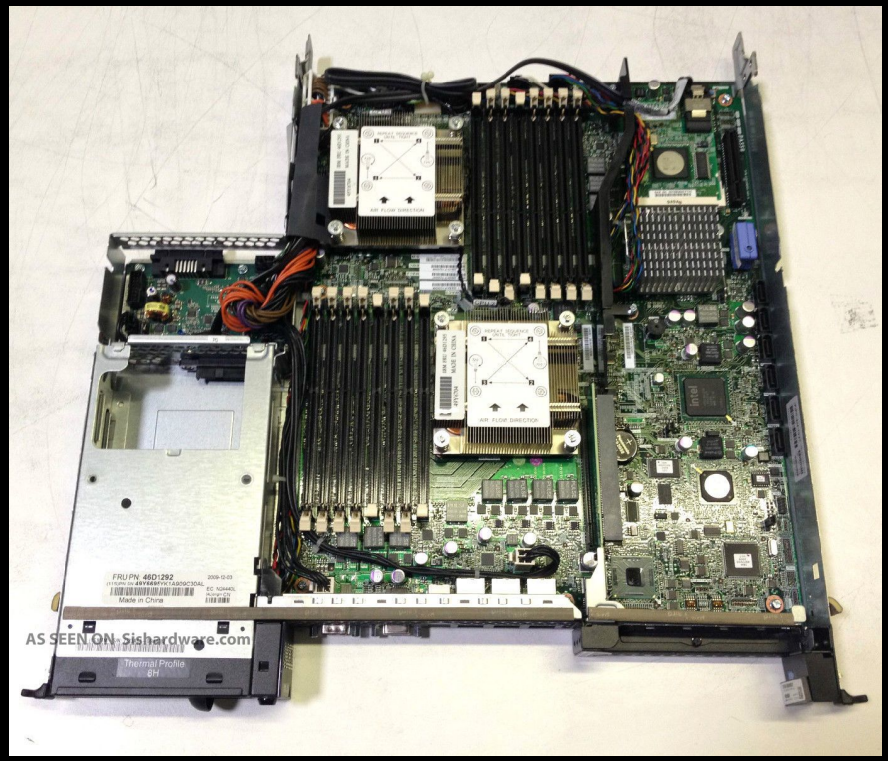
Which is faster?

Memory and Bandwidth Trends

Traditional (Current) multi-core node architecture

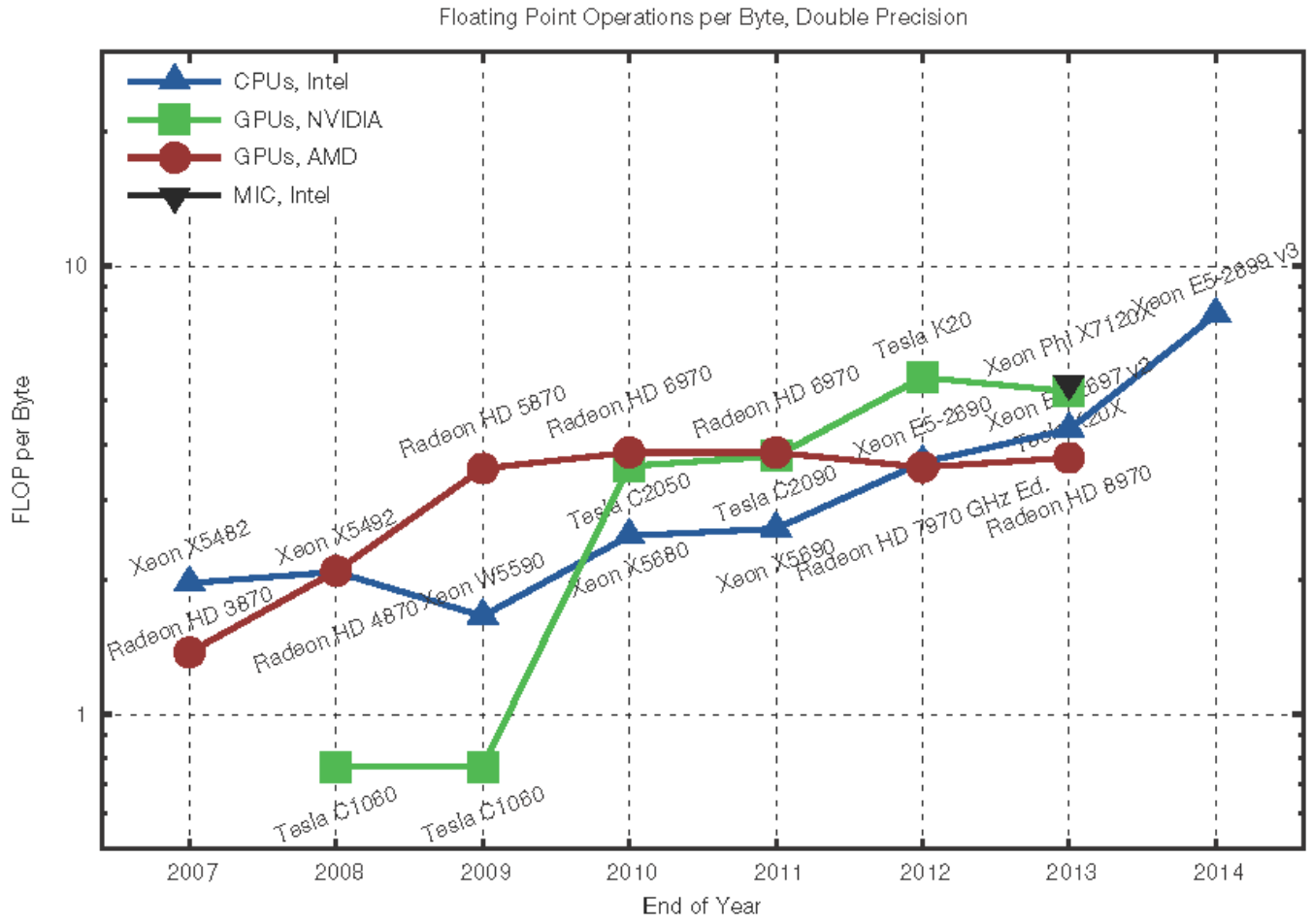


Sockets, cores, threads



Danse hall: memory segregated from processors

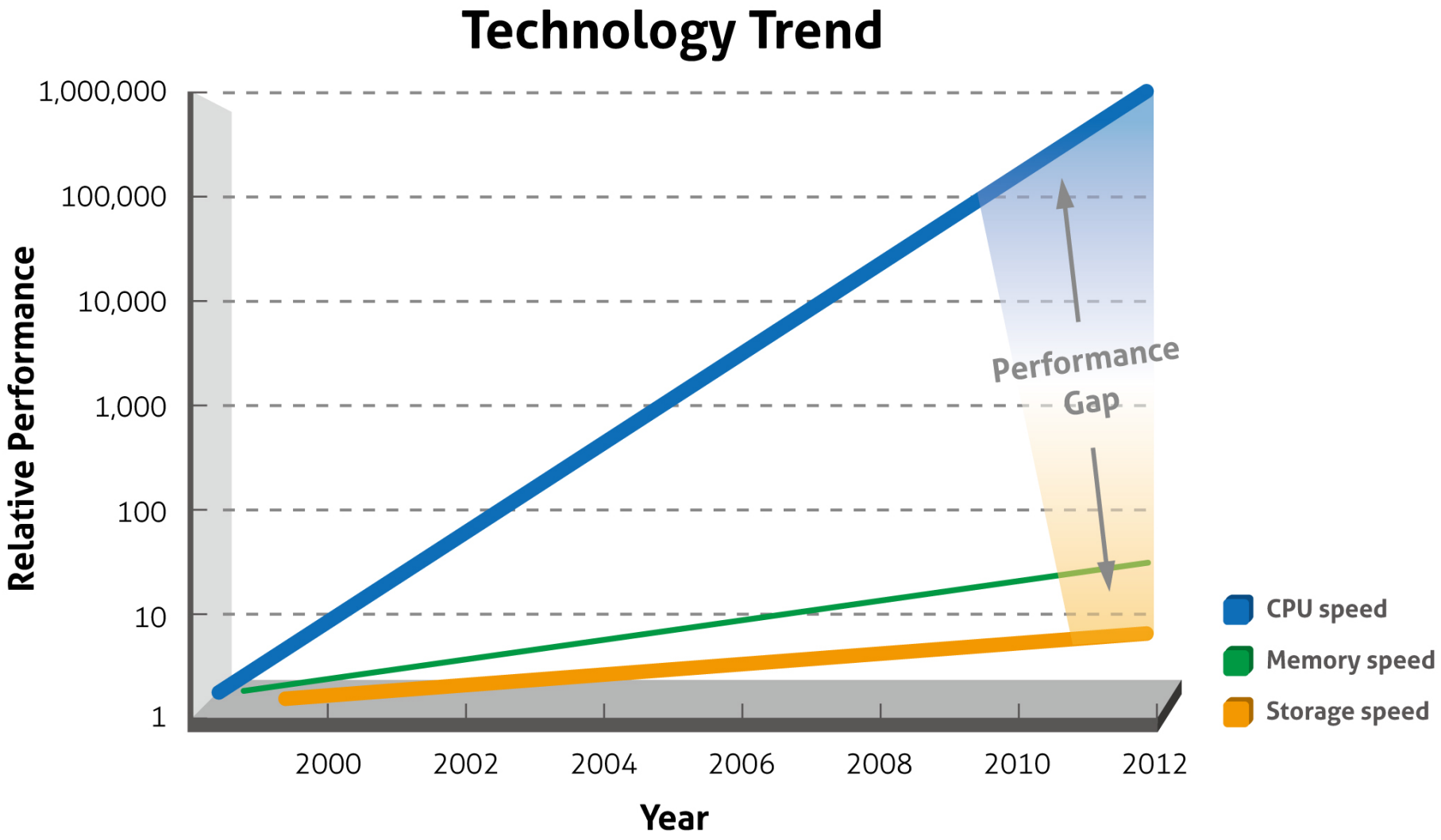
Peak flops is driving top 10 system design



[c/o Karl Rupp]

Slide courtesy of Jed Brown, CU

the performance gap



Power and Memory Bandwidth

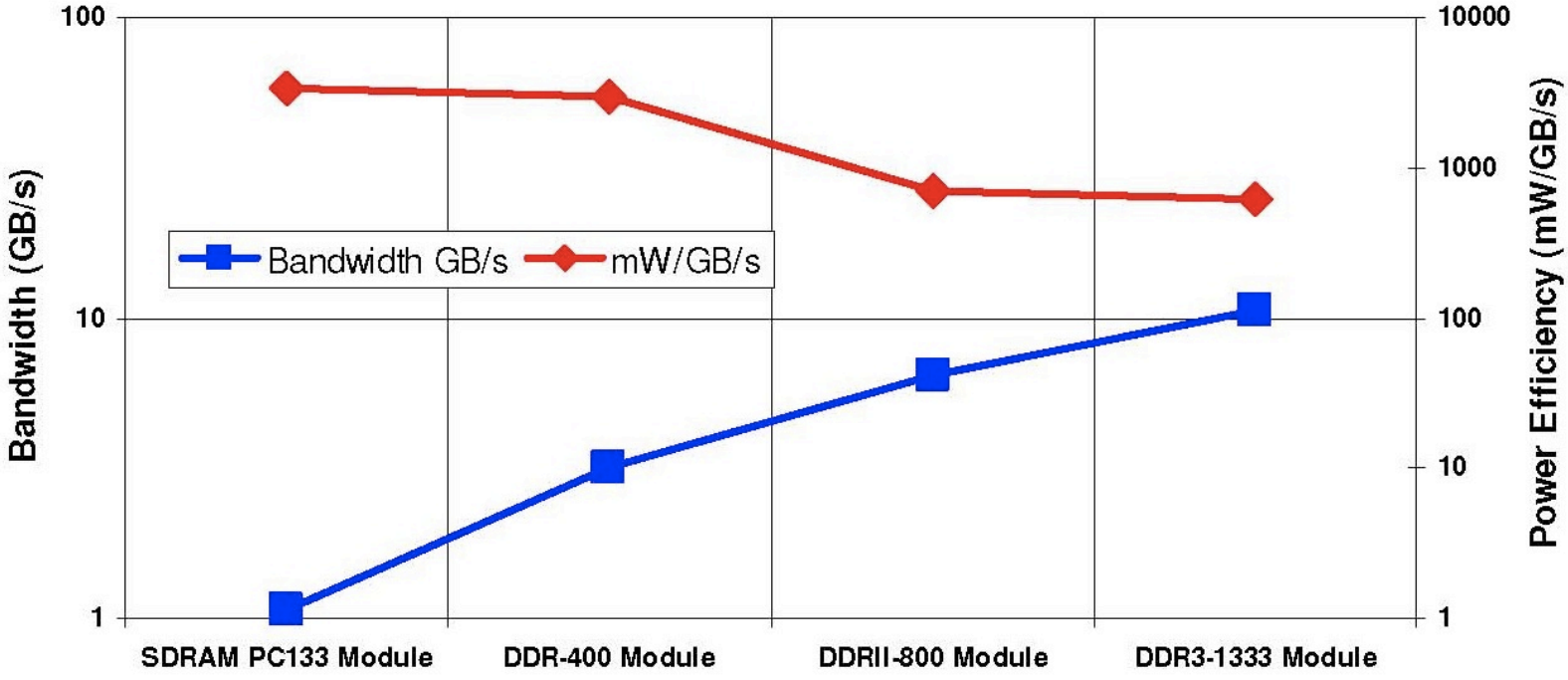


Figure 6.22: Commodity DRAM module power efficiency as a function of bandwidth.

O(1) watt/GB/s is relatively constant over time

Memory Bandwidth means power: consider an exaflops system

- Assume one, 8-byte word loaded from memory per flop
- To sustain 1% of peak, i.e. O(10) PFLOPS, we will need 80 PB/sec
- Using the DDR figure of merit, we the memory we need is ~80 Megawatts.
- This suggests a *new technology* is required to achieve a practical exascale capability.

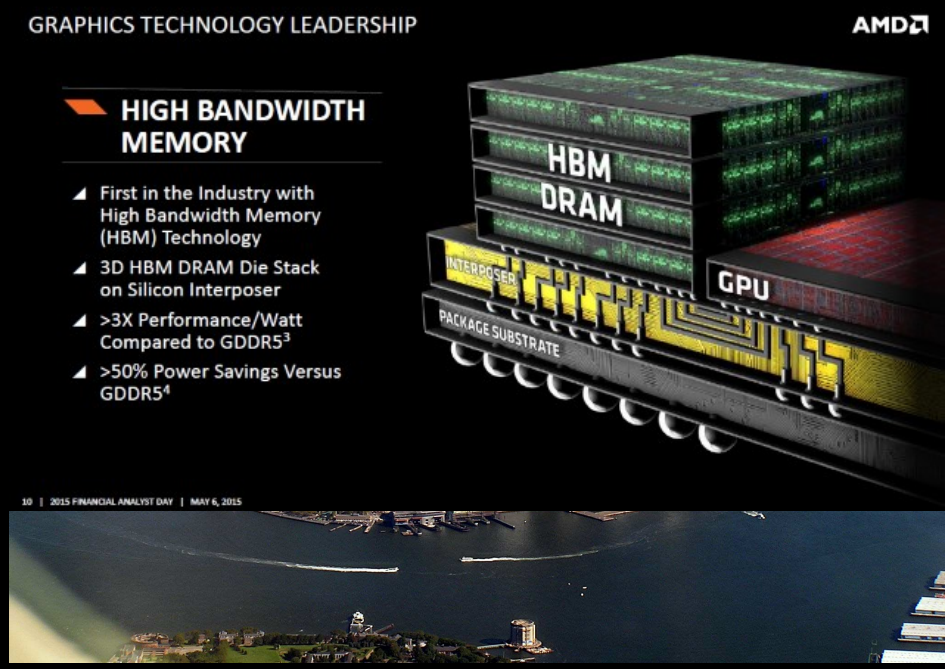
Off chip bandwidth limitations

- Off chip bandwidth = pin count x signaling rate
  - Pin count might increase at 5% per year
- Upping electrical signaling rate increases power consumption.
  - Intel considers a 10x improvement a grand challenge

Will we be trapped on a computational island?

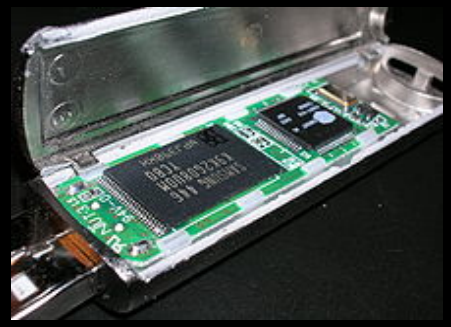
Going vertical: fuel-efficient memory bandwidth through 3-D Chips

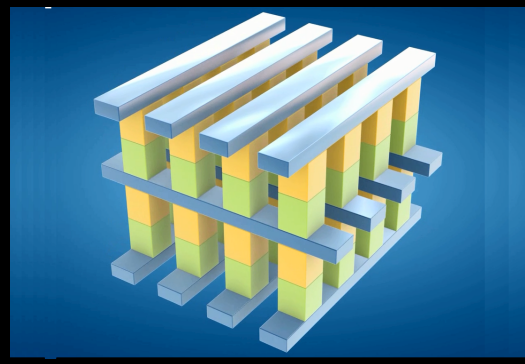
- Advantages
  - Increased density
  - Shorter wires
    - lower capacitance
    - lower power
    - higher GB/sec/Watt
- Disadvantages
  - Tighter thermal constraints on memory
  - Limits to capacity?



So-long disk? Non-volatile memory and NAND technologies

- Flash memory-based storage
  - Floating gate or NAND transistors
  - Limitations: density and endurance
  - Great for high IOPS/disk latency sensitive applications
- Nonvolatile storage technologies
  - 1000x faster than NAND
  - 1000x the endurance of NAND
  - 10x denser than DRAM



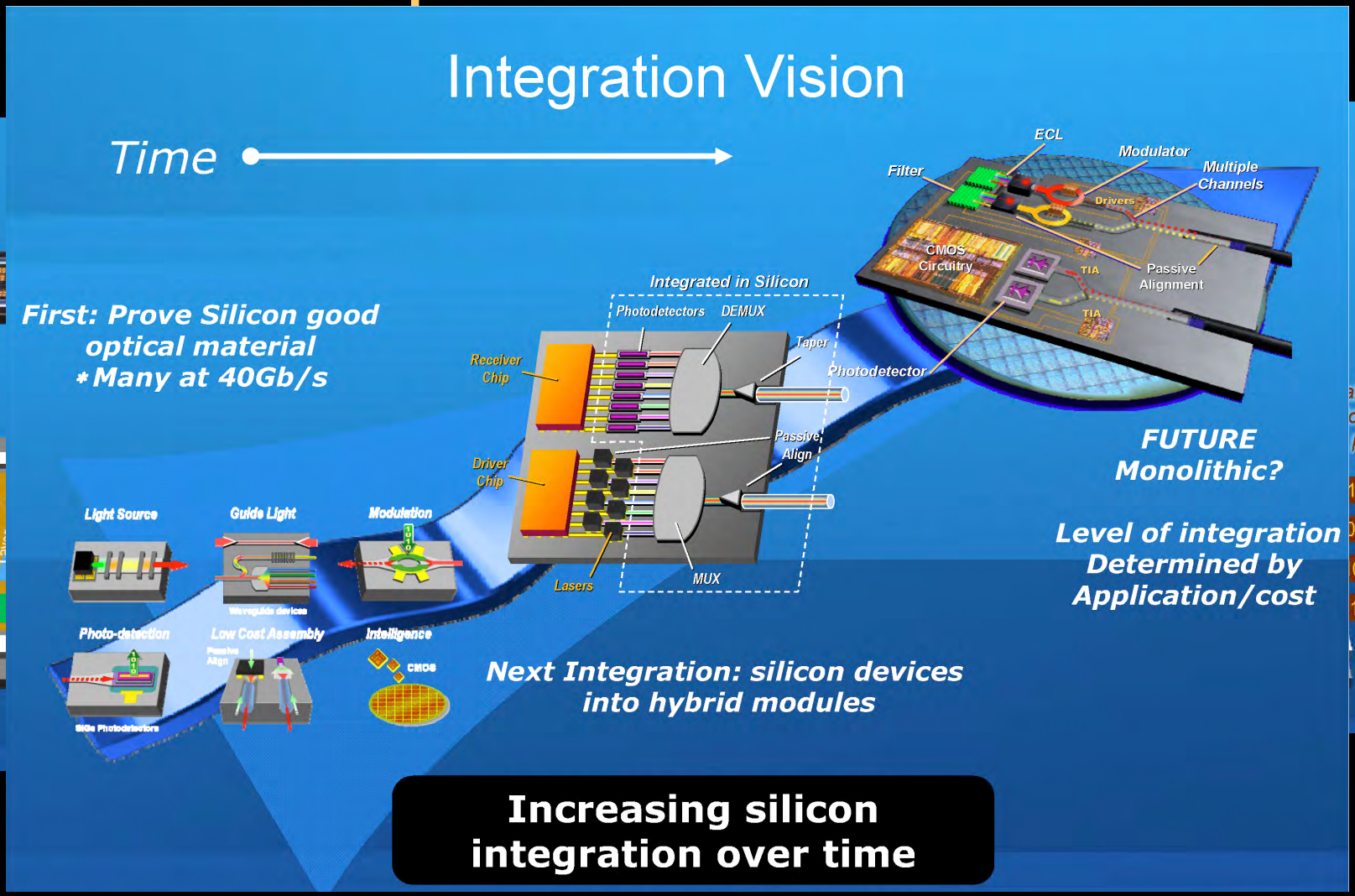


3D XPoint

Hybrid electro-optical systems

- Lots of bandwidth available in an optical fiber
- Parallelism: data can be carried on different wavelengths.
- Coupling optics to silicon is undergoing rapid development
  - Optically active silicon
  - Miniaturization of optical devices

Trends in micron-scale opto-electronics



Credit - Intel: Antuino, et al. Hot Chips 2010

Impacts on Atmospheric Models

Parallel programming models

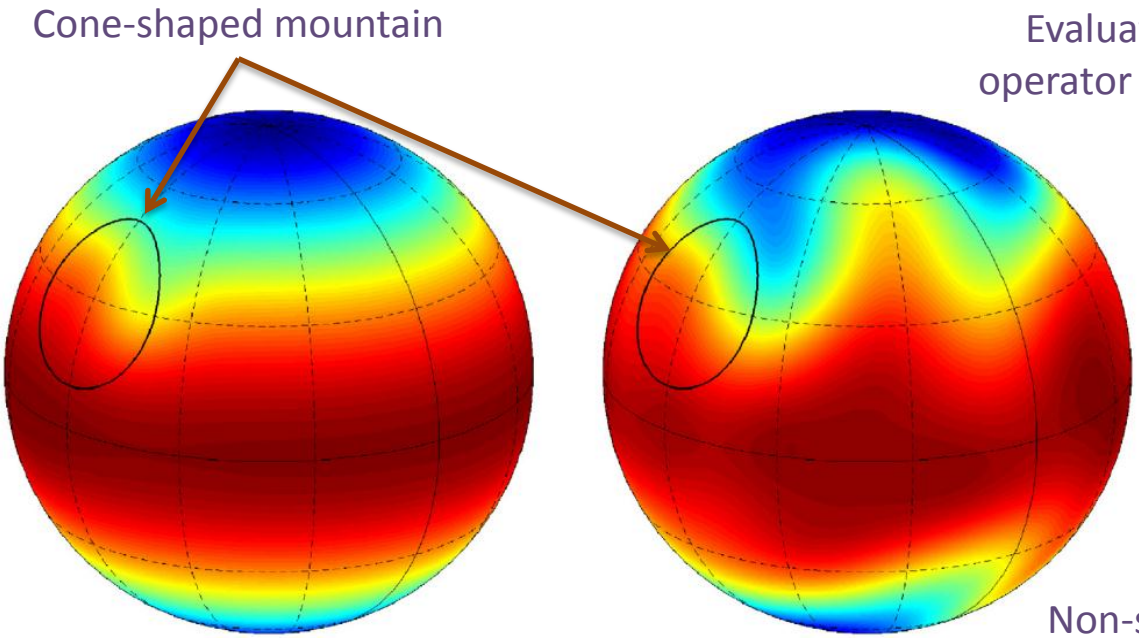
- MPI – distributed memory programming
  - Co-Array Fortran
- OpenMP – directive-based shared memory
  - CPU moving towards GPU support
- GPU programming models:
  - CUDA (NVIDIA's SIMD language)
  - OpenCL (SIMD language)
    - In theory portable to CPUs and GPUs
  - OpenACC (directive-based)
    - In theory portable to CPUs and GPUs

CPU-GPU intercomparison

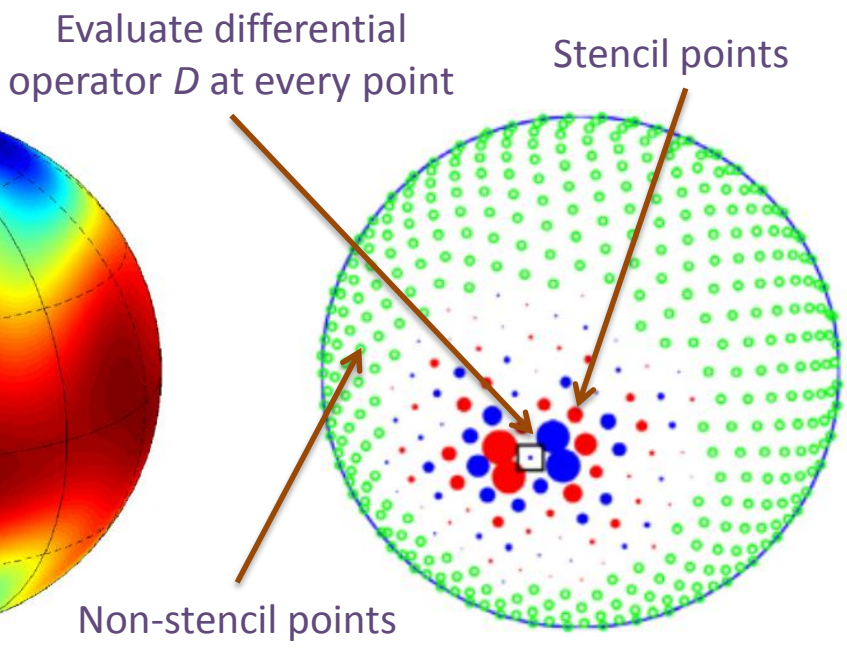
- Relevant algorithmic characteristics
  - Available parallelism
  - "SIMD-ness"
  - Computational intensity (flops/memory accesses)
  - Workload imbalances/conditionals
  - Communication overhead
- Other metrics of performance
  - Hardware cost-performance
  - Energy-performance (watts/flops)
  - Software cost-performance ($/SLOC optimized)

Which is Faster?
Benchmark Problem

- Shallow Water Equations (SWE)
  - A set of non-linear partial differential equations (PDE)
  - Capture features of atmospheric flow around the Earth
- Radial basis function-generated finite difference (RBF-FD) methods

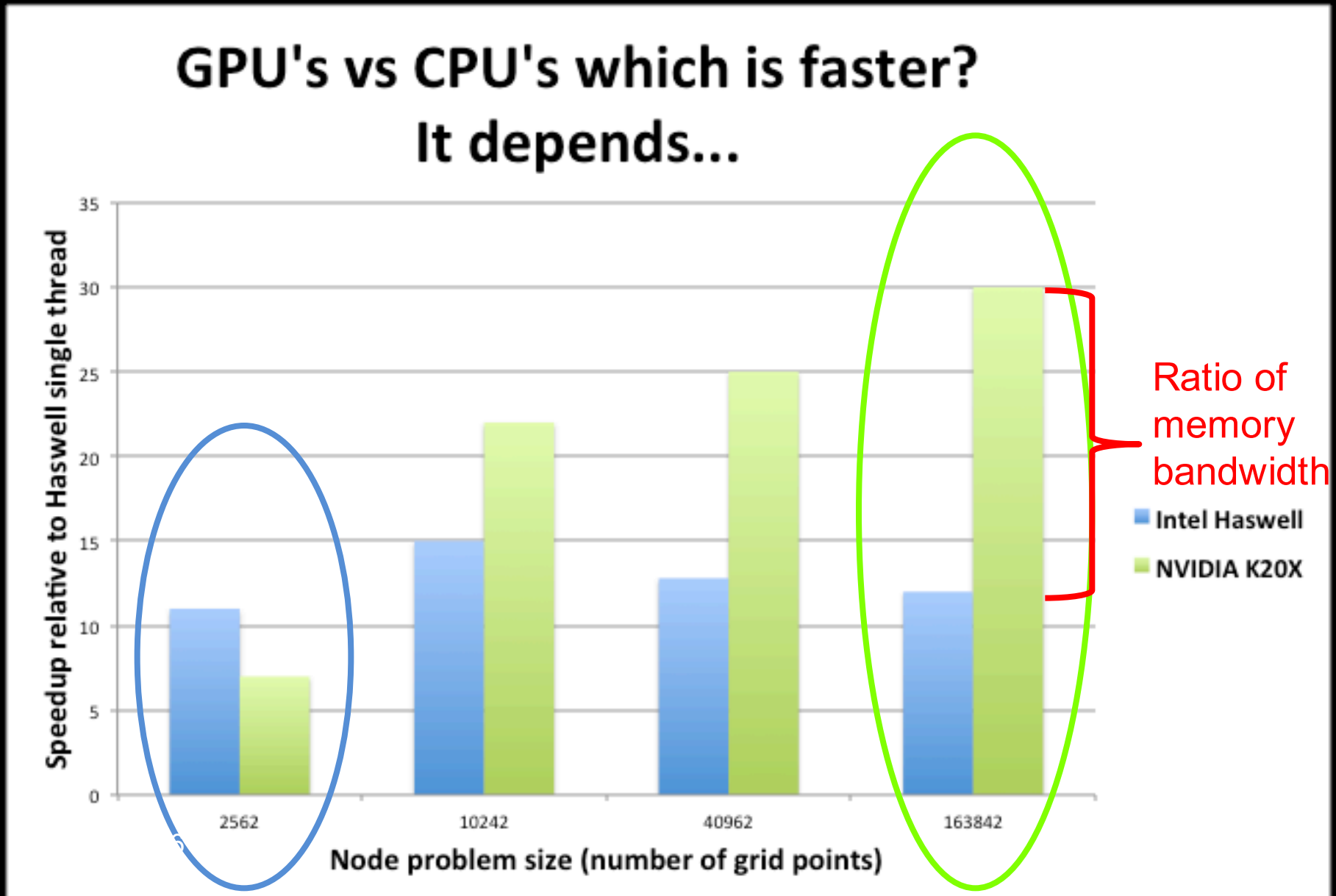


RBF-FD solution to SWE test case "Flow over an isolated mountain" using 655,532 points [1]



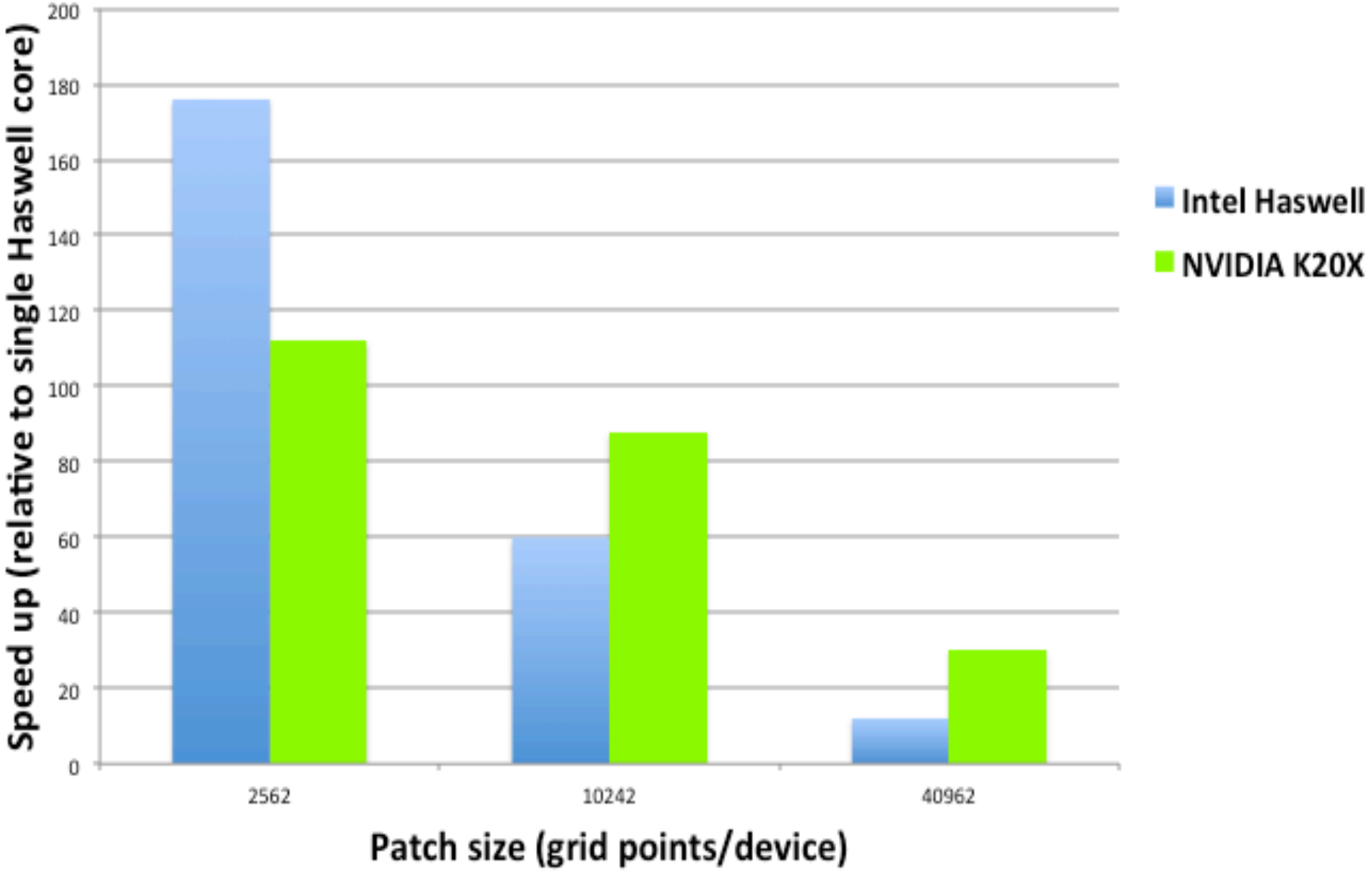
An example of 75-point stencil on a sphere [1]

Test case: shallow water equations, unstructured radial basis function grid, 31 point stencil



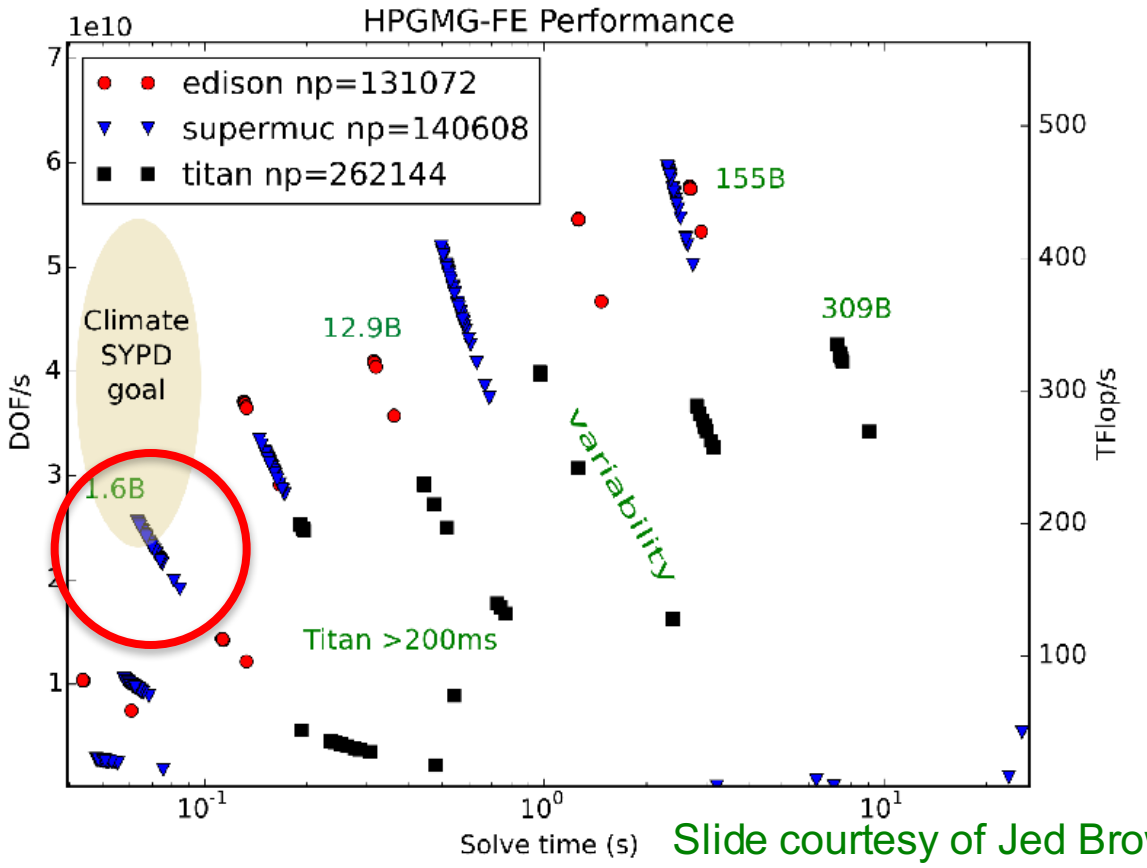
Multi-device performance: CPU vs GPU

note: projected ideal case (no comms; load imbalance)



Finite Element Benchmark Comparison of Large Systems

Scaling regime: HPGMG-FE on Edison, SuperMUC, Titan



Slide courtesy of Jed Brown, CU

Thanks! Question?

Dr. Richard Loft
NCAR
loft@ucar.edu
Feb 4, 2016

Image the size of supercomputer needed to match human brain!

Human Being

86 billion neurons
150 trillion synapses

~100,000 x honey bee
~20 x cat

=

Computer X?

1.7 million times faster than Yellowstone!

Cirrascale GX-8

8-way GPU cluster in a box