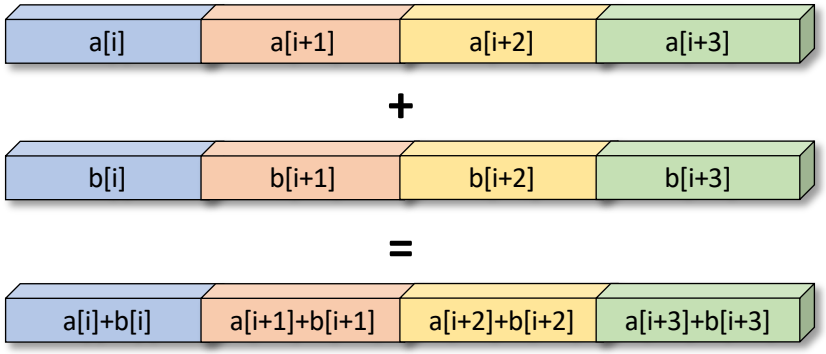
What Every Computational Researcher Should Know About Computer Architecture*



*Plus some other useful technology

Ian A. Cosden, Ph.D.
icosden@princeton.edu
Manager, HPC Software Engineering and Performance Tuning
Research Computing, Princeton University

RC Bootcamp
October 31, 2019

Outline

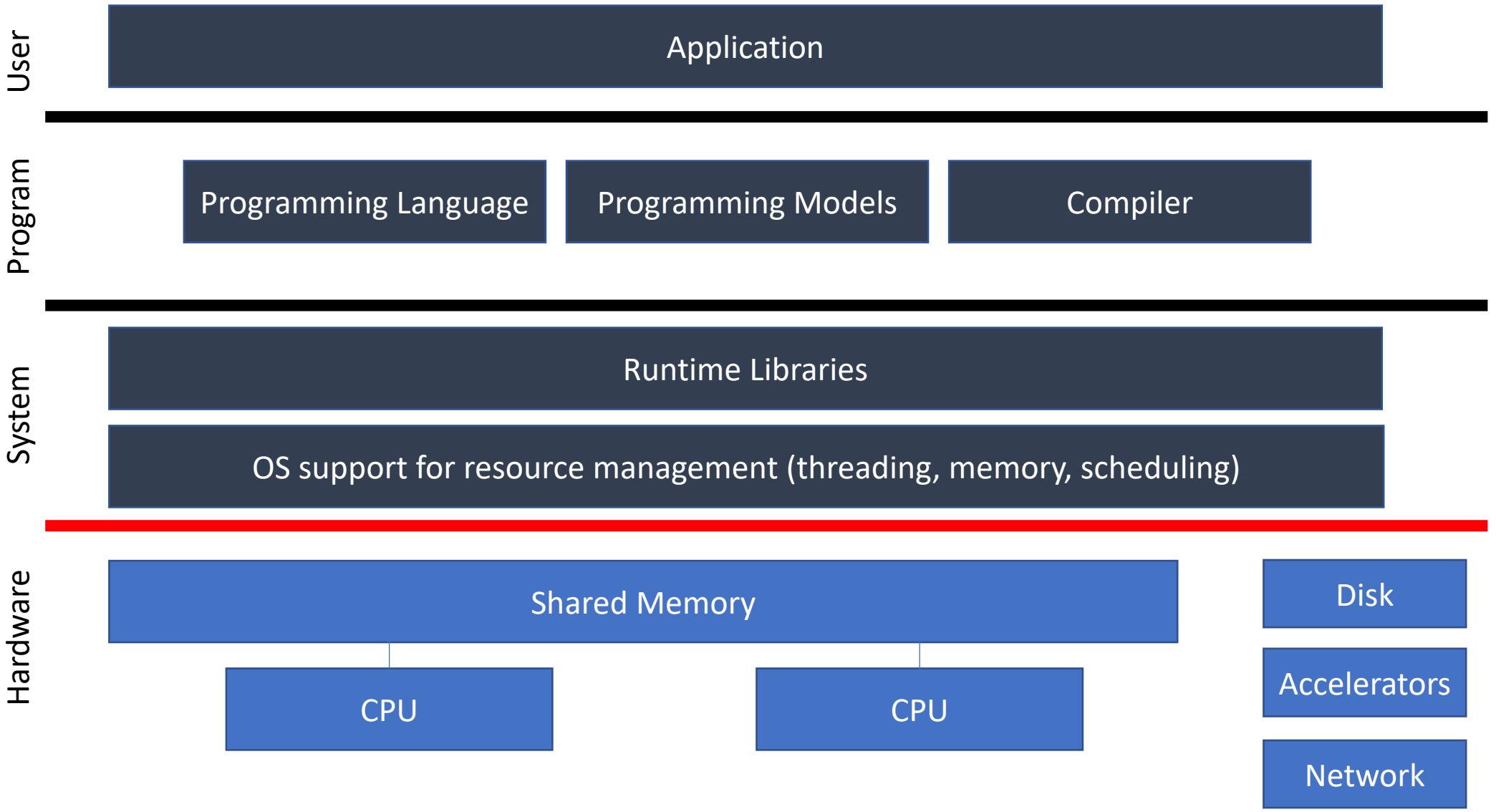
- Motivation
- HPC Cluster
- Compute
- Memory
- Disk
- Emerging Architectures

Slides are here: https://github.com/cosden/IntroCompArch

Why Care About the Hardware?

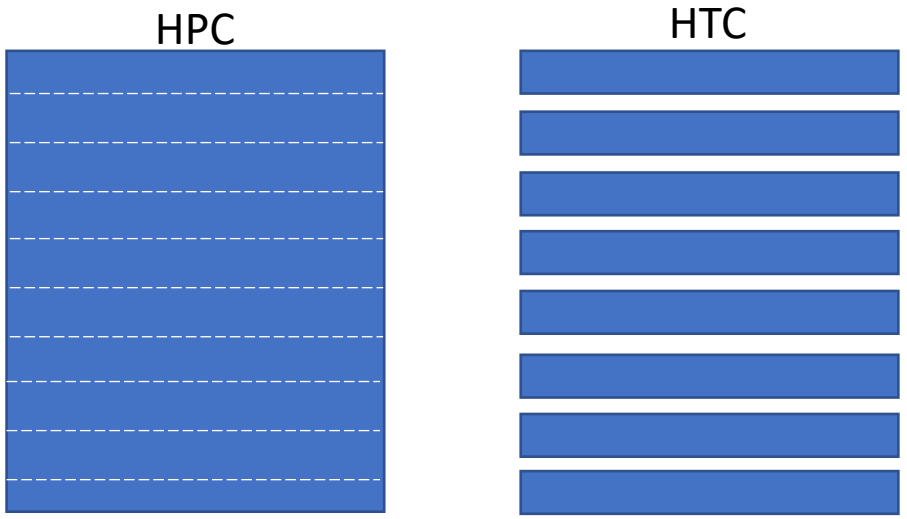
- Get the most out of this week
- Be able to ask better questions and understand the answers
- Architecture changes can force new design and implementation strategies
  - Prevent programming mistakes
  - Make educated decisions
- Exploit the hardware to your advantage (performance)

Hardware is the Foundation



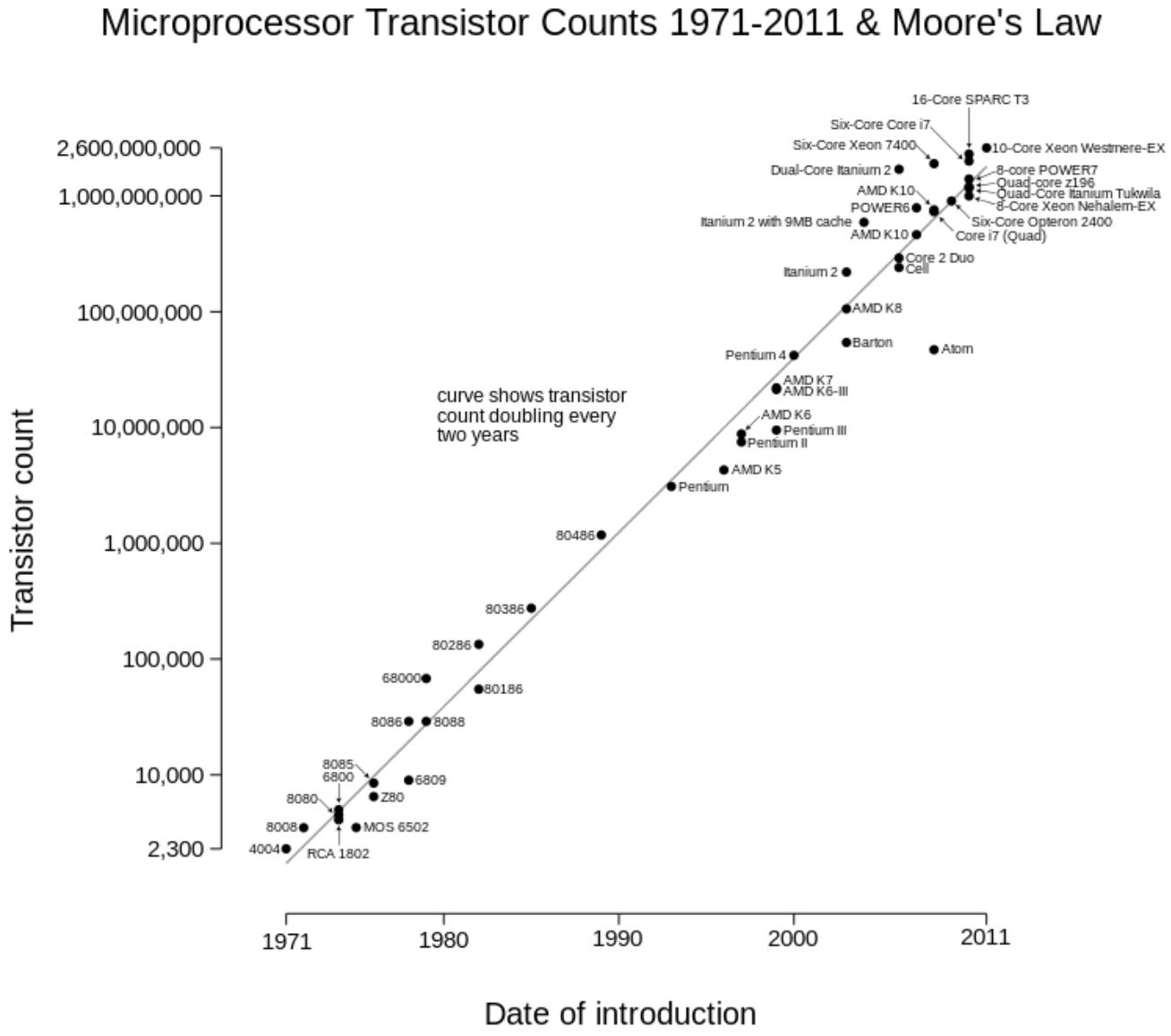
Why Care About HPC Hardware?

- High-Performance Computing (HPC)
  - Aggregate computing power in order to deliver far more capacity than a single computer
  - Supercomputers
- Some problems demand it from the onset
- Many problems evolve to need it
  - When you outgrow your laptop, desktop, departmental server
  - Need to know capabilities and limitations
  - Lots of High-Throughput Computing (HTC) jobs run on HPC clusters

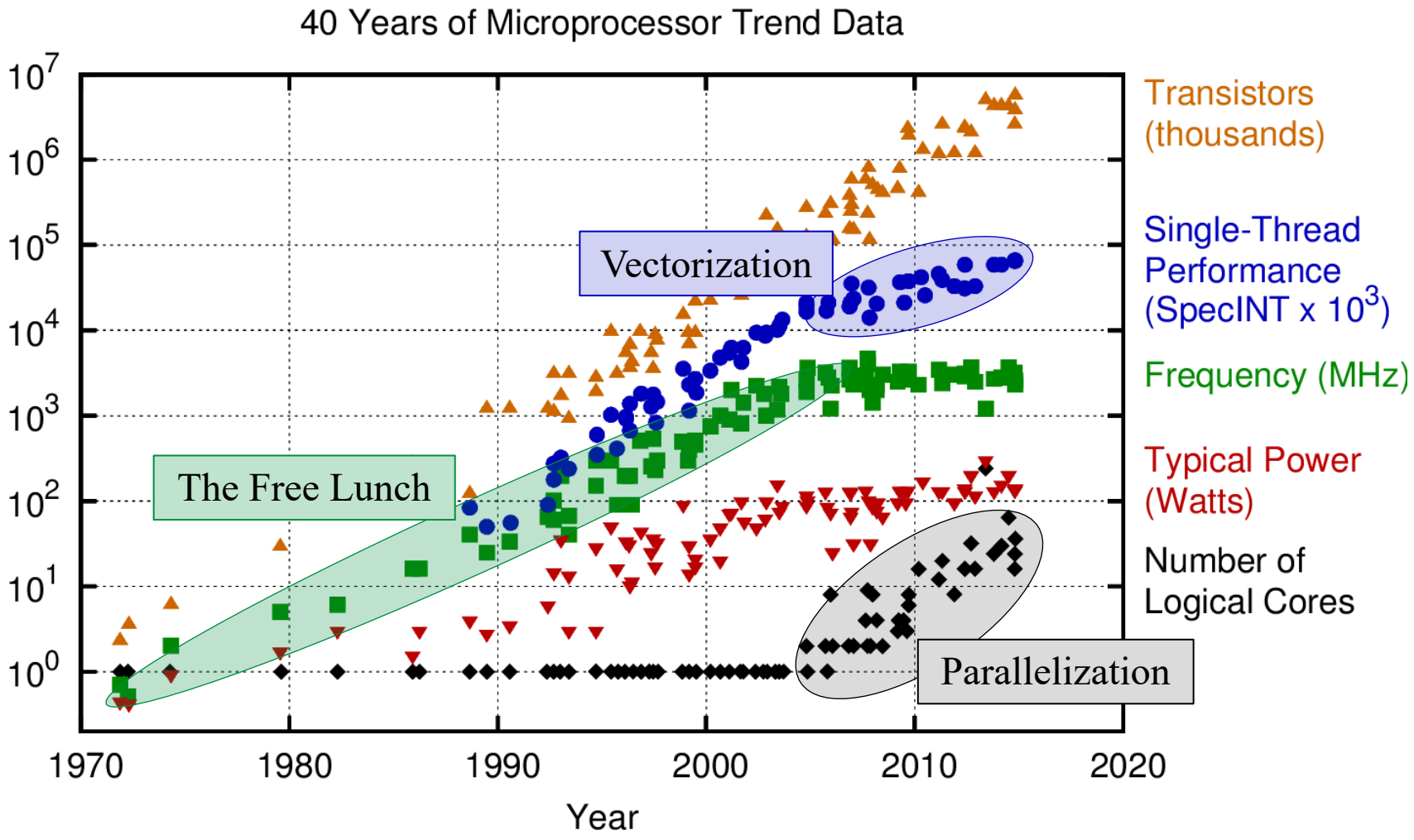


Moore's Law

- Number of transistors doubles ever 18-24 months
- More transistors = better, right?



The Free Lunch is Over

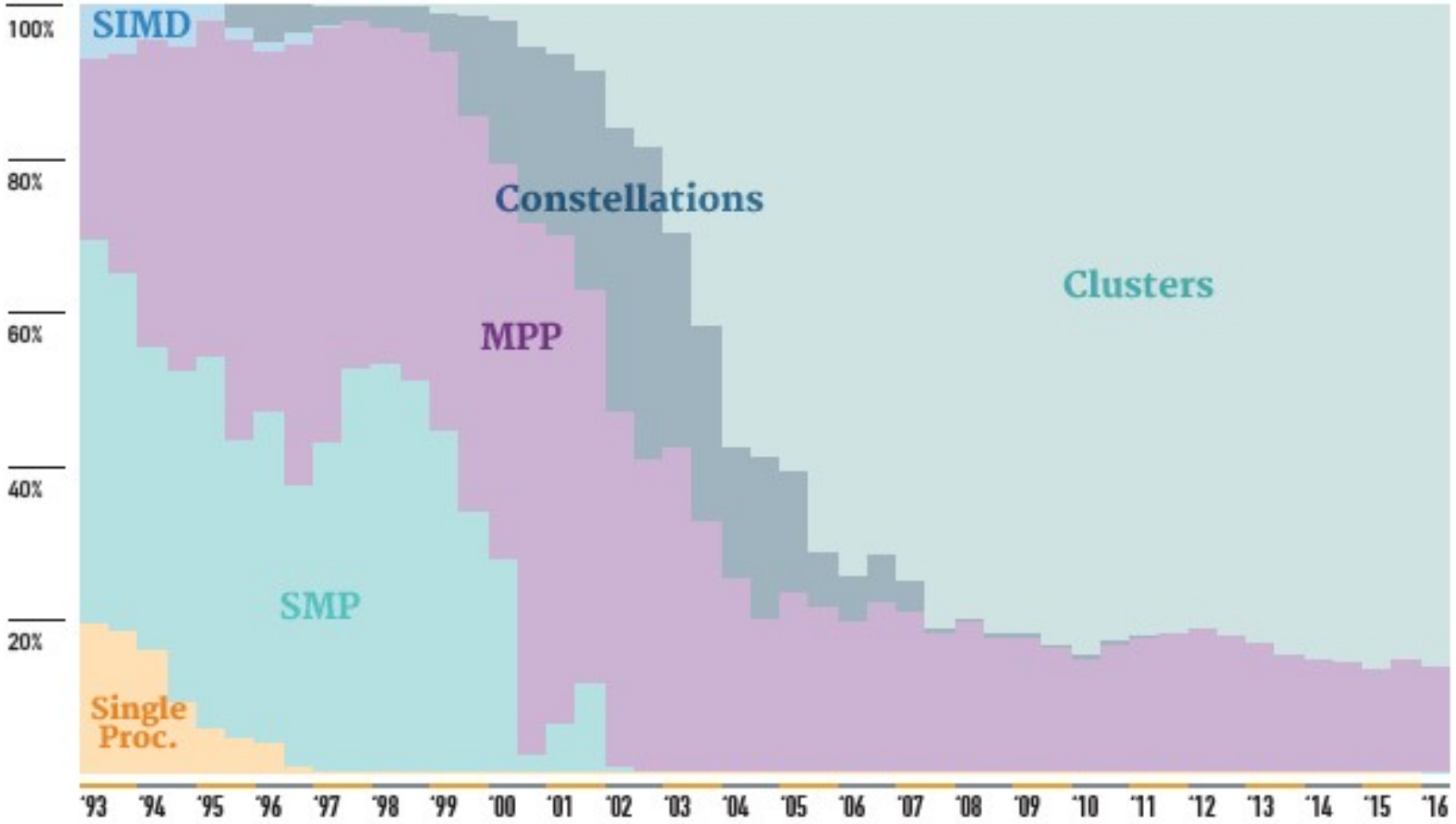


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

https://www.karlrupp.net/2015/06/40-years-of-microprocessor-trend-data/

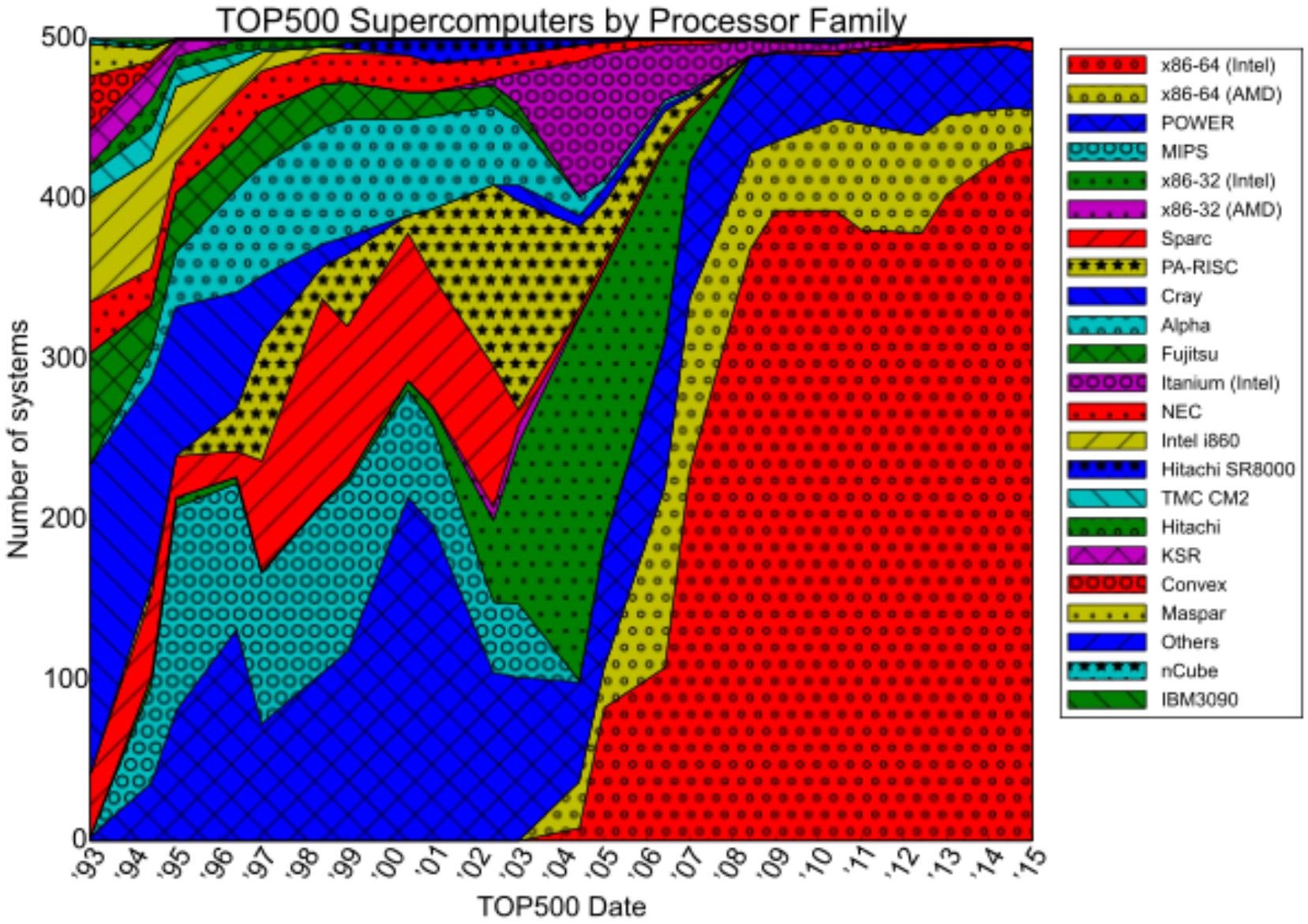
Top 500 Supercomputers

https://www.top500.org



https://www.nextplatform.com/2016/06/20/china-topples-united-states-top-supercomputer-user/

Why Talk About Intel Architecture?



https://www.revolvy.com/page/TOP500

Outline

- Motivation
- HPC Cluster
- Compute
- Memory
- Disk
- Emerging Architectures

Zoom In On A Cluster

- A cluster is just *connected* collection of computers called nodes
- A single-unconnected node is usually called a server
- Entire clusters and stand-alone servers are sometimes referred to as machines

Parallel Computers

HPC Cluster

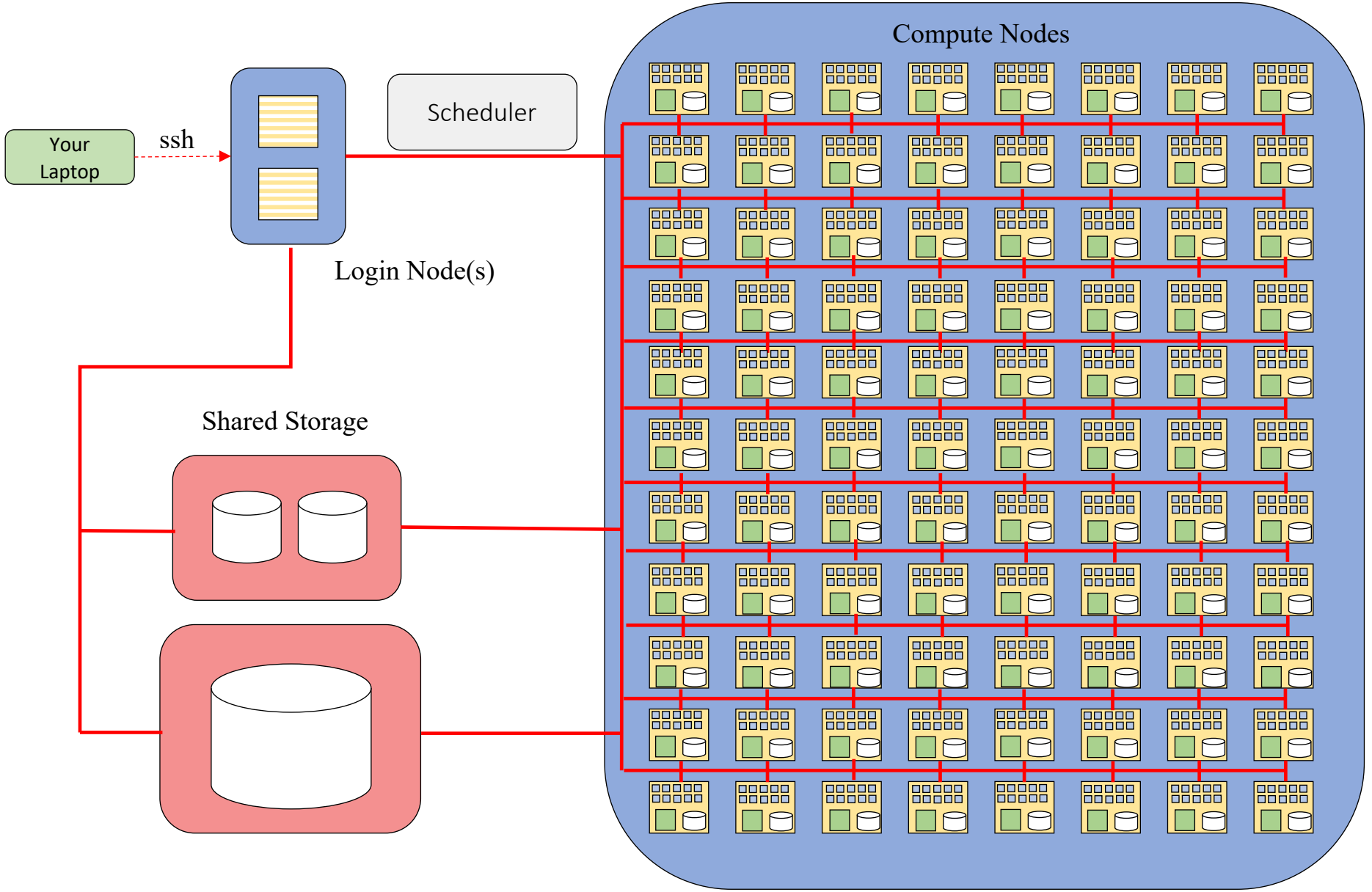
- A collection of servers connected to form a single entity
    - Each is essentially a self-contained computer
    - OS with CPU, RAM, hard drive, etc.
- Stored in racks in dedicated machine rooms
- Connected together via (low latency) interconnects
    - Ethernet < Infiniband, OPA (Intel's Omi-Path Architecture)
- Connected to storage
- Vast majority of HPC systems

SMP System

- Symmetric Multi-Processing
- CPUs all share memory – essentially one computer
- Expensive and serve unique purpose
    - Huge memory and huge OpenMP jobs

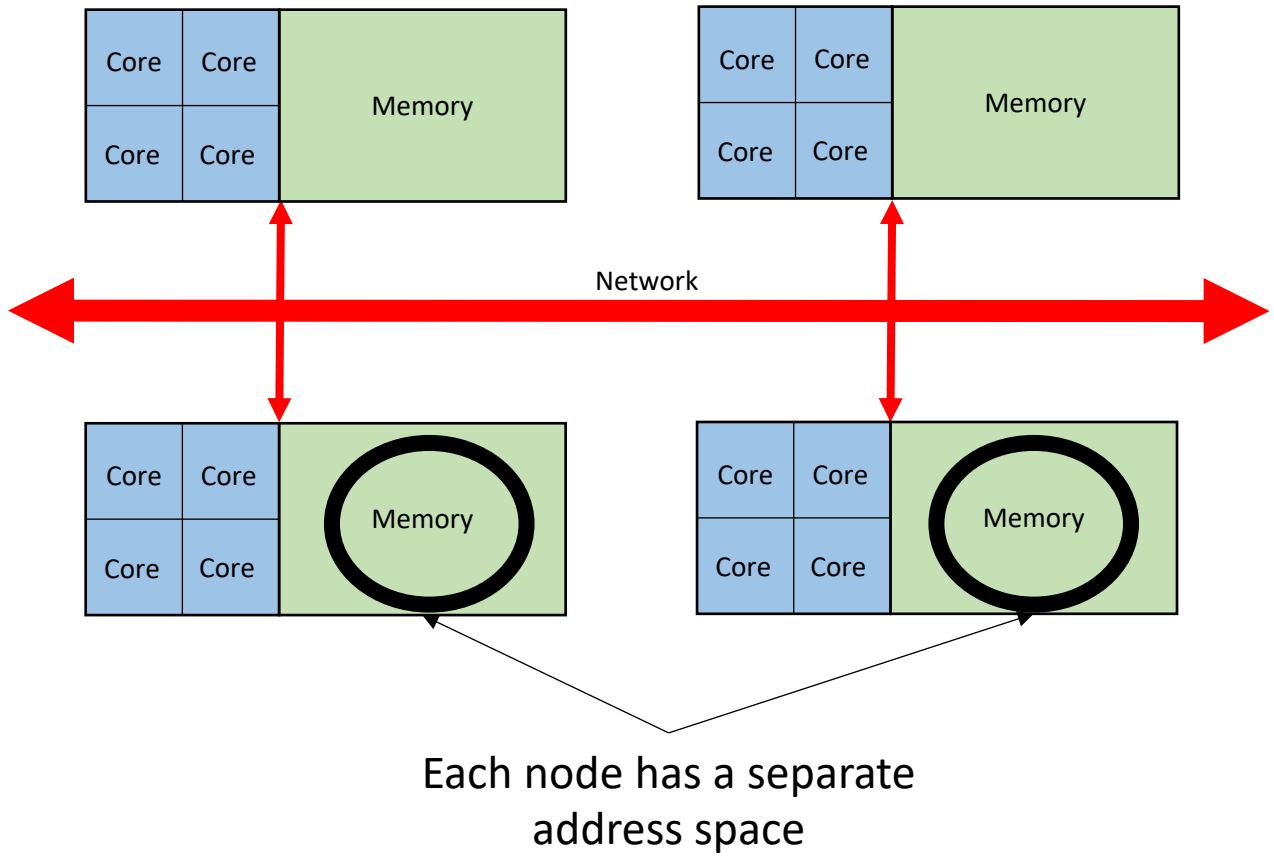
Princeton's Della Cluster

Basic Cluster Layout

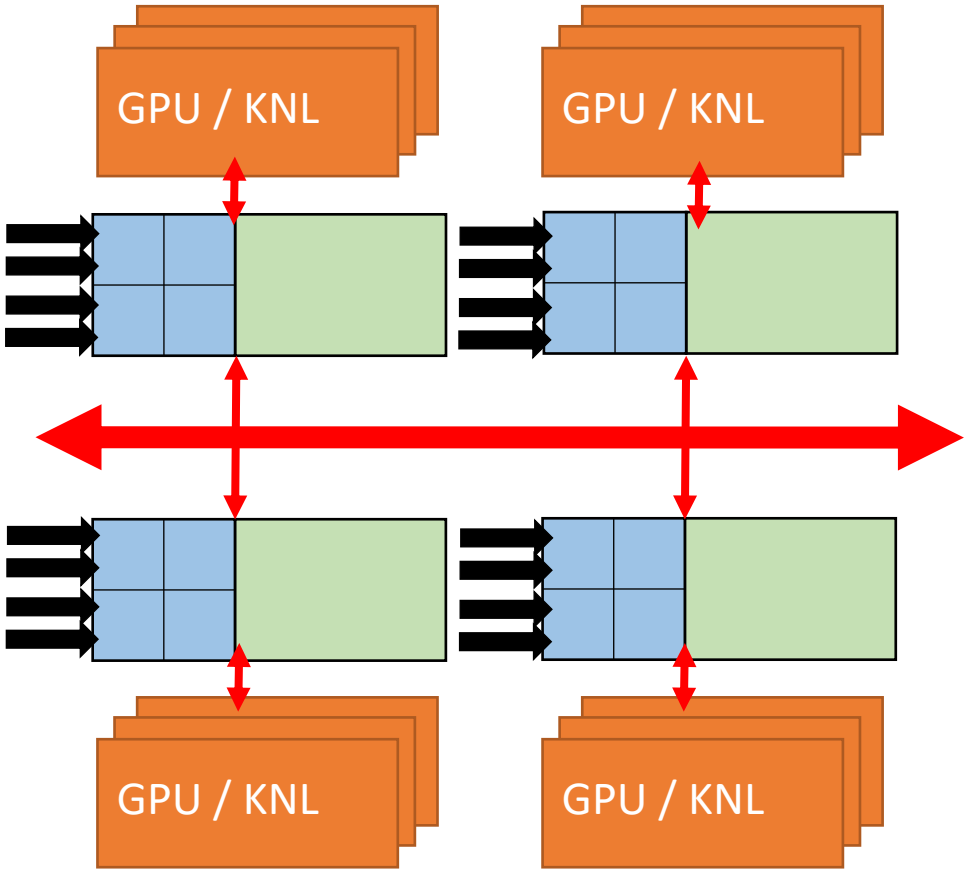


Nodes Are Discrete Systems

- Memory address space
  - Range of unique addresses
- Separate for each node
- Can’t just access all memory in cluster



Clusters With Accelerators



- Accelerators: GPUs or Xeon Phi (KNC, KNL)
- Programmable with MPI + x
  - Where x = OpenMP, OpenACC, CUDA, …
  - Or x = OpenMP + CUDA, …
- Increase computational power
  - Increase FLOPS / Watt
- Top500.org
  - Shift toward systems with accelerators a few years ago

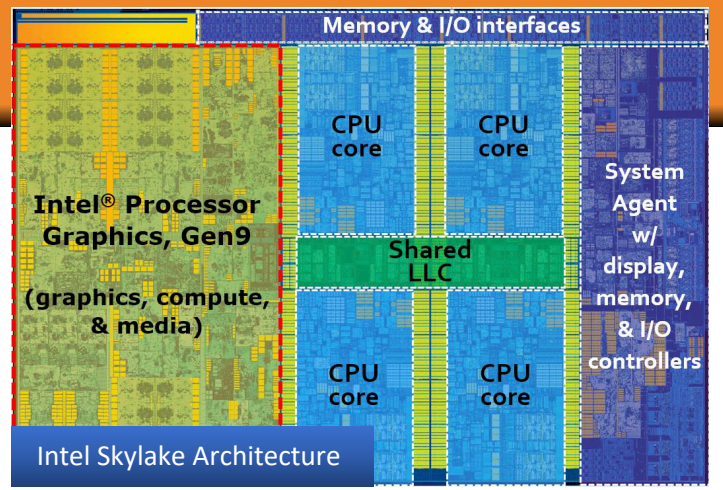
Outline

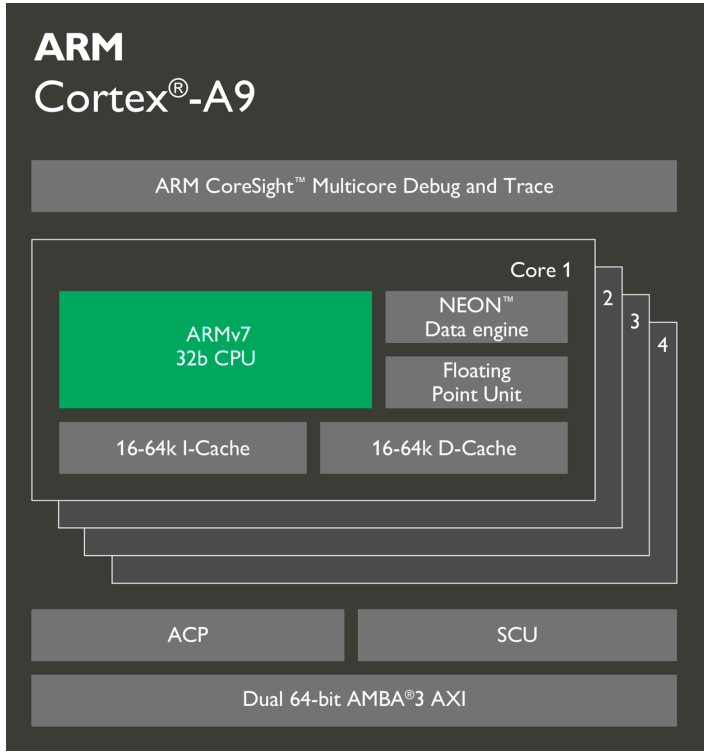
- Motivation
- HPC Cluster
- Compute
- Memory
- Disk
- Emerging Architectures

Multicore Processors

- (Most) modern processors are multicores
  - Intel Xeon
  - IBM Power
  - AMD Opteron
  - ARM processors
  - Mobile processors

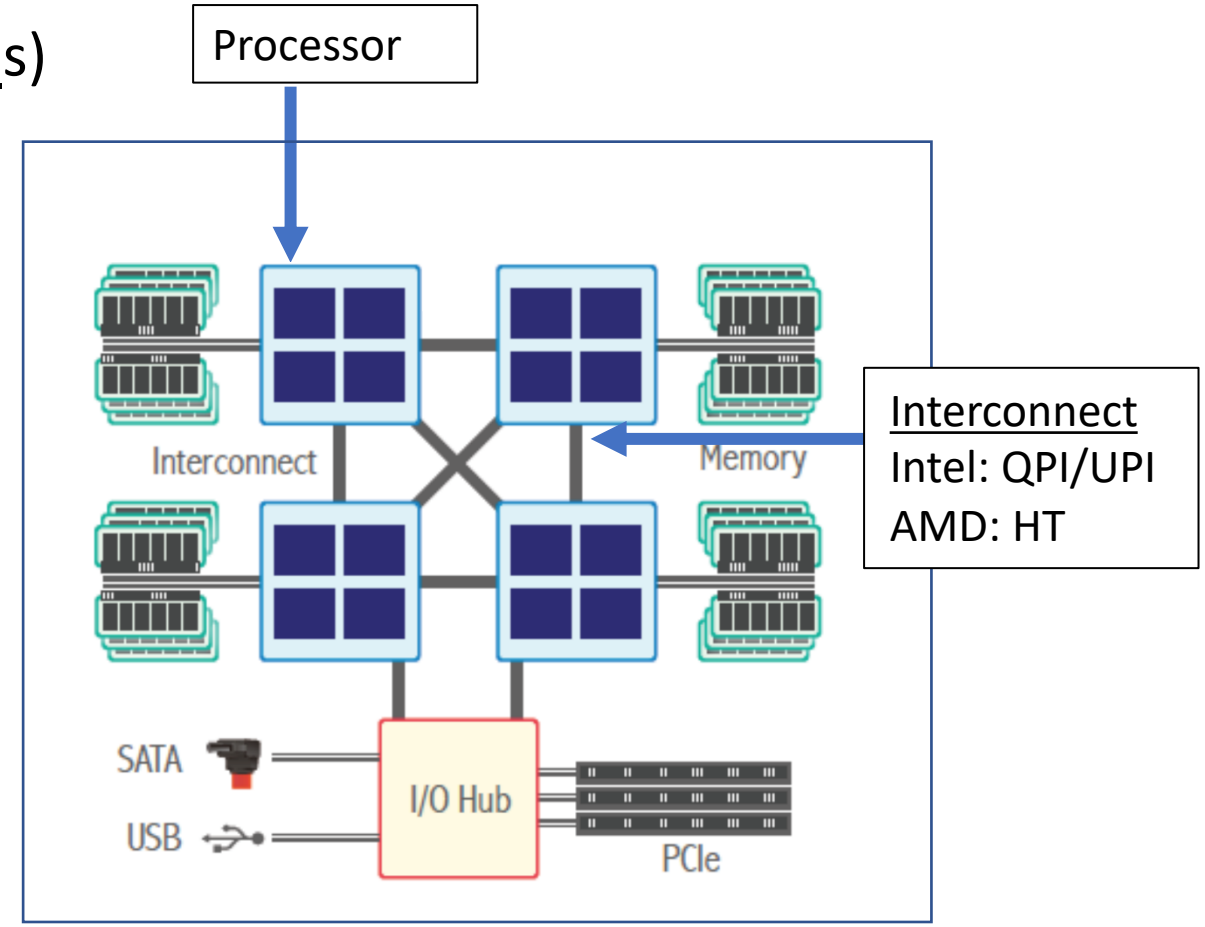
- Most have vector units
- Most have multiple cache levels
- Require special care to program (multi-threading, vectorization, …)
  - Free lunch is over!
- Instruction set
  - x86_64 – Intel and AMD
  - Power, ARM, NVIDIA, etc have different instruction sets





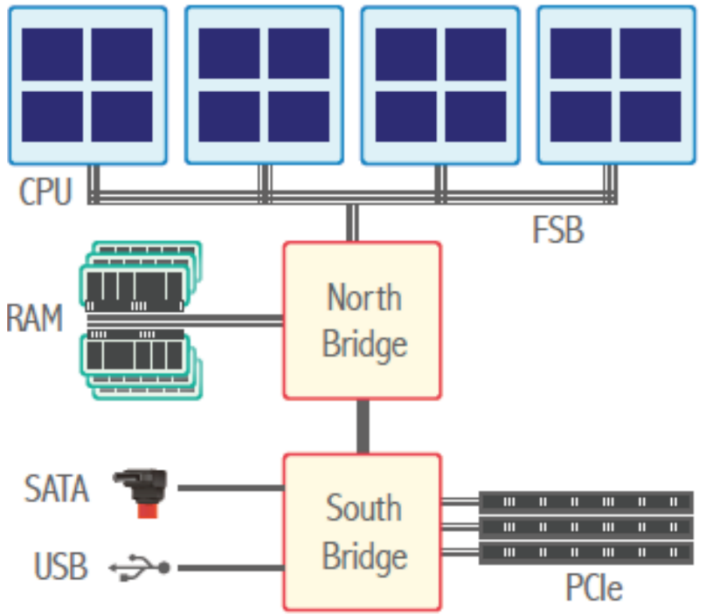
Motherboard Layout

- Typical nodes have 1, 2, or 4 Processors (CPUs)
  - 1: laptop/desktops/servers
  - 2: servers/clusters
  - 4: high end special use
- Each processor is attached to a socket
- Each processor has multiple cores
- Processors connected via interconnect
- Memory associated with processor
  - Programmer sees single node memory address space

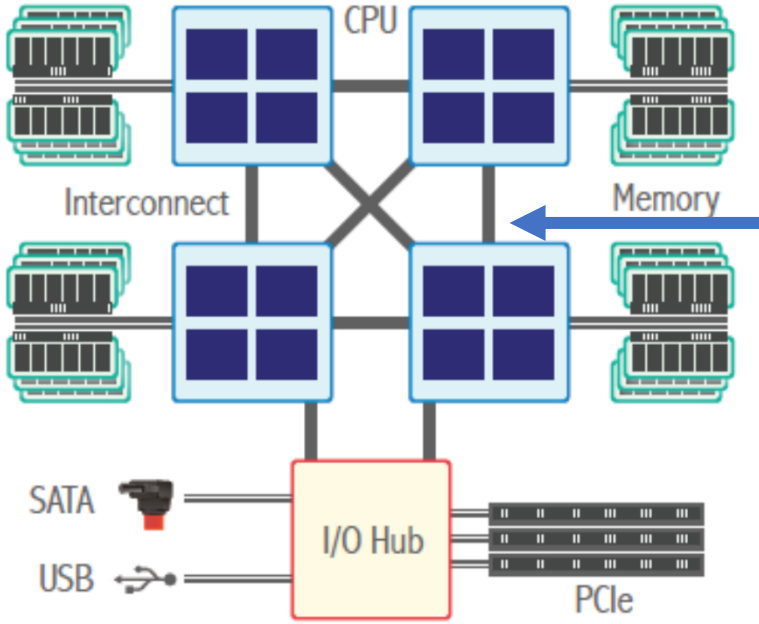


Motherboard Layouts

Elmer / Tuura / Lefebvre



Symmetric MultiProcessing (SMP)
front side bus (FSB) bottlenecks in SMP systems, device-to-memory bandwidth via north bridge, north bridge to RAM bandwidth limitations.



Non-Uniform Memory Access (NUMA)
physical memory partitioned per CPU, fast interconnect to link CPUs to each other and to I/O. Remove bottleneck but memory is no longer uniform – 30-50% overhead to accessing remote memory, up to 50x in obscure multi-hop systems.

Skylake Core Microarchitecture

Key Components:

Control logic
Register file
Functional Units

- ALU (arithmetic and logic unit)
- FPU (floating point unit)

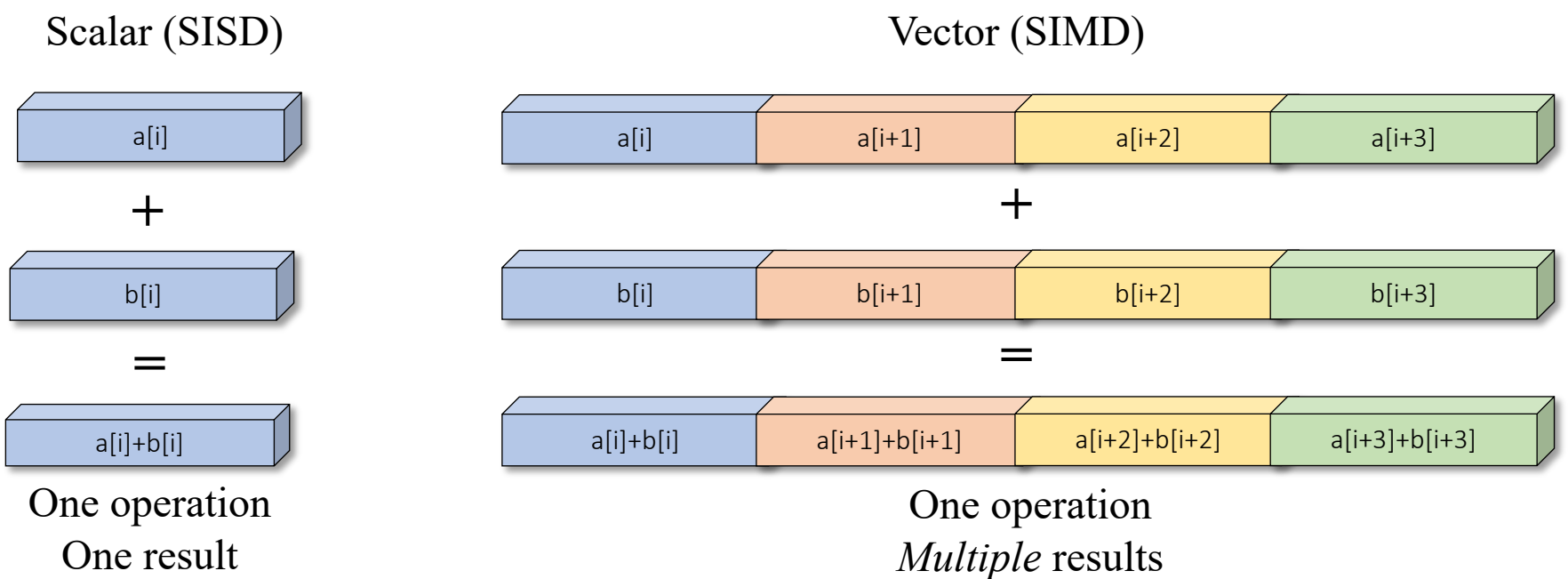
- Data Transfer
- Load / Store

Vectorization Terminology

- SIMD: Single Instruction Multiple Data
- New generations of CPUs have new capabilities
  - Each core has VPUs and vector registers
  - Machine (assembly) commands that take advantage of new hardware
    - Called “instruction set”
- E.g. Streaming SIMD Extensions (SSE) & AVX: Advanced Vector Extension (AVX)

At inner-most loop level

```
for (int i=0; i<N; i++) {
    c[i]=a[i]+b[i];  }
```

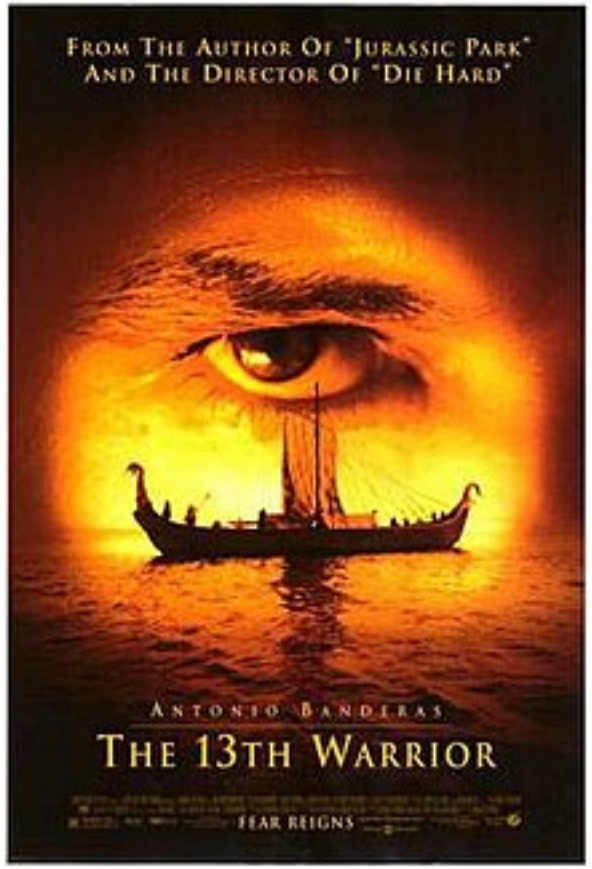


HPC Performance Metrics

- FLOP = Floating-point Operation
- FLOPs = Floating-point Operations
- FLOPS = Floating-point Operations Per Second
  - Often FLOP/sec to avoid ambiguity

- Calculation for single processor
  - Usually Double Precision (64-bit)
  - FLOPS = (Clock Speed) *(Cores)*(FLOPs/cycle)

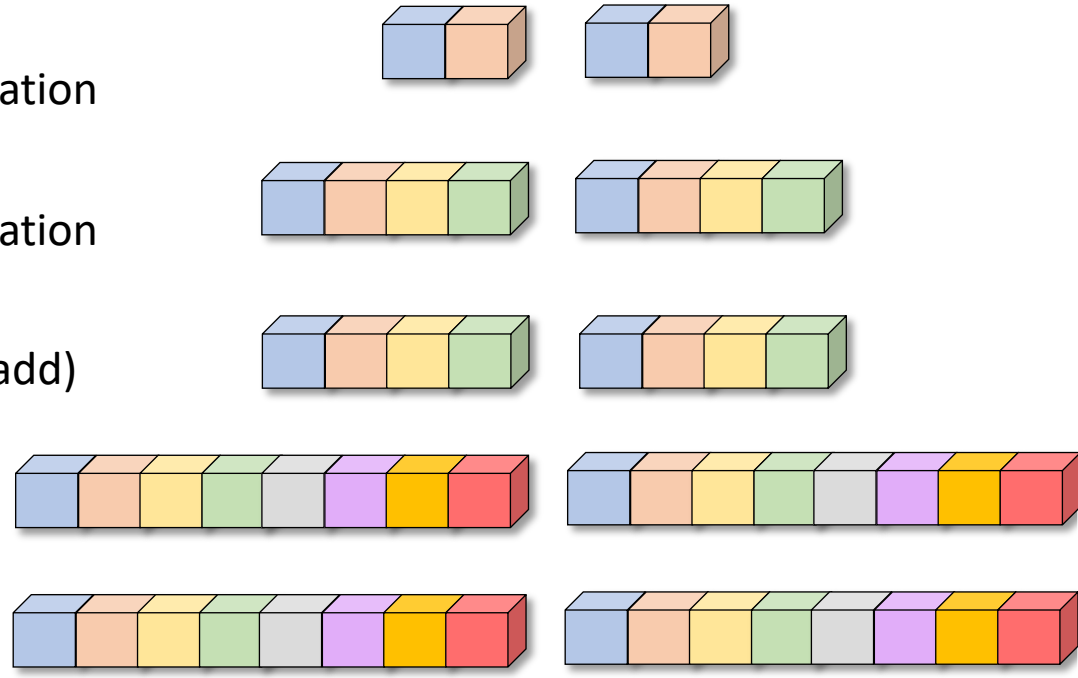
- Others:
  - FLOPS/$
  - FLOPS/Watt
  - Bandwidth: GB/s



N.B. - not to be confused with a Hollywood flop!

Intel Xeon (Server) Architecture Codenames

- Number of FLOP/s depends on the chip architecture
- Double Precision (64 bit double)
  - Nehalem/Westmere (SSE):
    - 4 DP FLOPs/cycle: 128-bit addition + 128-bit multiplication
  - Ivybridge/Sandybridge (AVX)
    - 8 DP FLOPs/cycle: 256-bit addition + 256-bit multiplication
  - Haswell/Broadwell (AVX2)
    - 16 DP FLOPs/cycle: two, 256-bit FMA (fused multiply-add)
  - KNL/Skylake (AVX-512)
    - 32 DP FLOPs/cycle: two*, 512-bit FMA
  - Cascade Lake (AVX-512 VNNI with INT8)
    - 32 DP FLOPs/cycle: two, 512-bit FMA



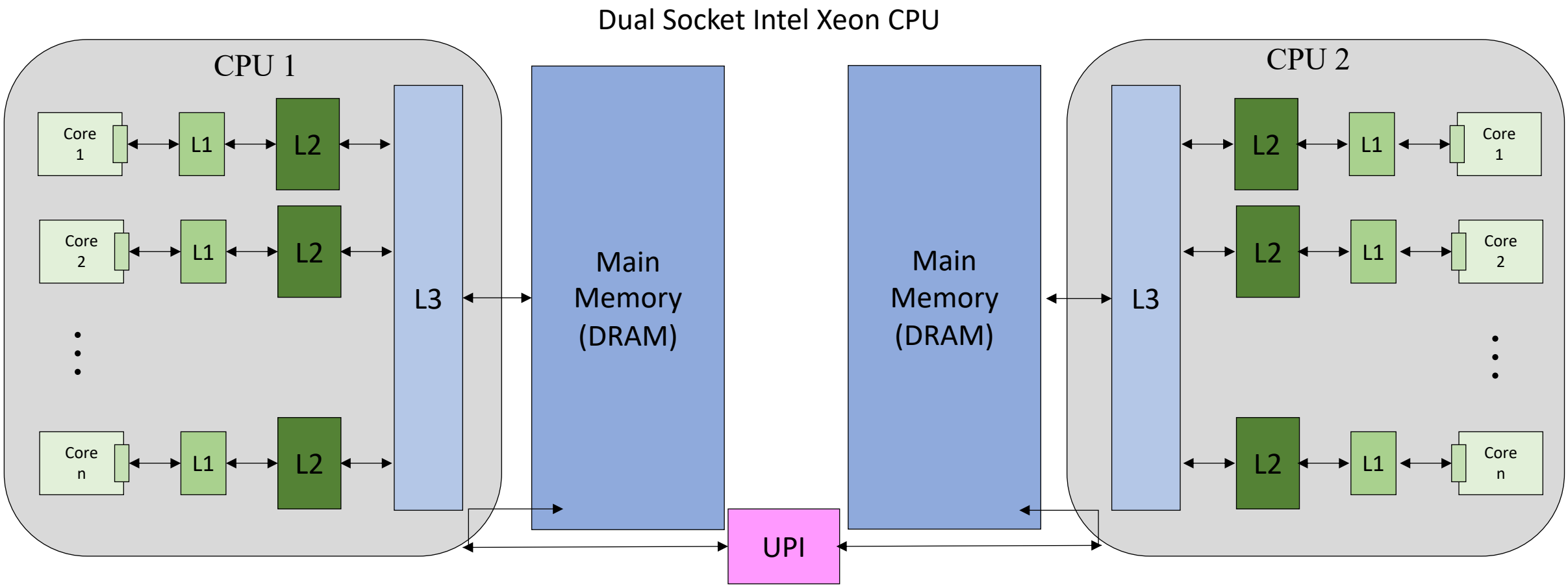
- FMA = (a × b + c)
- Twice as many if single precision (32-bit float)

*Some Skylake-SP have only one

Outline

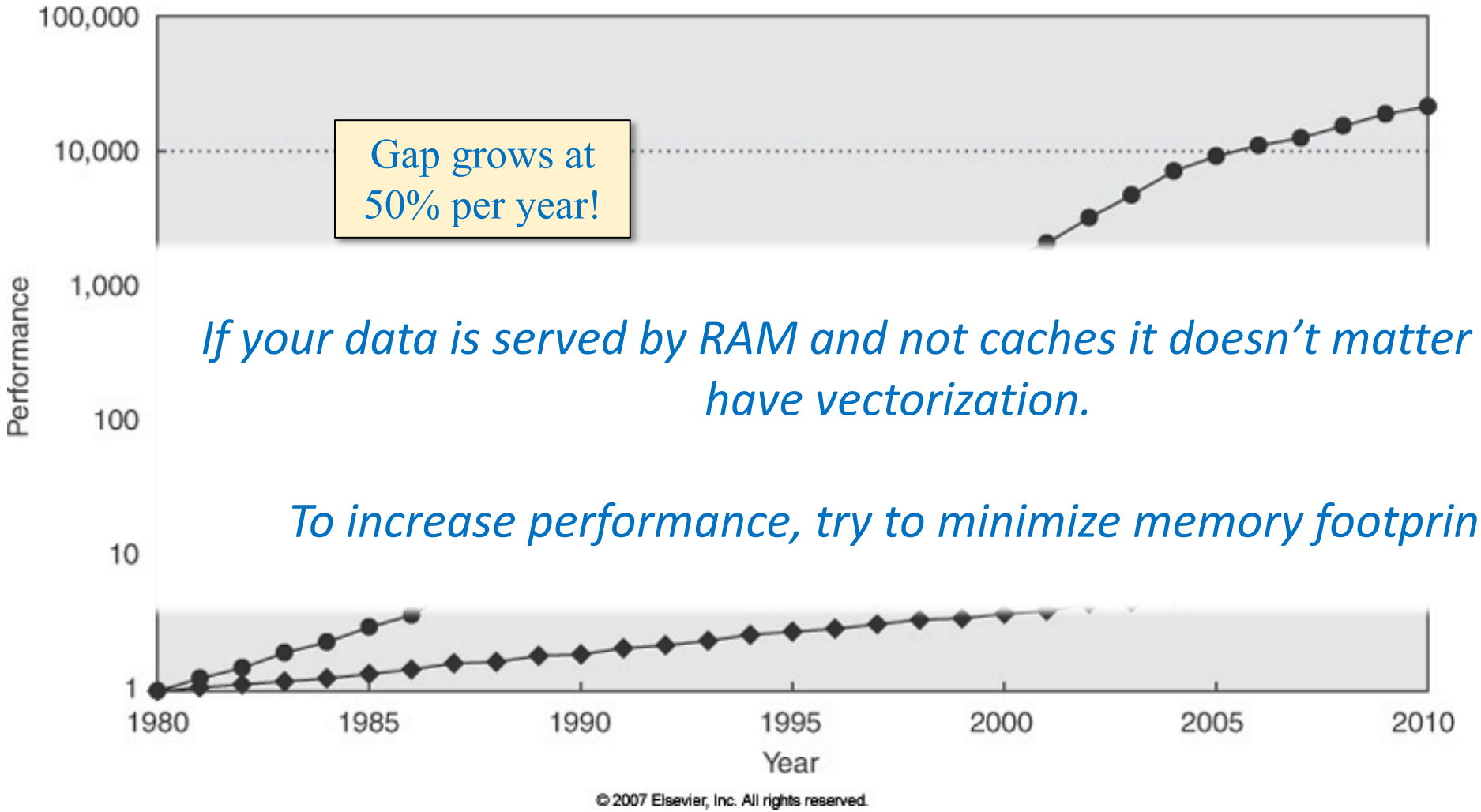
- Motivation
- HPC Cluster
- Compute
- Memory
  - Layout
  - Cache
  - Performance
- Disk
- Emerging Architectures

Memory Hierarchy



	Registers	L1 Cache	L2 Cache	L3 Cache	DRAM	Disk
Speed	1 cycle	~4 cycles	~10 cycles	~30 cycles	~200 cycles	10ms
Size	< KB per core	~32 KB per core	~256 KB per core	~35 MB per socket	~100 GB per socket	TB

Does It Matter?



If your data is served by RAM and not caches it doesn't matter if you have vectorization.

To increase performance, try to minimize memory footprint.

http://web.sfc.keio.ac.jp/~rdv/keio/sfc/teaching/architecture/architecture-2008/hennessy-patterson/Ch5-fig02.jpg

Outline

- Motivation
- HPC Cluster
- Compute
- Memory
- Disk
  - Types
  - Filesystems
- Emerging Architectures

File I/O

"A supercomputer is a device for turning compute-bound problems into I/O-bound problems."

-- Ken Batcher, Emeritus Professor of Computer Science at Kent State University

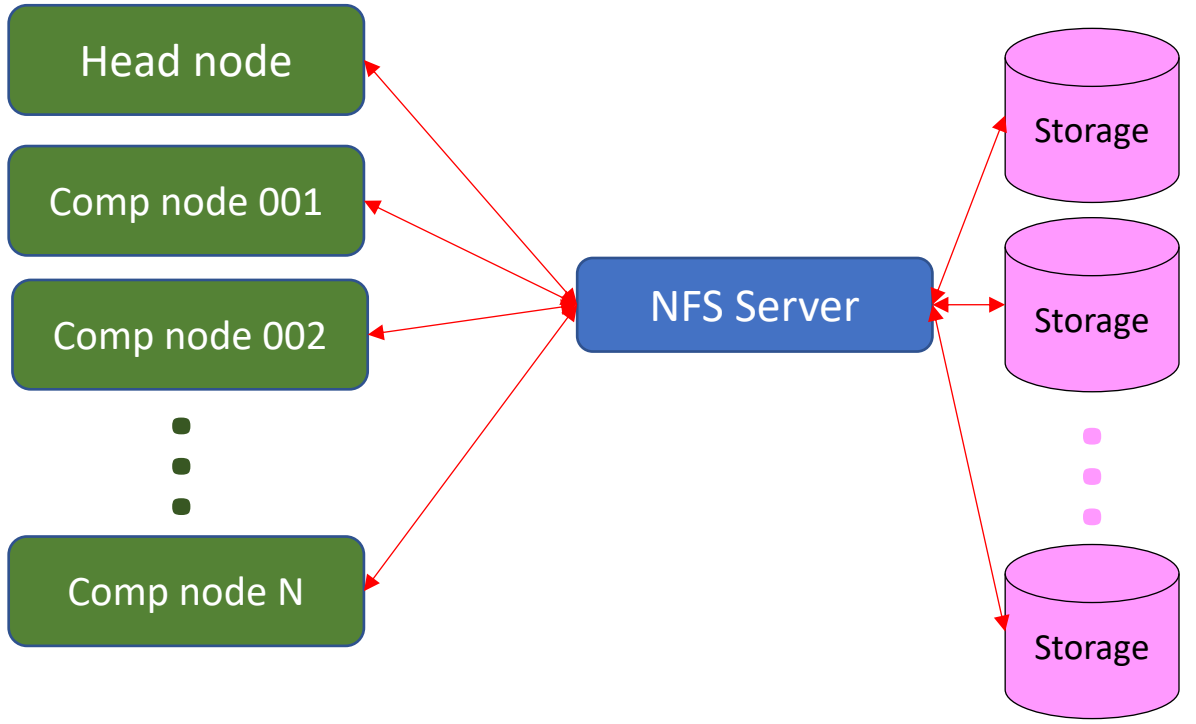
Types of Disk

- Hard Disk Drive (HDD)
  - Traditional Spinning Disk
- Solid State Drives (SSD)
  - ~5x faster than HDD
- Non-Volatile Memory Express (NVMe)
  - ~5x faster than SSD

Photo Credit: Maximum PC

Types of Filesystems - NFS

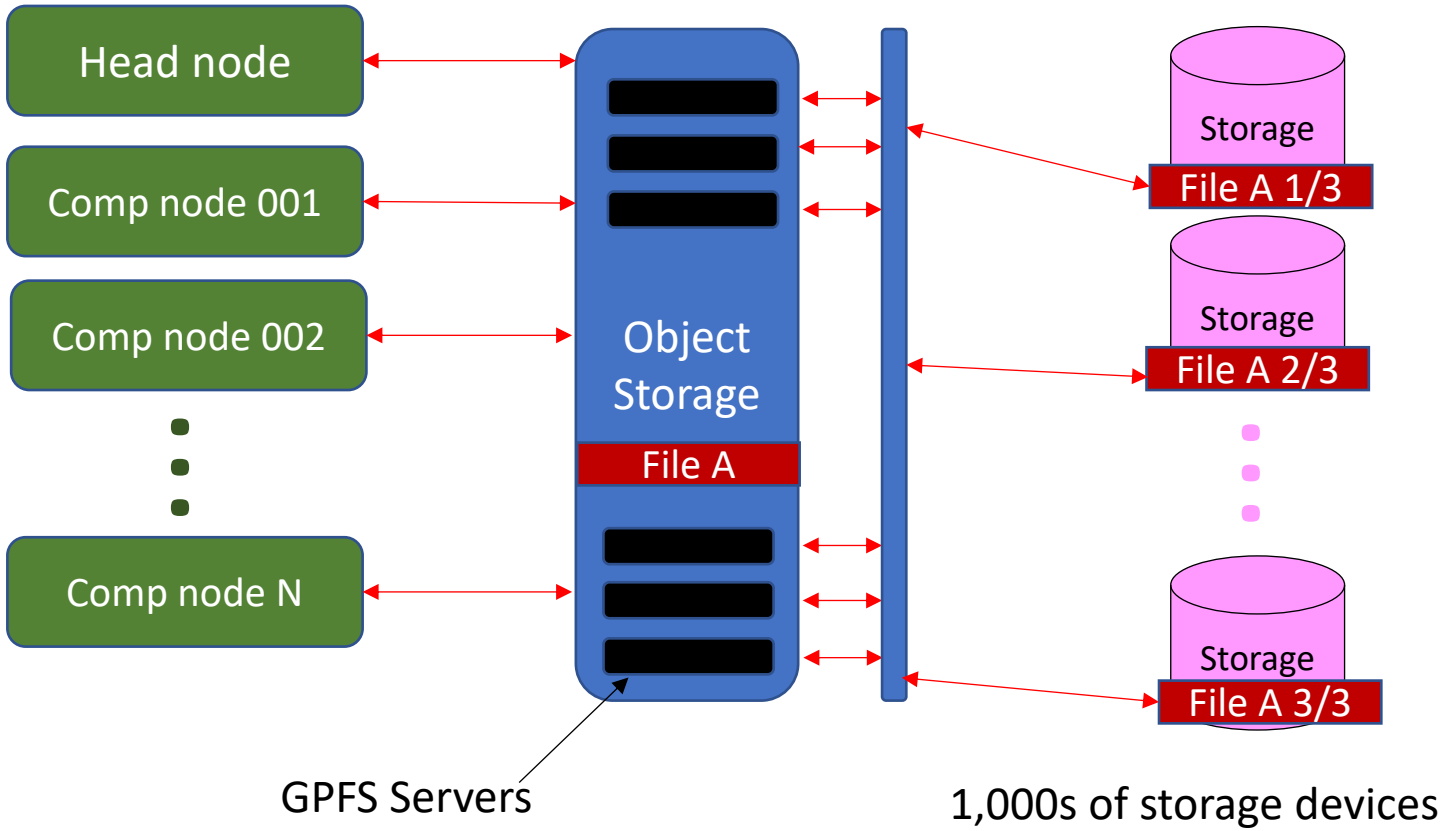
- NFS – Network File System
  - Simple, cheap, and common
  - Single filesystem across network
  - Not well suited for large throughput



≤ 8 storage devices

Parallel Filesystems

- GPFS (General Parallel Filesystem – IBM)
  - Designed for parallel read/writes
  - Large files spread over multiple storage devices
  - Allows concurrent access
  - Significantly increase throughput
- Lustre
  - Similar idea
  - Different implementation
- Parallel I/O library in software
  - Necessary for performance realization

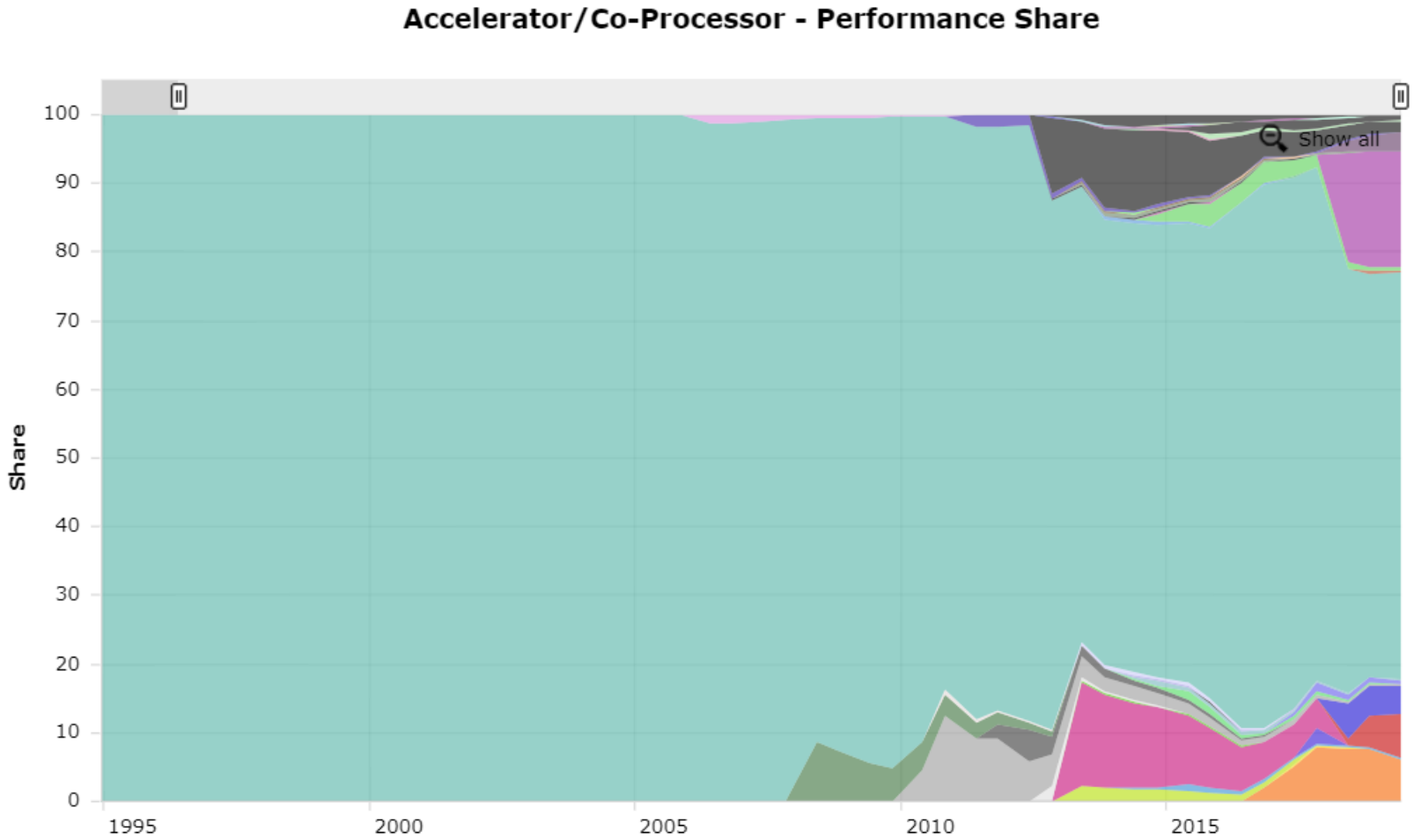


Outline

- Motivation
- HPC Cluster
- Compute
- Memory
- Disk
- Emerging Architectures
  - Xeon Phi
  - GPU
  - Cloud

Emerging Architectures

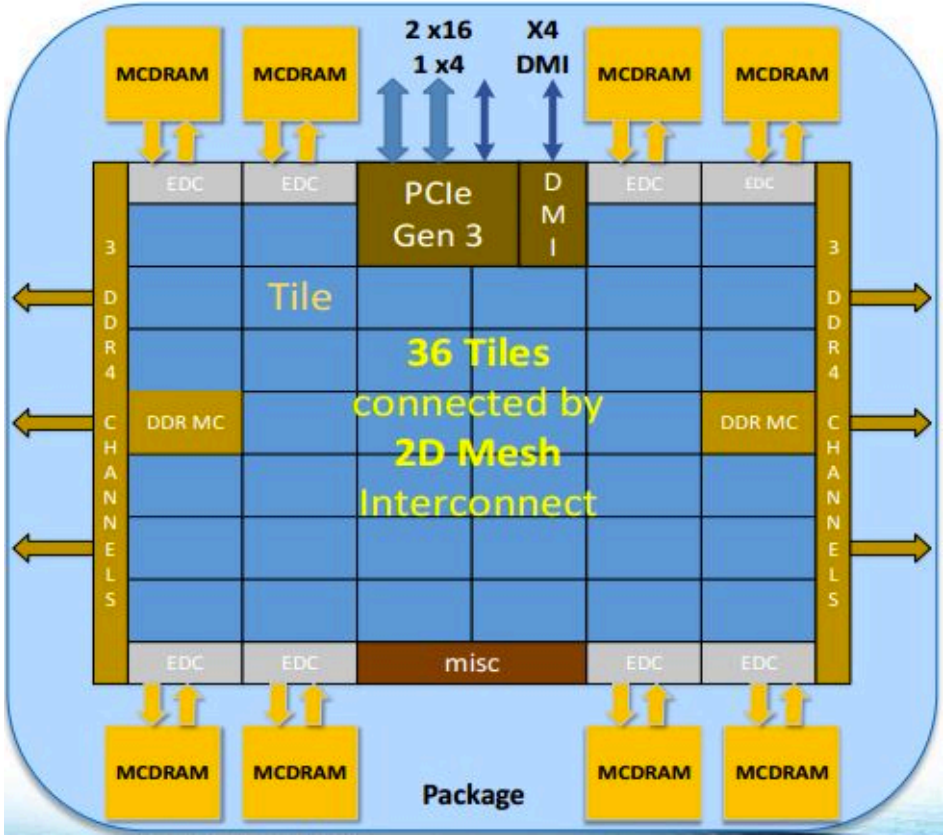
- The computing landscape changes rapidly
- 144 TOP500 systems have accelerators (June 2019)



https://www.top500.org/statistics

Intel Xeon Phi a.k.a. Intel MIC

- MIC: Many Integrated Cores
- x86-compatible multiprocessor architecture
- Programmable using
  - C, C++, Fortran
  - MPI, OpenMP, OpenCL, …
- Started as co-processor
  - Data travels over PCIe
- Became a main processor in 2nd generation
- Merged into Xeon line ~1 year ago
  - Intel announced end of Xeon Phi product line

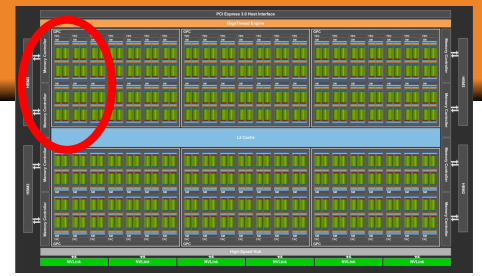


GPUs

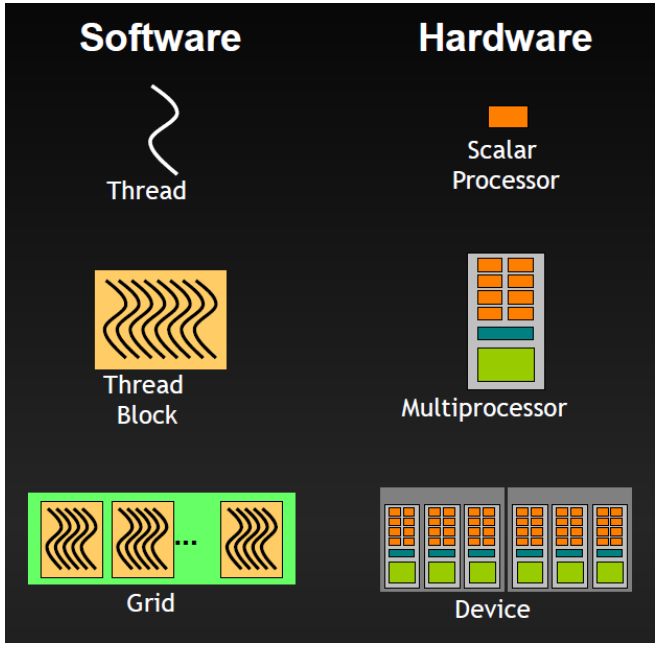
- Graphic Processing Unit (GPU)
  - General Purpose GPU (GPGPU)
- Exclusively a co-processor
  - Data travels over PCIe
  - NVLink is NVIDIAs new high-speed interconnect
- Not compatible with x86 library
- Programmable using:
  - CUDA, OpenCL
  - OpenACC, OpenMP
- NVIDIA GPUs
  - Fermi/Kepler/Maxwell/Pascal/Volta

NVIDIA Volta V100

NVIDIA Volta (V100) GPU



- Scalar Processors execute parallel thread instructions
- Streaming Multiprocessors (SMs) each contain:
  - 64 SP units (“CUDA cores”)
  - 32 DP unit
- 32 GB device memory
- Execution Model:



Accelerator Comparison

NB: Data from 2018

	Xeon Gold 5122	Xeon Platinum 8280M	**Xeon Phi 7290F**	**NVIDIA V100**
Cores	4	28	72	84 SMs
Logical Cores	8	56	288	5120 cores
Clock rate	3.6 – 3.7 GHz	2.5 – 3.8 GHz	1.5-1.7 GHz	1530 MHz
Theoretical GFLOPS (double)	460.8	2240	3456	7450
SIMD width	512 bit	512 bit	512 bit	Warp of 32 threads
Memory	--	--	16 GB MCDRAM 384 GB DDR4	32 GB
Memory B/W	127.8 GB/s	127.8 GB/s	400+ GB/s (MCDRAM) 115.2 GB/s	900 GB/s
Approx. Unit Price	$1,200	$12,000	$3,400	$9,000

Demystifying the Cloud

- “Regular” computers, just somewhere else
- Provide users with remote virtual machines or containers
- Can be used for anything:
  - Mobile-services, Hosting websites, Business Application, …
  - **Data Analysis, High Performance Computing**
- Providers
  - Major players: Amazon, Google, Microsoft, HP, IBM, Salesforce.com, …
  - Lots of others

Cloud Computing

- Advantages:
  - *Potentially* lower cost:
    - Pay as you go
  - *Potentially* lower cost:
    - Save on sysadmins and infrastructure
  - *Potentially* lower cost:
    - Economy of scale: providers
  - Scaling up or down as needed
    - Can be used to run overflow from a regular data center
  - Access to a wide range of hardware
- Additional challenges
  - Data movement
    - Expensive and time consuming
  - Security, privacy, …

What We Didn’t Talk About

- Microarchitecture details (ALU, FPU, pipelining…)
- Memory bandwidth
- Cache lines and cache coherence
- IBM (Power) or AMD
- FPGAs

Resources & References

- Very nice glossary: https://cvw.cac.cornell.edu/main/glossary
- J. Hennessy, D. Patterson, Computer Architecture: A Quantitative Approach, 6th edition (2017), ISBN 978-0128119051
- U. Drepper, What Every Programmer Should Know About Memory, http://people.redhat.com/drepper/cpumemory.pdf
- NVIDIA Volta Architecture Whitepaper https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf