Emerging Technologies in High Performance Computing

SDSC Summer Institute

Pietro Cicotti

August 5, 2016

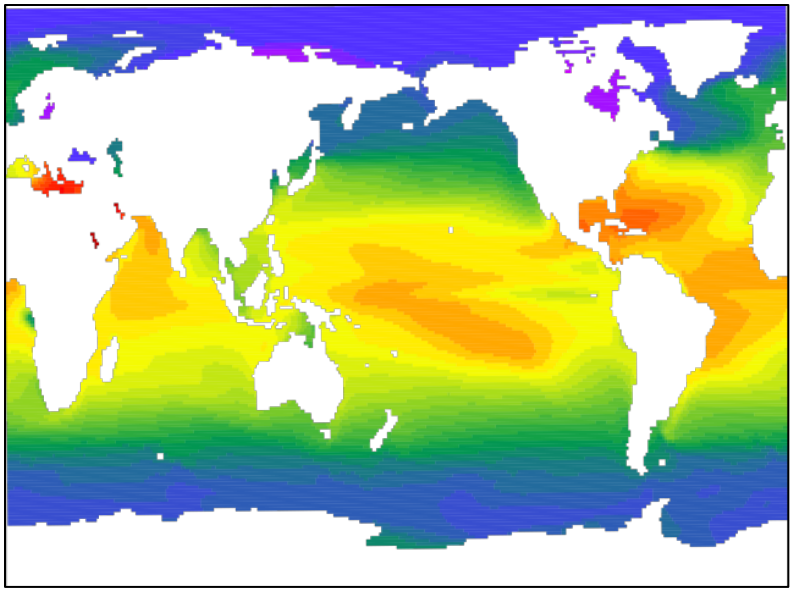
Purpose of this talk

Give a sense of the how the HPC technology and is evolving and what the implications are for developers and those who are involved with the operations and support of HPC systems.

Exascale applications will lead to technologies that are relevant for the 99%

Using current technology, power estimates for exascale systems based on todays technologies are hundreds to over a Giga Watt of power. This is driving technology in ways that will impact systems for the 99%.

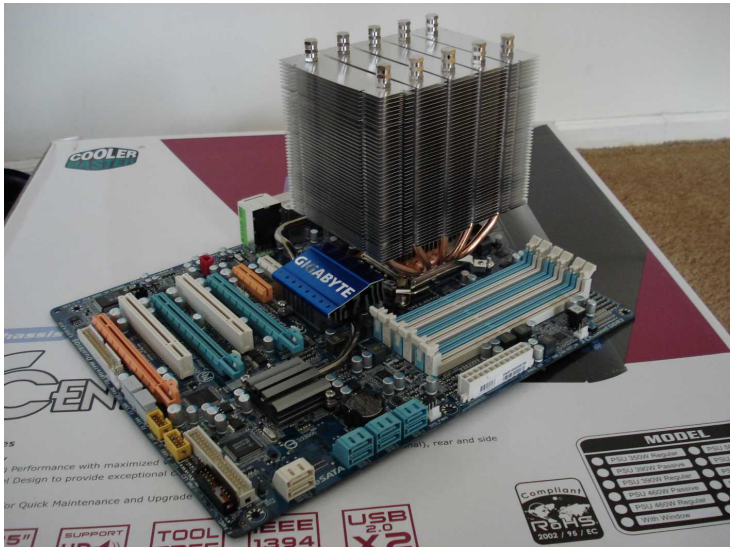
- **Exascale application drivers**
  - Climate modeling
  - Biological mechanism for human disease
  - Understanding the human brain
  - Galaxy formation
  - Analyzing genomic data
  - Turbulence modeling
  - And many others…



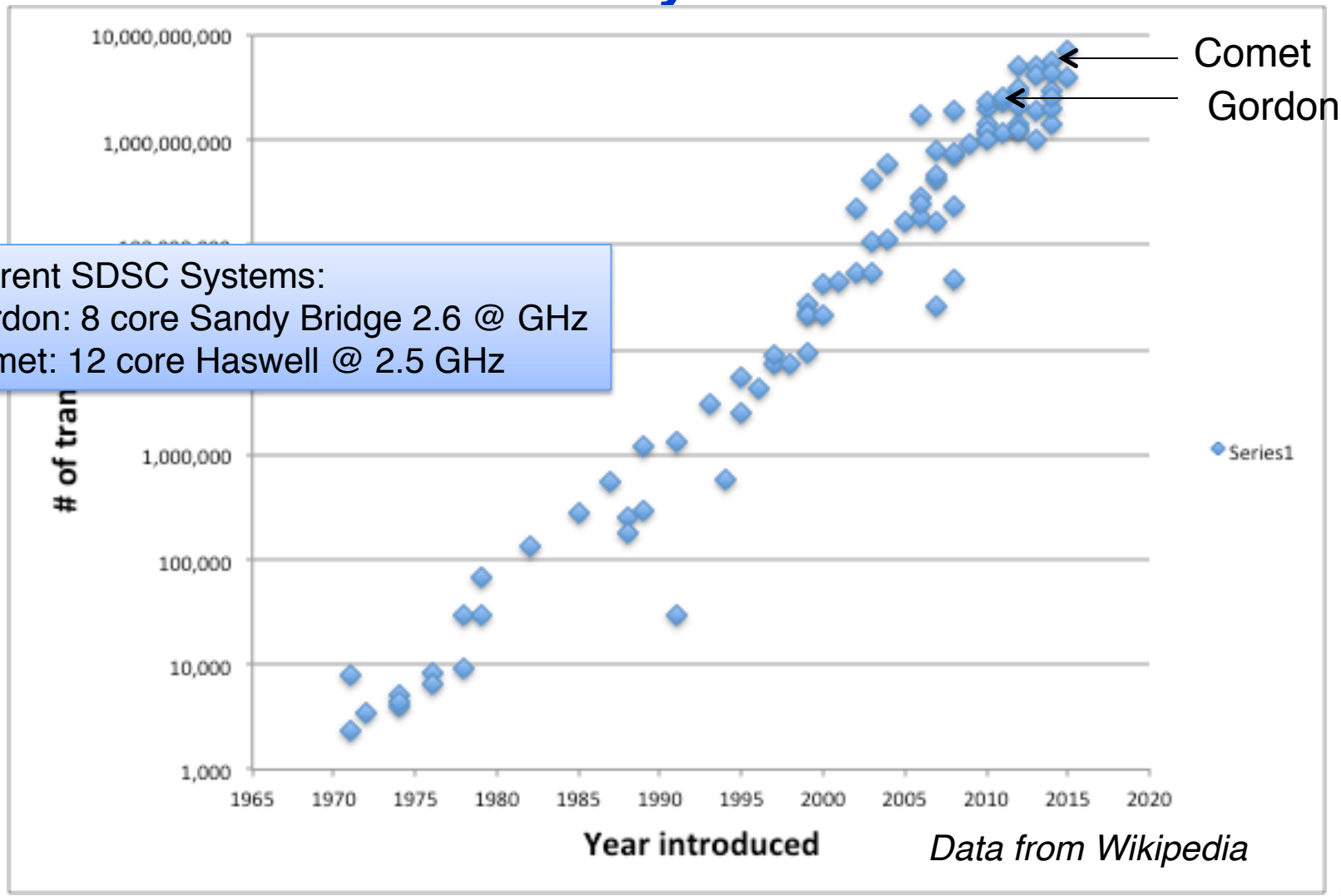
Processor

The end of Moore's Law?

- Moore's Law
  - transistor densities double roughly every 18 months
- Historically this has also meant that processor speeds have increased
- Dennard scaling breakdown
  - Power density not constant
  - More power
  - More heat

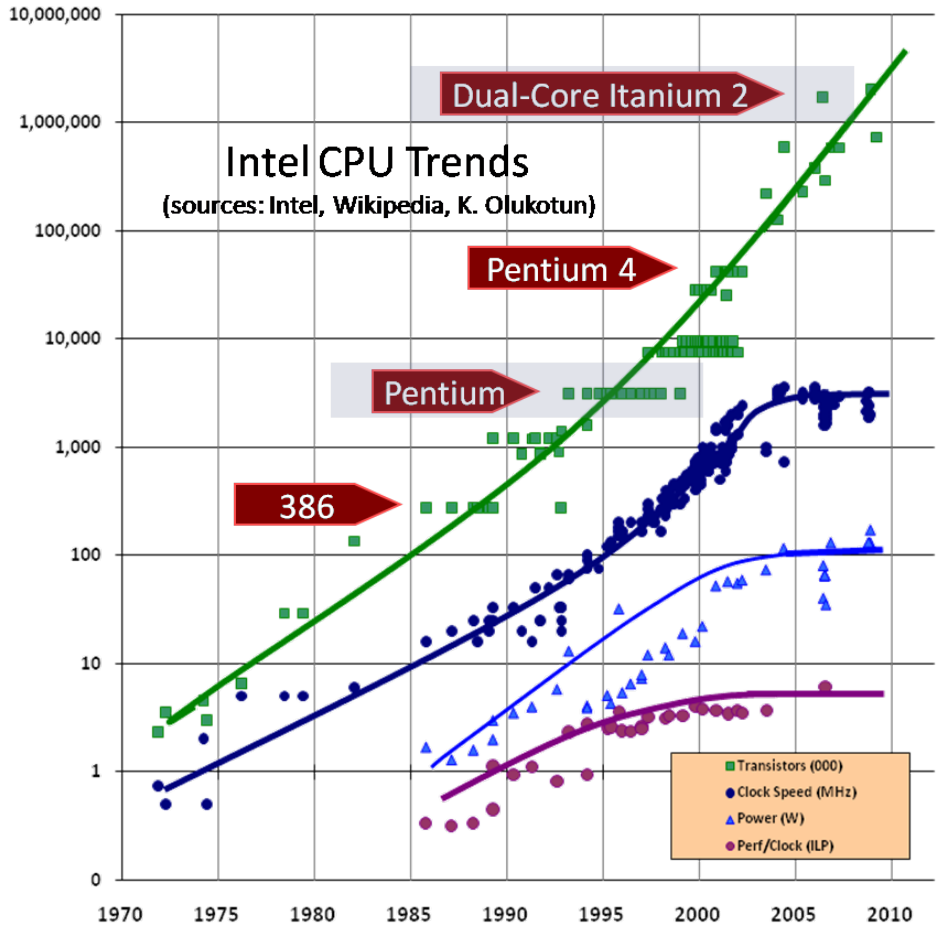


Moore's Law – a doubling of transistor count ever ~ 2 years



Current SDSC Systems:
Gordon: 8 core Sandy Bridge 2.6 @ GHz
Comet: 12 core Haswell @ 2.5 GHz

While the number of transistors has continued to increase, frequencies are topping out



Source: The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software. http://www.gotw.ca/publications/concurrency-ddj.htm

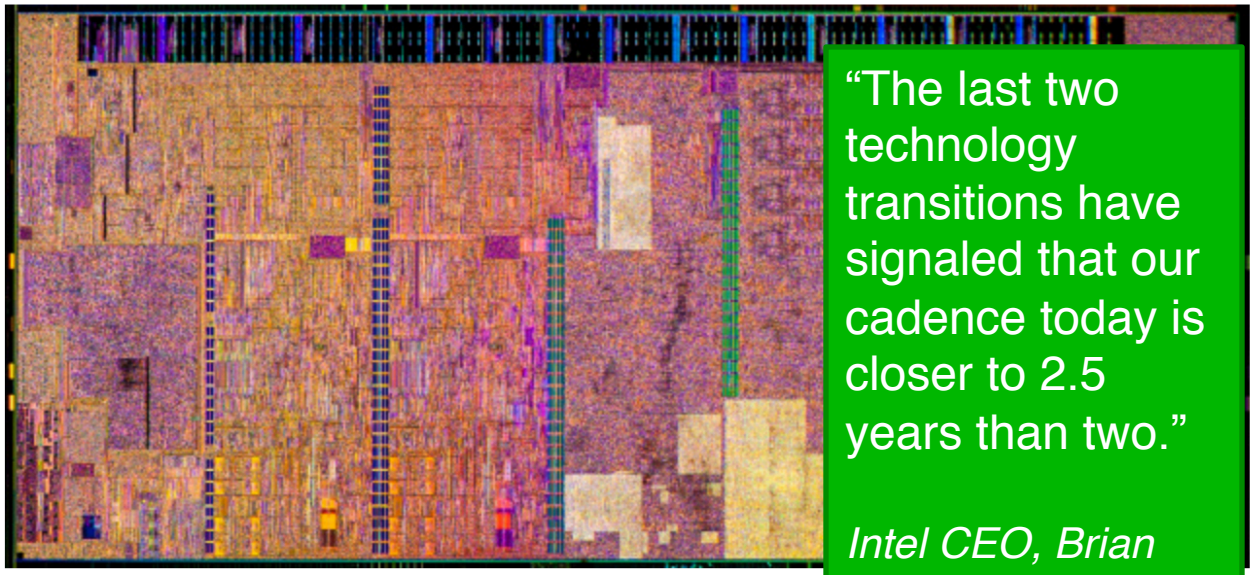
Moore's Law skips a tick…

Intel confirms tick-tock-shattering Kaby Lake processor as Moore's Law falters

Company will make three generations of 14nm processors, delaying the switch to 10nm.

by **Peter Bright** - Jul 15, 2015 5:52pm PDT

Share Tweet 191



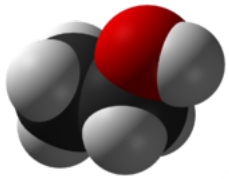
The Broadwell die, built on Intel's troublesome 14nm process.

Intel

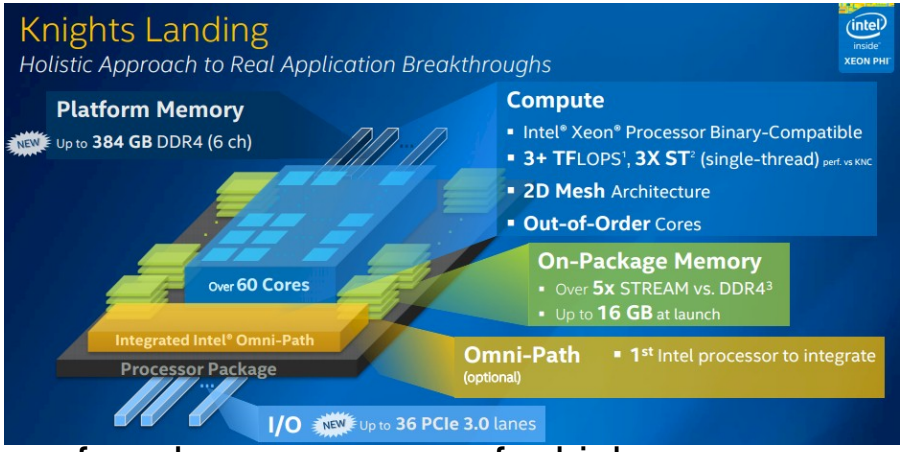
"The last two technology transitions have signaled that our cadence today is closer to 2.5 years than two."

Intel CEO, Brian Krzanich

65nm	P6	
	Core	Merom
45nm		Pennryn
	Nehalem	Nehalem
32nm		Westmere
	SandyBridge	SandyBridge
22nm		IvyBrdge
	Haswell	Haswell
14nm		Broadwell
	Skylake	Skylae
10nm		KabyLake
		CannonLake
	IceLake	IceLake
7nm		TigerLake
5nm		



Many core processors (Xeon Phi)



- Unmodified Xeon app will run on Xeon Phi (code mods required for performance!)
- 60+ cores
- Clock frequency of maybe 1.2-1.3 GHz
- ~3TF/node
- Self hosted (i.e., not an accelerator)
- ~16GB of on-package memory
- 1 socket per node – Aggregate high performance from low power cores for higher efficiency
- Focus on MPI/OpenMP hybrid applications and improving thread-level performance
- Investments in optimizing for many core will deliver improve Xeon performance.

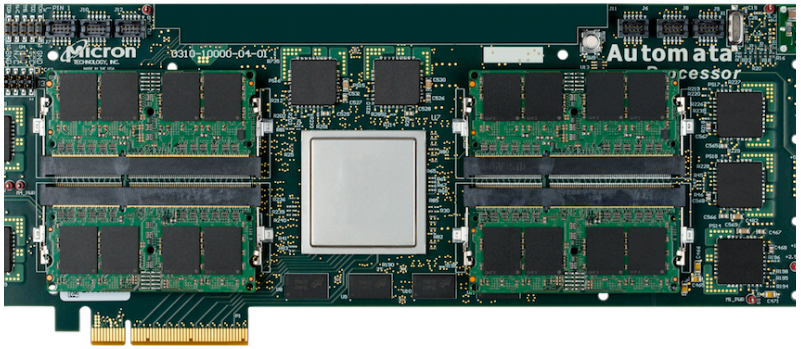
- Still not clear how the Xeon and Xeon Phi situation will manifest itself in future systems and processors.

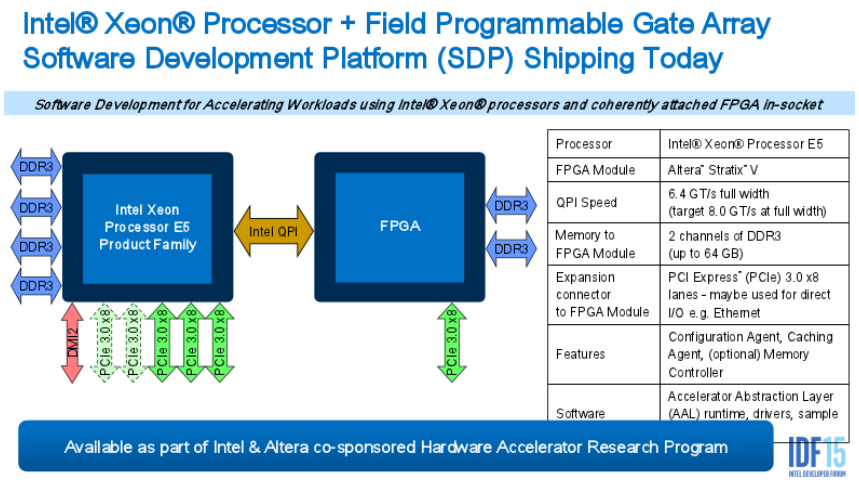
ARM architecture

- **RISC architecture**
- **Embedded systems**
  - Low power processors
- **Entering the datacenter and HPC market**
  - ARM v8
    - SIMD instructions
    - Higher clock frequency
- **Licensing model**
  - Cavium
  - Broadcom
  - Qualcomm

Specialization

- **Higher performance and efficiency**
  - Devices tailored to specific tasks
- **Micron's Automata Processor**
  - Engine for NFA
  - MISD processor
- **Processor+FPGA**
  - General purpose with reconfigurability

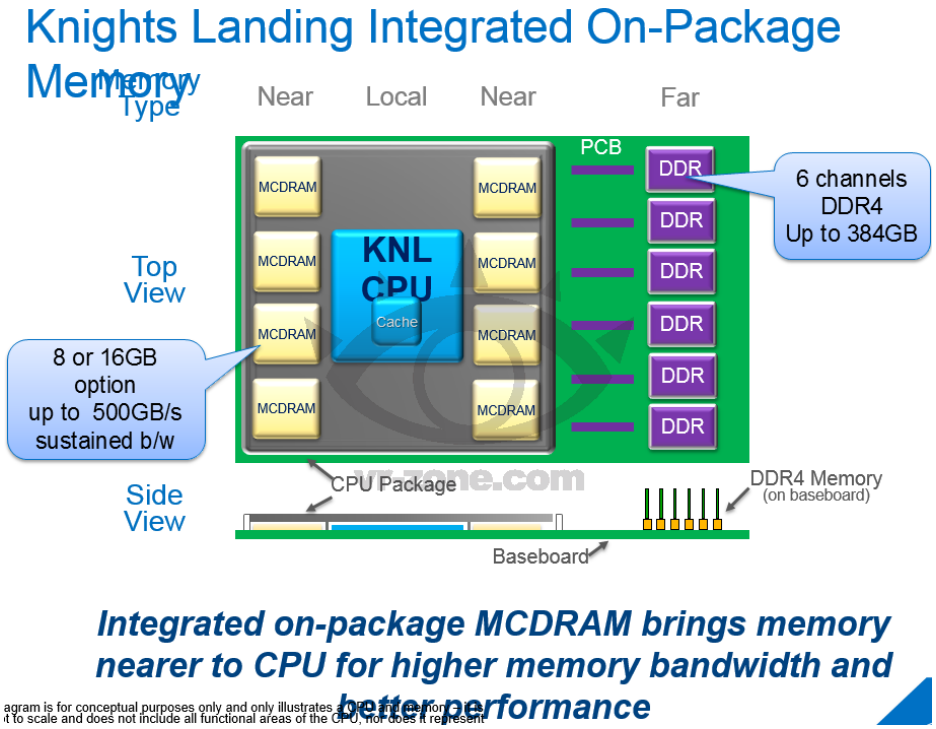




Memory

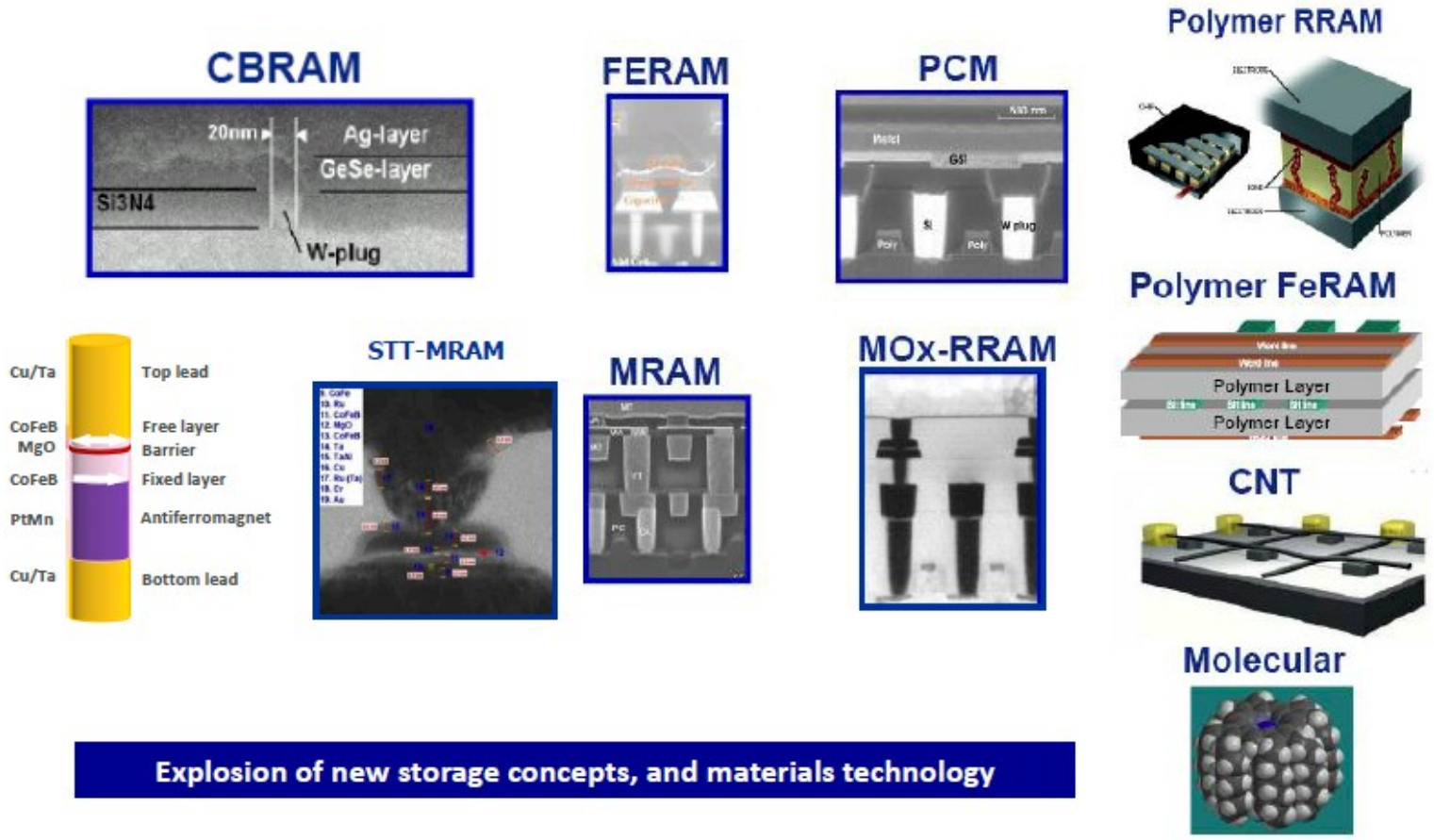
3D Stacking

- **Improve memory scaling and efficiency**
  - Uses TSV to connect planes
  - E.g. Micron's Hybrid Memory Cube (HMC)
- **Xeon Phi KNL's on-package MCDRAM**



Source: https://software.intel.com/en-us/forums/topic/499090

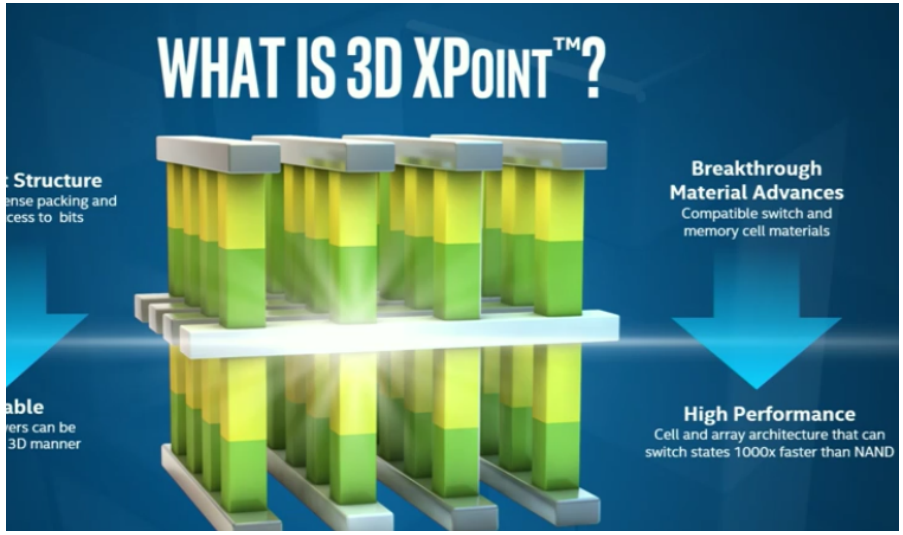
New non-volatile memory technologies improve latency and bandwidth.

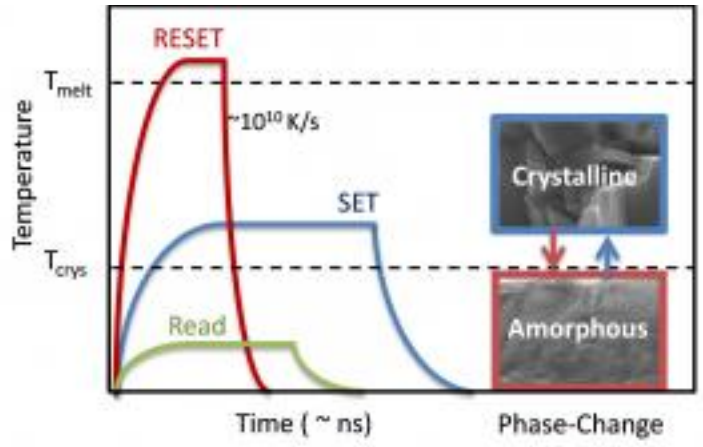


Source: http://www.theplatform.net/2015/07/29/scaling-the-growing-system-memory-hierarchy/

Recent announcement: 3D Xpoint (Intel & Micron)

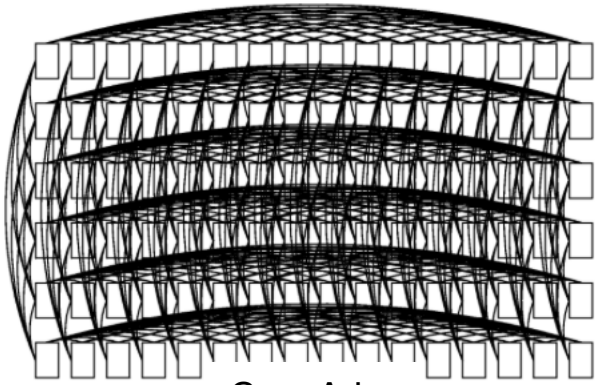
- 1000x the performance of NAND flash
- Non-volatile
- High endurance
- Hierarchically, sits between DRAM and NAND with features of both.
- Programming models and use cases are still emerging, but big data applications, those with unstructured data could be be winners.
- Underlying technology is not known for sure, but it appears to have the characteristics of phase change memory (PCM)



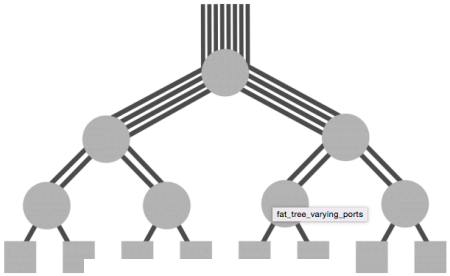


Interconnect

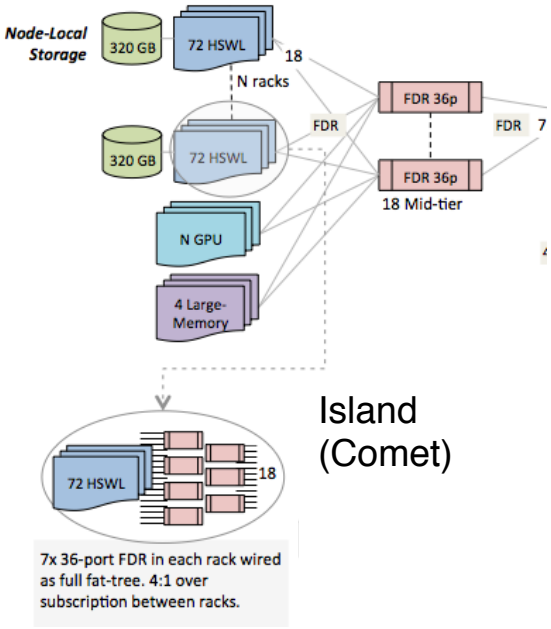
Interconnect topology refresher



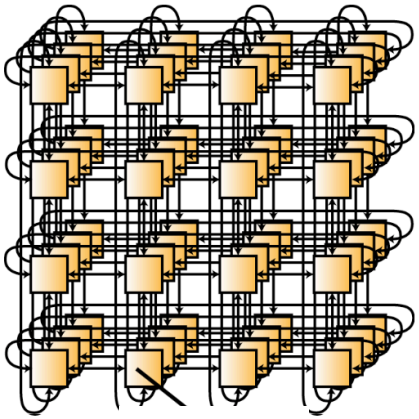
Cray Aries
(Edison)



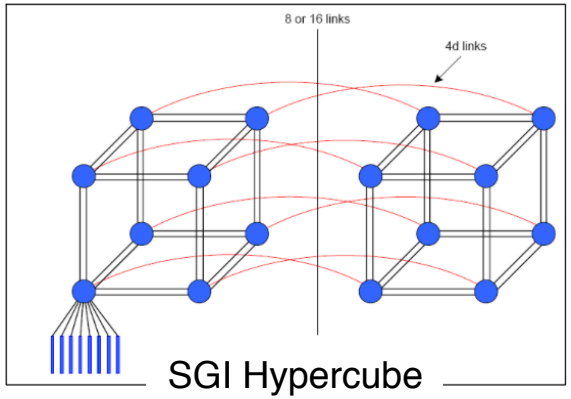
Full Fat Tree
(Yellowstone)



Island
(Comet)



3D Torus
(Gordon)



SGI Hypercube
(Pleides)

All of these but the Aries, typically use static routing

Today, InfiniBand is the predominant interconnect in HPC

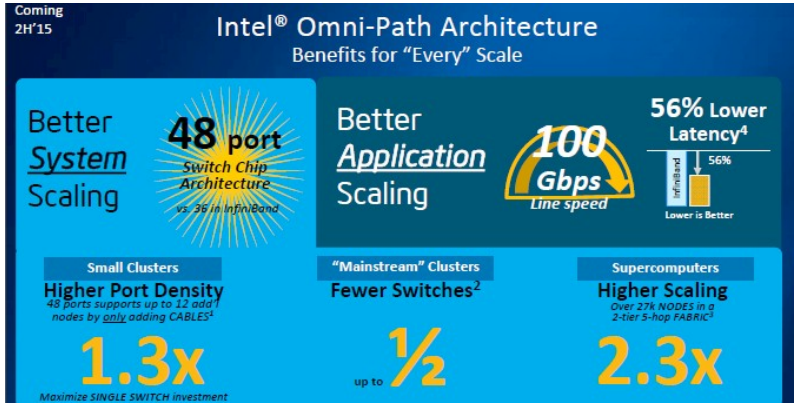
November 18, 2014
Mellanox InfiniBand Leads TOP500 as Most Used Interconnect Technology for HPC

"The number of InfiniBand-based supercomputers increased 8.7 percent year-over-year to 225, representing 45 percent of the TOP500."

- Cost for the interconnect (switches, cables, network adapters, software, etc), can be 20-30% of the system cost.
- This is partially why Comet uses an island approach. Alternative topologies and designs offer tradeoffs that allow system operators to reallocate resources to more important pieces of the design.

Intel OmniPath will feature adaptive routing and on-chip networking via silicon photonics

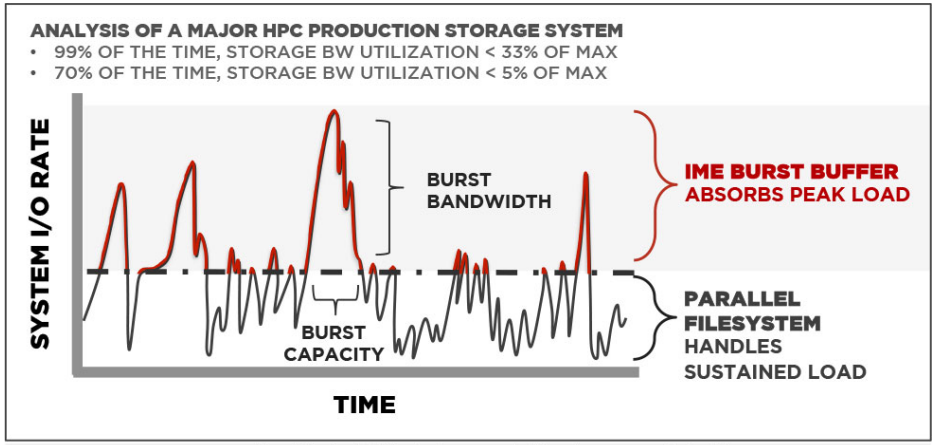
- Acquired Qlogic switch & HCA vendor
- Acquired Cray interconnect IP
- OPA
  - Adaptive routing
  - Silicon photonics on chip
  - Performance Scaled Messaging
  - 48 port switch architecture (vs. 36 port in IB)
  - Roughly equal to EDR in terms of latency and bandwidth
  - Will bring much-needed competition with IB/Mellanox.

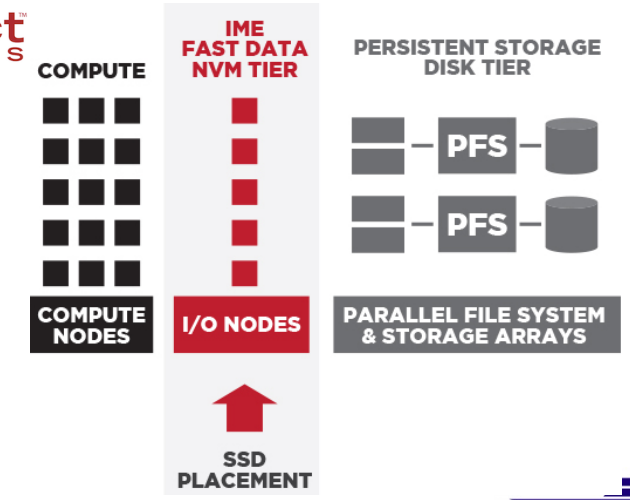


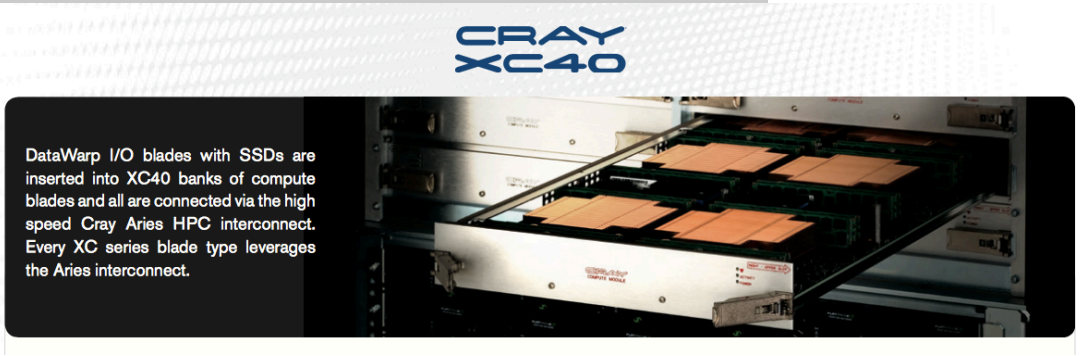
Storage

Hierarchical Storage and Burst Buffers

- I/O patterns are bursty and designing systems to support peak loads is expensive E.g., Comet can achieve 200 GB/s. But the average performance is probably in the 10's of GB/s
- Burst buffers act as capacitor for I/O, absorbing peak write activity using high-performance storage (like flash), then asynchronously moving it to lower performing storage. For reads, it can be used for staging, say prior to the start of a job, say prior to the start of a job.







NERSC Cori
Production deployment in mid 2016

- Intel many core processors (“Knight’s Landing”)
- 9,300 single socket nodes, at least **60 cores/socket**
- On-package memory will deliver 400GB/s, 5x that of the DRAM
- Cray XC-40 architecture based on the Aries dragonfly topology
- Cray DataWarp burst buffer flash-based storage provides 750 GB/s of I/O performance
- Significant application readiness effort via NESAP

Oak Ridge National Lab Summit System - Deployment in 2018

- IBM Power9 processor
- NVIDIA Volta GPUs
- 3400 compute nodes, up to 200 PF, 10MW
- Stacked memory will deliver 3xx PB/s of memory bandwidth
- Fabric based on Mellanox dual-rail EDR; NVLINK between CPU/GPU, and GPU/GPU
- Significant application readiness effort via Center for Accelerated Application Readiness (CAAR)

Argonne: Aurora
Production deployment in 2019

- Intel many core processors (“Knight’s Hill”)
- Minimum of 50,000 nodes, up to 200 PF
- On-package memory will deliver 30PB/s of memory bandwidth
- Fabric based on Intel’s OmniPath (Stormlake Gen2)
- Silicon photonics brings interconnect onto the processor
- Intel is prime, Cray is subcontractor as integrator
- Significant application readiness effort

More information: http://aurora.alcf.anl.gov

Summary

- HPC technology evolves in response to the computational challenges of the day
- There are new processor, memory, storage, and networking technologies on the horizon that look interesting and perhaps useful
- Understanding how and where this technology can be applied is a critical aspect supporting these challenges
- SDSC will continue its tradition of technology research, testing, and when the time is right, deploying these in future systems