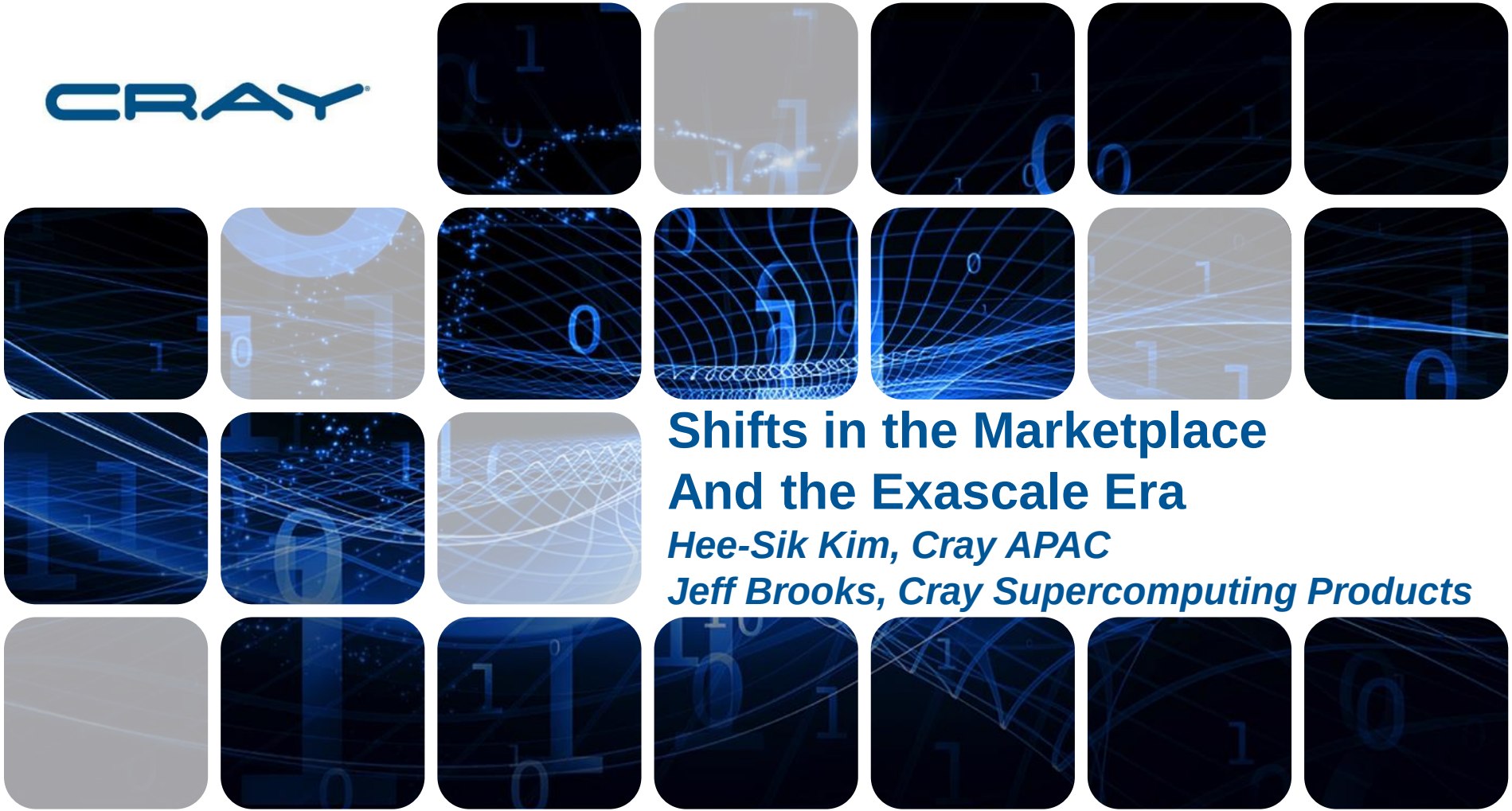
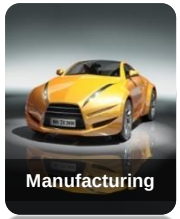


Shifts in the Marketplace And the Exascale Era

Hee-Sik Kim, Cray APAC
Jeff Brooks, Cray Supercomputing Products

Cray: a long history of supercomputing…

We build the world's fastest supercomputers to help solve "Grand Challenges" in science and engineering



Anything that can be simulated needs a Cray

Cray Today

Compute

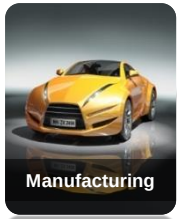
- Supercomputers
- Flexible Clusters
- Hybrid Architectures

Store

- Integrated Storage
- Data Management
- Tiered Storage Archive

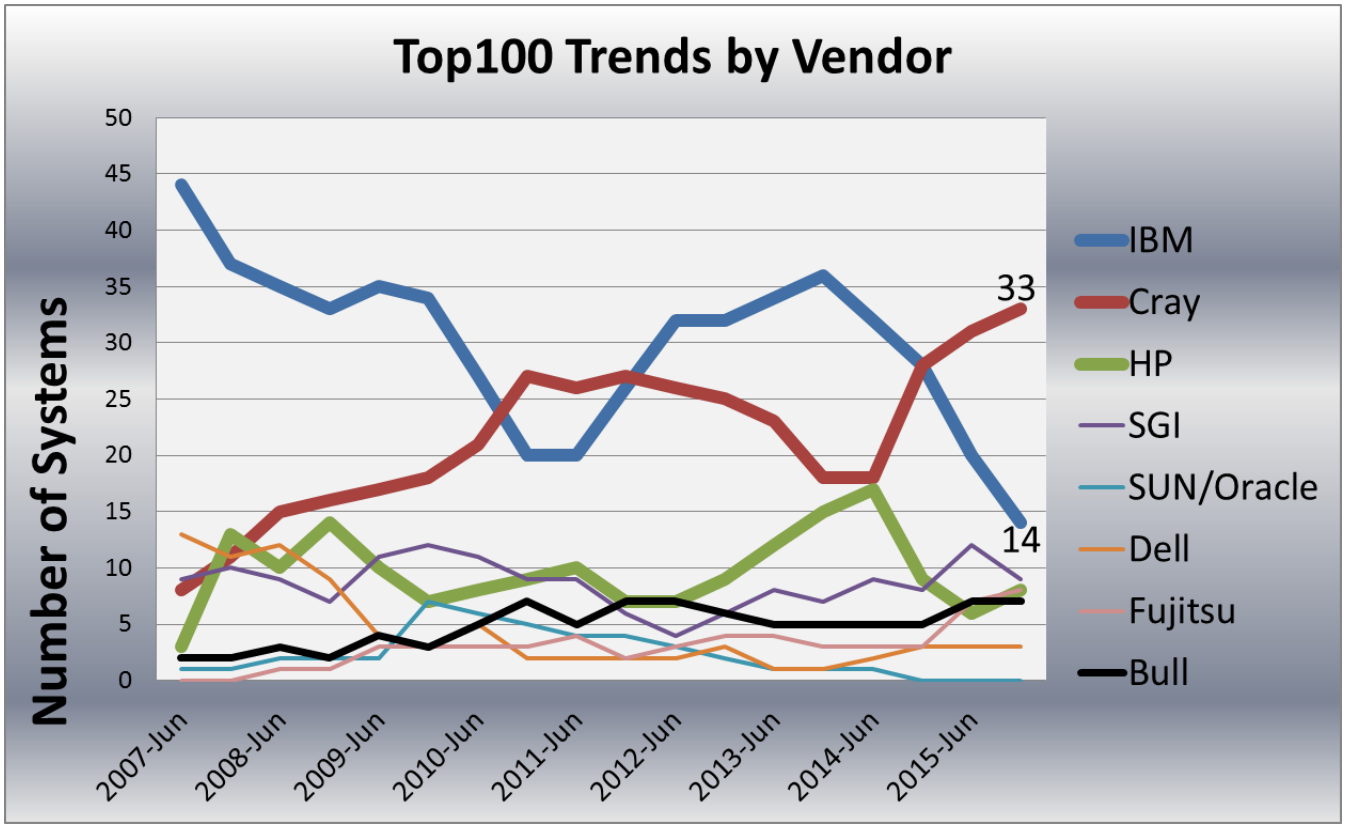
Analyze

- Graph Analytics
- Hadoop & Spark based Analytic Solutions



Anything that can be simulated needs a Cray

Top100 - Trends by Vendor

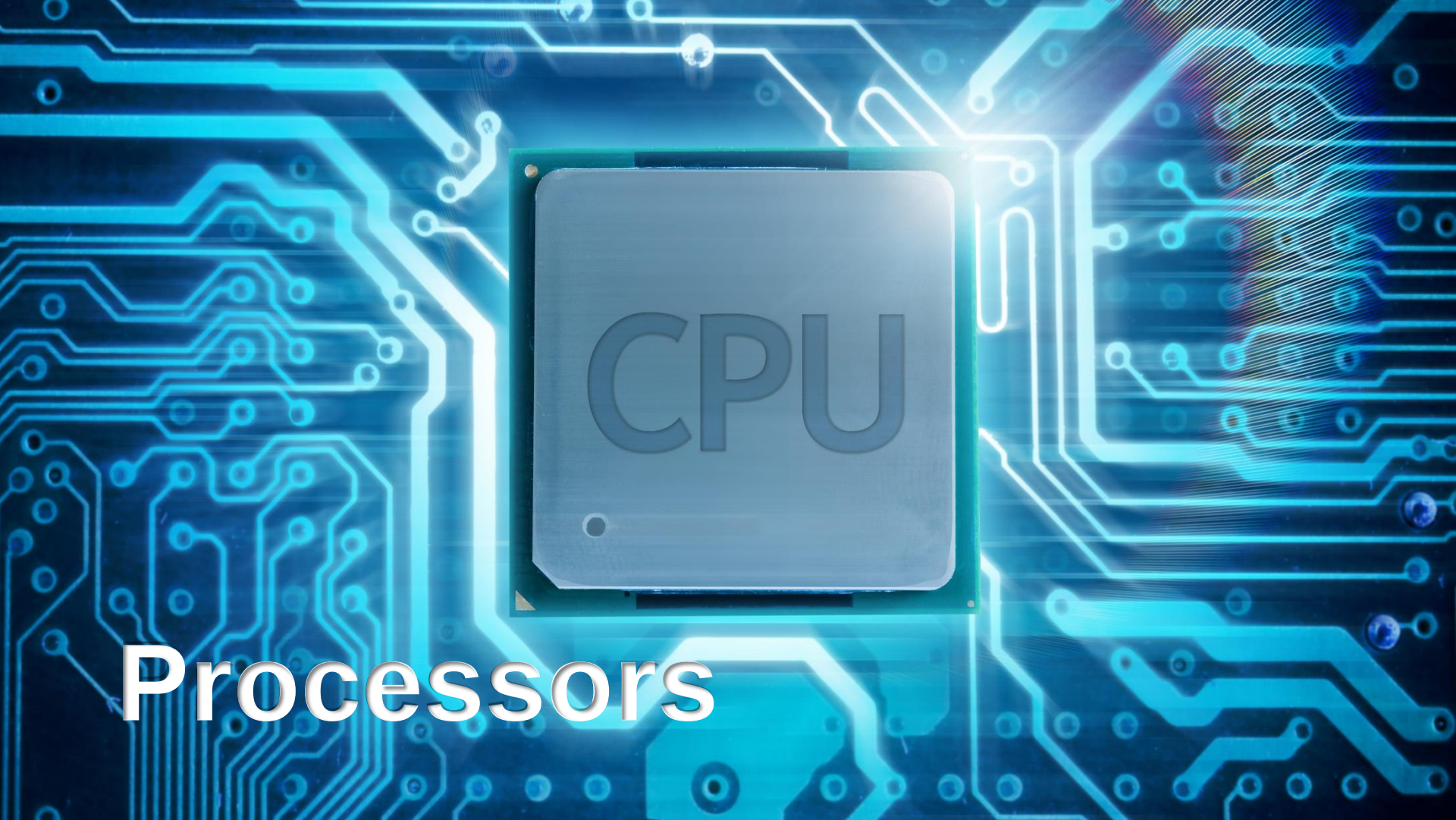


Petascale XC Series Installs & Orders

Technology Shifts and Responses…

- It has been 40 years since the first Cray-1 Supercomputer shipped to Los Alamos
- Optimal design has always been dependent on underlying technologies
  - Processors & Memories
  - Storage
  - Interconnects
- Shifts in these technologies can (and will) have large impacts on how systems look and how they are programmed

This talk will look forward to future technology shifts and their impact on system design and use



Processors

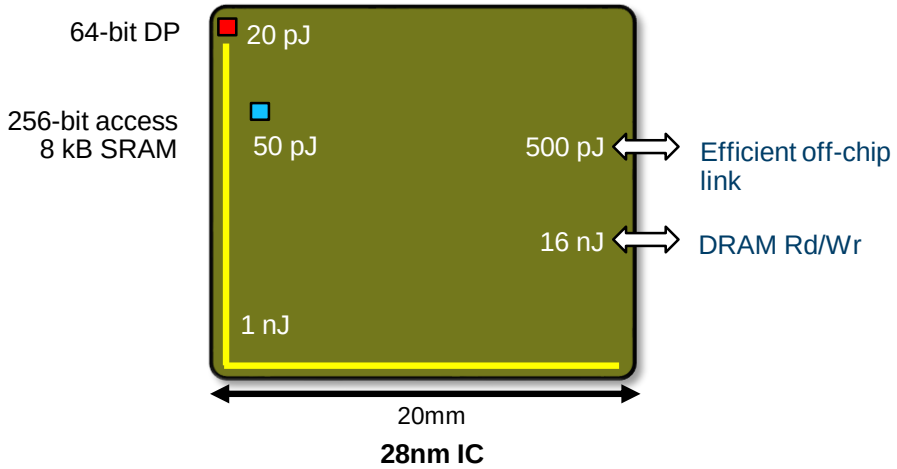
Post Dennard Scaling and the Power Wall
(2005 onward)

Voltage no longer drops with feature size

⇒ **perf/W/year** has slowed **dramatically (70% ➔ 20% CAGR)**

⇒ Have become *power* constrained

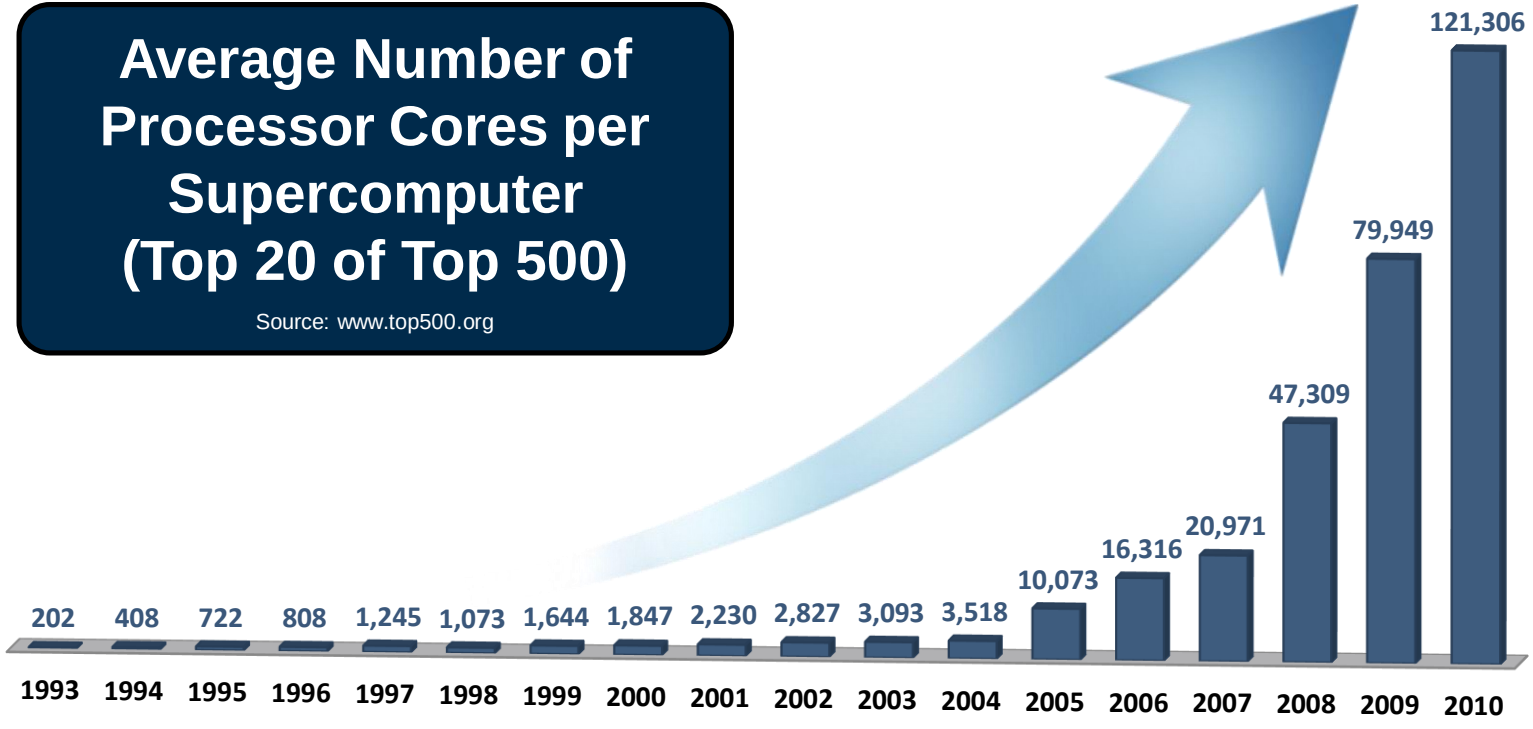
Communication much more expensive than computation



Architectural Response – Multicore Processors

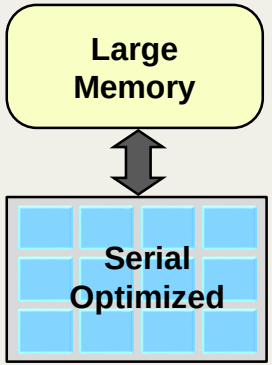
Average Number of Processor Cores per Supercomputer (Top 20 of Top 500)

Source: www.top500.org

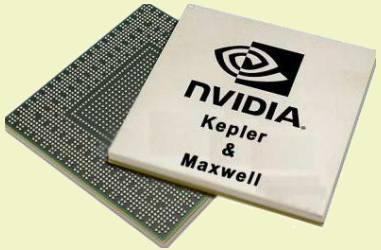


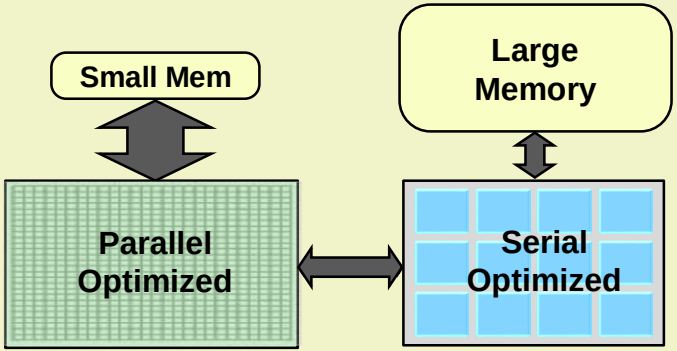
And a New Processor Landscape…

Multicore Processors
Xeon

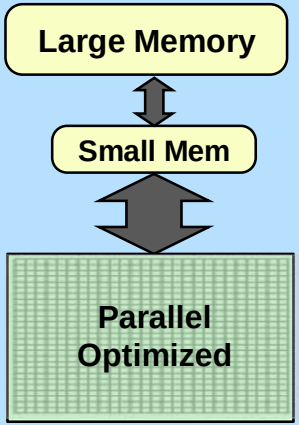


GPU computing (Nvidia Kepler)
Lots and lots of *much* simpler processors

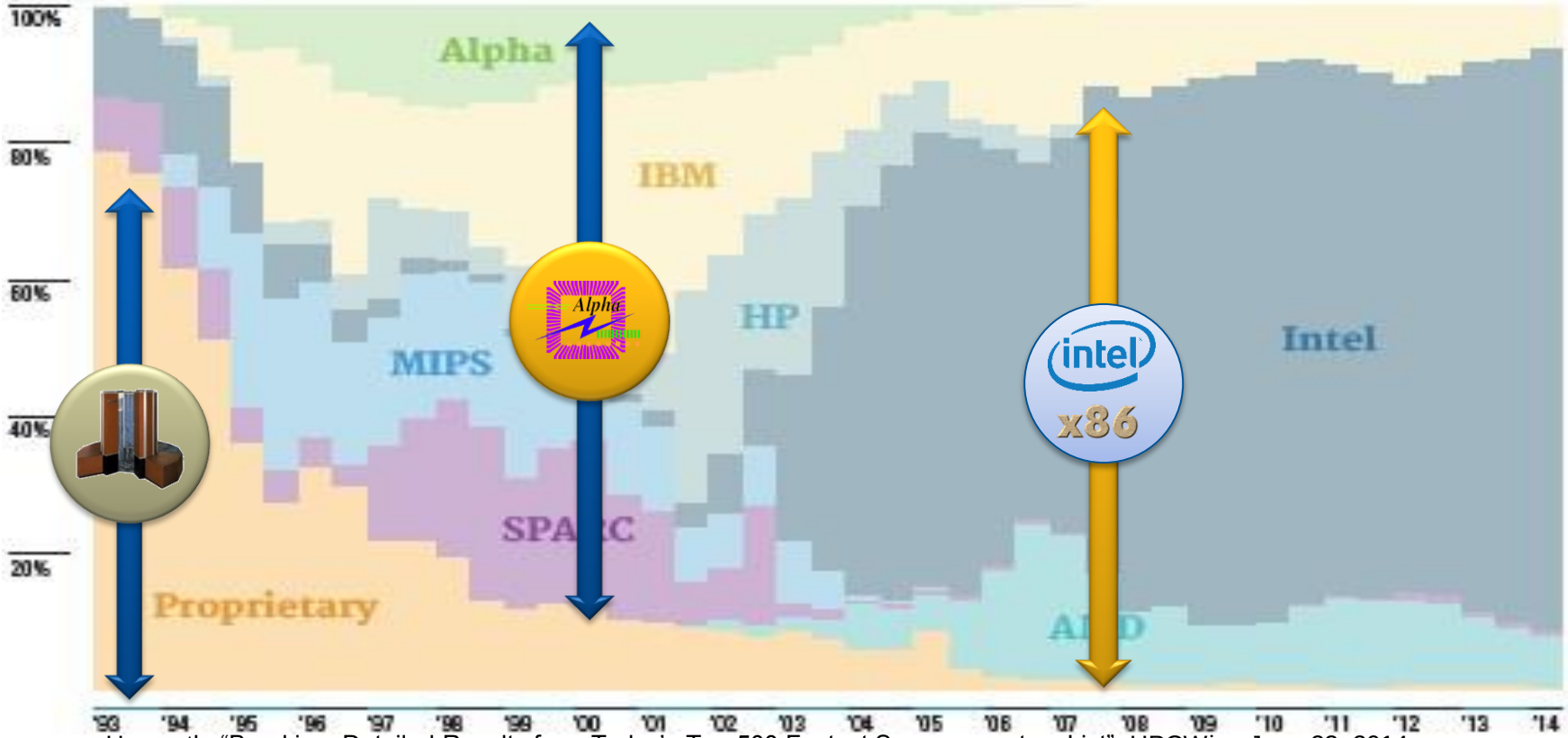




Many Core
Simpler cores with
Vector Processing



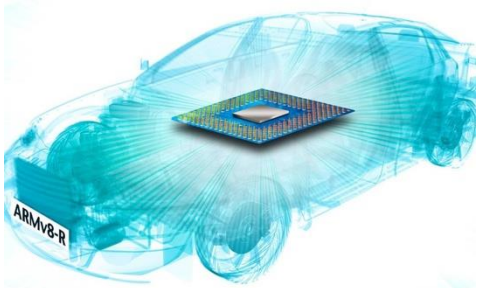
Disruptions in the Top500 Over Time

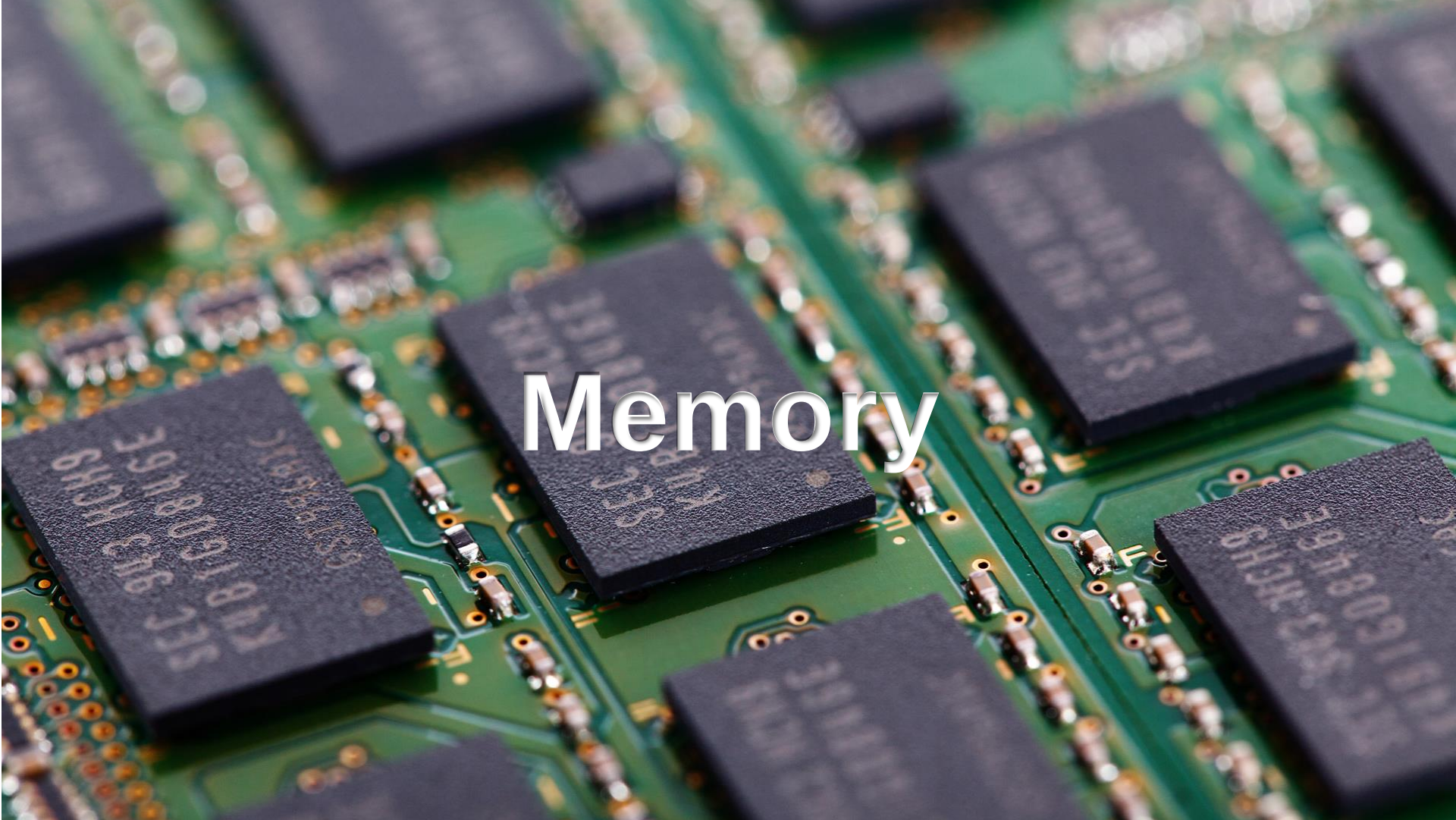


Hemsoth, "Breaking: Detailed Results from Today's Top 500 Fastest Supercomputers List". HPCWire, June 23, 2014

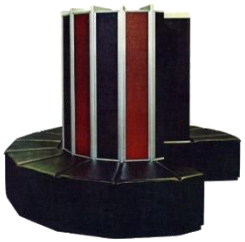
The next commodity disruption?

- ARM is *much* higher volume than x86, but is mostly used in consumer products (phones, tablets, automobiles)
- ARM64 is starting to ship its first real server parts
- ARM64 is targeting high-perf, high-volume markets




Memory

Memories Through the Decades



1980s

Small, Fast Memories

- Out-of-core programming
- Single Core
- Vectors

1990s

Large, DRAM-based memory, local memory

- Convert to in-core algorithms
- Still single-core
- Vectors

2000s

Distributed DDR Memory - caches

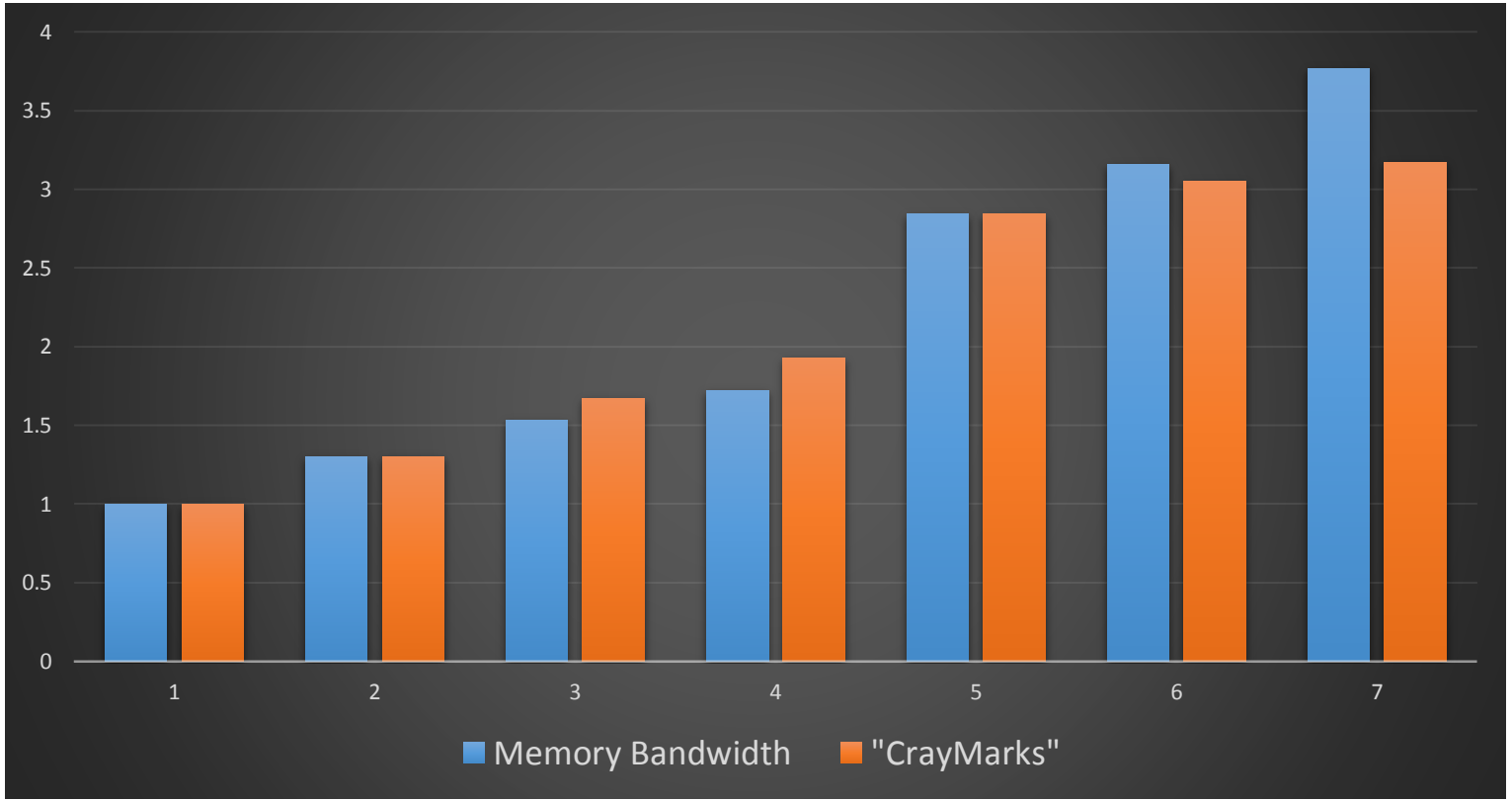
- Big Disruption! Convert to MPI
- Vectorization not so important

2010s

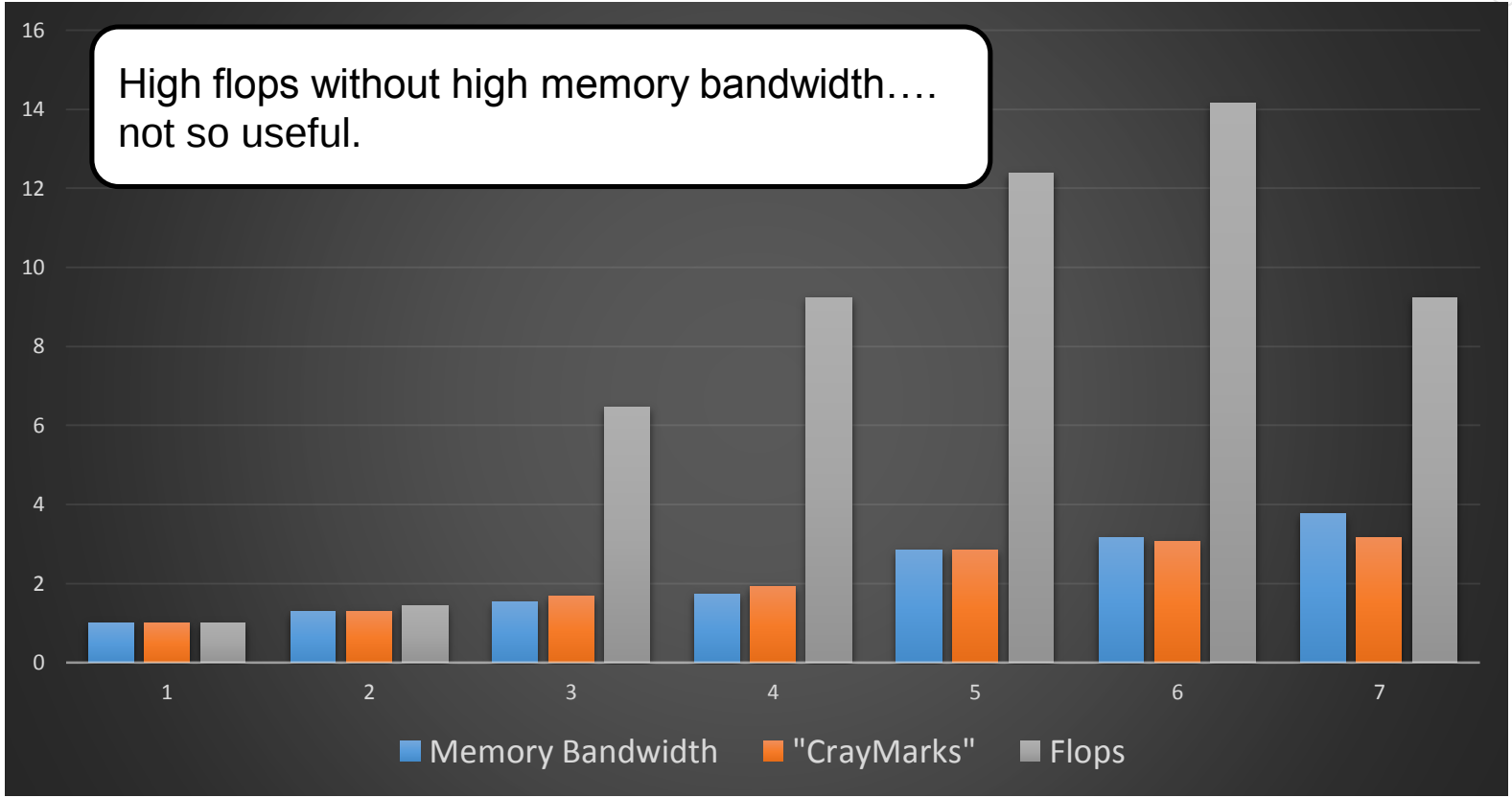
More cores per node- Some “hybrid” nodes with fast memories

- Some conversion to CUDA
- Some hybrid programming
- X86 “Vectors”

Application Performance Correlates with Bandwidth

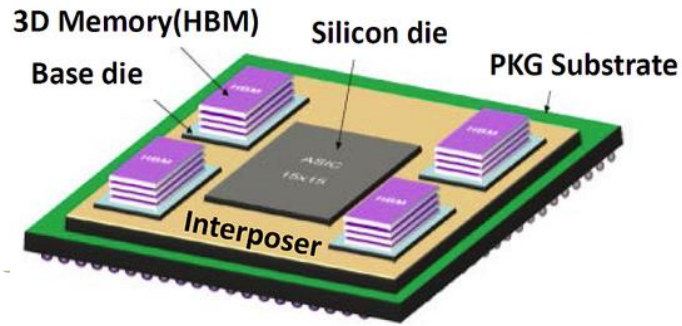


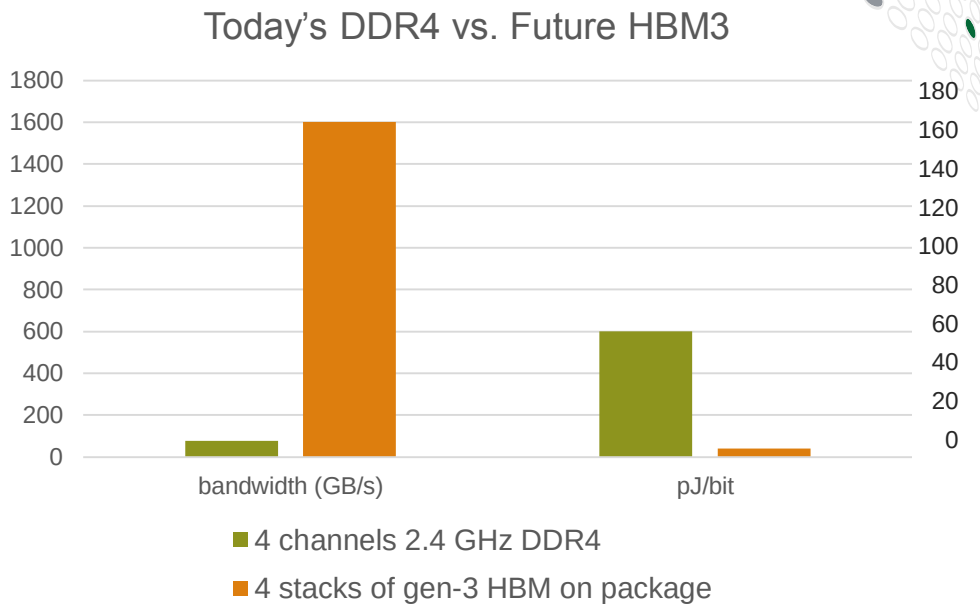
…Not FLOPS



Looking Forward - Expect a *Strong* Migration to HBM

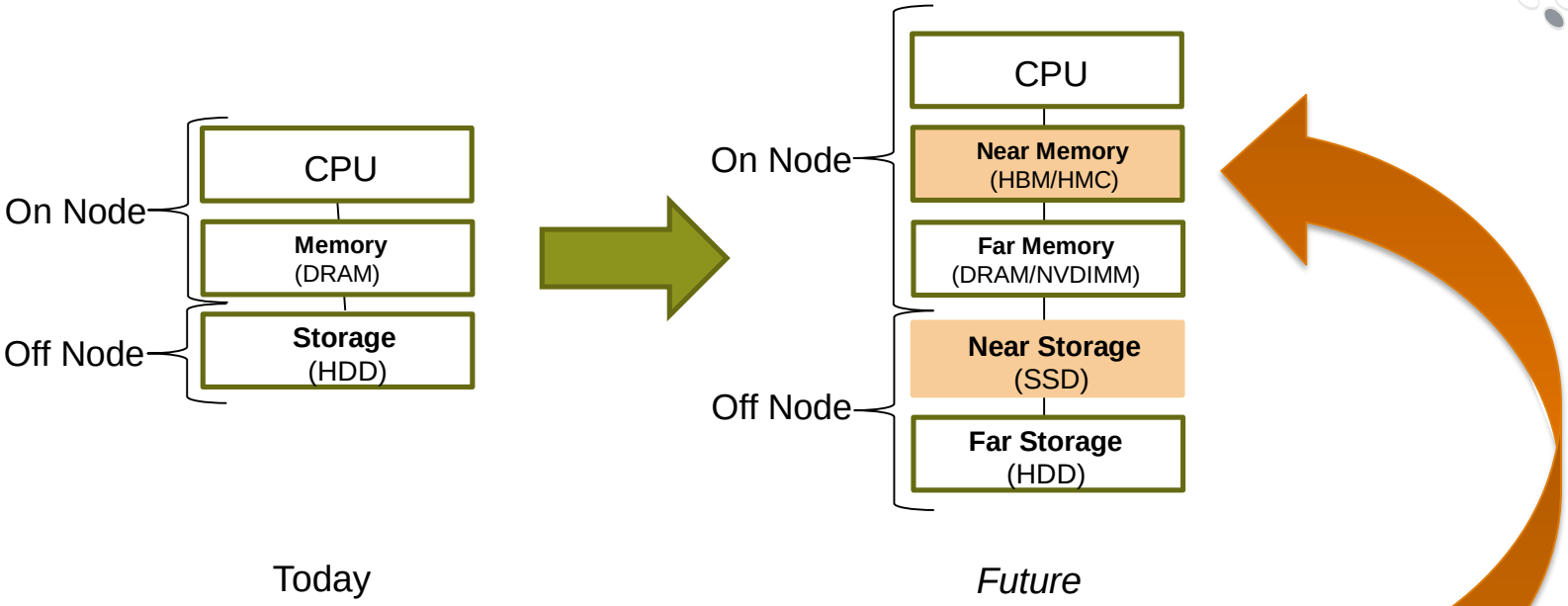
- Standard DDR memory BW has not kept pace with CPUs
- Many processors will be adopting stacked, on-package memory
- HBM:
  - 10x higher BW, 10x less energy/bit
  - Much lower latency
  - Costs less than 2x DDR4 per bit
  - JDEC standard with multiple sources





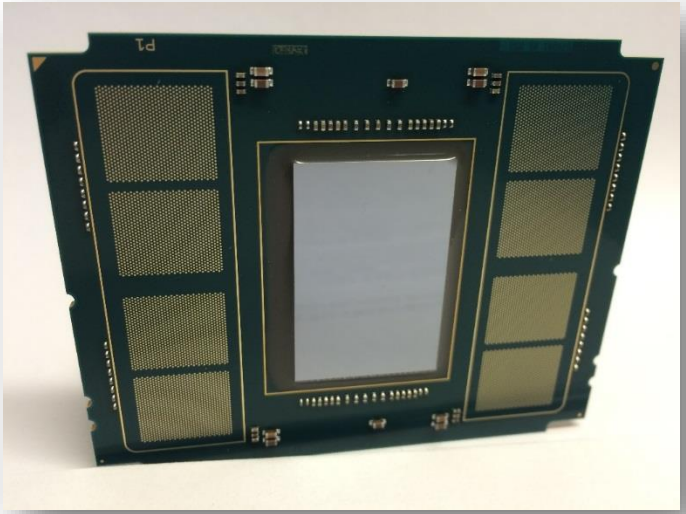
May want more, smaller nodes, with better BW and capacity per Flop

Exascale Computing Memory Trends

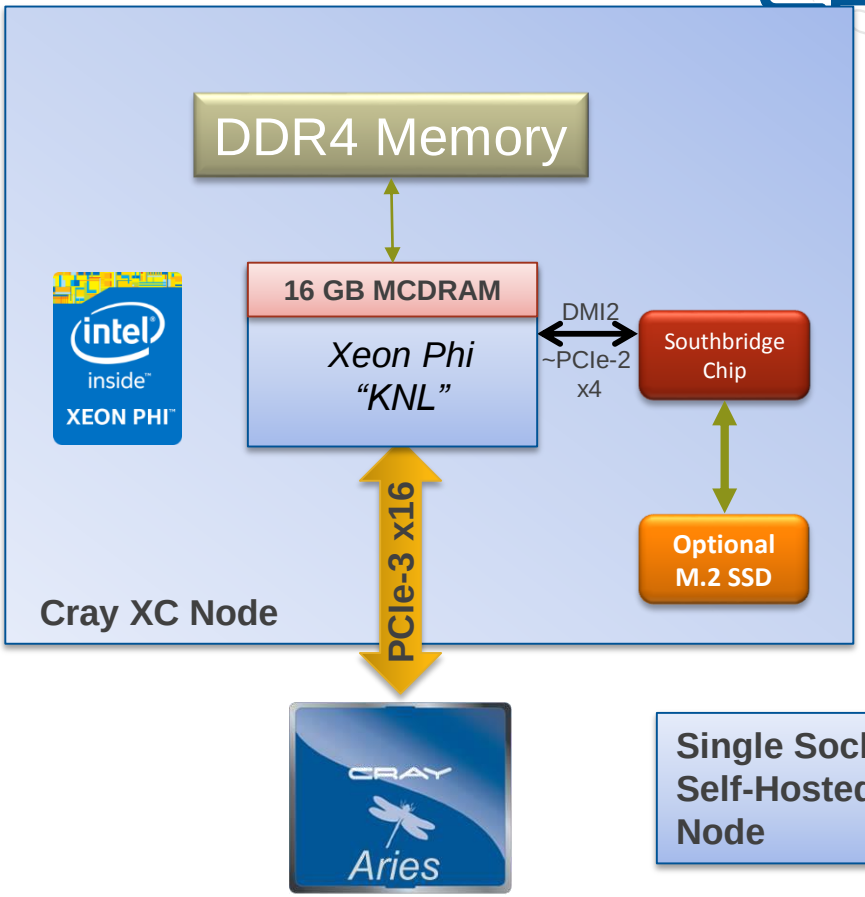


Recent NERSC8 and Trinity systems will contain "Near Memory" Layers

Xeon Phi KNL Node - 2016



- >60 Cores
- >3 Tflops
- Up to 16GB High Bandwidth Memory (~5X DRAM BW)
- Flexible Memory Modes



Single Socket, Self-Hosted Node

Manycore Parallelism – Likely more than MPI alone

Multicore vs. Manycore Parallelism

	Intel Ivy Bridge		Intel Xeon-Phi (Knights Landing)		
TLP	24	12 cores; 2 hardware threads each	TLP	240+	60+ cores; 4 hardware threads each
DLP	4	256 bit wide vector unit	DLP	8	512 bit wide vector unit

- TLP: Thread Level Parallelism
- DLP: Data Level Parallelism

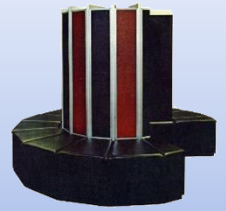
U.S. DEPARTMENT OF ENERGY | Office of Science

- 6 -

Computers are not getting faster…

Just wider

Programmers will need to deal with multi-level parallelism

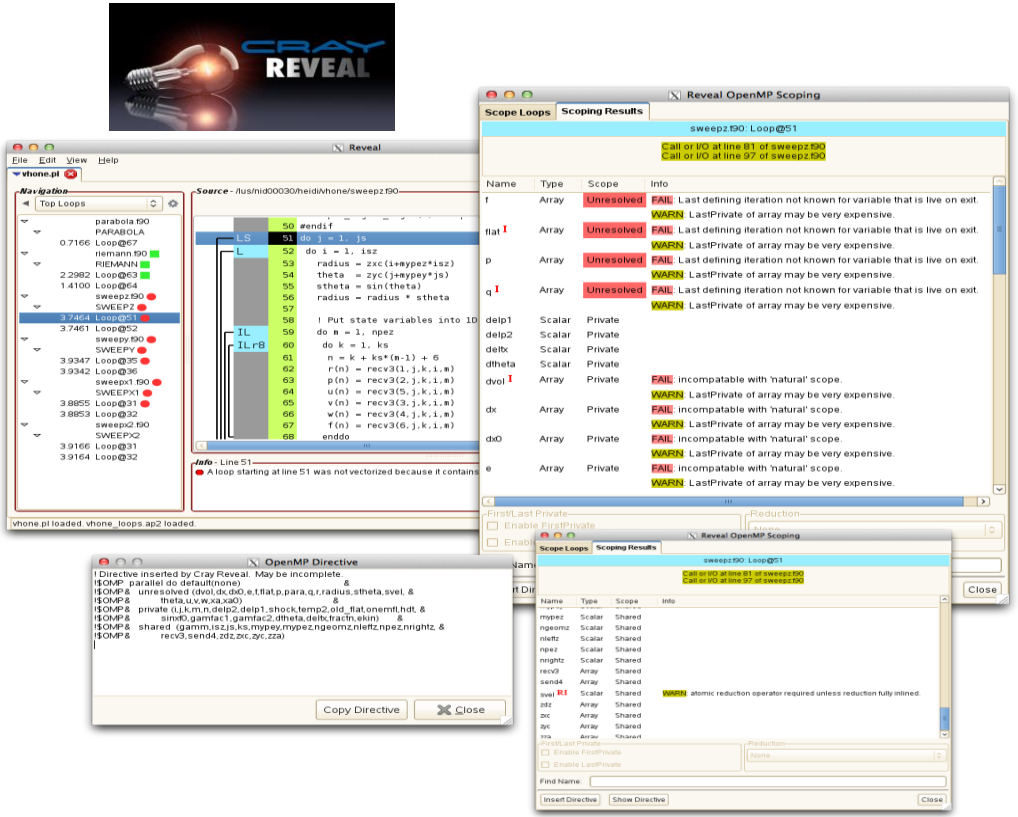


Vectors are back!

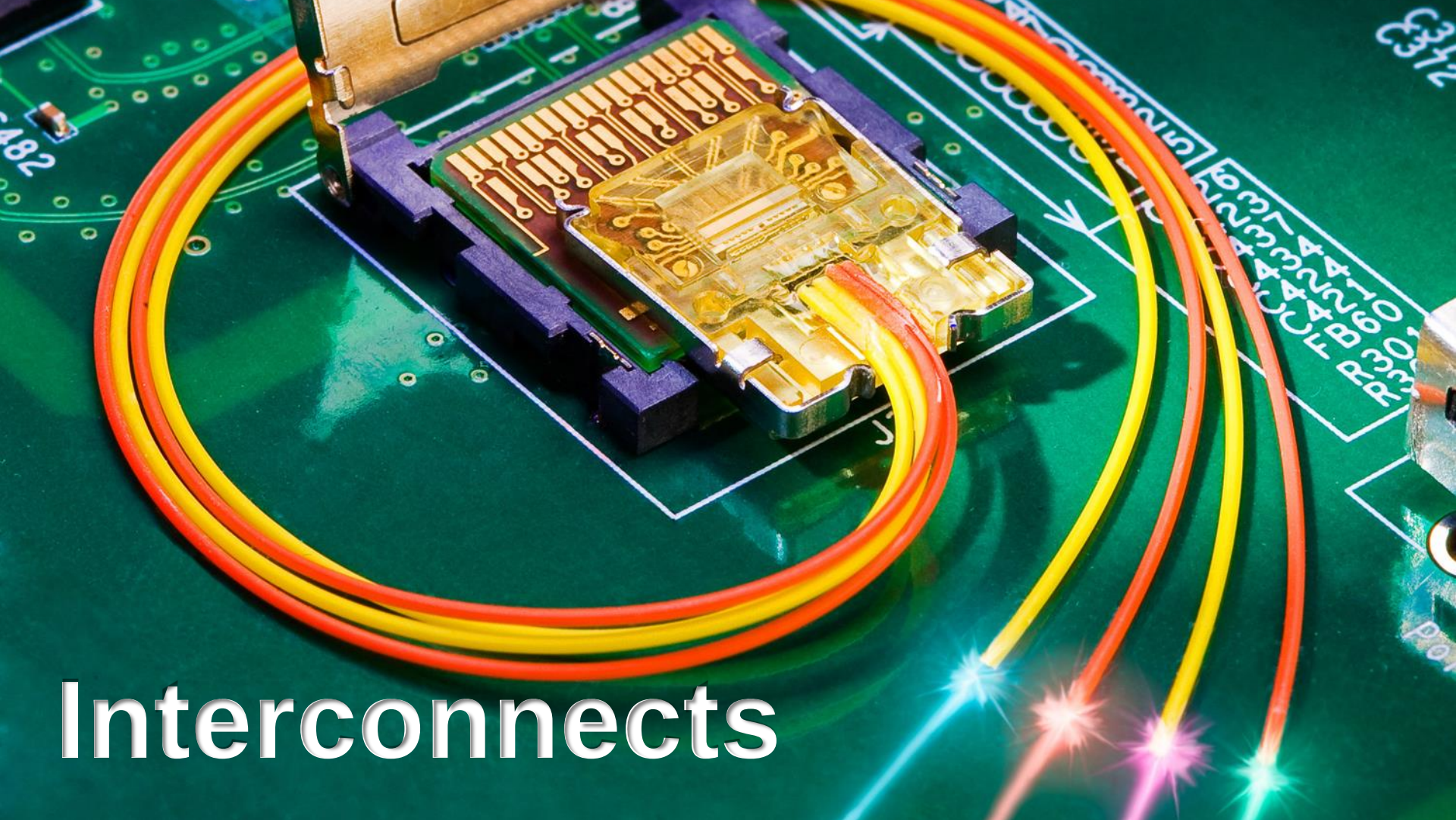
Slide courtesy of Scott French -NERSC User Services Group

ANALYZE

Simplifying Hybrid Conversion with Reveal



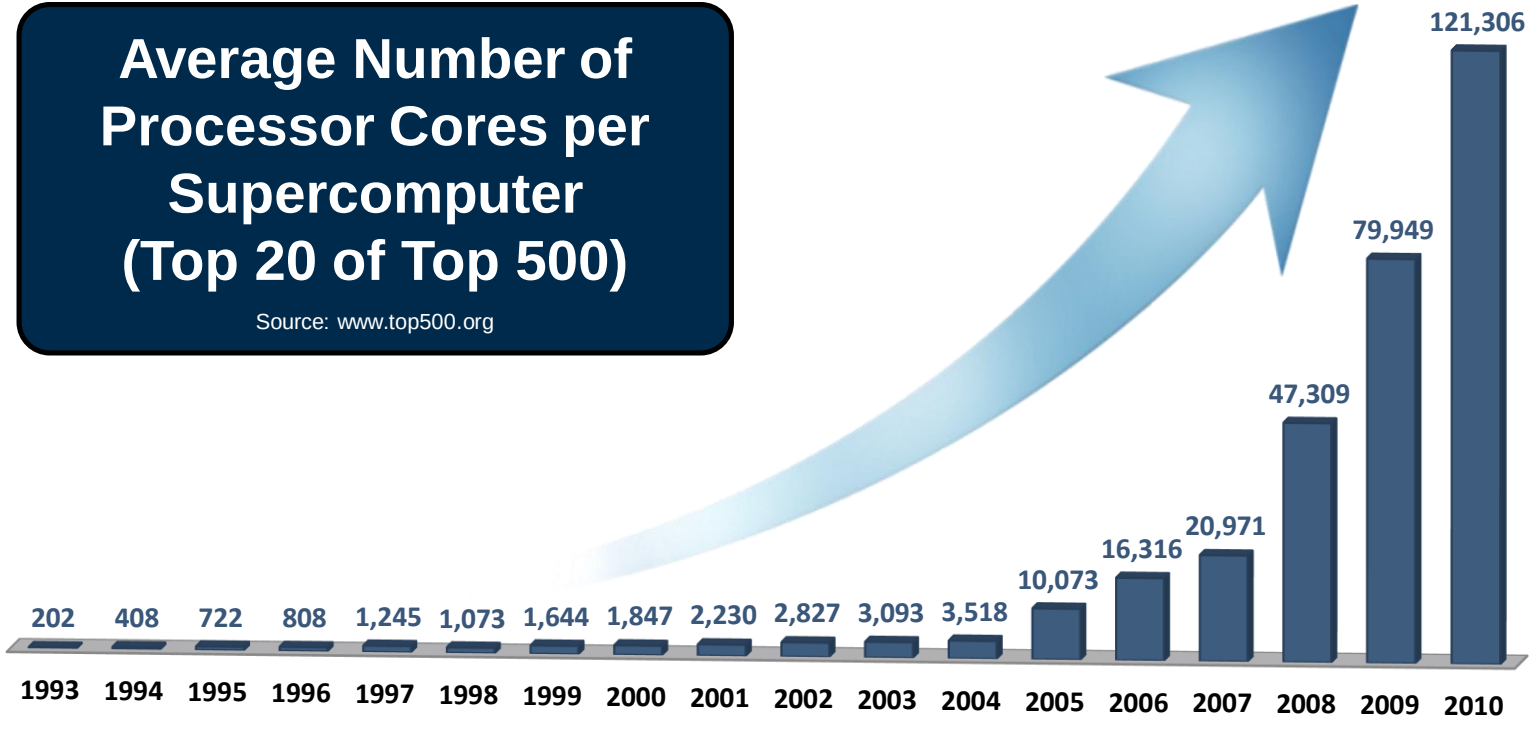
- Navigate to relevant loops to parallelize
- Identify parallelization and scoping issues
- Get feedback on issues down the call chain (e.g.: shared reductions)
- See vectorization and other compiler optimizations
- Optionally insert parallel directives into source
- Validate scoping correctness on existing directives


Interconnects

Interconnects Matter More and More…

Average Number of Processor Cores per Supercomputer (Top 20 of Top 500)

Source: www.top500.org



Cray Interconnects Over Time

1994

- **Cray T3E**
  - 3D Torus
  - E-Registers
  - PVM, Shmem and later MPI
  - Copper Cables

2003

- **Cray XT3 - SeaStar**
  - 3D Torus
  - Scalability – Connectionless protocol
  - Backplanes & Copper Cables

Cray Interconnects Over Time (continued...)

2010

- **Cray XE6 – Gemini**
  - Gemini Router
  - Still a 3D Torus!
  - Fast Message Rates
  - Resilience
  - Backplanes & Copper Cables

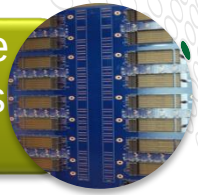
2012

- **Cray XC30 - Aries**
  - **Dragonfly – 5 Hop**
  - Global Bandwidth
  - Backplanes, Copper, **Active Optics – 14 Gpbs signaling**

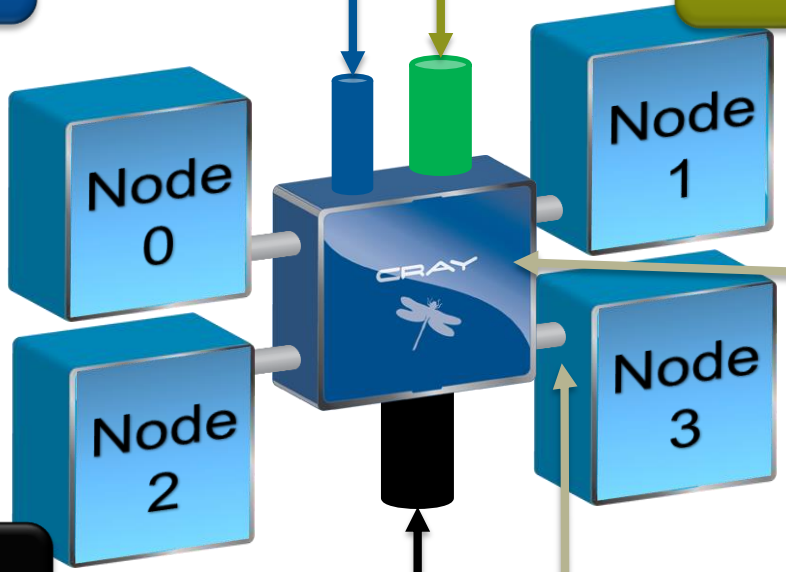
Dragonfly Topology – XC Compute Blade

Rank 3 – Optics
Cables x 10 links
12.5 Gbps

Rank 1 – Backplane
Cu Cables x 15 links
14 Gbps



4 compute nodes per blade



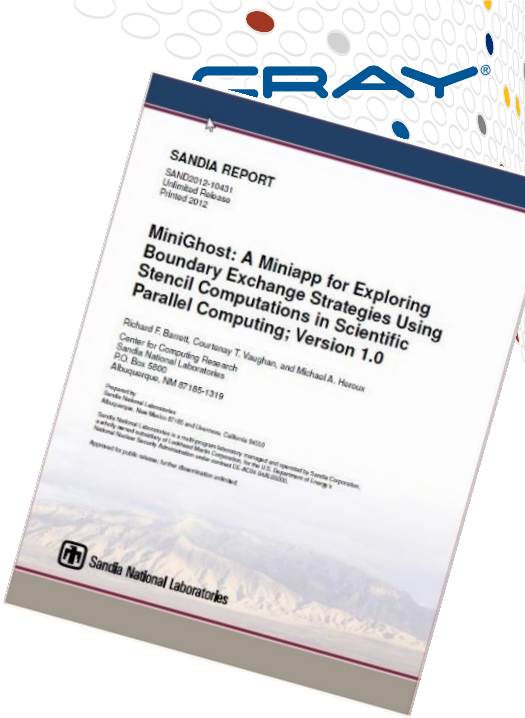
Aries 48-port Router
4 NICs, 2 router tiles each
40 router tiles for interconnect

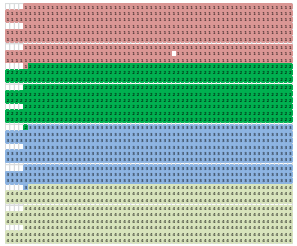
Rank 2 – Copper
Cables x 15 links
14 Gbps

PCIe-3
16 bits at 8.0 GT/s
per direction

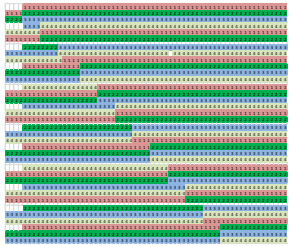
Dragonfly is placement insensitive

- Example: Sandia miniApp, miniGhost
- Running on 2256 node CSCS system (¼ global bandwidth)
  - Runtime in seconds for 100 cycles

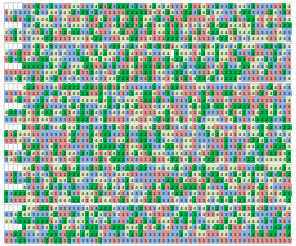




Contiguous Blocks of 512 nodes			
69.0	68.8	68.9	68.9



Random blocks of 64 nodes			
69.4	69.4	69.4	69.5



Random layout of nodes			
70.9	71.0	70.6	70.5

Perfect Placement

< 3% variance from best-case to worst-case placement

Worst-Case Placement

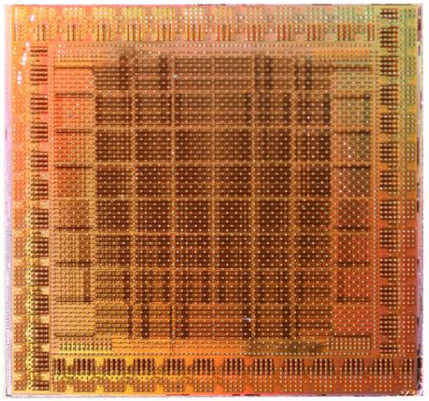
Two Recent Large Systems Compared

- Topology – 24x24x24 Torus
- Longest Minimal Hop Count – 36
- Global Bandwidth – 21 TB/sec

- Topology – Dragonfly
- Longest Minimal Hop Count – 5
- Global Bandwidth – 218 TB/sec

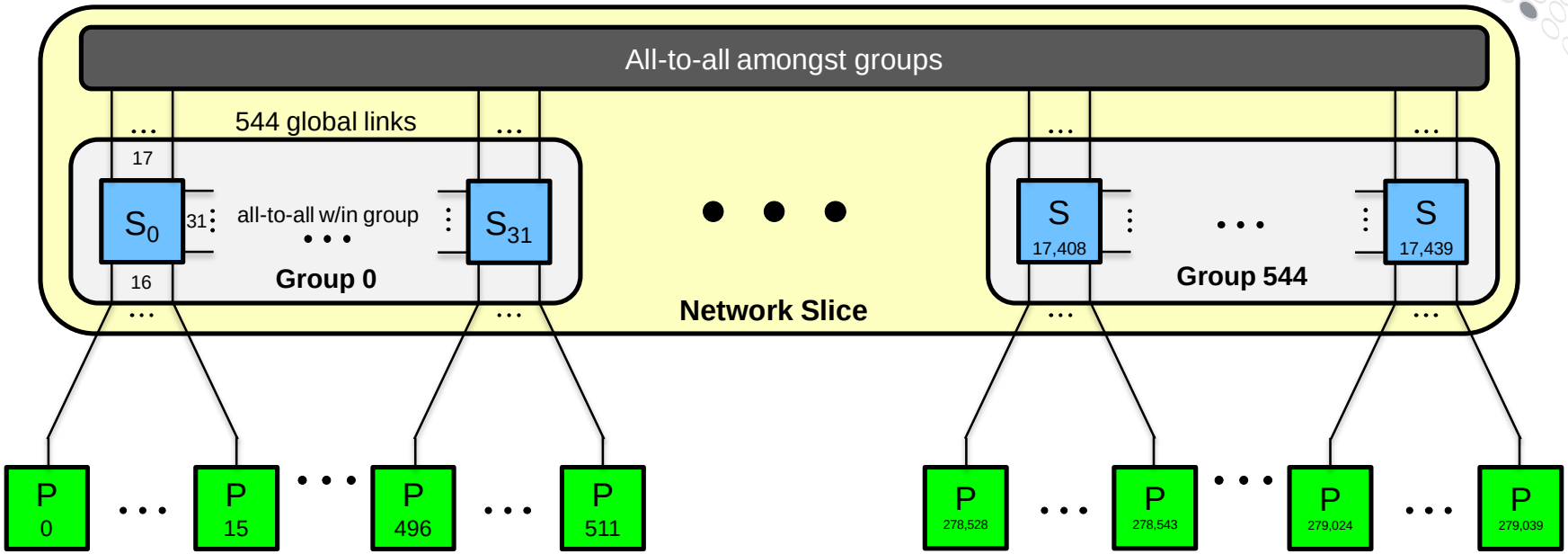
Predictions for the Future

- **Future interconnects will build on what we have done**
- **Expect to see these technologies in an "Exascale" system**
  - Higher Radix Routers
  - Higher Signaling rates
  - Continued mixed use of electrical and optical signaling
  - Lower Diameter Topologies



First 64 port router
Cray X2 (2005)

Example Dragonfly Network with a 64-port Switch

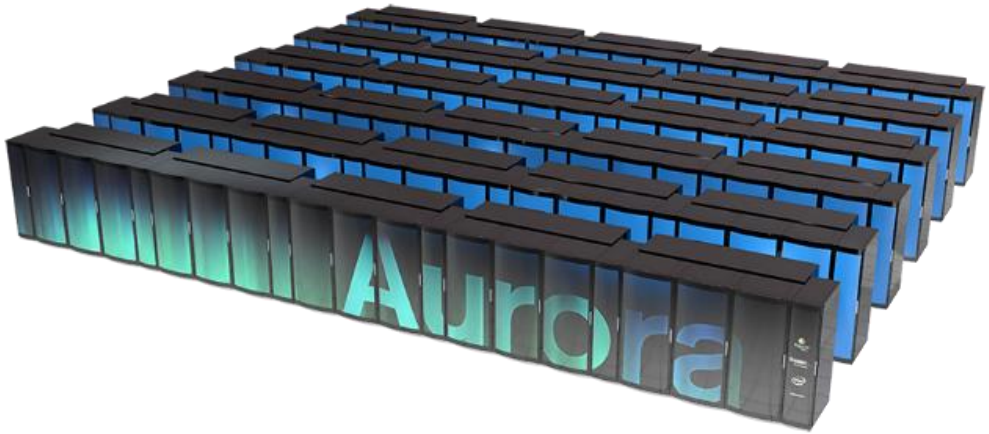


- Scales to 279K endpoints, with a network diameter of 3 hops!
- Only a *single* hop over a long (optical) link

Aurora – Our First Shasta Order

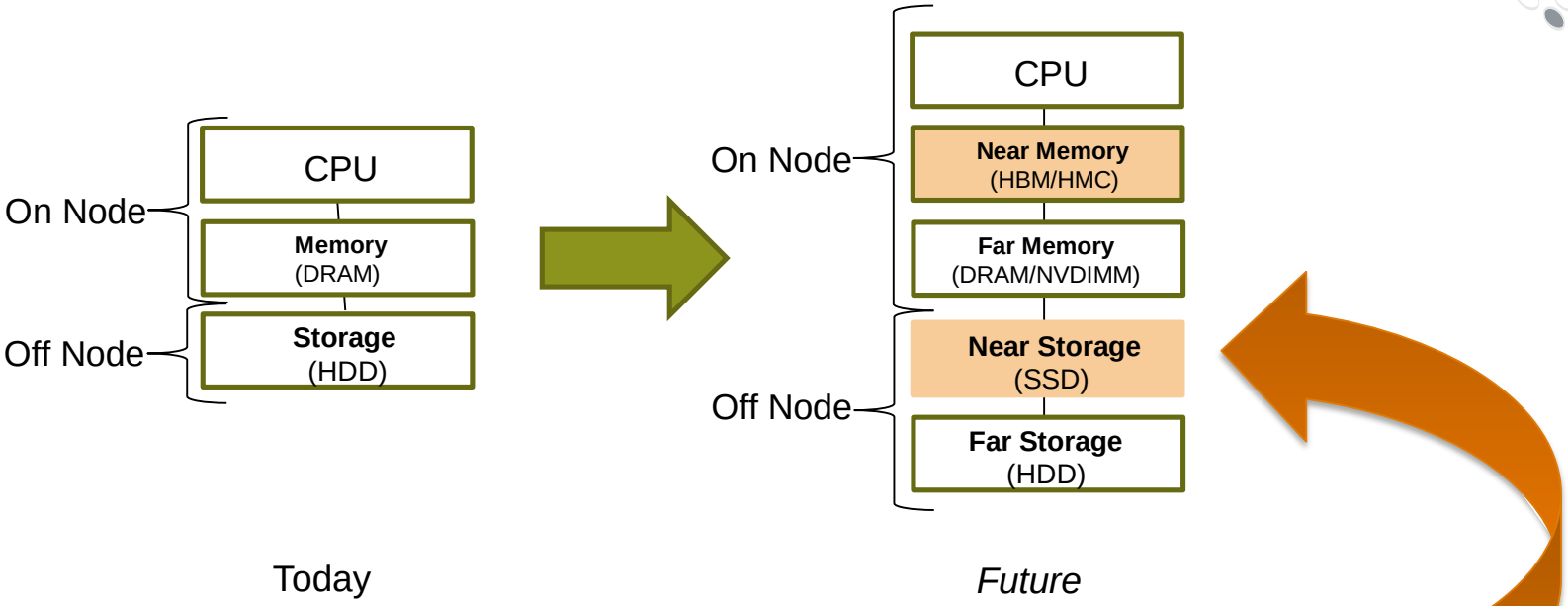
- **180 PetaFlop Peak Supercomputer Performance**
- **DOE/ANL ALCF (Argonne Leadership Computing Facility) Win based on Cray's:**
  - Scalable, adaptive supercomputing architecture collaboration and development with Intel
  - Targeting both supercomputing and analytics workloads

Argonne NATIONAL LABORATORY



Storage

Exascale Computing Memory Trends



Recent NERSC8, Trinity and KAUST orders will contain both of these "future" technologies

Cray DataWarp I/O Acceleration for Cray XC40

- **Pure performance**
  - 70 thousand to 40 million IOPS per system
  - Quality of Service to applications
- **Breakthrough efficiencies**
  - ***5x the bandwidth of disk at the same cost***
- **Flexible Usage Models**
  - Local and Shared I/O models
  - No application changes required

CRAY DATAWARP™

DataWarp overcomes the performance gap between compute and disk storage

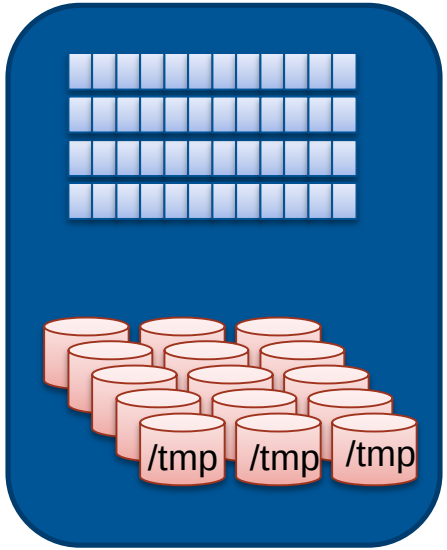
Use Case: Local Storage on Demand

Per Node Scratch

- Each compute node in a job is assigned a private part of the allocated SSD space
- Much faster than “faking it” with a parallel file system

Per Node Swap Space

- Compute node swap space
- Protects from accidental memory overflow

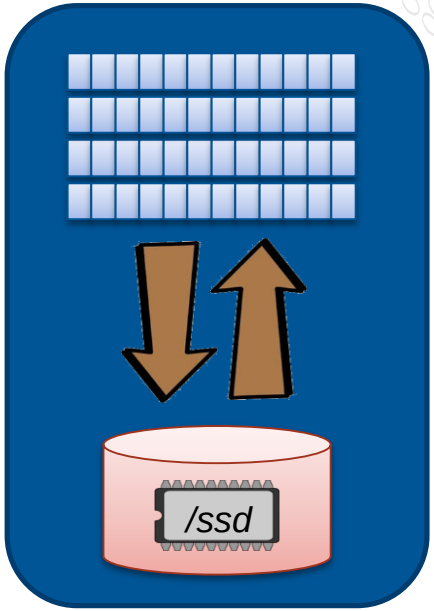


CRAY DATAWARP™

Use Case: Shared Fast /ssd

Shared Fast Scratch

- High Bandwidth access to shared files
- Files can be striped across multiple DataWarp Nodes
- Space can be temporary for the job, or be marked as persistent to work between jobs

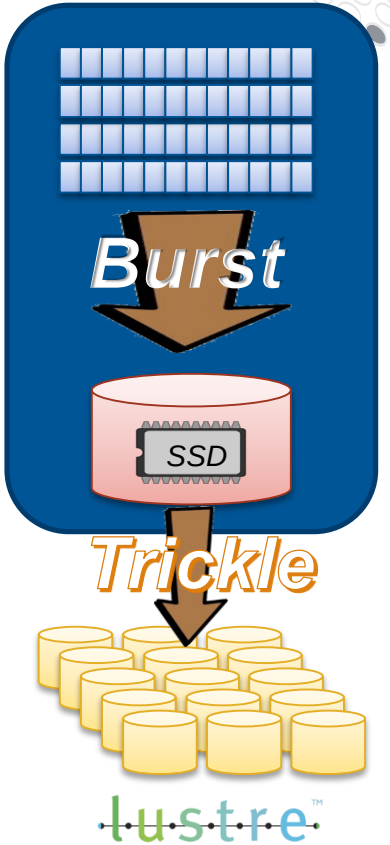


CRAY DATAWARP™

Use Case: Burst & Trickle

Burst & Trickle

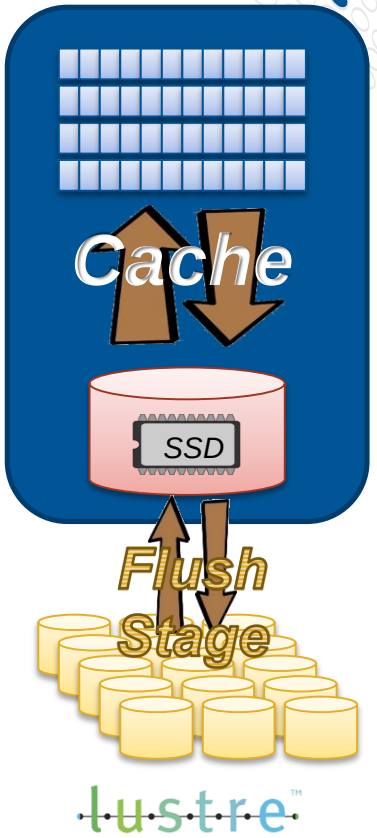
- User asks for enough SSD to cover the number of concurrently resident checkpoints
- High Bandwidth checkpoints are written to SSDs
- DataWarp library calls allow the user program to trigger asynchronous data motion at appropriate times



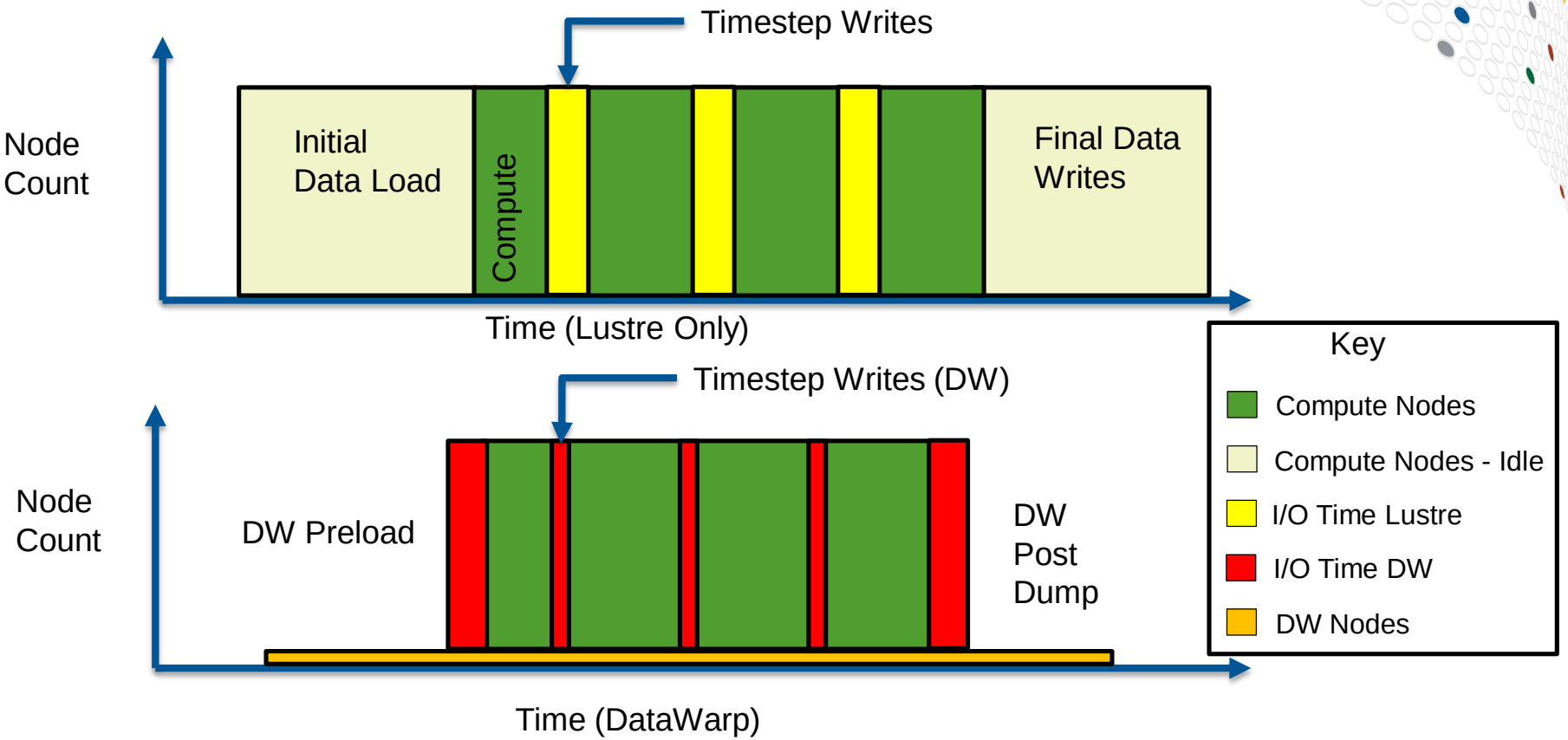
Use Case: File System Caching

Implicit File System Caching

- Global file system caching
- Both on-demand and transparent to the application
- Phase 2 Feature



DataWarp Notion – Minimize Compute Residence Time

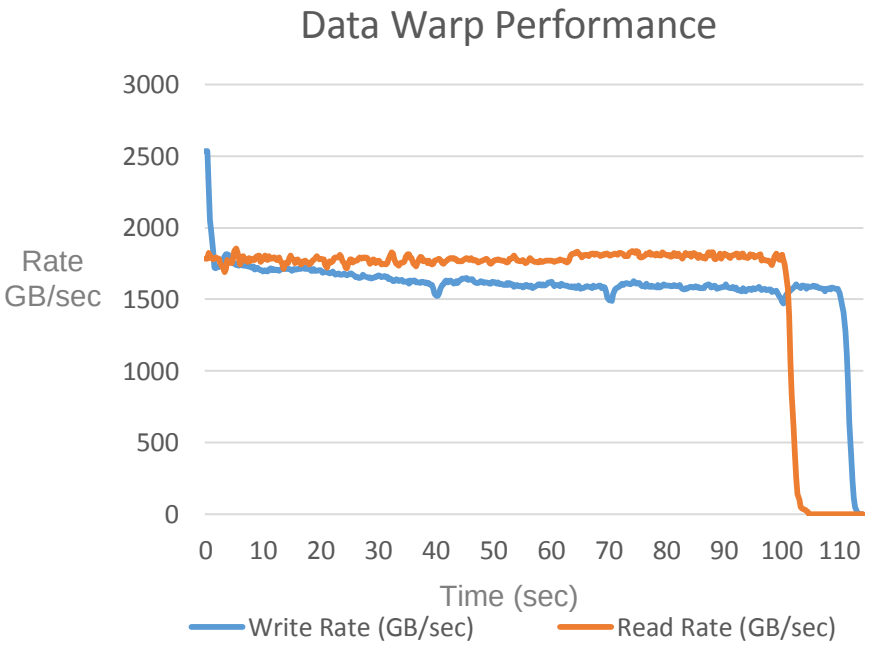


Workload Manager Integration is Key: Job With and Without DataWarp

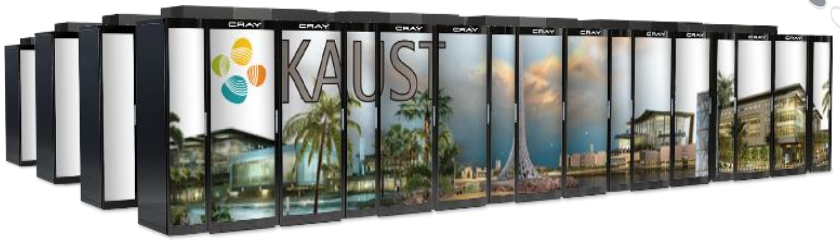
```
#!/bin/ksh
#SBATCH -n 3200 -t 2000

export TMPDIR=/lustre/my_dir

srun –n 3200 a.out
```

```
#!/bin/ksh
#SBATCH -n 3200 -t 2000

#DW jobdw type=scratch \
        access_mode=striped \
        capacity=1TiB
#DW stage_in type=directory
        source=/lustre/my_dir
        destination=$DW_JOB_STRIPED
#DW stage_out type=directory \
        destination=/lustre/my_dir \
        source=$DW_JOB_STRIPED

export TMPDIR=$DW_JOB_STRIPED

srun –n 3200 a.out
```

World Record IOR Bandwidth - KAUST



Test run with 4000 compute nodes against 264 DataWarp nodes

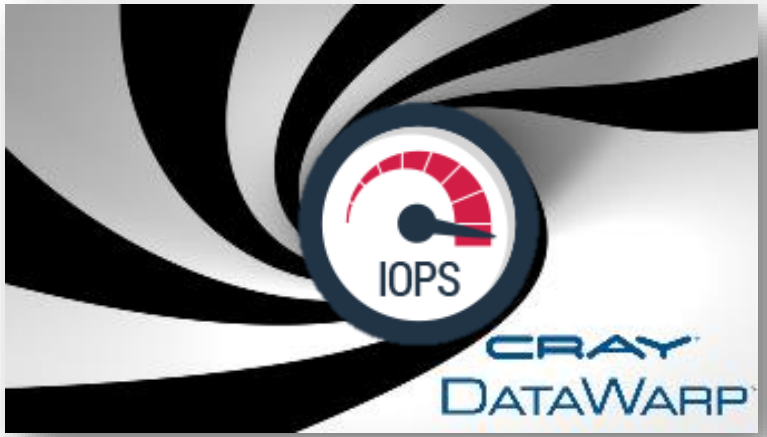


1.5 TB/sec Write
1.8 TB/sec Read

"This has been a learning experience as well as showing the new capabilities of our Cray XC40 system. Now we have all the ingredients to make scientific discoveries faster."

-Saber Feki, Computational Scientist

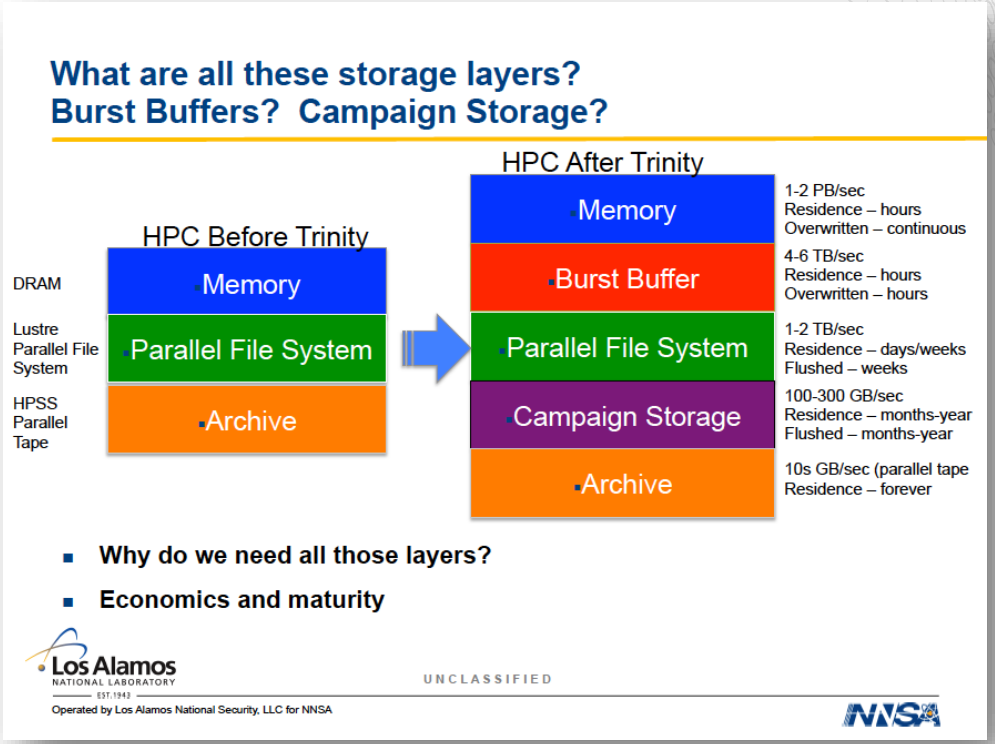
12 Million Random 4K IOPS!



140 DataWarp Nodes
4k random writes and reads
4480 1GiB Files

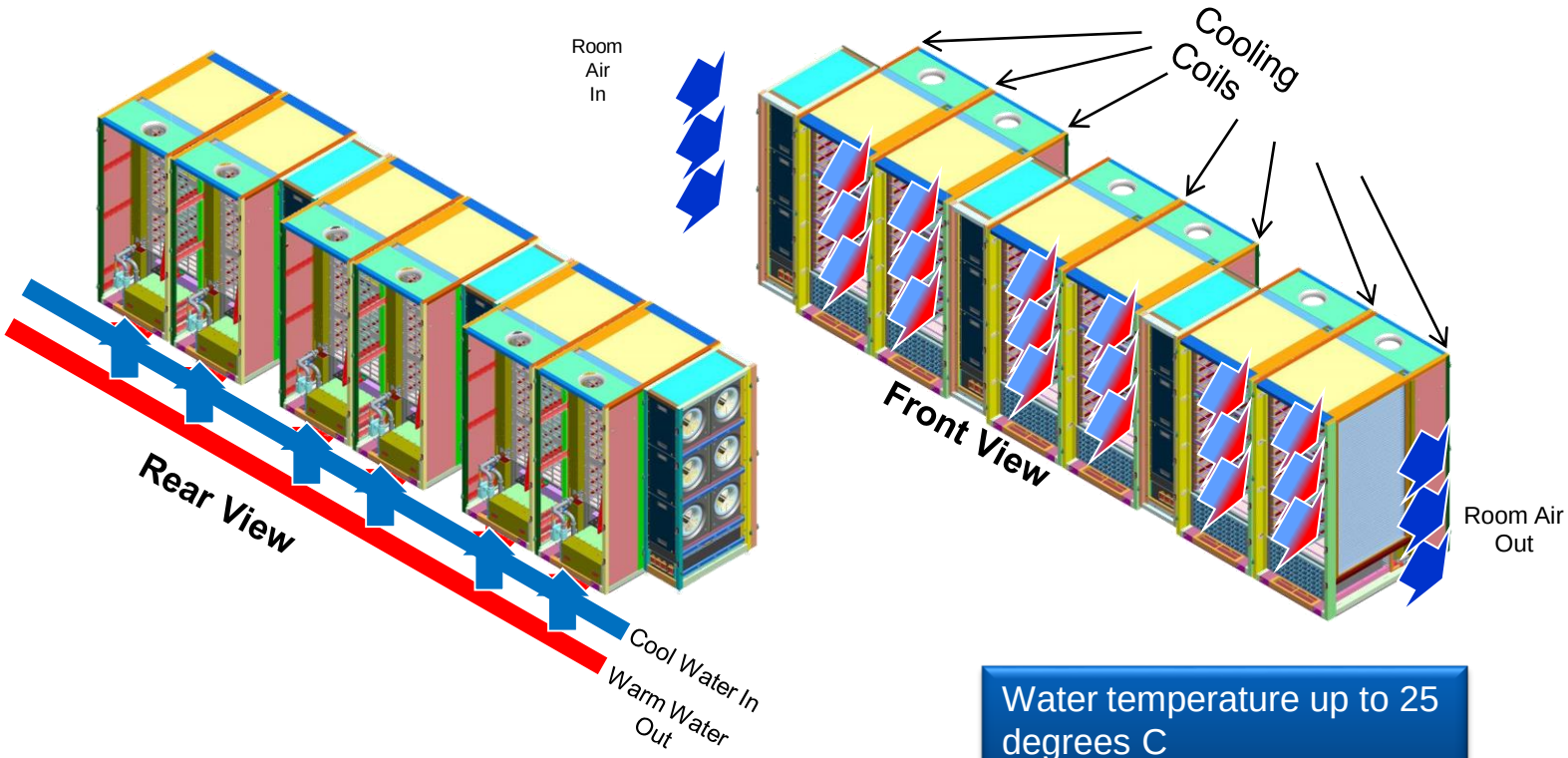
Looking Forward…

- DataWarp, or Burst Buffer will become more robust
- Cheaper, object storage will replace parallel file systems
- Clever software will automatically move data between layers with minimal user intervention




Power &
Cooling

Today XC Uses Transverse Liquid Cooling



Water temperature up to 25 degrees C

Power Management & Control

Power Monitoring

- Power Database
- Accounting plug-ins to record energy use by job
- System Level power, water flow and temperature monitoring

System Level Power Management

- System Level Power Capping
- Time of day system power adjustment
- Node Power Off / On policy
- Power charging policies

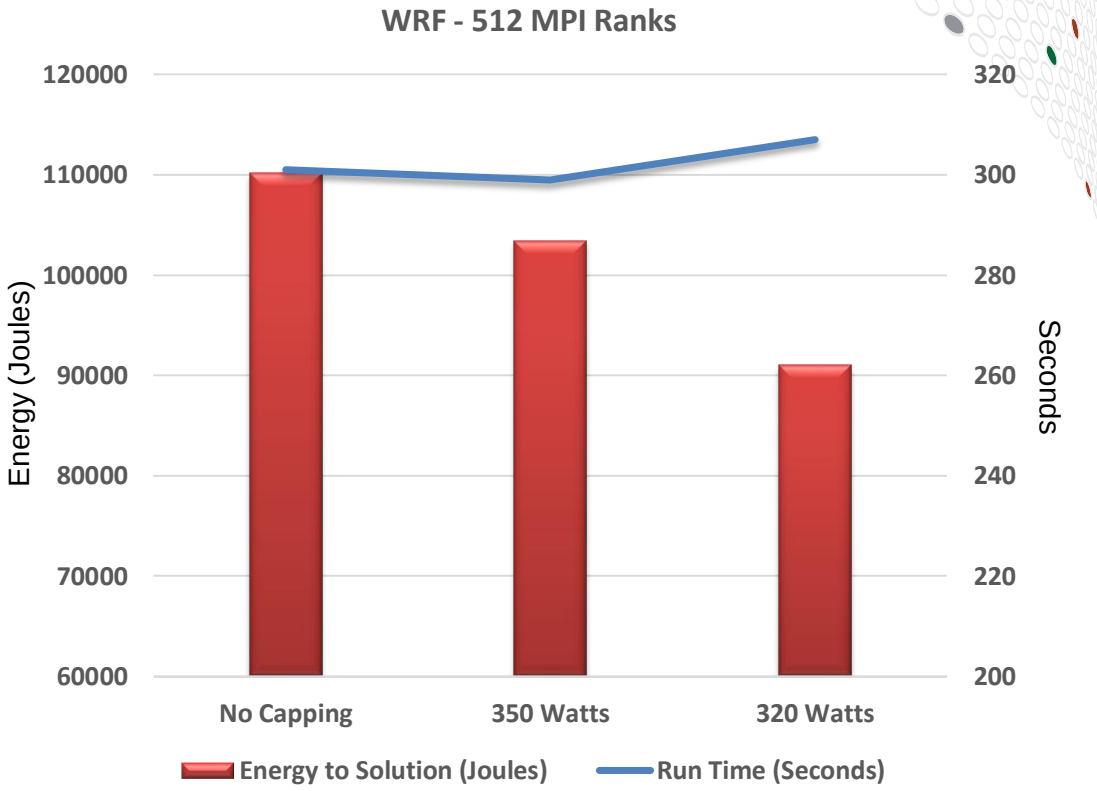
User Level Power Management

- Node power capping
- P-State Control
- Workload Manager integration
- Application power profiling

Cray Power Management Roadmap

Example - WRF

- **3 power levels tested**
  - High: No Capping
  - Medium: 350 Watts per Node
  - Low: 320 Watts per Node
- **Results**
  - Low power run required 17.4% less energy than the default settings
  - Runtime increased by just 2%



Looking Forward

- The first Exascale system will likely exceed 20MW of total power
- Direct liquid cooling will likely be required
- High voltage internal distribution may be needed to increase conversion efficiency
- 45 degree water with high flow rates will enable “free” cooling anywhere
- Power management features for ramp up and down will be required

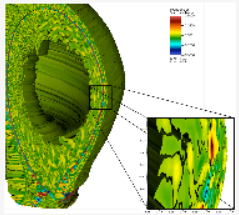

Workloads

Sustained Performance on Real World Applications - Running the largest jobs, Most Nodes, at High Utilization

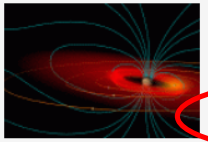
Machine Vitals

Edison	Cray XC30 **Peak TFlop/s:** 2,570 (2013)
Peak TFlop/s:	2570
Jobs running:	153
Jobs queued:	275
Cores in use:	128,784 (96%)
Backlog:	0.9 days

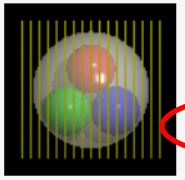
Top Jobs



Center for Edge Physics Simulation: SciDAC-3 Center
Office: Fusion Energy Sciences
Investigator: Choong-Seock Chang
Science Area: Fusion Energy
Cores: 30,720 (Edison)
Core Hours Used: 14,983.2

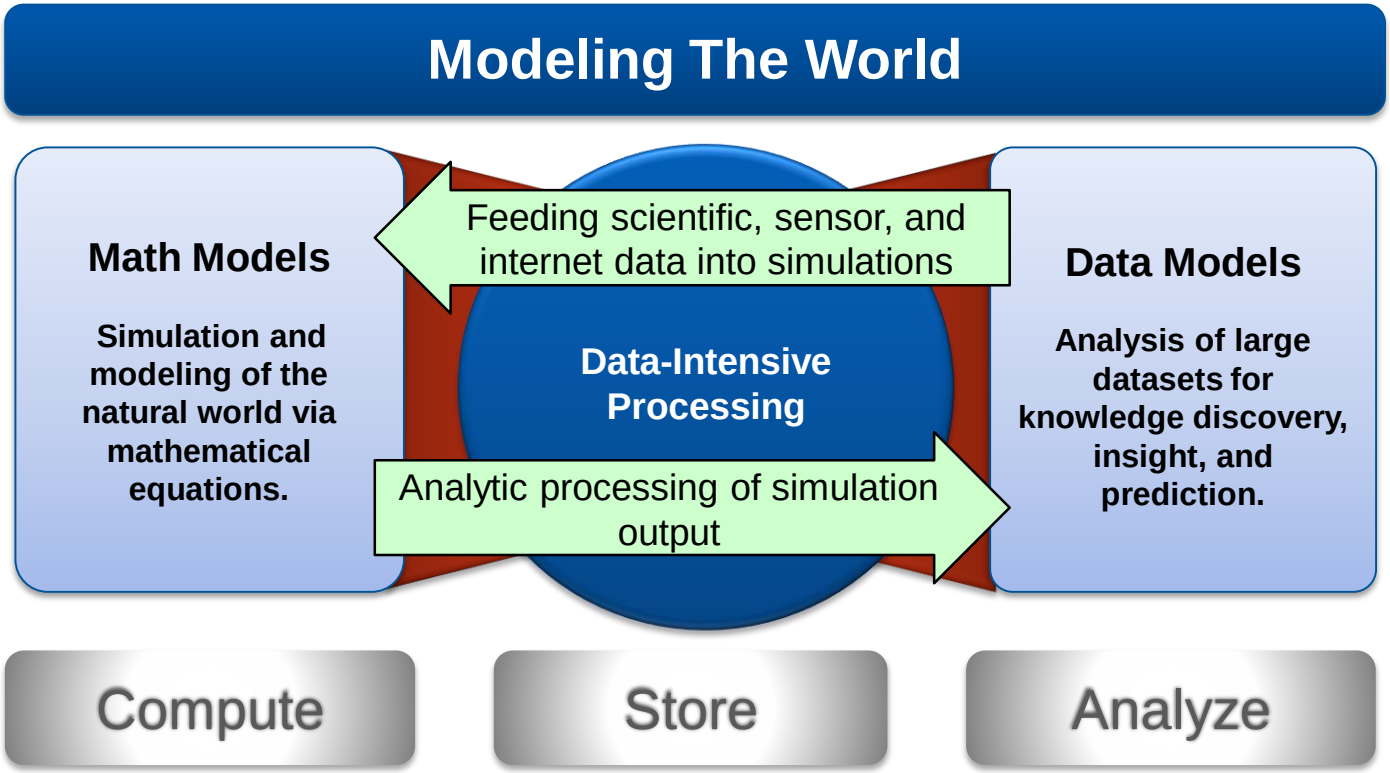


Computational studies in plasma physics and fusion energy
Office: Fusion Energy Sciences
Investigator: Abhay K. Ram
Science Area: Fusion Energy
Cores: 21,960 (Edison)
Core Hours Used: 330,521.4

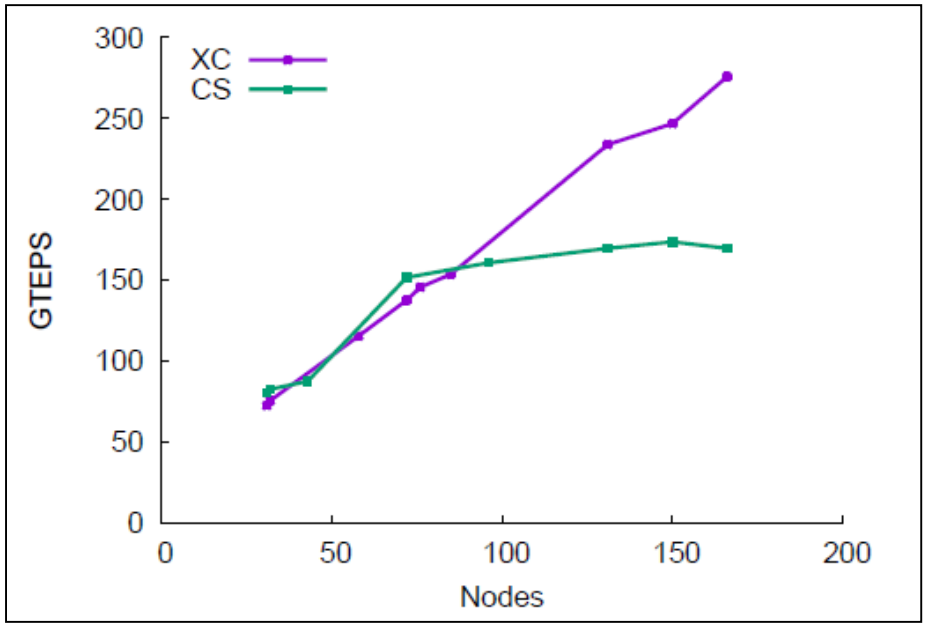


Quantum Chromodynamics with four flavors of dynamical quarks
Office: High Energy Physics
Investigator: Doug Toussaint
Science Area: Lattice QCD
Cores: 18,432 (Edison)
Core Hours Used: 90,039.5

Fusion of Supercomputing and Big (Fast) Data



Combining Analytics and Simulation on a Single Machine – ICM

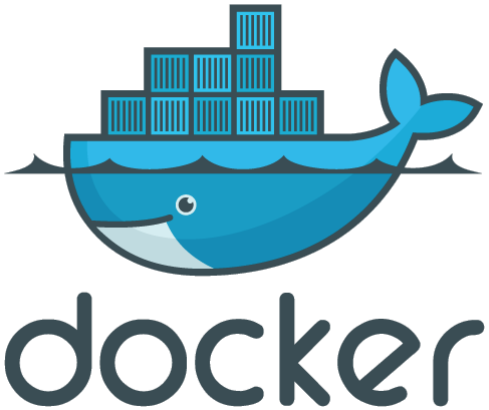


Graph500 Results – Aries (XC) compared with FDR (CS)

The Interdisciplinary Centre for Mathematical and Computational Modelling (ICM) will run simulation and analytics workloads on their new Cray XC40

HPC can learn from the Cloud

Improving portability and ease of use

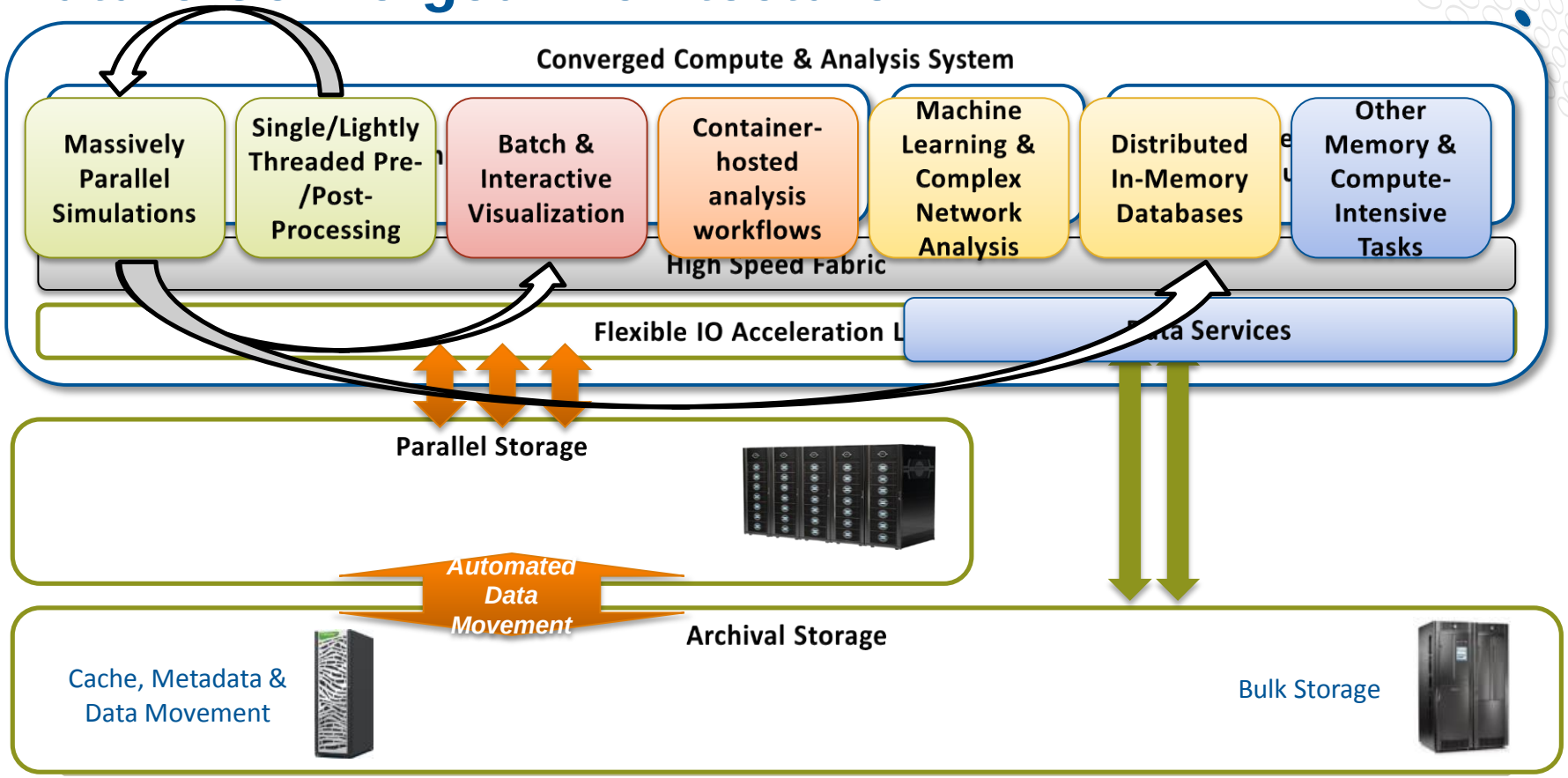


Build with Certified Software Stacks

Bundle libraries and dependencies

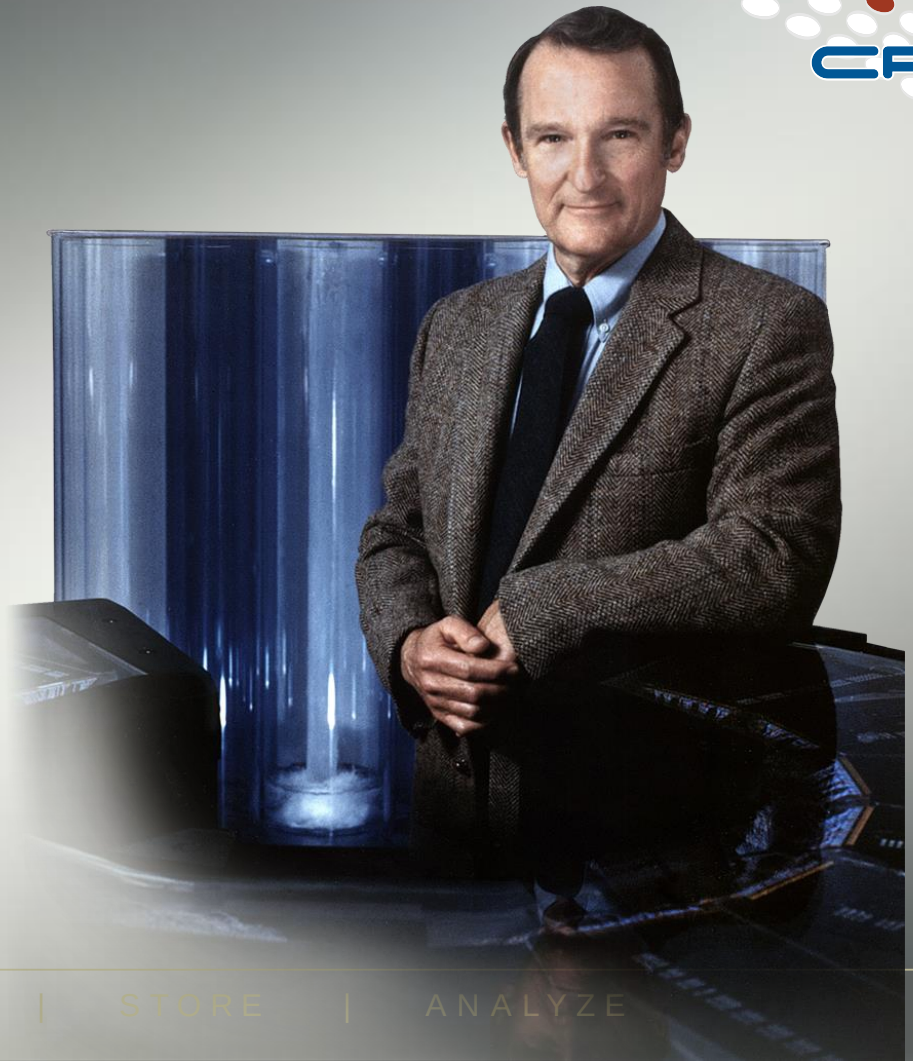
Build a consistent environment from desktop to supercomputer

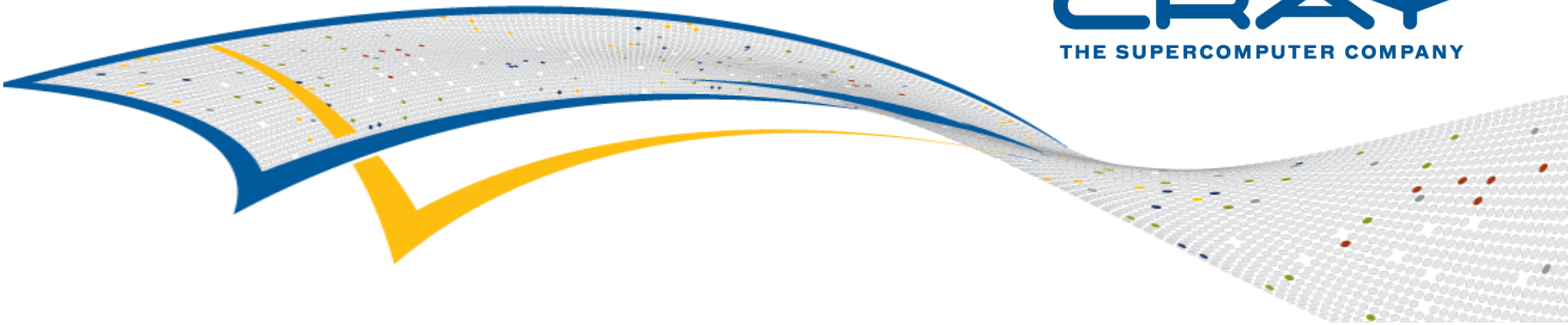
Future Converged Architecture



"The future is seldom the same as the past"

Seymour Cray
June 4, 1995





Thank You!

COMPUTE | STORE | ANALYZE