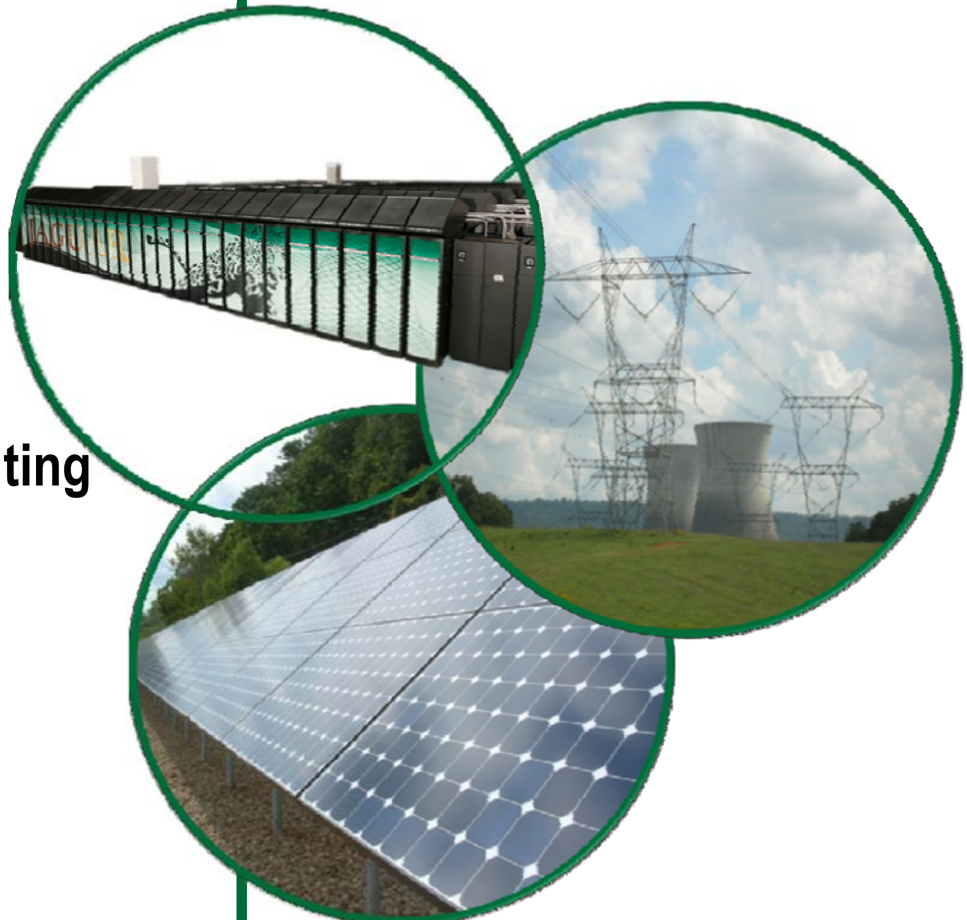
Jaguar: Powering and Cooling the Beast

Buddy Bland
2009 Conference on High-Speed Computing
The Salishan Lodge
Gleneden Beach, Oregon
April 30, 2009



U.S. DEPARTMENT OF ENERGY

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Outline

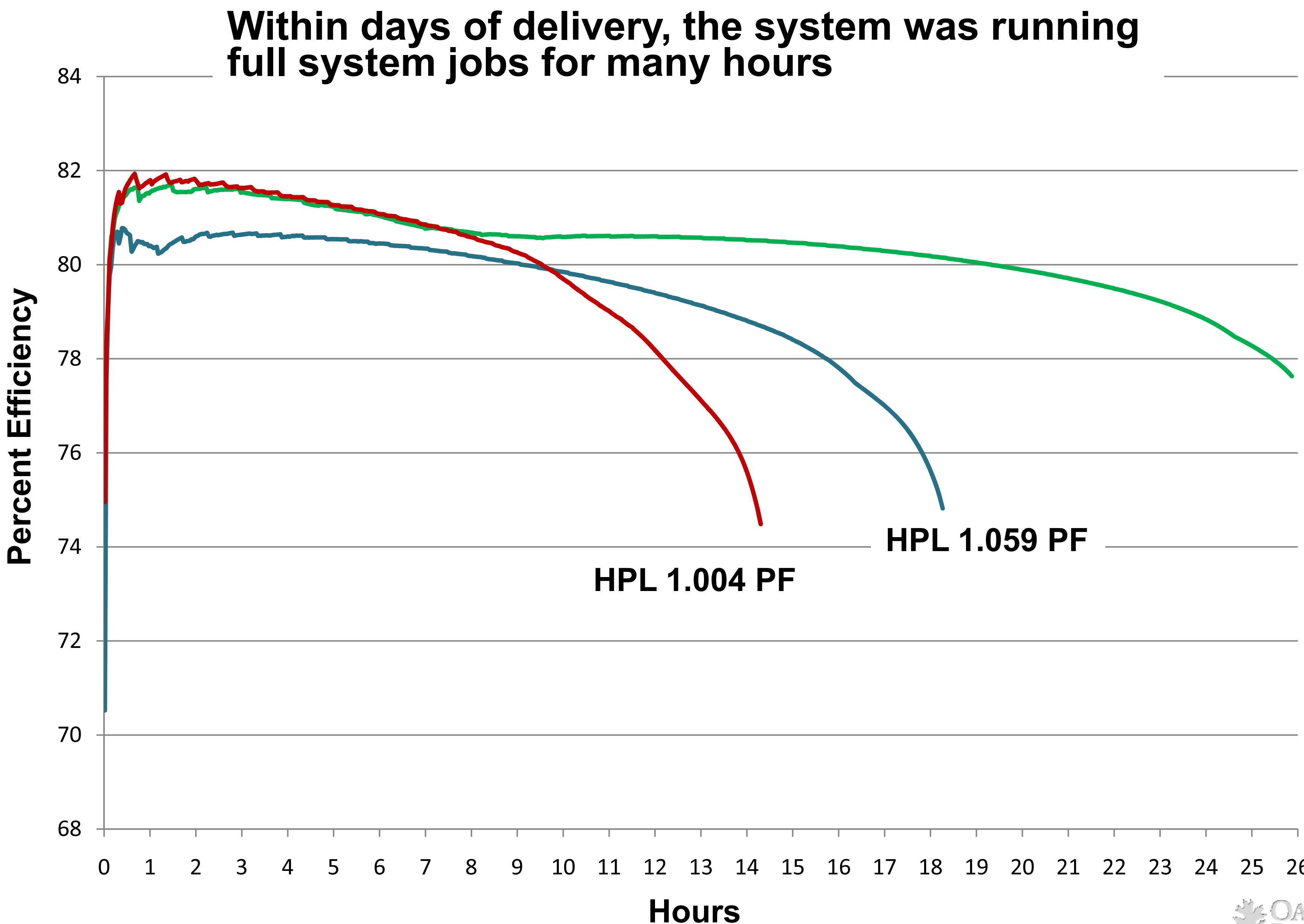
- Jaguar's features for performance and efficiency
- Historical overview of cooling systems on Cray's computers
- Implications for the future

Outstanding launch for petascale computing in Office of Science and ORNL at SC'08

Only 41 days after assembly of a totally new 150,000 core system

- **Jaguar beat the previous #1 performance on Top500 with an application running over 18 hours on the entire system**
- **Jaguar had two *real* applications running over 1 PF**
  - **DCA++ 1.35 PF Superconductivity problem**
  - **LSMS 1.05 PF Thermodynamics of magnetic nanoparticles problem**

Cray XT5 "Jaguar" is showing impressive stability

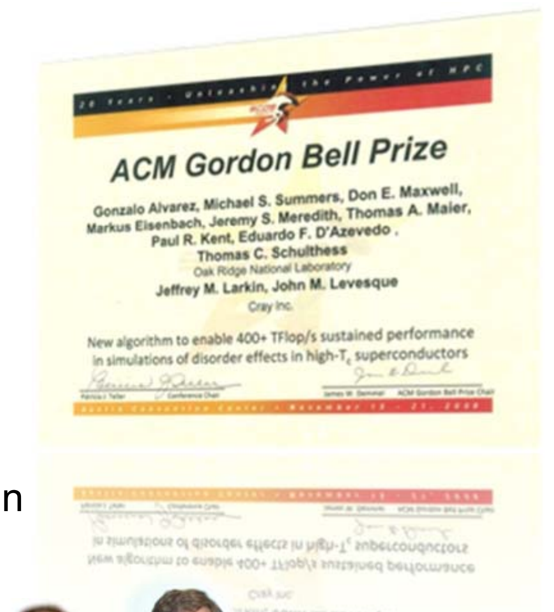


 Managed by UT-Battelle
for the U.S. Department of Energy

Gordon Bell prize awarded to ORNL team

Three of six GB finalist ran on Jaguar

- A team led by ORNL's Thomas Schulthess received the prestigious 2008 Association for Computing Machinery (ACM) Gordon Bell Prize at SC08
- For attaining fastest performance ever in a scientific supercomputing application
- Simulation of superconductors achieved 1.352 petaflops on ORNL's Cray XT Jaguar supercomputer
- By modifying the algorithms and software design of the DCA++ code, the team was able to boost its performance tenfold



Gordon Bell Finalists

✓	DCA++	ORNL
✓	LS3DF	LBNL
✓	SPECFEM3D	SDSC
•	RHEA	TACC
•	SPaSM	LANL
•	VPIC	LANL

HPC Challenge Awards

- HPC Challenge awards are given out annually at the Supercomputing conference
- Awards in four categories, result published for two others; tests many aspects of the computer's performance and balance
- Must submit results for all benchmarks to be considered
- Unfortunately, ORNL team only had two days on the machine to get the results. Got a better G-FFT number (5.804) the next day. ORNL submitted only baseline (unoptimized) results.

G-HPL (TF)		EP-Stream (GB/s)		G-FFT (TF)		G-Random Access (GUPS)		EP-DGEMM (TF)		PTRANS (GB/s)	
ORNL	902	ORNL	330	ANL	5.08	ANL	103	ORNL	1,257	SNL	4,994
LLNL	259	LLNL	160	SNL	2.87	LLNL	35.5	ANL	362	LLNL	4,666
ANL	191	ANL	130	ORNL	2.77↑	SNL	33.6	LLNL	162	LLNL	2,626

HPC CHALLENGE

http://icl.cs.utk.edu/hpcc/

Science Applications are Scaling on Jaguar

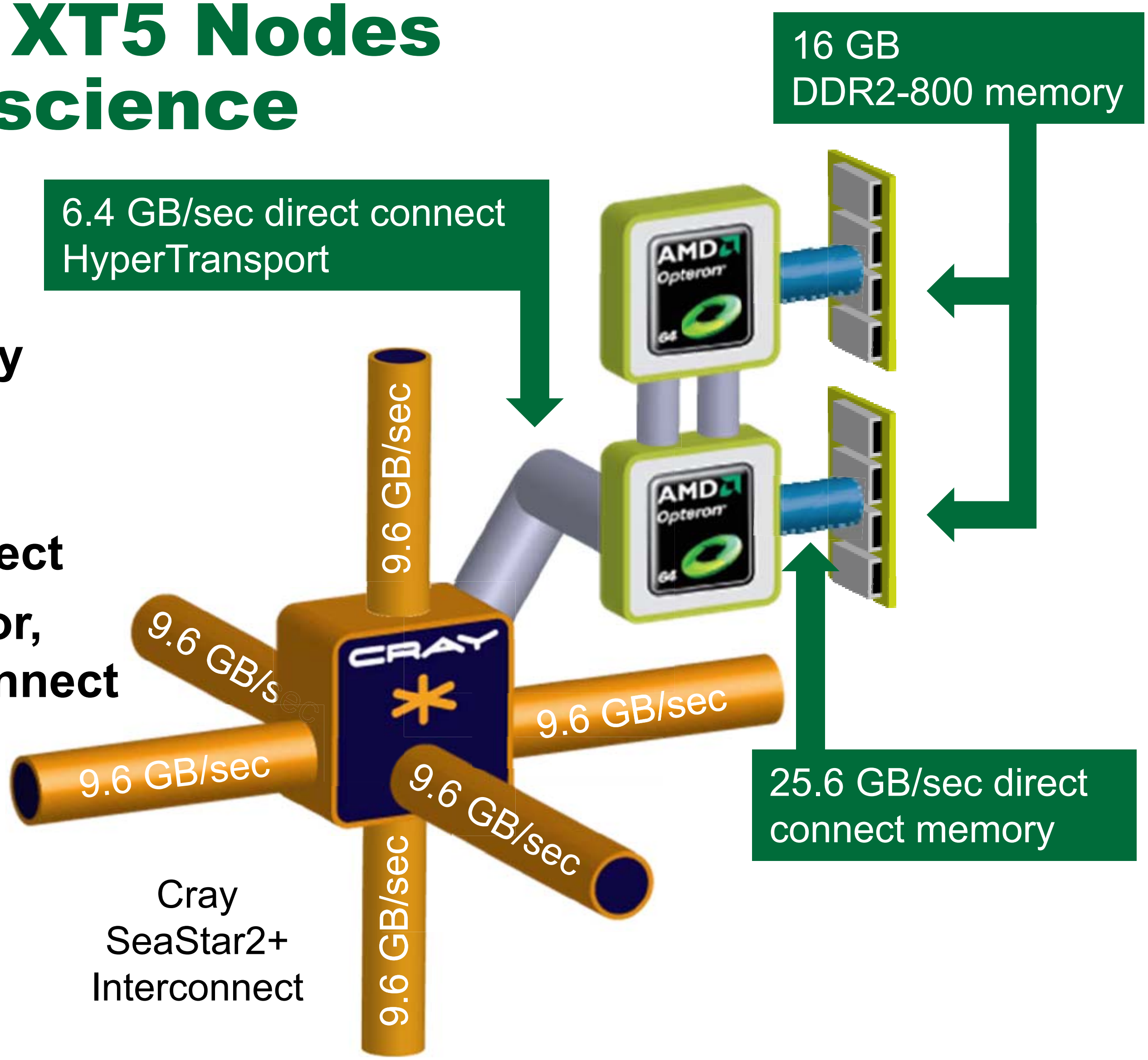
Science Area	Code	Contact	Cores	Total Performance	Notes
Materials	DCA++	Schulthess	150,144	1.3 PF*	Gordon Bell Winner
Materials	LSMS	Eisenbach	149,580	1.05 PF	
Seismology	SPECFEM3D	Carrington	149,784	165 TF	Gordon Bell Finalist
Weather	WRF	Michalakes	150,000	50 TF	
Climate	POP	Jones	18,000	20 sim yrs/ CPU day	
Combustion	S3D	Chen	144,000	83 TF	
Fusion	GTC	PPPL	102,000	20 billion Particles / sec	
Materials	LS3DF	Lin-Wang Wang	147,456	442 TF	Gordon Bell Winner
Chemistry	NWChem	Apra	96,000	480 TF	
Chemistry	MADNESS	Harrison	140,000	550+ TF	

Jaguar: World's most powerful computer Designed for science from the ground up

Peak performance	1.645 petaflops
System memory	362 terabytes
Disk space	10.7 petabytes
Disk bandwidth	200+ gigabytes/second

Jaguar's Cray XT5 Nodes Designed for science

- **Powerful node improves scalability**
- **Large shared memory**
- **OpenMP Support**
- **Low latency, High bandwidth interconnect**
- **Upgradable processor, memory, and interconnect**



GFLOPS	**76.3**
Memory (GB)	16
Cores	8
SeaStar2+	1

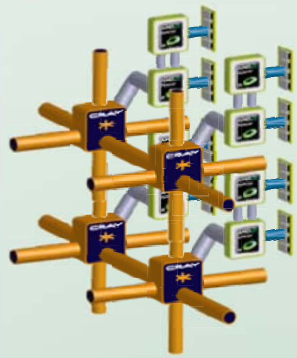
Building the Cray XT5 System

Node
73.6 GF
16 GB

1x1x1

4x

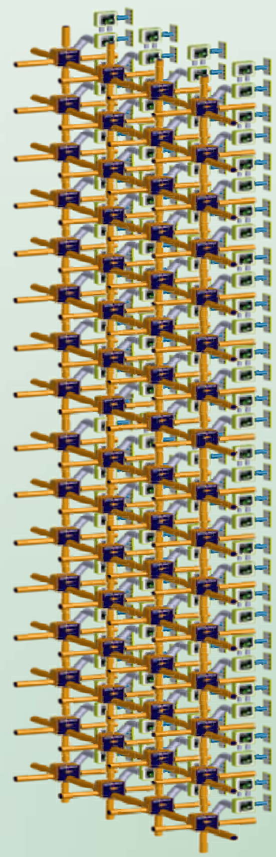
Blade
294 GF
64 GB



1x2x2

24x

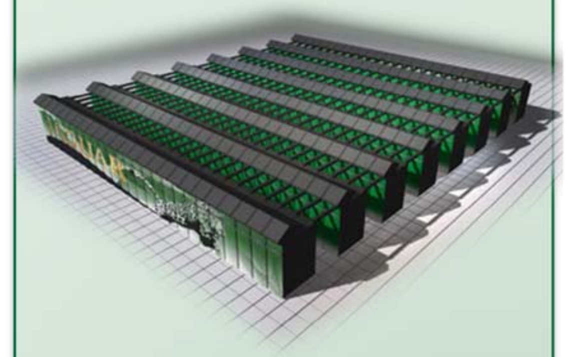
Rack
7.06 TF
1.54 TB



1x4x16

200x

System
1382 TF
300 TB

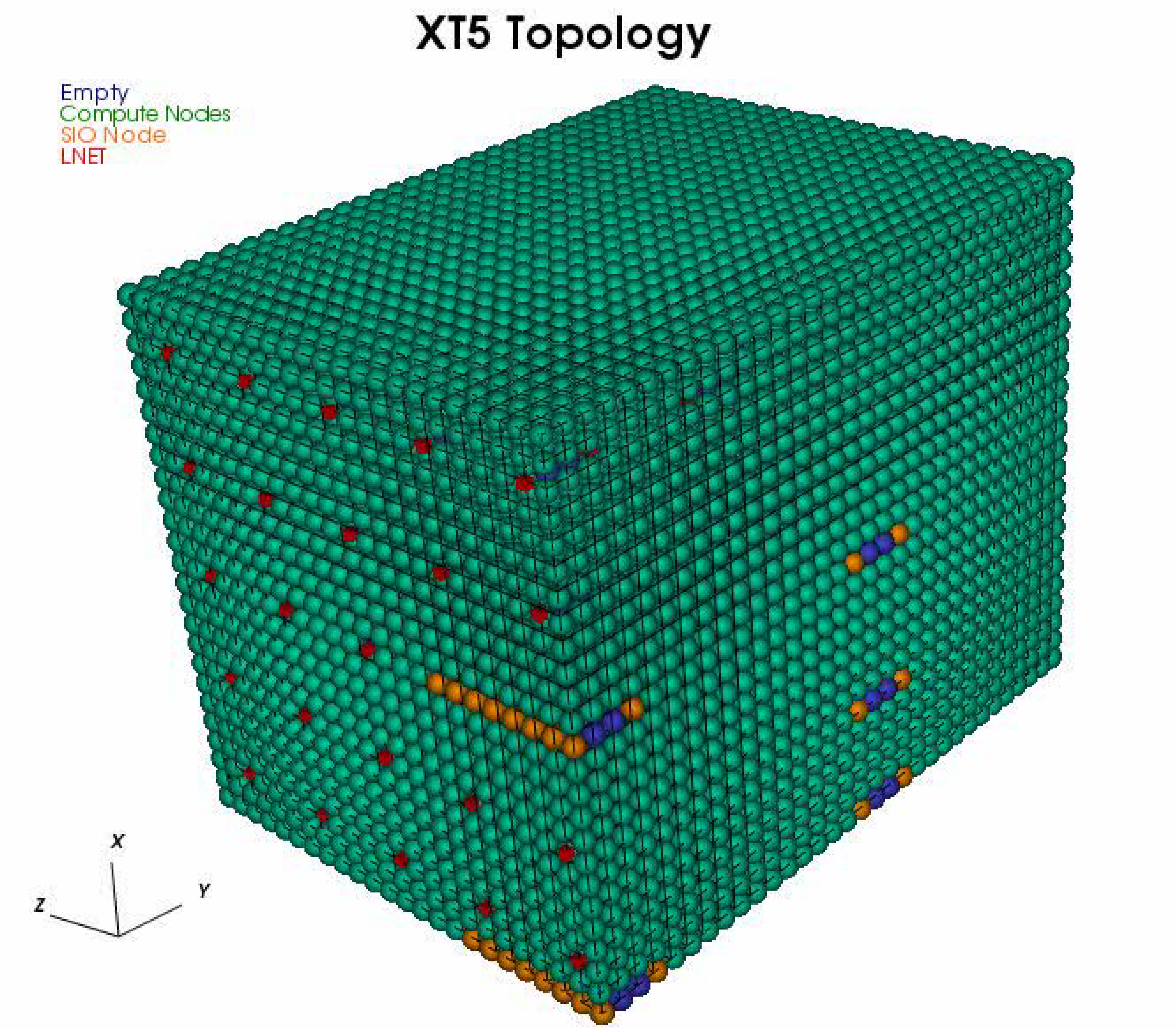


25x32x16

XT5 I/O Configuration
Driven by application needs

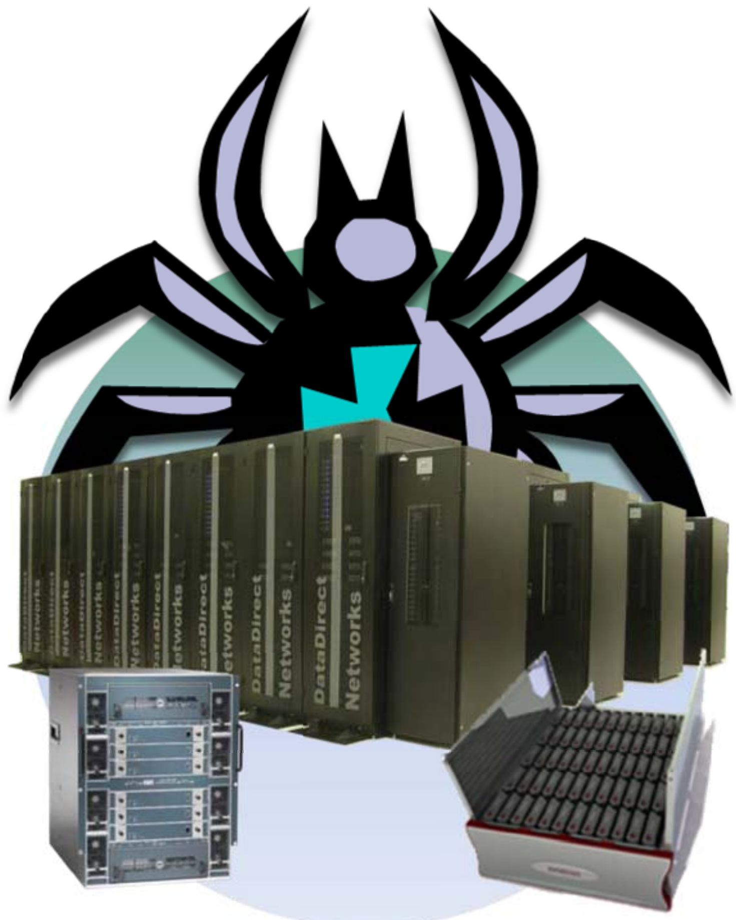
Features of I/O nodes

- 192 I/O nodes
- Each connected via non-blocking 4x DDR Infiniband to Lustre Object Storage Servers
- Fabric connections provides redundant paths
- Each OSS provide 1.25 GB/s
- I/O nodes spread throughout the 3-D torus to prevent hot-spots



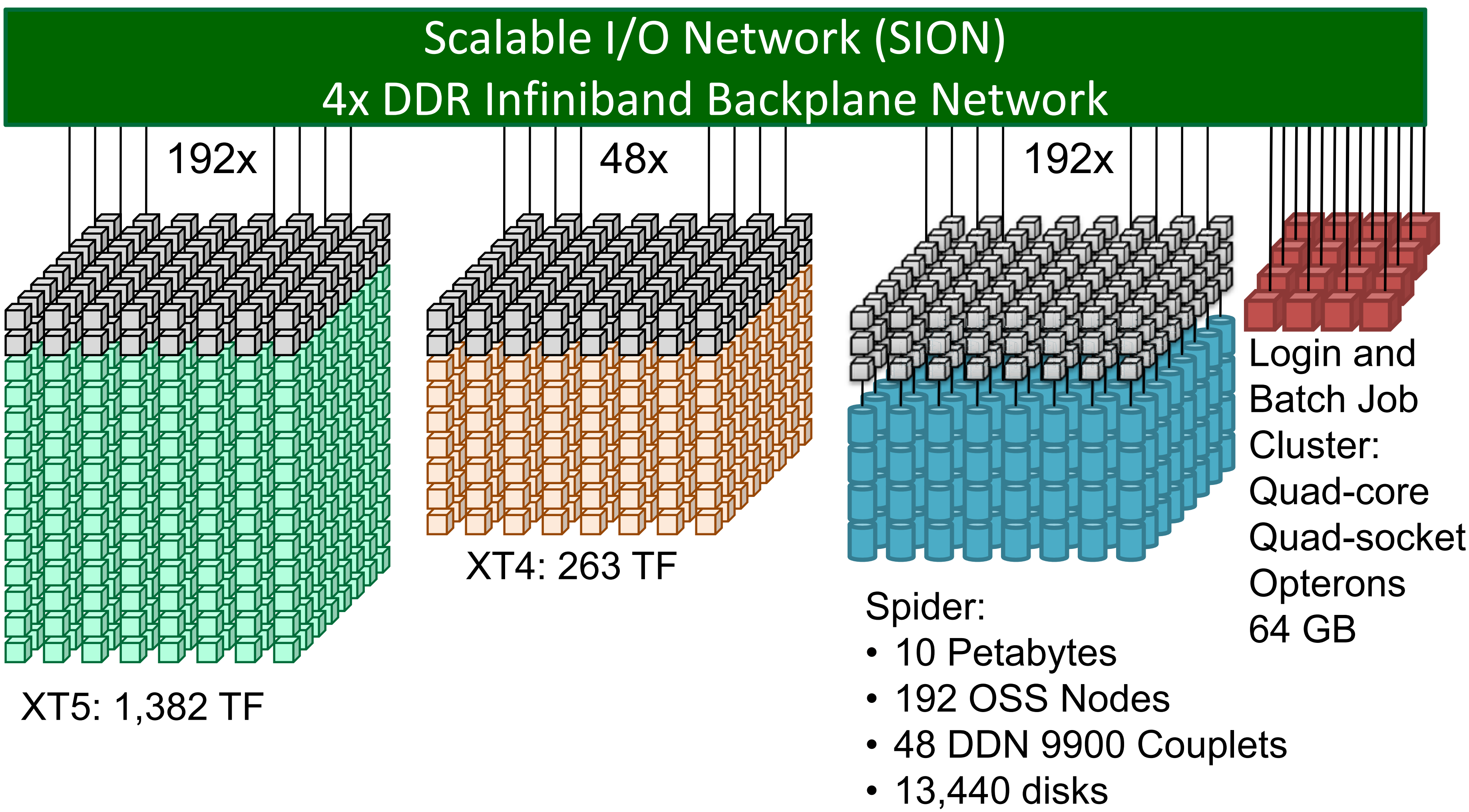
Movie of I/O node layout

Center-wide File System

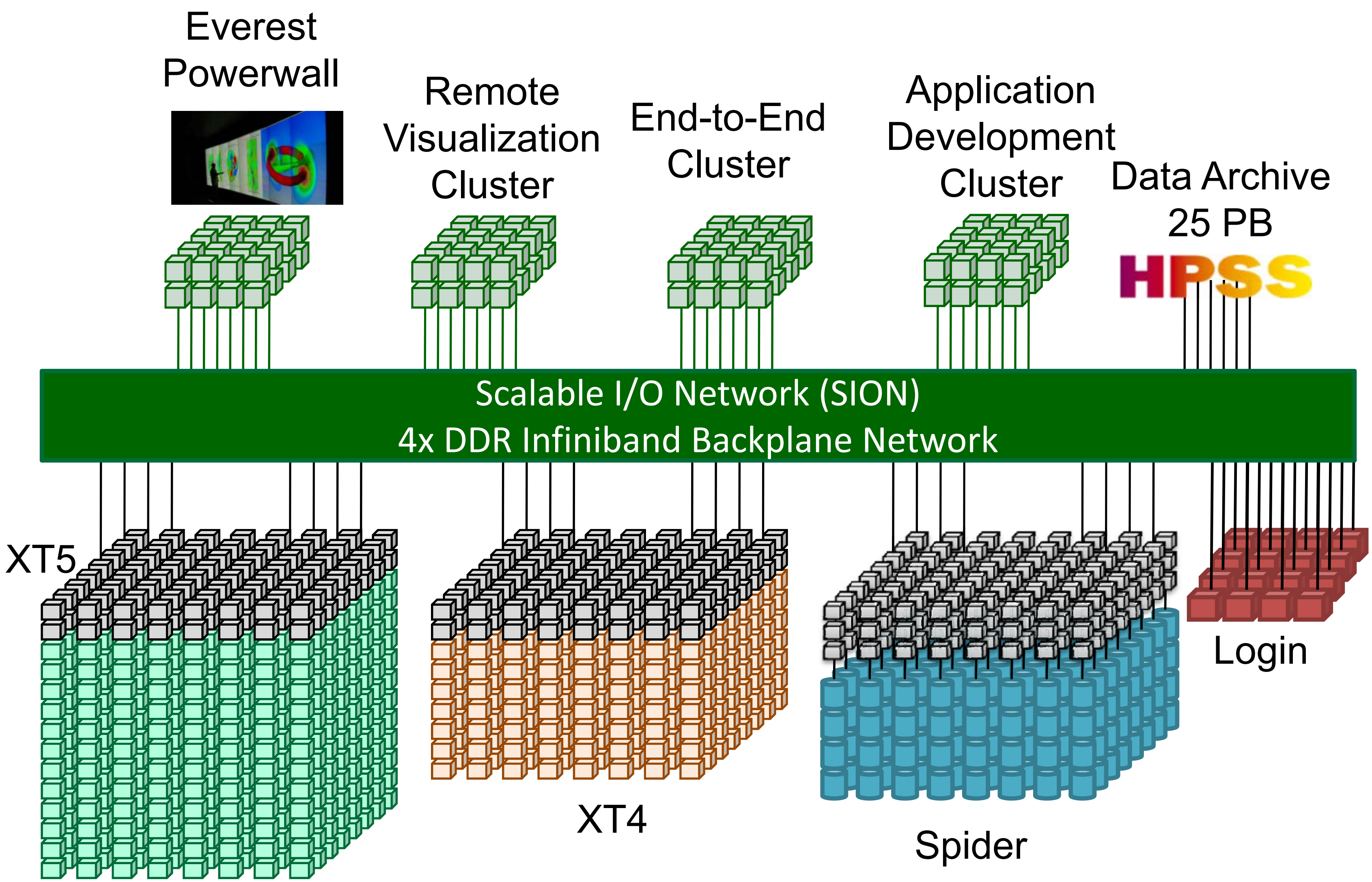


- "Spider" provides a shared, parallel file system for all systems
  - Based on Lustre file system
- Demonstrated bandwidth of over 200 GB/s
- Over 10 PB of RAID-6 Capacity
  - 13,440 1-TB SATA Drives
- 192 Storage servers
  - 3 TB of memory
- Available from all systems via our high-performance scalable I/O network
  - Over 3,000 InfiniBand ports
  - Over 3 miles of cables
  - Scales as storage grows
- Undergoing friendly user checkout with deployment expected in summer 2009

Combine the XT5, XT4, and Spider with a Login Cluster to complete Jaguar

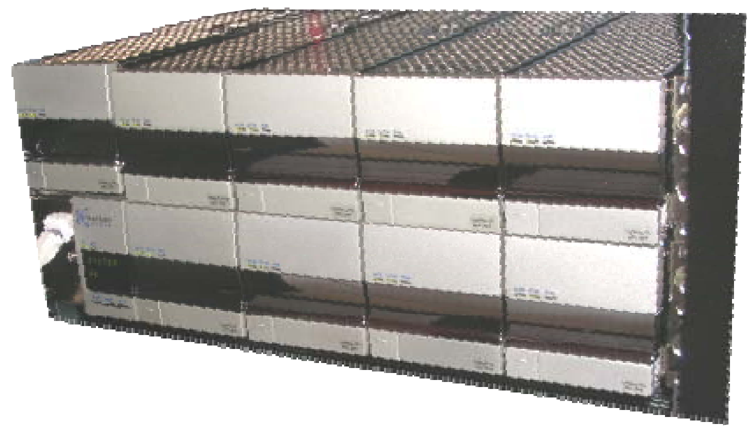


Completing the Simulation Environment to meet the science requirements

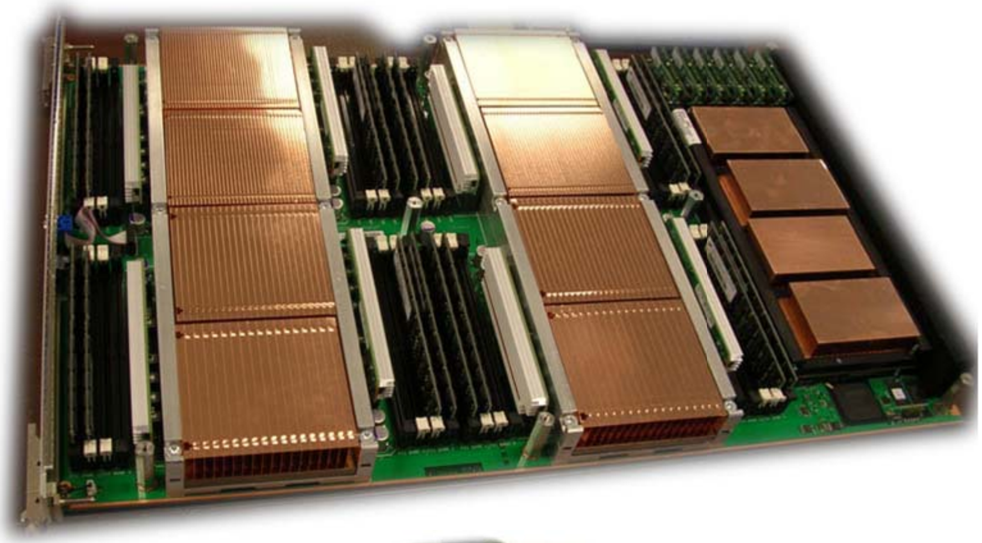


XT5 Innovations: 480 volt power to the cabinet

- Saved about $1M in site prep costs in copper and circuit breakers
- Saves in ongoing electrical power costs by reducing losses in transformers and wires
- Allows higher density cabinets which shrinks system size



High-density blades



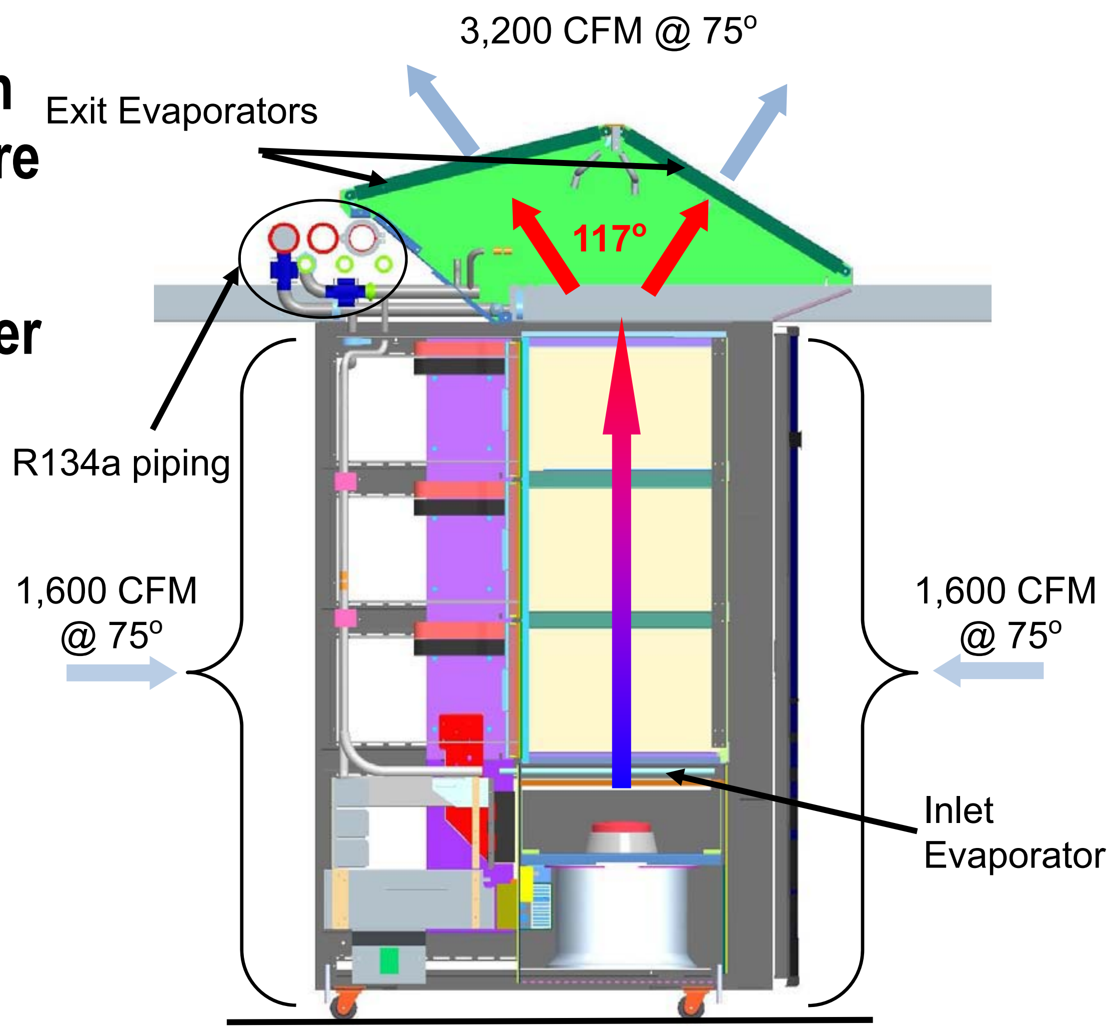
- Eight Opteron Sockets
- 32 DIMM slots
- 4 SeaStar2+ interconnect chips
- Variable pitch heat sinks

Single high-reliability fan

- Lower power than separate muffin-fans on each blade
- Higher reliability
- Custom designed turbine for high air-flow
- Variable speed

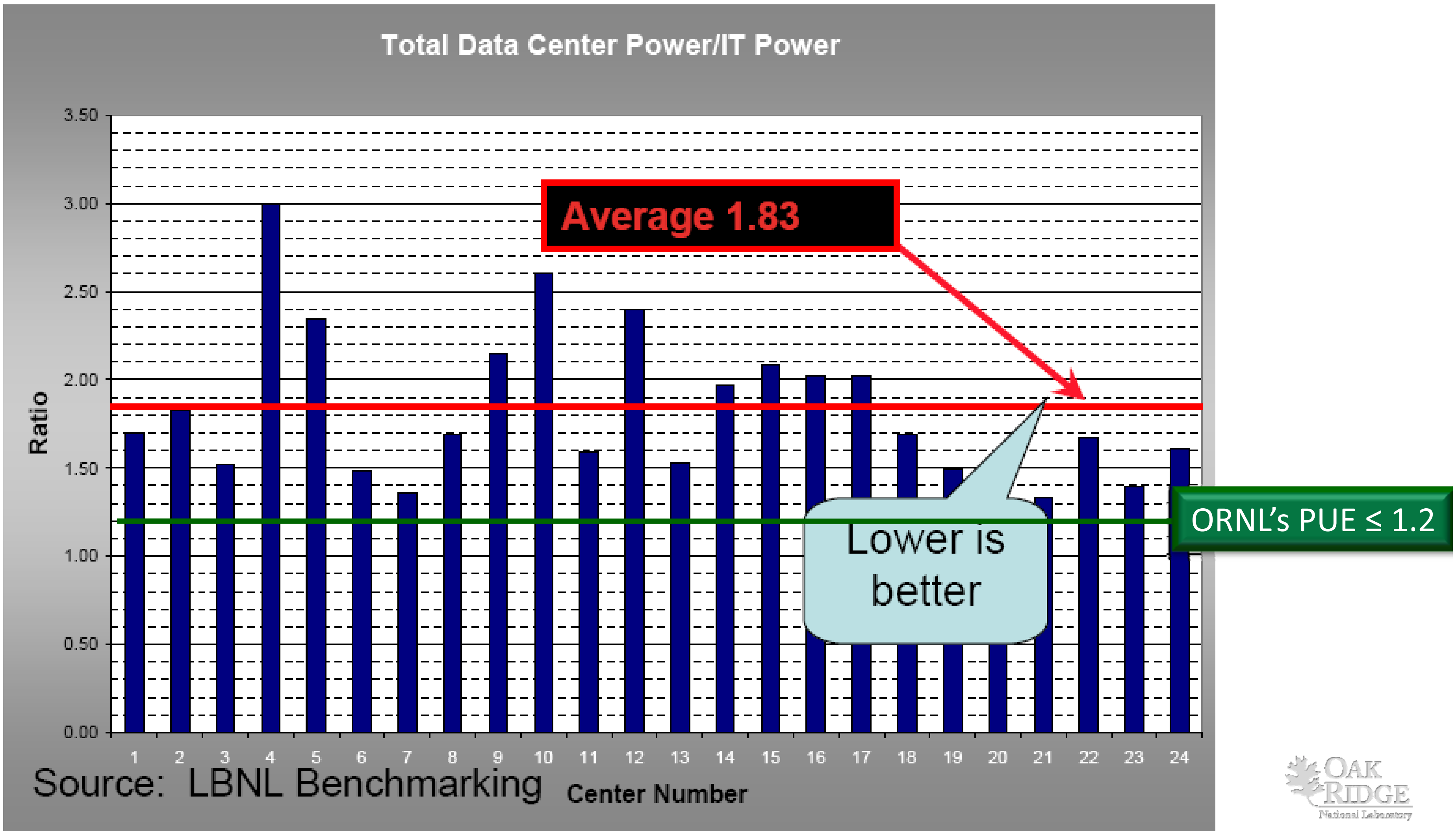
High Efficiency Liquid Cooling *Required* to build such a large system

- Newer Liquid Cooled design removes heat to liquid before it leaves the cabinet
- Saves about 900KW of power just in air movement and 2,500 ft^2 of floor space
- Phase change liquid to gas removes heat much more efficiently than water or air
- Each XDP heat exchanger replaces 2.5 CRAC units using one-tenth the power and floor space



Today, ORNL's facility is among the most efficient data centers

Power Utilization Efficiency (PUE) = Data Center power / IT equipment



Electrical Systems Designed for efficiency

13,800 volt power into the building saves on transmission losses

480 volt power to cabinets saves $1M in installation costs

High efficiency power supplies in the cabinets

Flywheel based UPS for highest efficiency

A bit of history about cooling and packaging

Power numbers in KW for a single CPU cabinet, not including SSD, IOS, HEU, or disks

Cray-1

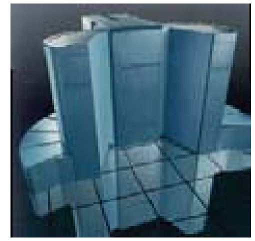
First Vector Supercomputer & first to utilize Freon cooling (150)

Cray X-MP

First vector multi-processor Supercomputer (160)

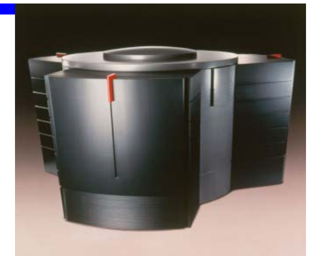
Cray-2

First Fluorinert Immersion cooled (200)



Cray Y-MP

First Supercomputer to sustain 1 GF, Fluorinert cold plates (145)



Cray C90

First Supercomputer with 1GF processor, Fluorinert cold plates (190)

Cray T90

First wireless supercomputer, Fluorinert immersion (345)

Cray T3E

First Supercomputer to sustain 1 TF, Fluorinert cold plate (45)

Cray X1/X1e

First Scalable Vector Supercomputer and first to utilize evaporative spray cooling (70)

Cray XT3/4

Highly scalable supercomputer, air cooled (20)

Cray XMT

First massively multithreaded supercomputer with extended memory semantics (25)

Cray XT5h

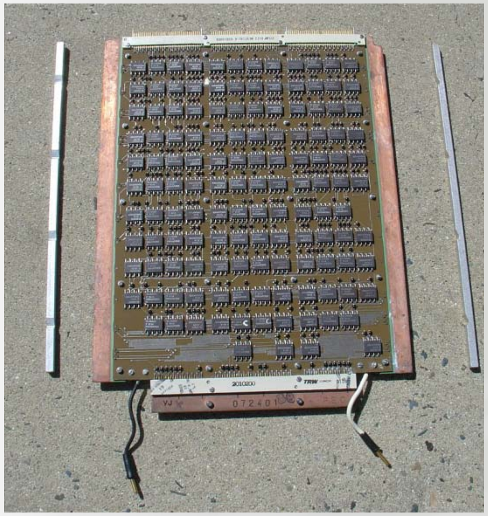
First Hybrid Supercomputer featuring scalable MPP , LC and Vector that utilized closed loop LC (45)

Cray XT5

First scalable system using R-134a cooling in top and bottom of the cabinet (40)

#1 Freon and Copper Cold Plates -1976

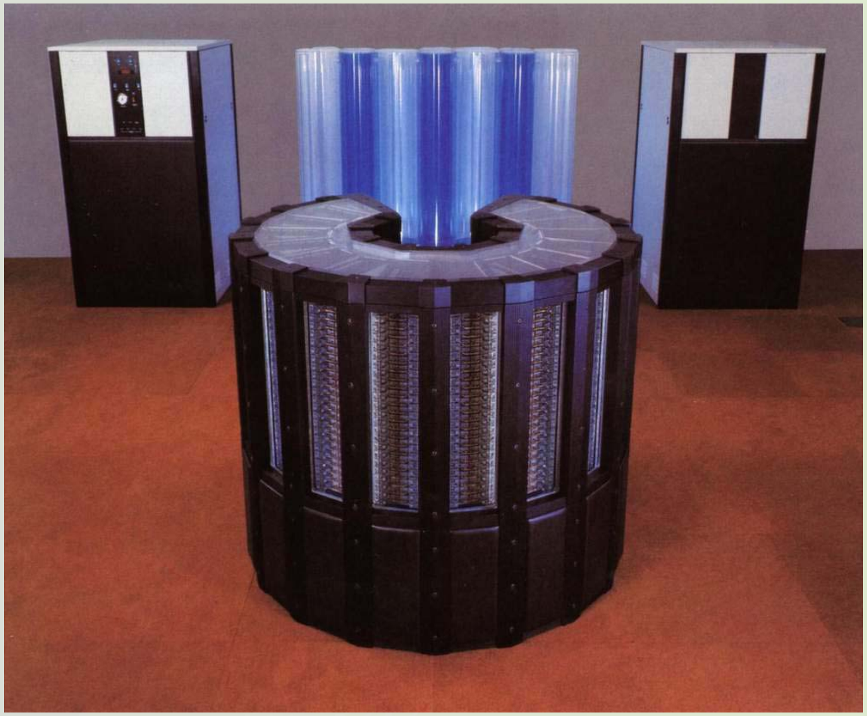
- Freon was used in conjunction with heat conducting plates
- Cray-1 and Cray XMP and I/O subsystems



Provided courtesy Cray Inc.

#2 Fluorinert Immersion -1986

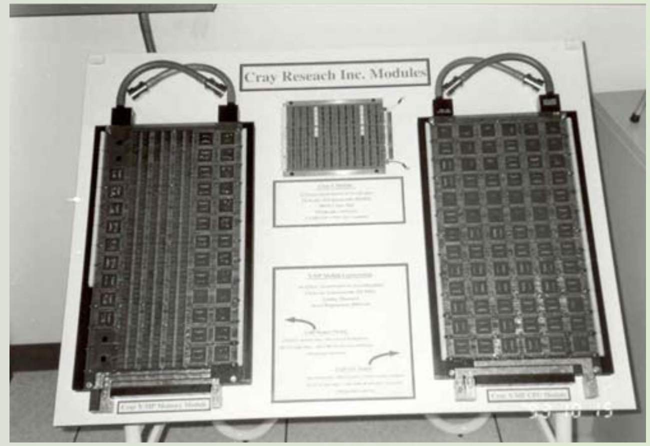
- Initially used on the Cray-2 system
- Later used on the Cray T90 system and the Cray-3
- Entire computer is immersed in liquid
- Allowed tightly packed, 3 dimensional modules

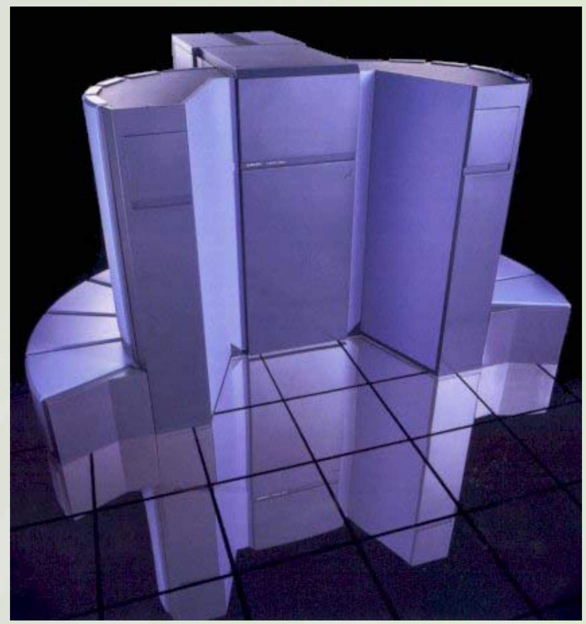


Provided courtesy Cray Inc.

#3 Captive Fluorinert Cold Plates

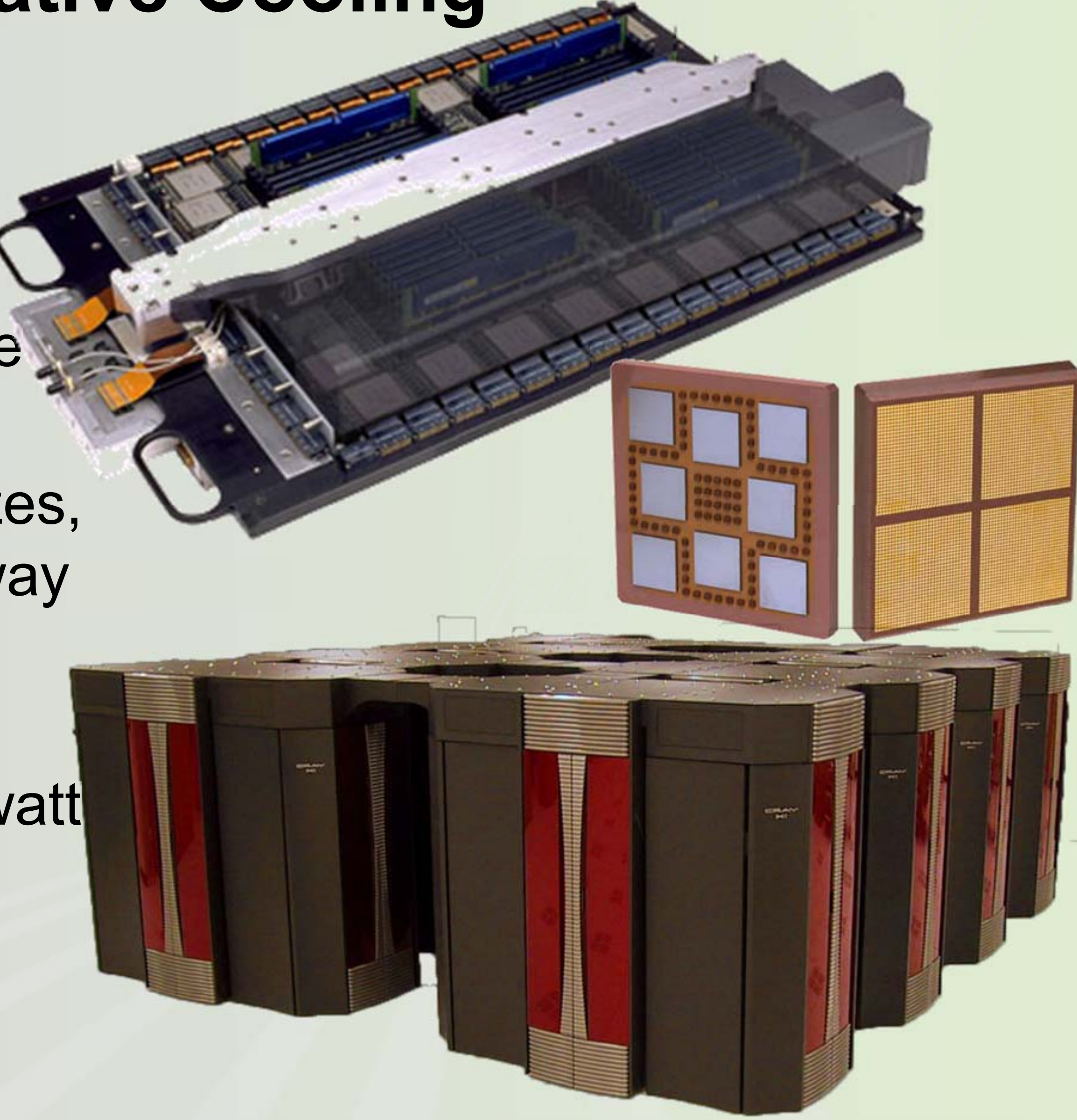
- Used on the Cray Y-MP, Cray C90, Cray T3D and Cray T3E Systems
- Fluorinert circulated through a hollow cold-plate
- Fluorinert was used to minimize the chances of damage to components when the snap fittings were disconnected for servicing modules





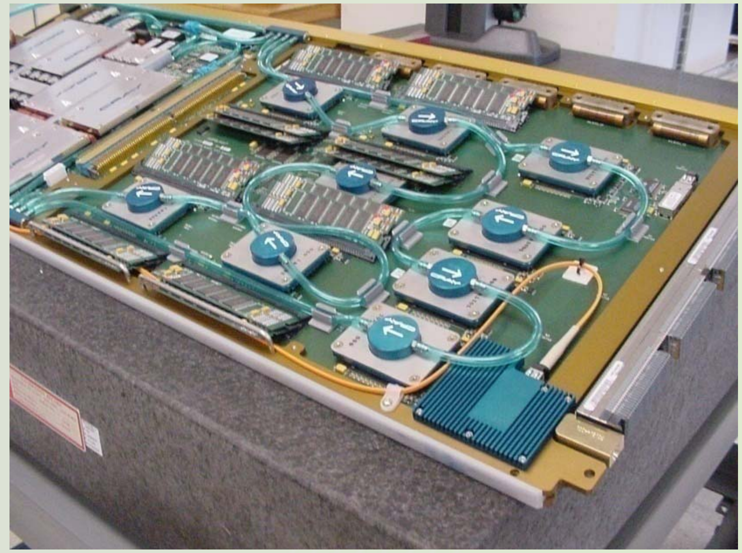
#4 Spray Evaporative Cooling

- Used on the Cray X1 processors
- A mist of Fluorinert is sprayed directly on the die
- The Fluorinert vaporizes, and heat is carried away via the latent heat of vaporization
- Used to cool a ~400 watt MCM



#5 Water Cap Cooling

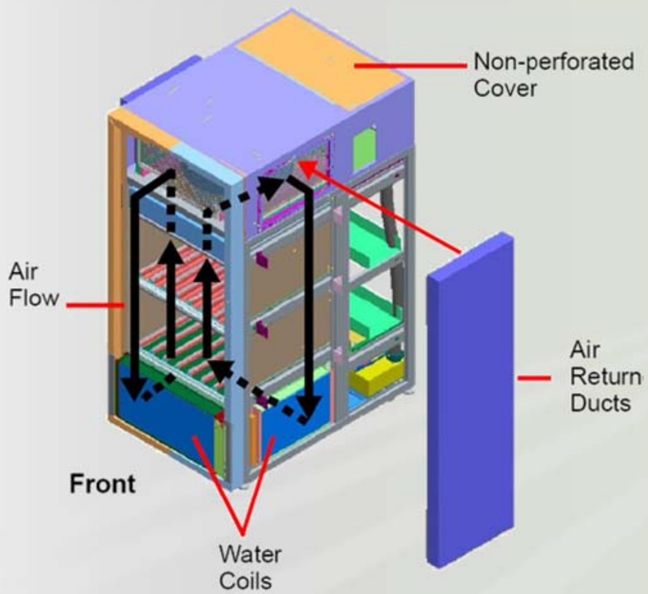
- A water-filled heat-sink is mounted directly on an ASIC
- Used on the Cray MTA-2
- Designed to cool the custom ASICs in the machine
- Originally ran with water
- Later changed to Fluorinert because of organic growth in the fluid (and electrical problems induced by water flowing over dissimilar metals)



Provided courtesy Cray Inc.

#6 Water Cooled Radiator

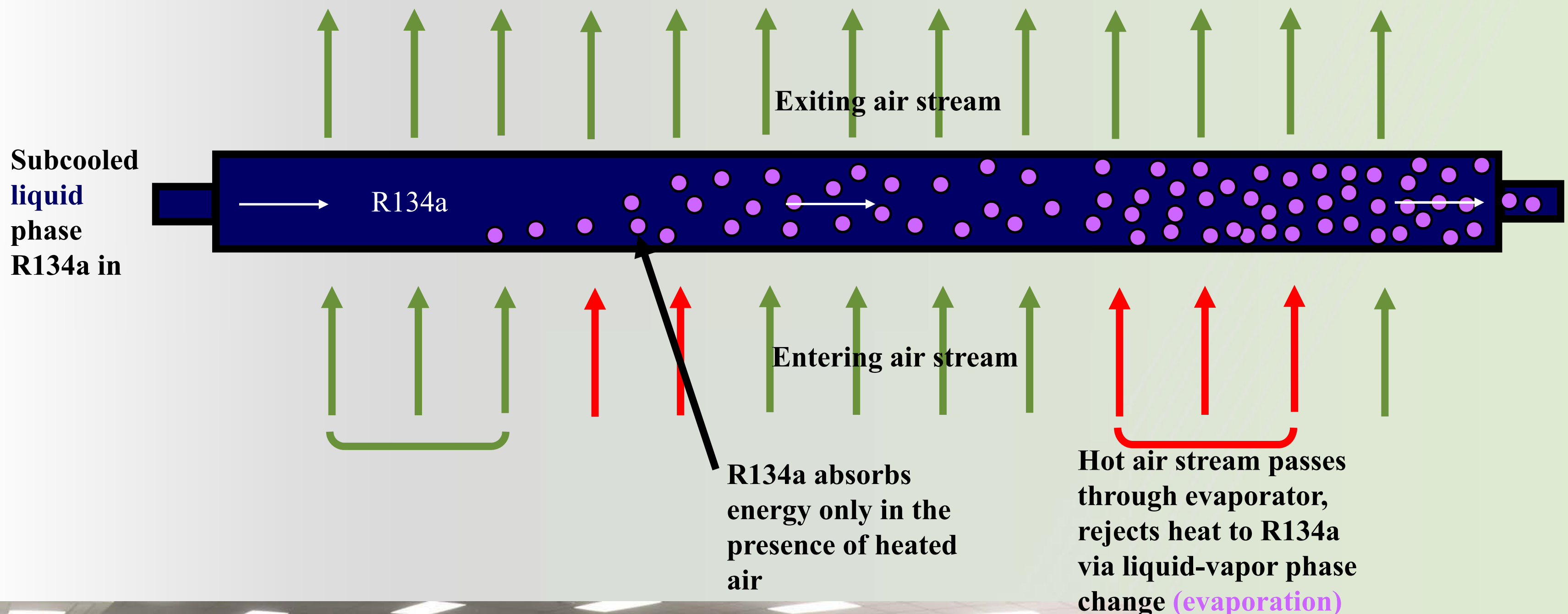
- Option on the Cray X2 vector processor cabinets
- Removes approximately 80% of the heat through chilled water
- Air is internally recirculated



Provided courtesy Cray Inc.

Cooling Method #7
R134A Phase Change Evaporative Cooling

- Available on Cray XT5



Over 10x more effective than a water coil of similar size (phase change much more effective method to remove heat)

Provided courtesy Cray Inc.

Trends in Data Centers

- Conventional data centers
  - Raised floor
  - Air or liquid cooled systems
  - Temperature typically controlled to 68°F

Does it really need to be that cold?

- Intel did a side by side test comparing traditional cooling to untreated outside air up to 90°F
  - 10 month test, 500 KW total load (both sides including cooling)
  - In Albuquerque, saved approx 67% of the annual cooling cost
  - Little difference between reliability of servers

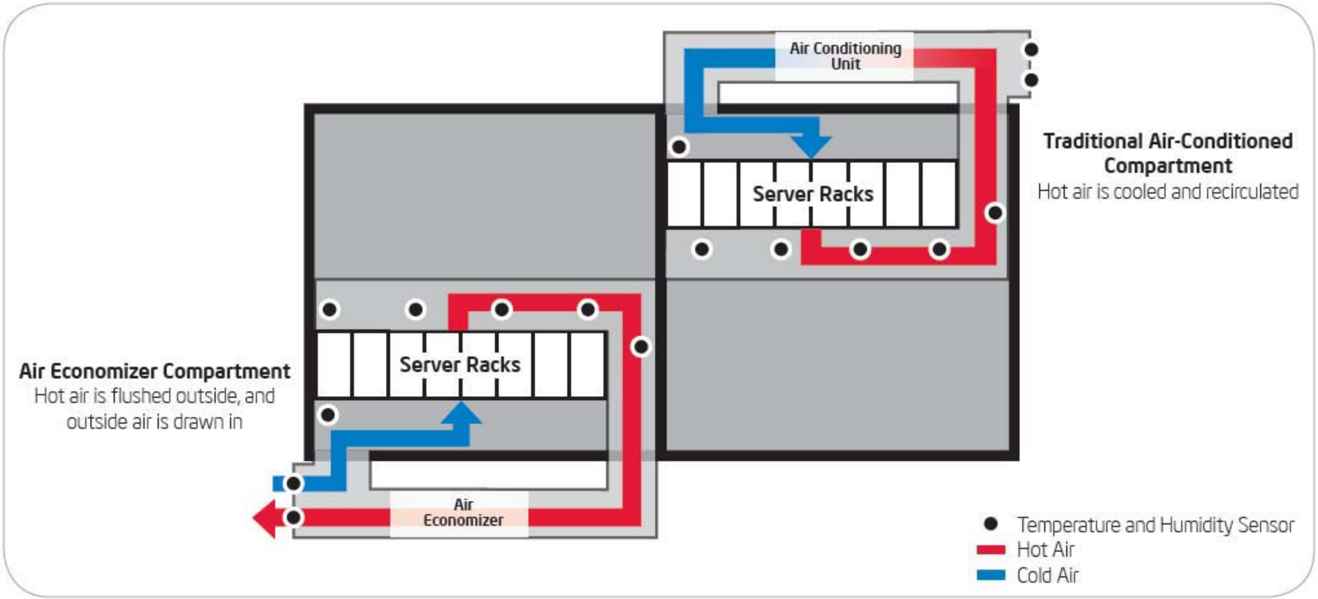


Figure 1. Proof of concept (PoC) data center environment.

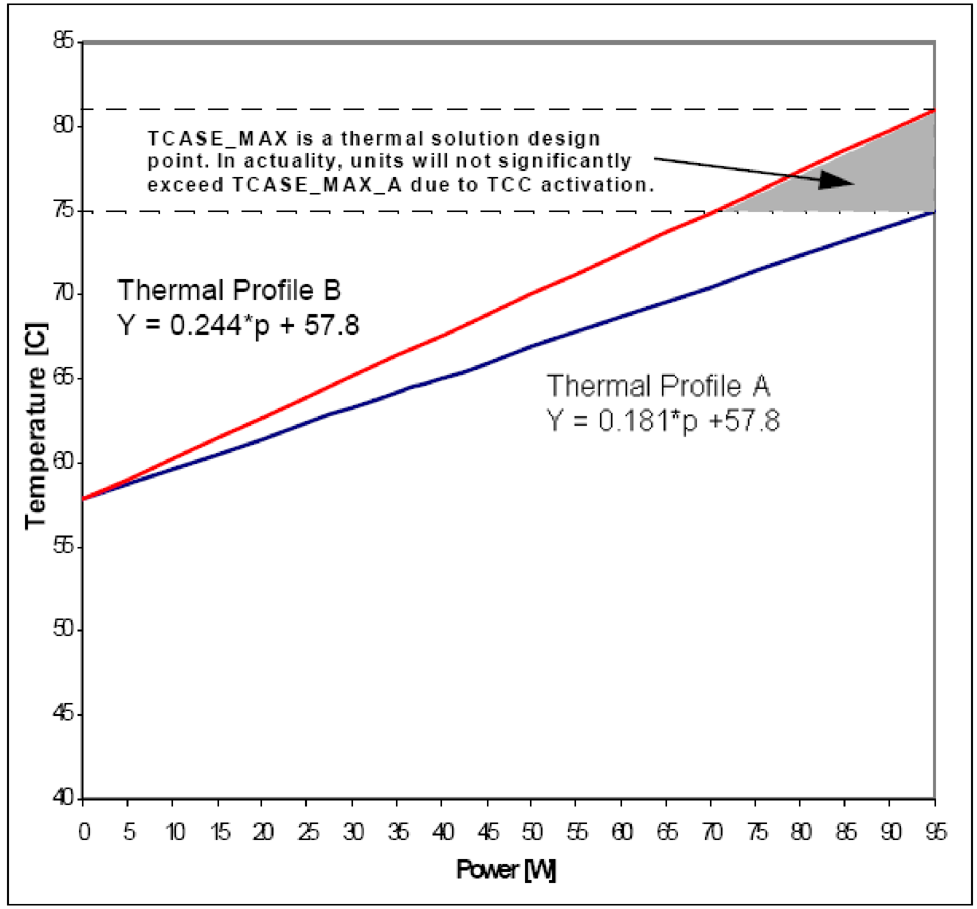
Provided courtesy Intel

IT@Intel

How hot is too hot?

- Intel Xeon 5500 (Nehalem) is specified at up to 75°C (167°F)
- Micron DDR3 memory is specified for 0 to 95°C
- DDN S2A9900 specs operations at 5 to 35°C (95°F)

Intel Xeon Processor 5500 Series Advanced SKU Thermal Profile



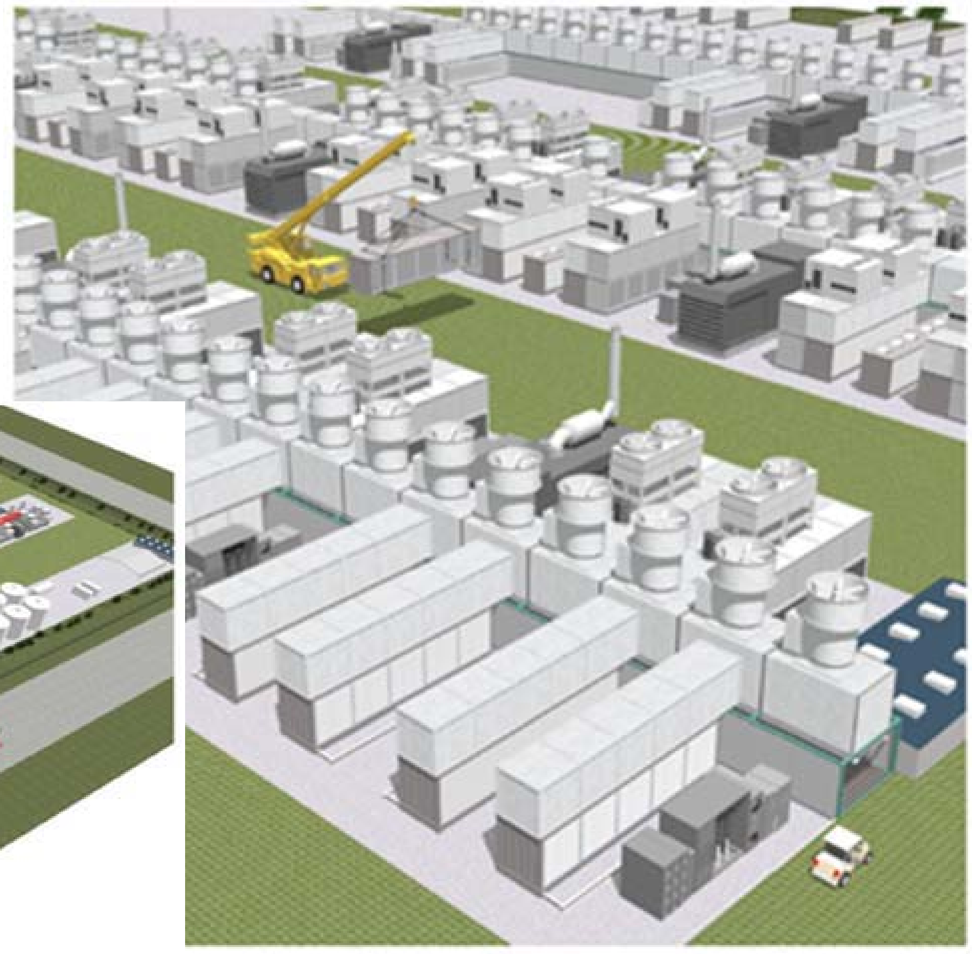
Intel® Xeon® Processor 5500 Series Datasheet, Volume 1
http://download.intel.com/design/xeon/datashts/321321.pdf (page 93)

What might a new data center look like?

- IBM, HP, Sun, Rackable, Verari, and others are building containerized data centers
- Custom configured with servers and disks
- Connect to power, chilled water, and networks and run
- Using standard 8’x8’x20’ or 8x8x40’ shipping containers
- Most are designed to run at up to 90°F with a PUE ≤ 1.3
- One container may contain up to 3,500 nodes or 12,000 disk drives
- Configured to order, 6-12 week delivery

But will it attract tornados?

Is a parking lot your next data center?



Illustrations from: ***Our Vision for Generation 4 Modular Data Centers - One way of Getting it just right . . .*****; Microsoft, 2008**

Questions?

"We finally have a true leadership computer that enables us to run calculations impossible anywhere else in the world. The huge memory and raw compute power of Jaguar combine to transform the scale of computational chemistry.

Now that we have NWChem and MADNESS running robustly at the petascale, we are unleashing a flood of chemistry calculations that will produce insights into energy storage, catalysis, and functionalized nano-scale systems."

Robert Harrison
ORNL and University of Tennessee