So...You Want to Build a Cray

Jeff Nichols, Oak Ridge National Lab

Cray XT3/XT4 Cabinet at KnoxMakers

Titan - Cray XT7 (successor to Jaguar)

The Numbers

- 18,688 16-Core AMD Opteron Processors = 299,008 Normal CPU Cores
- 18,688 Kepler K20X Computational GPUs with 2688 compute cores each
- 200 Cabinets
- Theoretical Max of 27 Quadrillion Floating Point Operations per Second (FLOPs). One PetaFLOP=1,000,000,000,000,000 FLOPs. Demonstrated 17 PetaFLOPs.

Wait! There's More

ORNL also is the host to two other world-class Crays:

~~Kraken (owned by University of Tennessee)~~

Partially replaced by a Cray XC30 (Intel Xeon Phi)

Gaea (owned by NOAA)

Server Room Tour


ECOphlex


35
KRAKEN


ckt 35


MSB-IT
ckt 33
MSB-IT
ckt 34

Fastest? Says Who?

- Top500.org lists published twice a year
- Rankings are by FLOPs achieved on matrix calculations (linear algebra)
- Jack Dongarra is an author of LINPACK and a Distinguished Scientist at ORNL and UT
- Chinese Military system at 34 PetaFLOPs is current No. 1. Titan is second at 17.59 PetaFLOPs.

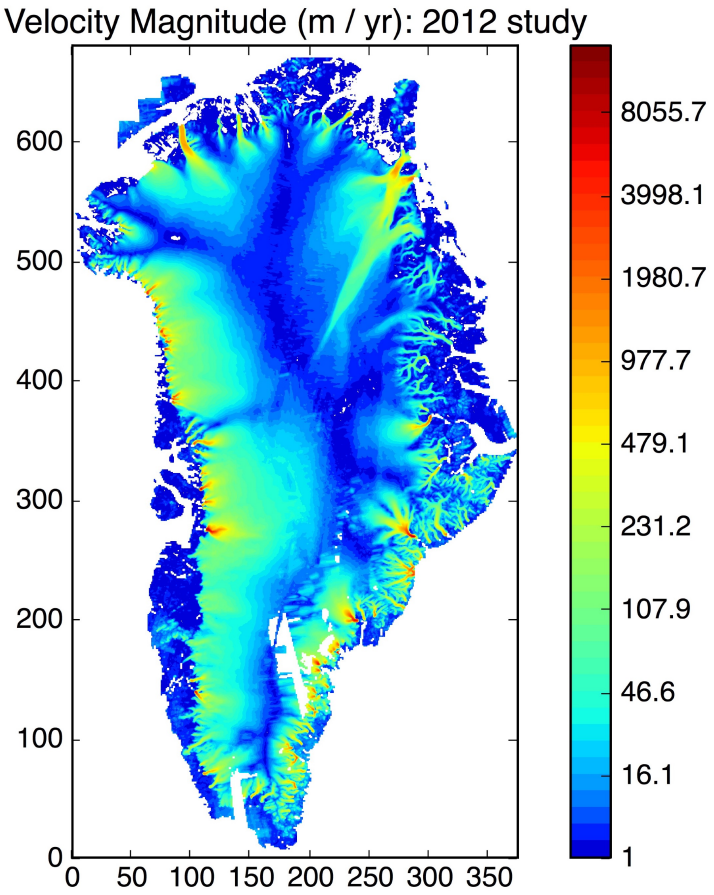
GLIMMER-CISM = COMMUNITY ICE SHEET MODEL

Glimmer Ice Sheet Model calculates a set of PDEs for ice thickness, movement, mass balance, melting, calving, land-ice interface,....

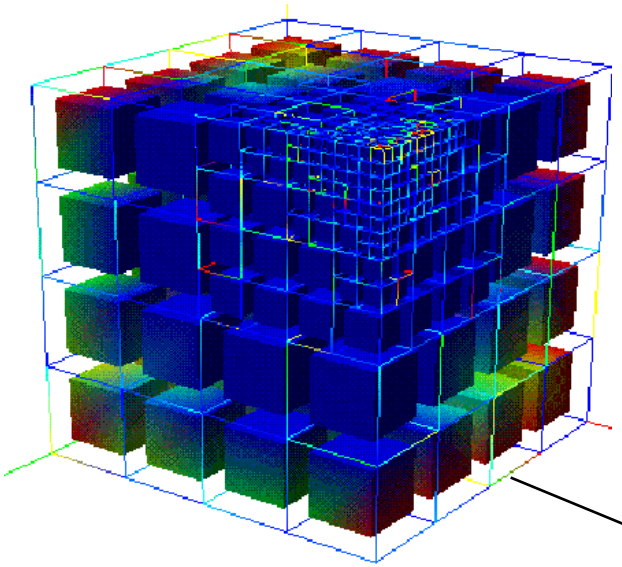
$$\nabla \cdot \mathbf{v} = 0$$

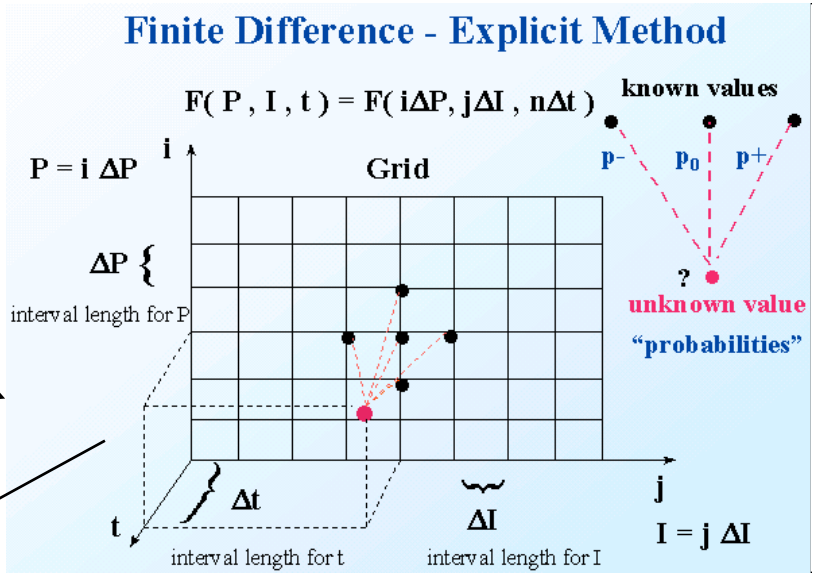
$$\frac{d\mathbf{v}}{dt} = \nabla \cdot \sigma + \rho \mathbf{g}$$

$$\rho \frac{c_P \theta}{dt} = \nabla (k_i \nabla \theta) + \Phi$$



Solving PDEs





	solution u at edge grid points (corner points not included)
west edge	$u_{i,1}^{k+1} = \frac{1}{4}\left(2h\,g_{i,1} + 2u_{i,2}^{k} + u_{i-1,1}^{k} + u_{i+1,1}^{k} - h^2 f_{i,1}\right),\ i = 2 \ldots n-1$
south edge	$u_{n,j}^{k+1} = \frac{1}{4}\left(2h\,g_{n,j} + 2u_{n-1,j}^{k} + u_{n,j-1}^{k} + u_{n,j+1}^{k} - h^2 f_{n,j}\right),\ j = 2 \ldots n-1$
east edge	$u_{i,n}^{k+1} = \frac{1}{4}\left(2h\,g_{i,n} + 2u_{i,n-1}^{k} + u_{i-1,n}^{k} + u_{i+1,n}^{k} - h^2 f_{i,n}\right),\ i = 2 \ldots n-1$
north edge	$u_{1,j}^{k+1} = \frac{1}{4}\left(2h\,g_{1,j} + 2u_{2,j}^{k} + u_{1,j-1}^{k} + u_{1,j+1}^{k} - h^2 f_{1,j}\right),\ j = 2 \ldots n-1$

$$u^{n+1} = \begin{pmatrix} 1-2\mu & \mu & & \\ \mu & \ddots & \ddots & \\ & \ddots & \ddots & \mu \\ & & \mu & 1-2\mu \end{pmatrix} u^{n}$$

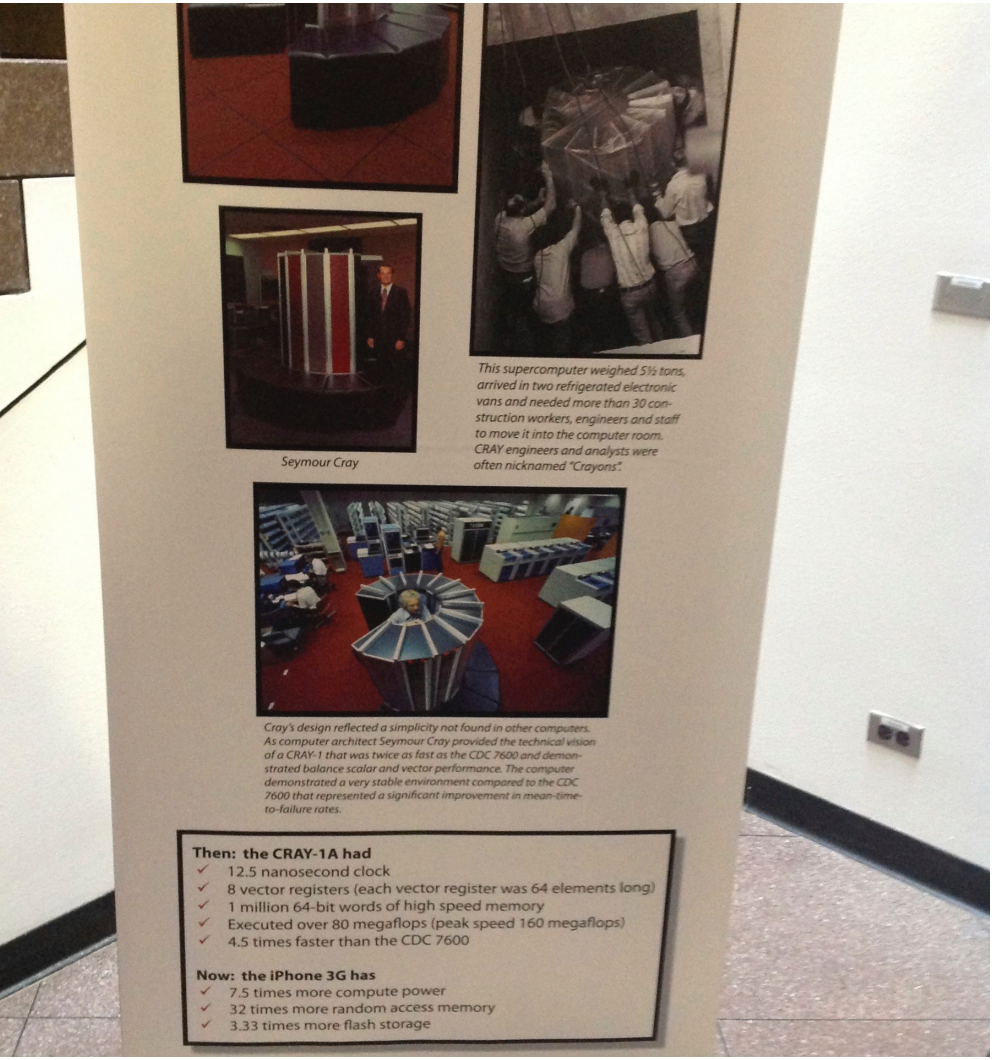
Components of a Cray

Processors

- I'm gonna speed up my gaming system by getting a Cray processor!
- Seymour Cray: "Anyone can build a fast CPU. The trick is to build a fast system."

Cray 1

The National Center for Atmospheric Research (NCAR) was first official customer of Cray Research in 1977, paying US$8.86 million ($7.9 million plus $1 million for the disks) for serial number 3. Over 80 machines were eventually sold.



Cray 1 Numbers

	Cray 1A	iPhone 3G
Clock	12.5 ns = 80Mhz	412Mhz
Vector Registers	8 each 64 elements long	
Memory	1M 64-bit words of memory	32x more
FLOPs	80-160 MFLOPs	7.5x faster
Weight	5.5 tons	1.46E-4 tons

Cray Processors Summary

- Seymour Cray: "Computers should obey a square law -- when the price doubles, you should get at least four times as much speed."
- Wanted to deliver 10x processor performance with each generation
- Forced development more exotic processors like Gallium-Arsenide
- Now commodity processors from AMD

Massively Parallel Computing

- Seymour Cray resisted the massively parallel. “If you were plowing a field, which would you rather use: Two strong oxen or 1024 chickens?”
- By mid-90s compiler technology had improved to support Massively Parallel programming
- Started a new company to design his own MP machine. Focused on communications and memory performance bottlenecks


THE CRAY-1 SUPERCOMPUTER
Featuring a central column surrounded by a padded, circular seat, the Cray-1 looked like no other computer. And performed like no other computer. It reigned as the world's fastest from 1976 to 1982.
Its distinctive design reflected Seymour Cray's innovative engineering solutions and theatrical flair. The round tower minimized wire lengths, while the distinctive bench concealed power supplies. Densely packed integrated circuits and a novel cooling system reflect Cray's attention to "packaging and plumbing."

1
Seymour Cray in a meeting, ca. 1987
Cray usually worked alone or with a small team. He wanted to focus on the engineering tasks and remain insulated from company bureaucracy. Several times, when the company encroached too much, he moved his engineering team to a new location.

2
Employees assembling a Cray backplane, ca. 1987
The wires were all installed by hand and cut to precise lengths to achieve a particular signal delay.

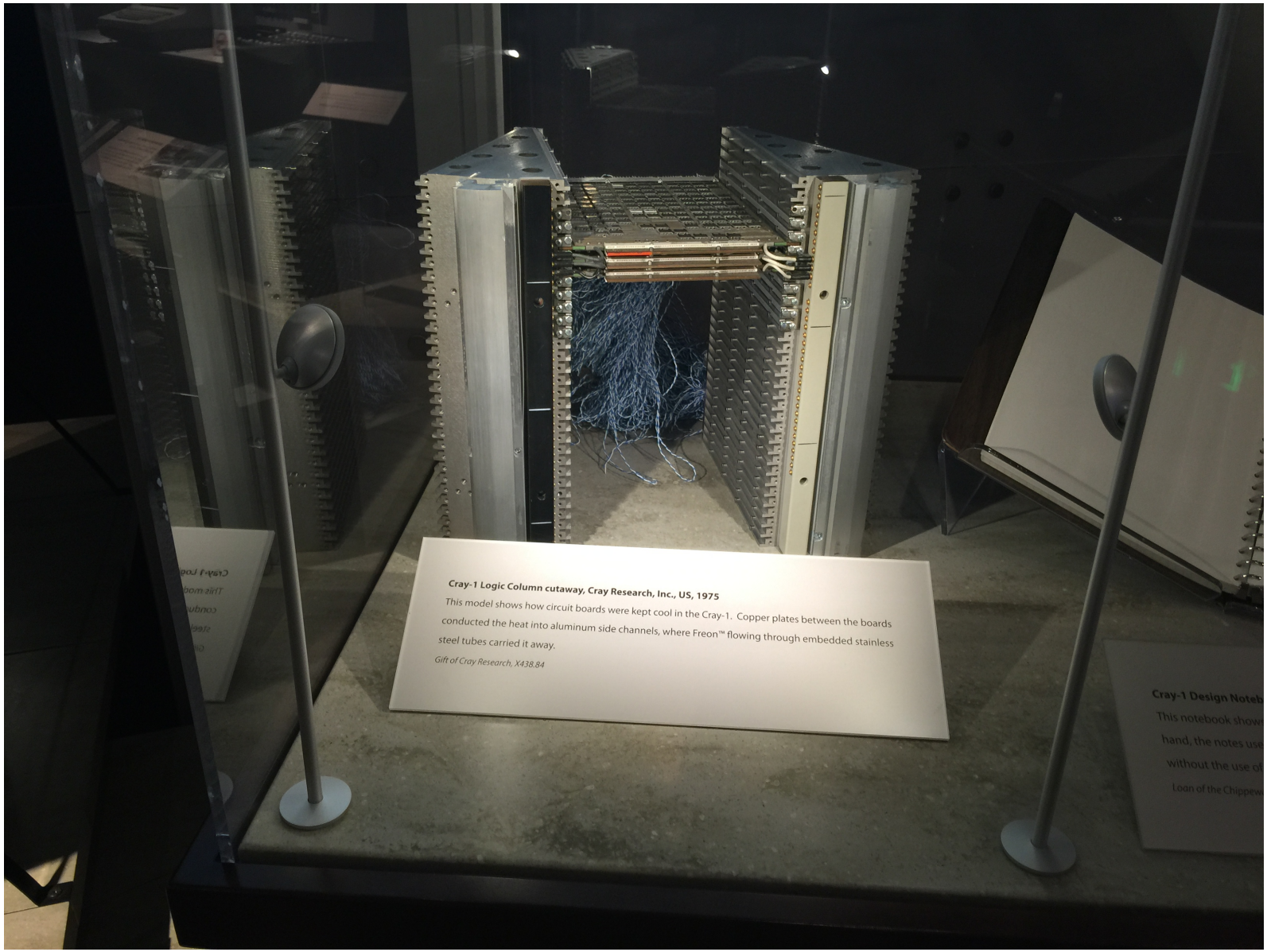
THE FATHER OF SUPERCOMPUTING

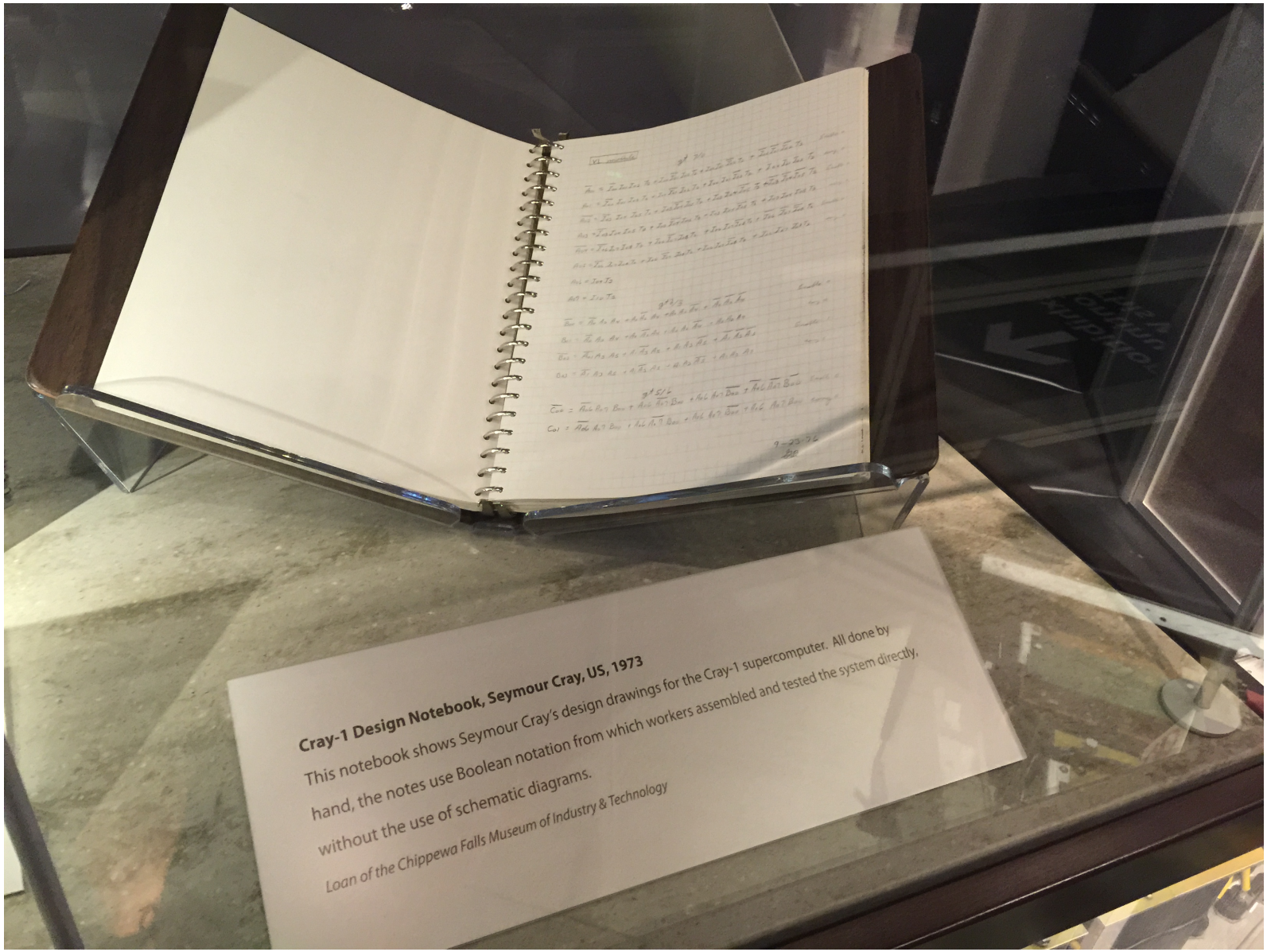
"One of my guiding principles" observed Seymour Cray, "is, 'don't do anything that other people are doing.'"

Cray was a brilliant, soft-spoken computer designer who made a career of building the world's fastest computers, time and again. He preferred working in small teams, undisturbed by managers. Or better still, working alone at night, free from interruptions.

Cray's quirky work habits were matched by his unusual diversions, which included digging tunnels near his house and, once, burning a boat he had built because he'd built a new one.


CRAY-1
CRAY RESEARCH, INC.

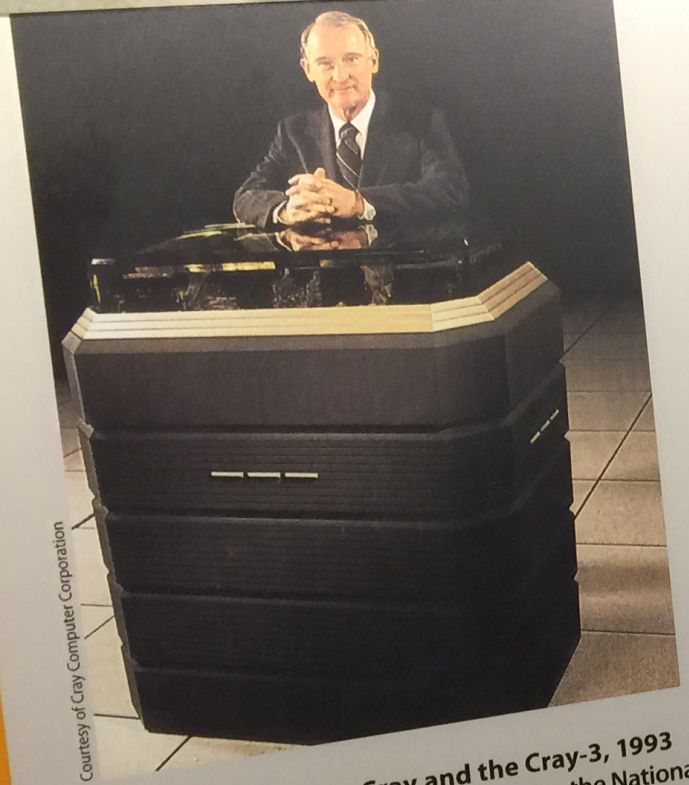

Cray-1 Logic Column cutaway, Cray Research, Inc., US, 1975
This model shows how circuit boards were kept cool in the Cray-1. Copper plates between the boards conducted the heat into aluminum side channels, where Freon™ flowing through embedded stainless steel tubes carried it away.
Gift of Cray Research, X438.84
Cray-1 Design Noteb
This notebook shows
hand, the notes use
without the use of
Loan of the Chippew



Cray-1 Design Notebook, Seymour Cray, US, 1973

This notebook shows Seymour Cray's design drawings for the Cray-1 supercomputer. All done by hand, the notes use Boolean notation from which workers assembled and tested the system directly, without the use of schematic diagrams.

Loan of the Chippewa Falls Museum of Industry & Technology



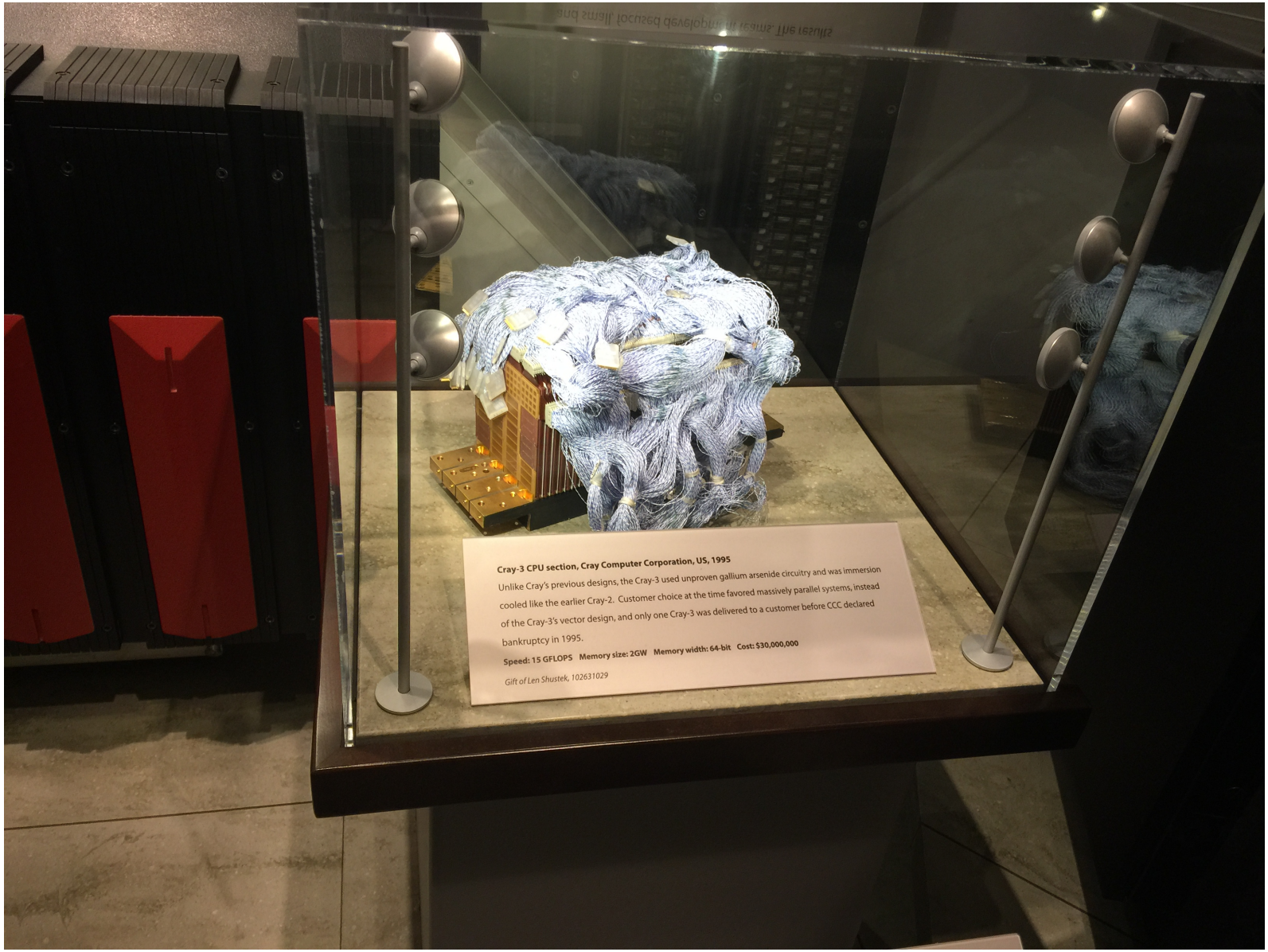
Courtesy of Cray Computer Corporation

Designer Seymour Cray and the Cray-3, 1993
The only Cray-3 made was delivered to the National Center for Atmospheric Research in 1993, but was soon taken out of service because it was unreliable.

HITTING THE WALL: THE CRAY-3

The Cray-3 was to be 10 times faster than the Cray-2. But Seymour Cray, unwilling to join the trend toward using many slower processors, switched to exotic gallium arsenide chips, packing more than a thousand in each 4-cubic-inch module.

This risky path to speed backfired. Cray sold only one partially completed machine.


Cray-3 CPU section, Cray Computer Corporation, US, 1995
Unlike Cray's previous designs, the Cray-3 used unproven gallium arsenide circuitry and was immersion cooled like the earlier Cray-2. Customer choice at the time favored massively parallel systems, instead of the Cray-3's vector design, and only one Cray-3 was delivered to a customer before CCC declared bankruptcy in 1995.
Speed: 15 GFLOPS Memory size: 2GW Memory width: 64-bit Cost: $30,000,000
Gift of Len Shustek, 102631029


Labs fought to own the coveted "Serial One" model.

Massively Parallel Computing

Processors	1 processor, dual processors	Message Passing Interface (MPI)	Library and Function Calls
Cores	4, 6, 8, 16-cores	Threads, ForkJoin, OpenMP	Compiler Extensions
Vector Units	2-4 per core	SSE extensions, Vectorization	CPU Instructions, Compilers
GPUs	1500+ Cores	CUDA, OpenACC	Compilers
Debugging/Profiling	Across tens of thousands of these processors		

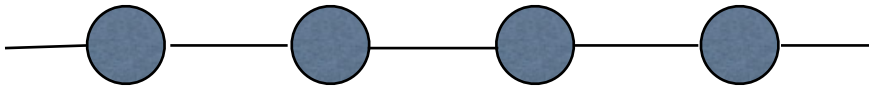
Scalable Problems or Algorithms can effectively utilize Massively Parallel Computing

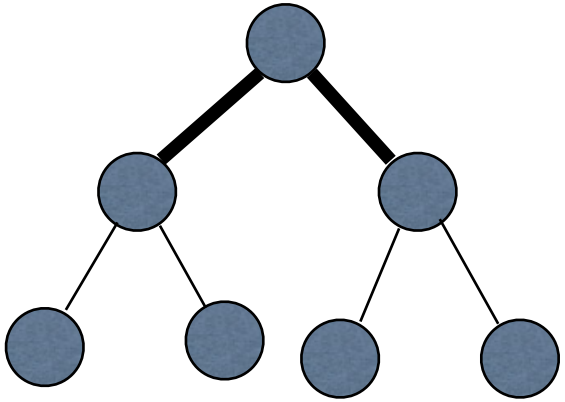
Parallel Summary

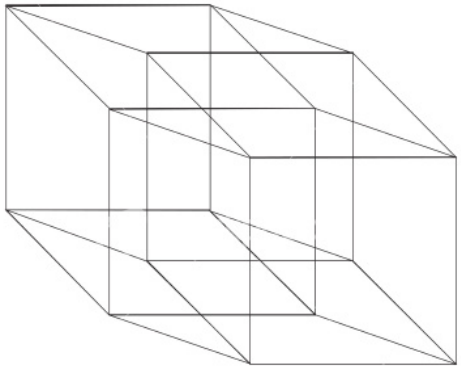
- "In 1978, the first standard software package consisting of the Cray Operating System (COS), the first automatically vectorizing Fortran compiler (CFT), and the Cray Assembler Language (CAL) were introduced."
- Cray compilers are tuned and applicable only to a Cray. Do remain cutting-edge; e.g. OpenACC.
- MPI, OpenMPI, MPICH
- OpenMP, pthreads, .NET Language extensions
- GCC Compiler Suite
- CUDA is free to download from nVidia. OpenCL is also free.

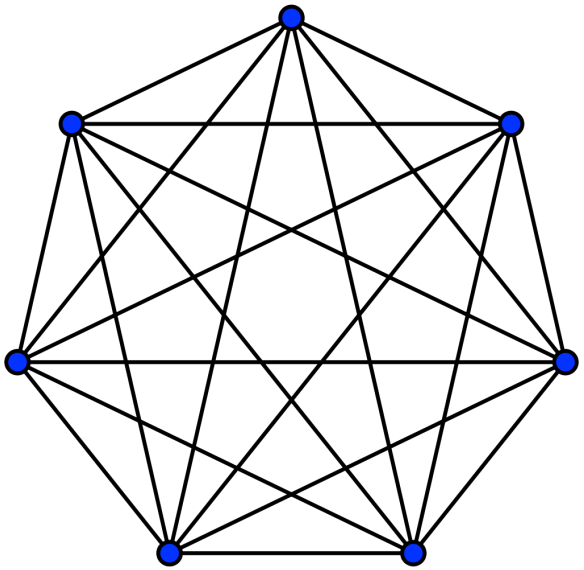
Interconnects

How to network and communicate between the parallel nodes?



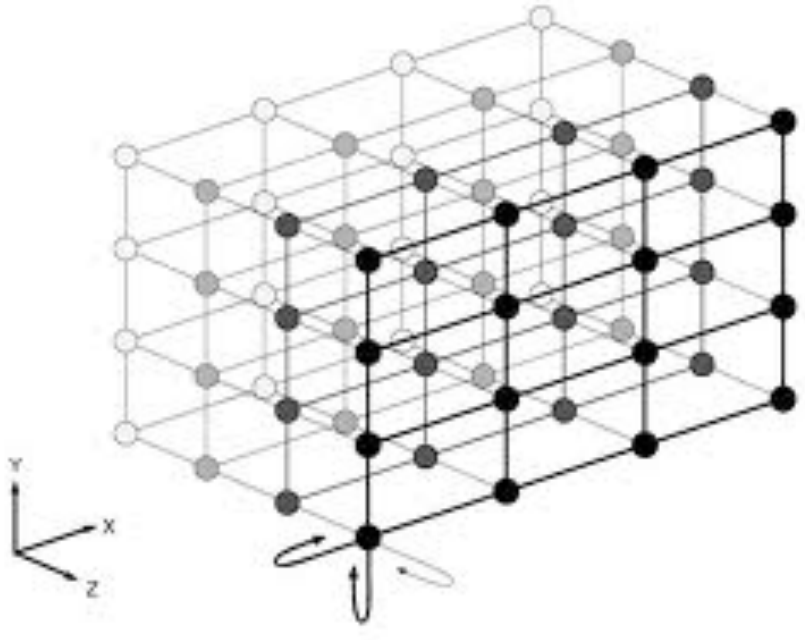






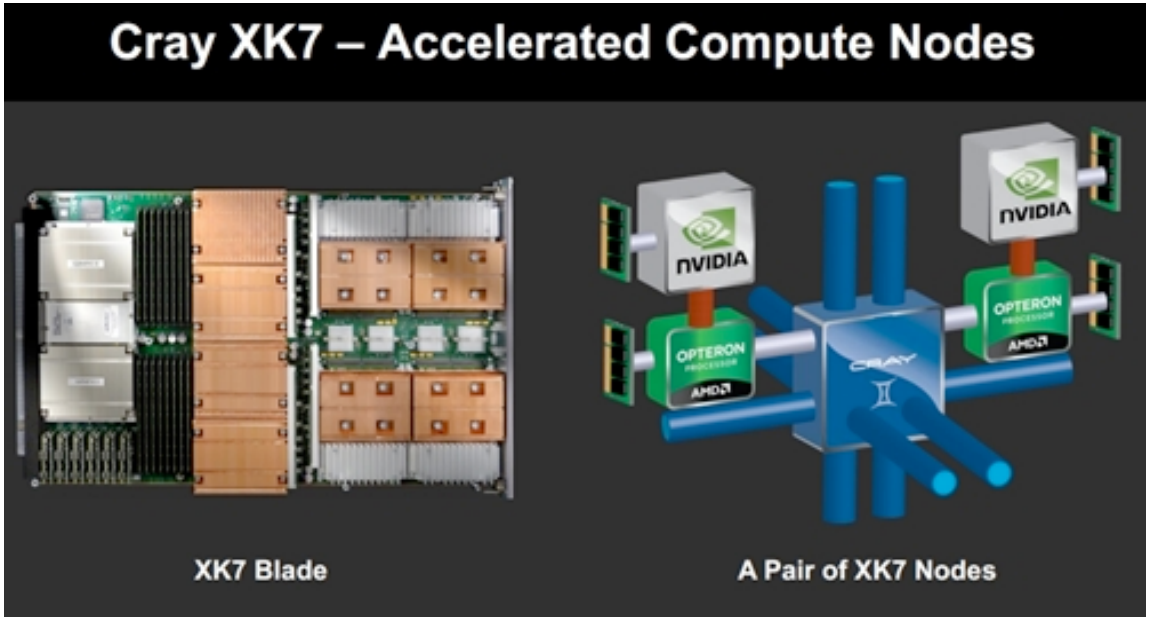
3D Torus Interconnect

Cray T3D - Torus is the mathematical name for the shape of a doughnut



Interconnects Summary

- Seastar previous generation. Gemini current generation.
- Exotic and Unique to Cray supercomputers
- Seastar failed lose whole machine. Gemini can reroute and recover.
- GigE switched network is good enough for most of us.



Data Storage

- Compute nodes are effectively netbooted and run from RAM disk (no local storage)
- Huge data sets required for initial data.
- Huge output storage and throughput requirements
- Huge potential bottleneck: Initial data loaded across all the compute nodes. 200,000 computers waiting for the hard disk!
- Parallel file systems active area of research. LUSTRE (lustre.org)

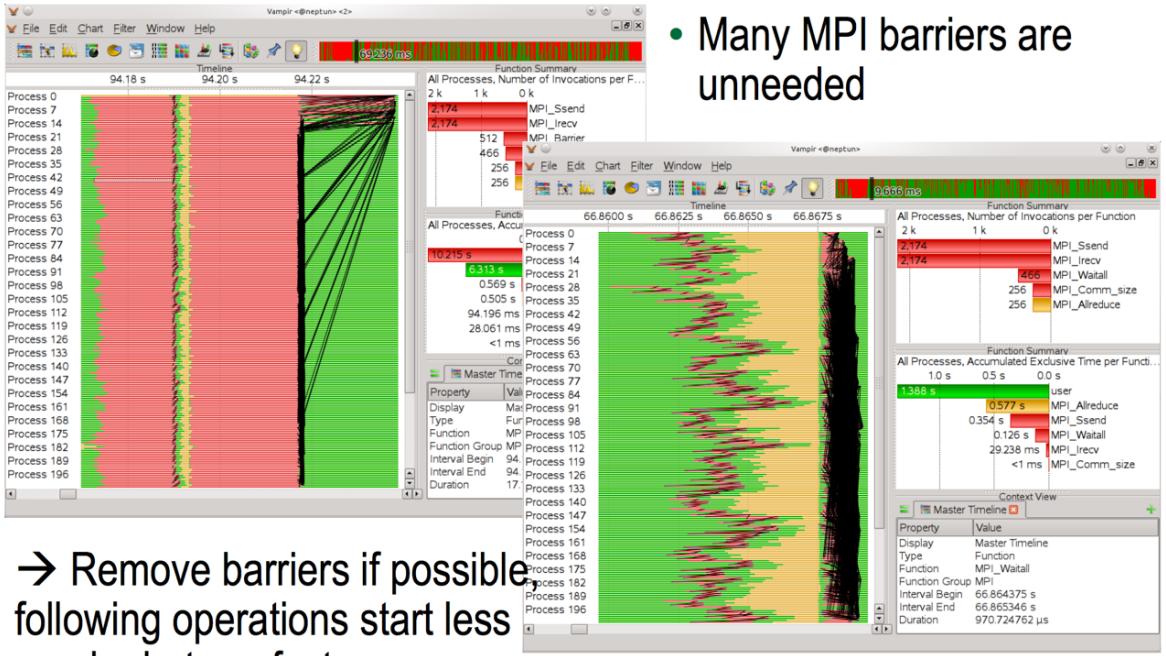
~~Current Filesystem:~~
~~14K SATA Drives~~
~~RAID6 8+2 redundancy~~
~~10-20 Hard Drive Failures per month~~

New Filesystem:
20,160 2TB SAS Drives
32 Petabytes
1.2TB/s transfer rate.


National
Center for
Computational
Sciences
INCITE

Programming/Operating Systems

- Operating System - Used to be proprietary, now a version of Linux, Cray Linux Environment (CLE)
- Cray Compilers
- PGI: Portland Group, HPC support.
- Intel Compilers. Multi-threading and vectorization. Getting most out of CPU.
- Libraries - Many are open-source or free from universities: OpenMPI, OpenMP/ pthreads, CUDA, gcc compiler suite
- Debugging, Profiling, Development Tools, Optimized Libraries and Applications, and Visualization Tools



Power/Cooling/Infrastructure

- Integrated Cooling - Seymour Cray design feature. Crays will not operate if the environment is unsatisfactory. Environmental Interlock: story of frozen computer.
- Power - 8.3 MW for the LINPACK Top500 Run. 8300 KW * 1 hour = 8300 KWh * $0.13/KWh = $1079/hr power bill ~ $26,000/day ~ $775K/month
- Thermal Efficiency
- Constraint to Exascale. 1000-fold increase in computing power from petaflop. 8300 MW = 8.3GW
- Infrastructure and Maintenance of space to hold it

Summary

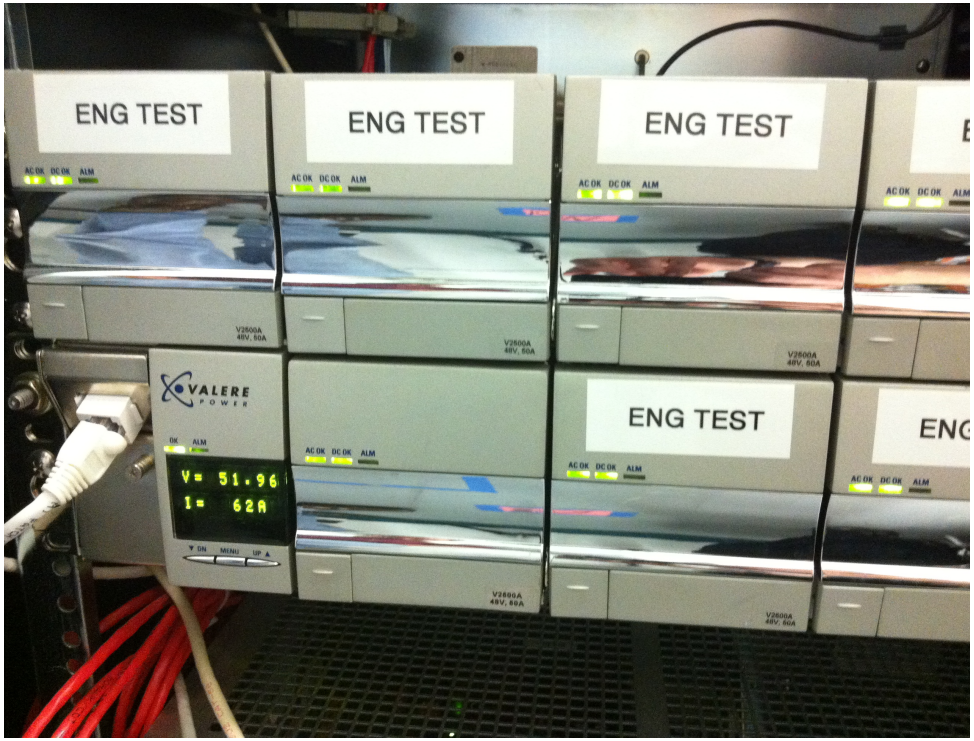
- Crays are wonderful machines, but are designed for scalable, scientific computation problems
- There are several exotic components, but the ones that impact us are commodities
- Our phones and computers are wonderful, little, parallel computers
- Most the software and resources used on the Cray are freely available
- Everybody can build their own Linux cluster and use the same software used on the Crays

Cray XT-3 at Makerspace

Cray XT-3 at the Dillow House

Power

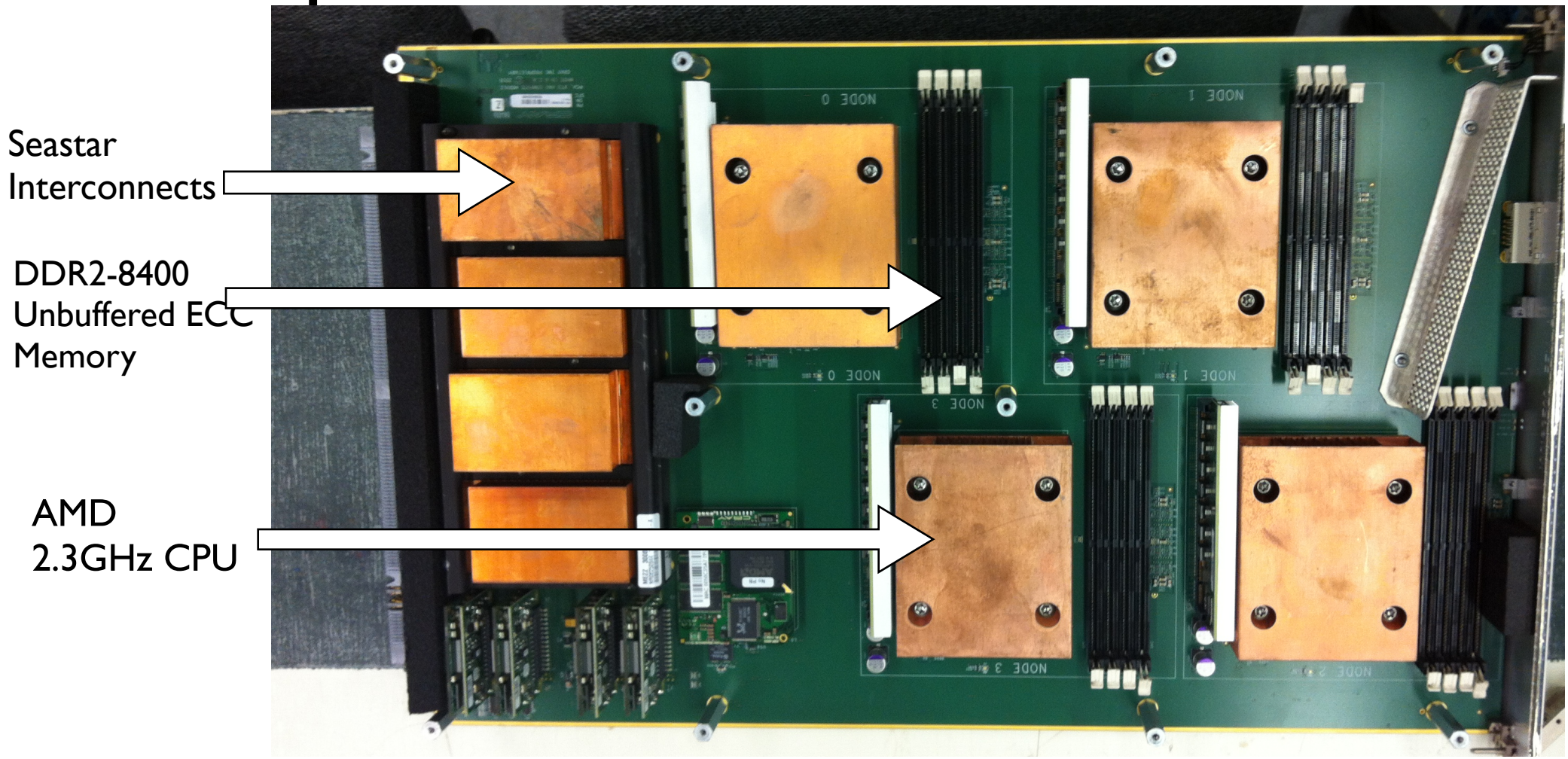
- 3-Phase 208V. Needed an electrician to install.
- Wants 100 Amp Circuit, We have 30 Amp circuit
- Uses about 3.5KW, so about $0.60/ hour to run, $14.30 per day, $432/ month



Cooling and Space

- Discussions of using single-phase car fan
- Using the original, 2200 CFM fan on 3-phase motor
- 15 degree Fahrenheit temperature rise between incoming and exhaust air
- We have a 25’ x 50’ x 20’ (25,000 cubic feet) space at the Oak Ridge Entrepreneurial Center

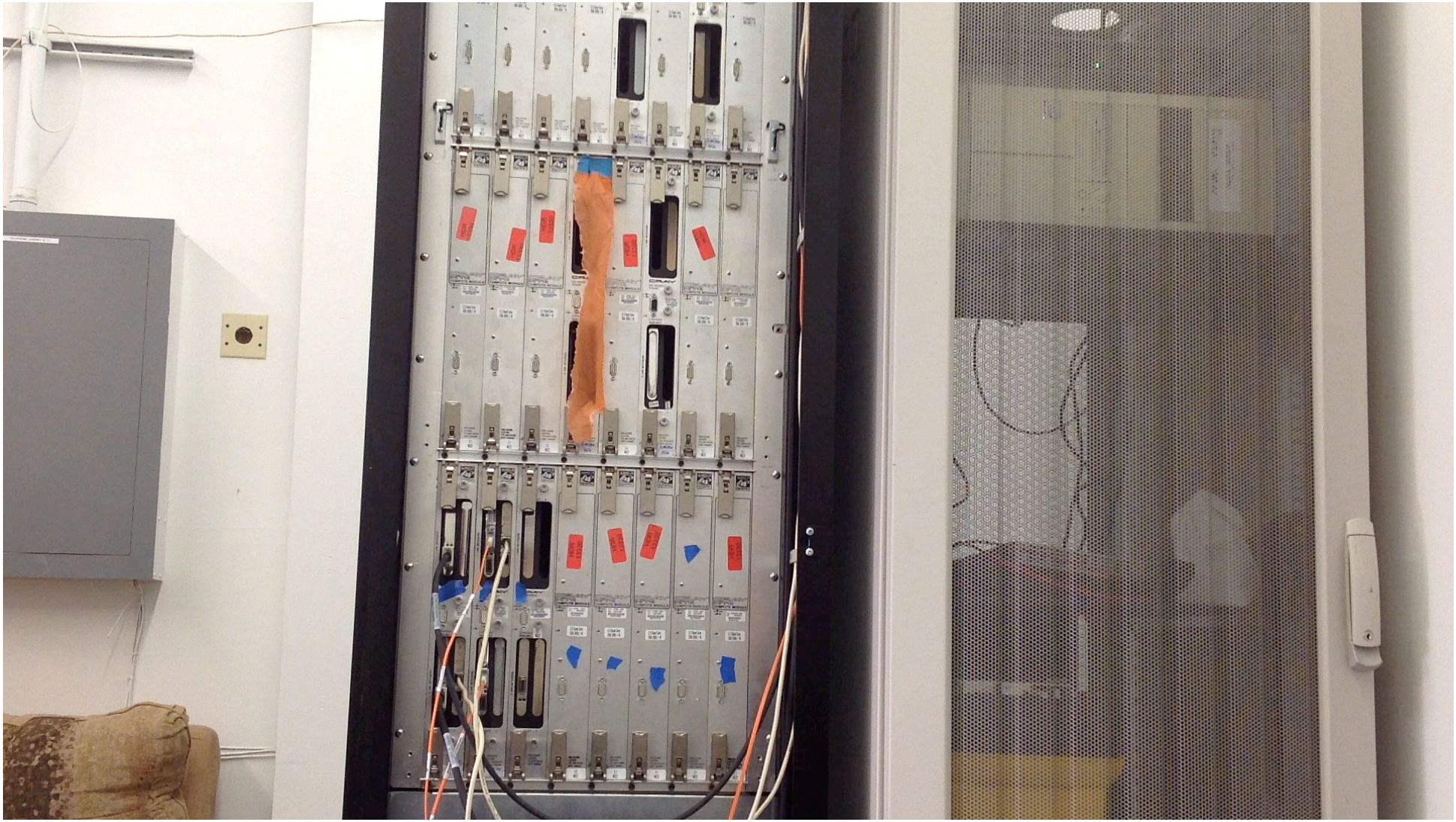
Compute Blades = 4 Modules/Nodes

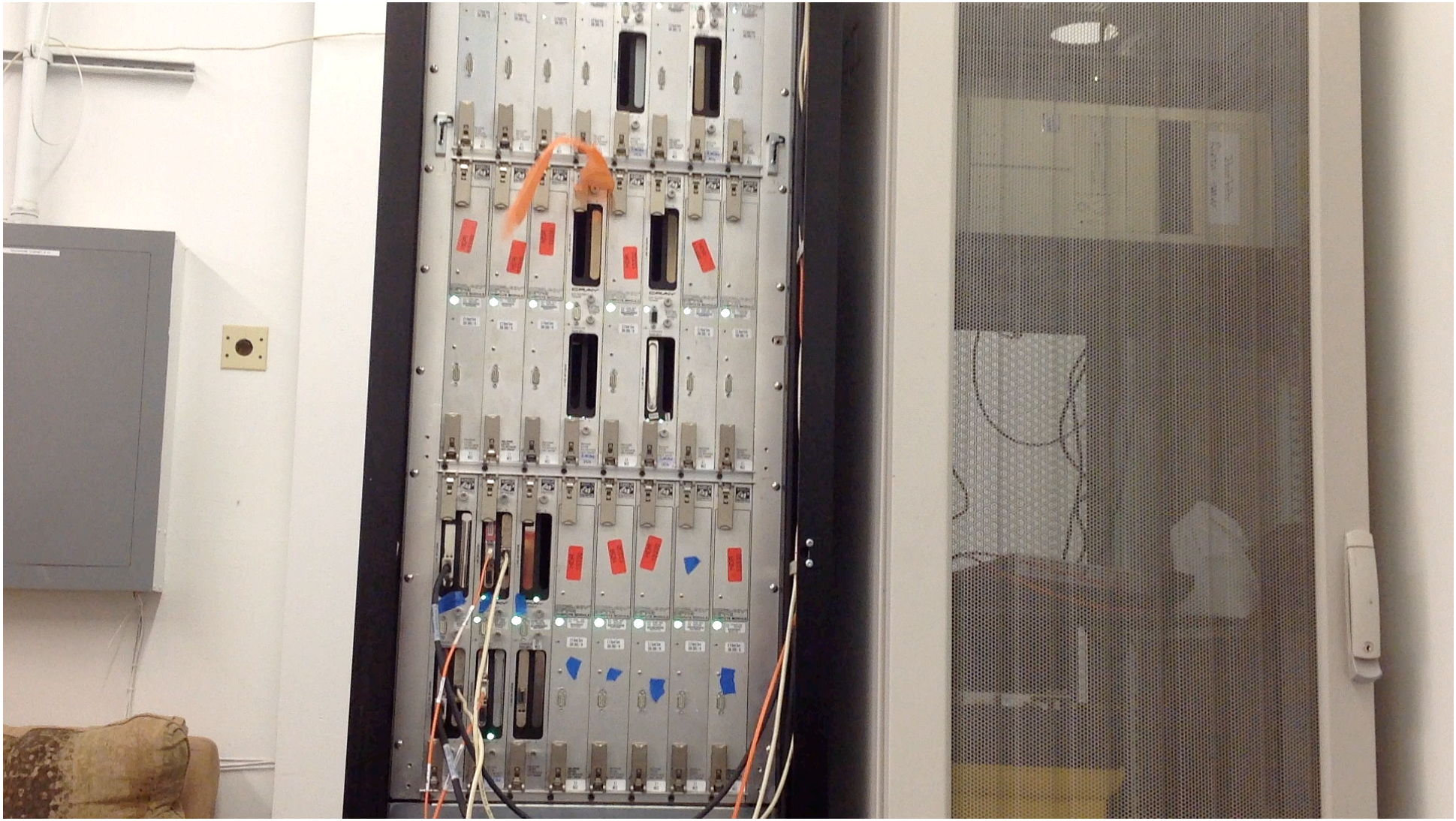


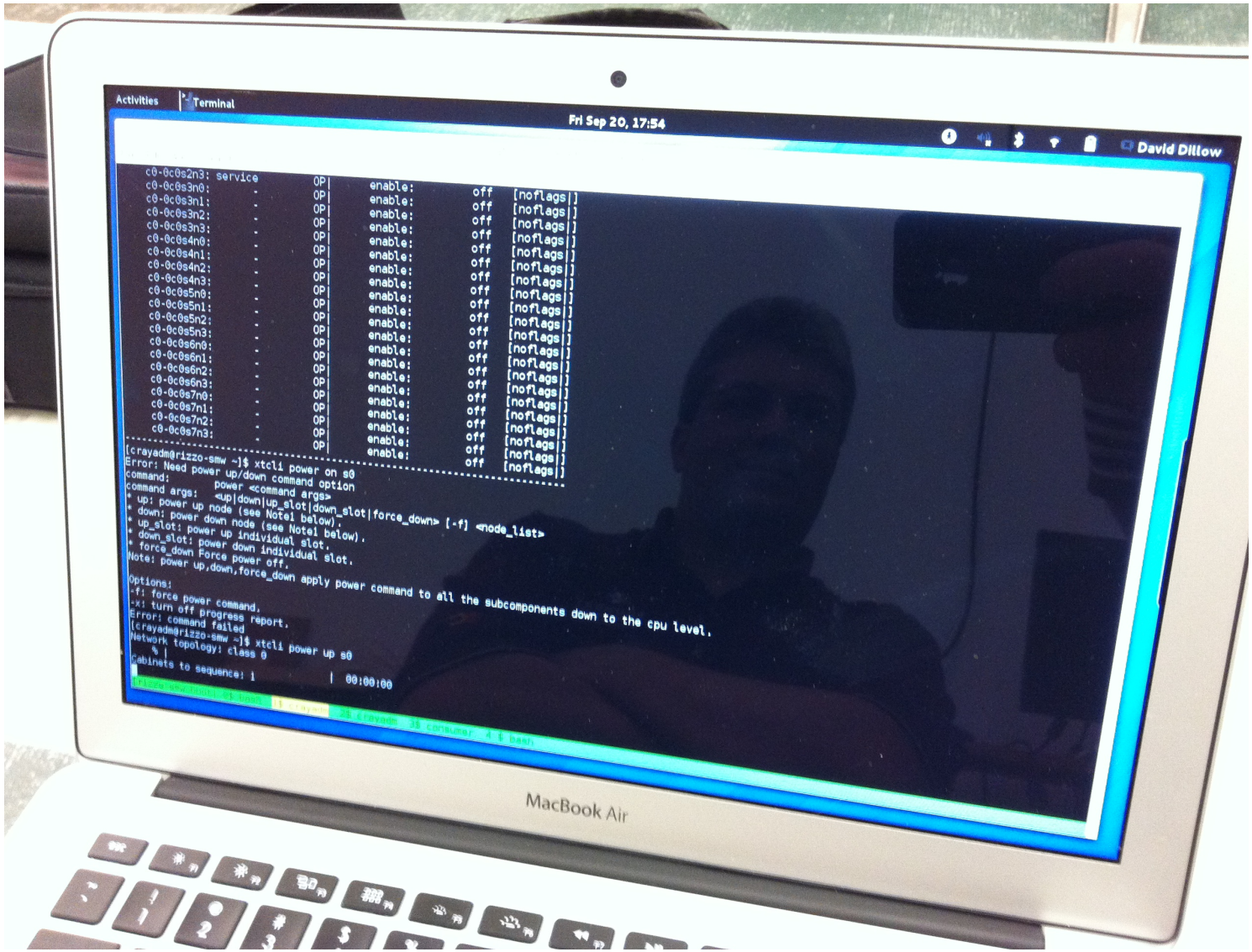
Interconnects

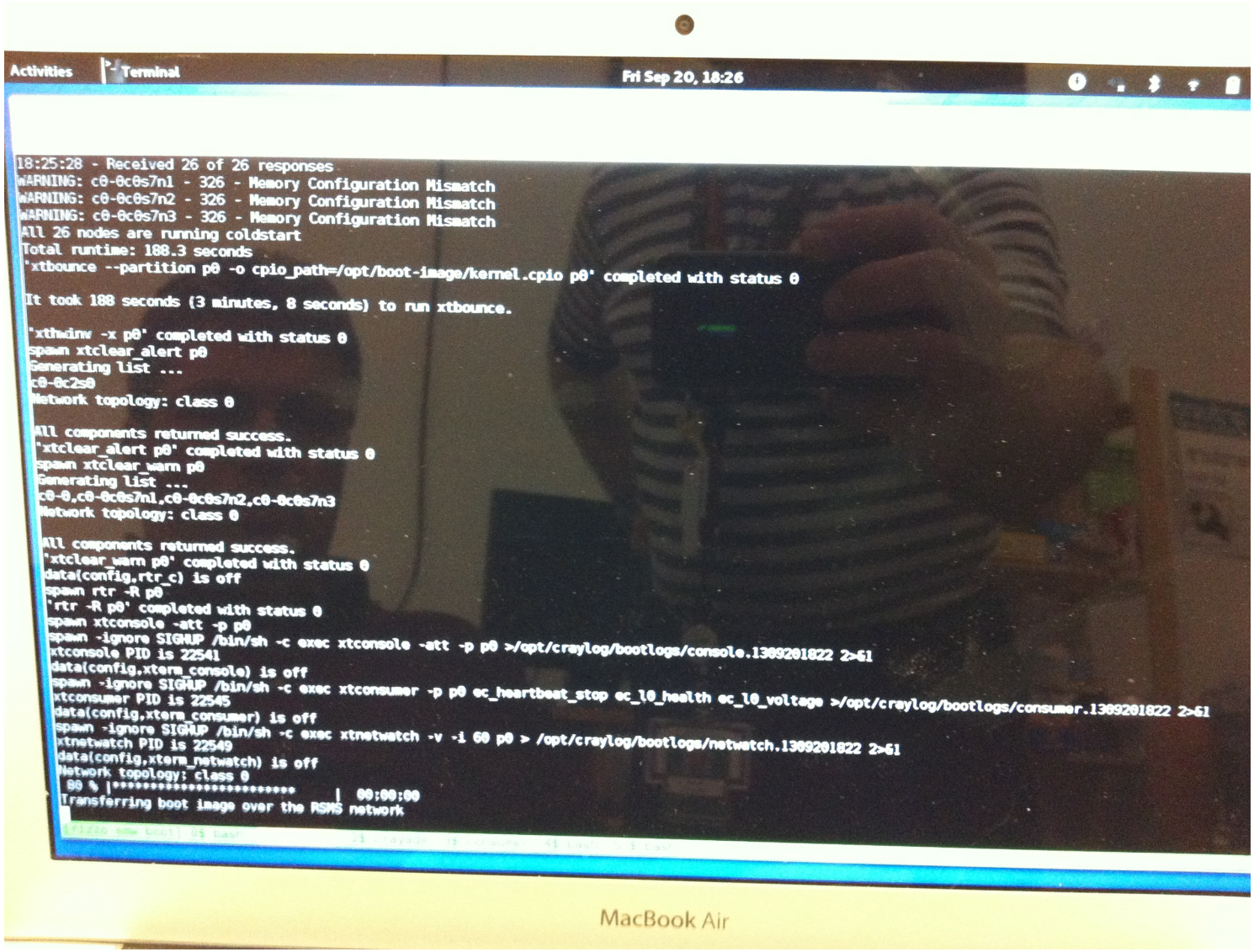

CRAY XT3
CRAY
XT3

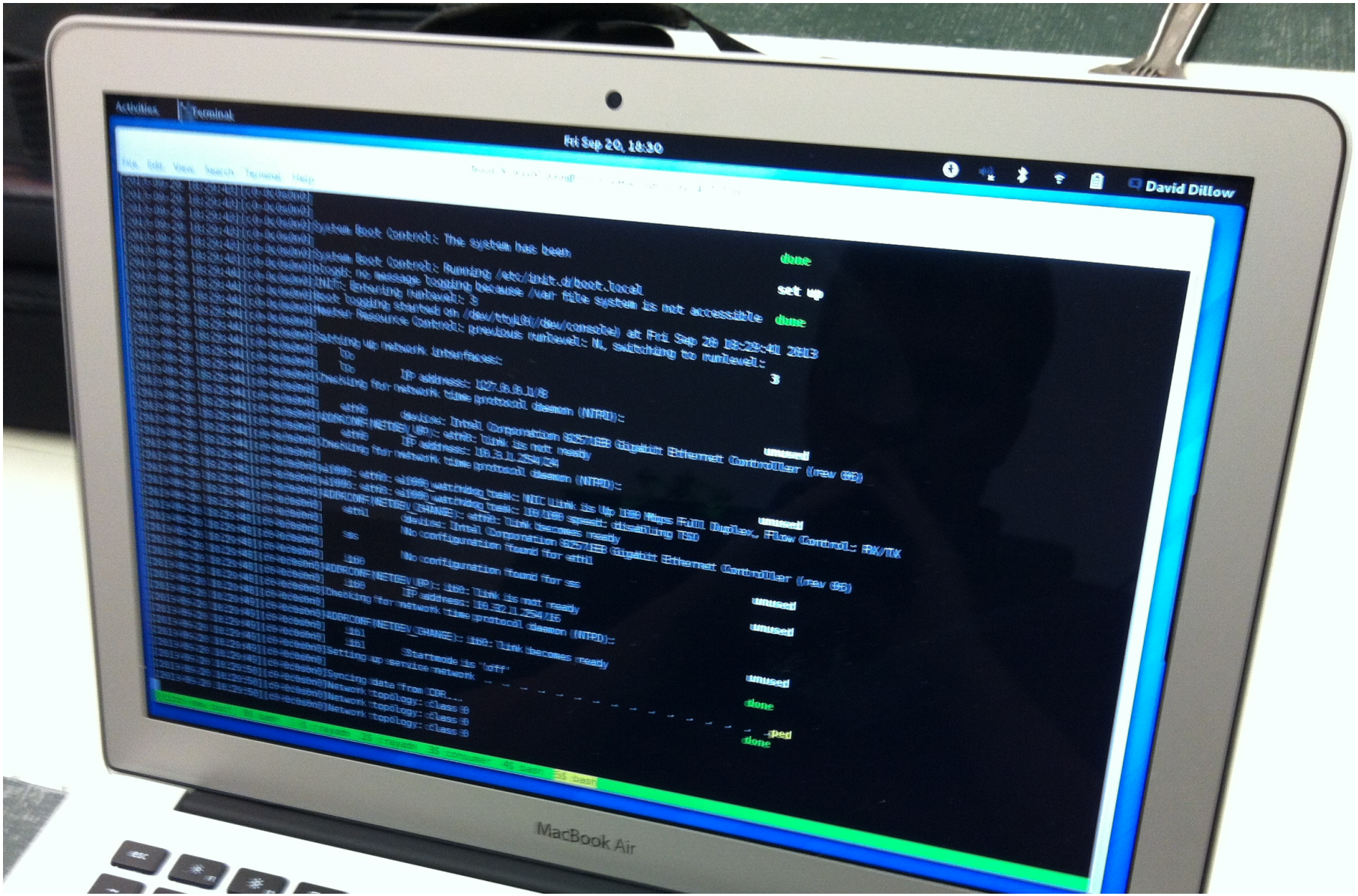
Starting it up...

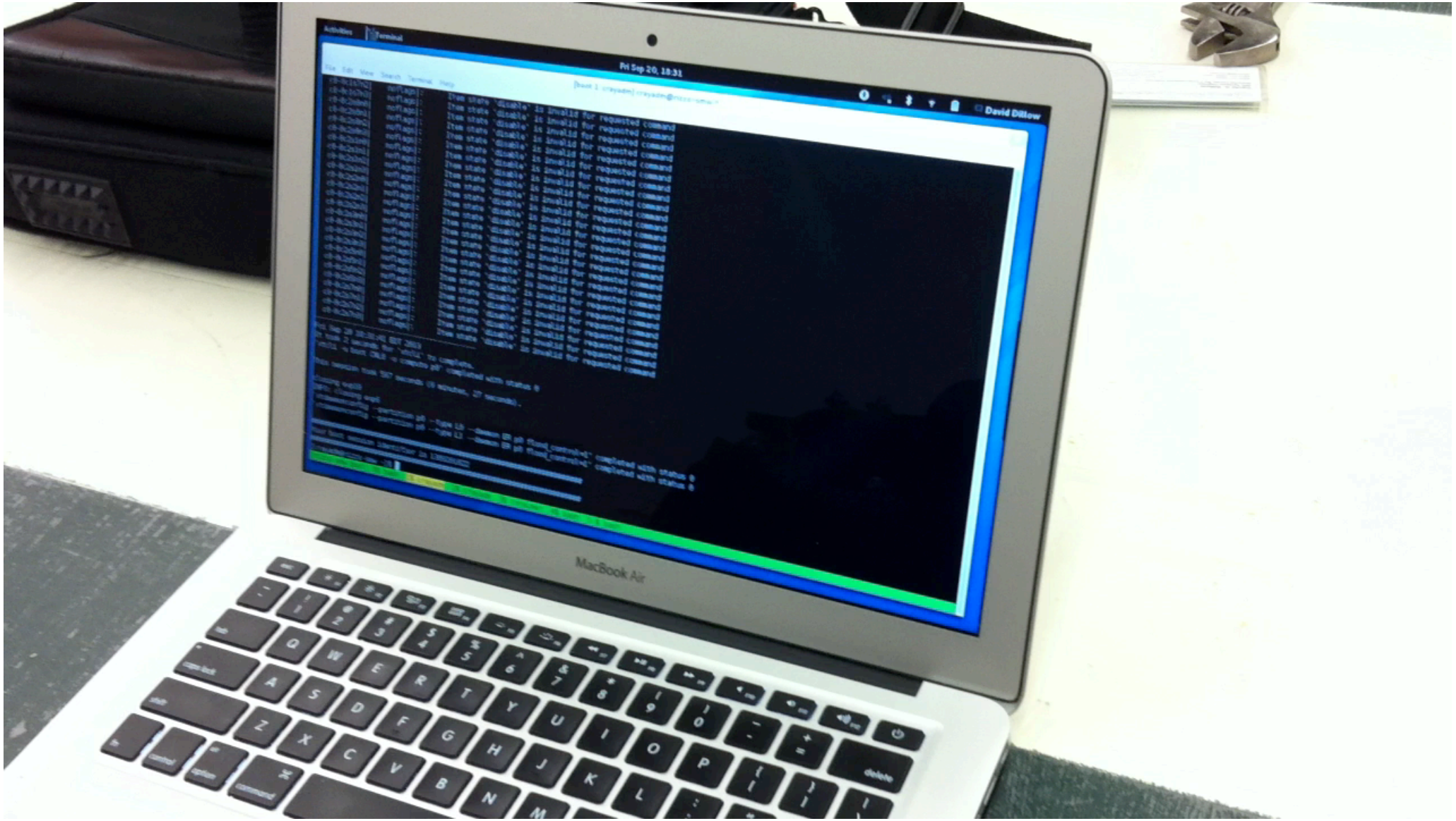





Activities
Terminal
Fri Sep 20, 17:54
David Dillow
c0-0c0s2n3: service OP| enable: off [noflags|]
c0-0c0s3n0: - OP| enable: off [noflags|]
c0-0c0s3n1: - OP| enable: off [noflags|]
c0-0c0s3n2: - OP| enable: off [noflags|]
c0-0c0s3n3: - OP| enable: off [noflags|]
c0-0c0s4n0: - OP| enable: off [noflags|]
c0-0c0s4n1: - OP| enable: off [noflags|]
c0-0c0s4n2: - OP| enable: off [noflags|]
c0-0c0s4n3: - OP| enable: off [noflags|]
c0-0c0s5n0: - OP| enable: off [noflags|]
c0-0c0s5n1: - OP| enable: off [noflags|]
c0-0c0s5n2: - OP| enable: off [noflags|]
c0-0c0s5n3: - OP| enable: off [noflags|]
c0-0c0s6n0: - OP| enable: off [noflags|]
c0-0c0s6n1: - OP| enable: off [noflags|]
c0-0c0s6n2: - OP| enable: off [noflags|]
c0-0c0s6n3: - OP| enable: off [noflags|]
c0-0c0s7n0: - OP| enable: off [noflags|]
c0-0c0s7n1: - OP| enable: off [noflags|]
c0-0c0s7n2: - OP| enable: off [noflags|]
c0-0c0s7n3: - OP| enable: off [noflags|]
[crayadm@rizzo-smw ~]$ xtcli power on s0
Error: Need power up/down command option
command: power <command args>
command args: <up|down|up_slot|down_slot|force_down> [-f] <node_list>
* up: power up node (see Note1 below).
* down: power down node (see Note1 below).
* up_slot: power up individual slot.
* down_slot: power down individual slot.
* force_down Force power off.
Note: power up,down,force_down apply power command to all the subcomponents down to the cpu level.
Options:
-f: force power command.
-x: turn off progress report.
Error: command failed
[crayadm@rizzo-smw ~]$ xtcli power up s0
Network topology: class 0
Cabinets to sequence: 1 | 00:00:00
MacBook Air


Activities
Terminal
Fri Sep 20, 18:26
18:25:28 - Received 26 of 26 responses
WARNING: c0-0c0s7n1 - 326 - Memory Configuration Mismatch
WARNING: c0-0c0s7n2 - 326 - Memory Configuration Mismatch
WARNING: c0-0c0s7n3 - 326 - Memory Configuration Mismatch
All 26 nodes are running coldstart
Total runtime: 188.3 seconds
'xtbounce --partition p0 -o cpio_path=/opt/boot-image/kernel.cpio p0' completed with status 0
It took 188 seconds (3 minutes, 8 seconds) to run xtbounce.
'xthwinv -x p0' completed with status 0
spawn xtclear_alert p0
Generating list ...
c0-0c2s0
Network topology: class 0
All components returned success.
'xtclear_alert p0' completed with status 0
spawn xtclear_warn p0
Generating list ...
c0-0,c0-0c0s7n1,c0-0c0s7n2,c0-0c0s7n3
Network topology: class 0
All components returned success.
'xtclear_warn p0' completed with status 0
data(config,rtr_c) is off
spawn rtr -R p0
'rtr -R p0' completed with status 0
spawn xtconsole -att -p p0
spawn -ignore SIGHUP /bin/sh -c exec xtconsole -att -p p0 >/opt/craylog/bootlogs/console.1389201822 2>&1
xtconsole PID is 22541
data(config,xterm_console) is off
spawn -ignore SIGHUP /bin/sh -c exec xtconsumer -p p0 ec_heartbeat_stop ec_l0_health ec_l0_voltage >/opt/craylog/bootlogs/consumer.1389201822 2>&1
xtconsumer PID is 22545
data(config,xterm_consumer) is off
spawn -ignore SIGHUP /bin/sh -c exec xtnetwatch -v -i 60 p0 > /opt/craylog/bootlogs/netwatch.1389201822 2>&1
xtnetwatch PID is 22549
data(config,xterm_netwatch) is off
Network topology: class 0
80 % |************************ | 00:00:00
Transferring boot image over the RSMS network
MacBook Air


Activities
Terminal
Fri Sep 20, 18:30
David Dillow
MacBook Air


Activities
Terminal
Fri Sep 20, 18:31
David Dillow
MacBook Air

Questions?