ARSC's New x86_64 Supercomputer

Dr. Greg Newby, Chief Scientist

Arctic Region Supercomputing Center

University of Alaska Fairbanks

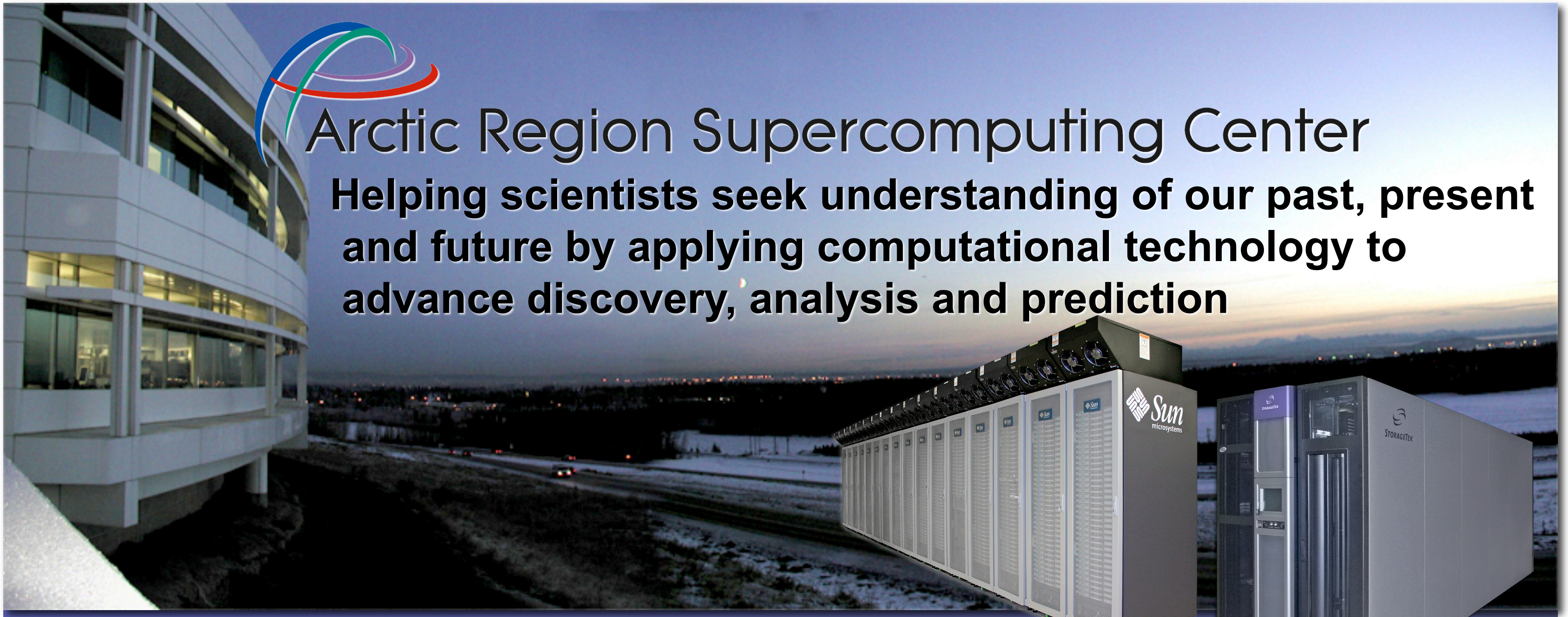
November 10 2007
Sun HPC Consortium, Reno

Presentation Outline

- ✓ **About ARSC**
- System overview
- Focus on nodes
- Focus on storage
- Applications, research
- Recommendations, findings

Arctic Region Supercomputing Center

Helping scientists seek understanding of our past, present and future by applying computational technology to advance discovery, analysis and prediction



ARSC's large, state of the technology systems provide 24/7 accessibility for high-performance computing needs

ARSC's large storage infrastructure provides redundant back-up, swift access and retrieval of archival data

UAF
UNIVERSITY OF
ALASKA
FAIRBANKS

ARSC is...

- A DoD HPCMP Allocated Distributed Center, established in 1993
- University of Alaska Fairbanks owned and operated
    - Provides HPC resources & support
    - Conducts and supports research with an emphasis on northern phenomena

Presentation Outline

- About ARSC
- ✓ System overview
- Focus on nodes
- Focus on storage
- Applications, research
- Recommendations, findings

Midnight, ARSC's Newest Supercomputer

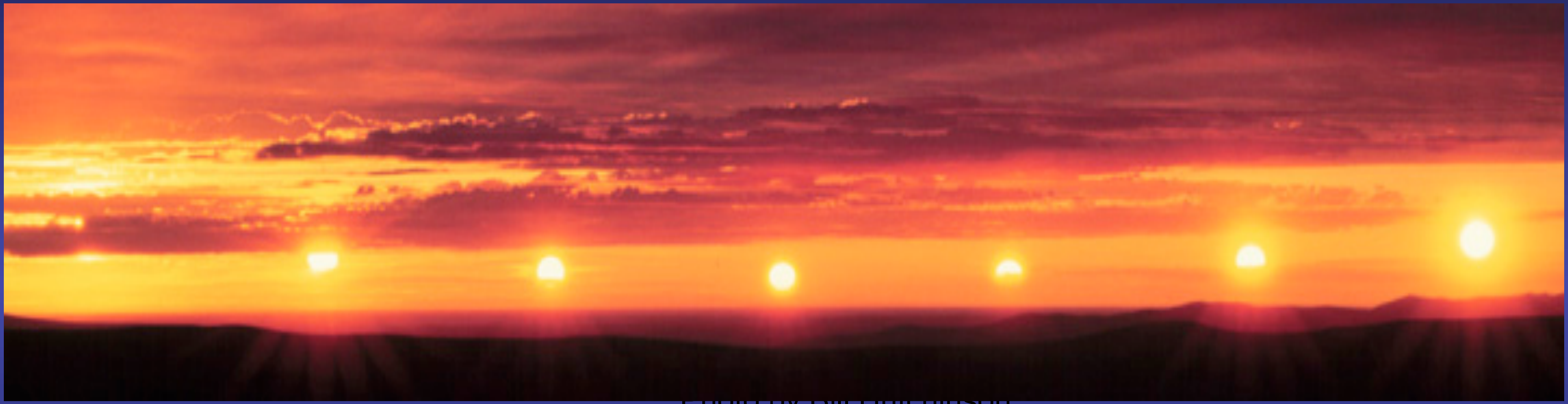


Photo by Bill Hutchinson

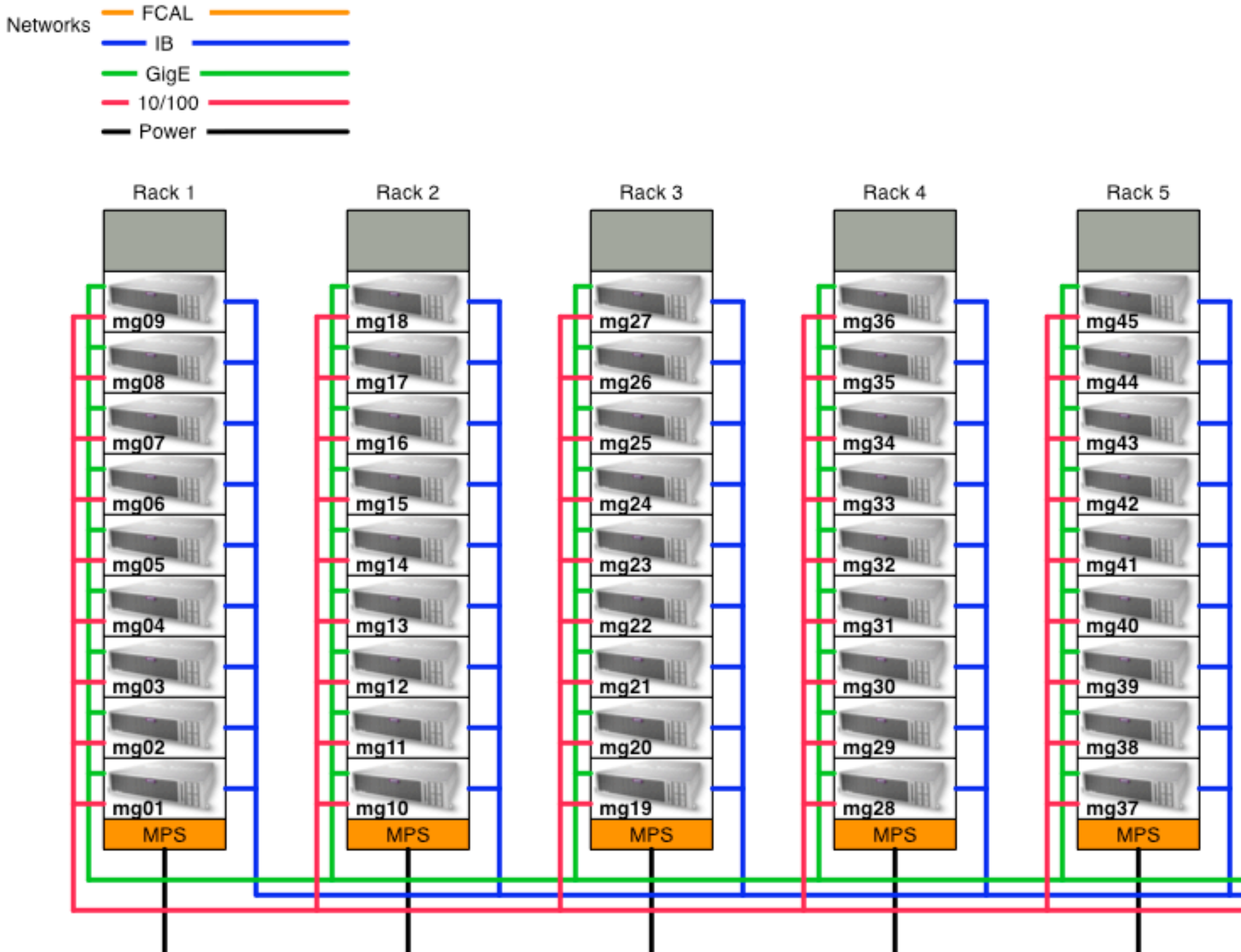
- **Built by Sun Microsystems**
- **Production status on July 16, 2007**
- **2312 computational cores on 413 nodes, 19 racks**
- **Theoretical aggregate peak of ~12.1 TFLOPS**

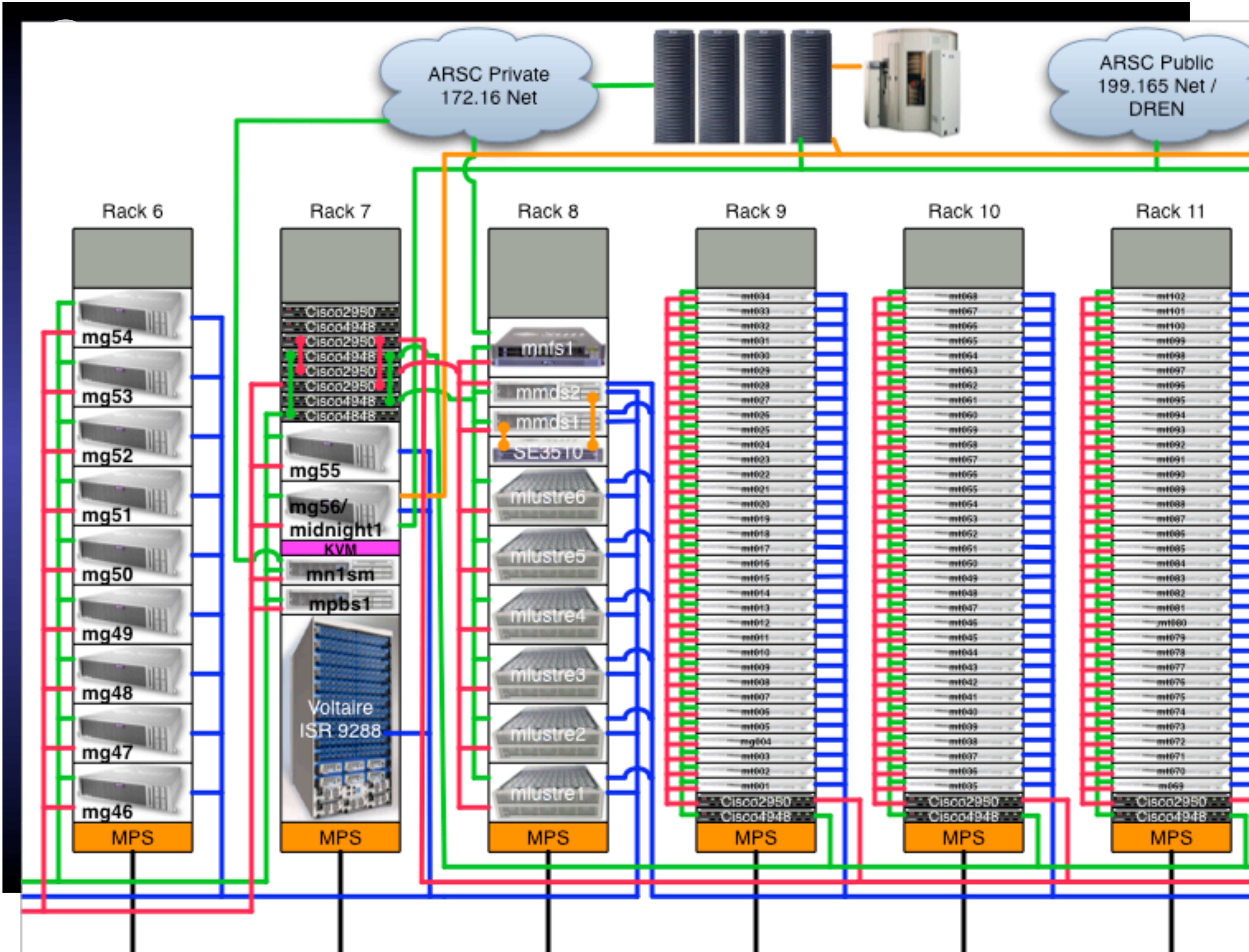
Physical Characteristics

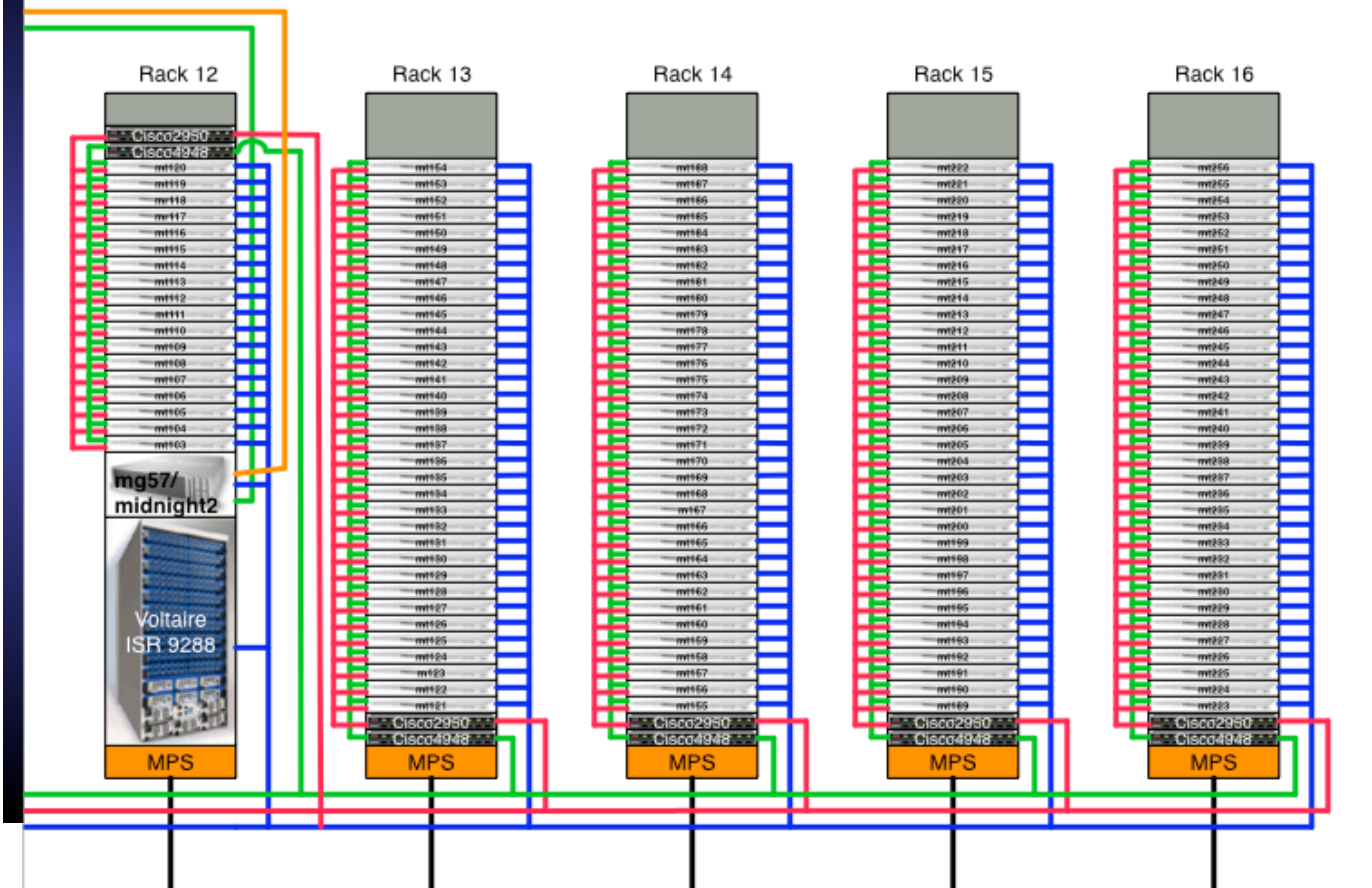
- Nineteen standard size (19”) cabinets with front and rear perforated doors
- Total power consumption of about 200KW (not including cooling)
- Cooling includes Liebert XDV overhead systems, as well as under-floor chilled air
- ~9000 s.f. data center at UAF has traditional hot/cold-aisle cooling, conditioned power, etc.

Midnight Network in Detail (next 3 slides)

- Infiniband (4X SDR): two Voltaire 9288 288-port switches. For node communication & data
- GigE: Multiple Cisco switches for private network
- 100MB Ethernet: For management
- FC/AL: to $ARCHIVE SAM/QFS


Networks
FCAL
IB
GigE
10/100
Power
Rack 1
mg09
mg08
mg07
mg06
mg05
mg04
mg03
mg02
mg01
MPS
Rack 2
mg18
mg17
mg16
mg15
mg14
mg13
mg12
mg11
mg10
MPS
Rack 3
mg27
mg26
mg25
mg24
mg23
mg22
mg21
mg20
mg19
MPS
Rack 4
mg36
mg35
mg34
mg33
mg32
mg31
mg30
mg29
mg28
MPS
Rack 5
mg45
mg44
mg43
mg42
mg41
mg40
mg39
mg38
mg37
MPS


ARSC Private
172.16 Net
ARSC Public
199.165 Net /
DREN
Rack 6
Rack 7
Rack 8
Rack 9
Rack 10
Rack 11
mg54
mg53
mg52
mg51
mg50
mg49
mg48
mg47
mg46
Cisco2950
Cisco4948
Cisco2950
Cisco4948
Cisco2950
Cisco2950
Cisco4948
Cisco4848
mg55
mg56/
midnight1
KVM
mn1sm
mpbs1
Voltaire
ISR 9288
mnfs1
mmds2
mmds1
SE3510
mlustre6
mlustre5
mlustre4
mlustre3
mlustre2
mlustre1
mt034
mt033
mt032
mt031
mt030
mt029
mt028
mt027
mt026
mt025
mt024
mt023
mt022
mt021
mt020
mt019
mt018
mt017
mt016
mt015
mt014
mt013
mt012
mt011
mt010
mt009
mt008
mt007
mt006
mt005
mg004
mt003
mt002
mt001
Cisco2950
Cisco4948
mt068
mt067
mt066
mt065
mt064
mt063
mt062
mt061
mt060
mt059
mt058
mt057
mt056
mt055
mt054
mt053
mt052
mt051
mt050
mt049
mt048
mt047
mt046
mt045
mt044
mt043
mt042
mt041
mt040
mt039
mt038
mt037
mt036
mt035
Cisco2950
Cisco4948
mt102
mt101
mt100
mt099
mt098
mt097
mt096
mt095
mt094
mt093
mt092
mt091
mt090
mt089
mt088
mt087
mt086
mt085
mt084
mt083
mt082
mt081
mt080
mt079
mt078
mt077
mt076
mt075
mt074
mt073
mt072
mt071
mt070
mt069
Cisco2950
Cisco4948
MPS
MPS
MPS
MPS
MPS
MPS


Rack 12
Rack 13
Rack 14
Rack 15
Rack 16
Cisco2950
Cisco4948
mt120
mt119
mt118
mt117
mt116
mt115
mt114
mt113
mt112
mt111
mt110
mt109
mt108
mt107
mt106
mt105
mt104
mt103
mg57/
midnight2
Voltaire
ISR 9288
MPS
mt154
mt153
mt152
mt151
mt150
mt149
mt148
mt147
mt146
mt145
mt144
mt143
mt142
mt141
mt140
mt139
mt138
mt137
mt136
mt135
mt134
mt133
mt132
mt131
mt130
mt129
mt128
mt127
mt126
mt125
mt124
mt123
mt122
mt121
Cisco2950
Cisco4948
MPS
mt188
mt187
mt186
mt185
mt184
mt183
mt182
mt181
mt180
mt179
mt178
mt177
mt176
mt175
mt174
mt173
mt172
mt171
mt170
mt169
mt168
mt167
mt166
mt165
mt164
mt163
mt162
mt161
mt160
mt159
mt158
mt157
mt156
mt155
Cisco2950
Cisco4948
MPS
mt222
mt221
mt220
mt219
mt218
mt217
mt216
mt215
mt214
mt213
mt212
mt211
mt210
mt209
mt208
mt207
mt206
mt205
mt204
mt203
mt202
mt201
mt200
mt199
mt198
mt197
mt196
mt195
mt194
mt193
mt192
mt191
mt190
mt189
Cisco2950
Cisco4948
MPS
mt256
mt255
mt254
mt253
mt252
mt251
mt250
mt249
mt248
mt247
mt246
mt245
mt244
mt243
mt242
mt241
mt240
mt239
mt238
mt237
mt236
mt235
mt234
mt233
mt232
mt231
mt230
mt229
mt228
mt227
mt226
mt225
mt224
mt223
Cisco2950
Cisco4948
MPS

Presentation Outline

- About ARSC
- System overview

✓ **Focus on nodes**

- Focus on storage
- Applications, research
- Recommendations, findings

The Larger Compute Nodes

- **X4600, with 8 dual-core processors (16 cores) and 64GB total memory**
  - Opteron 2.6Ghz, 64 bit
  - Hypertransport bus
  - 8GB memory “local” to each socket, 64GB memory total
  - Not SMP because HT is not symmetric to all memory addresses, but can be used as SMP
- **55 nodes, all on the same Infiniband switch**
  - Total of 880 cores

The Smaller Compute Nodes

- **X2200M2, with 2 dual-core processors (4 cores) and 16GB total memory**
  - Same processor speed as X4600 nodes
  - Hypertransport yields fewer “hops” to memory than 16-core system
- **256 nodes on one Infiniband switch**
- **102 nodes on the same switch as the X4600 nodes**
  - Total of 1430 cores

Regular multinode usage does not span all node types

- **Midnight jobs typically run on only one of the three main groups of nodes:**
  - X4600 (16 processor cores each). 55 nodes on one Infiniband switch
  - X2200M2 (4 processor cores each). 256 nodes on another Infiniband switch
  - X2200M2 (4 processor cores each). 102 nodes on the same switch as the X4600 nodes
- **Queuing policy structures this usage**
- **Special jobs can run across different groups of nodes, but do not have a full IB interconnect**

The Operating System, Libraries and Compilers

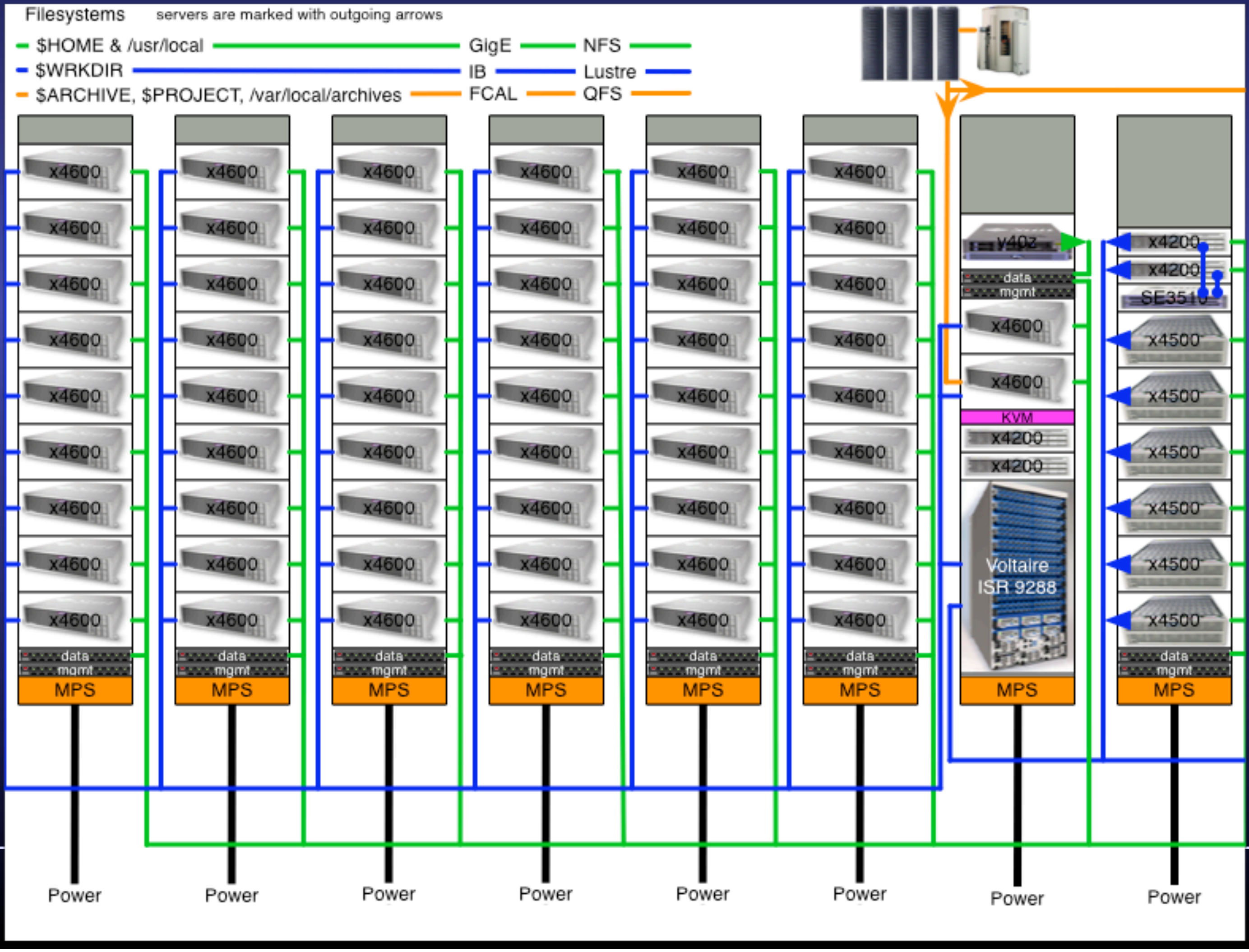
- **SuSE 9.3 Linux, with some modifications & localization (RedHat on X4500 nodes)**
- **Full suite of standard Linux tools**
- **Voltaire-provided MPI stack**
- **PBSPro workload manager**
- **Compilers:**
  - Pathscale: initial choice
  - PGI: recently added due to more community support
  - GCC: available (version 3.x)
  - SunStudio: available, good compatibility

Presentation Outline

- About ARSC
- System overview
- Focus on nodes
- ✓ **Focus on storage**
- Applications, research
- Recommendations, findings

Midnight Storage Network

(X4600 half; X2200 half is similar)



High Performance Cluster Storage

- **Six X4500 nodes forming a Lustre filesystem**
- **Each X4500 has 48 SATA disk drives with 500GB, yielding 24TB aggregate raw storage (144TB total)**
- **After RAID and Lustre overhead, about 68TB are available, split as follows:**
  - $WORKDIR: short-term shared scratch; purged
  - $DATADIR: long-term shared scratch; not purged. By request
  - Neither is backed up; quota is forthcoming (there is a disk hog monitor, currently)
- **Operates over the Infiniband network, with one IB card per node (4X SDR, PCI-X)**

Lustre over Infiniband

- **We are currently running Lustre 1.4.11 with Voltaire IB.**
- **We have a total of eight servers supporting our Lustre installation.**
  - Six provide the block storage and one provides the metadata storage with a warm standby.
  - Each OSS is composed of an x4500, running software RAID 5 with external ldiskfs journaling, on top of 48 500GB SATA disks on six SATA controllers.
  - We can lose a controller and/or several disks and continue to run.
  - Each OSS provides a total of 12TB of storage for a collective total of 72TB of storage.
  - Each MDS is configured with Fibre Channel connectivity back to a 3500 series Sun storage solution running RAID 1 that provides 365GB of metadata storage.
- **Sustained empty filesystem performance reaches 4.8GB/second. Sustained ~20% full filesystem performance is at 4.5GB/second.**

Use of NFS

- Midnight includes an NFS server machine (Sun v40z). This machine exports all $HOME file systems and all of the /usr/local file systems except to the central management node.
- NFS imports also are used for ARSC support file systems on the service and login nodes. These files reside on remote servers and are for administrative use only.

Shared QFS

- **SAM/QFS version 4.5.42**
- **Shared QFS**
  - We use the Linux Shared QFS client on login nodes for access to the $ARCHIVE and $PROJECTS filesystems
  - These filesystems reside on ARSC's SunFire 6800 systems (Solaris), tied to the StorageTek SL8500 tape silo using T10000 and 9940 tapes/drives
  - FC/AL interface
  - Re-exported via NFS to cluster compute nodes

Other temporary storage

- /tmp, /var/tmp: users should avoid
- $SCRATCH defined to /scratch on each node
  - X4600: about 64GB/node
  - X2200: about 216GB/node
- All scratch is shared on the node; purged at the end of a job

Presentation Outline

- About ARSC
- System overview
- Focus on nodes
- Focus on storage

✓ Applications, research

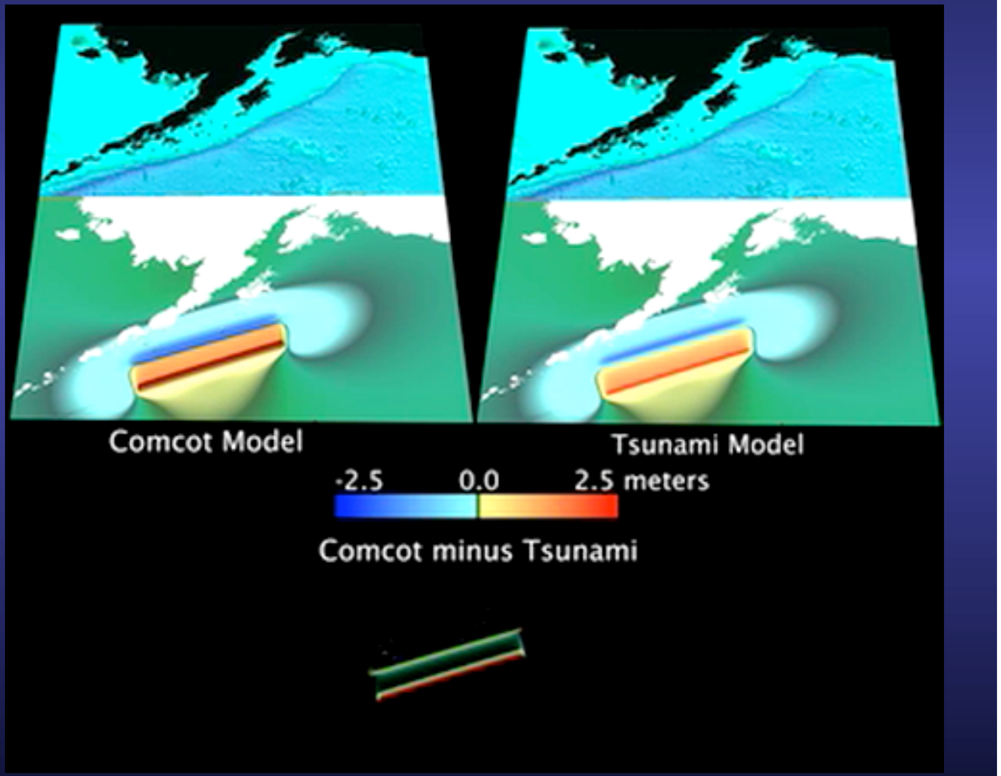
- Recommendations, findings

Research at ARSC

- **Open systems research (no classified or sensitive data)**
- **Deep expertise in geophysical models: climate, oceanography, weather, space physics, permafrost, etc.**
  - Ph.D. staff and other experts
  - Ties to UAF faculty & research groups
  - Mix of local, national & international users

Research at ARSC

- **Applications include:**
  - Locally-originated codes
  - Community codes
  - Analysis software (SAS…)
  - Packaged software (MATLAB, Fluent…)
- **Performance tuning & analysis**
- **Next-generation technology evaluation (acceleration technologies, compilers, systems)**



Presentation Outline

- About ARSC
- System overview
- Focus on nodes
- Focus on storage
- Applications, research
- **Recommendations, findings**
  - ✓Early & ongoing issues

Issues: Early

- **The ARSC testing procedures, derived from the HPCMP’s acquisition process, identified two critical problems. Each was difficult to diagnose (many partners, many hours) and required very significant fixes.**
- **Both problems already existed on other HPC systems, but had not been identified with sufficient clarity.**
  - Partner vendors, not Sun, were where the problems originated

Issues: Current

- **Sun support is not adequately integrated or aggressive for a large HPC system. Quoting an ARSC technical staff member,**
  - HPC requires integrated hardware and software support […] failures of any piece of the cluster can seldom be viewed in isolation. The cost of the failure of a single node is multiplied by the length of the production runs and the number of nodes involved when that one fails. Extended periods of downtime for diagnosis to correct system failures needs to be performed outside of the production environment and performed aggressively.

Presentation Outline

- About ARSC
- System overview
- Focus on nodes
- Focus on storage
- Applications, research

✓ Recommendations, findings

For the 2007 Sun HPC Forum: Recommendations

- **For supercomputing centers:**
  - Commodity clusters are not turnkey solutions
    - Allow for additional time for deployment, testing
    - Interacting with numerous vendors can be difficult. A strong "primary" vendor is important
  - Staff must be prepared to be deeply involved in system design
  - A detailed evaluation plan for a new system is immensely valuable

For the 2007 Sun HPC Forum: Recommendations

- **For potential customers**
  - Sun’s primary OS is Solaris. Detailed Linux support is not as deep
  - End-to-end system support is needed
  - Learn to love your support team, and figure out who is who
  - Ask Sun for your “off-menu” ideas: Sun listens!

For the 2007 Sun HPC Forum: Recommendations

- **For Sun**
  - HPC systems still need more visibility, a clearer identity within Sun
  - Develop more in-house expertise for HPC benchmarking, application tuning, etc.
  - Linux, Linux, Linux

In closing…

- ARSC's *Midnight* system is doing valuable work for researchers
- Deployment of *Midnight* was a learning process for both Sun & ARSC
- Sun's hardware offerings for HPC are fully-featured and capable. Software and support still has some gaps

See you at SC07!

- **ARSC is part of the HPCMP booth (near Sun's) #528**
- **See ARSC staff at the UAF/ARSC/IPY booth #2247**
- **Presentations by ARSC's Greg Newby & Ed Kornkven concerning Midnight:**
  - 2:00 Tuesday: Multicore performance (Sun's booth)
  - 3:30 Tuesday: Midnight overview (HPCMP booth)
  - 4:00 Tuesday: Multicore performance (HPCMP booth)