



CISL Update Operations and Services

CISL HPC Advisory Panel Meeting 03 May 2012

**Anke Kamrath
anke@ucar.edu
Operations and Services Division
Computational and Information Systems Laboratory**

A lot happening in the last 6 months...

- **Updates:**

- *Staffing*
- *NWSC*
- *Archive*
- *RDA*
- *USS*
- *New Supercomputer: Yellowstone*

OSD Staff Comings and Goings...

- **New Staff at ML**

- *Dan Nagle HPC/CSG Consultant*
 - Current chair of the US Fortran standards committee (PL22 and PL22.3)
 - NCAR rejoined InterNational Committee for Information Technology Standards (INCITS)
- *Wes Jeanette – Student Assistant with the VAPOR Team*
- *Kevin Marnross – EaSM DSS hire.*
 - From National Severe Storm Lab
- *Matthew Pasiewiecz*
 - Web Software Engineer – GIS Background

- **Departures**

- *Rob Perry – moved to GLOBE*
- *Alyson Ellerin – retirement*
- *Rocky Madden*
- *Lewis Johnson*

- **Openings**

- *1 Mechanical Technician (at NWSC)*
- *SSG Position (at NWSC)*
- *Temp Admin Support (at NWSC)*



NWSC Staff

• Relocations

- *Rebecca Bouchard (CPG → CASG)*
- *Raisa Leifer (CPG → CASG)*
- *Jirina Kokes (CPG → CASG)*
- *John Ellis (SSG)*
- *Jim VanDyke (NETS)*

• New Hires

- *Jenny Brennan – NWSC Admin*
 - Executive assistant Toyota subsidiary Detroit Michigan
- *Matthew Herring – NWSC Electrician*
 - Encore Electric subcontractor for NWSC
- *Lee Myers – NWSC CASG*
 - Computer Science Instructor - Eastern Wyoming College
- *Jonathan Roberts – NWSC CASG*
 - IBM Boulder
- *Jonathan Frazier – NWSC CASG*
 - DOD subcontractor
 - Served United States Air Force (Cheyenne)
- *David Read – NWSC CASG*
 - LOWES data center systems administration
 - LOWES distribution center (Cheyenne)



And here they are!



NWSC Update

- **So much has happened it is difficult to summarize**
- **Tape Archive Acceptance**
- **Network Complete to NWSC**
 - *LAN, WAN and Wireless*
- **Transition of NOC to TOC**
- **NWSC Facilities Management**
 - *First winter season*
 - *Fitup and transition of building*
- **HPSS Equipment Install and Preparation for Production cut-over**
- **Yellowstone Planning**
- **Electrical & Mechanical Install for Yellowstone**



Tape Archive Install



NWSC Oversight Committee becomes Transition Oversight Committee

- **Transition Oversight Committee (TOC)
Held December 15th**
 - *Smooth handoff from NOC (December 14th)*
 - *Three New Members added*
 - Thomas Hauser (CU)
 - Francesca Verdier (NERSC)
 - Henning Weber (DWD)
 - *Next Meeting Summer of 2012*
- **Exit Briefing (next slide)**

TOC Exit Briefing

- **General Comments**

- *"Nothing critical to say."*
- *"Presentations were great. Impressed with technical depth of staff."*
- *"Can't imagine how they thought of so much."*

- **General Recommendations**

- *Write whitepaper on NWSC design innovations to share with community (Aaron/Gary)*
- *Develop NWSC Space Strategy for potential partners*
 - Under development already (Anke, Aaron, Rich)
 - Separate NWSC Fiber Hut agreement in development
- *Consider having users involved during acceptance testing*
 - 21 day acceptance period. Bluefire/Yellowstone overlap short (~6 weeks)
- *Next meeting (August)*
 - High-level update on all topics
 - Overview of (increased) operational costs for NWSC over Mesa Lab
 - EOT Update (visitor center)
 - Grand Opening Planning Update

Lessons Learned (the expected)

- **Seasonal issues that affect the building**
 - *Wind*
 - *Weather*
 - *Cold*
- **Air sampling sensor froze causing a refrigerant alarm**
- **Air dryers for the pneumatic valves not installed correctly.**
 - *Condensation froze in the line*
- **Boilers went out on high winds**



Lessons Learned (the unexpected)

- **A ground fault within the utility (CLF&P) caused a power bump at NWSC**

- *Pumps in the business park surged*
- *Over pressurized the fire line*
- *Flow sensor identified water was flowing to computer room*
- *Triggered the EPO*
 - Should do this to protect electronics if water is flowing



- **Some additional fail safes required or pressure reduction**

- **Identified that not all new staff had all the right information**

- *Cell phone numbers*
- *Emergency procedures*
- *Where devices are located & how to respond*



Yellowstone Fitup - Electrical



Yellowstone Fitup - Electrical



Yellowstone Fitup - Mechanical



Yellowstone Fitup - Mechanical



Key Resource Transition Efforts

- **LAN/WAN Network Build-out**
 - **NEW** install at NWSC
 - **KEEP** current install at ML
- **Infrastructure Services (e.g., DNS, Nagios, DHCP)**
 - **NEW** install at NWSC
 - **KEEP** current install at ML
 - **MOVE** Data Collection Servers
- **Archival Services**
 - **NEW** install at NWSC
 - **MOVE** primary HPSS site from ML to NWSC
 - **MOVE** data from ML to NWSC
 - **KEEP** HPSS Disaster Recovery capability at ML
 - **DECOMISSION** Type B Tape Drives/Tapes at ML

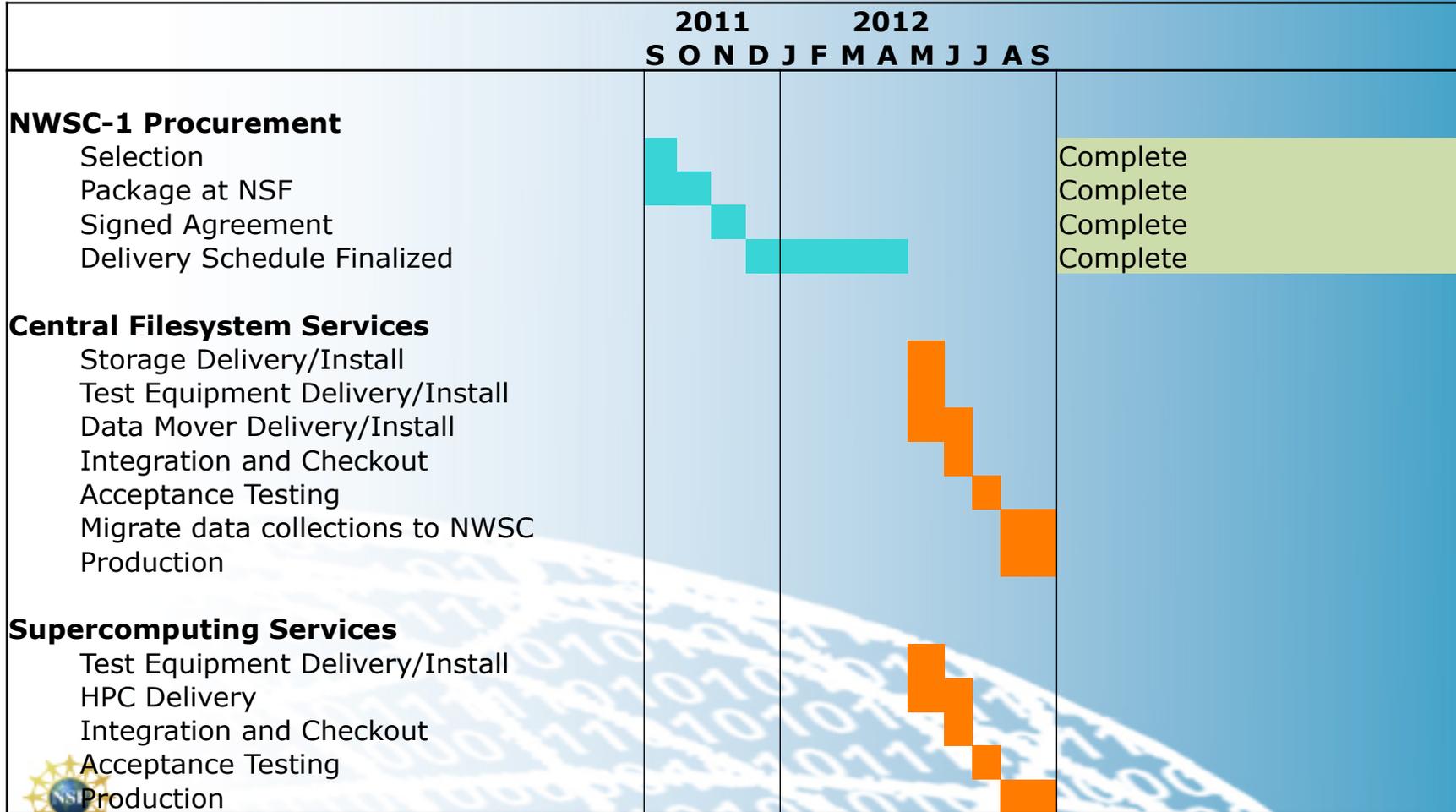
Resource Transition Efforts (cont)

- **Central Data Services**
 - **NEW** GLADE install at NWSC
 - **MOVE** (some) data (and possibly storage) from ML to NWSC
 - **DECOMMISSION** GLADE at ML
- **Supercomputing Services**
 - **NEW** install (Yellowstone) at NWSC
 - **DECOMMISSION** Bluefire at ML
 - **KEEP** Lynx at ML

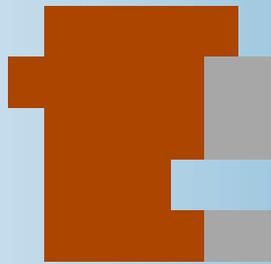
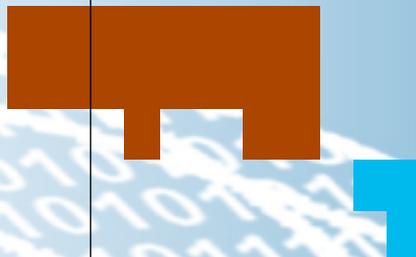
General Preparation Schedule

	2011				2012								
	S	O	N	D	J	F	M	A		M	J	J	A
HPC Fit Up													
Design and Pricing													Complete
Package at NSF													Complete
Equipment Ordering Installation													Complete
Staffing													
Onsite Management Team													Gary New, Ingmar T., Jim Van Dyke
Additional Elec/Mech													Complete
Admin/2 Sys Admins													Complete
2 Software Engineers													1 Existing, 1 New
Additional System Admins													Complete
Networking													
LAN													Complete
Wireless LAN													Complete
WAN N. Activation													Complete
WAN S. Activation													Complete
Data Center Switch Install													Complete
Enterprise Services													
Infrastructure Installation													Complete
Nagios Development													Training of SAI now monitoring
Accounting System Development and Deployment													Boulder from Cheyenne
													Mirroring Bluefire Accounting Now
Archival Services													
Tape Library Delivery/Install													Complete
Tape Archive Acceptance Test													Completed Nov 30
HPSS Servers/DataMovers/Cache													Complete
HPSS Primary at NWSC													1 day outage for cutover - May 23
Data Migration													Complete in 2014

Procurement & Installation Schedule



User Preparations & Allocations Schedule

	2011	2012											
	S	O	N	D	J	F	M	A	M	J	J	A	S
<p>Allocations</p> <ul style="list-style-type: none"> ASD Submission, Review, User Prep CSL Submission, Review, User Prep University Submission, Review Wyoming-NCAR Submission, Review NCAR Submission, Review, User Prep 													<ul style="list-style-type: none"> Due Mar 5, 2012 CSLAP Meet Mar 21-23 CHAP Meet May 3 WRAP Meet May 3-4
<p>User Preparation</p> <ul style="list-style-type: none"> Presentations (NCAR, SC, AGU, AMS, Wyoming) User survey, environment, documentation Training ASD and Early Use Production Use 											<ul style="list-style-type: none"> Weeklong Training Workshops Until mid-2016 		

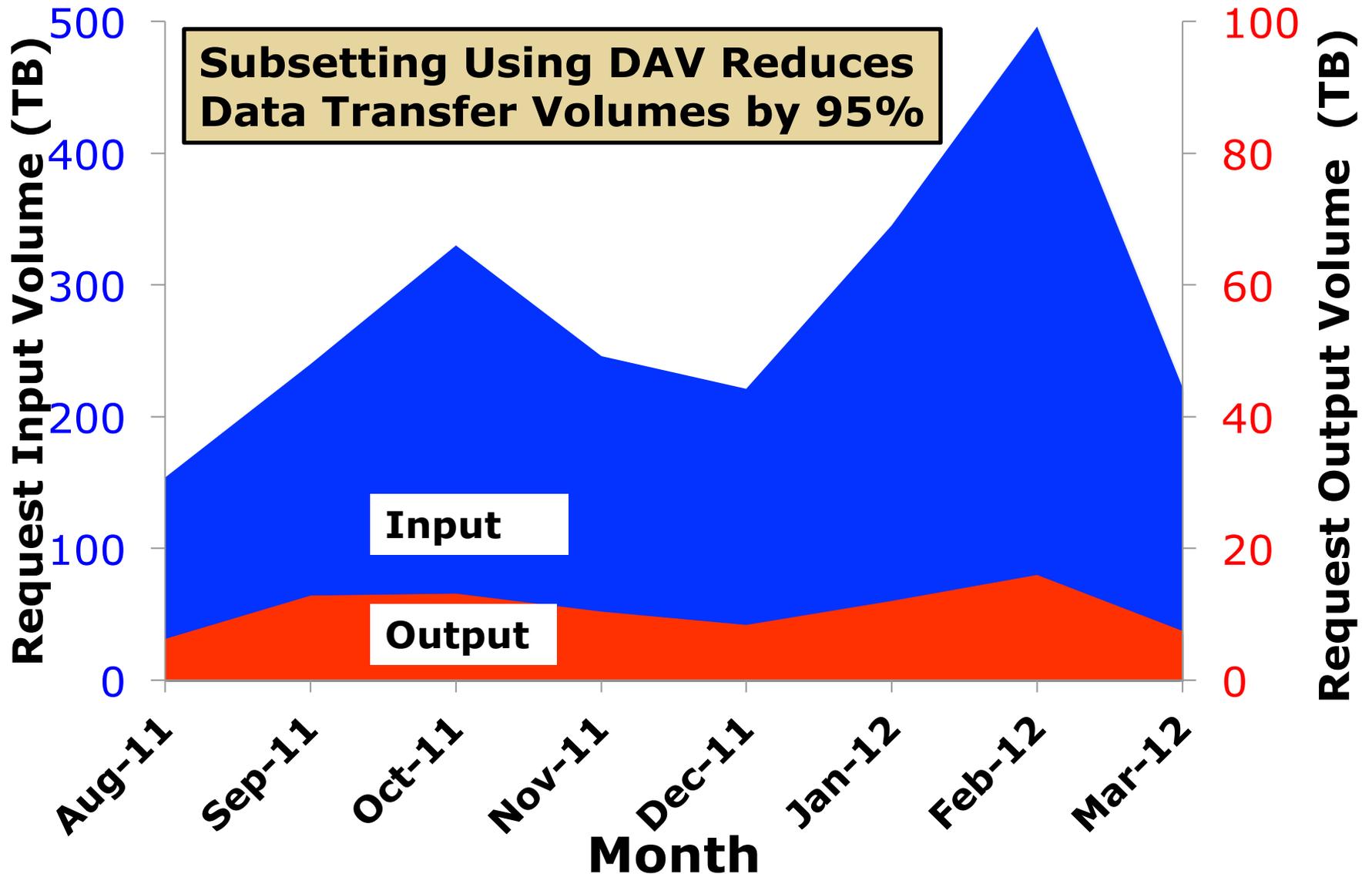
HPSS and AMSTAR Update

- **AMSTAR extended 2 years through December 2014.**
 - *5 TB per cartridge technology*
 - *30 PB capacity increase over 2 years*
 - *New tape libraries, drives, and media installed at NWSC in November 2011 for primary and second copies*
 - *New tape drives and media installed at ML in November 2011 for Disaster Recovery copies of selected data*
- **Will release an archive RFP in January 2013 after we have experience with NWSC-1 data storage needs.**
- **Additional HPSS server capacity delivered to NWSC and is being configured for May deployment.**
- **Production data and metadata will migrate to NWSC after CASG 24/7 support is available at NWSC in May.**

RDA Update

- **EaSM data support portal is now operational**
 - *6-hrly 20th Century Run & Four Future Projections*
- **RDA transition to NWSC plans**
 - *No users access downtime, increase amount of GLADE online data*
 - *Use NWSC DAV to offer more data subsetting services (current status on next slide)*

Monthly Impact of RDA Data Subsetting



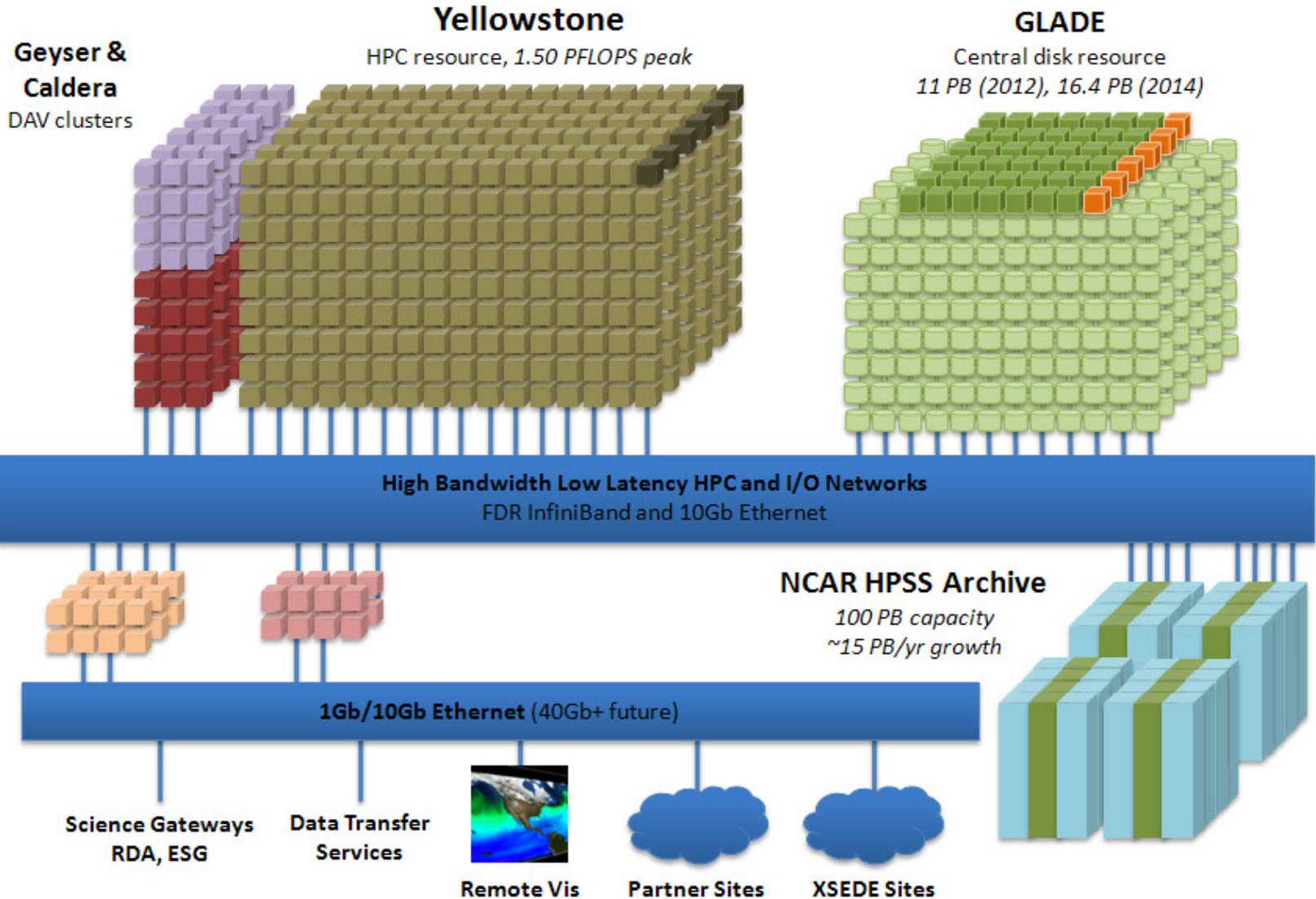
USS Update

- **New Daily Bulletin** rolled out last October (using Drupal).
- **Winter HPC workshop** in February to 13 onsite and 13 online attendees
- **CSG staff organized the first UCAR SEA (Software Engineering Assembly) Conference** in February
- **Globus Online service** deployed in production, now the recommended off-site data transfer method
- **CSG collaborated with NCO and netCDF developers** to resolve severe performance problem
- **Numerous documentation improvements**

NCAR's Data-Centric Supercomputing Environment: Yellowstone



Yellowstone Environment



GLADE: Globally Accessible Data Environment

NWSC Centralized Filesystems & Data Storage Resource

- **GPFS NSD Servers**

- 20 IBM x3650 M4 nodes; Intel Xeon E5-2670[†] processors w/AVX
- 16 cores, 64 GB memory per node; 2.6 GHz clock
- 91.8 GB/sec aggregate I/O bandwidth (4.8+ GB/s/server)

- **I/O Aggregator Servers (export GPFS, CFDS-HPSS connectivity)**

- 4 IBM x3650 M4 nodes; Intel Xeon E5-2670 processors w/AVX
- 16 cores, 64 GB memory per node; 2.6 GHz clock
- 10 Gigabit Ethernet & FDR fabric interfaces

- **High-Performance I/O interconnect to HPC & DAV Resources**

- Mellanox FDR InfiniBand full fat-tree
- 13.6 GB/sec bidirectional bandwidth/node

- **Disk Storage Subsystem**

- 76 IBM DCS3700 controllers & expansion drawers; 90 2 TB NL-SAS drives/controller initially; add 30 3 TB NL-SAS drives/controller 1Q2014
- 10.94 PB usable capacity (initially)
- 16.42 PB usable capacity (1Q2014)



Yellowstone

NWSC High Performance Computing Resource

• Batch Computation Nodes

- 4,518 IBM dx360 M4 nodes; Intel Xeon E5-2670[†] processors with AVX
- 16 cores, 32 GB memory, 2.6 GHz clock, 333 GFLOPs per node
- 4,518 nodes, 72,288 cores total - 1.504 PFLOPs peak
- 144.6 TB total memory
- 28.9 bluefire equivalents

• High-Performance Interconnect

- Mellanox FDR InfiniBand full fat-tree
- 13.6 GB/sec bidirectional bw/node
- <2.5 usec latency (worst case)
- 31.7 TB/sec bisection bandwidth



• Login/Interactive Nodes

- 6 IBM x3650 M4 Nodes; Intel Xeon E5-2670 processors with AVX
- 16 cores & 128 GB memory per node



• Service Nodes (LSF, license servers)

- 6 IBM x3650 M4 Nodes; Intel Xeon E5-2670 processors with AVX
- 16 cores & 32 GB memory per node

Geyser and Caldera

NWSC Data Analysis & Visualization Resource

- **Geyser: Large Memory System**
 - 16 IBM x3850 X5 nodes; Intel Westmere-EX processors
 - 40 2.4 GHz cores, 1 TB memory per node
 - 1 nVIDIA Quadro 6000 GPU[‡] per node
 - Mellanox FDR full fat-tree interconnect
- **Caldera: GPU-Computation/Visualization System**
 - 16 IBM dx360 M4 nodes; Intel Xeon E5-2670[†] processors with AVX
 - 16 2.6 GHz cores, 64 GB memory per node
 - 2 nVIDIA Tesla M2070Q GPUs[‡] per node
 - Mellanox FDR full fat-tree interconnect
- **Knights Corner System (November 2012 delivery)**
 - 16 IBM dx360 M4 Nodes; Intel Xeon E5-2670[†] processors with AVX
 - 16 Sandy Bridge cores, 64 GB memory, 2 Knights Corner adapters per node
 - Each Knights Corner contains 62 x86 processors
 - Mellanox FDR full fat-tree interconnect



[†] Codenamed "Sandy Bridge EP"

[‡] The Quadro 6000 and M2070Q are identical ASICs

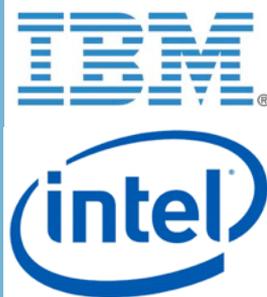
Yellowstone Software

- **Compilers, Libraries, Debugger & Performance Tools**

- *Intel Cluster Studio (Fortran, C++, performance & MPI libraries, trace collector & analyzer) 50 concurrent users*
- *Intel VTune Amplifier XE performance optimizer 2 concurrent users*
- *PGI CDK (Fortran, C, C++, pgdbg debugger, pgprof) 50 concurrent users*
- *PGI CDK GPU Version (Fortran, C, C++, pgdbg debugger, pgprof) for DAV systems only, 2 concurrent users*
- *PathScale EckoPath (Fortran C, C++, PathDB debugger) 20 concurrent users*
- *Rogue Wave TotalView debugger 8192 floating tokens*
- *Rogue Wave ThreadSpotter 10 seats*
- *IBM Parallel Environment (POE), including IBM HPC Toolkit*

- **System Software**

- *LSF-HPC Batch Subsystem / Resource Manager*
 - IBM has acquired Platform Computing, Inc. (developers of LSF-HPC)
- *Red Hat Enterprise Linux (RHEL) Version 6*
- *IBM General Parallel Filesystem (GPFS)*
- *Mellanox Universal Fabric Manager*
- *IBM xCAT cluster administration toolkit*



Erebus

Antarctic Mesoscale Prediction System (AMPS)

- **IBM iDataPlex Compute Cluster**
 - 84 IBM dx360 M4 Nodes; Intel Xeon E5-2670[†] processors
 - 16 cores, 32 GB memory per node; 2.6 GHz clock
 - 1,344 cores total - 28 TFLOPs peak
 - Mellanox FDR-10 InfiniBand full fat-tree interconnect
 - 0.54 bluefire equivalents
- **Login Nodes**
 - 2 IBM x3650 M4 Nodes; Intel Xeon E5-2670 processors
 - 16 cores & 128 GB memory per node
- **Dedicated GPFS filesystem**
 - 2 IBM x3650 M4 GPFS NSD servers
 - 57.6 TB usable disk storage
 - 9.6 GB/sec aggregate I/O bandwidth



[†] Codenamed "Sandy Bridge EP"

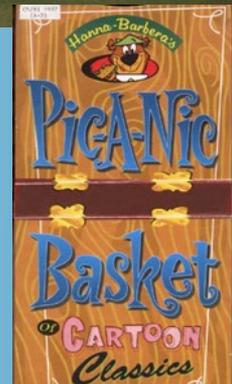


Erebus, on Ross Island, is Antarctica's most famous volcanic peak and is one of the largest volcanoes in the world – within the top 20 in total size and reaching a height of 12,450 feet. Small eruptions have been occurring 6-10 times daily since the mid-1980s.



Yellowstone Environment Test Systems

- **Yellowstone test system: “Jellystone” (installed @ NWSC)**
 - 32 IBM dx360 M4 Nodes; Intel Xeon E5-2670 processors
 - 1 login node, 2 management/service nodes
 - 2 36-port Mellanox FDR IB switches full fat-tree
 - 1 partially-filled iDataPlex rack & part of DAV 19” rack
- **GLADE test system: “Picanicbasket” (installed @ Mesa Lab, move to NWSC during 2H2012 & integrate)**
 - 2 IBM x3650 M4 Nodes; Intel Xeon E5-2670 processors
 - 1 DCS 3700 dual-controller & expansion drawer, ~150 TB storage
 - 1 IBM e3650 management/service node
 - 1 36-port Mellanox FDR IB switches full fat-tree
 - 2 partially-filled 19” racks
- **Geyser and Caldera test system: “Yogi” & “Booboo” (installed @ NWSC)**
 - 1 IBM e3850 large-memory node, 2 IBM dx360 M4 nodes GPU-computation nodes
 - 1 IBM e3650 management/service node
 - 36-port Mellanox FDR IB switch full fat-tree
 - 1 partially-filled 19” rack



Yellowstone Physical Infrastructure

Resource	# Racks
HPC	63 - iDataPlex Racks (72 nodes per rack) 10 - 19" Racks (9 Mellanox FDR core switches, 1 Ethernet switch) 1 - 19" Rack (login, service, management nodes)
CFDS	19 - NSD Server, Controller and Storage Racks 1 - 19" Rack (I/O aggregator nodes, management , IB & Ethernet switches)
DAV	1 - iDataPlex Rack (GPU-Comp & Knights Corner) 2 - 19" Racks (Large Memory, management , IB switch)
AMPS	1 - iDataPlex Rack 1 - 19" Rack (login, IB, NSD, disk & management nodes)

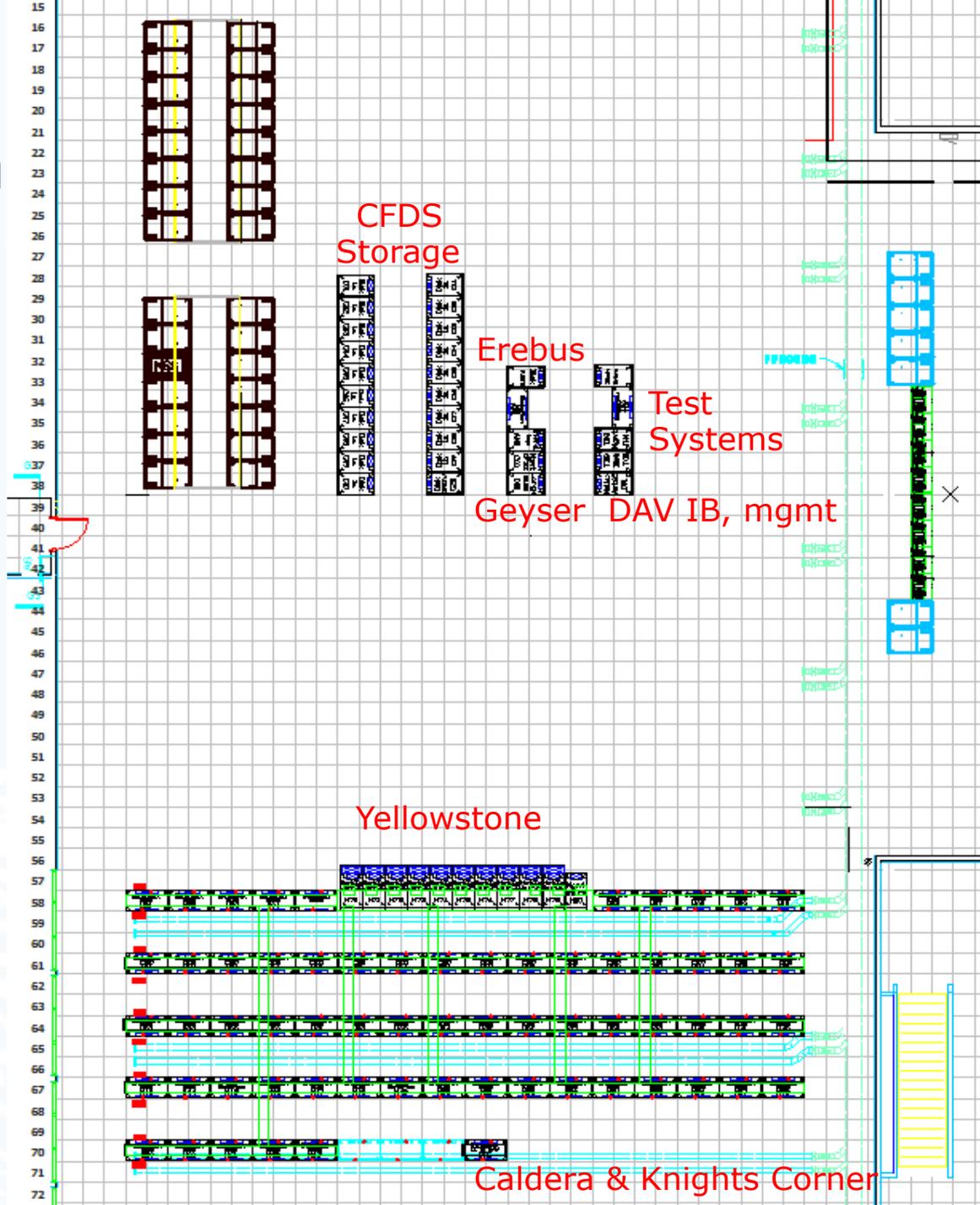
Total Power Required	~2.13 MW
HPC	~1.9 MW
CFDS	0.134 MW
DAV	0.056 MW
AMPS	0.087 MW



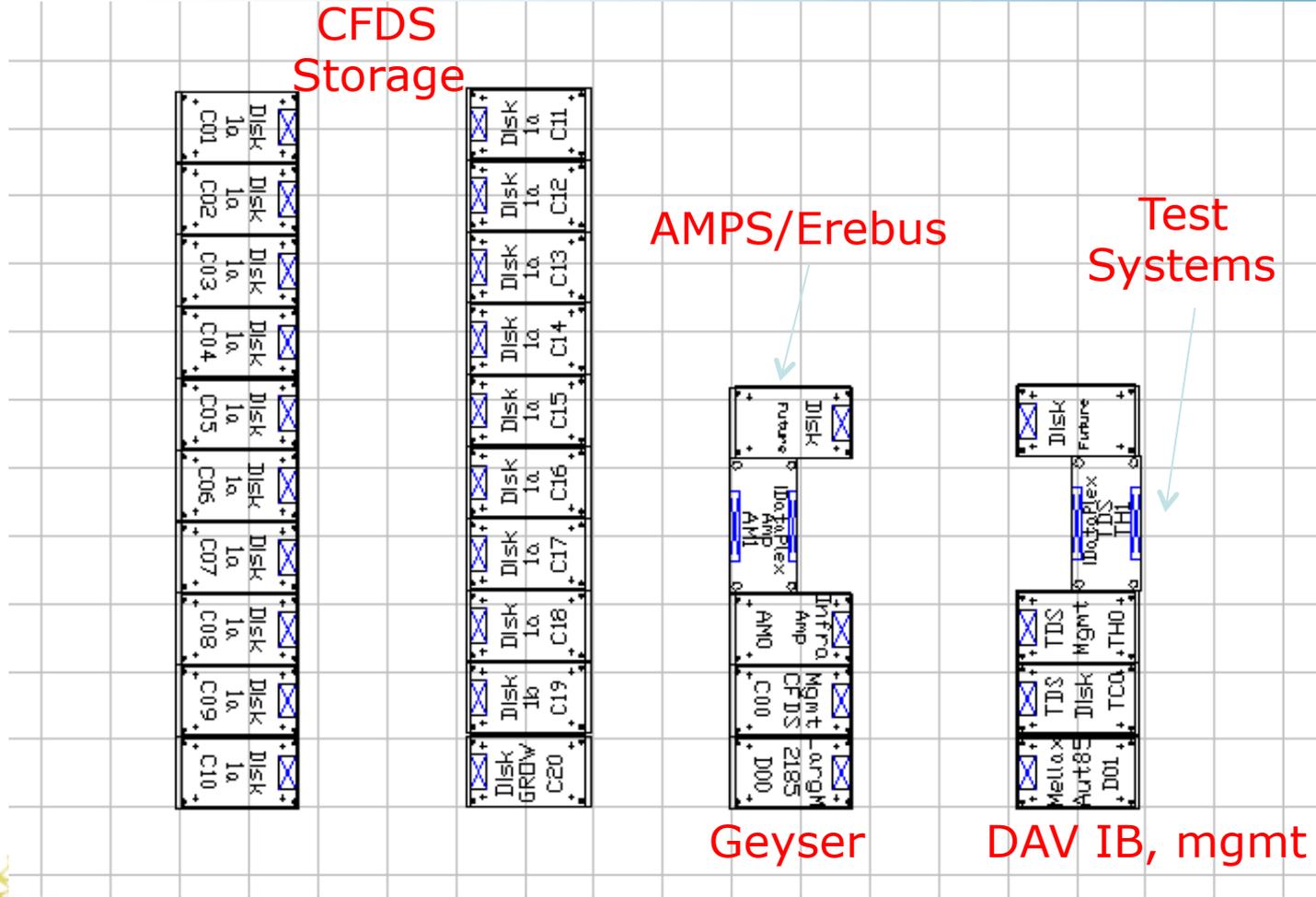


NWSC Floorplan

Viewing Area

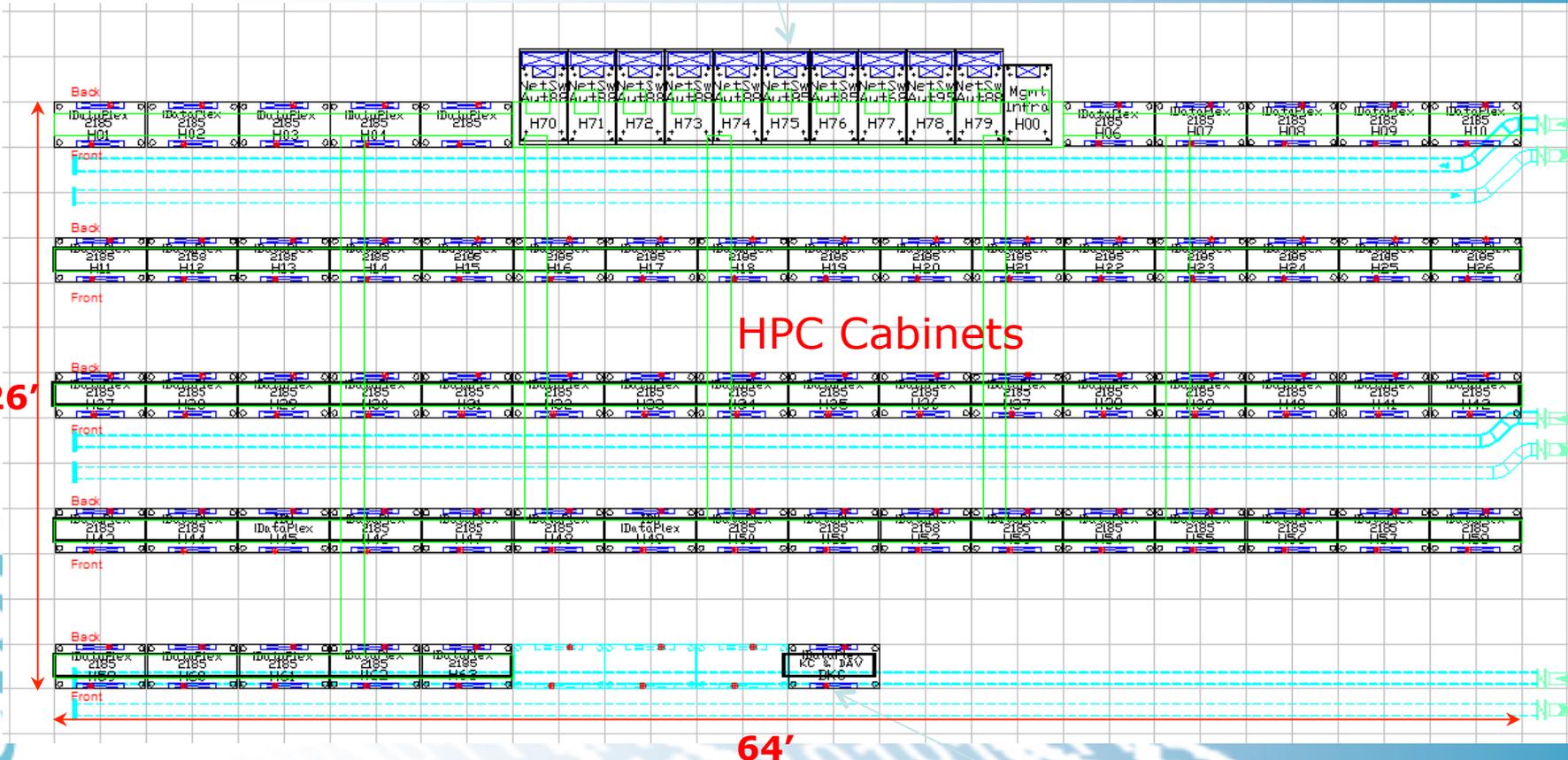


Yellowstone Systems (north half)



Yellowstone Systems (south half)

InfiniBand Switches



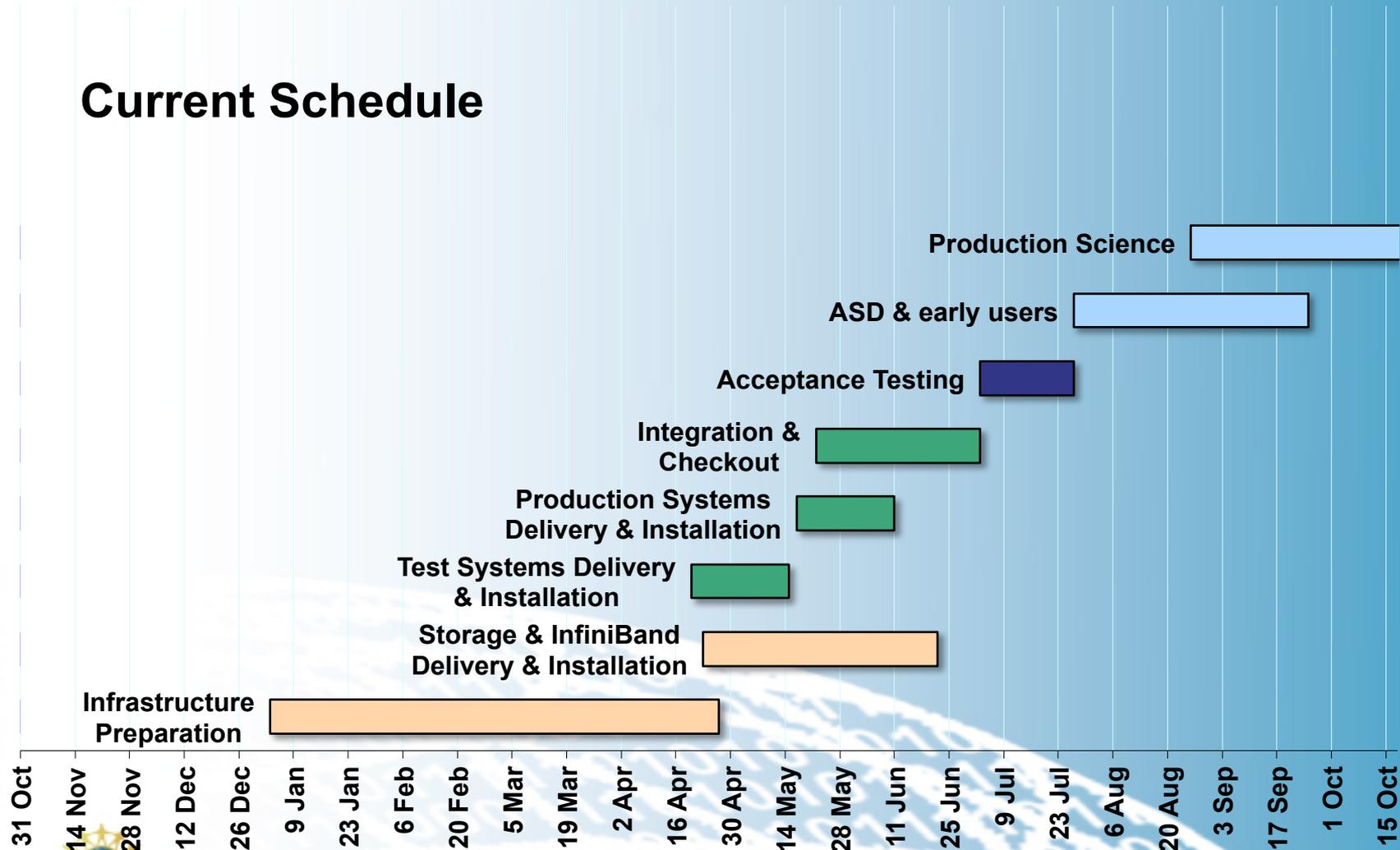
64'

26'

Caldera & Knights Corner

Yellowstone Schedule (100k')

Current Schedule



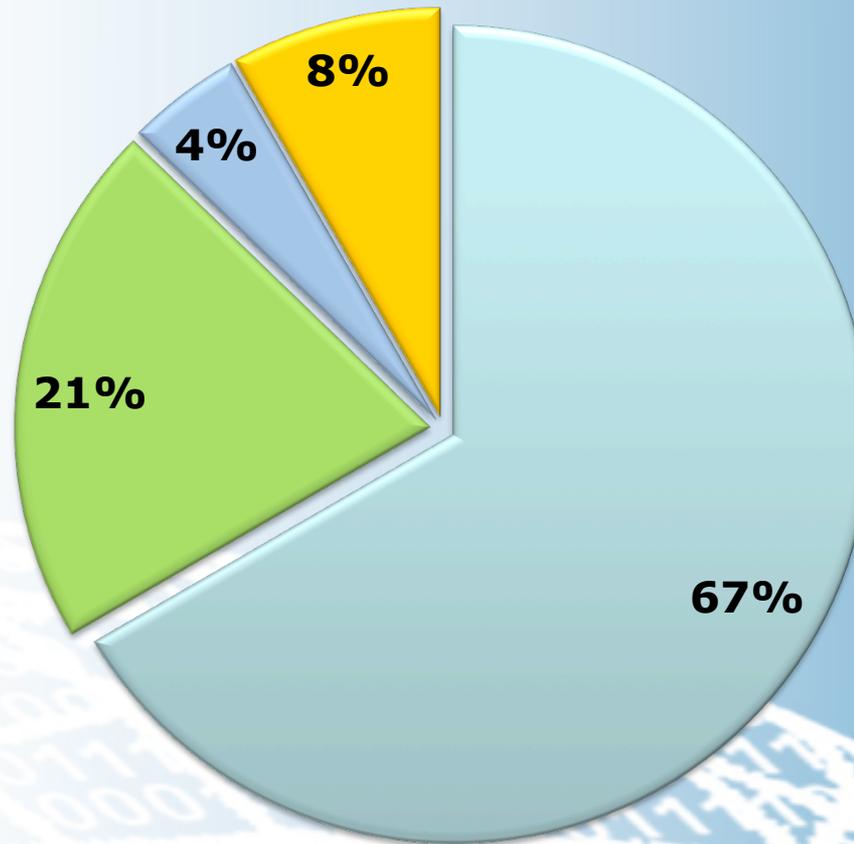
Equipment Delivery Schedule

Tentative schedule provided May 2: IBM still refining

Date		Equipment delivered to NWSC (unless otherwise indicated)
		<i>All dates are tentative - to be confirmed ~May 9</i>
Tue	15 May 2012	Mellanox InfiniBand racks
Thu	24 May 2012	CFDS test rack (deliver to Mesa Lab)
Thu	24 May 2012	HPC & DAV test racks
Tue	29 May 2012	AMPS System
Mon	11 Jun 2012	CFDS storage racks
Tue	19 Jun 2012	HPC management rack (H00) and 10 RDHX's
		Production HPC group 1
		Production HPC group 2
		Production HPC group 3
Fri	29 Jun 2012	All delivered over ~10 day period, finishing on this date
Mon	09 Jul 2012	Remainder of RDHX's

NWSC-1 Resource Allocation by Cost

■ HPC ■ CFDS ■ DAV ■ Maintenance



Questions?