NCAR Workload Analysis on Yellowstone

March 2015

V5.0

CISL Computational & Information Systems Laboratory

NCAR NSF

Purpose and Scope of the Analysis

- Understanding the NCAR application workload is a critical part of making efficient use of Yellowstone and in scoping the requirements of future system procurements.
- Analysis of application performance on Yellowstone is the first step in understanding the transition needed to move to new architectures.
- Primary sources of information for the analysis included:
  - *Science area*
  - *Application code*
  - *3rd party application usage*
  - *Algorithm*
  - *Job size*
  - *Memory usage*
  - *Threading usage*
  - *I/O patterns*

Yellowstone Environment

- **Yellowstone** ***(High-performance computing)***
  - IBM iDataPlex Cluster with Intel ‘Sandy Bridge’ processors
  - 1.5 PetaFLOPs; 4,536 nodes; 72,576 Xeon E5-2670 cores
  - 145 TB total memory
  - Mellanox FDR InfiniBand quasi fat-tree interconnect
- **GLADE** ***(Centralized file systems and data storage)***
  - GPFS file systems, 16.4 PB capacity, >90 GB/s aggregate I/O bandwidth
- **Geyser & Caldera** ***(Data analysis and visualization)***
  - Large-memory system – Geyser:
    16 nodes, 640 Westmere-EX cores, 1 TB/node, 16 NVIDIA K5000 GPUs
  - GPU computation/visualization system – Caldera:
    16 nodes, 256 Xeon E5-2670 cores, 64 GB/node, 32 NVIDIA K20X GPUs
- **Pronghorn** ***(Intel Phi testbed system)***
  - 16 nodes, 256 Xeon E5-2570 cores; 64 GB/node
  - 32 Intel Phi 5110P adapters (Knight’s Corner)
- **Erebus** ***(Antarctic Mesoscale Prediction System, AMPS)***
  - 84 nodes; 1,344 Xeon E5-2570 cores; 32 GB/node; 2 login nodes; 58 TB dedicated GPFS file system capacity, 9.6 GB/s aggregate bandwidth

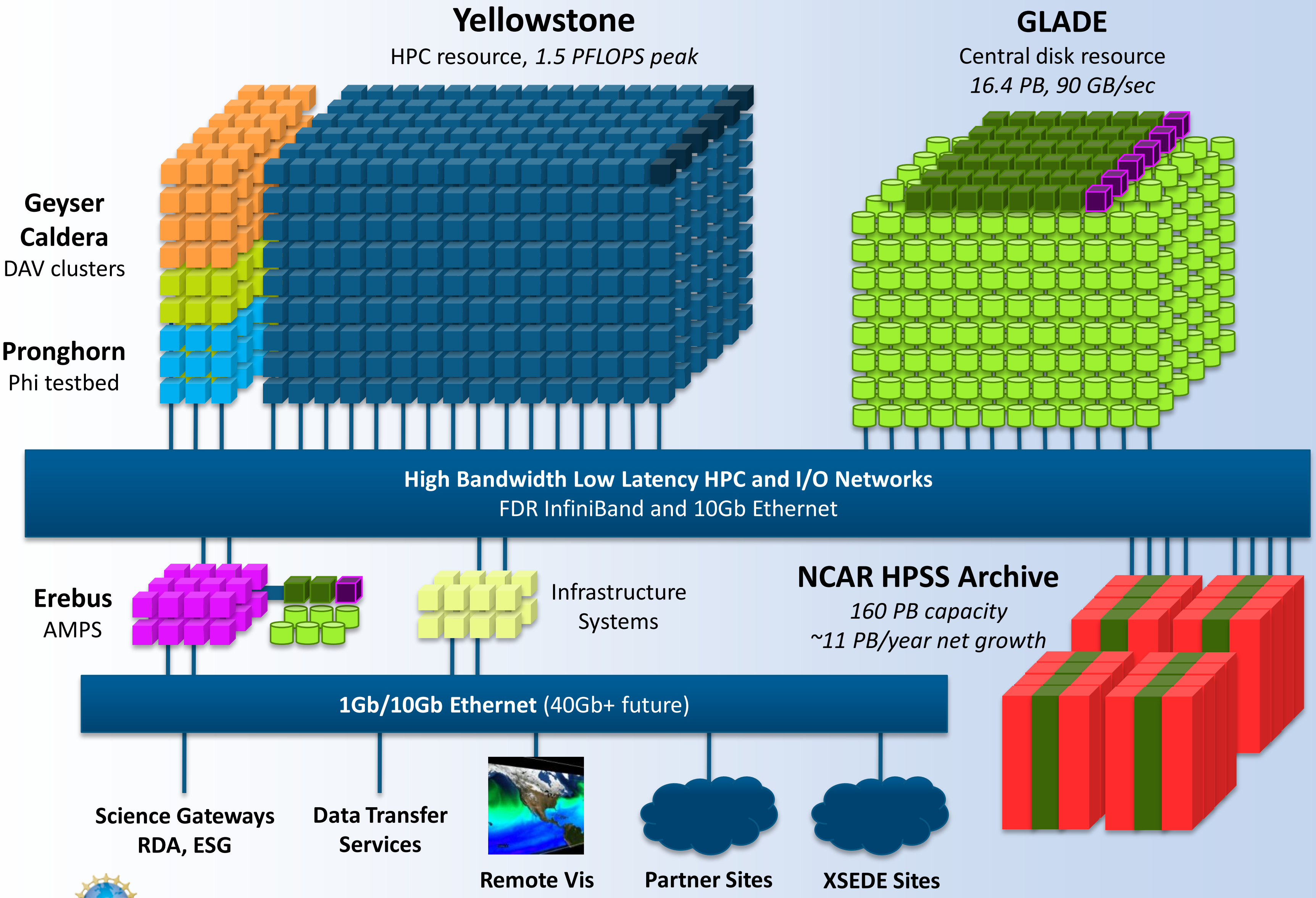
Yellowstone Physical Infrastructure

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)
GLADE	19 - NSD Server, Controller and Storage Racks 1 - 19” Rack (I/O aggregator nodes, management , InfiniBand & Ethernet switches)
DAV	1 - iDataPlex Rack (Caldera & Pronghorn) 2 - 19” Racks (Geyser, management , InfiniBand switch)
AMPS	1 - iDataPlex Rack 1 - 19” Rack (login, InfiniBand, NSD, disk & management nodes)

Total Power Required	~1.7 MW
HPC	~1.4 MW
GLADE	0.134 MW
DAV	0.056 MW
AMPS	0.087 MW

Computational & Information Systems Laboratory

Yellowstone Environment



Yellowstone
HPC resource, *1.5 PFLOPS peak*

GLADE
Central disk resource
16.4 PB, 90 GB/sec

Geyser
Caldera
DAV clusters

Pronghorn
Phi testbed

High Bandwidth Low Latency HPC and I/O Networks
FDR InfiniBand and 10Gb Ethernet

Erebus
AMPS

Infrastructure
Systems

NCAR HPSS Archive
160 PB capacity
~11 PB/year net growth

1Gb/10Gb Ethernet (40Gb+ future)

Science Gateways
RDA, ESG

Data Transfer
Services

Remote Vis

Partner Sites

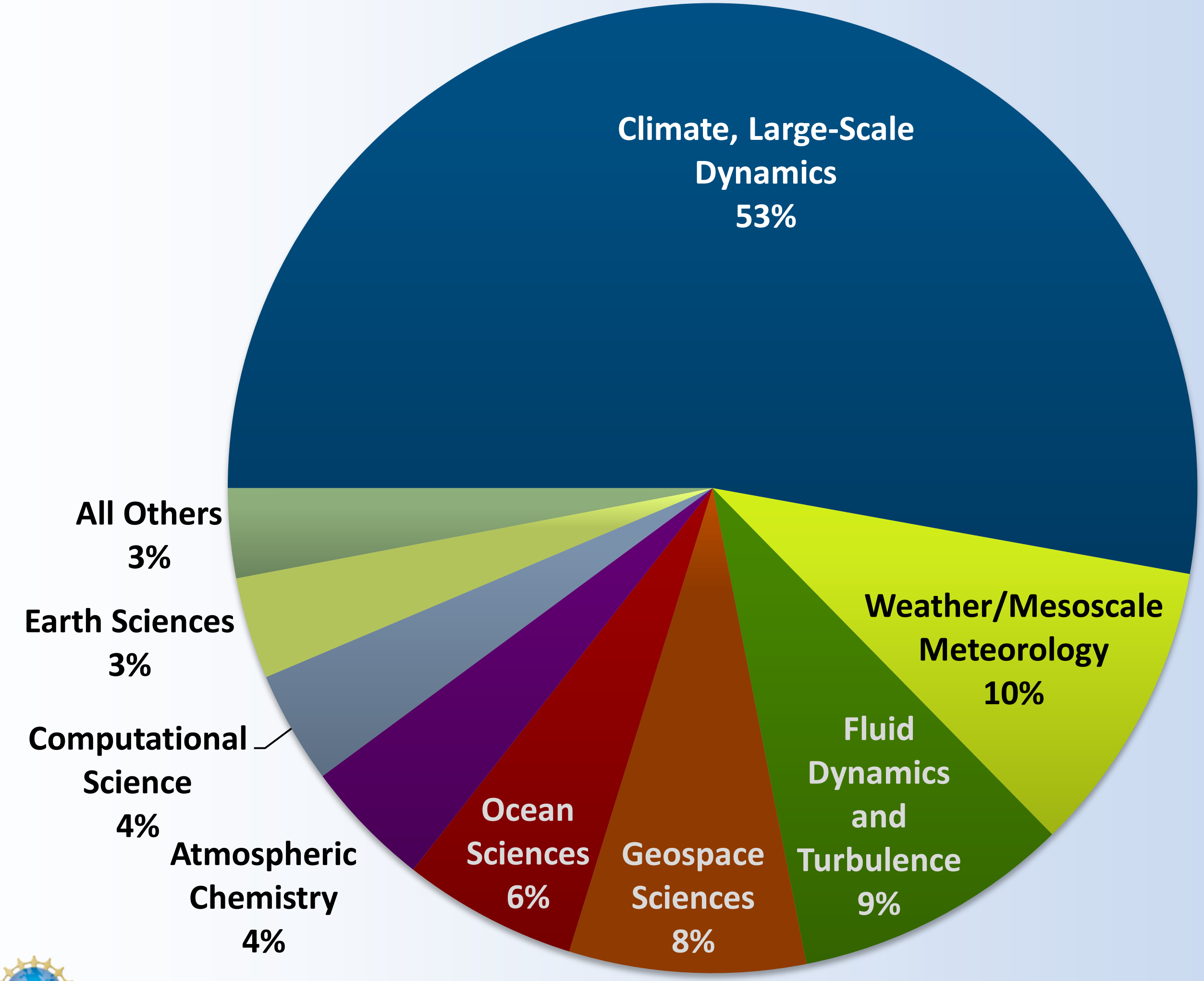
XSEDE Sites

User Communities

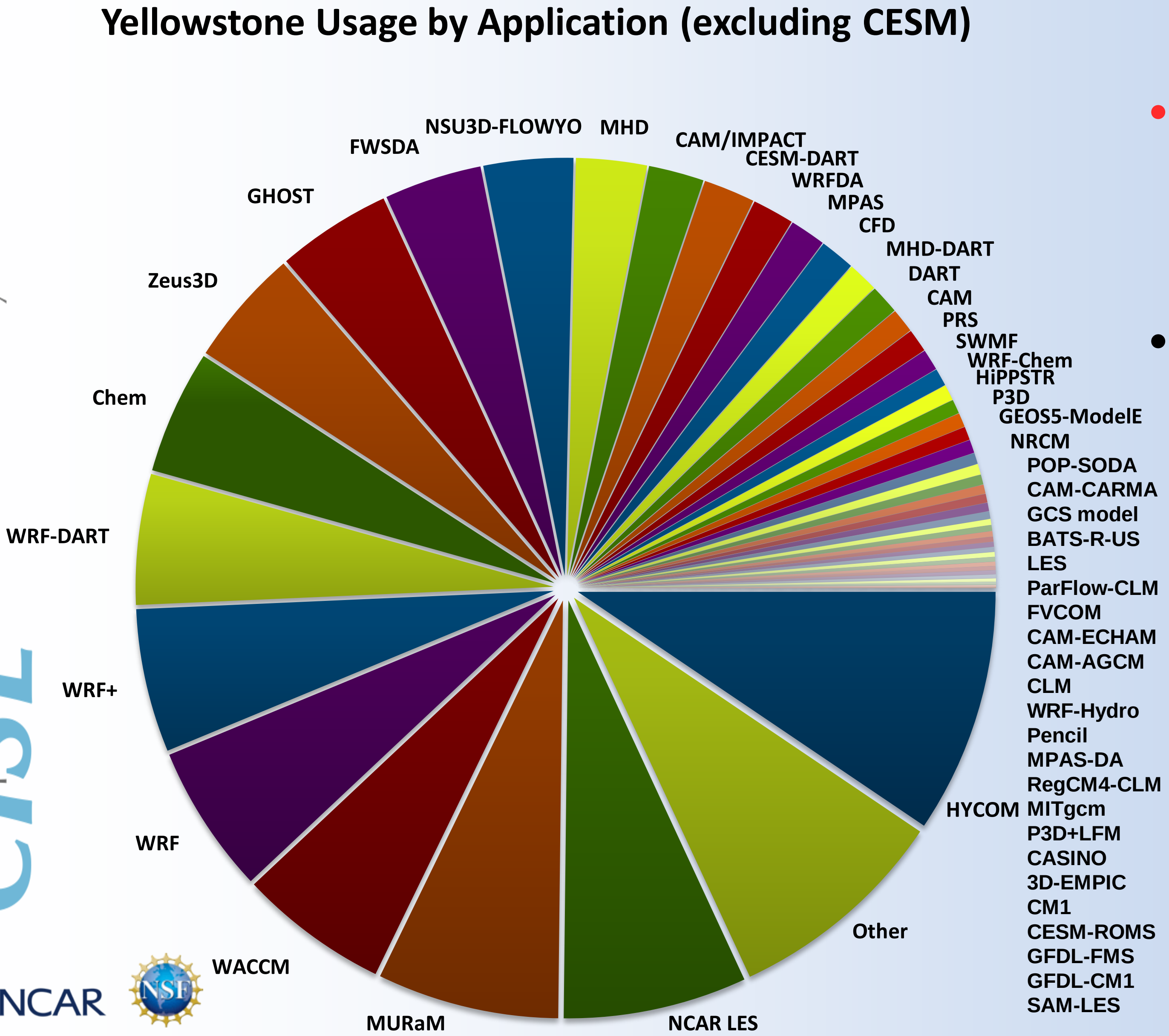
1,160 HPC users in the last 12 months — more than 475 distinct users each month
612 projects in the last 12 months — more than 275 distinct projects each month

- **NCAR staff (29%)**
  - Roughly equal use by CGD, MMM, ACD, HAO, RAL
  - Smaller use by CISL, EOL, other programs
- **University (29%)**
  - Larger number of smaller scale projects
  - Many graduate students, post-docs
- **Climate Simulation Laboratory (28%)**
  - Small number (<6) large-scale climate-focused projects
  - Large portion devoted to CESM community
- **Wyoming researchers (13%)**
  - Smaller number of activities from a broader set of science domains

Yellowstone usage reflects its mission to serve the atmospheric sciences



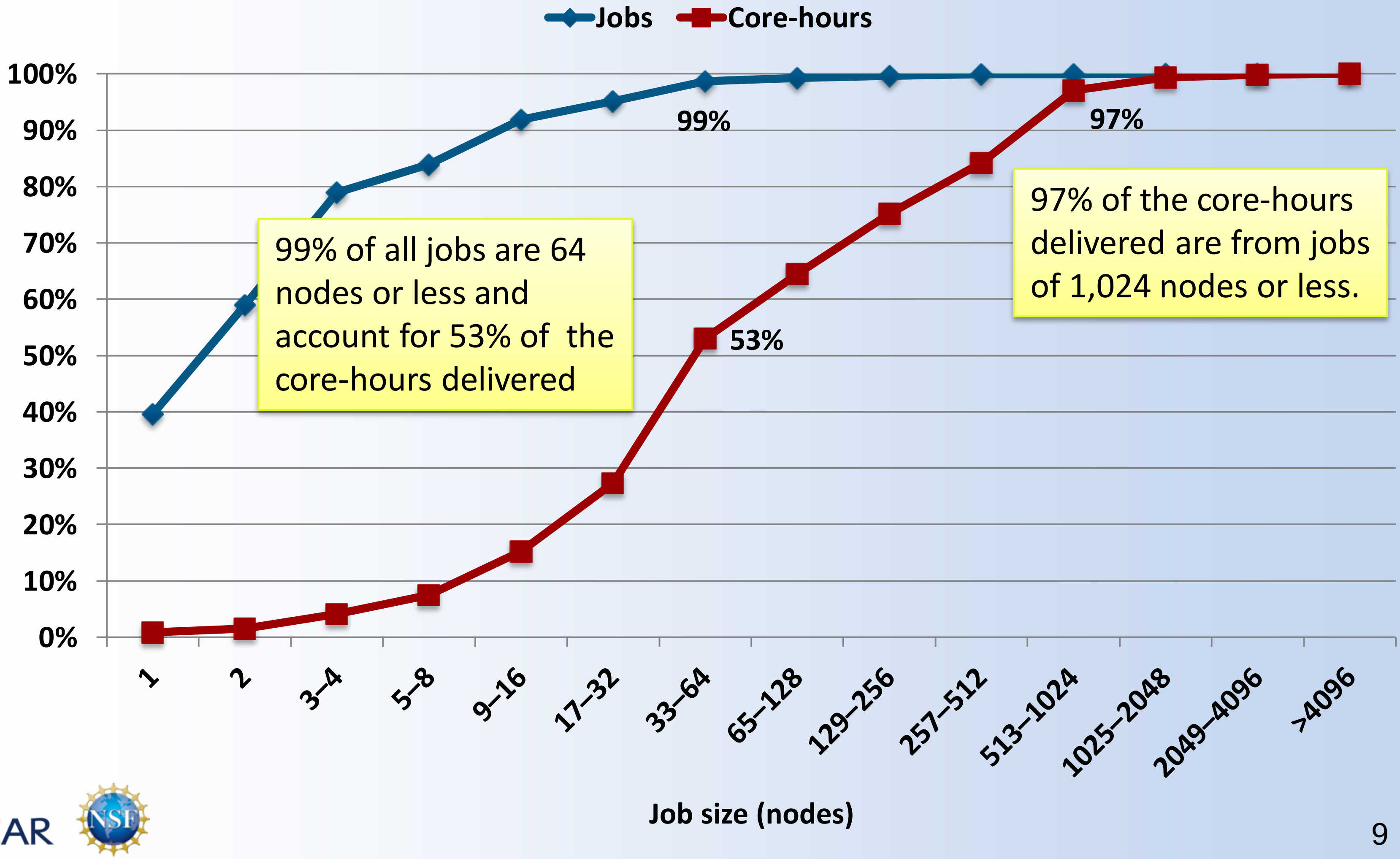
Applications used on Yellowstone



- *50+% of use from CESM (not shown on this chart)*
- **52+ other apps/models identified in 171 projects, representing 95% of resource use**

Computational & Information Systems Laboratory

Most Yellowstone jobs are ‘small’ while ~50% of core-hours are consumed by jobs using >64 nodes

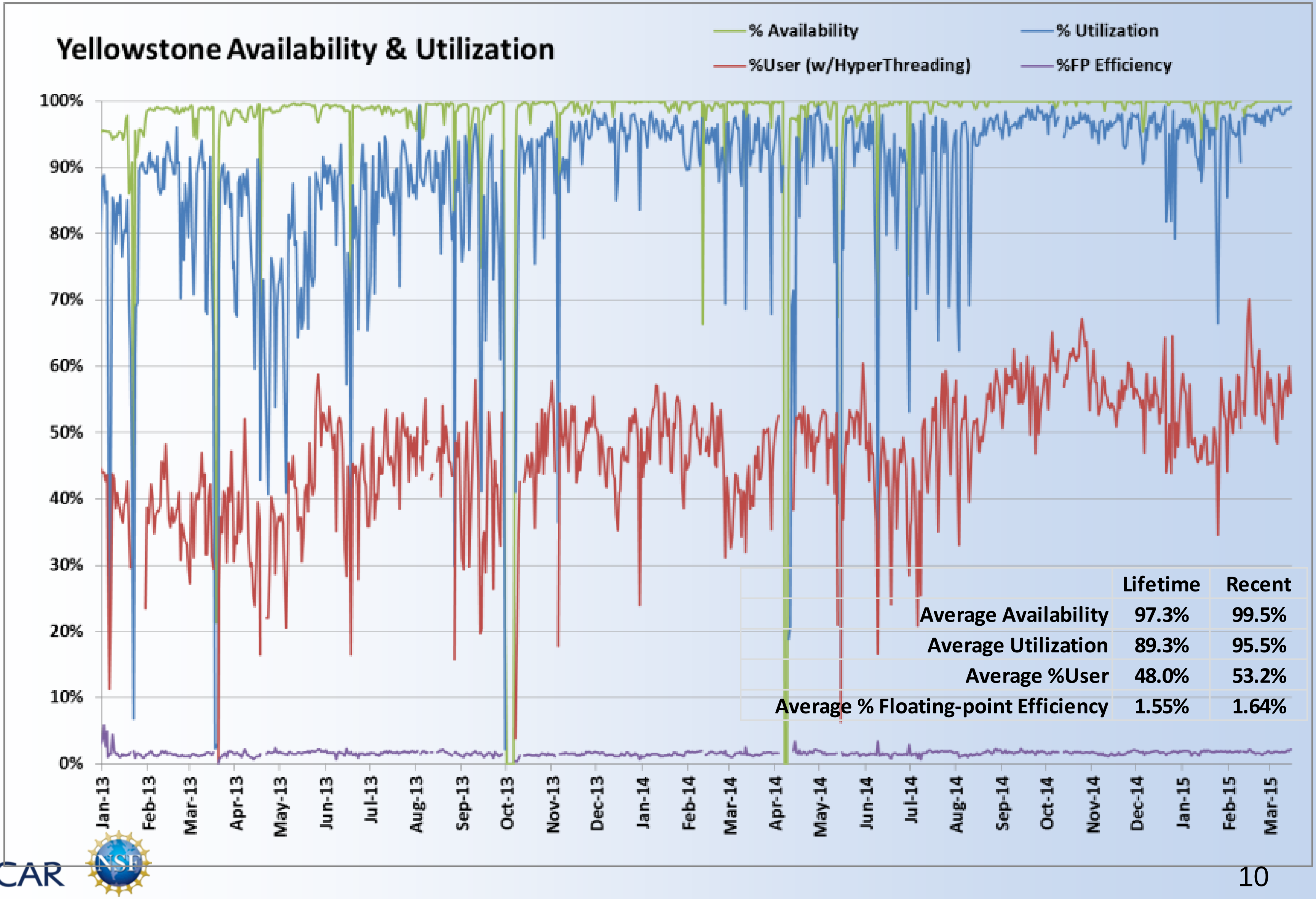


99% of all jobs are 64 nodes or less and account for 53% of the core-hours delivered

97% of the core-hours delivered are from jobs of 1,024 nodes or less.

Yellowstone:

High Availability, High Utilization



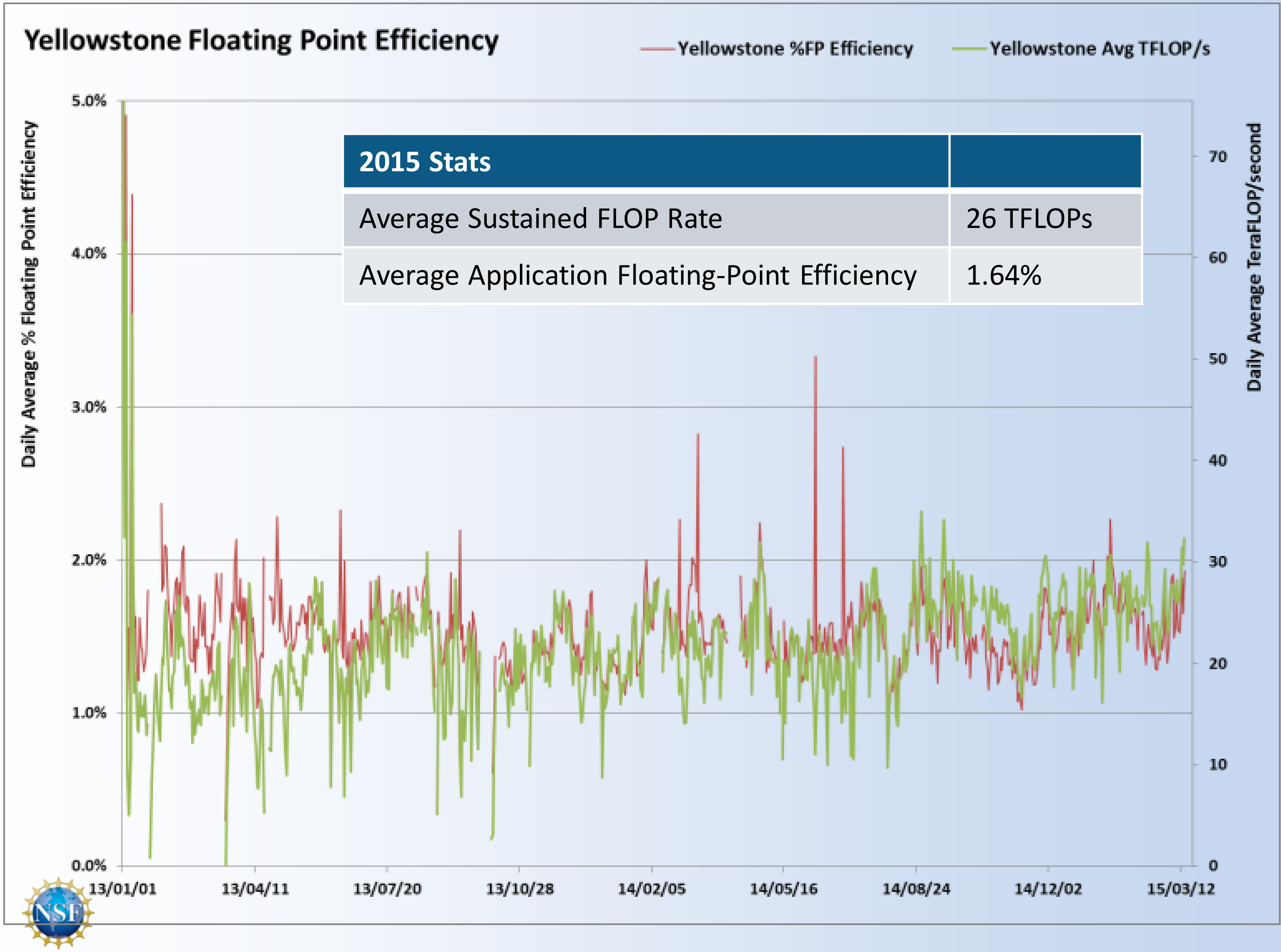
	Lifetime	Recent
Average Availability	97.3%	99.5%
Average Utilization	89.3%	95.5%
Average %User	48.0%	53.2%
Average % Floating-point Efficiency	1.55%	1.64%

CISL Computational & Information Systems Laboratory

NCAR

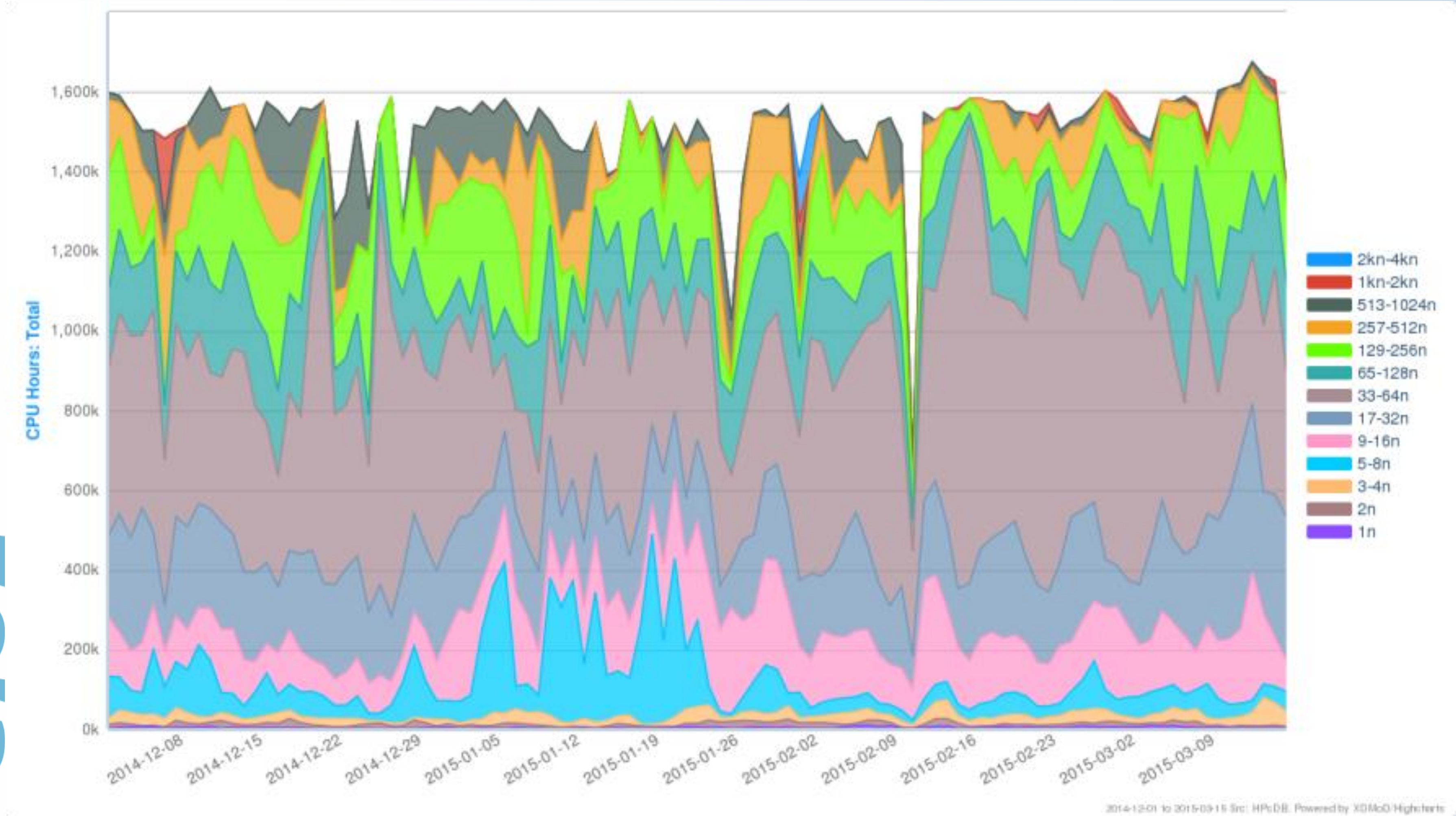
NSF

Average Yellowstone floating point efficiency is low relative to theoretical peak



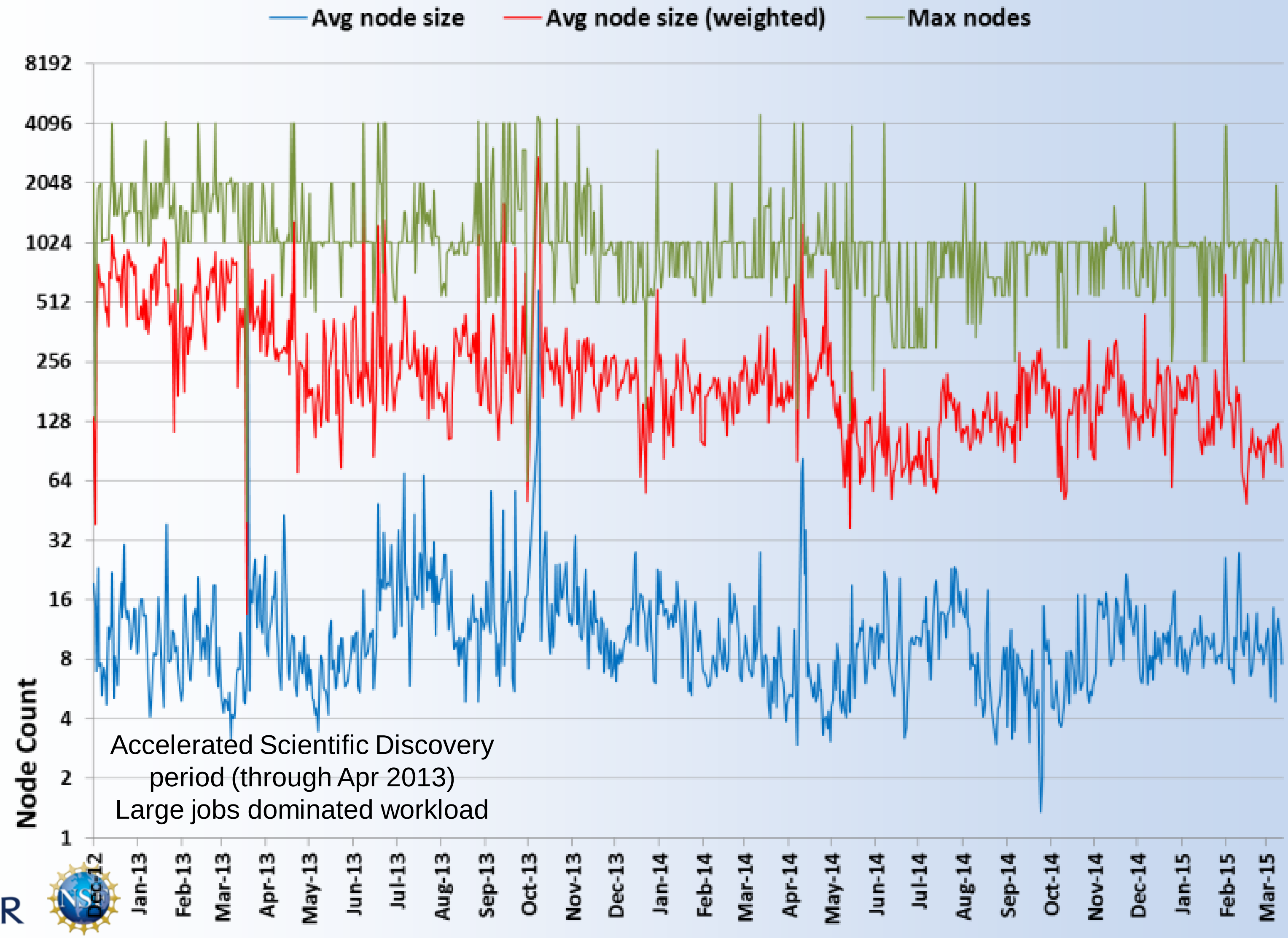
2015 Stats	
Average Sustained FLOP Rate	26 TFLOPs
Average Application Floating-Point Efficiency	1.64%

Recent production workload is dominated by jobs using 9-128 nodes



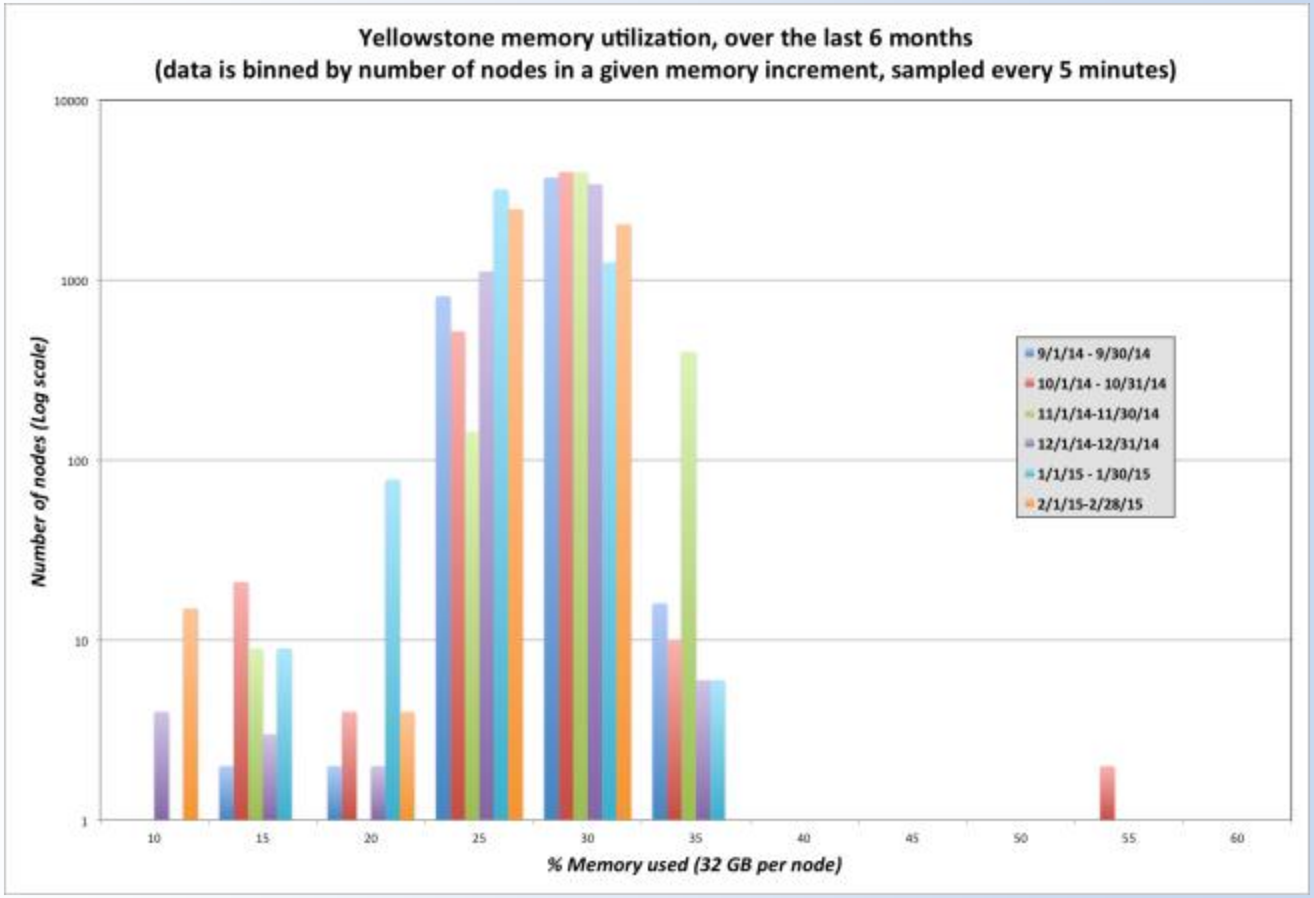
2014-12-01 to 2015-03-15 Src: HPcDB. Powered by XDMoD/Highcharts

Historical trends in job size (average, weighted, max) show no dramatic shifts



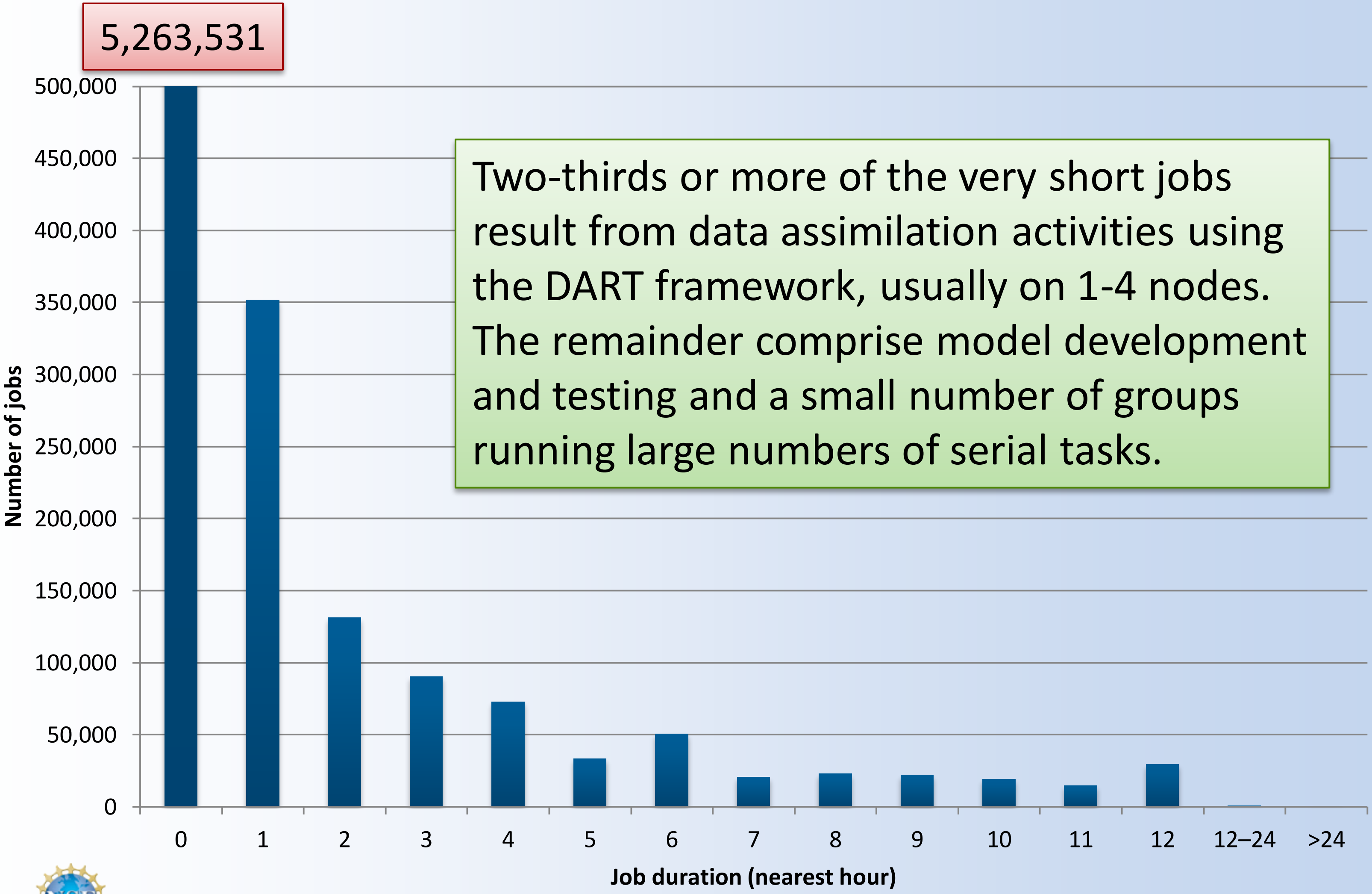
Accelerated Scientific Discovery
period (through Apr 2013)
Large jobs dominated workload

On average, applications use about 30% of Yellowstone's available node-level memory



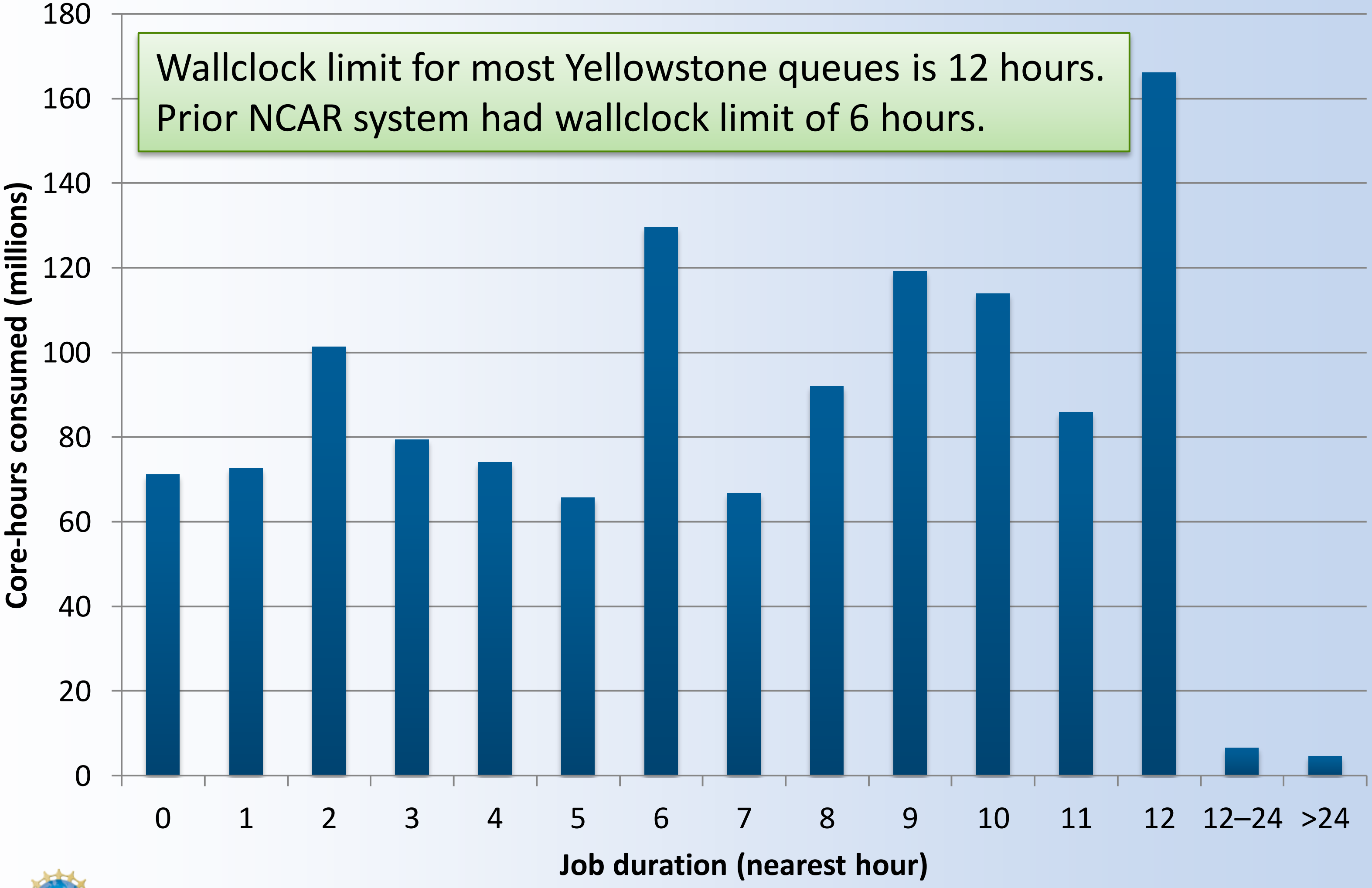
- Yellowstone has 32 GB of memory per node which is 2 GB/core
- Memory use is collected from each node every 5 minutes, then averaged over time.
- There has been a slight uptick in node memory utilization over time

When looking at runtimes, most jobs consume 30 minutes or less



Two-thirds or more of the very short jobs result from data assimilation activities using the DART framework, usually on 1-4 nodes. The remainder comprise model development and testing and a small number of groups running large numbers of serial tasks.

When looking at core-hours consumed, distribution of runtimes is fairly uniform



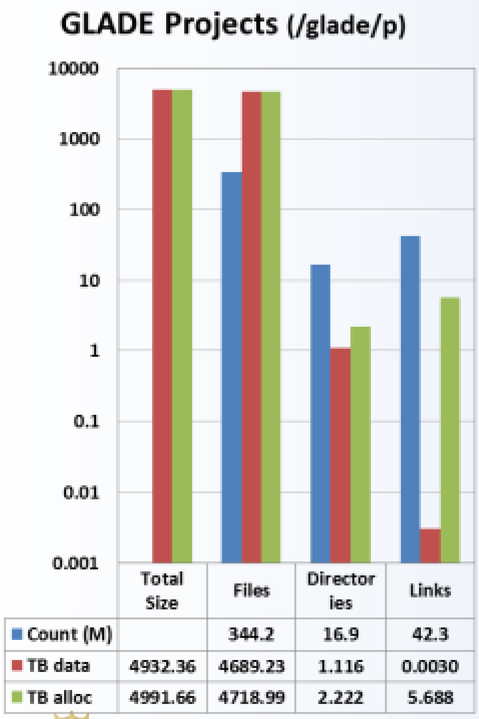
Wallclock limit for most Yellowstone queues is 12 hours.
Prior NCAR system had wallclock limit of 6 hours.

GLADE: GLobally Accessible Data Environment

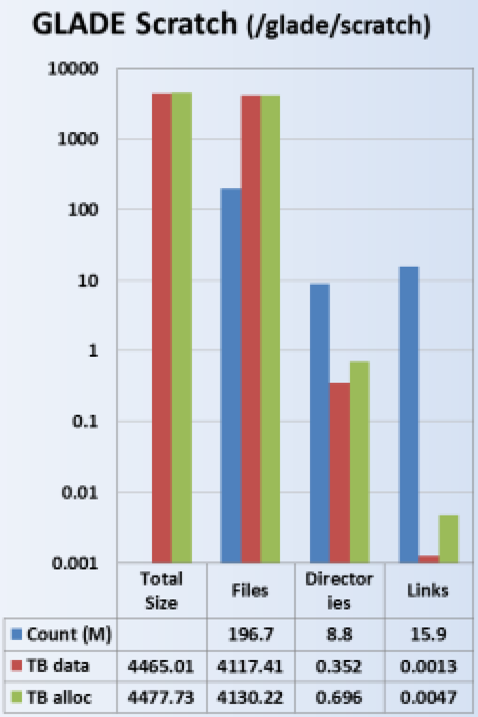
- **GPFS NSD Servers**
  - 20 IBM x3650 M4 nodes; Intel Xeon E5-2670 processors w/AVX
  - 16 cores, 64 GB memory per node; 2.6 GHz clock
  - 91.8 GB/sec aggregate I/O bandwidth (4.8+ GB/s/server)
- **I/O Aggregator Servers (export GPFS, GLADE-HPSS connectivity)**
  - 4 IBM x3650 M4 nodes; Intel Xeon 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 each populated with 90 3 TB NL-SAS drives/controller
  - 16.42 PB usable capacity

GLADE Filesystems Snapshot (March 2015)

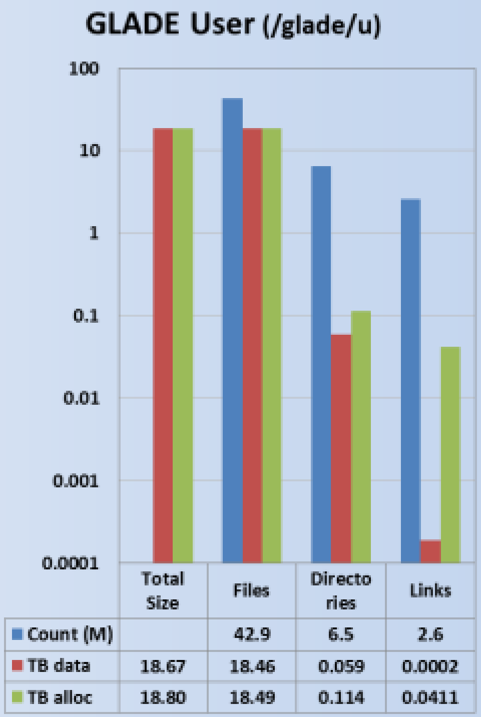
File System	Intended use	Capacity (PB)	Used (PB)	Sub-block/ Block size
/glade/u	User program files; environment	.8	.02	16KB / 512KB
Projects	Allocated project space; not purged	9	5.0	128KB / 4MB
Scratch	Scratch space; purged (90 day retention)	5	4.5	128KB / 4MB


GLADE Projects (/glade/p)

	Total Size	Files	Directories	Links
🟦 Count (M)		344.2	16.9	42.3
🟥 TB data	4932.36	4689.23	1.116	0.0030
🟩 TB alloc	4991.66	4718.99	2.222	5.688


GLADE Scratch (/glade/scratch)

	Total Size	Files	Directories	Links
🟦 Count (M)		196.7	8.8	15.9
🟥 TB data	4465.01	4117.41	0.352	0.0013
🟩 TB alloc	4477.73	4130.22	0.696	0.0047

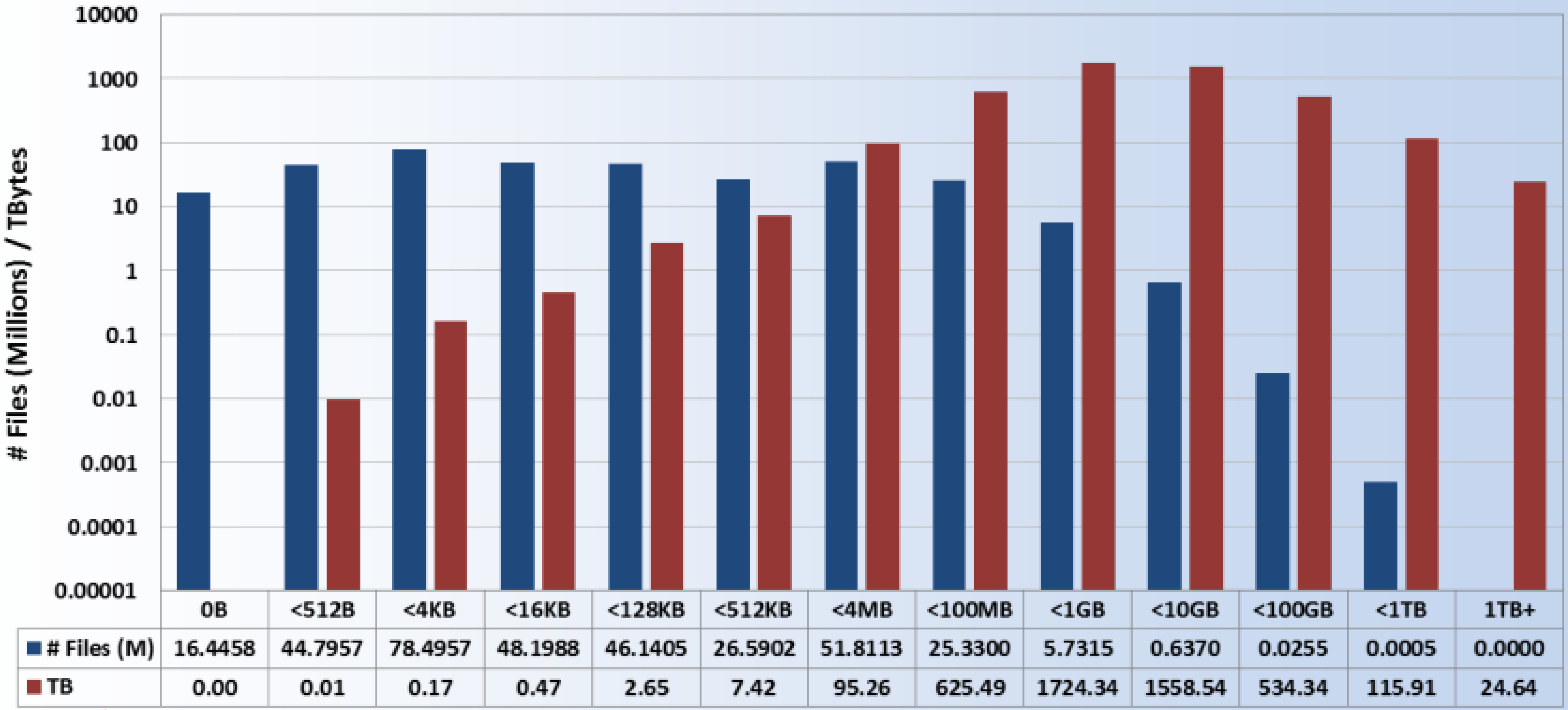

GLADE User (/glade/u)

	Total Size	Files	Directories	Links
🟦 Count (M)		42.9	6.5	2.6
🟥 TB data	18.67	18.46	0.059	0.0002
🟩 TB alloc	18.80	18.49	0.114	0.0411

Project file system is dominated by a large number of small files

TB Used	# Files	# Dirs	# Links
4,992	344.2 M	16.9 M	42.3 M

/glade/p File System (Project Space) [4MB block, 128kB subblock]

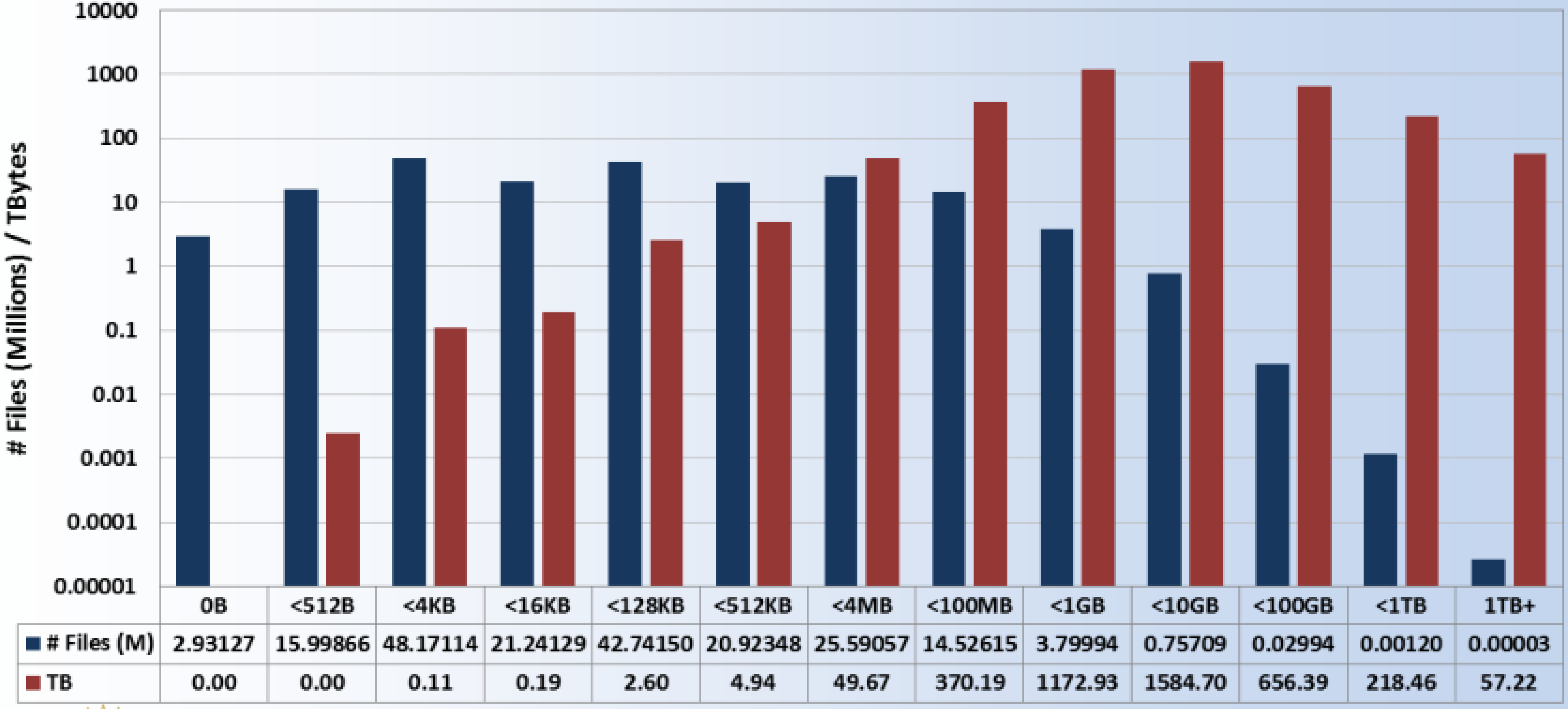


	0B	<512B	<4KB	<16KB	<128KB	<512KB	<4MB	<100MB	<1GB	<10GB	<100GB	<1TB	1TB+
🟦 # Files (M)	16.4458	44.7957	78.4957	48.1988	46.1405	26.5902	51.8113	25.3300	5.7315	0.6370	0.0255	0.0005	0.0000
🟥 TB	0.00	0.01	0.17	0.47	2.65	7.42	95.26	625.49	1724.34	1558.54	534.34	115.91	24.64

Scratch file system exhibits similar usage pattern as *Projects* space

TB Used	# Files	# Dirs	# Links
4,478	196.7 M	8.8 M	15.9 M

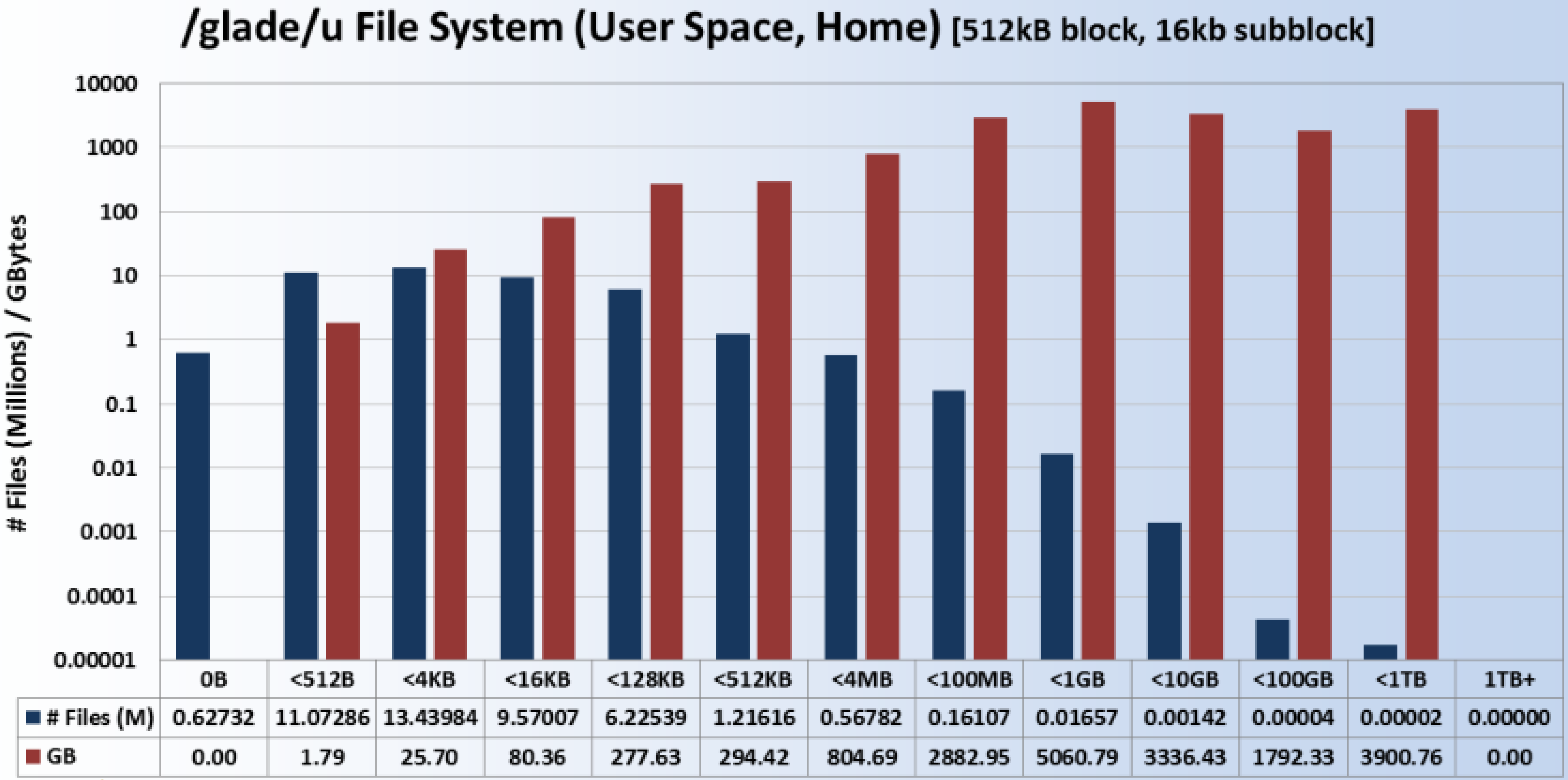
/glade/scratch File System (Scratch Space) [4MB blck, 128kB subblock]



	0B	<512B	<4KB	<16KB	<128KB	<512KB	<4MB	<100MB	<1GB	<10GB	<100GB	<1TB	1TB+
🟦 # Files (M)	2.93127	15.99866	48.17114	21.24129	42.74150	20.92348	25.59057	14.52615	3.79994	0.75709	0.02994	0.00120	0.00003
🟥 TB	0.00	0.00	0.11	0.19	2.60	4.94	49.67	370.19	1172.93	1584.70	656.39	218.46	57.22

/glade/u file system is used for home file system, applications & tools directories

TB Used	# Files	# Dirs	# Links
18.5	42.9 M	6.5 M	2.6 M


/glade/u File System (User Space, Home) [512kB block, 16kb subblock]

	0B	<512B	<4KB	<16KB	<128KB	<512KB	<4MB	<100MB	<1GB	<10GB	<100GB	<1TB	1TB+
■ # Files (M)	0.62732	11.07286	13.43984	9.57007	6.22539	1.21616	0.56782	0.16107	0.01657	0.00142	0.00004	0.00002	0.00000
■ GB	0.00	1.79	25.70	80.36	277.63	294.42	804.69	2882.95	5060.79	3336.43	1792.33	3900.76	0.00

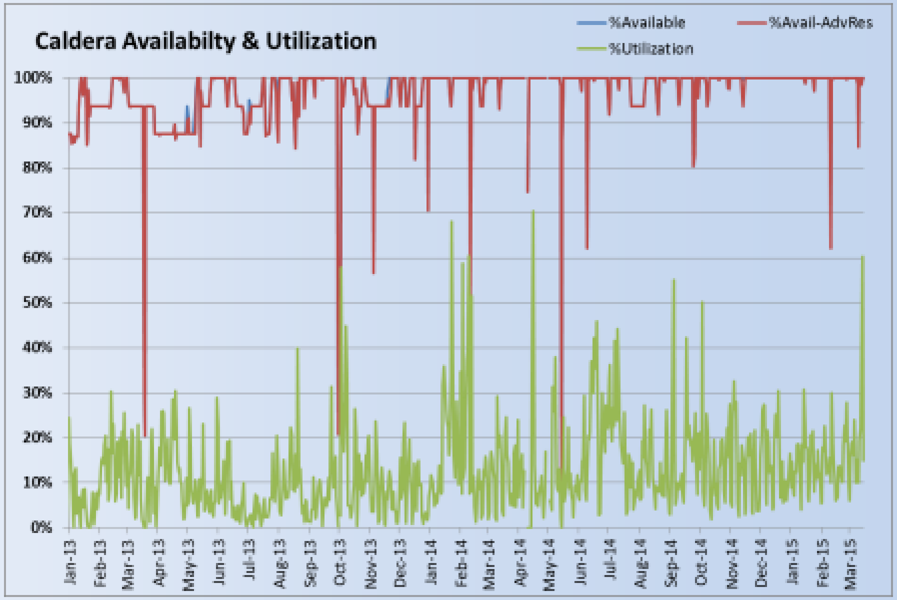
DAV Resource Utilization

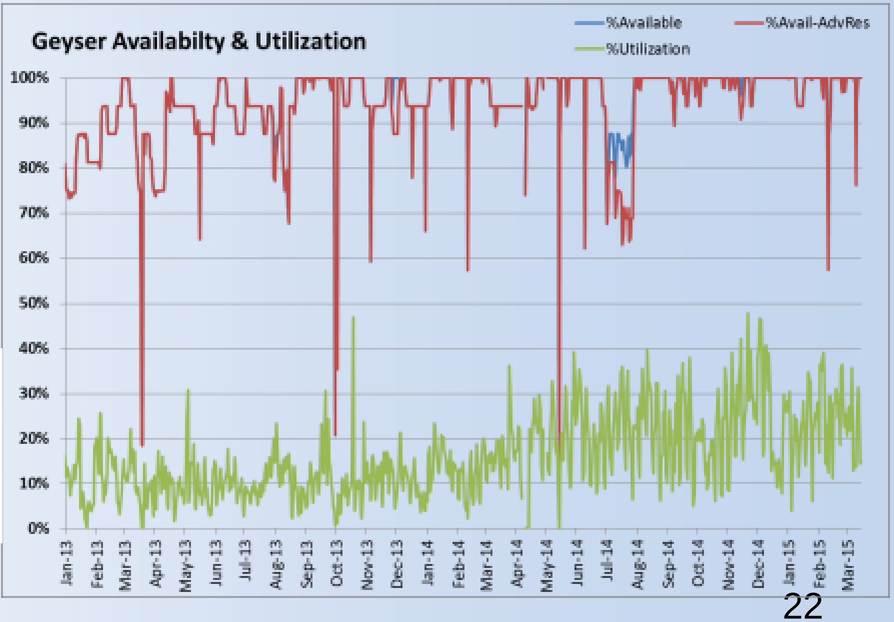
Lifetime average utilization:
Caldera 12.8%
Geyser 16.0%

There has been a slight uptrend in utilization of both DAV systems in recent months.

While the DAV resources are, in part, meant to be used interactively (and thus should not be routinely running at high %utilization), they remain relatively underutilized – particularly the caldera GPU-accelerated computational system.

	# Node	mem/node	GPU/node
Caldera	16	64 GB	2 NVIDIA K20X
Geyser	16	1 TB	1 NVIDIA K5000





Profile of a “typical” CESM run

- Between 3.54 GB per *node* (2° resolution) and 7 GB per *node* (¼° resolution)
- 15 cores, 2 threads per core (not all CESM models are threaded, however). Best Yellowstone configuration for modest-sized runs (may not be true for all machines).
- The use of 16 cores appears to result in high OS noise (jitter) that reduces performance below the 15 core configuration. Active area of investigation.
- Largest cases may not use threading (affects on scalability being investigated)

I/O pattern of a typical CESM run shows lots of small files doing small block I/O

- Opens 400-750 files (depending on configuration)
- Has aggregate I/O performance of 350-450 MB/s
- Spends 3%-8% of runtime in I/O
- Most I/O operations are very small (< 100 Byte) POSIX file operations, but model output is written as ~512 kB chunks
- Analysis of GLADE/GPFS performance shows no bottlenecks in metadata, disk, or network I/O traffic