Cheyenne Workload and Usage Analysis

Version 1.1

RFP000074 Attachment 5

Prepared for the NWSC-3 Procurement Process

March 2020

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Purpose and scope of this analysis

- Understanding the NCAR application workload is a critical part of scoping the requirements of future system procurements.
- Analysis of application performance on Cheyenne 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

Section 1

Hardware Environment & User Community

Cheyenne

Planned production: January 2017 – July 2022

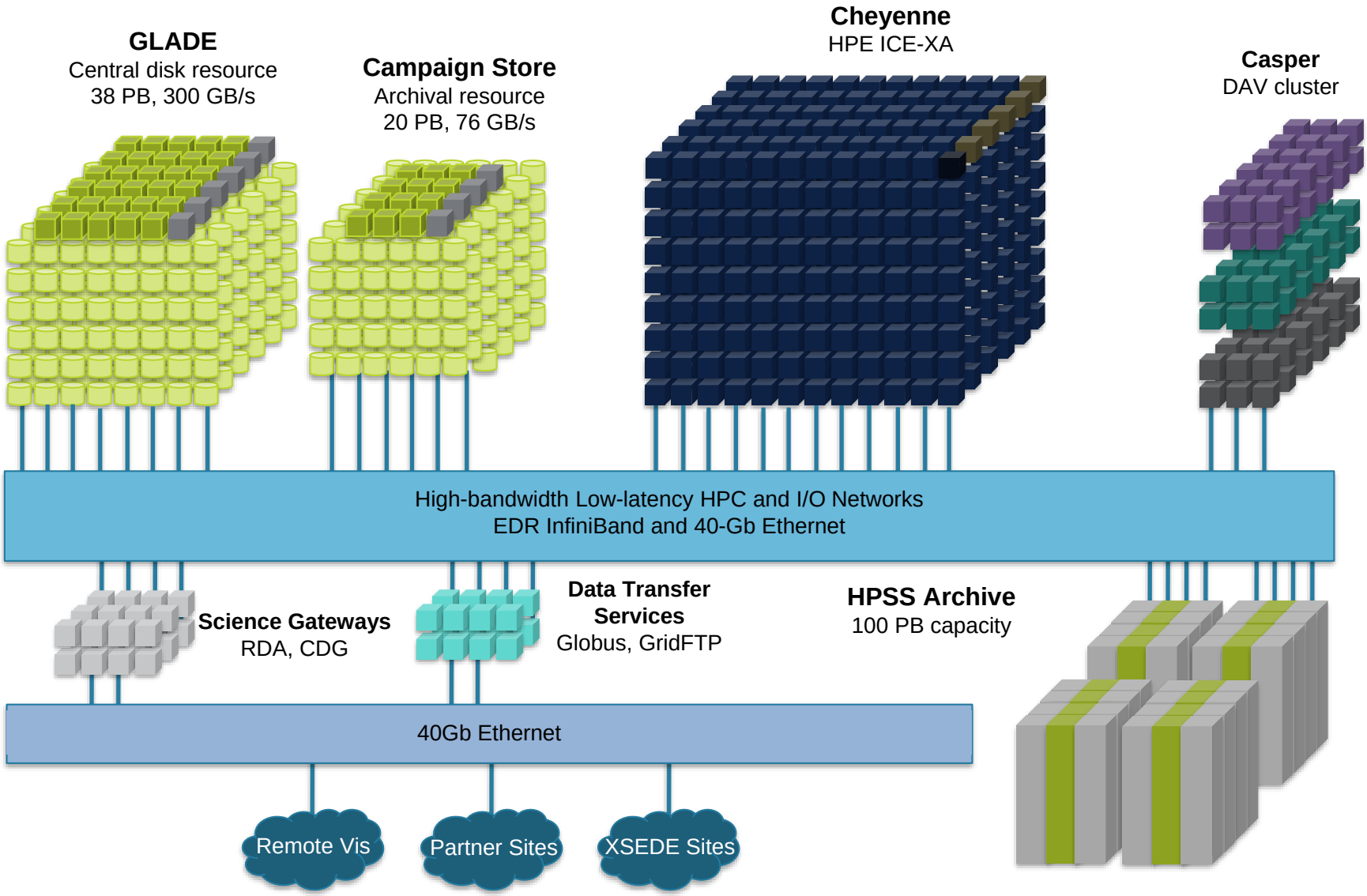
- Scientific Computation Nodes
  - HPE/SGI ICE XA cluster
  - 4,032 dual-socket nodes
  - 18-core, 2.3-GHz Intel Xeon E5-2697v4 processors
  - 145,152 Broadwell cores total
  - 5.34 PFLOPs peak
  - 313 TB total memory (3,164 64-GB and 864 128-GB nodes)
- High-Performance Interconnect
  - Mellanox EDR InfiniBand
  - 9-D enhanced hypercube topology
  - 97 Gbps link bandwidth — 0.5 µs latency
  - 36 TB/s bisection bandwidth
  - 224 36-port switches, no director switches
- GLADE — Central file systems and storage
  - 38 PB usable
  - 8x DDN SFA14KXe each with 10x 84-slot drive chassis
    - 32 embedded NSD servers
    - 6,580 8-TB SAS disk drives
    - 160 4-TB SSD drives
  - ~300 GB/s aggregate I/O bandwidth for new capacity
  - IBM Spectrum Scale (GPFS) file system

Cheyenne physical infrastructure

Resource	# Racks	
Cheyenne	14	ICE XA E-Cells each containing 2 water-cooled E-Racks & heat exchanger, and 16 Mellanox 36-port EDR InfiniBand switches
	2	air-cooled storage & service racks including login nodes
GLADE	8	DDN SFA14KXe racks containing 32 NSD servers and storage

Total Power	~2.0 MW
HPC	1.75 MW
GLADE	0.21 MW

Cheyenne environment

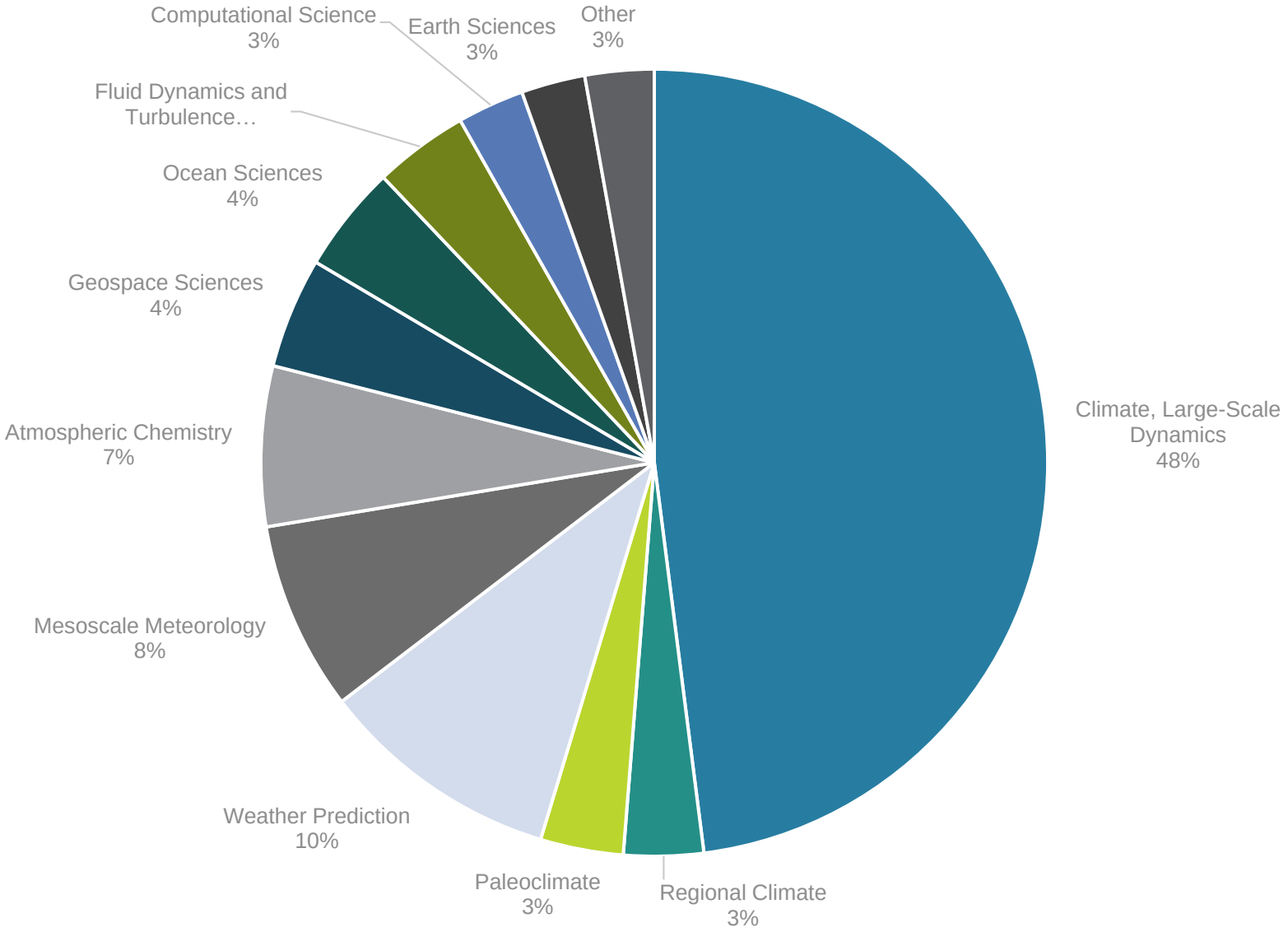


User communities

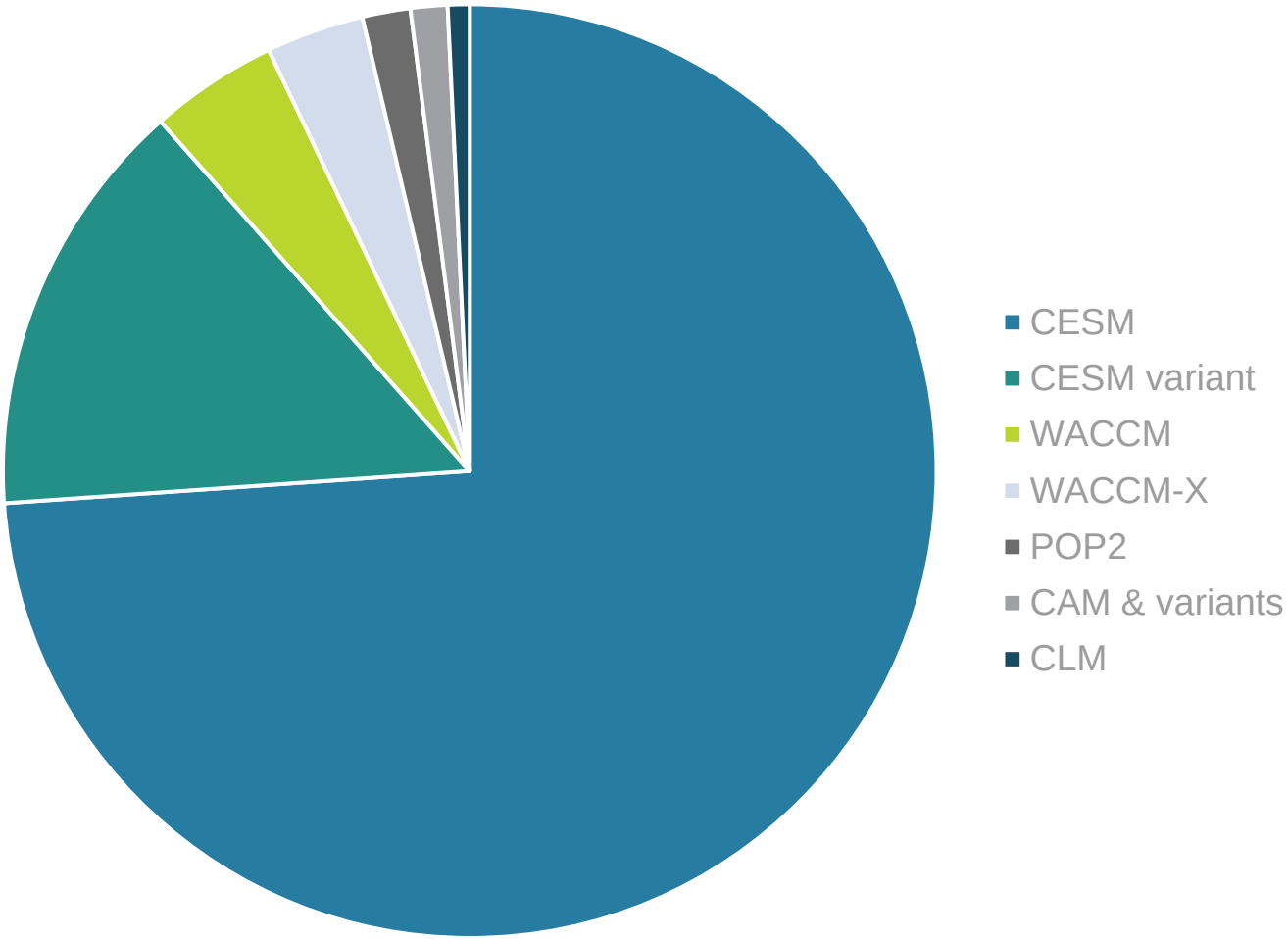
1,700 HPC users in the last 12 months — averaging 700 users each month
850 projects in the last 12 months — averaging more than 380 projects each month

- **NCAR staff (29%)**
  - Larger activities proposed by lab researchers and reviewed by panel of NCAR scientists
  - Small- medium-scale use managed by labs
- **University (29%)**
  - Large-scale projects reviewed by panel
  - Small-scale projects for PIs and grad students/post-docs
- **Climate Simulation Laboratory (28%)**
  - Large portion devoted to CESM community
  - Support for large-scale climate-focused projects from University community
- **Wyoming researchers (13%)**
  - Smaller number of activities from a broader set of science domains

Cheyenne usage reflects its mission for the atmospheric sciences

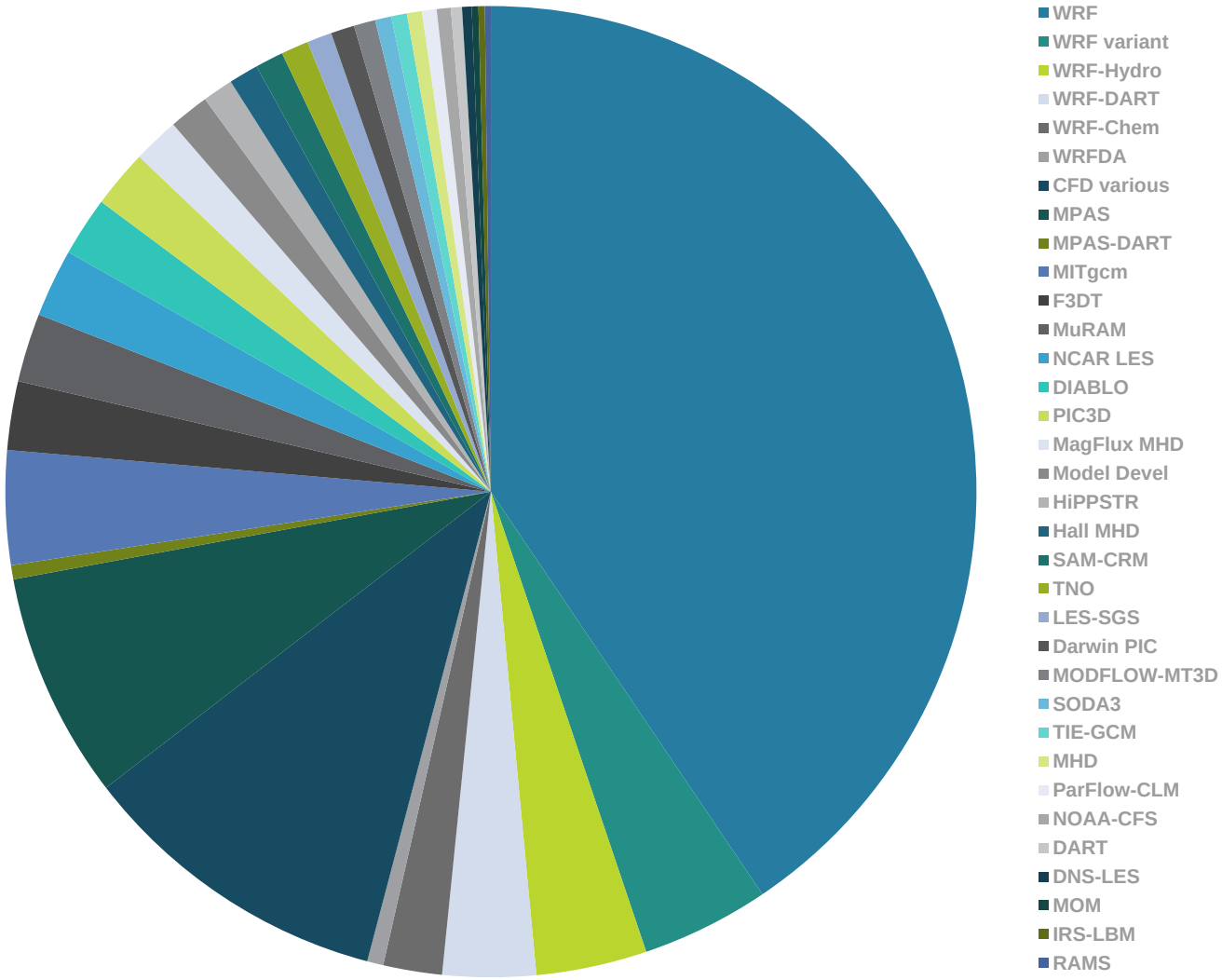


CESM represents 50% of Cheyenne use



50% of Cheyenne use is from CESM, CESM variants, and component models

50% of Cheyenne use *not* from CESM

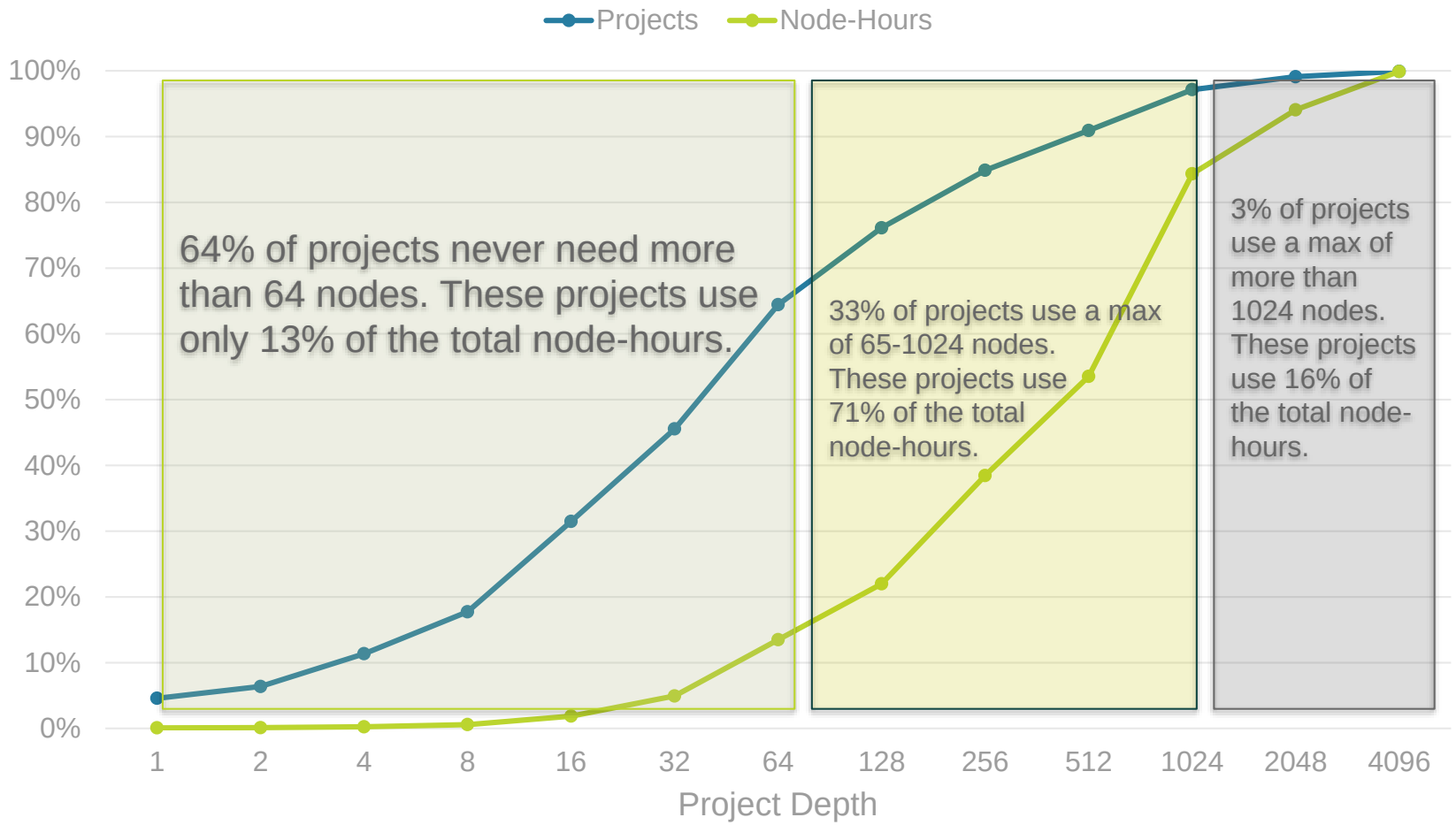


Besides CESM, 30+ other applications and models were identified.

Section 2

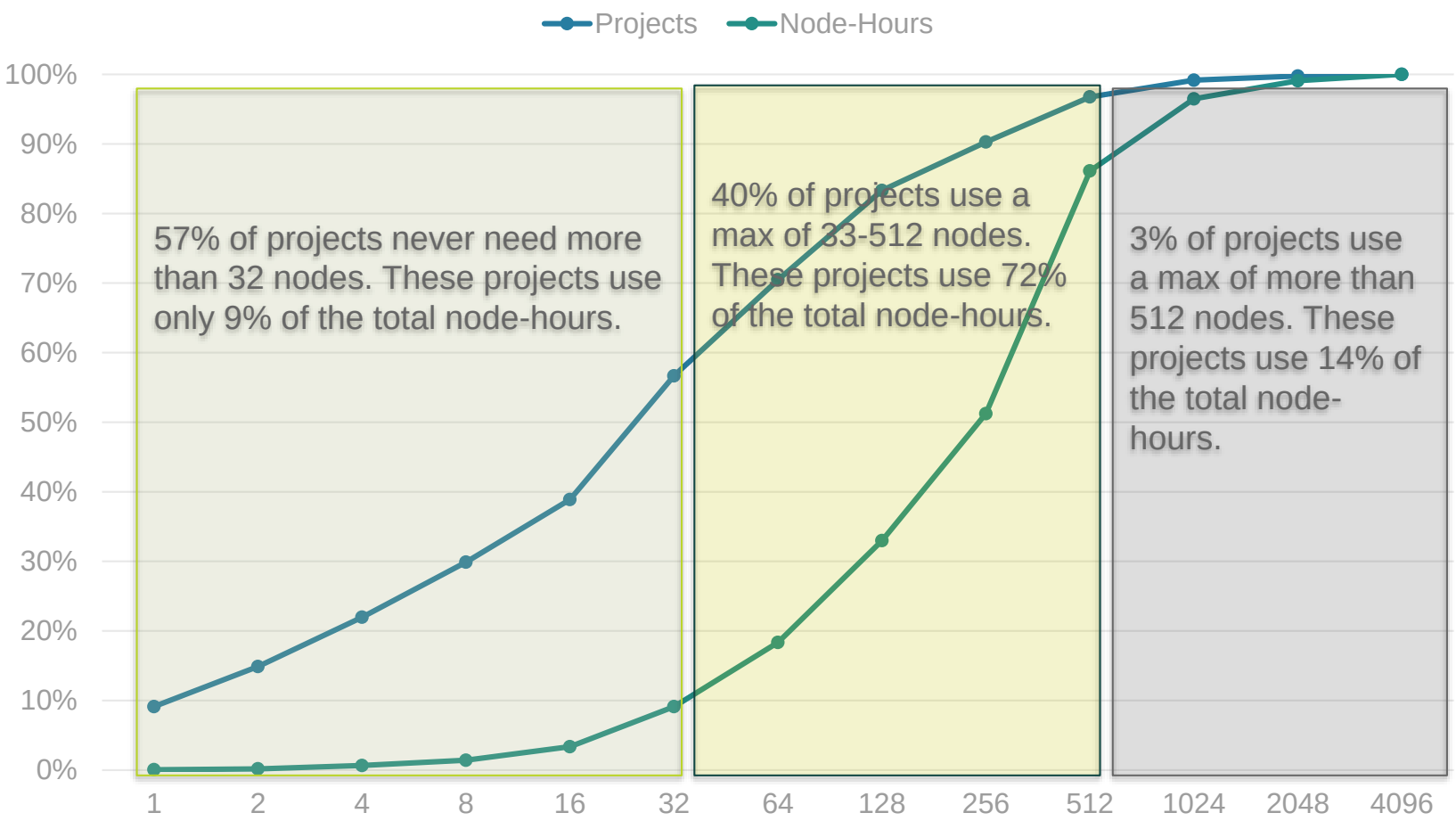
Cheyenne Job Patterns & Workload

Yellowstone projects and job sizes



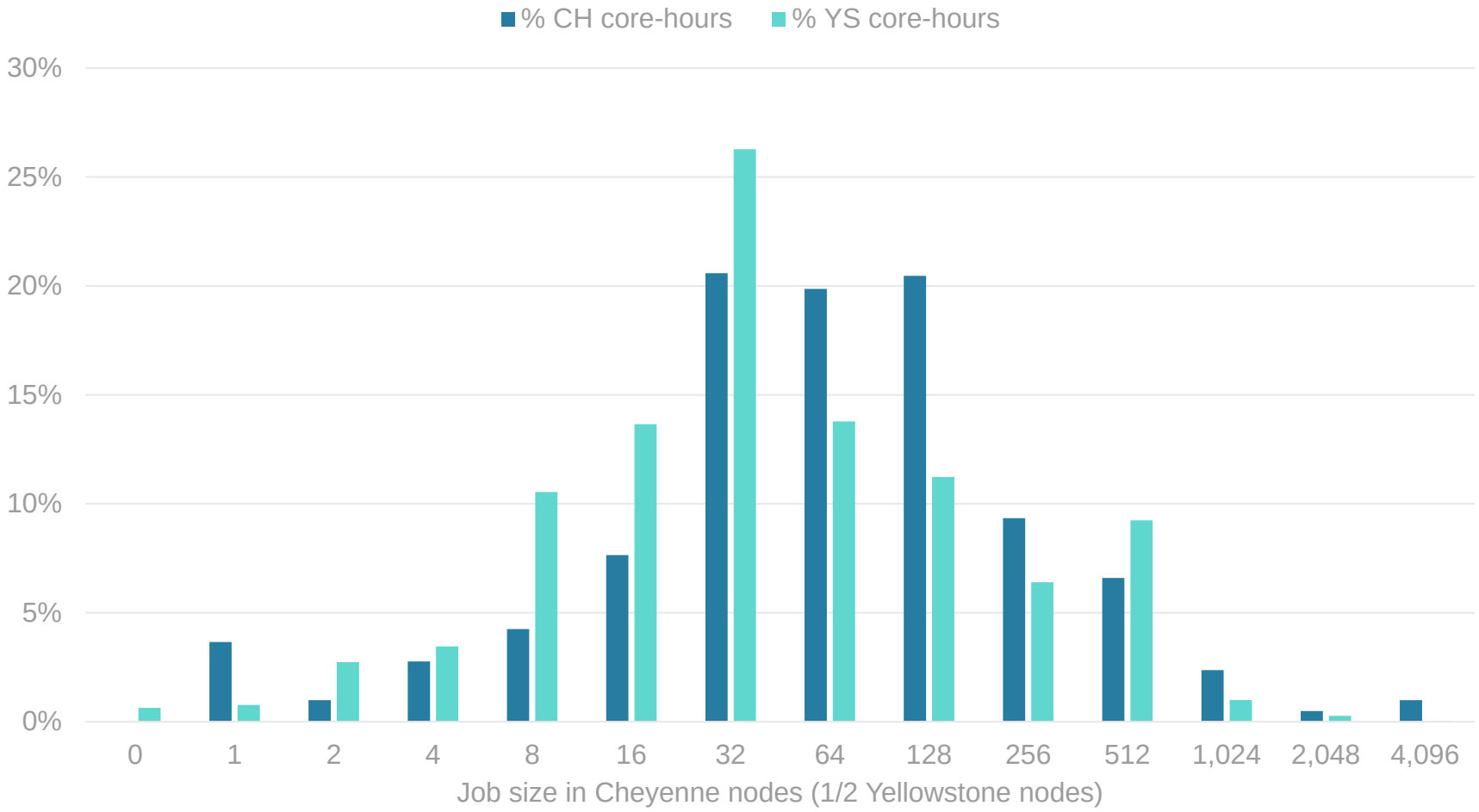
A project's depth is defined as the size of its largest job in nodes (rounded up to the next greater power of 2). Note: Yellowstone nodes had 16 cores.

Cheyenne projects and job sizes

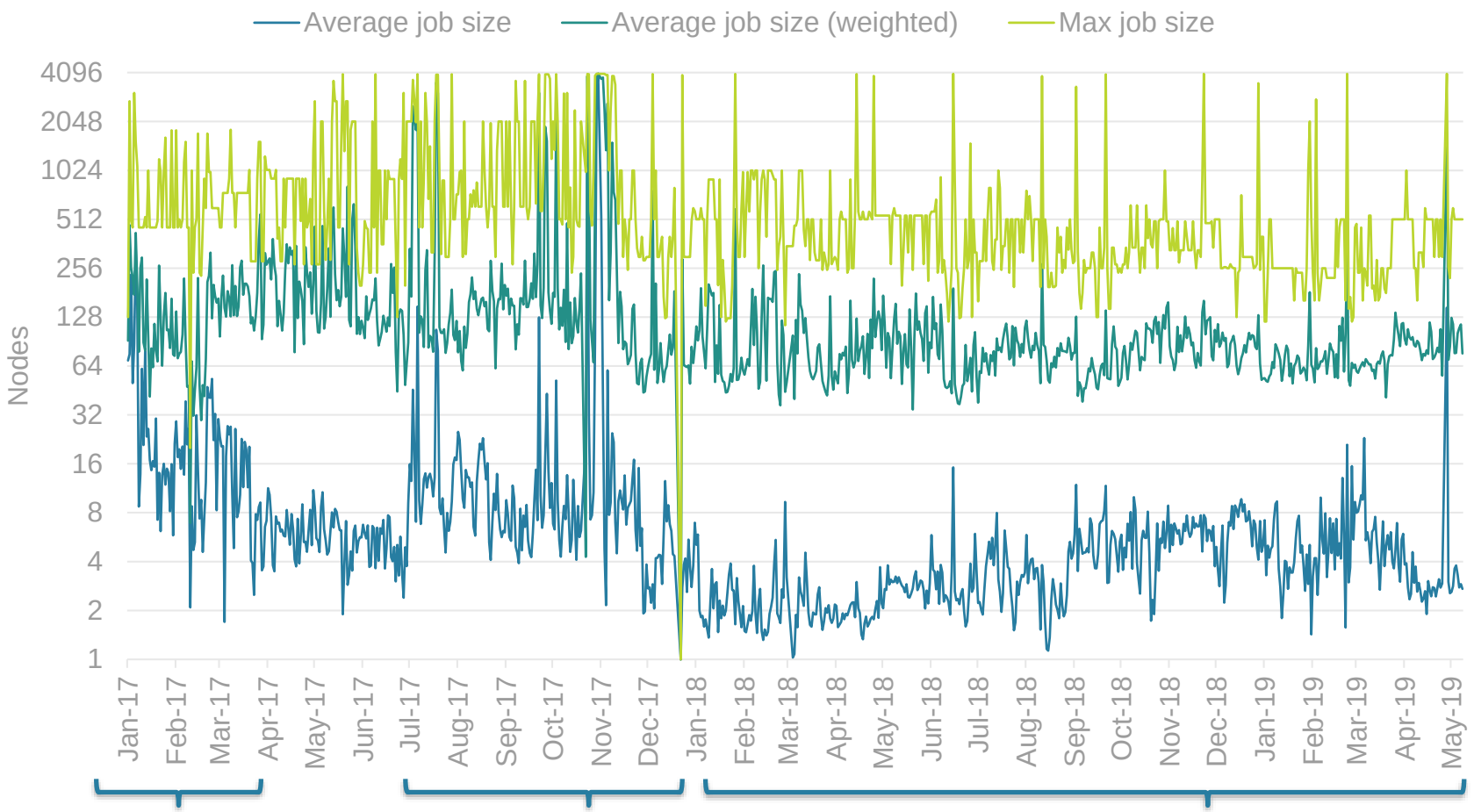


A project's depth is defined as the size of its largest job (rounded to the next greater power of 2). Note the similarity in usage by projects compared to Yellowstone use.

Cheyenne has seen a shift toward 64–128 node jobs compared to Yellowstone workload



Historical trends in Cheyenne job size show no dramatic shifts

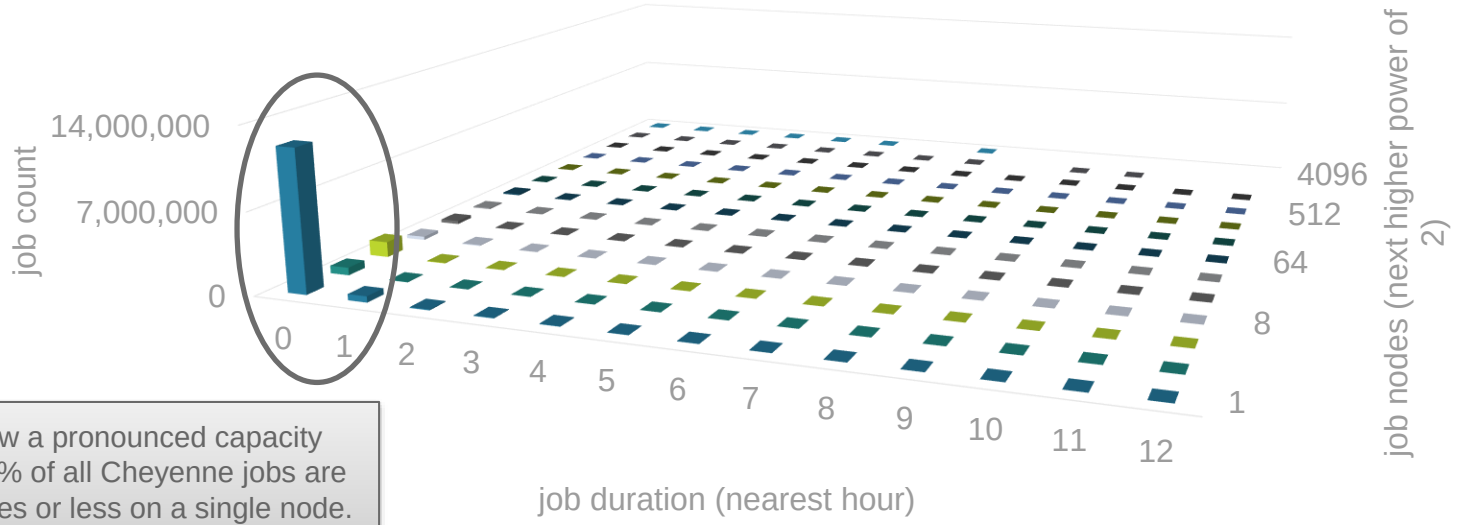


Slightly larger average job size during Accelerated Scientific Discovery project period (Jan–Mar 2017)

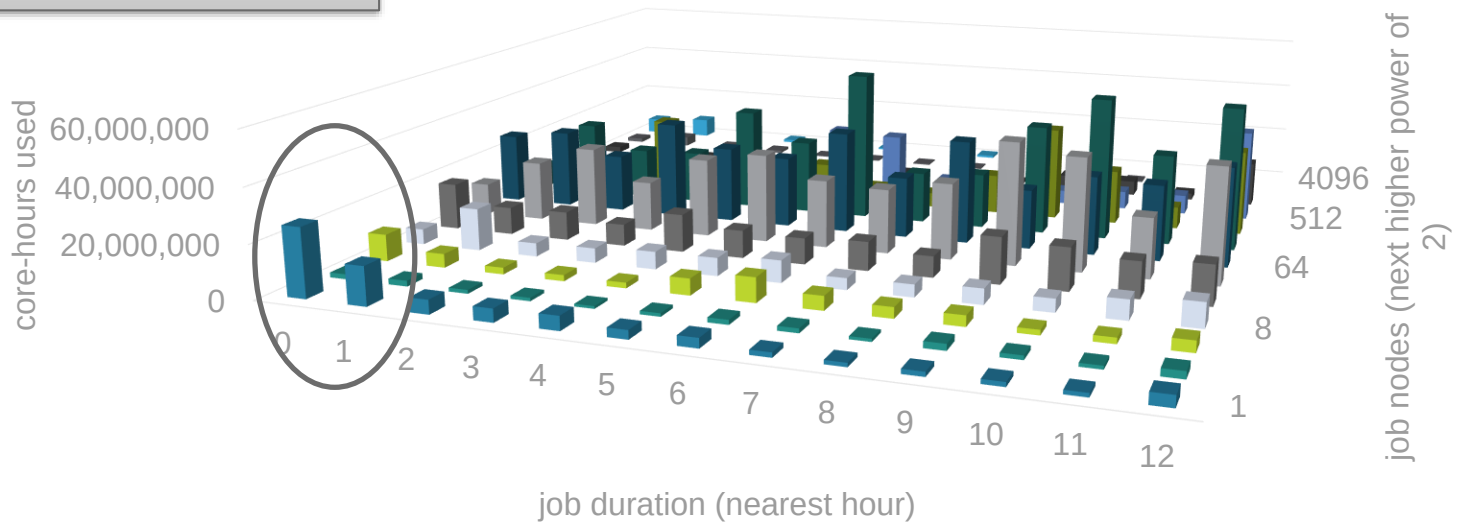
Larger average job size resumed July–Dec 2017 while Yellowstone remained in production and with economy queue discount charging in effect.

Average job size, raw and weighted, decreased further post-Yellowstone. Average (raw) job size increased again during the 2nd half 2018

Cheyenne users need more than HPC



Job data show a pronounced capacity use case: 75% of all Cheyenne jobs are last 30 minutes or less on a single node. They consume 1.3% of total core-hours.

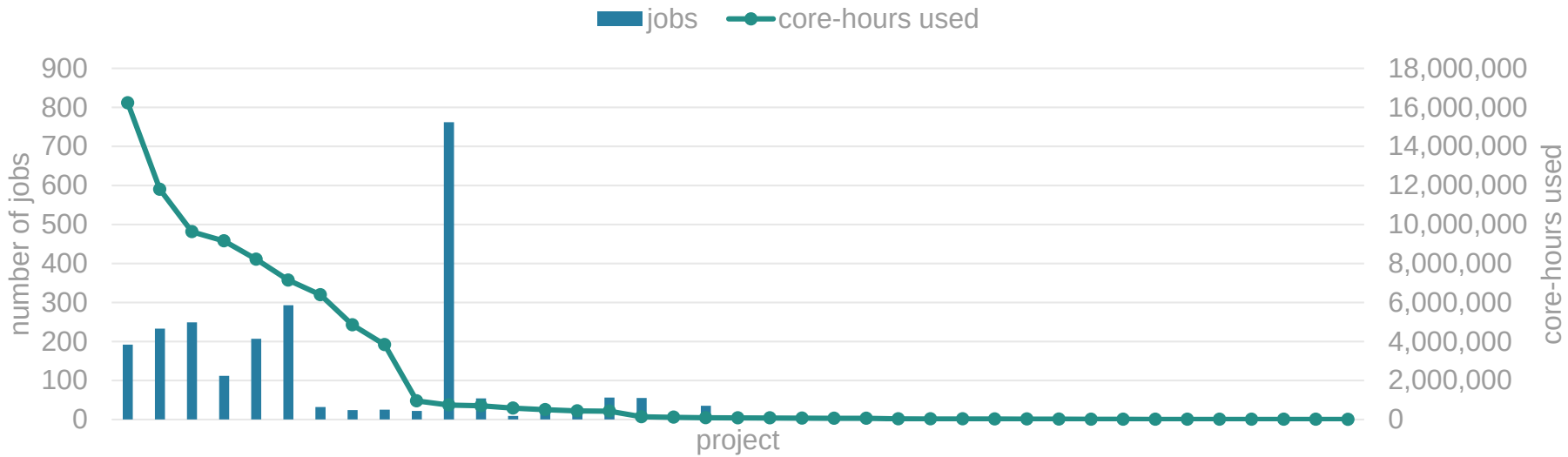


Cheyenne's "pure capacity" workload

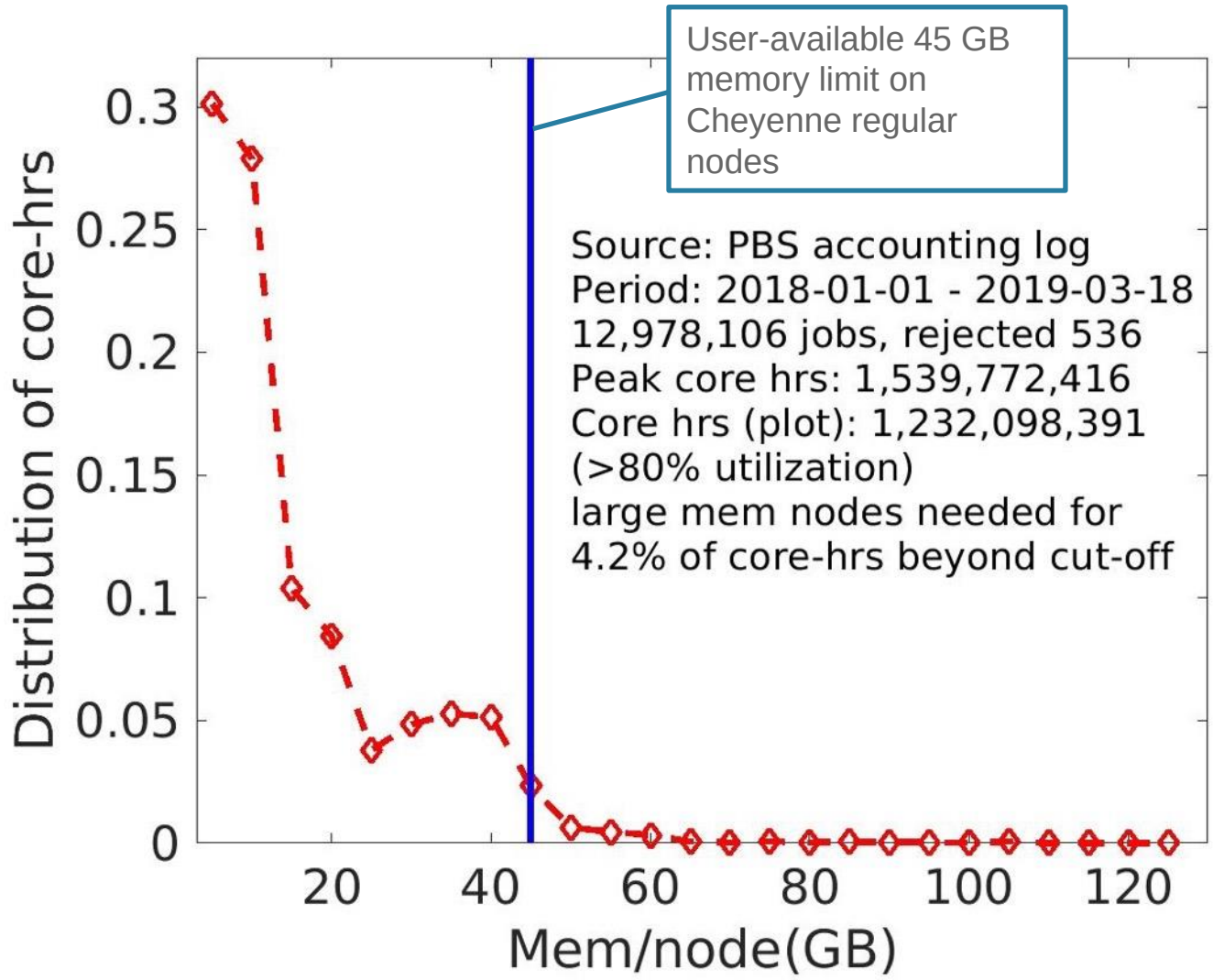
- Single-node, short duration (<30 minutes) jobs dominate the scheduler effort
  - 12.1M jobs, 75% of all jobs submitted
    - *Another 2M (12%) of all jobs are 2-4 nodes, <30 min*
- Single-node, short jobs consumed 25.7M core-hours
  - 1.2% of all core-hours delivered
- 15 projects responsible for 80% of these jobs
  - *top 3 projects responsible for 42% of these jobs*
- 6 projects responsible for 75% of the core-hours
  - *1 project responsible for 35% of the core-hours*
- 25 projects responsible for 90% of these jobs and 91% of the core-hours
- *~40 Cheyenne nodes could have handled all these jobs (assuming they were spread evenly over Cheyenne's production period)*
  - *About 2%-3% of Cheyenne nodes (~120 nodes) could have handled most of the high level "bursts" of activity in this regime*

Cheyenne “very large job” workload

- 39 projects have run at least 1 job on more than 512 nodes
- 2,500 jobs altogether, for 82.4M core-hours (4% of Cheyenne usage)
- 9 projects responsible for 94% of that usage—almost quarter of which was for CISL benchmarking work
- Only 18 projects ran more than 9 such jobs



Cheyenne memory use



More than 95% of Cheyenne jobs fit within 45 GB limit on regular nodes (1.25 GB/core).

Users can idle cores to use more memory per core.

Large-memory nodes have 109 GB usable (3 GB/core).

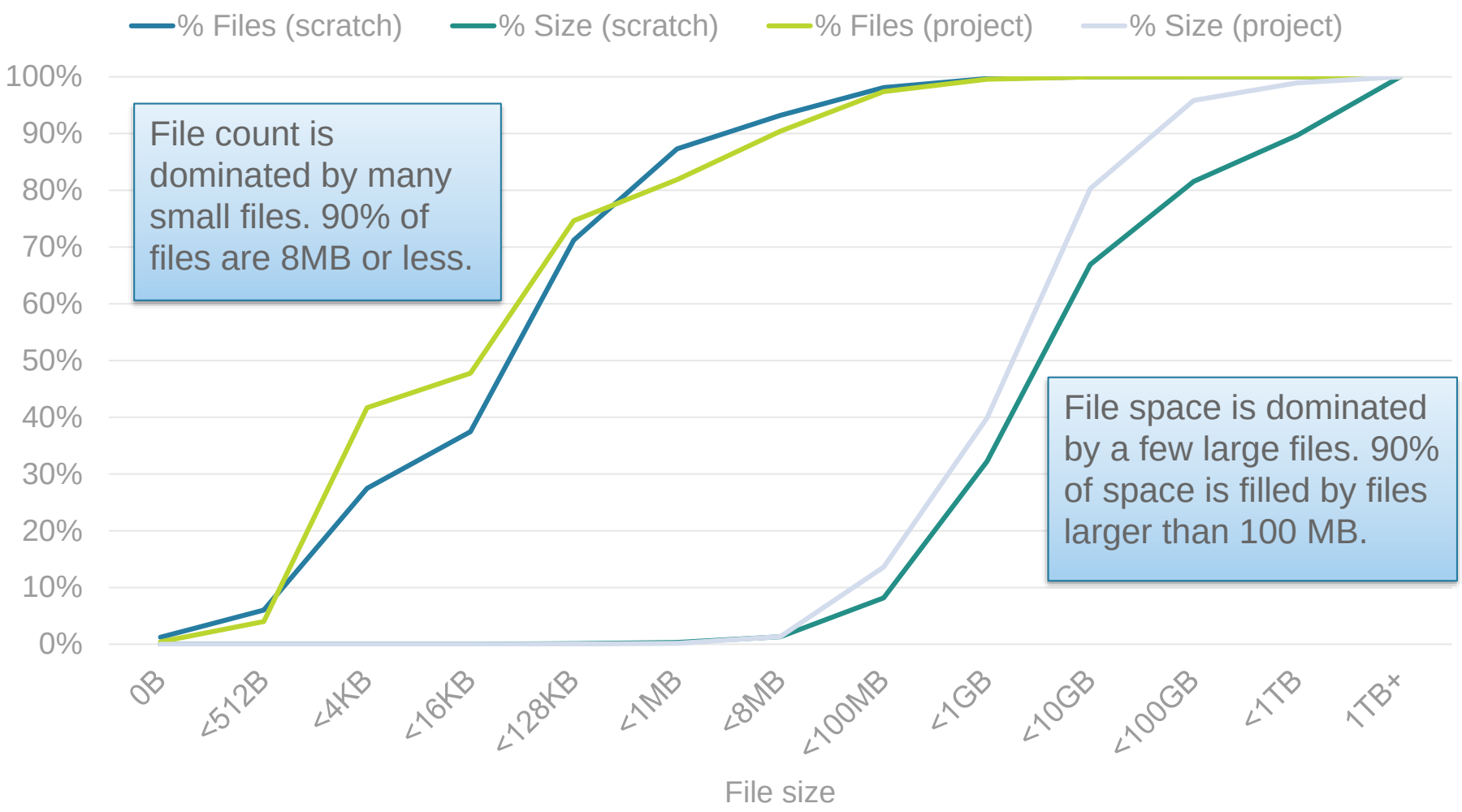
Section 3

File System Usage Patterns

GLADE file systems and purposes

- Scratch
  - 10 TB default per user, increased upon request/need
  - Temporary storage for work in progress
  - 90-day retention from last access, automatic purge
- Project space
  - Allocated via review processes
  - Longer-term storage for shared access by teams of users (projects)
  - 12-month retention from last access, automatic purge
- User work space
  - 1 TB space default per user for various user defined needs
  - No purge
- Flash
  - Access by request
  - SSD file system, temporary storage for high IOPS work
  - 1 week retention
- User home directories
  - 10 GB default per user, backed up

Current GLADE usage: Scratch and project space



Current file distribution: Scratch

Size	# Files	% of Files	Total Size	% of Size
0B	4,619,820	1.23%	0 b	0.00%
<512B	17,939,994	4.78%	2.31 GB	0.00%
<4KB	80,591,848	21.47%	219.14 GB	0.00%
<16KB	37,466,966	9.98%	355.48 GB	0.00%
<128KB	**126,620,017**	**33.73%**	**7.52 TB**	**0.10%**
<1MB	60,384,162	16.09%	16.01 TB	0.21%
<8MB	22,257,298	5.93%	76.41 TB	0.99%
<100MB	18,369,013	4.89%	520.97 TB	6.72%
<1GB	5,873,855	1.56%	1.8307 PB	24.20%
<10GB	**1,162,743**	**0.31%**	**2.6438 PB**	**34.94%**
<100GB	52,301	0.01%	1.1136 PB	14.72%
<1TB	3,076	0.00%	616.06 TB	7.95%
1TB+	50	0.00%	788.04 TB	10.17%

Total size	408,862,855	7.5662 PB	6.8735 PB allocated
Total # Files	375,342,154	7.5590 PB	6.8730 PB allocated
Total # Dirs	8,891,792	74.02 GB	86.68 GB allocated
Total # Links	24,628,909	1.88 GB	0 b allocated

Current file distribution: User work space

Size	# Files	% of Files	Total Size	% of Size
0B	1,373,060	1.07%	0 b	0.00%
<512B	17,957,232	13.99%	3.33 GB	0.00%
<4KB	**32,130,950**	**25.03%**	58.26 GB	0.02%
<16KB	**30,171,925**	**23.51%**	250.01 GB	0.07%
<128KB	**27,957,924**	**21.78%**	1.24 TB	0.38%
<1MB	11,997,461	9.35%	4.11 TB	1.26%
<8MB	4,578,913	3.57%	12.85 TB	3.94%
<100MB	1,788,632	1.39%	46.06 TB	14.13%
<1GB	348,598	0.27%	88.93 TB	27.29%
<10GB	50,257	0.04%	103.10 TB	31.64%
<100GB	2,622	0.00%	58.16 TB	17.85%
<1TB	57	0.00%	11.08 TB	3.40%
1TB+		0.00%	0 b	0.00%

Total size	151,649,121	325.89 TB	325.27 TB allocated
Total # Files	128,357,631	325.84 TB	325.24 TB allocated
Total # Dirs	13,410,454	60.50 GB	21.72 GB allocated
Total # Links	9,881,036	651.69 MB	0 b allocated

Current file distribution: User home directories

Size	# Files	% of Files	Total Size	% of Size
0B	1,161,823	1.10%	0 b	0.00%
<512B	22,152,372	21.00%	3.63 GB	0.02%
<4KB	**36,449,737**	**34.55%**	64.55 GB	0.30%
<16KB	**24,525,463**	**23.24%**	193.85 GB	0.91%
<128KB	**16,385,084**	**15.53%**	679.69 GB	3.20%
<1MB	3,750,761	3.55%	1.14 TB	5.49%
<8MB	836,203	0.79%	2.18 TB	10.51%
<100MB	223,561	0.21%	**5.15 TB**	**24.85%**
<1GB	22,576	0.02%	**5.70 TB**	**27.49%**
<10GB	1,984	0.00%	**4.49 TB**	**21.68%**
<100GB	67	0.00%	1.08 TB	5.23%
<1TB		0.00%	0 b	0.00%
1TB+		0.00%	0 b	0.00%

Total size	126,588,211	20.72 TB	38.82 TB allocated
Total # Files	105,509,631	20.65 TB	38.81 TB allocated
Total # Dirs	15,247,857	65.45 GB	16.35 GB allocated
Total # Links	5,830,723	305.59 MB	0 b allocated

Current file distribution: Project spaces

Size	# Files	% of Files	Total Size	% of Size
0 B	1,300,752	0.44%	0 b	0.00%
< 512 B	10,586,176	3.57%	2.11 GB	0.00%
< 4 KB	111,584,548	3.57%	193.01 GB	0.00%
< 16 KB	18,030,519	6.09%	152.74 GB	0.00%
< 128 KB	**79,585,287**	**26.87%**	2.46 TB	0.04%
< 1 MB	21,405,926	7.23%	7.69 TB	0.12%
< 8 MB	25,316,460	8.55%	76.36 TB	1.17%
< 100 MB	20,710,144	6.99%	787.72 TB	12.06%
< 1 GB	6,400,388	2.16%	**1.6892 PB**	26.47%
< 10 GB	1,254,344	0.42%	**2.5996 PB**	40.74%
< 100 GB	51,300	0.02%	**999.7 TB**	15.30%
< 1 TB	1,062	0.00%	198.37 TB	3.04%
1 TB+	23	0.00%	69.06 TB	1.06%

Total size	314,585,747	6.3803 PB	6.4094 PB allocated
Total # Files	296,226,929	6.3803 PB	6.4094 PB allocated
Total # Dirs	7,076,159	59.09 PB	66.57 GB allocated
Total # Links	11,282,659	1.05 PB	0 b allocated

Scratch and Flash purge statistics

Scratch purge stats:

- 2018 data is a mix of scratch2 and fs1/scratch)
  - Files so far in 2018: 80,978,786
  - Data so far in 2018: 2,247,479,460,032 (2.0 PB)
    - low because nothing got purged off glade2/scratch2 for a few months

- 2017 data is a mix of the old glade_scratch and scratch2
  - Files in 2017: 724,608,993
  - Data in 2017: 1,6719,388,844,928 (15.2 PB)

Flash purge stats:

- Files purged so far in 2018: 377,898
- Data purged so far in 2018: 625,900,251,920 (582.92 TB)

Top 25 file types: Scratch

# Files	Type	%	# Files	Type	%
137,501,463	**.nc**	**33.60%**	3,895,146	.optrpt	0.95%
48,264,079	.snd	11.79%	3,777,610	.silo	0.92%
14,751,444	.CHANOBS_DOMAIN1	3.60%	1,825,510	.bin	0.45%
13,407,862	.txt	3.28%	1,735,901	0	0.42%
13,318,810	.CHRTOUT_DOMAIN1	3.25%	1,722,213	.xy	0.42%
9,742,989	.gz	2.38%	1,697,002	.grib2	0.41%
9,361,260	.LDASIN_DOMAIN1	2.29%	1,691,760	0.1	0.41%
9,183,164	.png	2.24%	1,667,669	.TS	0.41%
9,029,047	.mod	2.21%	1,456,316	.dat	0.36%
7,267,003	.o	1.78%	1,331,339	.inc	0.33%
6,367,592	.LAKEOUT_DOMAIN1	1.56%	1,260,178	.data	0.31%
5,771,426	.vtu	1.41%	1,204,073	.grb2	0.29%
5,382,260	.3520_nc	1.32%			

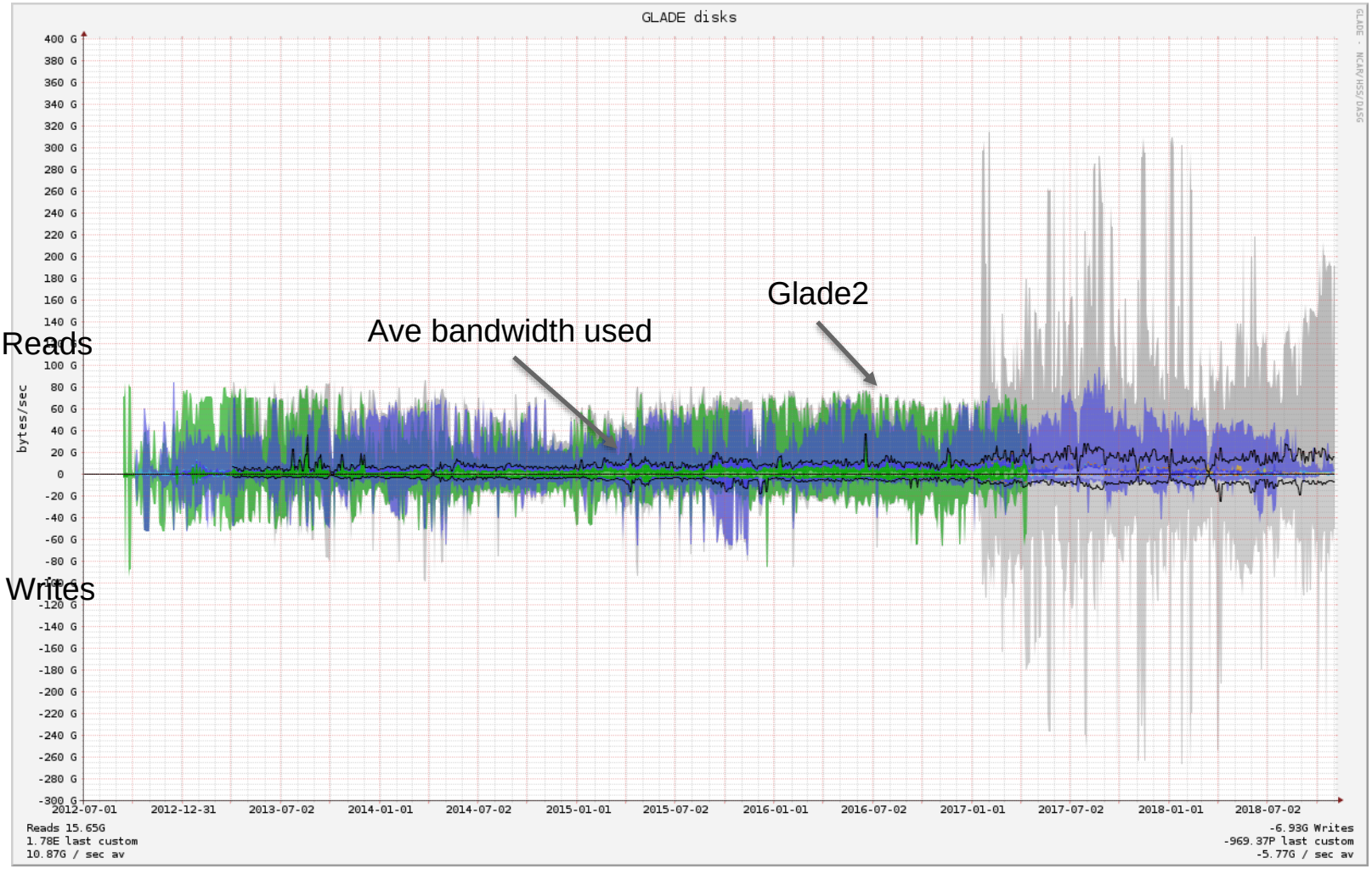
Top 25 file types: Project space

# Files	Type	%	# Files	Type	%
38,573,673	**.txt**	**12.28%**	1,636,277	.o	0.52%
34,568,913	**.nc**	**11%**	1,201,149	.2640_nc	0.38%
20,233,666	.gz	6.44%	1,045,327	.mod	0.33%
6,805,719	.CHRTOUT_DOMAIN1	2.17%	947,770	.LAKEOUT_DOMAIN1	0.30%
6,794,697	.LDASIN_DOMAIN1	2.16%	756,374	.h	0.24%
6,185,881	.3520_nc	1.97%	689,865	.f90	0.22%
6,059,155	.svn-base	1.93%	678,318	.nc4	0.22%
4,341,467	.CHANOBS_DOMAIN1	1.38%	662,682	.f	0.21%
3,170,301	.png	1.01%	657,678	.c	0.21%
2,160,150	.grb	0.69%	636,302	.inc	0.20%
1,980,806	.F90	0.63%	607,447	.html	0.21%
1,729,203	.bin	0.55%	601,816	.GWOUT_DOMAIN1	0.22%
1,655,377	.dat	0.53%			

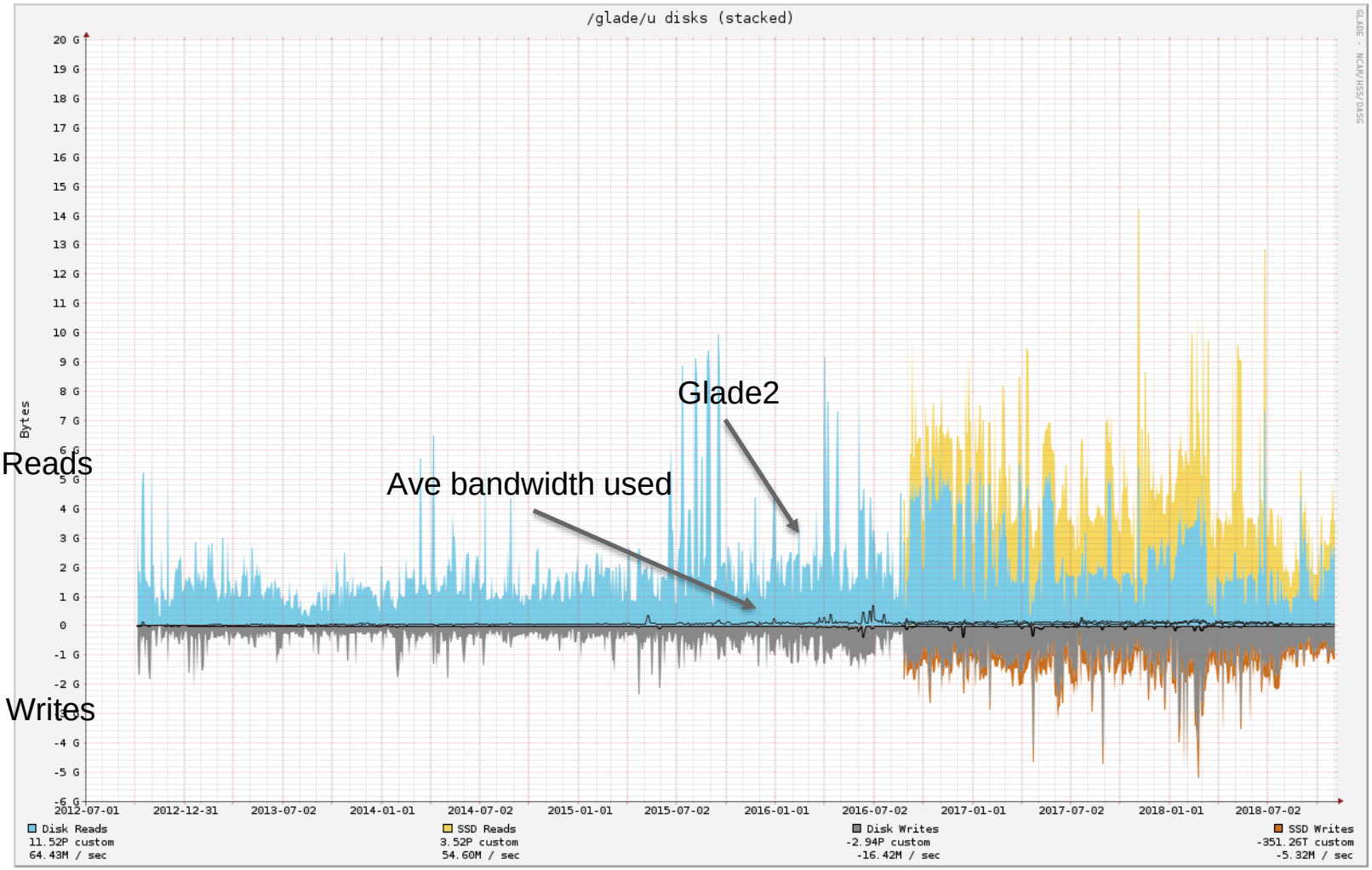
Top 25 file types: User work space

# Files	Type	%	# Files	Type	%
16,072,270	**.nc**	**10.69%**	1,728,250	.CHANOBS_DOMAIN1	1.15%
11,981,382	.svn-base	7.97%	1,717,899	.f	1.14%
8,293,219	.gz	5.52%	1,668,536	.h	1.11%
8,111,727	.LDASIN_DOMAIN1	5.40%	1,651,495	.c	1.10%
4,902,243	.F90	3.26%	1,571,004	.sav	1.05%
4,240,555	.txt	2.82%	1,511,560	.xml	1.01%
3,947,097	.o	2.63%	1,493,292	.dat	0.99%
3,648,052	.png	2.43%	1,490,321	.pyc	0.99%
2,821,208	.F	1.88%	1,339,240	.2640_nc	0.89%
2,644,600	.inc	1.76%	1,125,195	.csh	0.75%
2,619,960	.py	1.74%	1,117,575	.csv	0.74%
2,124,497	.mod	1.41%	865,520	.html	0.58%
2,081,540	.f90	1.38%			

GLADE performance over past six years



User home directory performance over the past six years



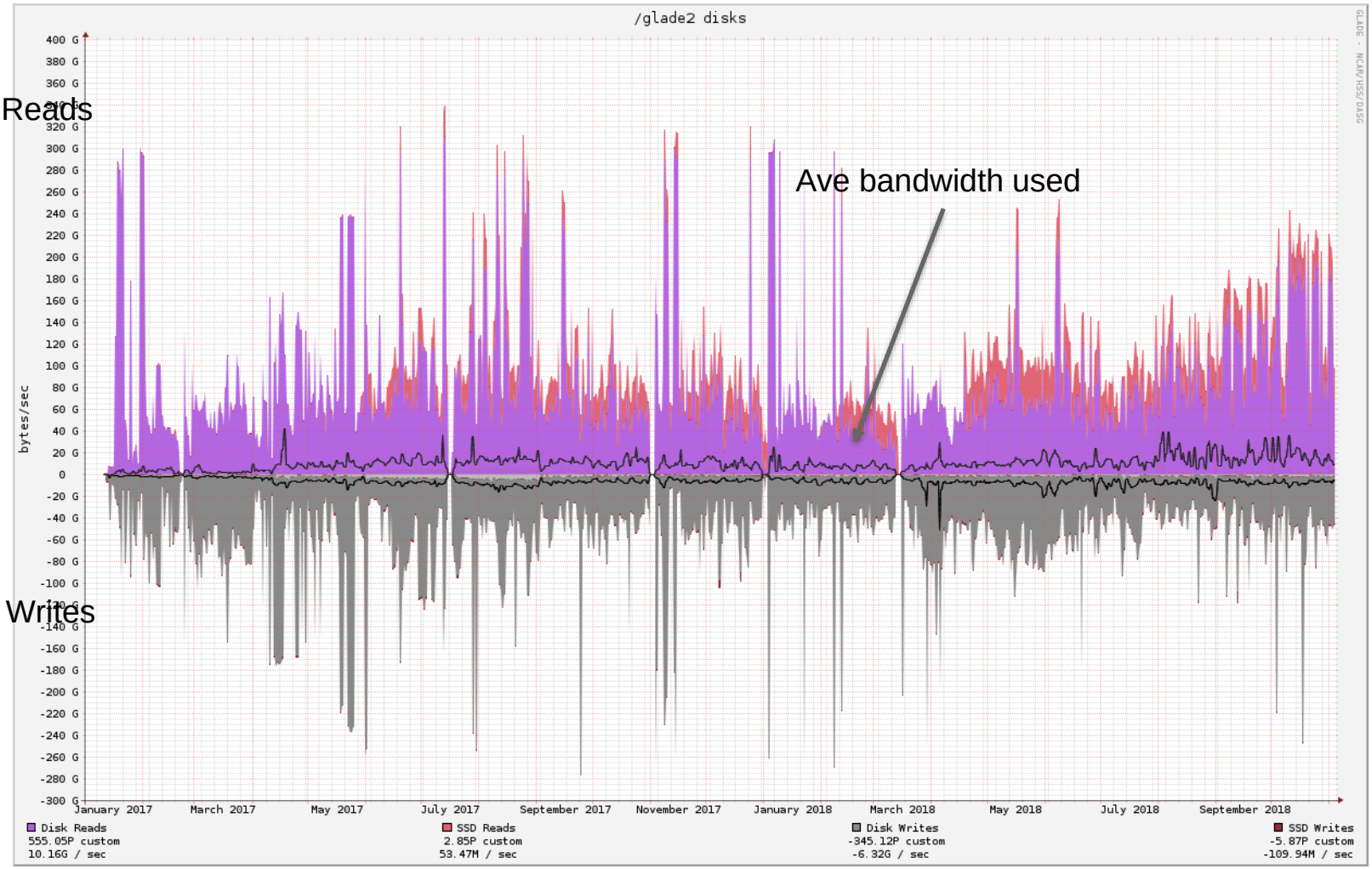
Glade2

Ave bandwidth used

Reads

Writes

GLADE performance over the past two years



For more information about data contained in this presentation, please send e-mail to:

cislhelp@ucar.edu