Trends in Small HPC Center Management

Birds of a Feather Session at Supercomputing 13

November 19, 2013

Notes by David Stack
david@uwm.edu

Before the SC13 conference, a survey was distributed via email and social media to solicit information that would inform the discussion at the Birds of a Feather session.

The survey questions were very similar to those of the previous two years so that trends could be identified.

A discussion took place on Twitter as to what the term "small HPC" actually means, i.e., is it a small computational resource or is it mean a small team regardless of the size of the resources they oversee?

The question arose because the smallest category for machine size on the pre-conference survey was <1,000 CPUs. Some people were reporting that they were getting a lot of science with clusters on the order of 100 CPUs.

There were 63 responses to the survey, which was similar to 2012. The response rate in 2011 was about half of that.

Please see the detailed survey results (below) for additional information.

Comparisons from 2011-2013

GPUs

The number of respondents who report that their cluster has no GPUs has trended steadily downward over the years, i.e., 61%, 51% and 32%. Attendees reported that GPUs were mostly used for teaching and for running pre-packaged codes, especially for molecular dynamics. Not many researchers are porting their existing codes to run on GPUs.

CPU Cores

The number of CPU cores per HPC cluster has trended steadily upwards. In 2011, approximately 11% reported clusters of 5,000 cores or larger. By 2013, that number had risen to 23%.

Memory Per Node

Similarly, the amount of memory per physical server, i.e., per compute node, has also trended upwards over the last three years. In 2011, only 37% of clusters were reported as having nodes with more than 64 Gb. In 2013, 73% of machines had such nodes. BoF attendees recommended 4 Gb per core as a base line that did not cost significantly more than 2 GB per core.

Low Latency Network

Compared to 2011, the use of Infiniband for low latency networking was down somewhat, i.e., 61% as compared to 74%. In 2013, the type of Infiniband was broken out according to QDR (33%), FDR (22%) and DDR (6%). This was the first year that there was an "other" category for low latency networking which garnered 10% of the responses and may have resulted in the overall percentage for Infiniband being lower than in previous years.

Schedulers

TORQUE and SLURM appear to be slowly increasing in popularity. The change in scheduler usage is not precipitous because change is resisted by the user population. Recent versions of TORQUE were described as unreliable. Difficulties with managing GPUs with TORQUE were also noted. LSF usage has dropped significantly since IBM took it over with resulting changes in pricing and support. A couple attendees reporting writing scheduler wrappers for research groups such as chemists running Gaussian.

Serial Jobs

There was a slight decrease in the percentage of respondents who said they allowed serial jobs on their cluster, i.e., 90% in 2013 versus 95% in 2011. Conversely, the amount of memory allowed per serial job increased. In 2011, only about 7% of serial jobs could use 24 Gb or more. That number has risen to 34% in 2013. Some of this may simply be the result of cluster configurations getting larger and sysadmins not imposing any limits on the sizes of serial jobs other than the size of the hardware itself.

Home Directories

Some 33% of clusters have home directories on a parallel file system, which is comparable to the previous two years.

Scratch Storage

64% of clusters have high performance scratch storage on a parallel file system compared with 100% in 2011. There were contrasting increases in the percentages reporting that scratch was handled via local disk on the nodes (38%), NFS (31%) and various other schemes (13%).

Medium Performance Storage

In 2011, 72% of the respondents reported that they did not have a medium performance storage system for holding working files and the like. By 2013, that number was down to 33% which probably reflects increased expectations from new user groups that are less technically savvy.

Backup

In 2013, 75% reported backing up users home directories as compared to only 53% in 2011. Similarly, the backup of scratch storage was up to 13% as compared to 5% in 2011. Again, this probably reflects expectations and demands from new user groups. The attendees noted that it isn't feasible to do backups at the Petabyte scale and the decision of whether to do backups may depend upon where the users' home directories are located. There is also a distinction between back ups for disaster recovery purposes versus backing up so that individual files can be recovered if necessary.

Solid State Disks

For the first time, a question was asked as to whether or not solid state disks were used on the clusters. 25% responded affirmatively as follows:

- 15% local storage on compute nodes
- 3% home directories
- 13% high performance / scratch space
- 2% medium performance / online storage

Sharing

Another new question for 2013 was whether the cluster was shared among multiple departments at the institution. 92% responded affirmatively. The attendees noted that there may have been some selection bias influencing the responses because those who run clusters for "only" a single department might be been less like to see themselves as an HPC "center" and therefore less likely to respond. It is hard to argue against sharing a cluster with other departments if it is not 100% used. There are various local models for condo clusters and the loaning or reserving of nodes for specific research groups.

36% of respondents reported sharing a cluster with other institutions. The discussion included the various practices for onboarding researchers from other institutions. Some institutions push external collaborators through an institutional guest account process before letting them on the cluster. Others simply create and ID and password on the cluster itself. It is frustrating when a principal investigator says a collaborator needs to be given an account and it subsequently becomes clear that the person needs considerable support.

Biggest Struggles

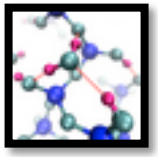
Other than funding, survey respondents reported that their biggest struggles were with user support and staffing. These were seen as two sides of the same coin because user support is time consuming. Central institution help desks can do little to support HPC users.

Trends in Small HPC Center Management

Birds of a Feather Session at SC13 Denver, CO

Participate in annual survey of small #HPC cluster config's tinyurl.com/smallhpc13 Results on web & #SC13

Derekgottlieb Wow. Definition of small is larger than I'd expect. <1000 cores is smallest bin?



Glennklockwood That surprised me too. "small HPC center" >> "small HPC cluster" I guess

Trends in Small HPC Center Management

Birds of a Feather Session at SC13 Denver, CO

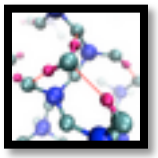
HPC_Guru Agree with both Derek & Glenn - maybe @davidstack can comment. #HPC

davidstack 8% of 2012 #HPC Survey respondents were <1,000 cores Expect fewer in 2013 ow.ly/qoP6k

kehoste Doesn't surprise me. Can you still call it HPC nowadays if you only have a couple 100s?

Trends in Small HPC Center Management

Birds of a Feather Session at SC13 Denver, CO



glennklockwood YES. Depends on the scientific output. Most of our #hpc jobs are < 100 cores.

derekgottlieb There's a lot of science done on our clusters using 16 or fewer cores per job.

derekgottlieb I know Unis with central HPC clusters grown over 5 yrs that are still <1k cores total.

Trends in Small HPC Center Management

Birds of a Feather Session at SC13 Denver, CO

kehoste Then I wonder how much HPC people are doing on their new shiny multi-core laptops...

ajdecon "HPC" is a wonderful moving target, used for techniques as much as scale.

ajdecon I've been using "technical" and/or "cluster" computing more often with laymen.

GPUs

Birds of a Feather Session at SC13 Denver, CO

How many GPU cores does this cluster have?

Answer	2011	2012	2013
None	61%	51%	32%
1-999	33%	29%	32%
1,000 - 4,999	6%	6%	11%
5,000 - 9,999	0%	2%	11%
10,000 - 24,999	0%	2%	3%
25,000 or more	0%	10%	11%

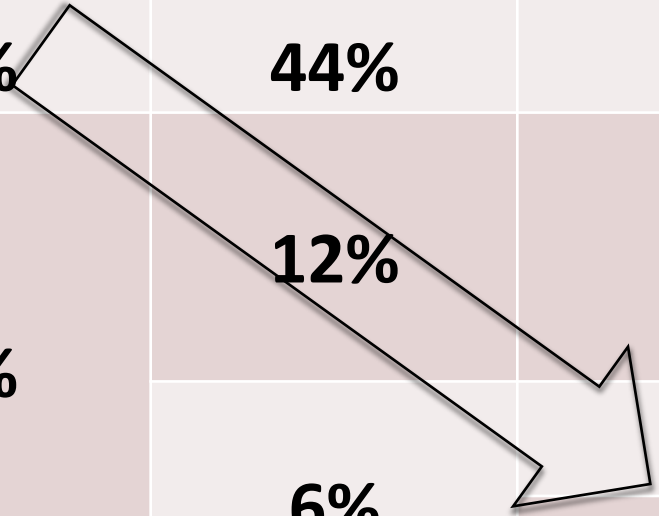
Decreasing trend in those replying “None”

Hardware Configuration

Birds of a Feather Session at SC13 Denver, CO

How many CPU cores are in this HPC cluster?

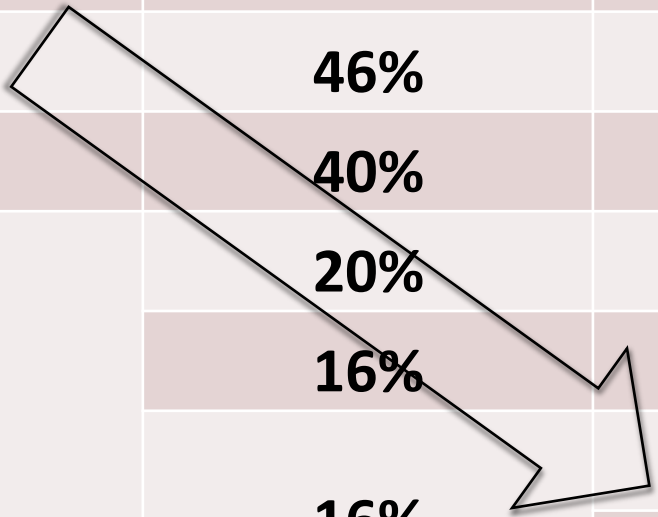
Answer	2011	2012	2013
< 1,000	37%	38%	27%
1,000-4,999	~52%	44%	49%
5,000- 9,999	~11%	12%	14%
10,000 - 14,999		6%	6%
15,000 +			3%



Trend toward greater numbers of cores

How much memory is there per physical server, i.e., per compute node?

Answer	2011	2012	2013
0 -16 Gb	48%	30%	22%
17 - 32 Gb	32%	46%	43%
33 - 64 Gb	58%	40%	49%
65 – 128 Gb	37%	20%	30%
129 – 256 Gb		16%	22%
257 – 512 Gb		16%	11%
> 512 Gb			10%
Unsure	5%	0%	0%



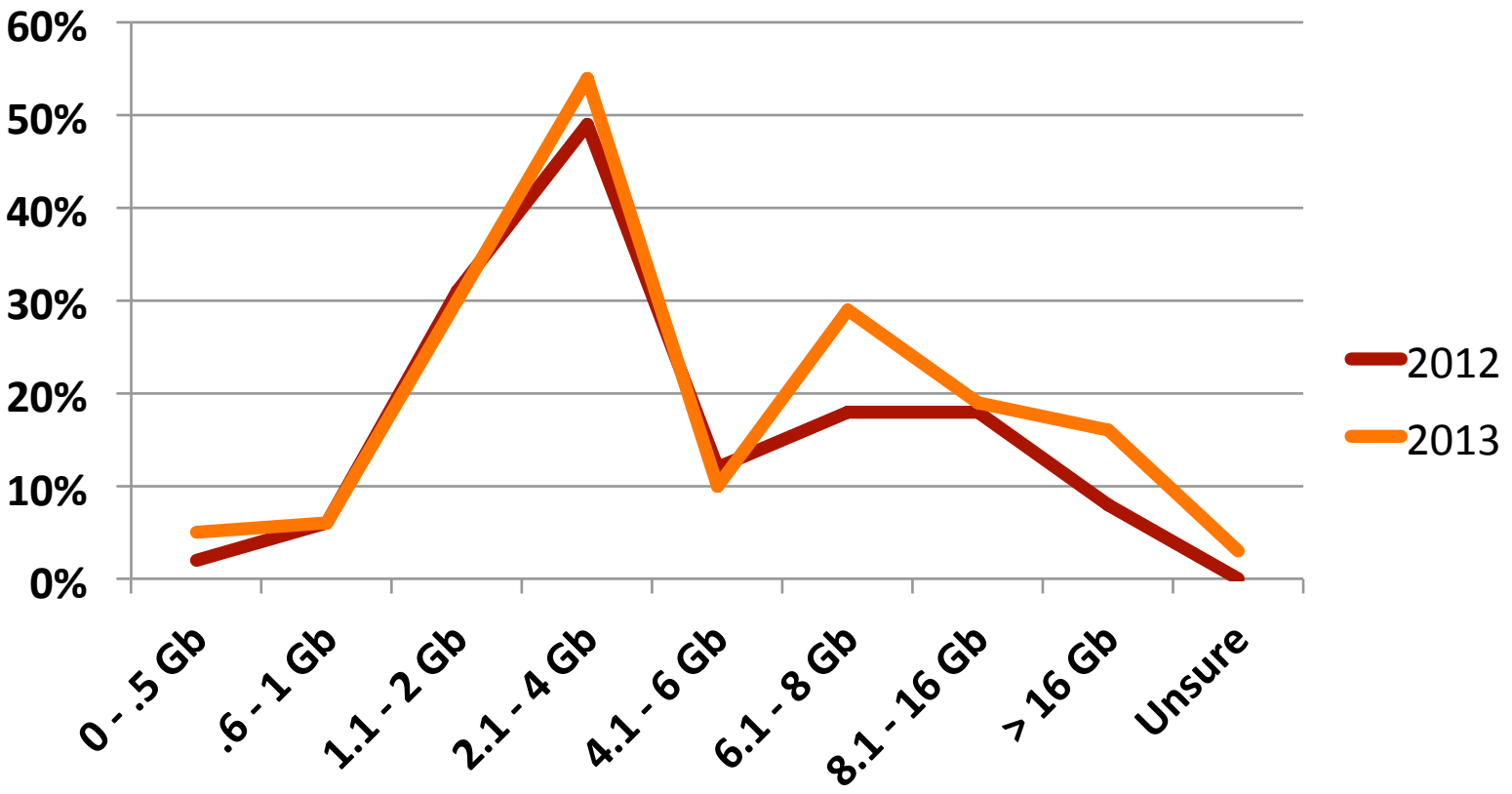
Trend toward greater memory per compute node

How much memory is there per core?

Answer	2012	2013
0 - .5 Gb	2%	5%
.6 - 1 Gb	6%	6%
1.1 - 2 Gb	31%	30%
2.1 - 4 Gb	49%	54%
4.1 - 6 Gb	12%	10%
6.1 - 8 Gb	18%	29%
8.1 - 16 Gb	18%	19%
> 16 Gb	8%	16%
Unsure	0%	3%

Long tails toward the high side

How much memory is there per core?



What low latency network for MPI communication among CPU cores?

Answer	2011	2012	2013
Infiniband QDR	74%	75%	33%
Infiniband FDR			22%
Infiniband DDR			6%
10 Gigabit Ethernet, 10 GbE	11%	12%	13%
1 Gigabit Ethernet, 1 GbE	11%	12%	16%
Other	---	---	10%
Unsure	5%	1%	0%

Trend is flat

What low latency network for MPI communication among CPU cores?

Answer	2013
Other:	**10%**
• Mix of infiniband QDR and DDR • Cray Gemini • Proprietary IBM Blue Gene/P networking • BlueGene/P proprietary interconnect • Both infiniband QDR, FDR • NUMAlink 5	

Scheduler Questions

Birds of a Feather Session at SC13 Denver, CO

What is the scheduler? 1 of 3

Answer	2011	2012	2013
TORQUE	33%	42%	---
TORQUE and Maui			16%
TORQUE and MOAB			30%
Maui/MOAB	39%	44%	---
SLURM	0%	14%	17%

46%

TORQUE and SLURM slowly increasing

What is the scheduler? 2 of 3

Answer	2011	2012	2013
PBS	6%	4%	3%
SGE (Oracle/Sun Grid Engine)	28%	22%	22%
LSF	33%	6%	10%
Lava	0%	2%	0%
Condor	6%	0%	5%
Other	0%	8%	10%

LSF is the oddball

What is the scheduler? 3 of 3

Answer	2013
Other:	**10%**
• LoadLeveler (2) • Univa GE • PBS Pro (3)	

Do you allow serial jobs on this cluster?

Answer	2011	2012	2013
Yes	95%	96%	90%
No	5%	4%	10%

Slight decrease in serial jobs in 2013

What is the maximum amount of memory allowed per serial job?

Answer	2011	2012	2013
No maximum enforced	72%	65%	63%
<= 16 GB	~ 12%	9%	4%
17 - 24 GB	~ 0%	4%	0%
More than 24 GB	~ 7%	22%	34%

Trend toward more memory allowed per serial job

What is the maximum amount of memory allowed per multi-core (mp or mpi) job?

Answer	2011	2012	2013
No maximum enforced	74%	59%	67%
16 GB or less	5%	0%	2%
17 - 32 GB	~ 0%	9%	2%
33 - 48 GB	~ 11%	0%	0%
49 – 64 GB	11%	32%	3%
More than 64 GB			26%

Difficult to see any trend

Storage

Birds of a Feather Session at SC13 Denver, CO

Where do users' home directories reside?

Answer	2011	2012	2013
Local disk	0%	0%	5%
NFS	47%	56%	46%
Parallel file system	47%	31%	---
Lustre	---	---	2%
IBM GPFS	---	---	16%
GFS	---	---	2%
ZFS	---	---	5%
PanFS	---	---	8%
Unsure	0%	2%	0%
Other	5%	10%	16%

33%

Where do users' home directories reside?

Answer	2013
Other:	**16%**
• Direct attached storage node • Gluster (2) • AFS • panfs and NFS • Partly NFS partly AFS • DDN with GPFS (3) • IBRIX	

What type of high performance storage/scratch space?

Answer	2011	2012	2013
Local disk on nodes	26%	35%	38%
NFS	16%	33%	31%
Parallel file system	100%	67%	---
Lustre	---	---	13%
IBM GPFS	---	---	28%
GFS	---	---	3%
ZFS	---	---	5%
PanFS	---	---	15%
Unsure	0%	0%	0%
Other	0%	19%	13%

64%

What type of high performance storage/scratch space?

Answer	2013
Other:	**13%**
• Direct attached storage node (via Infiniband QDR) • Gluster (2) • Isilon X • DDN with GPFS (3) • IBRIX	

Do you have an online, medium performance data storage service?

Answer	2011	2012	2013
No	72%	44%	33%
Yes	28%	56%	
Local disk on nodes			0%
NFS			30%
Lustre			0%
PanFS			0%
IBM GPFS			13%
GFS			2%
ZFS			8%
Other			15%

Trend is increasing

68%

Do you have an online, medium performance data storage service?

Answer	2013
Other:	**15%**
• SAM/QFS (2) • AFS • Isilon NL • AFS • NFS mount of a GPFS filesystem • DDN with GPFS (3)	

Which of the following storage environments on this cluster do you back up?

Answer	2011	2012	2013
Home directories	53%	67%	75%
High performance / scratch space	5%	8%	13%
Medium performance, online storage	11%	40%	28%
None	47%	29%	18%
Unsure	0%	2%	0%

Trend toward increasing backup of home directories

Do you have solid state disks on this cluster?

Answer	2013
No	75%
Local Storage on Compute Nodes	15%
Home Directories	3%
High Performance / Scratch Space	13%
Medium Performance / Online Storage	2%

Sharing

Birds of a Feather Session at SC13 Denver, CO

Is this cluster shared among multiple departments at your institution?

Answer	2013
Yes	92%
No	8%

Is this cluster shared among multiple institutions?

Answer	2013
Yes	36%
No	64%

Other than funding, what is your biggest struggle?

Answer (N>2)	2013
User Support	17
Staffing	9
Scheduling / Prioritization	6
Storage	6
Applications	6
Space	5
Power	4
Legacy Hardware	3

Other than funding, what is your biggest struggle?

Answer (N>2)	2013
User Support	17
Staffing	9
Scheduling / Prioritization	6
Storage	6
Applications	6
Space	5
Power	4
Legacy Hardware	3

Two sides of the same coin?

Other than funding, what is your biggest struggle?

Answer	2013
User Support:	17
- User training / user attitude and ignorance	
- Avoiding User Wars	
- Job conflicts across groups – esp. when cluster is busy	
- Bringing new users up to speed from PCs to clusters	
- No dedicated staff for this	
- Need both Comp Sci and scientific domain expertise	

Small HPC BoFs Contact Information

2013 Survey	**tinyurl.com/smallhpc13**
Website	**https://sites.google.com/site/smallhpc/**
Email List	**See link at above website**
Beth Anderson	**beth.anderson@intel.com**
David Stack	**david@uwm.edu**

Legacy slides from the 2012 BoF

Current Directions (2012)

If you were buying new compute nodes today, how many cores per node?

2011	%
4	0%
8	5%
12	21%
16	32%
> 16	16%
Unsure	26%

2012	%
4	0%
8	9%
12	5%
16	50%
24	0%
32	14%
Unsure	14%
Other	9%

If you were buying new compute nodes today, how much memory per node?

2011	%
0-8 GB	0%
9-16 GB	0%
17-24 GB	0%
25-48 GB	41%
>48 GB	35%
Unsure	24%

2012	%
0-8 GB	8%
9-16 GB	4%
17-24 GB	2%
25-48 GB	27%
49-64 GB	19%
More than 64 GB	33%
Unsure	6%

Staffing (2012)

How many different individuals, excl. students, involved in operation, support, development?

Answer	2012
1 individual	15%
2-3 individuals	50%
4-5 individuals	21%
6-8 individuals	13%
9-10 individuals	2%
11-15 individuals	0%
More than 15 individuals	0%

Approximately how many FTE, incl. students, operate the cluster to maintain the status quo (excluding user support)?

Answer	2012
< 1 FTE	31%
1.1 – 2 FTE	42%
2.1 – 4 FTE	23%
4.1 – 6 FTE	4%
6.1 – 8 FTE	0%
More than 8 FTE	0%

Approximately how many FTE, incl. students, support users of the cluster?

Answer	2012
< 1 FTE	40%
1.1 – 2 FTE	27%
2.1 – 4 FTE	27%
4.1 – 6 FTE	4%
6.1 – 8 FTE	2%
More than 8 FTE	0%

Approximately how many FTE, incl. students, are involved in hardware/software development efforts related to the cluster?

Answer	2012
< 1 FTE	52%
1.1 – 2 FTE	30%
2.1 – 4 FTE	15%
4.1 – 6 FTE	2%
6.1 – 8 FTE	0%
More than 8 FTE	0%

Inward facing staff versus outward facing staff?

Answer	2012
There is a **clear separation**.	10%
There is **some separation**.	31%
There is **almost no separation**.	58%

Small HPC BoFs Contact Information

2012 Survey	**tinyurl.com/smallHPC**
Website	**https://sites.google.com/site/smallhpc/**
Email List	**See link at above website**
Roger Bielefeld	**roger.bielefeld@cwru.edu**
David Stack	**david@uwm.edu**