UNIVERSITY OF BIRMINGHAM

Birmingham Status and Plans

HEPSYSMAN Meeting, 14th June, 2017

Mark Slater, Birmingham University

UNIVERSITY OF BIRMINGHAM

Current Status - Tier 2

As of right now, the Birmingham Grid Site consists of:

- ~1500 Cores providing ~17K HS06
- 980TB Storage with another 200TB being prepared

Cluster management is done with a global (T2+T3) Puppet/Foreman instance.
Current running services include:

- Torque/CREAM batch system
- ALICE Storage on XRootD install
- All other storage on DPM
- Squid, BDII, APEL, ARGUS, VO Box
- 10 Gb/s link (soon to be 20 Gb/s)

The divide between experiments is ALICE 60%, ATLAS 30%, LHCb 5%, Other 5%

Current Status - Tier 3

The current state of our local (Tier 3) systems is shown below:

- Batch Cluster Farm, 8 nodes, 32 (logical) cores, 48GB per node
- 180TB 'New' Storage + 160TB 'Older' Storage + 160TB 'Oldest' Storage
- ~80 (mostly) Fedora 24 Desktops
- SL6 image (access through chroot)
- Two F24 login nodes
- Two Web servers
- DHCP, Mail and LDAP servers
- Share 1 10Gb/s link with the university

UNIVERSITY OF BIRMINGHAM

Operational Status

The biggest change in recent months has been the loss of Matt Williams

There were a few teething problems during autumn due to this, mostly because of teaching commitments

However, thanks to improvements in monitoring and changes to some services, things are on a more even keel now!

With this in mind, I have been moving forward with plans for the future taking into account the reduced manpower:

- Making progress on server room rearrangement
- Integrating all monitoring into Grafana
- Switching to using VAC on Grid
- Switching to using ZFS for the storage
- Shifting all Tier3 storage to MooseFS

In recent weeks I have even managed to do some Ganga work!

Switching to ZFS

After several discussions with people, I have decided to (gradually!) move all our Tier 2 storage from hardware RAID 6 to ZFS and not buy RAID6 cards for new storage

From my point of this has several benefits:

- Easy to monitor disk health across all systems
- Can use ~any disks in the RAID
- Cheaper to buy new hardware

I have currently moved 100TB (prev. 40TB) of storage over to ZFS on Tier2 and have had no issues at present...

UNIVERSITY OF BIRMINGHAM

MooseFS on Tier3 Storage

Since I took over 3 (4?) years ago, I have been attempting to find a good distributed solution for our Tier 3 storage to move away from basic NFS mounted RAIDs

I originally tried lustre but had no end of trouble trying to keep it working with the modern kernels that Fedora ships with

I eventually found 'MooseFS' which offered everything I needed:

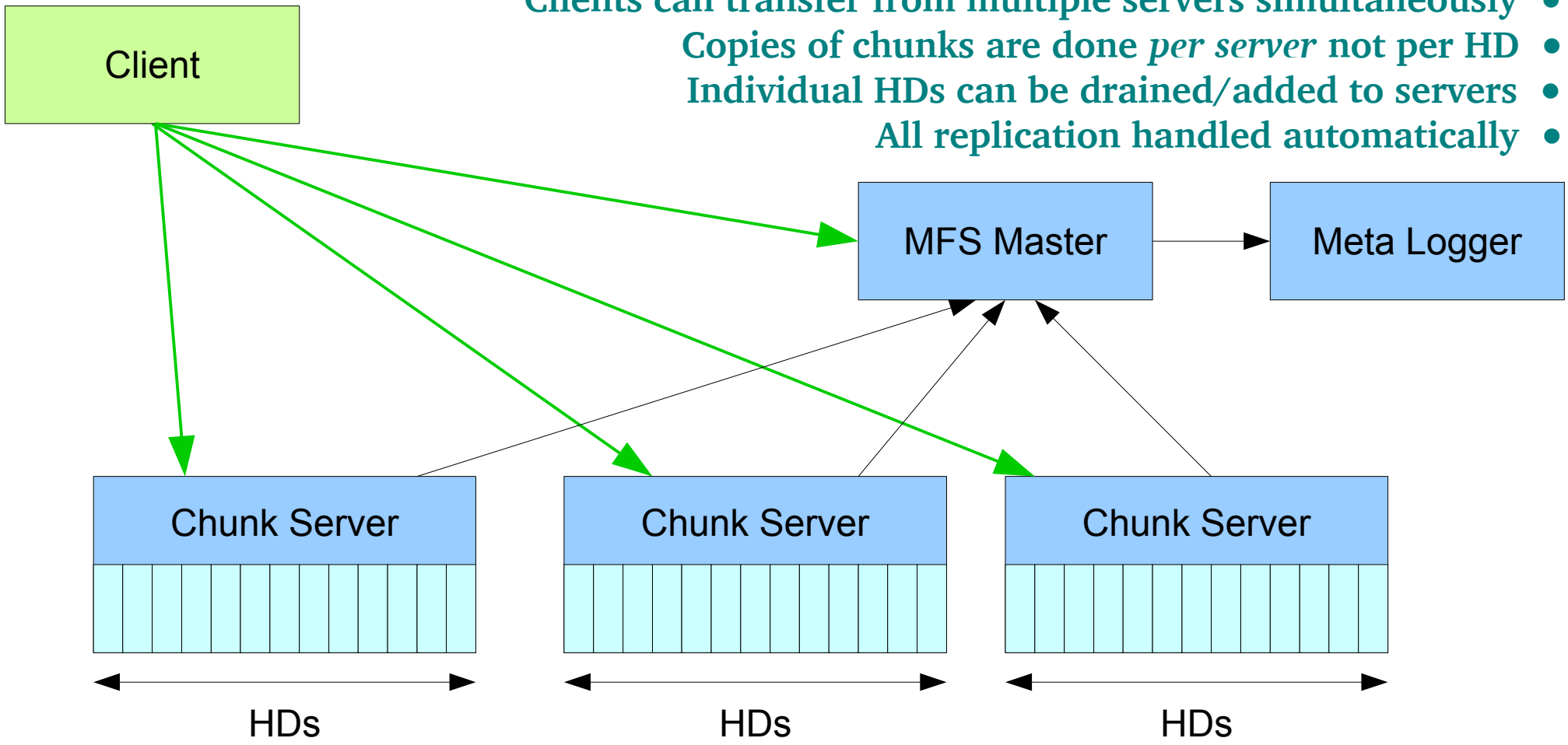
- Very easy to setup and administer
- Very configurable
- Redundancy built in
- All done through fuse
- Can keep using disks until they die

We currently have ~470TB giving (with my setup) 235TB usable space

MooseFS on Tier3 Storage

Moose works by dividing every file in to 'chunks' and then copying these chunks to the 'chunk servers' given the appropriate policy:

- Can have 1+ copies of each chunk
- Clients can transfer from multiple servers simultaneously
- Copies of chunks are done *per server* not per HD
- Individual HDs can be drained/added to servers
- All replication handled automatically

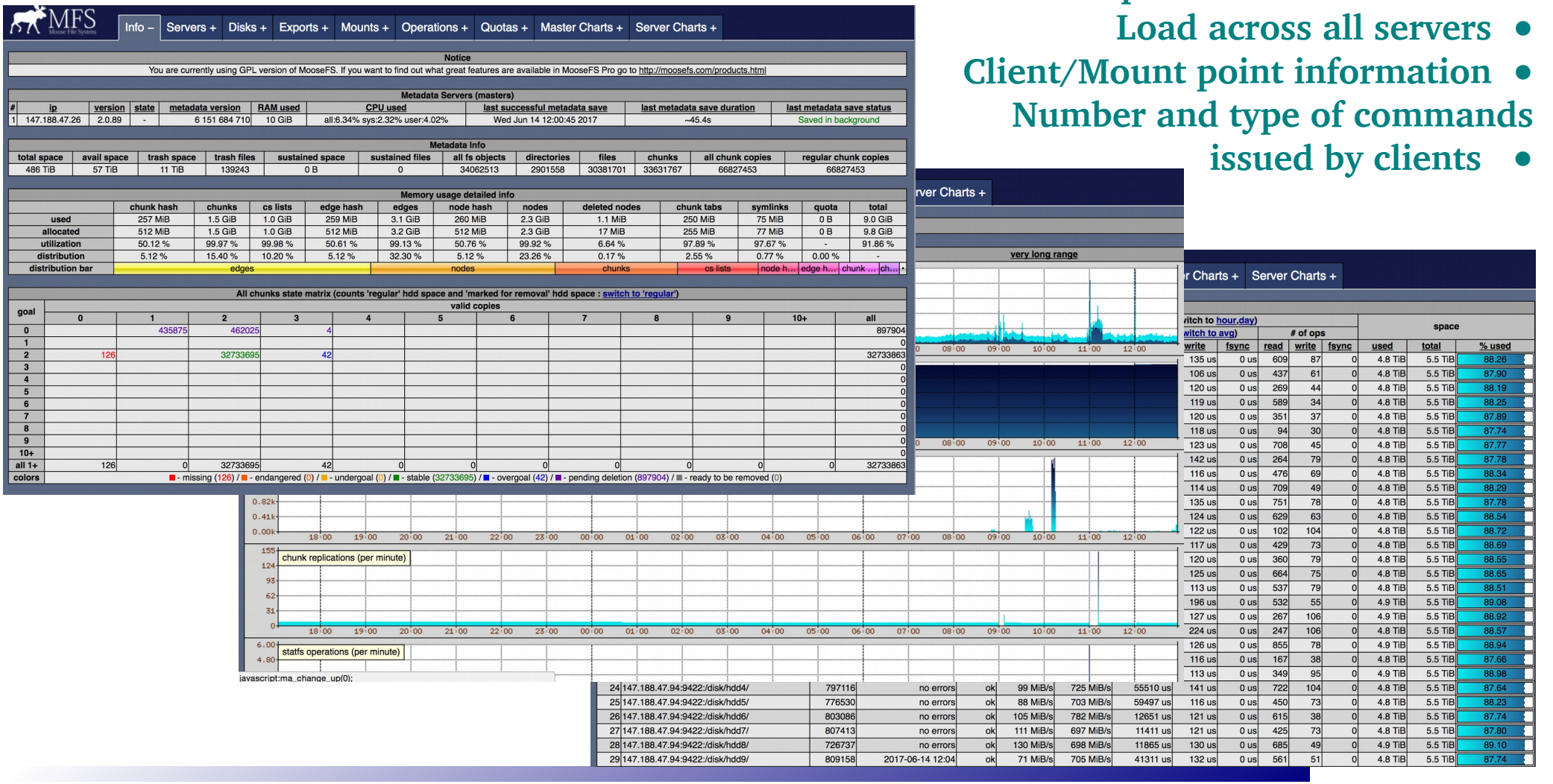


UNIVERSITY OF BIRMINGHAM

MooseFS on Tier3 Storage

Monitoring is handled through a very good web interface that gives:

- Overview of chunk replication progress (e.g. under/over goal)
- Individual disk space across all servers
- Load across all servers
- Client/Mount point information
- Number and type of commands issued by clients



MFS Moose File System

Info – | Servers + | Disks + | Exports + | Mounts + | Operations + | Quotas + | Master Charts + | Server Charts +

Notice

You are currently using GPL version of MooseFS. If you want to find out what great features are available in MooseFS Pro go to http://moosefs.com/products.html

Metadata Servers (masters)

#	ip	version	state	metadata version	RAM used	CPU used	last successful metadata save	last metadata save duration	last metadata save status
1	147.188.47.26	2.0.89	-	6 151 684 710	10 GiB	all:6.34% sys:2.32% user:4.02%	Wed Jun 14 12:00:45 2017	~45.4s	Saved in background

Metadata Info

total space	avail space	trash space	trash files	sustained space	sustained files	all fs objects	directories	files	chunks	all chunk copies	regular chunk copies
486 TiB	57 TiB	11 TiB	139243	0 B	0	34062513	2901558	30381701	33631767	66827453	66827453

Memory usage detailed info

	chunk hash	chunks	cs lists	edge hash	edges	node hash	nodes	deleted nodes	chunk tabs	symlinks	quota	total
used	257 MiB	1.5 GiB	1.0 GiB	259 MiB	3.1 GiB	260 MiB	2.3 GiB	1.1 MiB	250 MiB	75 MiB	0 B	9.0 GiB
allocated	512 MiB	1.5 GiB	1.0 GiB	512 MiB	3.2 GiB	512 MiB	2.3 GiB	17 MiB	255 MiB	77 MiB	0 B	9.8 GiB
utilization	50.12 %	99.97 %	99.98 %	50.61 %	99.13 %	50.76 %	99.92 %	6.64 %	97.89 %	97.67 %	-	91.86 %
distribution	5.12 %	15.40 %	10.20 %	5.12 %	32.30 %	5.12 %	23.26 %	0.17 %	2.55 %	0.77 %	0.00 %	-
distribution bar	edges				nodes			chunks	cs lists	node h…	edge h…	chunk … ch…

All chunks state matrix (counts 'regular' hdd space and 'marked for removal' hdd space : switch to 'regular')

goal	valid copies											
	0	1	2	3	4	5	6	7	8	9	10+	all
0		435875	462025	4								897904
1												0
2	126		32733695	42								32733863
3												0
4												0
5												0
6												0
7												0
8												0
9												0
10+												0
all 1+	126	0	32733695	42	0	0	0	0	0	0	0	32733863
colors	■ - missing (126) / ■ - endangered (0) / ■ - undergoal (0) / ■ - stable (32733695) / ■ - overgoal (42) / ■ - pending deletion (897904) / ■ - ready to be removed (0)											

rver Charts +

very long range

r Charts + | Server Charts +

itch to hour,day)
witch to avg)

space			# of ops					
used	total	% used	read	write	fsync	write	fsync	
4.8 TiB	5.5 TiB	88.26	609	87	0	135 us	0 us	
4.8 TiB	5.5 TiB	87.90	437	61	0	106 us	0 us	
4.8 TiB	5.5 TiB	88.19	269	44	0	120 us	0 us	
4.8 TiB	5.5 TiB	88.25	589	34	0	119 us	0 us	
4.8 TiB	5.5 TiB	87.89	351	37	0	120 us	0 us	
4.8 TiB	5.5 TiB	87.74	94	30	0	118 us	0 us	
4.8 TiB	5.5 TiB	87.77	708	45	0	123 us	0 us	
4.8 TiB	5.5 TiB	87.78	264	79	0	142 us	0 us	
4.8 TiB	5.5 TiB	88.34	476	69	0	116 us	0 us	
4.8 TiB	5.5 TiB	88.29	709	49	0	114 us	0 us	
4.8 TiB	5.5 TiB	87.78	751	78	0	135 us	0 us	
4.8 TiB	5.5 TiB	88.54	629	63	0	124 us	0 us	
4.8 TiB	5.5 TiB	88.72	102	104	0	122 us	0 us	
4.8 TiB	5.5 TiB	88.69	429	73	0	117 us	0 us	
4.8 TiB	5.5 TiB	88.55	360	79	0	120 us	0 us	
4.8 TiB	5.5 TiB	88.65	664	75	0	125 us	0 us	
4.8 TiB	5.5 TiB	88.51	537	79	0	113 us	0 us	
4.9 TiB	5.5 TiB	89.08	532	55	0	196 us	0 us	
4.9 TiB	5.5 TiB	88.92	267	106	0	127 us	0 us	
4.8 TiB	5.5 TiB	88.57	247	106	0	224 us	0 us	
4.9 TiB	5.5 TiB	88.94	855	78	0	126 us	0 us	
4.8 TiB	5.5 TiB	87.66	167	38	0	116 us	0 us	
4.9 TiB	5.5 TiB	88.98	349	95	0	113 us	0 us	
4.8 TiB	5.5 TiB	87.64	722	104	0	141 us	0 us	
4.8 TiB	5.5 TiB	88.23	450	73	0	116 us	0 us	
4.8 TiB	5.5 TiB	87.74	615	38	0	121 us	0 us	
4.8 TiB	5.5 TiB	87.80	425	73	0	121 us	0 us	
4.9 TiB	5.5 TiB	89.10	685	49	0	130 us	0 us	
4.9 TiB	5.5 TiB	87.74	561	51	0	132 us	0 us	

chunk replications (per minute)

statfs operations (per minute)

javascript:ma_change_up(0);

24	147.188.47.94:9422:/disk/hdd4/	797116	no errors	ok	99 MiB/s	725 MiB/s	55510 us	
25	147.188.47.94:9422:/disk/hdd5/	776530	no errors	ok	88 MiB/s	703 MiB/s	59497 us	
26	147.188.47.94:9422:/disk/hdd6/	803086	no errors	ok	105 MiB/s	782 MiB/s	12651 us	
27	147.188.47.94:9422:/disk/hdd7/	807413	no errors	ok	111 MiB/s	697 MiB/s	11411 us	
28	147.188.47.94:9422:/disk/hdd8/	726737	no errors	ok	130 MiB/s	698 MiB/s	11865 us	
29	147.188.47.94:9422:/disk/hdd9/	809158	2017-06-14 12:04	ok	71 MiB/s	705 MiB/s	41311 us	

UNIVERSITY OF BIRMINGHAM

Making Monitoring Easier

Before Matt left, he installed Grafana and started setting up monitoring pages. I've been continuing this work to cover both T2 and T3 machines:

- Graphite/Carbon system is incredibly easy to setup
- Can monitor everything I want and easily add more
- Grafana makes setting up dashboards trivial

Moving Workers to VAC

The biggest ongoing change is that I'm switching all the workers from Torque/CREAM to VAC. Again, this has many benefits for us:

- Very easy to setup (after initial teething problems)
- Don't have to worry as much about OS updates, etc.
- Minimal ongoing administration required
- Don't have to run CREAM, Torque, APEL
- Reduces complexity of other services (Squid, BDII, Argus)
- Overall a *significant* reduction in manpower required

Drawbacks I've currently encountered:

- Initial setup did have problems (mostly because of me!)
- Much harder to overprovision due to HD and memory reqs being 'enforced'
- I found I needed a Squid per VM factory/Worker

Current status of VAC at Bham is that I have shifted ~50% of the site over

Many Thanks to Andrew McNab for helping me through the setup!

UNIVERSITY OF BIRMINGHAM

Installing/Configuring VAC

Generally, the install and setup of VAC was very easy. I just followed the instructions on the web page:

https://www.gridpp.ac.uk/vac/admin-guide.html

Fundamentally though, after installing appropriate libvirt tools, it's just a case of installing a single RPM

The configuration is managed through a small handful of easy-to-understand config files.

The only issues/gotchas I encountered were:

Firewall:

As I use puppet to manage iptables, it took a few tries to get every rule put in correctly

HS06, GOCDB entry:

VAC is able to send accounting records directly, however you must remember to add an appropriate GOCDB entry and the HS06 values for each worker node

Squid:

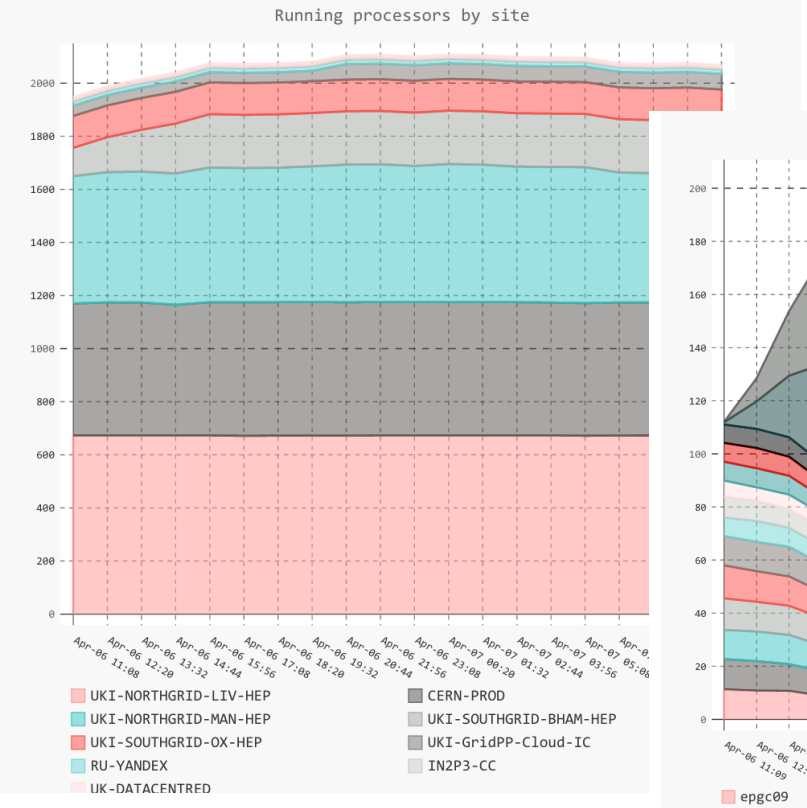
You will probably need multiple squids to cover the additional load because, as far as the squid is concerned, you will have a worker node per core. Our squid couldn't handle this and so I went to a squid per factory. Hopefully I can reduce this in the future.

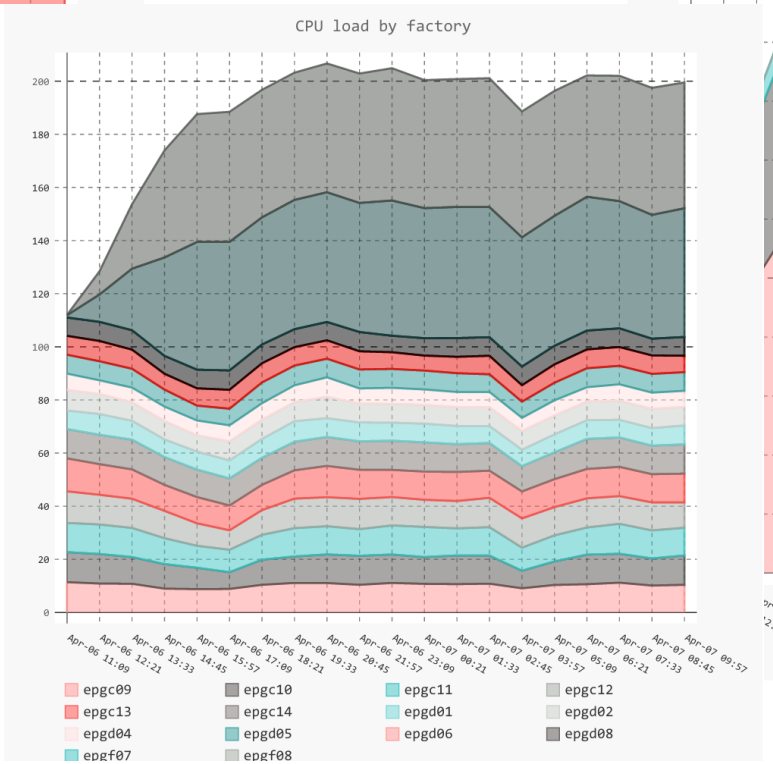
Monitoring VAC

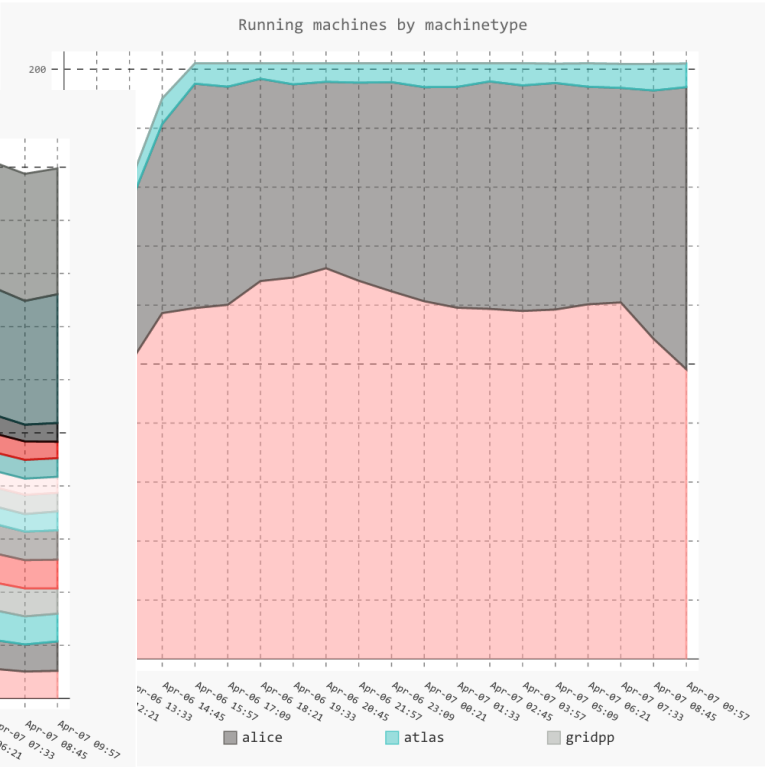
There is very good overall VAC monitoring available here:

http://vacmon.gridpp.ac.uk/1f4:15180::/

From this you can drill down to your site and individual workers







Conclusions

In recent months I've been concentrating on putting things in place to make sys-admin tasks as easy as possible:

- Reorganised server room allowing for ease of installation and expansion
- Easy monitoring of all aspects of the site via Grafana
- Switching to ZFS over HW RAID 6
- Moving all workers to VAC
- Switching to MooseFS from Lustre

I hope to have completed all these tasks by the end of the year and will be in a much better position to keep on top of both Tier3 and Tier2 machines with the reduced manpower