ScotGrid Glasgow - Site Update

Dr. Emanuele Simili

Outline

- Glasgow site capacity & performances
- Current Cluster & Network
- Upcoming Data Center
- New cluster & new network prototype
- HTCondor-CE (brief experience)
- CEPH (by Sam)
- Documentation & Inventory

ScotGrid Glasgow: Gareth, Sam, Gordon + me (Emanuele)

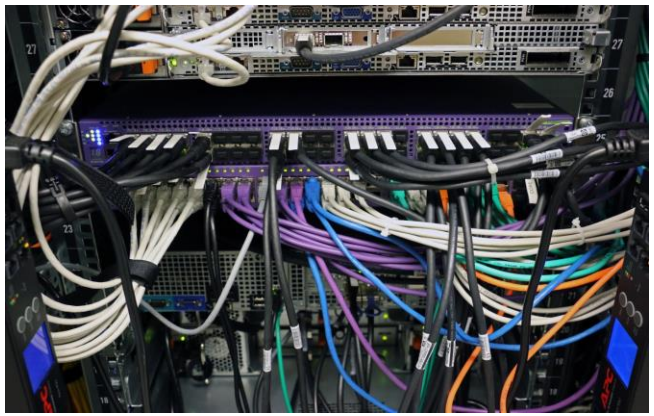
me: NOT an IT expert, but ex physicist, teacher, programmer …
my role: learn, ask questions, organize the information → Wiki! (see last slide)

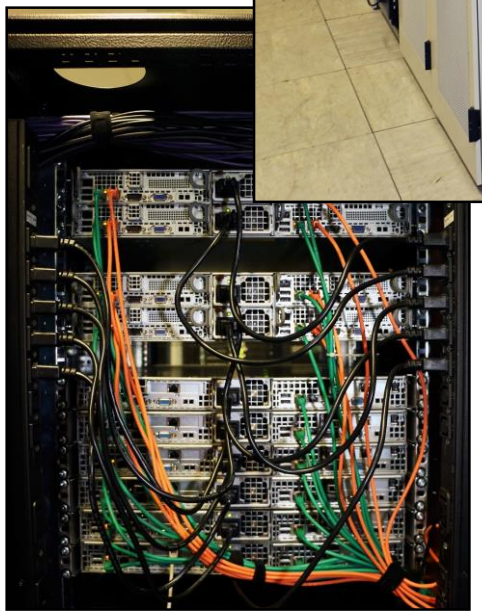
UKI-SCOTGRID-GLASGOW

- Part of the GridPP collaboration providing resources to the Worldwide LHC Compute Grid (WLCG).
- One of 19 institutions comprising 4 distributed Tier-2 sites (SCOTGRID, NORTHGRID, SOUTHGRID and LT2).
- Part of the SCOTGRID Distributed Tier-2 including Glasgow, Edinburgh and Durham Universities.

Our Current Capacity

- At present consist of:
  - 4864 CPU cores (with 1000 core uplift by end if 2016)
  - 7,348,500 MB of RAM
  - 2.4 PB of Storage (with 0.5 PB being commissioned)
  - 160 Gb/s internal network bandwidth





Site Performance

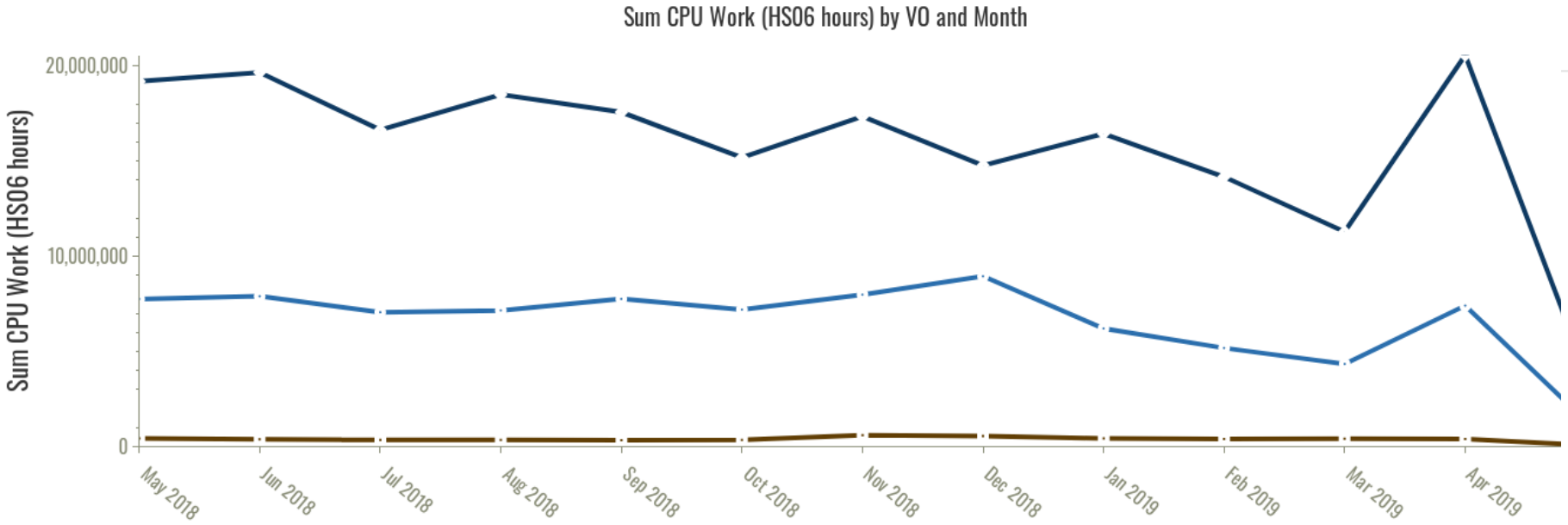
Resource Centre uki-scotgrid-glasgow — Sum CPU Work (HS06 hours) by VO and Month (LHC VOs)

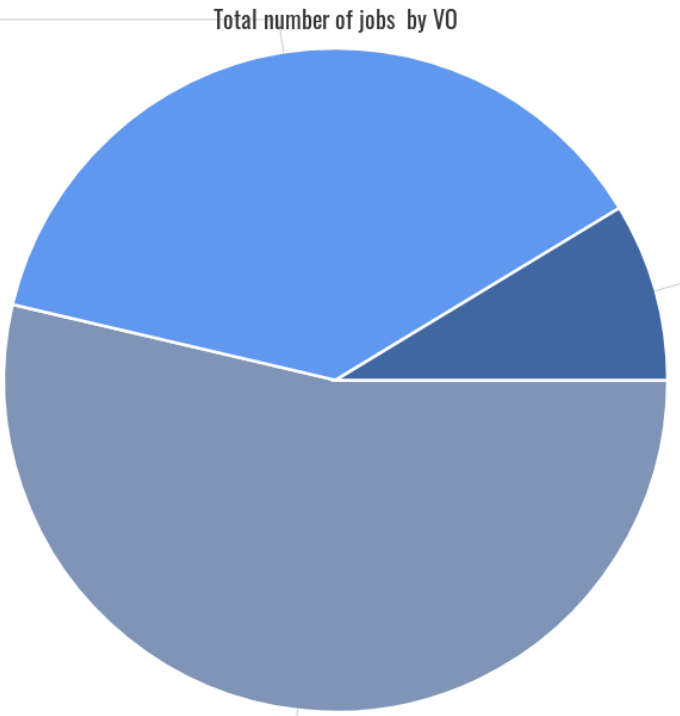
VO	May 2018	Jun 2018	Jul 2018	Aug 2018	Sep 2018	Oct 2018	Nov 2018	Dec 2018	Jan 2019
atlas	19,195,618	19,663,581	16,640,824	18,500,603	17,571,458	15,182,736	17,351,971	14,774,078	16,447,470
cms	430,785	374,041	344,363	348,997	335,017	346,565	596,845	548,861	420,125
lhcb	7,755,369	7,899,550	7,058,324	7,154,368	7,762,070	7,195,556	7,981,783	8,952,423	6,203,194
Total	27,381,772	27,937,171	24,043,511	26,003,968	25,668,546	22,724,856	25,930,599	24,275,362	23,070,789
Percent	9.23%	9.42%	8.11%	8.77%	8.66%	7.66%	8.74%	8.19%	7.78%

1 - 3 of 3 results ‹ 1 › Number of rows per page 30

Download JSON Data / Download CSV Data

The information in the previous table is also shown in the following graph.

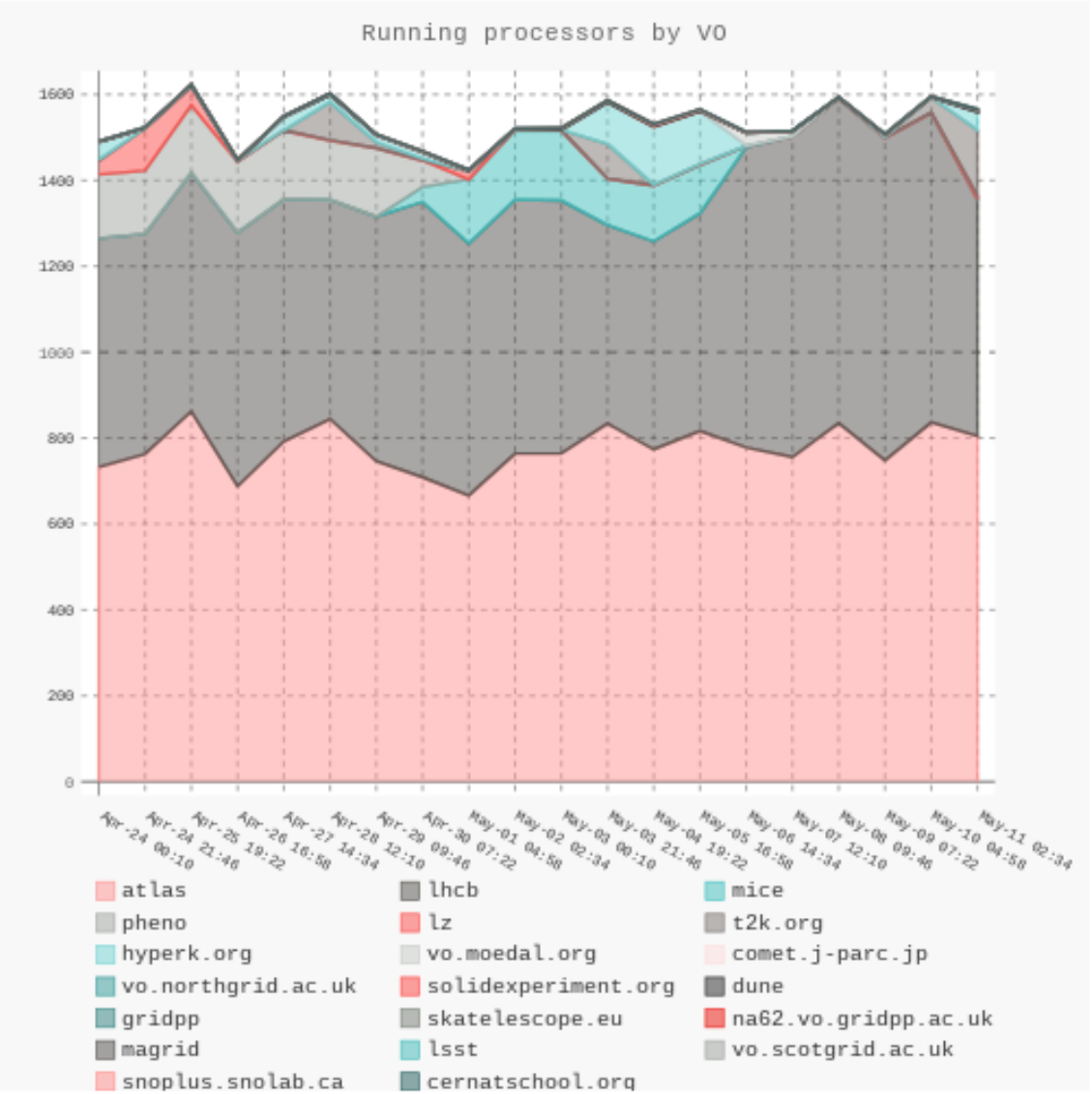


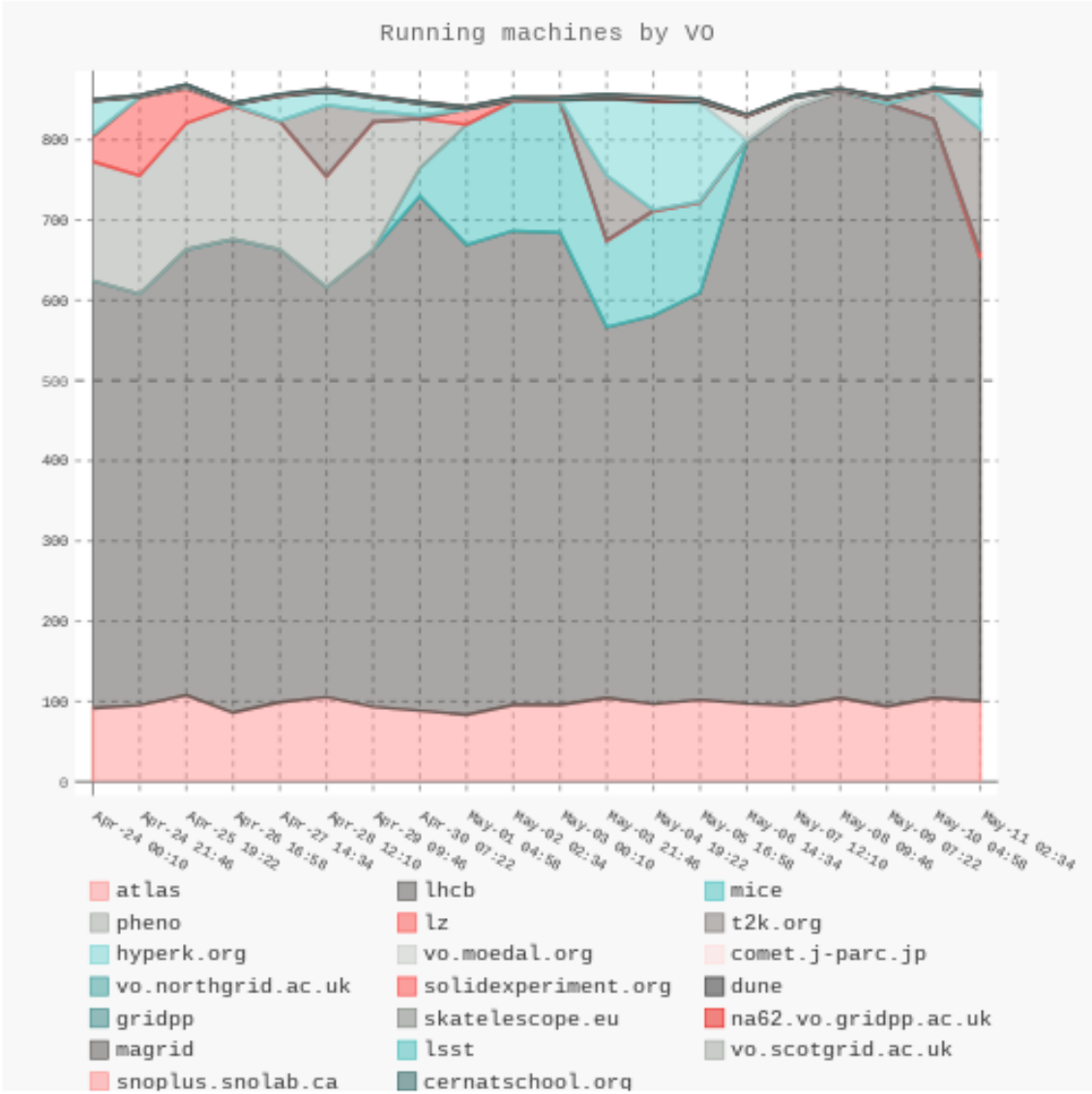


We support large VOs via the standard ARGUS/ARC architecture.

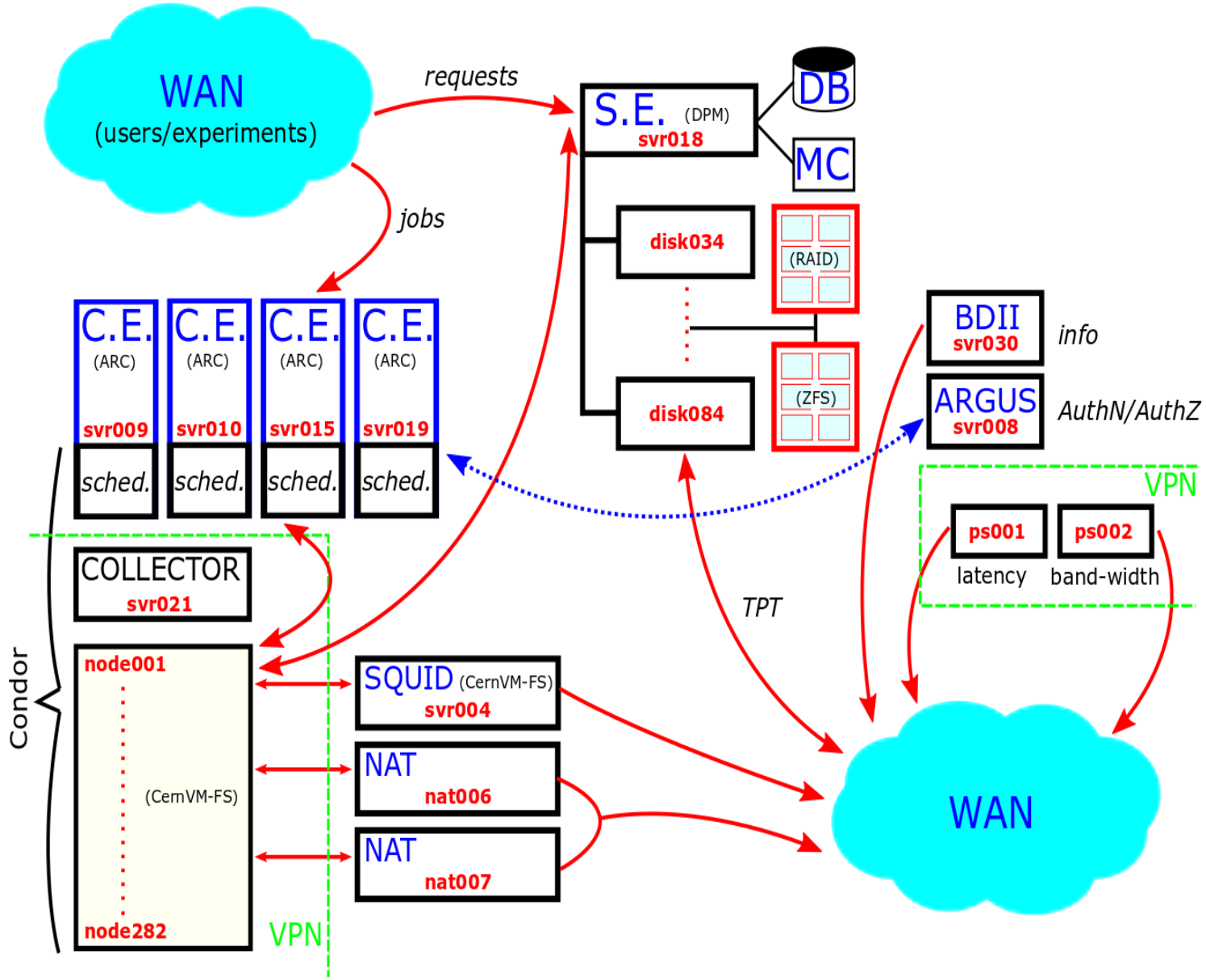
VACs

We continue to run VAC, which are exceedingly useful to support small VO.



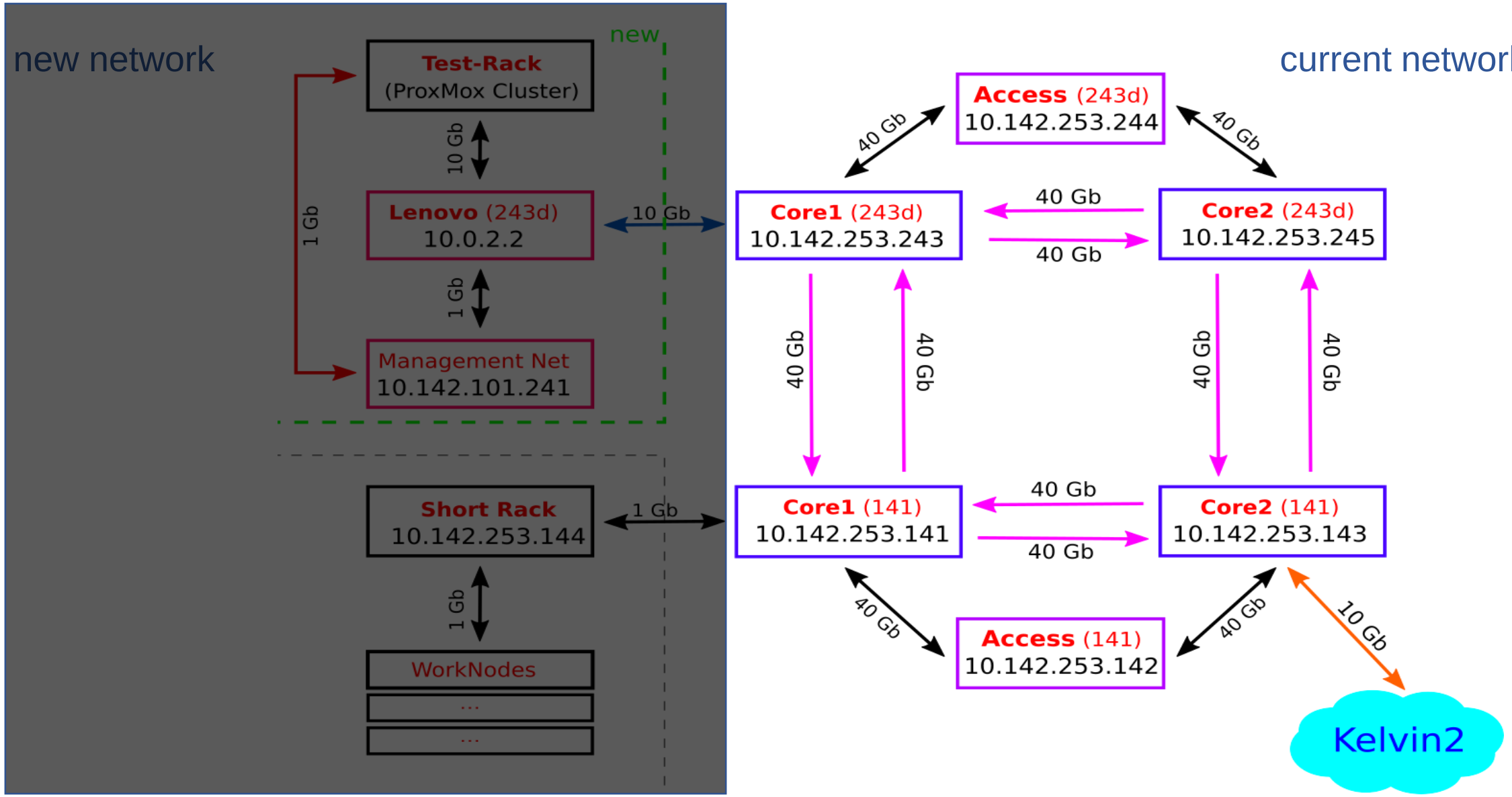


Cluster Map



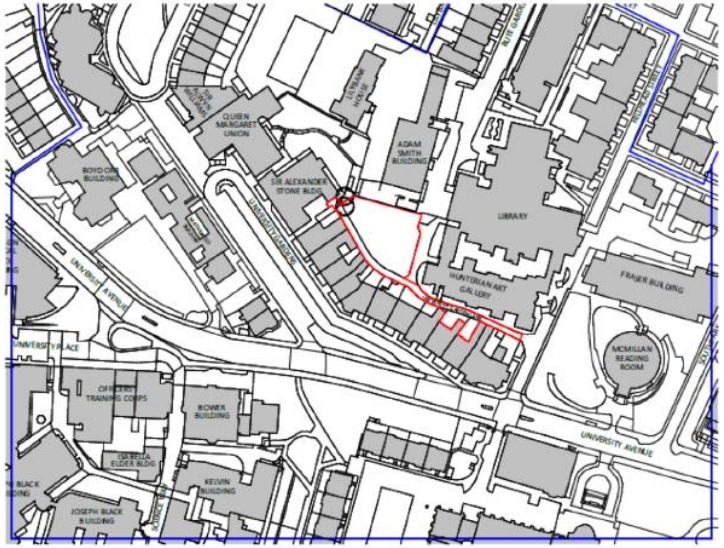
Service	Host	IPv4 (internal)	IPv4 (external)	VM
ARGUS	svr008	10.141.255.8	130.209.239.8	KVM
BDII-Site	svr030	10.141.255.30	130.209.239.30	
ARC-CE	svr009	10.141.255.9	130.209.239.9	
	svr010	10.141.255.10	130.209.239.10	
	svr015	10.141.255.15		
	svr019	10.141.255.19	130.209.239.19	
DPM	svr018	10.141.255.18		
Condor Collector	svr021	10.141.101.42		oVirt
Condor Master	condor-master	10.141.101.44		
Squid	svr004	10.141.255.24	130.209.239.24	
		10.141.201.1		
NAT	nat006	10.141.246.5	130.209.239.5	
	nat007	10.141.246.6	130.209.239.6	
perfSONAR	ps001	10.141.255.123	130.209.239.123	
	ps002	10.141.255.124	130.209.239.124	
Cobbler	provision	10.141.100.1		
		10.142.100.1		
VPN	svr024	10.141.255.24	130.209.239.24	
		10.141.201.1		
Node(s)	node001 - node282	10.141.0.1 - 10.141.1.29		
Disk(s)	disk034 - disk084	10.141.245.34 - 10.141.245.89	130.209.239.34 - 130.209.239.89	
Vac(s)	vac001 - vac044	10.141.213.1 - 10.141.213.44		

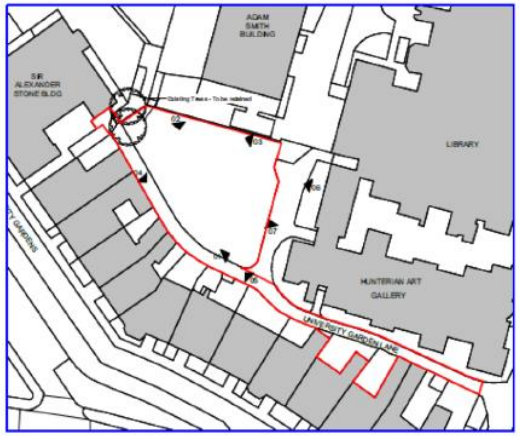
Network Map



All IPs refer to the internal network configuration.

Data Center





View 01

View 04

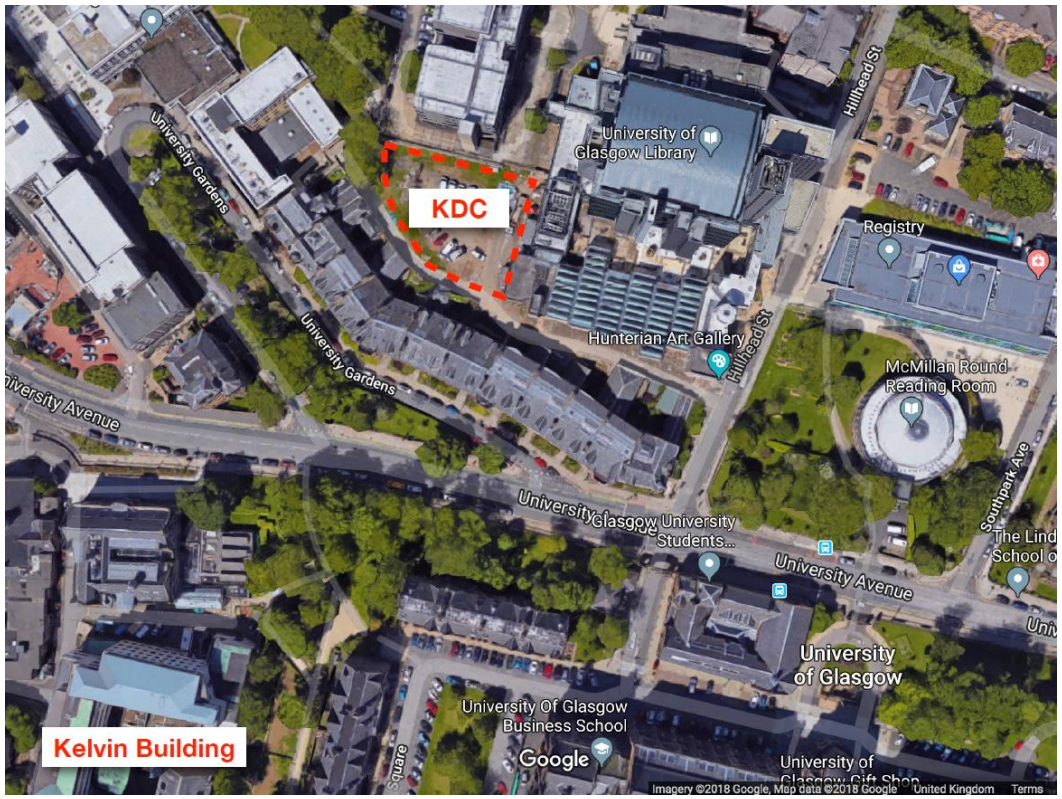
View 05

View 06

View 07

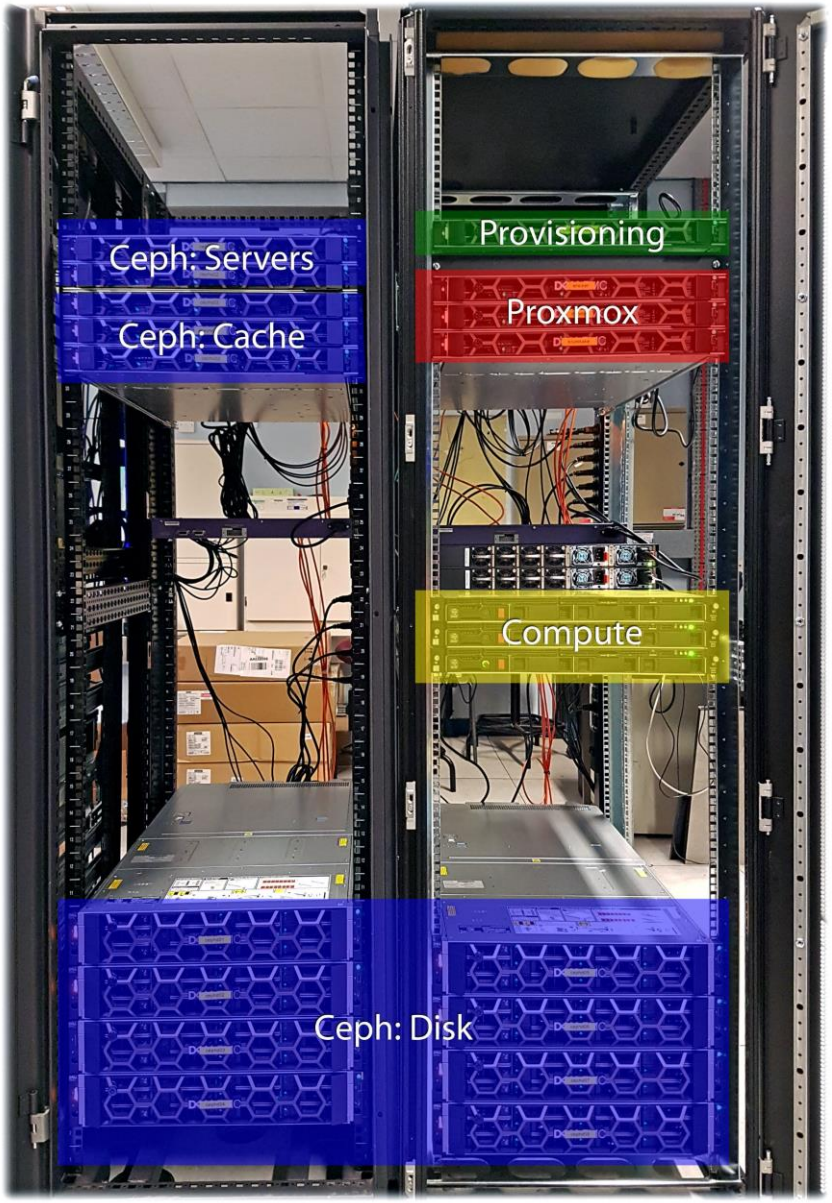
View 02

View 03



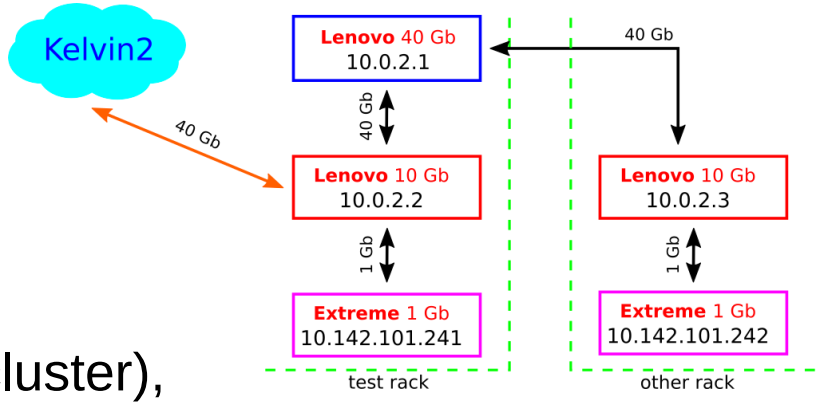
The new data center is being completed and we hope to move the first servers in July …

Data Center Prototype



We have built the first prototype of the new Data Center by re-furbishing 2 racks with:

- brand new DELL 440, 640 & 740 servers,
- 3 recycled HP ProLiant worknodes,
- 1, 10 & 40 Gb switches.



And we have installed:

- provisioning server,
- VM Hypervisor (ProxMox cluster),
- HTCondor batch system,
- CEPH storage,
- networking (DNS, DHCP, NAT),
- services.

(see details on the next slides)

New Cluster

These are the machine and switches installed in the two new racks.

Hardware	Name	Model	IPv4	Service
10Gb Switch		Lenovo RackSwitch G8272	10.0.2.3	
Dell Server	CephSvr1	DELL R440	10.1.50.1	CEPH server
Dell Server	CephSvr2	DELL R440	10.1.50.2	CEPH server
Dell Server	CephSvr3	DELL R440	10.1.50.3	CEPH server
Dell Server	Cache11	DELL R440 +	10.1.50.11	CEPH cache
Dell Server	Cache12	DELL R440 +	10.1.50.12	CEPH cache
1Gb Switch		Extreme Summit x440-48t	10.142.101.242	*mngm*
Dell Server		DELL R740	10.1.50.21	CEPH disk
Dell Server		DELL R740	10.1.50.22	CEPH disk
Dell Server		DELL R740	10.1.50.23	CEPH disk
Dell Server		DELL R740	10.1.50.24	CEPH disk

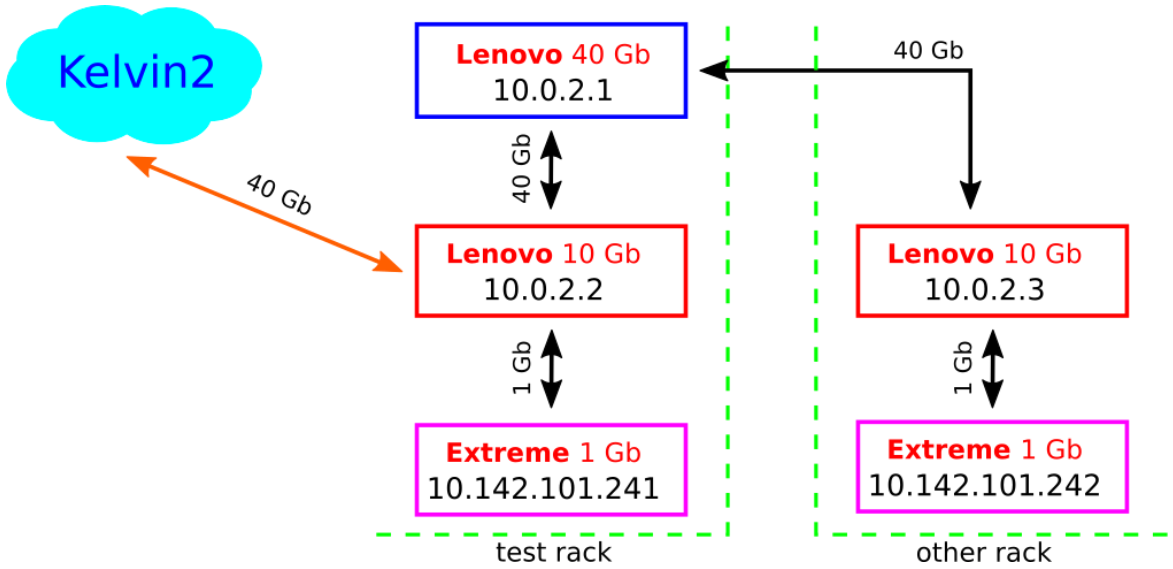
Hardware	Name	Model	IPv4	Service
40Gb Switch		Lenovo RackSwitch G8332	10.0.2.1	
10Gb Switch		Lenovo RackSwitch G8272	10.0.2.2	
Dell Server	Croquembouche	DELL R640	10.1.10.1	provision
Dell Server	Snicker	DELL R440	10.1.20.1	ProxMox
Dell Server	Doodle	DELL R440	10.1.20.2	
Dell Server	Krumkake	DELL R440	10.1.20.3	
1Gb Switch		Extreme Summit x440-48t-10G	10.142.101.241	*mngm*
10Gb Switch		Lenovo RackSwitch G8272	10.0.2.4	
10Gb Switch		Lenovo RackSwitch G8272	10.0.2.5	
HP Server		HPE ProLiant DL60 Gen9	10.1.60.1	WorkNode
HP Server		HPE ProLiant DL60 Gen9	10.1.60.2	WorkNode
HP Server		HPE ProLiant DL60 Gen9	10.1.60.3	WorkNode
Dell Server		DELL R740	10.1.50.25	CEPH disk
Dell Server		DELL R740	10.1.50.26	CEPH disk
Dell Server		DELL R740	10.1.50.27	CEPH disk
Dell Server		DELL R740	10.1.50.28	CEPH disk

New Internal Network

- 4 separate VLANs dedicated to different purposes
- fresh set of IPv4 addresses which will be assigned following some rules

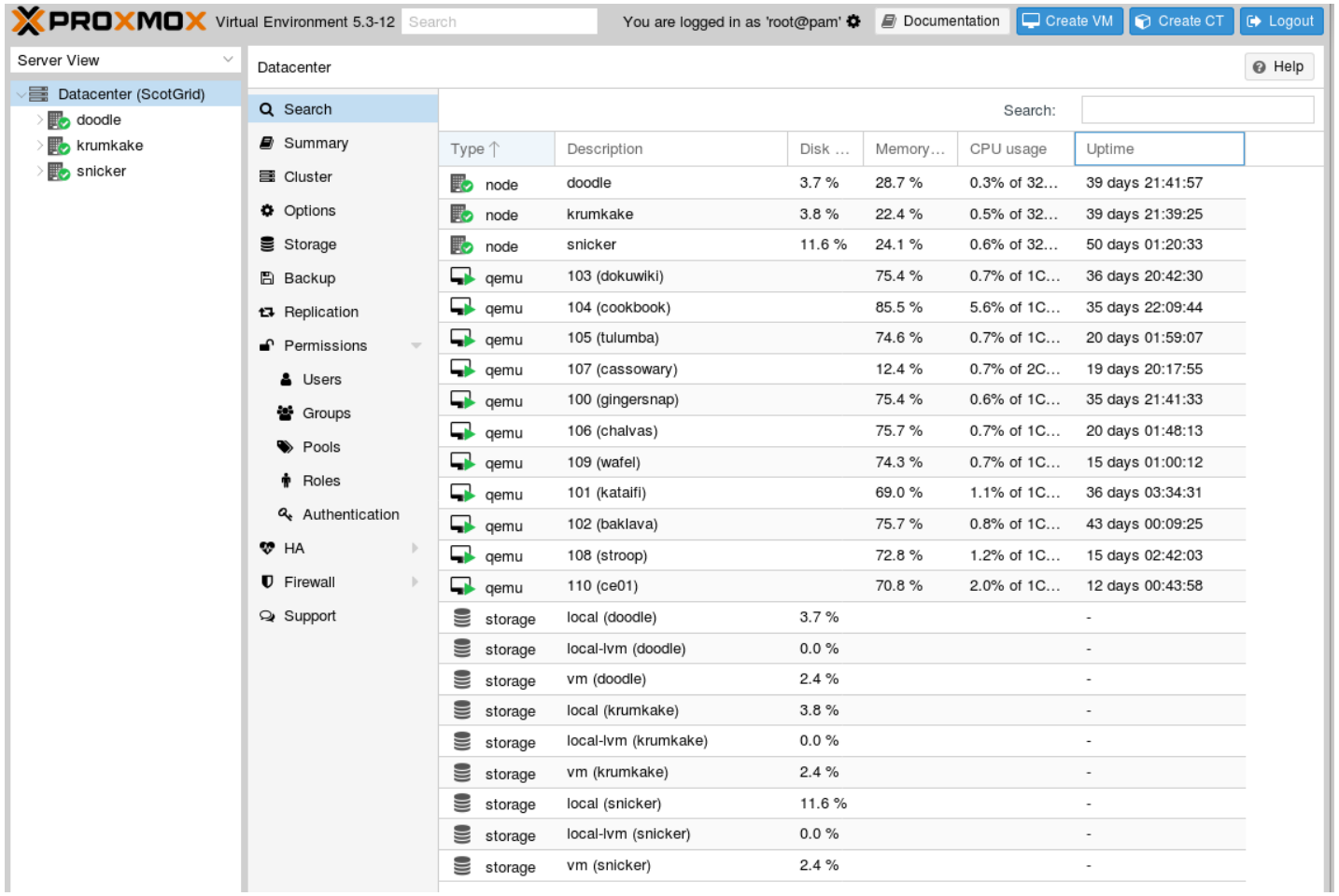
IPv4	VLAN	Network	X range
10.0.X.*	IPMI	management	1 → pdu
			2 → switches
			3 → raid-card
			4 → environmental
10.1.X.*	MAIN	internal	10 → bare-metal
			20 → hypervisor (vmhost)
			30 → vpn
			40 → services
			50 → storage
			60+ → compute
130.209.239.*	EXT	external	only as needed
192.168.*.*	CEPH	Ceph FS	only as needed

VLAN	scope	Tag
MAIN	internal	100
EXT	external	200
IPMI	management	300
CEPH	Ceph FS	400



VM Hypervisor

The ProxMox cluster is composed of 3 DELL 440 servers. It works great, was easy to install and configure, and it is already running several VMs dedicated to various services (see tables):



Network Services (10.1.40.*)

Name	Service	IPv4
Kataifi	DHCP	10.1.40.1
Baklava	DNS	10.1.40.11
Tulumba	DNS	10.1.40.12
Chalvas	DNS	10.1.40.13
Stroop	NAT	10.1.40.21
Wafel	NAT	10.1.40.22

Miscellaneous Services (10.1.41.*)

Name	Service	IPv4
Dokuwiki	Wiki	10.1.41.1
Cookbook	Gitlab	10.1.41.2

Blah blah blah... (10.1.42.*)

Name	Service	IPv4
Gingersnap		10.1.42.1

HTCondor (10.1.43/44.*)

Name	Service	IPv4
Cassowary	Manager	10.1.43.1
Ce01	CE	10.1.44.1

New HTCondor Cluster

From 3 recycled HP ProLiant machines (decommissioned from the old cluster) we have built the new HTCondor batch system prototype (CentOS7).

We have done the provisioning with the new PpePixie, and wrote Ansible roles for the complete set-up of Worknodes and Manager (repositories, packages, configuration).

Name	Service	IPv4	Hardware
wn001	WorkNode	10.1.60.1	HP ProLiant
wn002	WorkNode	10.0.60.2	HP ProLiant
wn003	WorkNode	10.0.60.3	HP ProLiant
Cassowary	Manager	10.1.43.1	ProxMox VM
Ce01	CE	10.1.44.1	ProxMox VM

Node type	Daemons
Worker Node	MASTER, STARTD
Manager Node	COLLECTOR, MASTER, NEGOTIATOR, SCHEDD
CE Node	MASTER, SCHEDD

A default user has been created and the HTCondor batch system was then tested with the Eratostene’s sieve to find prime numbers … so far so good.

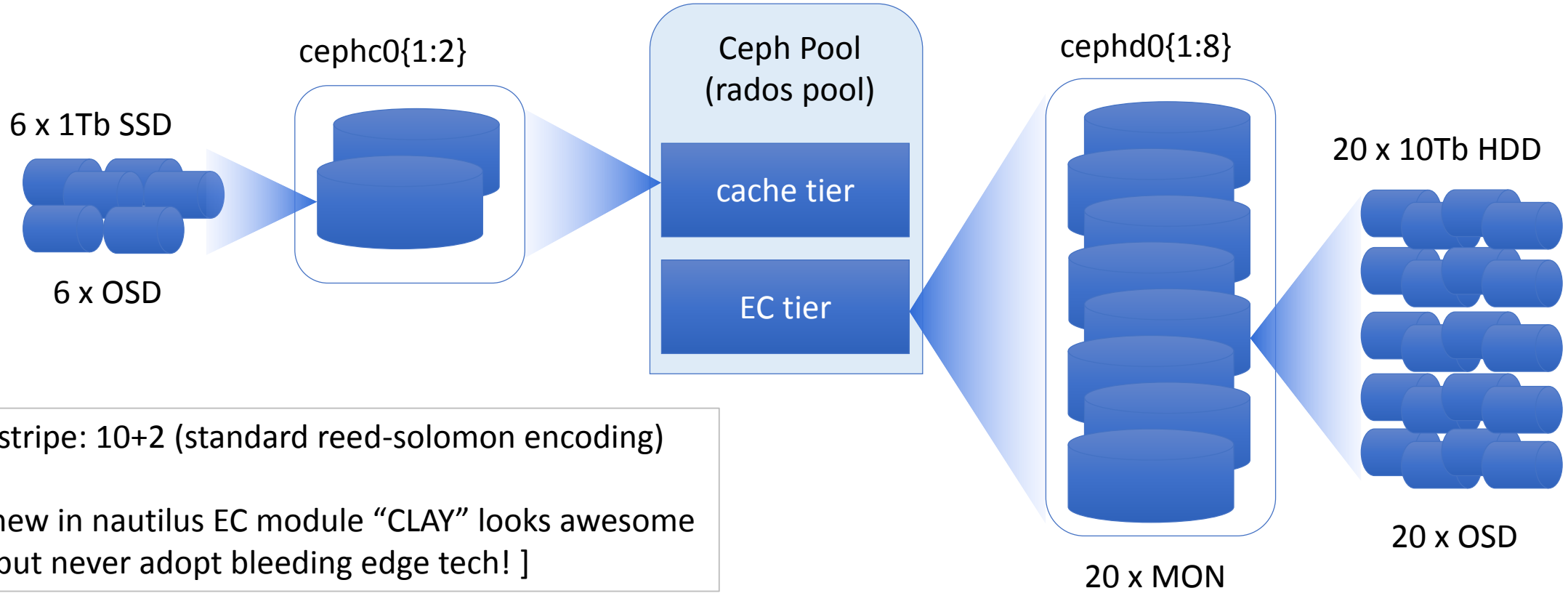
HTCondor-CE Attempt

We investigated deployment of an HTCondor-CE, using details from Steve Jones (see GridPP talk) and the OSG documentation …
No documentation was an exact match for our set-up: multiple CEs (four at present), with a separate central manager. In particular, the configuration of security between the CE and the manager proved tricky (it might be easier if we turned all security off, but we're reluctant to do that!).

We still have a lot of work to do prior to our impending Data Center, which restricts the amount of time we can spend. Plus, it appeared there would be a lot of background research to do in areas of Condor we're not familiar with before we became comfortable with the system.

So … we have decided to stick with ARC CEs at the moment, as we are familiar with the configuration and have confidence in the technology. We will revisit HTCondor-CEs at a later stage, when we have a bit more time and experience!

CEPH



EC stripe: 10+2 (standard reed-solomon encoding)

[new in nautilus EC module “CLAY” looks awesome but never adopt bleeding edge tech!]

- Basic config for ceph pool at GLA
- Config via ceph-ansible [which has some quirks if you use the Centos 7 Storage SIG Repo, as there’s one or two odd packaging dependencies]
- Plus: xrootd, gridftp daemons using libradosstriper for data placement and access.
  - libradosstriper splits incoming files into stripes of uniform chunks (chunk size is invariant) – chunks are then EC coded into stripes in the EC tier.

CEPH (2)

- Differences to RAL
  - RAL caching is in xrootd / per WN
    - Reconstructs entire libradosstriper stripe + caches file
    - Caches are local to WN (so less shared cache efficiency)
  - GLA caching is in ceph cache tier / per pool
    - Reconstructs EC stripe + caches libradosstriper chunks
    - (There's still a reconstruction overhead per libradosstriper chunk for accessing entire file)
    - Caches are shared for entire cluster (better shared cache efficiency)

- Plan A – as in these slides, ATLAS space entirely in ceph (migrate our other Nx10TB disk nodes also into the ceph pool)
- Plan B – add cephfs "posix" layer on top, export as "neoclassical SE"
- Plan C – try EOS (without EC as EOS EC performance is ?poor?, so efficiency loss)

For any question related to Storage and CEPH, please ask Sam

Documentation

Cluster information and administrative procedures (for technicians) are being organized in a shiny DokuWiki format.

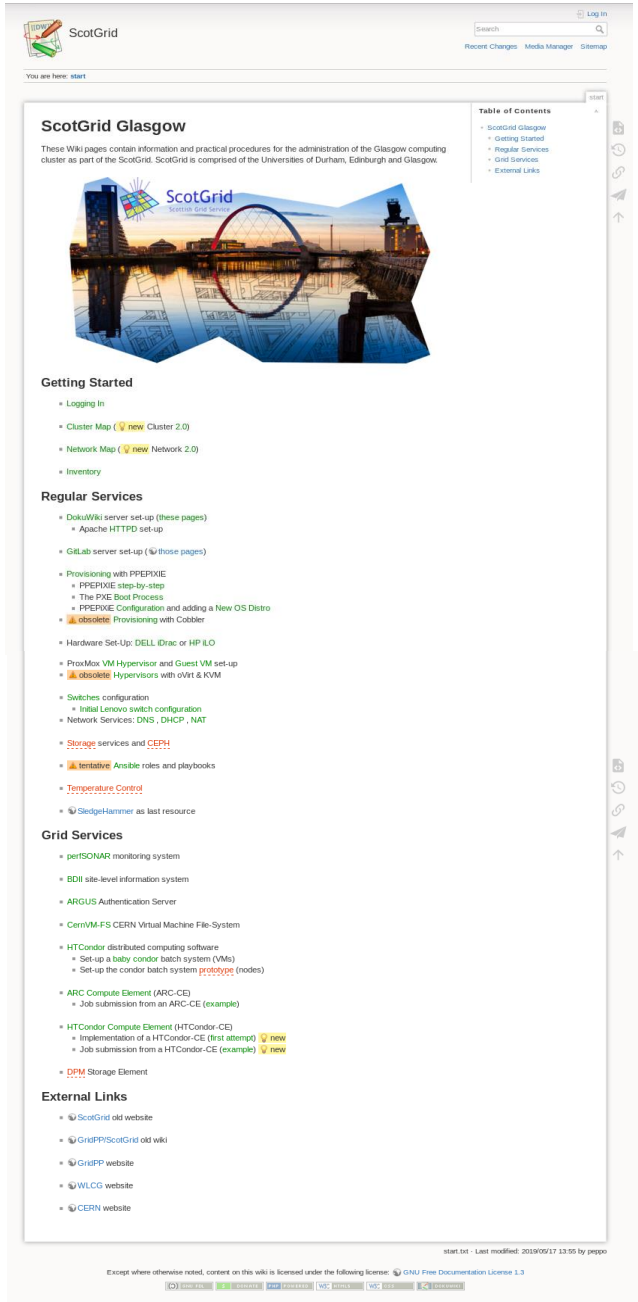
Information such as:
- cluster & network map,
- list of machines (inventory),
- services status.

http://dokuwiki.beowulf.cluster/

Procedures such as:
- routers configuration,
- servers provisioning,
- VMs creation,
- services set-up.

Since the information is very site-specific (and security-sensitive), these pages are currently available only to us (Glasgow).

But we are ready to share our procedures and Ansible roles to whoever asks.

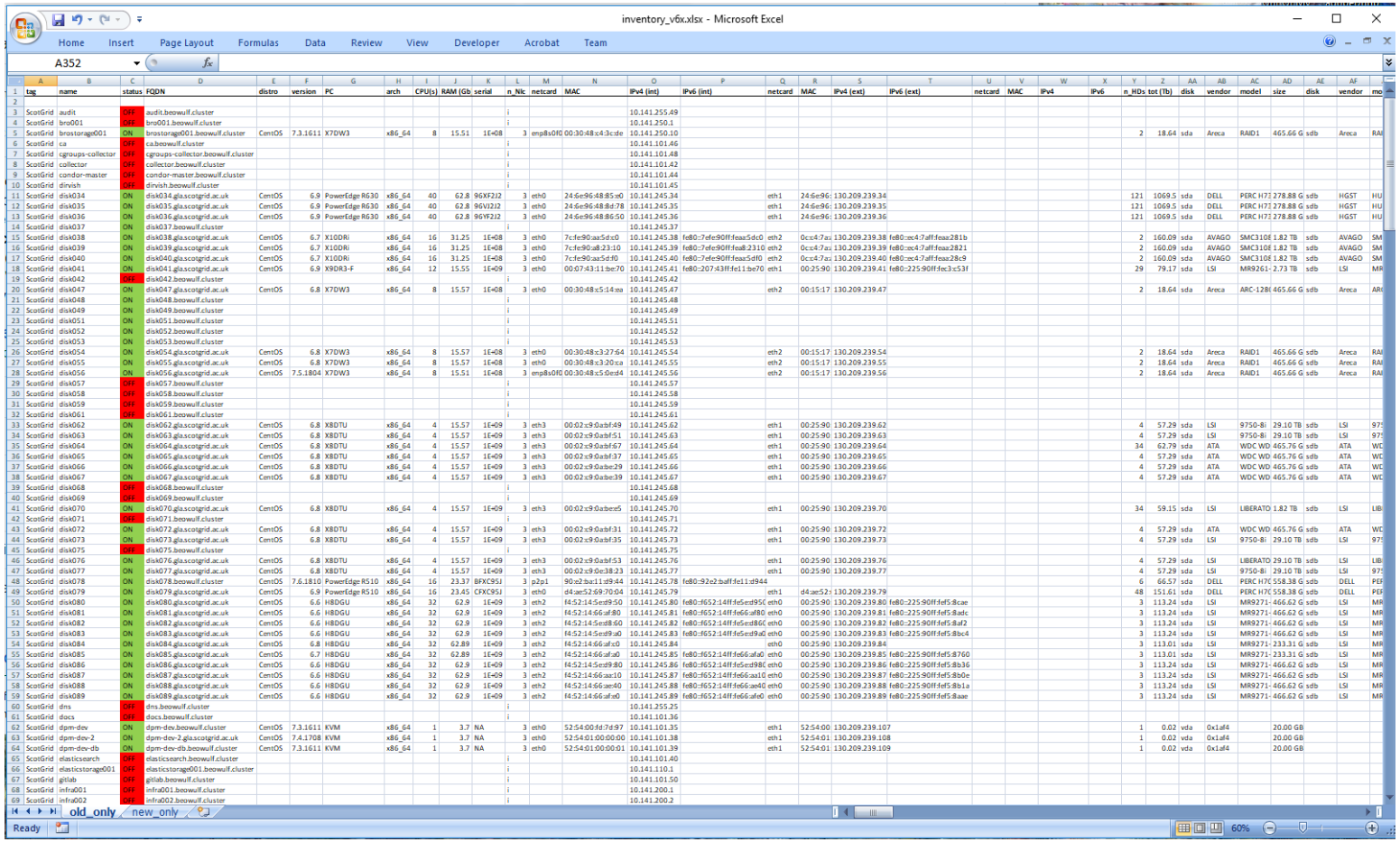


Inventory Tool

A useful tool I have developed at the very beginning of my contract is a sort of inventory script, which uses Ansible to collect information from all machines on the cluster and python script to organize them in a CSV file.

The procedure is easy to run and can be done every month to keep the machine inventory up to date.

The starting point is anyway the ansible inventory, which need to be maintained by hand - unless you wanna play with *nmap* and *dig* and try to blindly re-map the whole network (a futile exercise which I have also done …).



Show the latest update: inventory_v6x.xlsx

ScotGrid Glasgow - Site Update

END