CHPC WLCG Tier2 Facility and SA Grid Operations

Sean Murray
ALICE
CHPC
CSIR

October 6, 2016

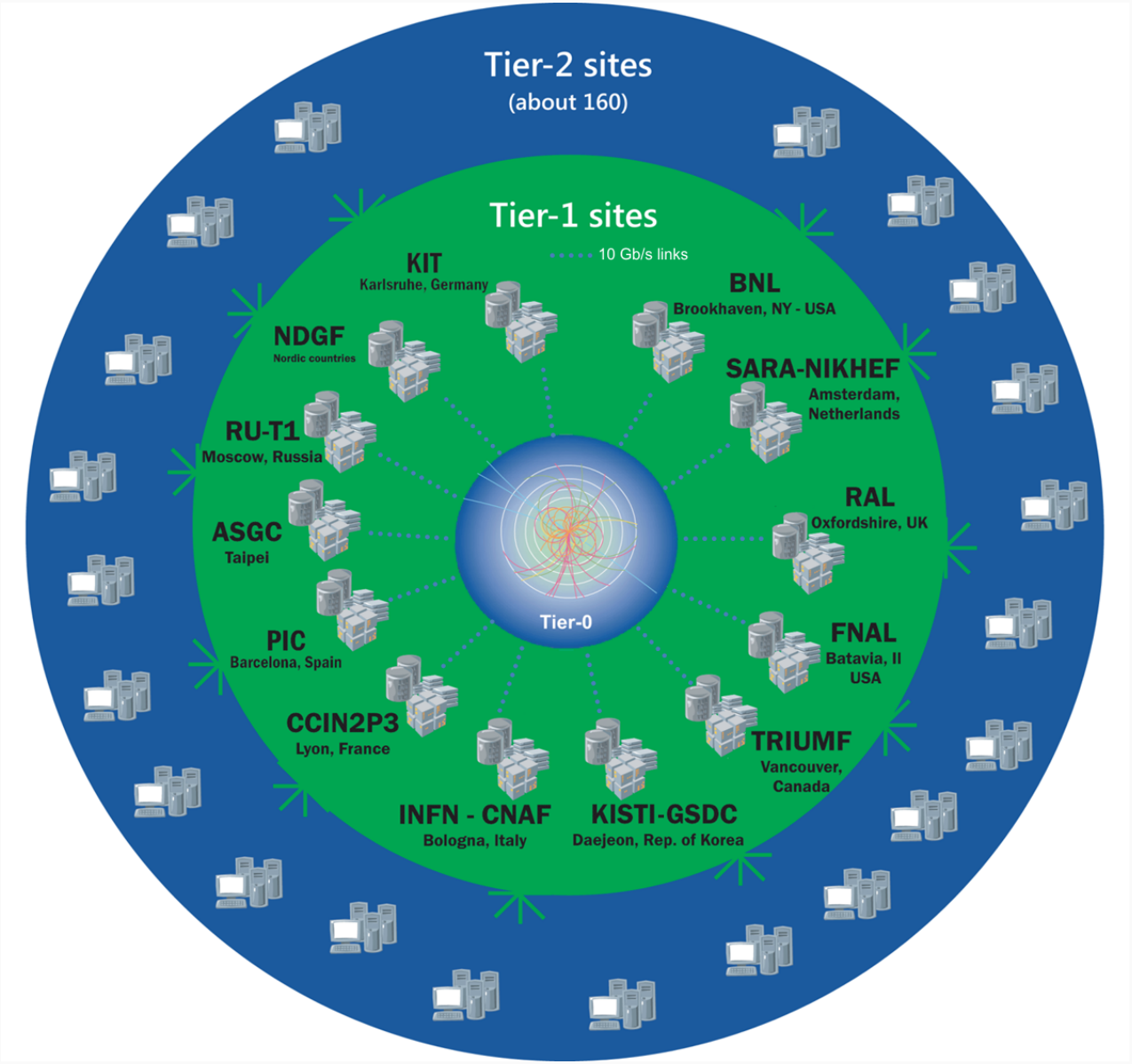
Outline

WLCG

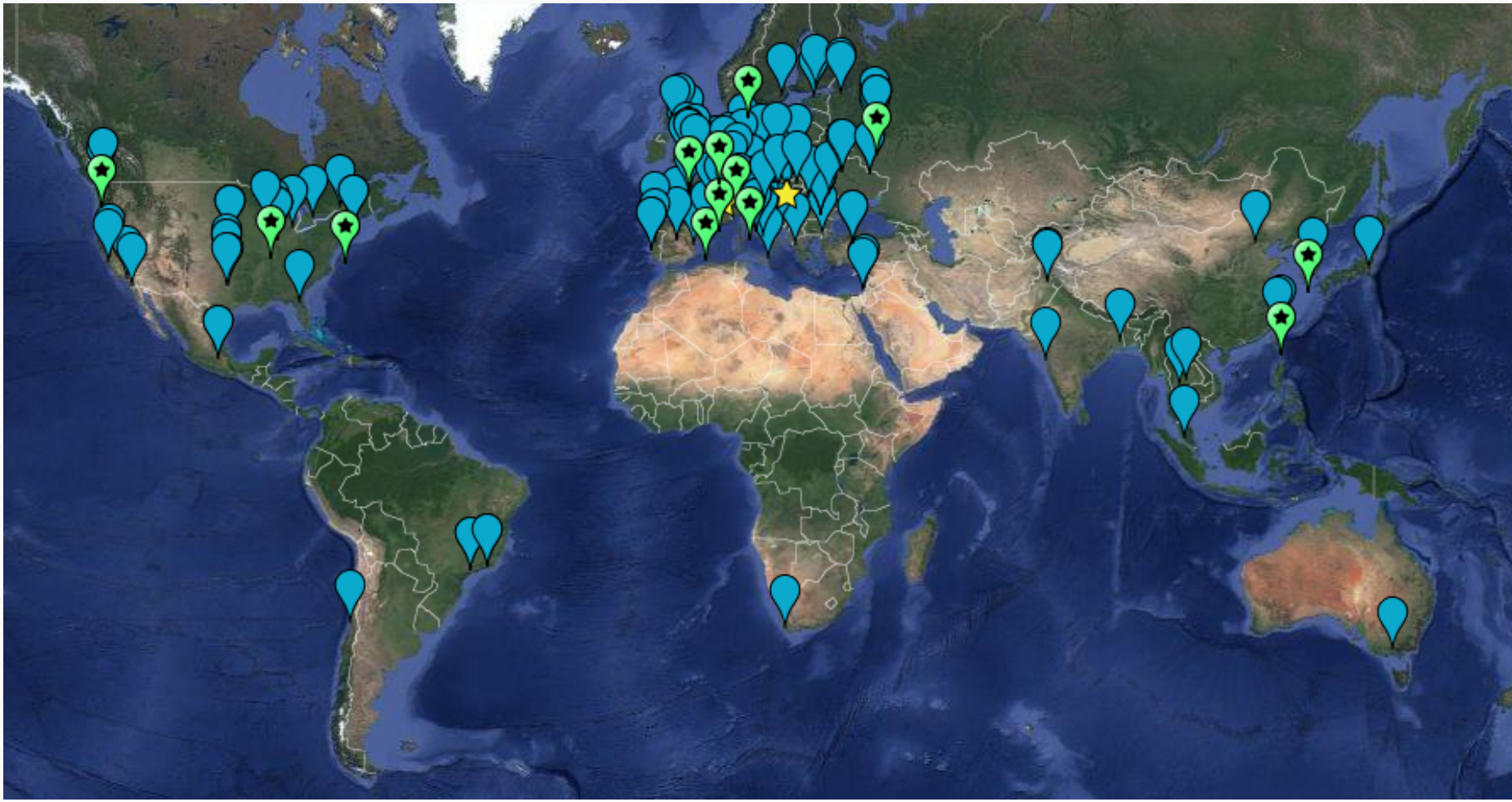
SAGrid, user analysis

Backup Slides

What is the WLCG



WLCG Map of Sites



WLCG MOU Signed

28 April 2015

Commitments

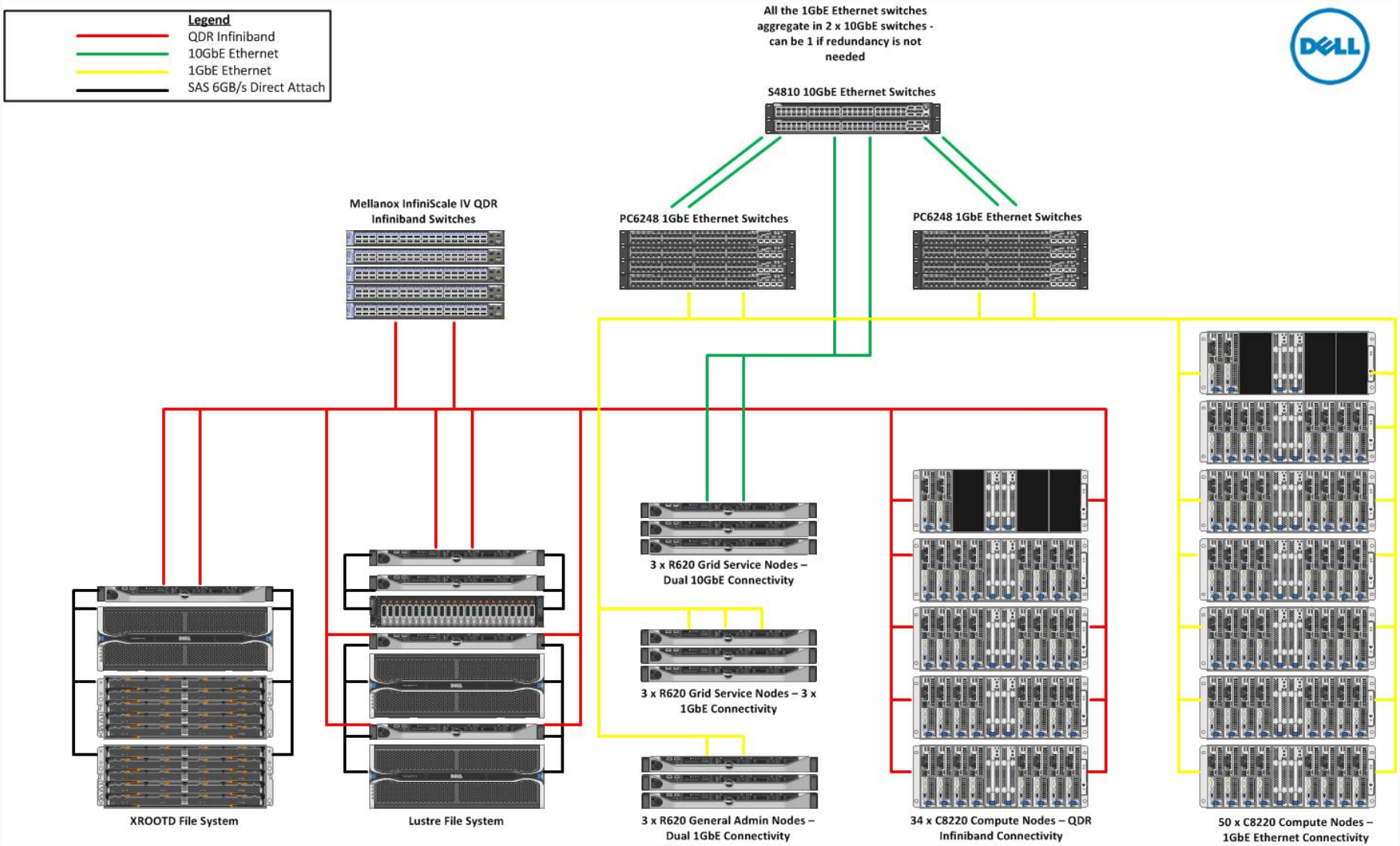
According to Tender :

- ALICE 600 cores
- ATLAS 600 cores
- ALICE 400TB
- ATLAS 400TB

According to :
https://wlcg-rebus.cern.ch/apps/pledges/resources/

- 6000 HEPSPEC06 cores (c. 560 of our cores)
- 100TB storage
- All ALICE.
- now equal ALICE and ATLAS (?)

Computing Infrastructure



Current hardware

- 50 nodes of 48 cores 192GB RAM and 2x800TB of SSD, 1G ethernet
- 34 nodes of 48 cores 96GB RAM and 1TB SATA, FDR infiniband, 6 “stolen”
- c.100TB of Lustre on the 34 nodes with FDR infiniband.
- 9 management servers, lower spec
  - compute element (head node,ce),
  - storage element 2 redirectors, 2 storage nodes with direct attached multipath storage
  - authentication, user interface (gone), monitoring, provisioning.

Current Storage

- 383TB EOS for ALICE, down from 440TB
- 252 TB EOS for ATLAS, down from 400TB
- 107 TB lustre for 34 nodes.
- 104 TB EMC for ATLAS, not going to be used.
- 400TB ALICE
- 400TB ATLAS
- 400TB General EOS

Reduction in data sizes is due to reorganisation for reliability.

Availability / Reliability

So who monitors us :

- WLCG
- EGI
- ALICE
- ATLAS
- me via zabbix/grafana and racktables

There are a lot of eyes, ignoring the ones in this room.

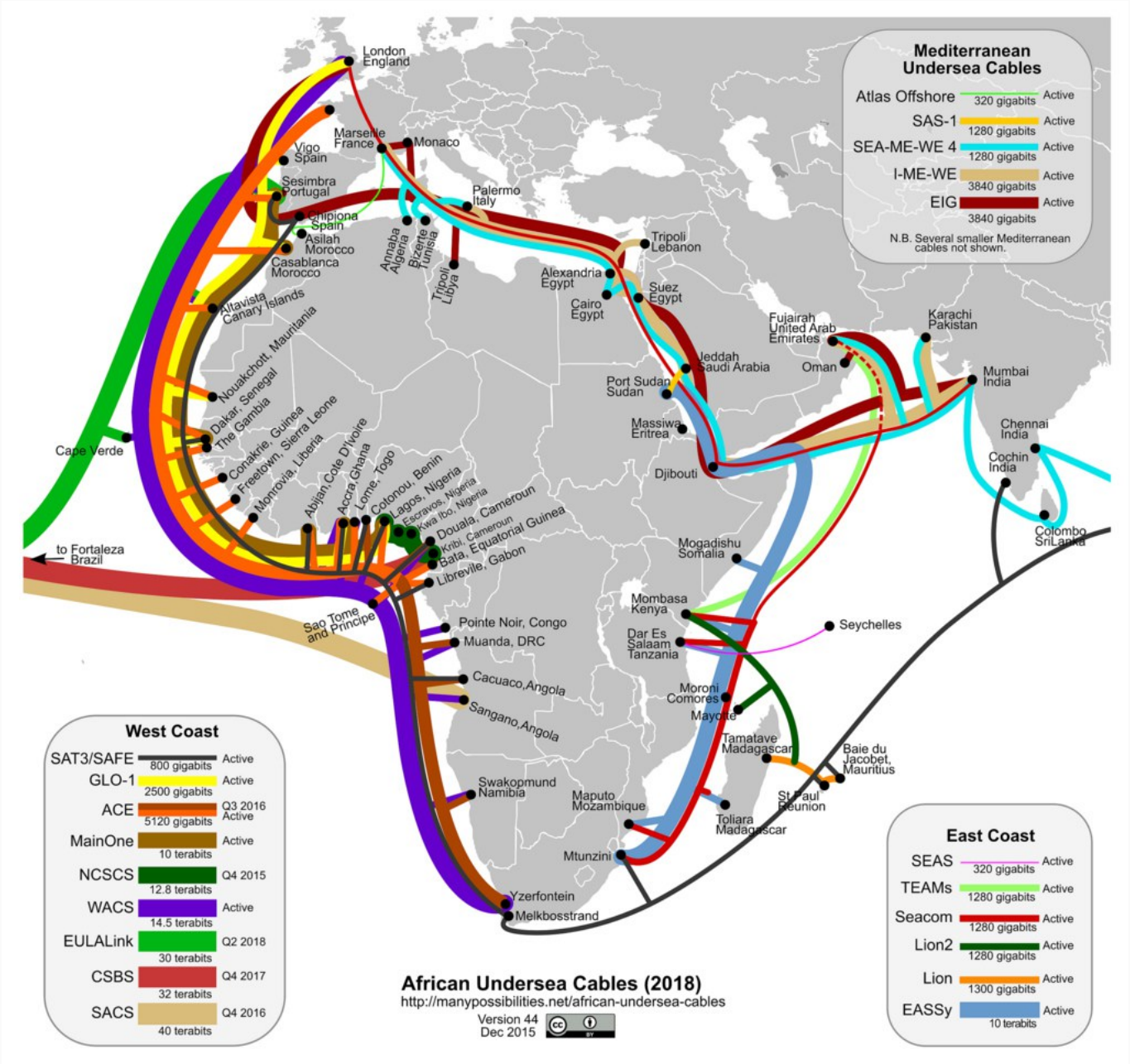
Function	July	August	September
Availability	100	100	100
Reliability	100	100	100

I am soooo going to regret this slide !

Grafana ALICE

This is currently being expanded to pull in more from MonaLisa, and more experiment and storage specific metrics to help to trivialy diagnose and forewarn problems.

Our Data's Scenic Tour



Tier2 Processing

- ATLAS processing direct std batch jobs.
- ALICE processing, job agents and leave it to central services, vobox.
- ops jobs fail due to starvation from alice. currently a dedicated machine
- all software is distributed via cvmfs, both experiments.
- ALICE has something similar to code-rade, an entire ecosystem of modules, versions, builds, etc.
- still no official validation on CC7.

Tier2 Storage operations

Thankfully both experiments use xrootd.

- ALICE has its own authentication system, controlled centrally from CERN.
- ALICE has one big block of storage.
- ATLAS uses grid pool accounts, more standard grid.
- ATLAS has various storage's on site, cut up locally.
- ATLAS on puppet, ALICE will go on puppet when upgraded.

Philosophy

- Keep It Simple Simpleton.
- Only use liberally licensed software.
- Stand on whom evers shoulders I can.
- I dont want to do sys admin, I dont enjoy it.
- Abstract sys admin to a Software Engineering problem.
- Let me get back to the things that I enjoy ALICE O2 and HLT which I enjoy, computational physics..

Simple Solution

Switch everything off for 4 or 5 months while i redo **everything**

the plan

- Keep site up and processing maintaing our Tier 2 status.
- incrementally replace systems.
- the return of 28 grid nodes, on infiniband, a place to test configurations when batch is “free”.

Improvements

- Fixed ALICE error rates, mostly.
- ALICE concurrent jobs from 700 to 1800.
- max out bandwidth now regularly.
- ATLAS running pilot jobs.
- Storage cleaned up and reduced.
- New storage quoting
- Plan in place to overhaul, and being tested.
- Work defined to Redo entire internal network.

upgrades

- attempts to delay till CC7 validation have failed.
- Puppify, to auto site deployment, r10k an issue.
- Transition to foreman from xcat.
- upgrade monitoring server to zabbix3, via docker this time.
- Rewire whole network, re-power to monitored pdu, and monitor all.
- Add inherent redundancy into 10G interfaces on ce,se,se2.
- Reinstall while TRYING to keep A/R. problems are vobox and ce.
- Storage, we need an additional 750TB for ALICE and ATLAS just to maintain Tier2.

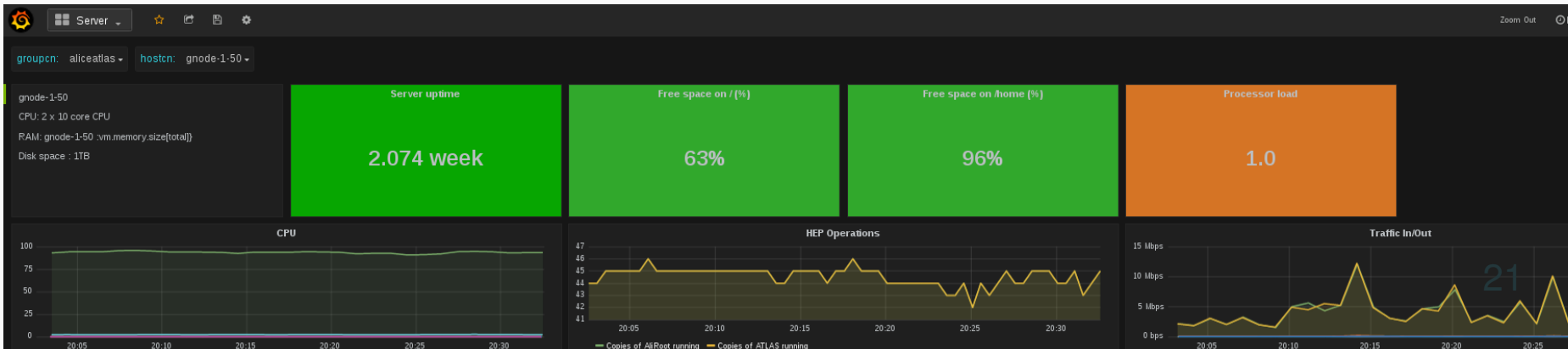
Transparency

Historically this has not been great, so ...

- Federated logins to zabbix and grafana, i.e. You
- All code on github in line with AAROC
- All issues PUBLIC on github, this has been a serious uhming and ahring session.
- Still have GGUS for normal tickets.
- Some training on the user analysis facility (hopefully online)
- A couple of things remain private like network diagrams, obviously, and passwords. github encrypted repos (todo)
- AAROC slac channels.


ALICE
Zoom Out
Job Agents Running
1800
Batch Jobs Running
1800
ALICE Jobs Running
1797
Jobs Running/Queued
Batch Jobs Running
Batch Jobs Queued
Running Jobs
Jobs Agents
Job Statuses
alice_failed_jobs
alice_saved_jobs
alice_error_v_jobs
alice_error_jb_jobs
alice_error_e_jobs
alice_zombie_jobs
ALICE EOS
No datapoints
ALICE Zombies Running
ALICE Running


ATLAS
Zoom Out
Total ATLAS Jobs Running
3
Jobs Running/Queued
Total ATLAS Jobs Running
Total ATLAS Jobs Queued


Server
Zoom Out
groupcn: aliceatlas
hostcn: gnode-1-50
gnode-1-50
CPU: 2 x 10 core CPU
RAM: gnode-1-50 :vm.memory.size[total]}
Disk space : 1TB
Server uptime
2.074 week
Free space on / (%)
63%
Free space on /home (%)
96%
Processor load
1.0
CPU
HEP Operations
Copies of AliRoot running
Copies of ATLAS running
Traffic In/Out

Outline

WLCG

SAGrid, user analysis

Backup Slides

28 nodes

July 28 of the 34 nodes came back for SAGrid and HEP user analysis.

- Go back to SAGrid to support anybody on SAGrid VO.
- HEP user analysis, based on federated identities, no user account admin.
- either HEP user supplies their own eco system with job or build their custom ecosystem via code-rade.
- code based on CODE-RADE, or LHC experiments.
- HLT replica for me and others. (no gpu)
- O2 dev for O2 development with dds/mesos/pbs. (no gpu/fpga)
- Local Storage for users, eos and lustre.
- 24/4 sl6/centos7 in prep for container migration.

deployment/provisioning

Foreman . . .

1. deployment
2. configuration, puppet, chef, salt, ansible.
3. monitoring.

Not using the fully integrated puppet/foreman as i prefer to be tool agnostic, or at least strive for it.

puppet

A method/system to maintain a system in a known state.
hiera, a key value store used by puppet.
We store all site specific information in hiera, tightly integrated into puppet..

AAROC Devops and r10k

- r10k branches are mapped to environments (dev, testing, production).
- Cant be under AAROC, so chpctier2, production branch will be synced to aaroc/devops/chpc.
- 3 branches, dev, testing and production, currently only use production.

Testing

Devops requires some form of automated testing.

- Unit testing via rspec.
- Function testing.
- Fact testing.
- System testing.

Monitoring

- zabbix all data ends up here.
- grafana on zabbix.
- racktables with zabbix link.
- foreman
- monalisa
- have singular sources of authority. idrac mostly.
- all zabbix templates, items etc. on github, not really devops, and deployed by node type.

network

bandwidth

We were graciously donated a 500Mbps test.

We used it all.

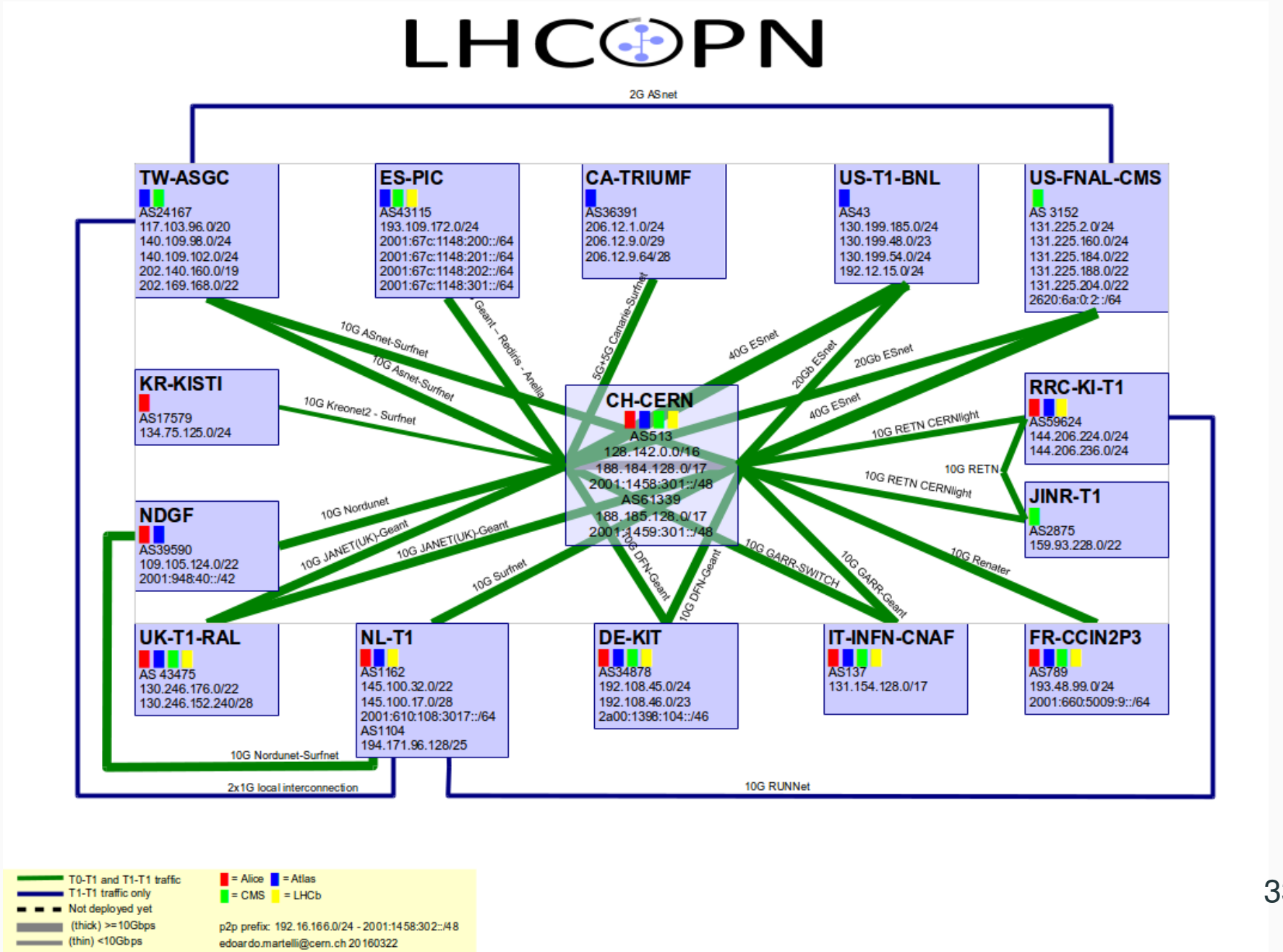
We need more. . . . . .

Summary

- remove as much human contact as possible.
- abstract as much away to bot and psuedo bots
- let the science take priority.

Outline

LHCOPN


LHCOPN
2G ASnet
TW-ASGC
AS24167
117.103.96.0/20
140.109.98.0/24
140.109.102.0/24
202.140.160.0/19
202.169.168.0/22
ES-PIC
AS43115
193.109.172.0/24
2001:67c:1148:200::/64
2001:67c:1148:201::/64
2001:67c:1148:202::/64
2001:67c:1148:301::/64
CA-TRIUMF
AS36391
206.12.1.0/24
206.12.9.0/29
206.12.9.64/28
US-T1-BNL
AS43
130.199.185.0/24
130.199.48.0/23
130.199.54.0/24
192.12.15.0/24
US-FNAL-CMS
AS 3152
131.225.2.0/24
131.225.160.0/24
131.225.184.0/22
131.225.188.0/22
131.225.204.0/22
2620:6a:0:2::/64
KR-KISTI
AS17579
134.75.125.0/24
CH-CERN
AS513
128.142.0.0/16
188.184.128.0/17
2001:1458:301::/48
AS61339
188.185.128.0/17
2001:1459:301::/48
RRC-KI-T1
AS59624
144.206.224.0/24
144.206.236.0/24
JINR-T1
AS2875
159.93.228.0/22
NDGF
AS39590
109.105.124.0/22
2001:948:40::/42
UK-T1-RAL
AS 43475
130.246.176.0/22
130.246.152.240/28
NL-T1
AS1162
145.100.32.0/22
145.100.17.0/28
2001:610:108:3017::/64
AS1104
194.171.96.128/25
DE-KIT
AS34878
192.108.45.0/24
192.108.46.0/23
2a00:1398:104::/46
IT-INFN-CNAF
AS137
131.154.128.0/17
FR-CCIN2P3
AS789
193.48.99.0/24
2001:660:5009:9::/64
10G ASnet-Surfnet
10G Asnet-Surfnet
Geant - Rediris - Anella
5G+5G Canarie-Surfnet
40G ESnet
20Gb ESnet
20Gb ESnet
40G ESnet
10G Kreonet2 - Surfnet
10G RETN CERNlight
10G RETN
10G RETN CERNlight
10G Nordunet
10G JANET(UK)-Geant
10G JANET(UK)-Geant
10G Surfnet
10G DFN-Geant
10G DFN-Geant
10G GARR-SWITCH
10G GARR-Geant
10G Renater
10G Nordunet-Surfnet
2x1G local interconnection
10G RUNNet
T0-T1 and T1-T1 traffic
T1-T1 traffic only
Not deployed yet
(thick) >=10Gbps
(thin) <10Gbps
= Alice
= Atlas
= CMS
= LHCb
p2p prefix: 192.16.166.0/24 - 2001:1458:302::/48
edoardo.martelli@cern.ch 20160322

IOU table

three or more other institutes providing amongst them 52-week coverage).

The following parameters define the minimum levels of service. They will be reviewed by the operational boards of the WLCG Collaboration.

Service	*Maximum delay in responding to operational problems*			*Average availability[6] measured on an annual basis*	
	Service interruption	Degradation of the capacity of the service by more than 50%	Degradation of the capacity of the service by more than 20%	During accelerator operation	At all other times
Raw data recording	4 hours	6 hours	6 hours	99%	n/a
Event reconstruction or distribution of data to Tier-1 Centres during accelerator operation	6 hours	6 hours	12 hours	99%	n/a
Networking service to Tier-1 Centres during accelerator operation	6 hours	6 hours	12 hours	99%	n/a
All other Tier-0 services	12 hours	24 hours	48 hours	98%	98%
All other services[7] – prime service hours[8]	1 hour	1 hour	4 hours	98%	98%
All other services[7] – outwith prime service hours[8]	12 hours	24 hours	48 hours	97%	97%

Annex 3.2. Tier-1 Services

Each Tier1 Centre[9] forms an integral part of the central data handling service of the LHC Experiments. It is thus essential that each such centre undertakes to provide its services on a long-term basis (initially at least 5 years) and to make its best efforts to upgrade its installations steadily in order to keep pace with the expected growth of

mou table

reviewed by the operational boards of the WLCG Collaboration.

Service	*Maximum delay in responding to operational problems*			*Average availability[6] measured on an annual basis*	
	Service interruption	Degradation of the capacity of the service by more than 50%	Degradation of the capacity of the service by more than 20%	During accelerator operation	At all other times
Acceptance of data from the Tier-0 Centre during accelerator operation	12 hours	12 hours	24 hours	99%	n/a
Networking service to the Tier-0 Centre during accelerator operation	12 hours	24 hours	48 hours	98%	n/a
Data-intensive analysis services, including networking to Tier-0, Tier-1 Centres outwith accelerator operation	24 hours	48 hours	48 hours	n/a	98%
All other services[7] – prime service hours[10]	2 hour	2 hour	4 hours	98%	98%
All other services[7] – outwith prime service hours[10]	24 hours	48 hours	48 hours	97%	97%

The response times in the above table refer only to the maximum delay before action