
CERN

CERN Site Report*

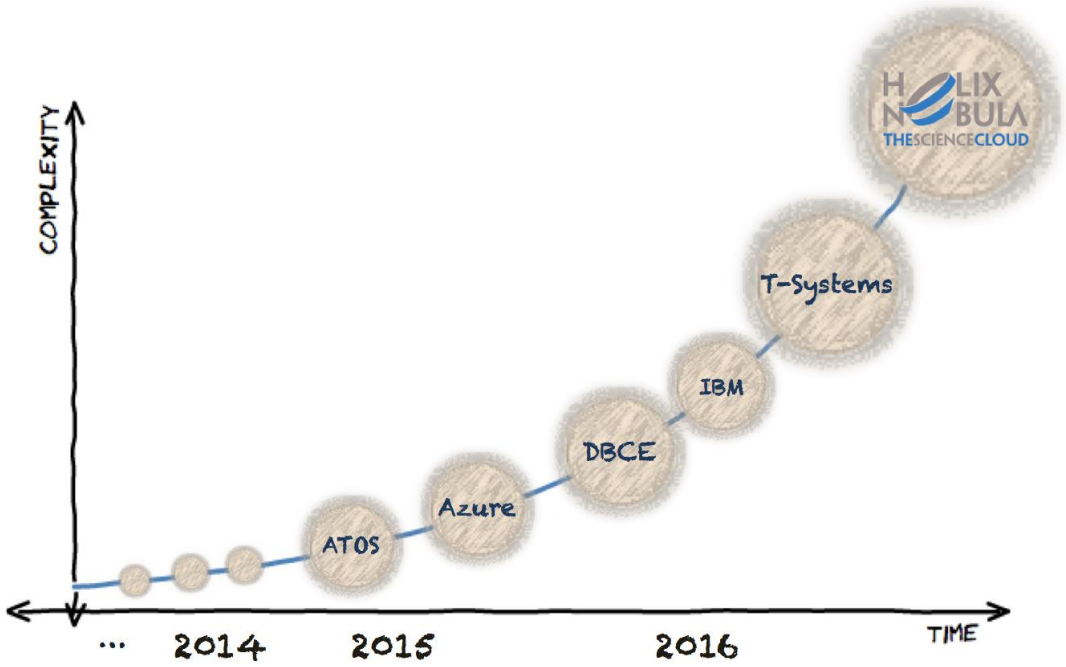
Oliver Keeble & Andrea Manzi on behalf of the DPM team

* Yes, really!

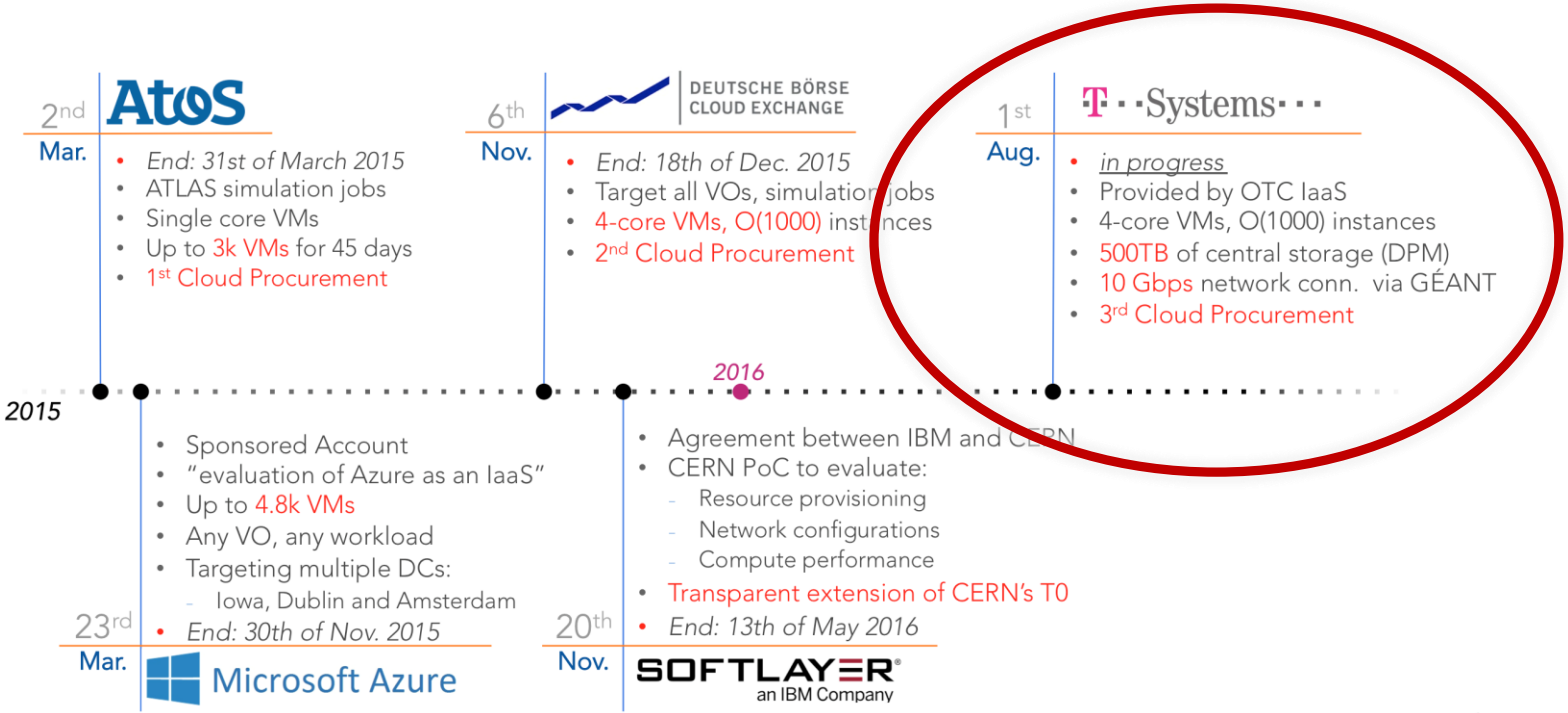
Helix Nebula & CERN cloud

Started in 2011 with the EC funded project Helix-Nebula

Since 2015, series of short CERN procurement projects of increasing size and complexity



CERN cloud projects



The procurement

- 90 days
- 4000 cores
- 1000 VMs
- 500TB block storage
- 10Gb/s uplink to CERN

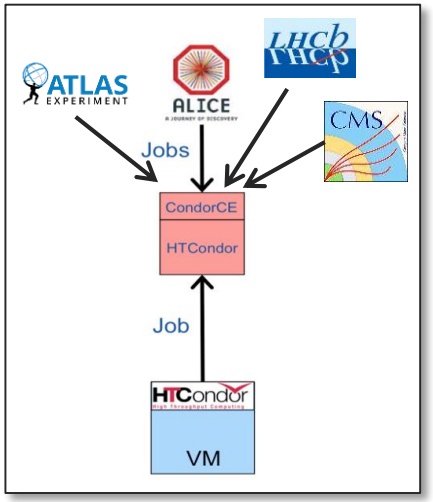
Transparent Extension of CERN Resources

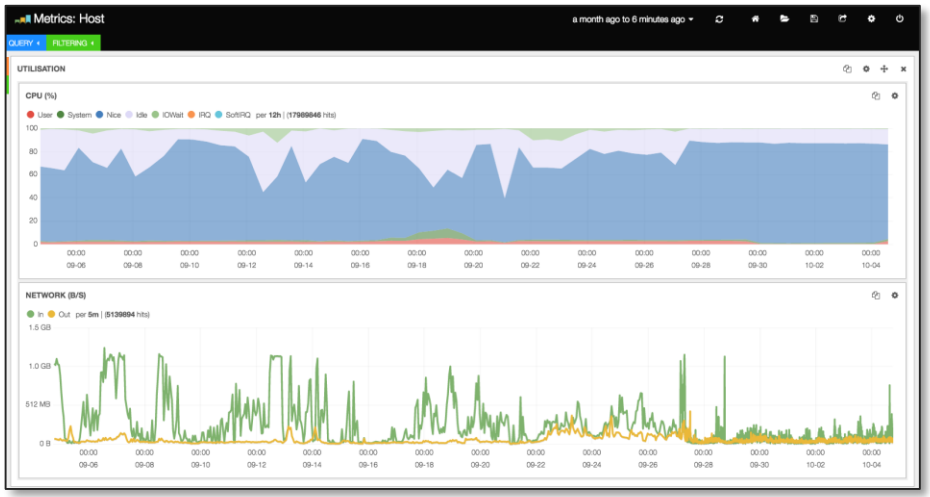
Consolidate the strategies adopted in the past cloud activities

- Manage and exploit external resources using same toolset and entry points as CERN on premises resources
  - *Puppet* configuration
  - HTCondor for scheduling and match-making
  - Infrastructure monitoring

[see CHEP p-22]

- Adopted *Terraform* for VM lifecycle management (N.B.: looking for long VM lifetime)
  - Open source toolkit, supports several cloud providers





The standard VM

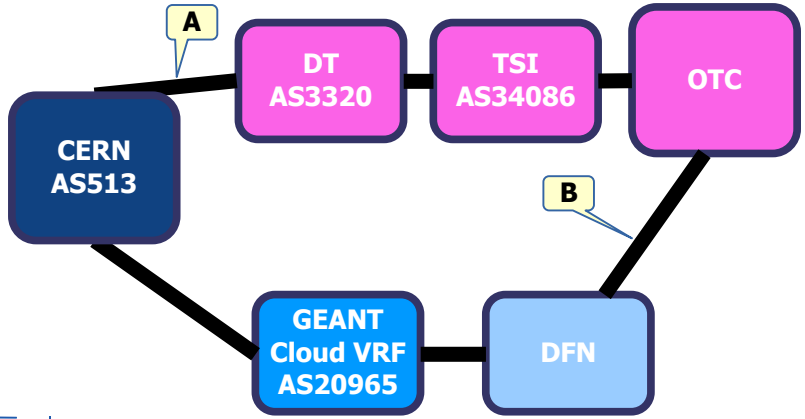
- 4 cores
- 100GB disk (networked block device)
  - 1k iops, 50MB/s streaming
- 1Gb/s “east-west” (ie LAN to workers)
- 500Mb/s inbound from CERN
- 300Mb/s outbound to CERN
- For data, “SSD”s were available
  - 20k iops, >200MB/s streaming

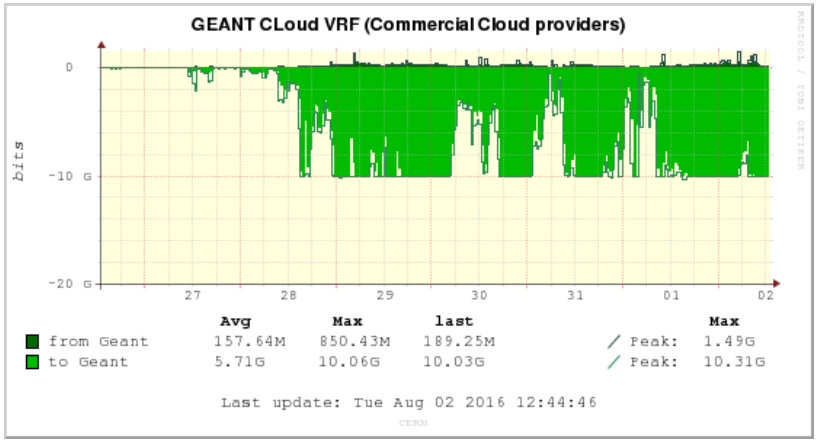
WAN connectivity over GÉANT

Requirement for CSP since the first procurement (early ’15)

GÉANT Cloud VRF is currently connecting CERN and T-Systems (via DFN)

- 10 Gbps of total reserved peak bandwidth available
- The VRF is configured to only allow traffic between CPs and NRENs; no CP-CP traffic is allowed

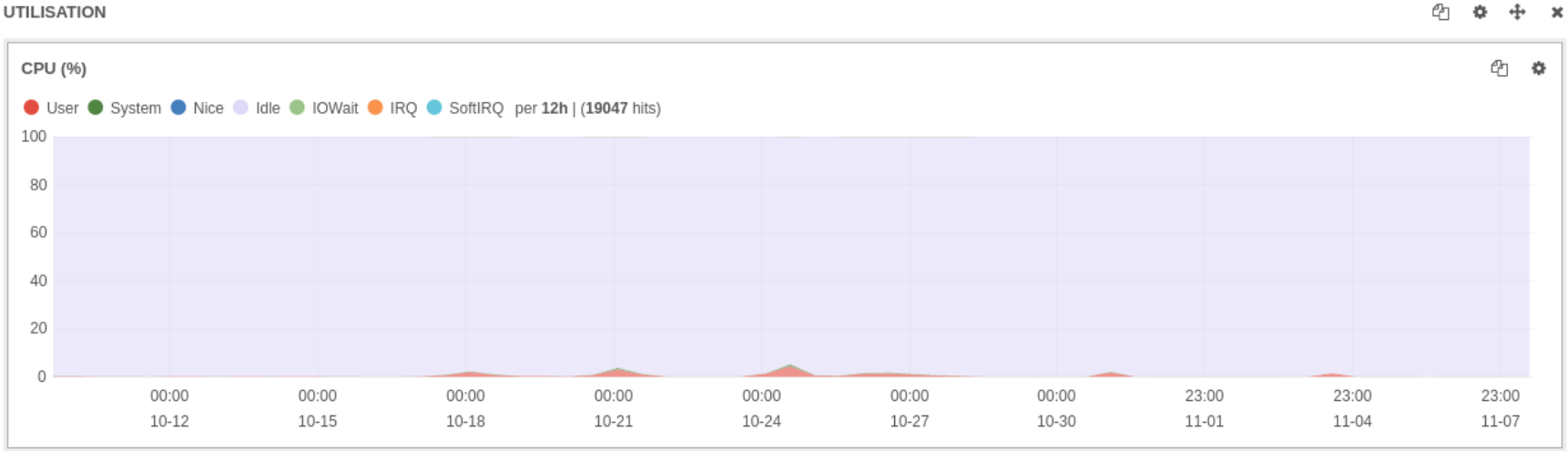




Dimensioning the service

- Assume ¾ of WNs access data
- 48 disk servers ~ 16Mb/s per WN core
  - ~2MB/s
  - ~ The largest figure we had been quoted was 2.5MB/s (Alice)
- 500TB/48 per disk server
  - 2 4.9TB block devices
- IOPS…
  - Hmm, put the db on an SSD and hope for the best
- Alice
  - Needed special monitoring (apmon) so a separate instance for them

xdpmhn01 cpu %

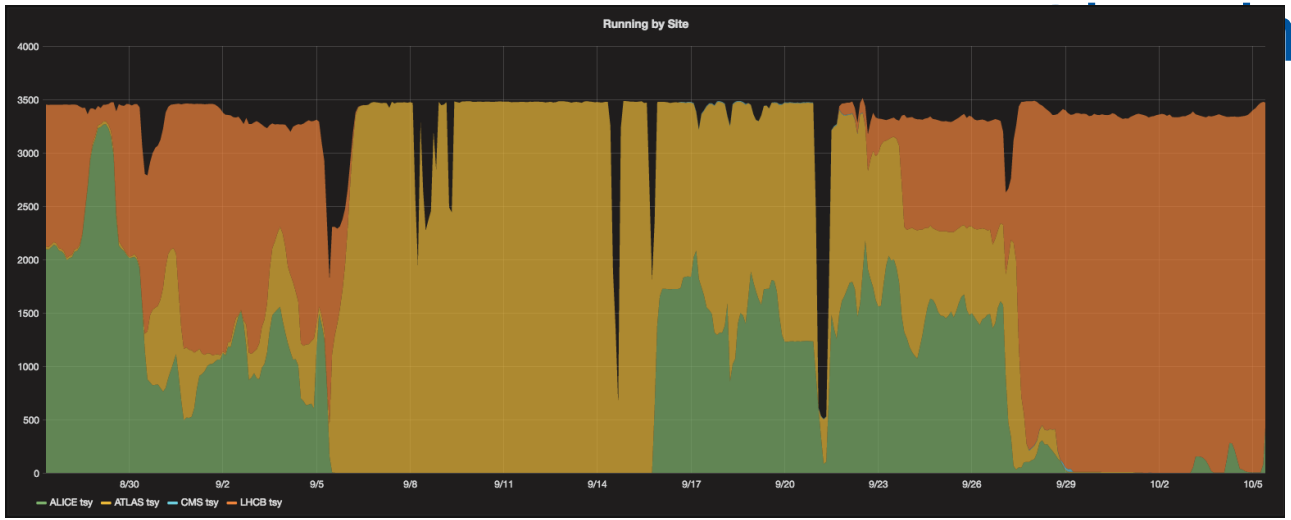


Commissioning

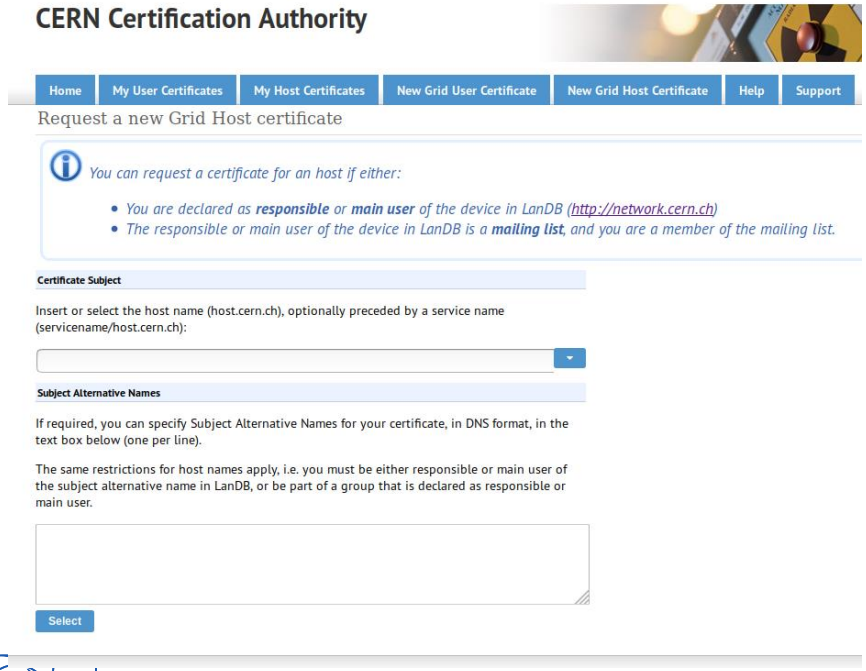
- Everything puppetised
  - gridFTP redirection ON
  - Ext4 (was this the right choice)?
  - Single head node with everything on it
    - SSD for db
  - Still some final tweaks required
    - Atlas STs, ACLs etc

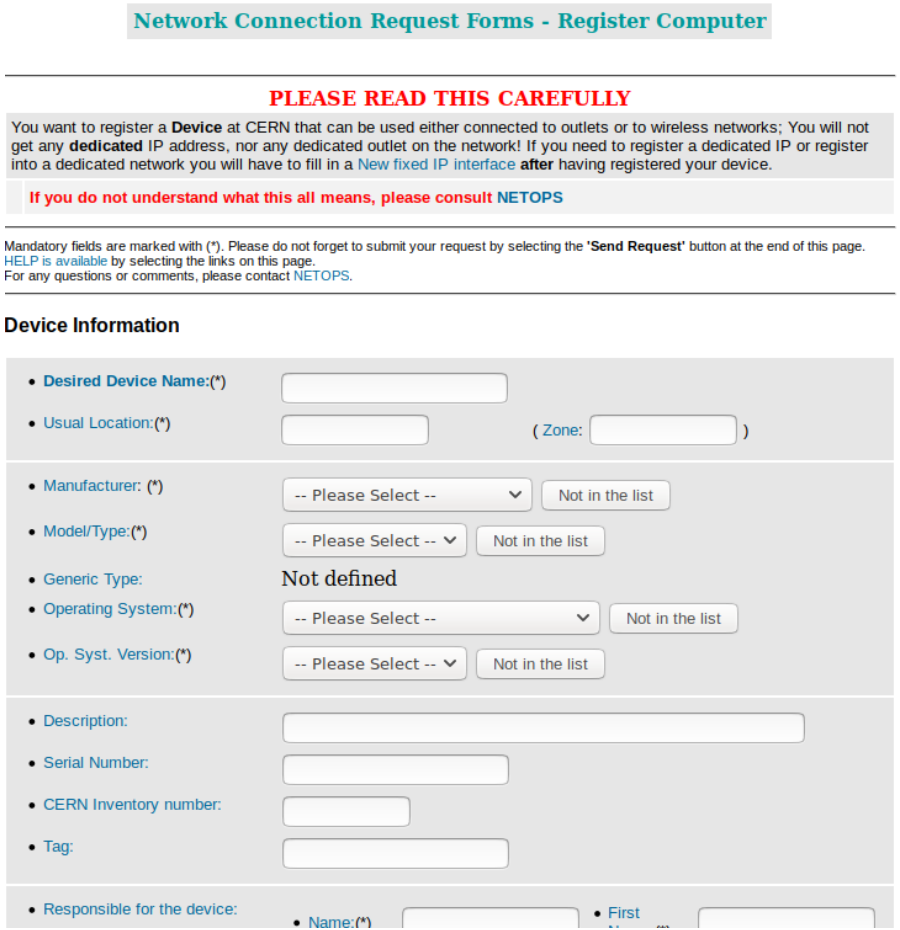
The use cases

- Alice – sim + reco
- Atlas – sim + digireco
- CMS – sim + digireco



Challenge #1: subdomains





Challenge #2 : NAT

- T-systems implemented one-to-one NAT
  - Each host has its own public/private mapping
- xrootd
  - gfal-ls root://xdpmhn01.tsy.cern.ch/dpm/tsy.cern.ch/home/dteam
  - xrdfs root://xdpmhn01.tsy.cern.ch/ ls /dpm/cern.ch/home/dteam
    - kXR_locate request, which results in their redirection to a private IP address
- GridFTP
  - Extra config for redirection

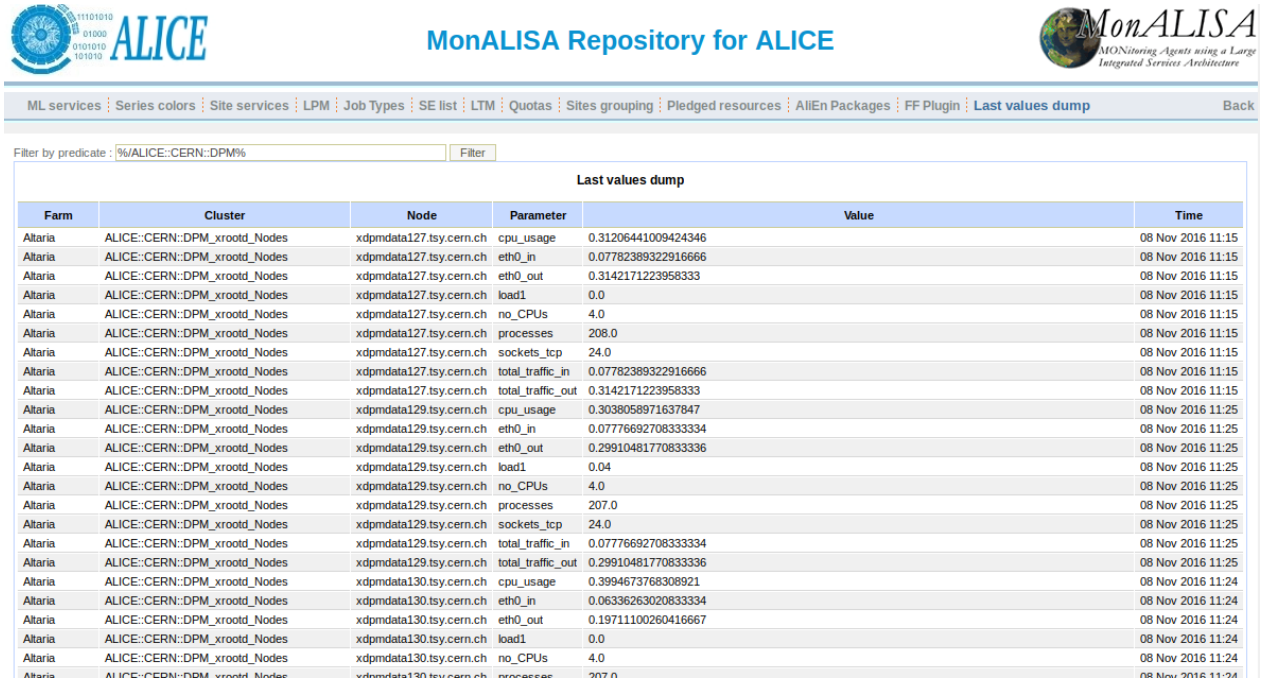
Challenge #3 : reverse DNS

- GSI (Globus security) requires reverse DNS to be configured for servers
- We needed 4 DNS services in 3 places
  - Forward and reverse private IP
    - -> deployed in the cloud
  - Forward public IP
    - -> CERN DNS
  - Reverse public IP
    - -> T-systems
    - …they weren’t expecting this!

Challenge #4 : Configuration

- Reboots of services had side effects
  - Losing hostname
    - Frontends didn’t start
  - Related to learning the VM management API, tackling cloud-init, managing DNS…
- Stabilised in the end
  - … but puppet doesn’t erase history
    - Nodes are only “eventually identical”

Challenge #5 : Monitoring



MonALISA Repository for ALICE

ML services | Series colors | Site services | LPM | Job Types | SE list | LTM | Quotas | Sites grouping | Pledged resources | AliEn Packages | FF Plugin | Last values dump | Back

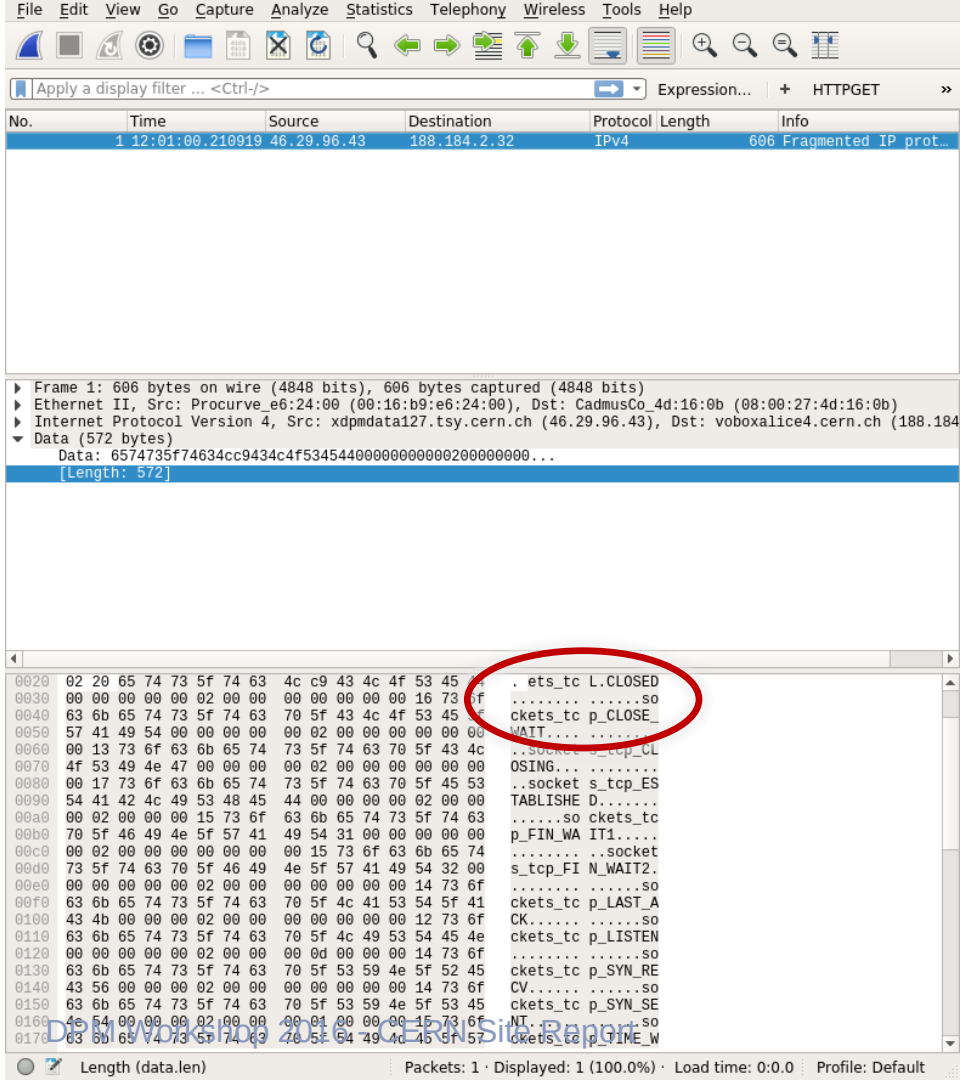
Filter by predicate : %/ALICE::CERN::DPM% Filter

Last values dump

Farm	Cluster	Node	Parameter	Value	Time
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata127.tsy.cern.ch	cpu_usage	0.3120644100942434 6	08 Nov 2016 11:15
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata127.tsy.cern.ch	eth0_in	0.07782389322916666	08 Nov 2016 11:15
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata127.tsy.cern.ch	eth0_out	0.3142171223958333	08 Nov 2016 11:15
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata127.tsy.cern.ch	load1	0.0	08 Nov 2016 11:15
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata127.tsy.cern.ch	no_CPUs	4.0	08 Nov 2016 11:15
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata127.tsy.cern.ch	processes	208.0	08 Nov 2016 11:15
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata127.tsy.cern.ch	sockets_tcp	24.0	08 Nov 2016 11:15
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata127.tsy.cern.ch	total_traffic_in	0.07782389322916666	08 Nov 2016 11:15
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata127.tsy.cern.ch	total_traffic_out	0.3142171223958333	08 Nov 2016 11:15
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata129.tsy.cern.ch	cpu_usage	0.3038058971637847	08 Nov 2016 11:25
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata129.tsy.cern.ch	eth0_in	0.07776692708333334	08 Nov 2016 11:25
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata129.tsy.cern.ch	eth0_out	0.29910481770833336	08 Nov 2016 11:25
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata129.tsy.cern.ch	load1	0.04	08 Nov 2016 11:25
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata129.tsy.cern.ch	no_CPUs	4.0	08 Nov 2016 11:25
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata129.tsy.cern.ch	processes	207.0	08 Nov 2016 11:25
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata129.tsy.cern.ch	sockets_tcp	24.0	08 Nov 2016 11:25
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata129.tsy.cern.ch	total_traffic_in	0.07776692708333334	08 Nov 2016 11:25
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata129.tsy.cern.ch	total_traffic_out	0.29910481770833336	08 Nov 2016 11:25
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata130.tsy.cern.ch	cpu_usage	0.3994673768308921	08 Nov 2016 11:24
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata130.tsy.cern.ch	eth0_in	0.06336263020833334	08 Nov 2016 11:24
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata130.tsy.cern.ch	eth0_out	0.19711100260416667	08 Nov 2016 11:24
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata130.tsy.cern.ch	load1	0.0	08 Nov 2016 11:24
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata130.tsy.cern.ch	no_CPUs	4.0	08 Nov 2016 11:24
Altaria	ALICE::CERN::DPM_xrootd_Nodes	xdpmdata130.tsy.cern.ch	processes	207.0	08 Nov 2016 11:24

Challenge #5 : Monitoring

- Both apmon and xrootd send UDP packets
- These were disappearing – traced to corruption from virtual switch on hypervisor
- Fixed with hypervisor patch



Challenge #6 : network i/o

- Why does xdpmdata104 have hundreds of clients connected while the others have 1 or 2?
- Why is its network throughput 50MB/s rather that 100MB/s?
- Something happened to this node previously and it built up a huge queue… but what?
- …unresolved

Challenge #7 : checksums

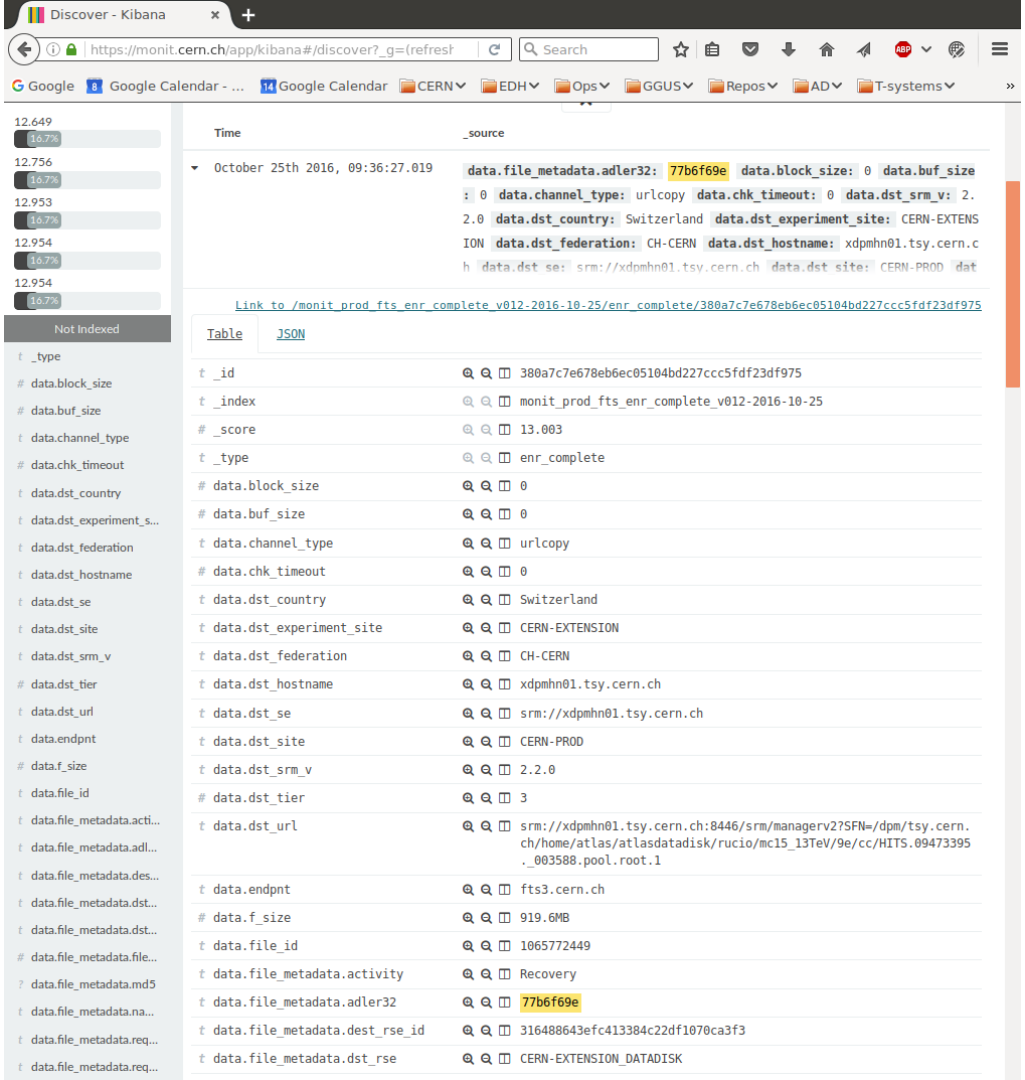
- Atlas reported checksum problems copying from DPM to the Worker

```
!!WARNING!!2990!! Remote and local
checksums (of type adler32) do not match
for HITS.09458365._000184.pool.root.1
(cd88ab28 != a522d6aa)
```

- There were around 20 problematic files, all had been transferred in during the week before (via both gridFTP & xroot)

Challenge #7 :

- We could verify
  - File was transferred successfully with checksum
  - mtime on disk is the same as upload time
  - mtime in DPM db is the same as upload time
- But…
  - A single 4096 byte block was different!
- Status: traced to defective SSD



Things I didn’t mention

- Experiment experience integrating a new system
  - A slightly different kind of thing, as compute was transparently extended, storage was not
- All the work done by IT-CM to integrate batch, puppet, monitoring…

Conclusions

- DPM is deployable in the cloud
  - Even with NAT, subdomains etc it's possible
- However, a cloud is not your own computer centre
  - Debugging can involve numerous parties
- It takes a while to amortise the overheads of commissioning a storage system
  - One has to consider carefully how best to spend "cloud money" on CPU/storage/network