



# CERN Site Report\*

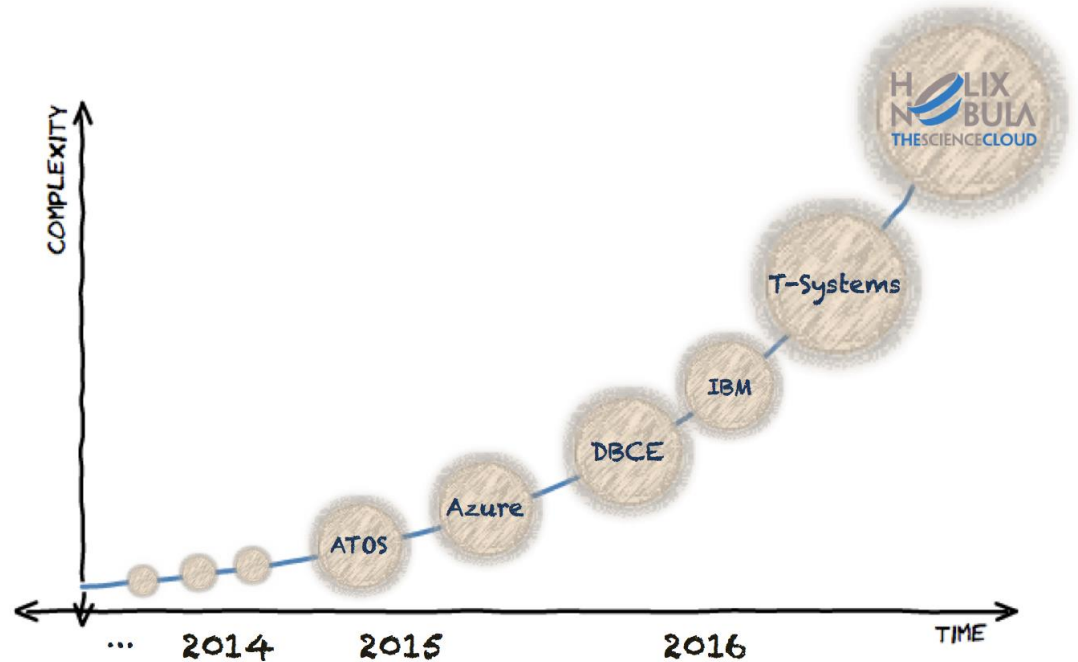
Oliver Keeble & Andrea Manzi on behalf of the DPM team

\* Yes, really!

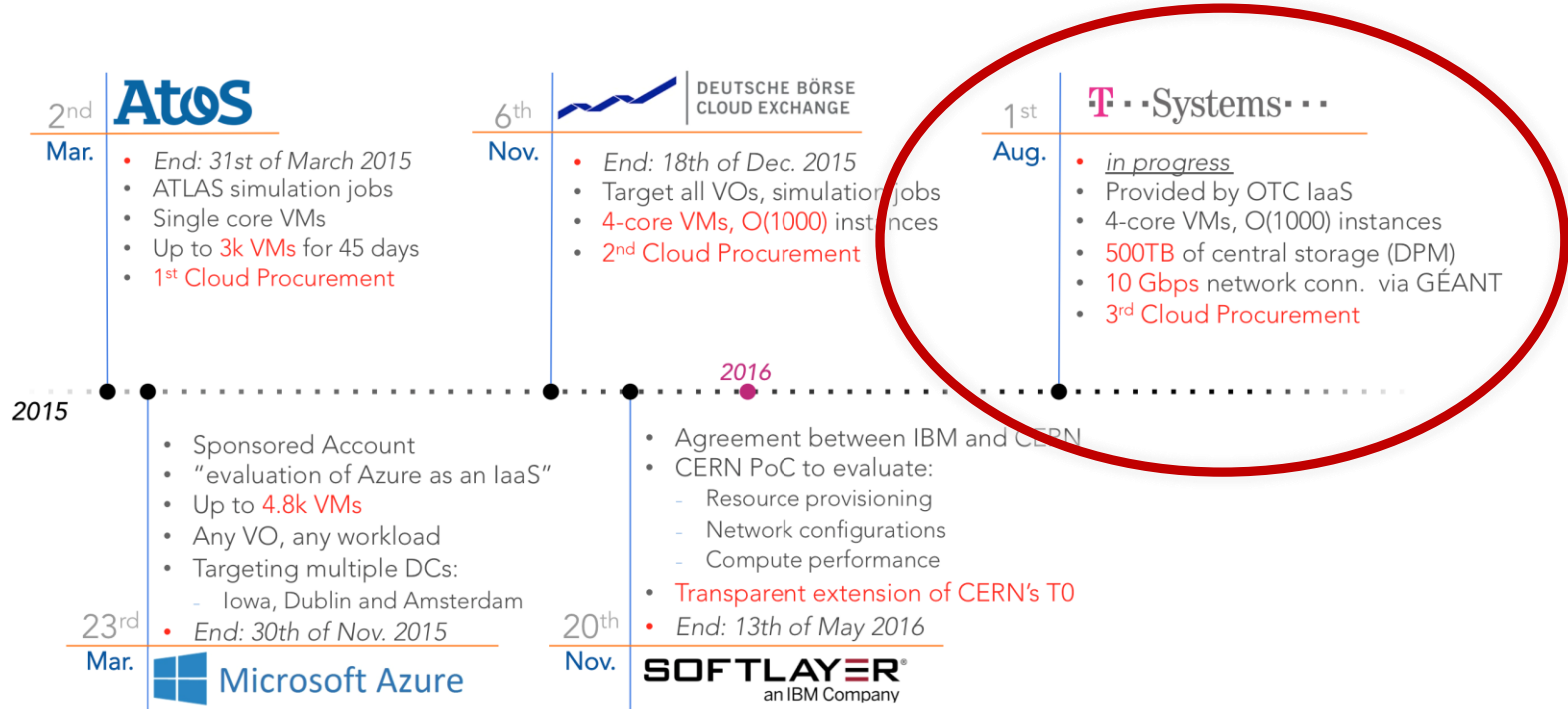
# Helix Nebula & CERN cloud

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

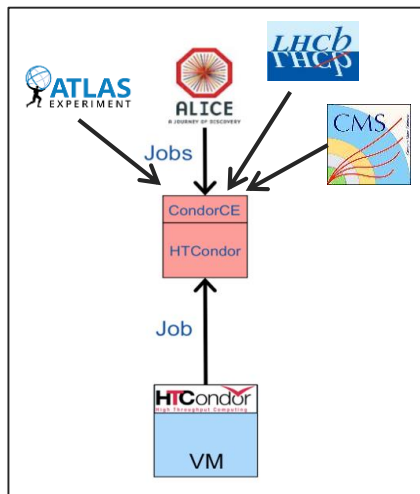
- 90 days
- 4000 cores
- 1000 VMs
- 500TB block storage
- 10Gb/s uplink to CERN

# Transparent Extension of CERN Resources

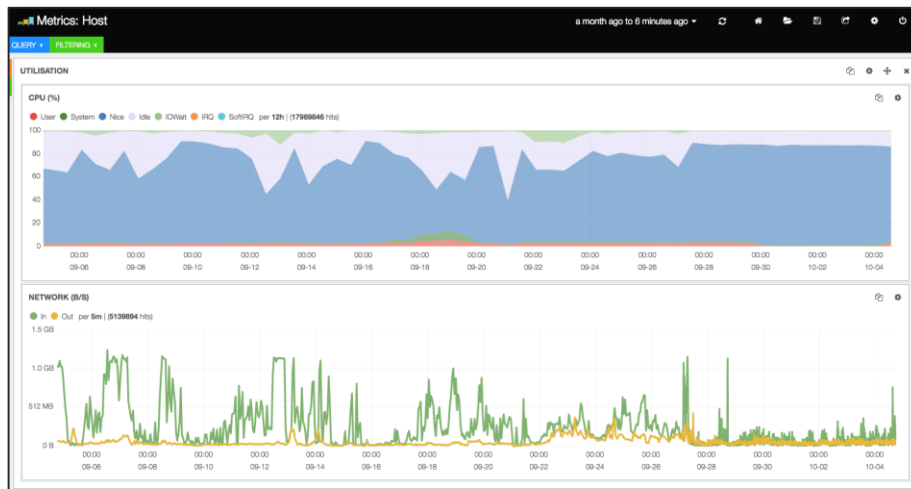
Consolidate the strategies adopted in the past cloud activities

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

[see CHEP p-22]



23/11/2016



DPM Workshop 2016 - CERN Site Report

# The standard VM

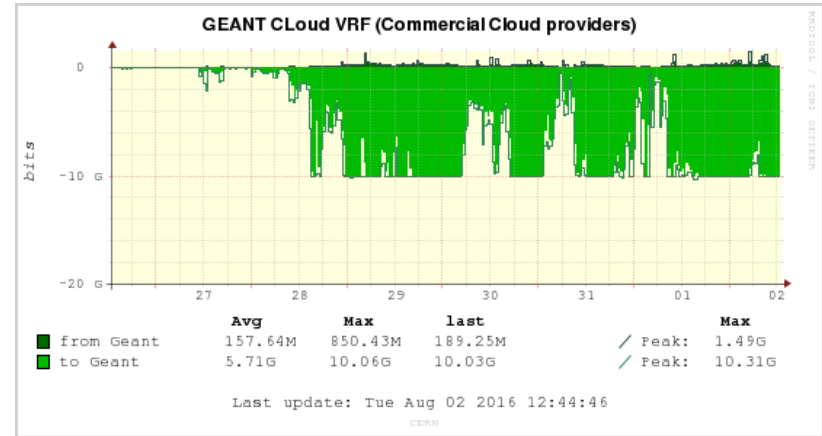
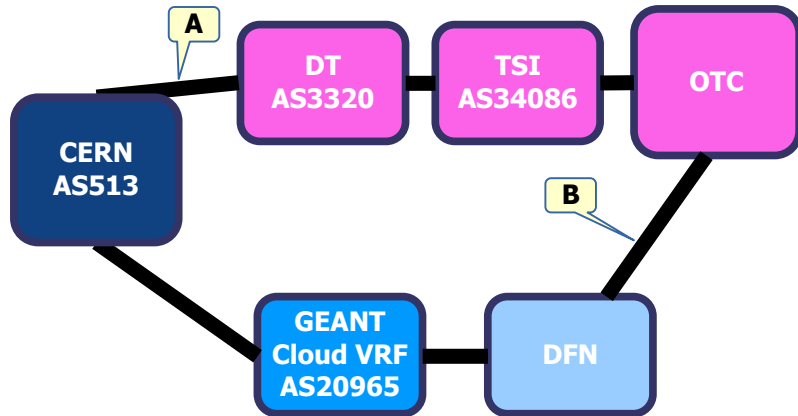
- 4 cores
- 100GB disk (networked block device)
  - 1k iops, 50MB/s streaming
- 1Gb/s “east-west” (ie LAN to workers)
- 500Mb/s inbound from CERN
- 300Mb/s outbound to CERN
- For data, “SSD”s were available
  - 20k iops, >200MB/s streaming

# 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



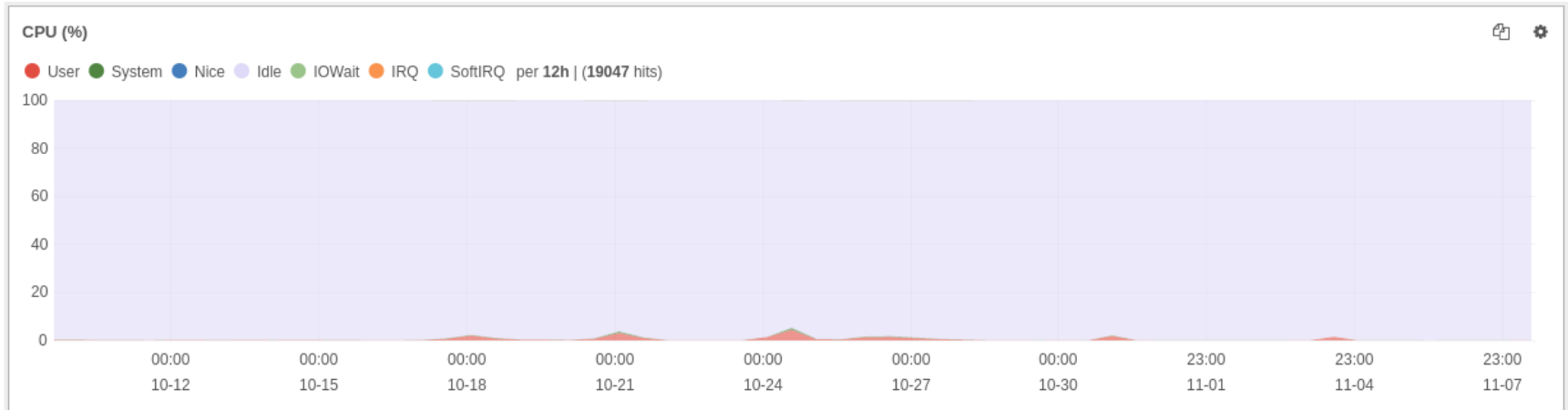


# Dimensioning the service

- Assume  $\frac{3}{4}$  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 %

## UTILISATION

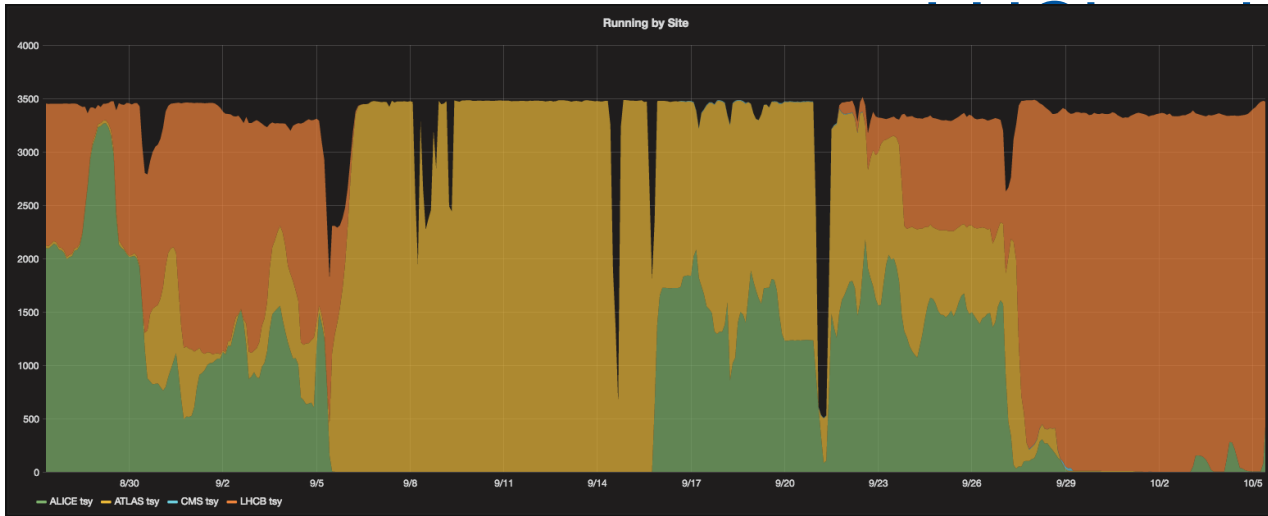


# Commissioning

- 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
- LHCb – sim + digireco



# Challenge #1: subdomains

## CERN Certification Authority

Home My User Certificates My Host Certificates New Grid User Certificate New Grid Host Certificate Help Support

Request a new Grid Host certificate



You can request a certificate for an host if either:

- You are declared as **responsible** or **main user** of the device in LanDB (<http://network.cern.ch>)
- The responsible or main user of the device in LanDB is a **mailing list**, and you are a member of the mailing list.

### Certificate Subject

Insert or select the host name (host.cern.ch), optionally preceded by a service name (servicename/host.cern.ch):

### Subject Alternative Names

If required, you can specify Subject Alternative Names for your certificate, in DNS format, in the text box below (one per line).

The same restrictions for host names apply, i.e. you must be either responsible or main user of the subject alternative name in LanDB, or be part of a group that is declared as responsible or main user.

Select

## Network Connection Request Forms - Register Computer

### PLEASE READ THIS CAREFULLY

You want to register a **Device** at CERN that can be used either connected to outlets or to wireless networks; You will not get any **dedicated** IP address, nor any dedicated outlet on the network! If you need to register a dedicated IP or register into a dedicated network you will have to fill in a [New fixed IP interface](#) **after** having registered your device.

**If you do not understand what this all means, please consult NETOPS**

Mandatory fields are marked with (\*). Please do not forget to submit your request by selecting the 'Send Request' button at the end of this page. [HELP](#) is available by selecting the links on this page. For any questions or comments, please contact [NETOPS](#).

### Device Information

• <b>Desired Device Name: (*)</b>	<input type="text"/>
• <b>Usual Location: (*)</b>	<input type="text"/> ( Zone: <input type="text"/> )
• <b>Manufacturer: (*)</b>	-- Please Select -- <input type="button" value="Not in the list"/>
• <b>Model/Type: (*)</b>	-- Please Select -- <input type="button" value="Not in the list"/>
• <b>Generic Type:</b>	<b>Not defined</b>
• <b>Operating System: (*)</b>	-- Please Select -- <input type="button" value="Not in the list"/>
• <b>Op. Syst. Version: (*)</b>	-- Please Select -- <input type="button" value="Not in the list"/>
• <b>Description:</b>	<input type="text"/>
• <b>Serial Number:</b>	<input type="text"/>
• <b>CERN Inventory number:</b>	<input type="text"/>
• <b>Tag:</b>	<input type="text"/>
• <b>Responsible for the device:</b>	• <b>Name: (*)</b> <input type="text"/> • <b>First</b> <input type="text"/>



23/11/2016

DPM Workshop 2016 - CERN Site Report

# 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 | AllEn 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	0.31206441009424346	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

The screenshot shows a Wireshark interface with a single packet selected. The packet list pane shows:

No.	Time	Source	Destination	Protocol	Length	Info
1	12:01:00.210919	46.29.96.43	188.184.2.32	IPv4	606	Fragmented IP prot...

The packet details pane shows:

- ▶ Frame 1: 606 bytes on wire (4848 bits), 606 bytes captured (4848 bits)
- ▶ Ethernet II, Src: Procurve\_e6:24:00 (00:16:b9:e6:24:00), Dst: CadmusCo\_4d:16:0b (08:00:27:4d:16:0b)
- ▶ Internet Protocol Version 4, Src: xdpmdat127.tsy.cern.ch (46.29.96.43), Dst: voboxalice4.cern.ch (188.184.2.32)
- ▶ Data (572 bytes)
- Data: 6574735f74634cc9434cf534544000000000200000000...
- [Length: 572]

The packet bytes pane shows a hex dump and ASCII representation. A red circle highlights the ASCII text: `..ets_tc L.CLOSED`.

At the bottom of the interface, the status bar shows: Length (data.len) | Packets: 1 · Displayed: 1 (100.0%) · Load time: 0:0:0 | Profile: Default



# 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 than 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

The screenshot shows a web browser window with the URL `https://monit.cern.ch/app/kibana#/discover?_g=(refres`. The page displays a search result for a file with the following metadata:

```
data.file_metadata.adler32: 77b6f69e data.block_size: 0 data.buf_size: 0 data.channel_type: urlcopy data.chk_timeout: 0 data.dst_srm_v: 2.2.0 data.dst_country: Switzerland data.dst_experiment_site: CERN-EXTENSION data.dst_federation: CH-CERN data.dst_hostname: xdpmhn01.tsy.cern.ch data.dst_se: srm://xdpmhn01.tsy.cern.ch data.dst_site: CERN-PROD data...
```

Below the metadata, there is a link to the file: `Link to /monit_prod_fts_enr_complete_v012-2016-10-25/enr_complete/380a7c7e678eb6ec05104bd227ccc5fdf23df975`.

The table below shows the file metadata details:

Table	JSON
<code>t_id</code>	<code>380a7c7e678eb6ec05104bd227ccc5fdf23df975</code>
<code>t_index</code>	<code>monit_prod_fts_enr_complete_v012-2016-10-25</code>
<code>#_score</code>	<code>13.003</code>
<code>t_type</code>	<code>enr_complete</code>
<code>#data.block_size</code>	<code>0</code>
<code>#data.buf_size</code>	<code>0</code>
<code>t data.channel_type</code>	<code>urlcopy</code>
<code>#data.chk_timeout</code>	<code>0</code>
<code>t data.dst_country</code>	<code>Switzerland</code>
<code>t data.dst_experiment_site</code>	<code>CERN-EXTENSION</code>
<code>t data.dst_federation</code>	<code>CH-CERN</code>
<code>t data.dst_hostname</code>	<code>xdpmhn01.tsy.cern.ch</code>
<code>t data.dst_se</code>	<code>srm://xdpmhn01.tsy.cern.ch</code>
<code>t data.dst_site</code>	<code>CERN-PROD</code>
<code>t data.dst_srm_v</code>	<code>2.2.0</code>
<code>#data.dst_tier</code>	<code>3</code>
<code>t data.dst_url</code>	<code>srm://xdpmhn01.tsy.cern.ch:8446/srm/manager/v2?SFN=/dpm/tsy.cern.ch/home/atlas/atlasdatadisk/rucio/mc15_13TeV/9e/cc/HITS.09473395_003588.pool.root.1</code>
<code>t data.endpnt</code>	<code>fts3.cern.ch</code>
<code>#data.f_size</code>	<code>919.6MB</code>
<code>t data.file_id</code>	<code>1065772449</code>
<code>t data.file_metadata.activity</code>	<code>Recovery</code>
<code>t data.file_metadata.adler32</code>	<code>77b6f69e</code>
<code>t data.file_metadata.dest_rse_id</code>	<code>316488643efc413384c22df1070ca3f3</code>
<code>t data.file_metadata.dst_rse</code>	<code>CERN-EXTENSION_DATADISK</code>



23/11/2016

# 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