

### **CERN Site Report\***

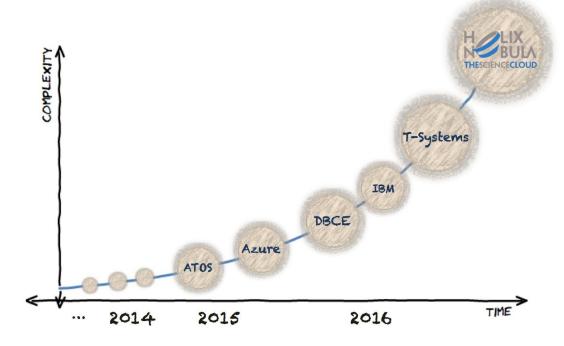
Oliver Keeble & Andrea Manzi on behalf of the DPM team

\* Yes, really!

### Helix Nebula & CERN cloud

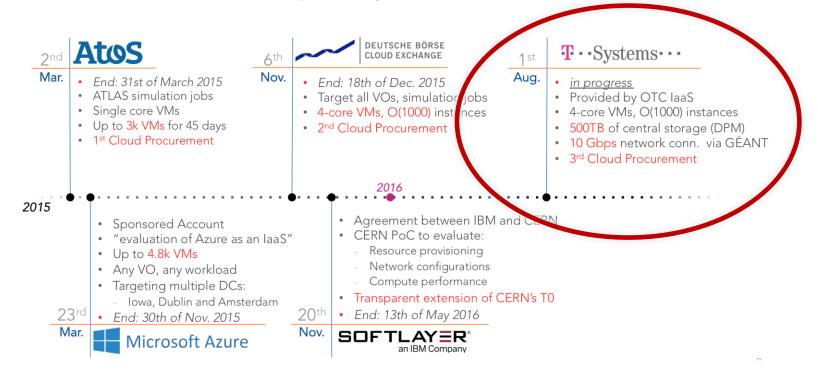
# Started in 2011 with the EC funded project Helix-Nebula

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



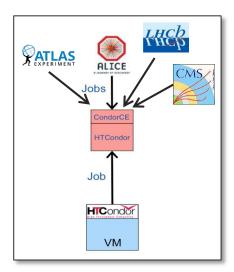
### **Transparent Extension of CERN Resources**

### Consolidate the strategies adopted in the past cloud activities

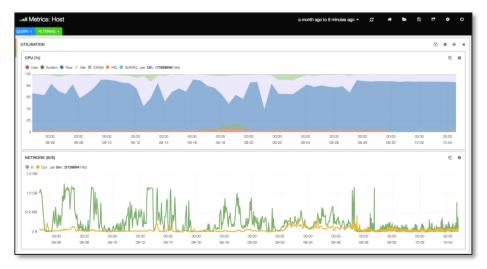
- Manage and exploit external resources using same toolset and entry points as CERN on premises resources
  - *Puppet* configuration
  - HTCondor for scheduling and match-making
  - Infrastructure monitoring

[see CHEP p-22]

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



23/11/2016





### The standard VM

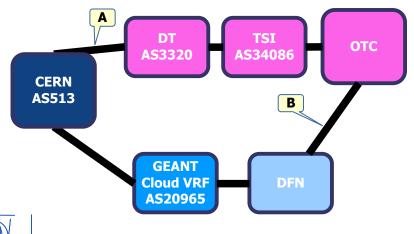
- 4 cores
- 100GB disk (networked block device)
  - 1k iops, 50MB/s streaming
- 1Gb/s "east-west" (ie LAN to workers)
- 500Mb/s inbound from CERN
- 300Mb/s outbound to CERN
- For data, "SSD"s were available
  - 20k iops, >200MB/s streaming

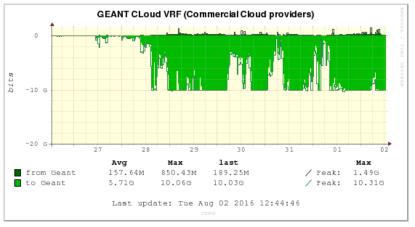


# WAN connectivity over GÉANT

Requirement for CSP since the first procurement (early '15) <u>GÉANT Cloud VRF</u> 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





DPM Workshop 2016 - CERN Site Report



23/11/2016

# 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 %

### ሮካ UTILISATION 20 CPU (%) ÷ ● User ● System ● Nice ● Idle ● IOWait ● IRQ ● SoftIRQ per 12h | (19047 hits) 100 80 60 40 20 0 00:00 00:00 00:00 00:00 23:00 00:00 00:00 00:00 23:00 23:00 10-12 10-15 10-18 10-21 11-01 11-04 10-24 10-27 10-30 11-07



### 23/11/2016

# Commissioning

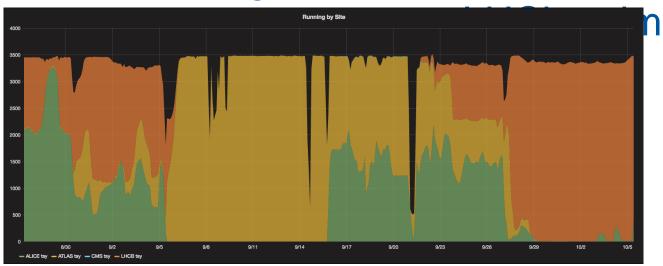
- Everything puppetised
  - gridFTP redirection ON
  - Ext4 (was this the right choice)?
  - Single head node with everything on it
    - SSD for db
  - Still some final tweaks required
    - Atlas STs, ACLs etc



### The use cases

- Alice sim + reco
- Atlas sim + digireco

CMS – sim +
 digireco





23/11/2016

# Challenge #1: subdomains

### **CERN Certification Authority**



Help

Support

New Grid Host Certificate

### Home My User Certificates My Host Certificates

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.

New Grid User Certificate

### Certificate Subject

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

### Subject Alternative Names

Select

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.

| Network Connection Request Forms - Register Compute |
|---|
|---|

### 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

| Desired Device Name:(*)     Usual Location:(*) | (Zone: )                        |
|--|---------------------------------|
| Manufacturer: (*)                              | Please Select   Not in the list |
| Model/Type:(*)                                 | Please Select V Not in the list |
| Generic Type:                                  | Not defined                     |
| Operating System:(*)                           | Please Select   Not in the list |
| Op. Syst. Version:(*)                          | Please Select V Not in the list |
| Description:                                   |                                 |
| Serial Number:                                 |                                 |
| CERN Inventory number:                         |                                 |
| • Tag:   |                                 |
| Responsible for the device:                    | • Name:(*) • First              |

### DPM Workshop 2016 - CERN Site Report

23/11/2016

# 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



| er by predic     | ate : %/ALICE::CERN::DPM%     | Filter                  |                   |                     |               |  |
|------------------|-------------------------------|-------------------------|-------------------|---------------------|---------------|--|
| Last values dump |                               |                         |                   |                     |               |  |
| Farm             | Cluster                       | Node                    | Parameter         | Value               | Time          |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata127.tsy.cern.ch | cpu_usage         | 0.31206441009424346 | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata127.tsy.cern.ch | eth0_in           | 0.07782389322916666 | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata127.tsy.cern.ch | eth0_out          | 0.3142171223958333  | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata127.tsy.cern.ch | load1             | 0.0                 | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata127.tsy.cern.ch | no_CPUs           | 4.0                 | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata127.tsy.cern.ch | processes         | 208.0               | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata127.tsy.cern.ch | sockets_tcp       | 24.0                | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata127.tsy.cern.ch | total_traffic_in  | 0.07782389322916666 | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata127.tsy.cern.ch | total_traffic_out | 0.3142171223958333  | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata129.tsy.cern.ch | cpu_usage         | 0.3038058971637847  | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata129.tsy.cern.ch | eth0_in           | 0.07776692708333334 | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata129.tsy.cern.ch | eth0_out          | 0.29910481770833336 | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata129.tsy.cern.ch | load1             | 0.04                | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata129.tsy.cern.ch | no_CPUs           | 4.0                 | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata129.tsy.cern.ch | processes         | 207.0               | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata129.tsy.cern.ch | sockets_tcp       | 24.0                | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata129.tsy.cern.ch | total_traffic_in  | 0.07776692708333334 | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata129.tsy.cern.ch | total_traffic_out | 0.29910481770833336 | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata130.tsy.cern.ch | cpu_usage         | 0.3994673768308921  | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata130.tsy.cern.ch | eth0_in           | 0.06336263020833334 | 08 Nov 2016 1 |  |
| taria            | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata130.tsy.cern.ch | eth0_out          | 0.19711100260416667 | 08 Nov 2016 1 |  |
| taria            | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata130.tsy.cern.ch | load1             | 0.0                 | 08 Nov 2016 1 |  |
| Itaria           | ALICE::CERN::DPM_xrootd_Nodes | xdpmdata130.tsy.cern.ch | no_CPUs           | 4.0                 | 08 Nov 2016 1 |  |
| Itoria           | ALICE-CEDN-DDM vrootd Nodes   | vdnmdata130 tev cern ch | nrocaceae         | 207.0               | 08 Nov 2016 1 |  |



### 23/11/2016

# 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

|   |  |   | <u>A</u> nalyze <u>S</u>   | tatistics Telep   | hon <u>y W</u> ireless   | <u>T</u> ools <u>H</u> elp  |           |                |    |
|---|--|---|--|---|--|---|-----------|----------------|----|
|   |  | 6101<br>0108<br>0108<br>0109  | × 6  | < 🔶   | 2 주 👱  | ∎∎€   | $\Theta$  | 3              |    |
| App   | oly a display filter                                 | <ctrl-></ctrl->   |  |   |  | Express   | on +      | HTTPGET        |    |
| No.   | Time   |   | Source   | Destinat  |  | Protocol Length   | In        |                |    |
|   | 1 12:01:   | 00.210919   | 46.29.96.4   | 13 188.184  | .2.32  | IPv4  | 606 FI    | ragmented IP p | ro |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
| L Era   | me 1: 606 byte                                       | s on wire   | (4848 hits   | a) 606 hytes  | cantured (4848   | R hite)   |           |                |    |
| Eth   | ernet II, Src:                                       | Procurve_   | e6:24:00 (   | 00:16:b9:e6:2   | 4:00), Dst: Ca   | admusCo_4d:16:0   |           |                |    |
|   |  | Version 4   | , Src: xdp   | omdata127.tsy.  | cern.ch (46.29   | 9.96.43), Dst:  | voboxalic | e4.cern.ch (18 | 8. |
|   | a (572 bytes)<br>Data: 6574735f                      | 74634cc943  | 4c4f534544   | 000000000000000000000000000000000000000   | 000000   |   |           |                |    |
|   | [Length: 572]  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   |  |   |  |   |           |                |    |
|   |  |   | 4 0 - 40   |   |  | 010055  |           |                |    |
| 0020  | 02 20 65 74 73<br>00 00 00 00 00                     |   |  | 4c 4f 53 45   |  | L.CLOSED  |           |                |    |
| 0020<br>0030<br>0040  | 00 00 00 00 00<br>63 6b 65 74 73                     | ) 02 00 00<br>3 5f 74 63  | 00 00 00<br>70 5f 43   | 00 00 16 73<br>4c 4f 53 45  | ff<br>ckets_tc   | so<br>p_CLOSE_  |           |                |    |
| 0020<br>0030<br>0040<br>0050  | 00 00 00 00 00<br>63 6b 65 74 73<br>57 41 49 54 00   | ) 02 00 00<br>3 5f 74 63<br>) 00 00 00  | 00 00 00<br>70 5f 43<br>00 02 00   | 00 00 16 73<br>4c 4f 53 45<br>00 00 00 00 0   | f<br>€ ckets_tc<br>00 WAIT   | p_CLOSE_  |           |                |    |
| 0020<br>0030<br>0040  | 00 00 00 00 00<br>63 6b 65 74 73                     | 0 02 00 00<br>5f 74 63<br>0 00 00 00<br>8 6b 65 74  | 00 00 00<br>70 5f 43<br>00 02 00<br>73 5f 74   | 00 00 16 73<br>4c 4f 53 45  | f<br>f ckets_tc<br>00 WAIT<br>cSocket  | so<br>p_CLOSE_<br>  |           |                |    |
| 0020<br>0030<br>0040<br>0050<br>0060<br>0070<br>0080  | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 0 02 00 00<br>5 f 74 63<br>0 00 00 00<br>6 b 65 74<br>7 00 00 00<br>8 6b 65 74  | 00 00 00<br>70 5f 43<br>00 02 00<br>73 5f 74<br>00 02 00<br>73 5f 74   | 00 00 16 73<br>4c 4f 53 45<br>00 00 00 00 00<br>63 70 5f 43<br>00 00 00 00 00<br>63 70 5f 45  | ckets_tc<br>WAIT<br>CSOCKET<br>OO OSING<br>S3socket  | so<br>p_CLOSE_<br><br>s_tcp_CL<br>s_tcp_ES  |           |                |    |
| 0020<br>0030<br>0040<br>0050<br>0060<br>0060<br>0070<br>0080<br>0080<br>0090  | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 0 02 00 00<br>5 f 74 63<br>0 00 00 00<br>6 6 65 74<br>7 00 00 00<br>8 6b 65 74<br>9 65 65 74<br>9 53 48 45  | 00 00 00<br>70 5f 43<br>00 02 00<br>73 5f 74<br>00 02 00<br>73 5f 74<br>44 00 00   | 00         00         16         73           4c         4f         53         45           00         00         00         00           63         70         5f         43           00         00         00         00           63         70         5f         43           00         00         00         00           63         70         5f         45           00         00         00         00           63         70         5f         45           00         00         02         00 | of   | so<br>p_CLOSE_<br><br>s_tcp_CL<br>s_tcp_ES<br>D   |           |                |    |
| 0020<br>0030<br>0040<br>0050<br>0060<br>0070<br>0080  | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 0 02 00 00<br>3 5f 74 63<br>0 00 00 00<br>3 6b 65 74<br>7 00 00 00<br>3 6b 65 74<br>9 00 00 00<br>3 6b 65 74<br>9 53 48 45<br>0 15 73 6f  | 00 00 00<br>70 5f 43<br>00 02 00<br>73 5f 74<br>00 02 00<br>73 5f 74<br>44 00 00<br>63 6b 65   | 00 00 16 73<br>4c 4f 53 45<br>00 00 00 00 00<br>63 70 5f 43<br>00 00 00 00 00<br>63 70 5f 45  | ckets_tc<br>WAIT<br>CSOCKET<br>00 OSING<br>33Socket<br>00 TABLISHE<br>53so   | so<br>p_CLOSE_<br><br>s_tcp_CL<br><br>s_tcp_ES<br>D<br>ckets_tc   |           |                |    |
| 0020<br>0030<br>0040<br>0050<br>0060<br>0070<br>0080<br>0090<br>0090<br>0090<br>0020                                | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 0 02 00 00<br>5 5 74 63<br>0 00 00 00<br>6 6 65 74<br>7 00 00 00<br>8 6 65 74<br>9 53 48 45<br>9 15 73 6 f<br>5 57 57 41<br>0 00 00 00  | 00 00 00<br>70 5f 43<br>00 02 00<br>73 5f 74<br>00 02 00<br>73 5f 74<br>44 00 00<br>63 6b 65<br>49 54 31<br>00 15 73   | $\begin{array}{cccccccccccccccccccccccccccccccccccc$  | ckets_tc<br>WAIT<br>00 OSING<br>03socket<br>04 TABLISHE<br>05so<br>06 p_FIN_WA   | socket  |           |                |    |
| 0020<br>0030<br>0040<br>0050<br>0060<br>0070<br>0080<br>0090<br>0080<br>0090<br>0000<br>0000<br>000                 | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | ) 02 00 00<br>5 f 74 63<br>0 00 00 00<br>6 b 65 74<br>7 00 00 00<br>8 6b 65 74<br>9 53 48 45<br>9 15 73 6f<br>9 15 73 6f<br>9 5f 57 41<br>9 00 00 00<br>9 5f 46 49  | 00 00 00<br>70 5f 43<br>00 02 00<br>73 5f 74<br>00 02 00<br>73 5f 74<br>44 00 00<br>63 6b 65<br>49 54 31<br>05 73<br>4e 5f 57  | $\begin{array}{cccccccccccccccccccccccccccccccccccc$  | C ckets_tc<br>0 WAIT<br>10 OSING<br>10 OSING<br>10 CALL<br>10 TABLISHE<br>11 CALL<br>10 P_FIN_WA<br>12 CALL<br>10 S_tcp_FI | so<br>p_CLOSE_<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>CL<br>_CL   |           |                |    |
| 0020<br>0030<br>0040<br>0050<br>0060<br>0070<br>0080<br>0080<br>0080<br>0090<br>0000<br>0000<br>000                 | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 02       00       00         3       5f       74       63         0       00       00       00         3       6b       65       74         0       00       00       00         3       6b       65       74         0       00       00       00         3       6b       65       74         0       15       73       6f         2       5f       57       41         0       00       00       00         0       5f       46       49         0       20       00       00  | $\begin{array}{cccccccccccccccccccccccccccccccccccc$   | $\begin{array}{cccccccccccccccccccccccccccccccccccc$  | if   | so<br>p_CLOSE_<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>tcp_CL<br>t |           |                |    |
| 0020<br>0030<br>0040<br>0050<br>0060<br>0070<br>0080<br>0080<br>0080<br>0090<br>0000<br>0000<br>000                 | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 02     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     15     73       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00       00     00     00   | 00 00 00<br>70 5f 43<br>00 02 00<br>73 5f 74<br>00 02 00<br>73 5f 74<br>44 00 00<br>63 6b 65<br>49 54 31<br>00 15 73<br>4e 5f 57<br>00 00 00<br>70 5f 4C<br>00 00 00                         | $\begin{array}{cccccccccccccccccccccccccccccccccccc$  | if   | so<br>p_CLOSE_<br>CL<br>CL<br>s_tcp_CL<br><br>ckets_tc<br>IT1<br>so<br>p_LAST_A<br>so   |           |                |    |
| 0020<br>0030<br>0040<br>0050<br>0060<br>0070<br>0080<br>0090<br>0080<br>0000<br>0000<br>0000<br>000                 | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 02       00       00         3       5f       74       63         0       00       00       00       00         3       6b       65       74         7       00       00       00       00         8       6b       65       74         9       53       48       45         9       53       48       45         9       00       00       00         9       00       00       00         9       00       00       00         9       00       00       00         9       00       00       00         9       00       00       00         9       02       00       03         9       02       00       03         9       02       03       3         9       02       03       3       5   | $\begin{array}{cccccccccccccccccccccccccccccccccccc$   | $\begin{array}{cccccccccccccccccccccccccccccccccccc$  | if   | so<br>p_CLOSE_<br>tcp_CL<br><br>s_tcp_ES<br>D<br>ckets_tc<br>IT1<br>so<br>p_LAST_A<br>so<br>p_LISTEN  |           |                |    |
| 0020<br>0030<br>0040<br>0050<br>0060<br>0070<br>0080<br>0090<br>0020<br>0020<br>0000<br>0000<br>0000<br>000         | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 02         00         00           00         00         00         00           00         00         00         00           00         00         00         00           00         00         00         00           00         00         00         00           00         00         00         00           00         00         00         00           00         00         00         00           00         00         00         00           01         01         00         00           02         00         00         00           02         00         00         00           03         5f         74         63           04         5f         74         63           02         00         00         00  | 00 00 00<br>70 5f 43<br>00 02 00<br>73 5f 74<br>00 02 00<br>73 5f 74<br>44 00 00<br>63 6b 65<br>49 54 31<br>00 15 73<br>4e 5f 57<br>00 15 73<br>00 15 74<br>00 00 00<br>70 5f 4c<br>00 00 00 | $\begin{array}{cccccccccccccccccccccccccccccccccccc$  | if   | so<br>p_CLOSE_<br>socket<br>s_tcp_CL<br>s_tcp_CL<br>s_tcp_ES<br>D<br>sockets_tc<br>IT1<br>socket<br>N_WAIT2.<br>so<br>p_LAST_A<br>so<br>p_LISTEN<br>so  |           |                |    |
| 0020<br>0030<br>0040<br>0050<br>0060<br>0070<br>0080<br>0090<br>0040<br>0060<br>0060<br>0060<br>0060<br>0060<br>006 | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 02       00       00       00         3       5f       74       63         0       00       00       00       03         3       6b       65       74       70         4       00       00       00       00         5       55       74       63       15         5       57       41       15       73       6f         9       02       00       00       00       00       00         0       05       74       63       02       00 <td< td=""><td><math display="block">\begin{array}{cccccccccccccccccccccccccccccccccccc</math></td><td><math display="block">\begin{array}{cccccccccccccccccccccccccccccccccccc</math></td><td>if        </td><td>so<br/>p_CLOSE_<br/>CL<br/>CL<br/>CL<br/>CL<br/>CL<br/>DLS<br/>DSC<br/>Ckets_tc<br/>IT1so<br/>ckets_tc<br/>IT1so<br/>p_LAST_A<br/>so<br/>p_LISTEN<br/>so<br/>p_SYN_RE</td><td></td><td></td><td></td></td<>   | $\begin{array}{cccccccccccccccccccccccccccccccccccc$   | $\begin{array}{cccccccccccccccccccccccccccccccccccc$  | if   | so<br>p_CLOSE_<br>CL<br>CL<br>CL<br>CL<br>CL<br>DLS<br>DSC<br>Ckets_tc<br>IT1so<br>ckets_tc<br>IT1so<br>p_LAST_A<br>so<br>p_LISTEN<br>so<br>p_SYN_RE  |           |                |    |
| 0020<br>0030<br>0040<br>0050<br>0060<br>0080<br>0090<br>0080<br>0090<br>0000<br>0000<br>000                         | $\begin{array}{cccccccccccccccccccccccccccccccccccc$ | 00         01         15         73         67         55         57         41         0         00 </td <td><math display="block">\begin{array}{cccccccccccccccccccccccccccccccccccc</math></td> <td><math display="block">\begin{array}{cccccccccccccccccccccccccccccccccccc</math></td> <td>if        </td> <td>so<br/>p_CLOSE_<br/>so<br/>s_tcp_ES<br/>D<br/>ckets_tc<br/>IT1so<br/>p_LAST_A<br/>so<br/>p_LAST_A<br/>so<br/>p_LISTEN<br/>p_SYN_RE<br/>so<br/>p_SYN_SE</td> <td></td> <td></td> <td></td> | $\begin{array}{cccccccccccccccccccccccccccccccccccc$   | $\begin{array}{cccccccccccccccccccccccccccccccccccc$  | if   | so<br>p_CLOSE_<br>so<br>s_tcp_ES<br>D<br>ckets_tc<br>IT1so<br>p_LAST_A<br>so<br>p_LAST_A<br>so<br>p_LISTEN<br>p_SYN_RE<br>so<br>p_SYN_SE  |           |                |    |



### 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



| 2 | 3/ | 1 | 1/ | 2 | 0 | 1 | 6 |
|---|----|---|----|---|---|---|---|
|   |    |   |    |   |   |   |   |

| Discover - Kibana                      | × +   |  |
|--|---|--|
| 🗲 🛈 🔒   https://monit.c                | ern.ch/app/kibana#/discover?_g=(refree              | sh 🕑 🔍 Search 🔂 🖨 😎 🖡 🏫 🐗 🚳 🛩 🛞 🚍  |
| Google 🔋 Google Cale                   | endar 👖 Google Calendar 📄 CER                       | NY ■EDHY ■OpsY ■GGUSY ■ReposY ■ADY ■T-systemsY »                                       |
| 12.649                                 | Time  | _source  |
| 12.756                                 | <ul> <li>October 25th 2016, 09:36:27.019</li> </ul> | data.file metadata.adler32: 77b6f69e data.block size: 0 data.buf size                  |
| 16.7%                                  |   | : 0 data.channel type: urlcopy data.chk timeout: 0 data.dst srm v: 2.                  |
| 12.953                                 |   | 2.0 data.dst_country: Switzerland data.dst_experiment_site: CERN-EXTENS                |
| 12.954                                 |   | ION data.dst_federation: CH-CERN data.dst_hostname: xdpmhn01.tsy.cern.c                |
| 16.7%                                  |   | h data.dst se: srm://xdpmhn01.tsy.cern.ch data.dst site: CERN-PROD dat                 |
| 12.954                                 | link to /monit prod fts enr co                      | mplete v012-2016-10-25/enr complete/380a7c7e678eb6ec05104bd227ccc5fdf23df975           |
| Not Indexed                            | Table JSON  |  |
| t_type                                 | t_id  | Q Q Ⅲ 380a7c7e678eb6ec05104bd227ccc5fdf23df975   |
| # data.block_size                      | t_index   | © □ monit_prod_fts_enr_complete_v012-2016-10-25  |
| # data.buf_size                        | # score   |  |
| t data.channel_type # data.chk timeout | t_type  | © ○ □ enr complete   |
| t data.dst_country                     | <pre># data.block size</pre>                        | Q Q II 0   |
| t data.dst_experiment_s                | <pre># data.buf_size</pre>                          | Q Q II 0   |
| t data.dst_federation                  | t data.channel type                                 | Q Q II urlcopy   |
| t data.dst_hostname                    | # data.chk_timeout                                  |  |
| t data.dst se                          | t data.dst country                                  | Q Q 🔲 Switzerland  |
| t data.dst_site                        | <pre>t data.dst_experiment_site</pre>               | Q Q II CERN-EXTENSION  |
| tv                                     | t data.dst_federation                               | Q Q II CH-CERN   |
| # data.dst_tier                        | t data.dst_hostname                                 | Q Q Ⅲ xdpmhn01.tsy.cern.ch   |
| t_data.dst_url                         | t data.dst_se                                       | Q Q Ⅲ srm://xdpmhn01.tsy.cern.ch   |
| t data.endpnt                          | t data.dst site                                     | Q Q II CERN-PROD   |
| # data.f_size                          | t data.dst srm v                                    | Q Q II 2.2.0   |
| t_data.file_id                         | # data.dst tier                                     | Q Q II 3   |
| t data.file_metadata.acti              | t data.dst url                                      | Q Q □ srm://xdpmhn01.tsy.cern.ch:8446/srm/managerv2?SFN=/dpm/tsy.cern.                 |
| t data.file_metadata.adl               |   | ch/home/atlas/atlasdatadisk/rucio/mc15_13TeV/9e/cc/HITS.09473395<br>003588.pool.root.1 |
| t data.file_metadata.des               | t data.endpnt                                       | Q Q 🔲 fts3.cern.ch   |
| t data.file_metadata.dst               | # data.f size                                       | <b>Q</b> , <b>Q</b> , Ⅲ 919.6MB  |
| t data.file_metadata.dst               | t data.file id                                      | Q Q II 1065772449  |
| # data.file_metadata.file              | <pre>t data.file_metadata.activity</pre>            | Q Q II Recovery  |
| ? data.file_metadata.md5               | <pre>t data.file_metadata.adler32</pre>             | Q Q III 77b6f69e   |
| t data.file_metadata.na                | <pre>t data.file_metadata.dest_rse_id</pre>         | Q Q II 316488643efc413384c22df1070ca3f3  |
| t data.file_metadata.req               | t data.file metadata.dst rse                        |  |
| t data.file_metadata.req               |   | · · -  |

# 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

