• Organisation
• Hardware
• Networking
• Storage
• Facilities
• Miscellaneous

Thanks to colleagues for contributions
New UK Research Organisation

- UK Research and Innovation, launched 1st April 2018, is the new funding organisation for research and innovation in the UK
- It brings together the seven UK research councils, Innovate UK and a new organisation, Research England, working closely with its partner organisations in the devolved administrations
  - Includes STFC, which runs RAL
- UK Research and Innovation intends to be an outstanding organisation that ensures the UK maintains its world-leading position in research and innovation
- Rising funding profile through 2020/21
- Not expecting any changes at facilities level

15/05/2018
Tier1 Hardware

- **CPU:** ~236k HS06 (~22k cores)
  - FY17/18: procurement ~91k HS06 (Dell, XMA)
- **Castor:** ~16.5 PB useable
  - Dropping as older hardware is retired
- **Ceph:** ~20PB raw / ~13PB configured
  - FY 17/18:
    - 74 x Supermicro 24 bay units -> 19.5PB raw / 14.2PB configured
    - Acceptance testing
- **Tape:** 10k slot SL8500 (one of two in system)
  - 80PB capacity (T10KD), ~30PB physics data
Networking

• Tier1 WAN/LAN
  - No significant changes

• IPv6
  - IPv6 available on Tier1 network
  - All required services for WLCG now IPv6 dual stack
    • Not Castor

• RAL Site
  - Firewalls replaced, can now do IPv6 in ASIC, better performance
  - Recently, issues revealed in internal switching and border routing configuration, particularly with IPv6
    • Working to understand and provide long term fixes
  - 100Gbs connectivity to site early summer
• CASTOR: disk-only service run-down continues
  - ~10PB data remains
  - Planning to rationalise 4 instances to one (for tape)

• ECHO: disk-only service - expansion continues
  - Possibly the largest CEPH Cluster using erasure coding
  - 30 more storage nodes added (24 x 8TB), total useable space now 13PB
  - Data held: Atlas @ 4.2PB, CMS @ 1.6PB
  - Working on improving disk replacement workflow for large clusters
• **UK’s leading environmental science supercomputer**
  - Supports UK and European climate and earth-system science communities
  - Access to very large environmental data sets
  - Power to process data very rapidly

• **2017: ~20PB useable, ~5000 cores**
  - Mostly Panasas HPC storage
    - world’s largest ‘realm’, largest single site installation
  - CPU split ~50/50:
    - batch computing and cloud (Openstack)
    - Virtualised environments (VMware)
• Added 30PB (useable) software defined scale out parallel file system storage from Quobyte
  - Dell and Supermicro Hardware
  - ~5PB targeted interchangeably for File or Object (S3)
  - Expected to deliver 200G-400GBytes/sec file throughput
    • all from HDDs
• Additional 5PB of a more traditional dedicated object store (S3 but with a unique direct NFS interface to the object data), from Caringo on Supermicro hardware
• JASMIN4 CLOS network spine
  - 8 x 32 x 100Gb port Mellanox switches w/Cumulus/BGP
• Additional ~5000 cores (Dell) for Openstack (RHEL/KVM)
• ~500TB useable high availability PURE all flash “FlashBlade” NAS
  - home and scratch small file/compilation/metadata heavy work loads
New Data Centre Network

• Added a third layer to the routed CLOS
  - Connecting as many JASMIN CLOS networks together at near line rate
  - All existing SCD STFC services connected at many 100’s of Gbit/sec
    • 16 x 32 x 100Gb port Mellanox switches @ Data Centre layer
    • Will include Tier1
  - Total (so far) ~15Tbit/s spread over:
    • 1,600x 10Gb, 80x 50Gb, 100x 40Gb, ~500x 25Gb, 16x 100Gb server side
      connections all linked with 290x 100Gb links

• New high bandwidth pipe between new and old data centres
  - At least 3 x 144 fibre (72 link) 100Gb/s
  - Shared by JASMIN, Tier1, Site network
• **OpenNebular:**
  - reducing from ~700 to ~300 cores

• **OpenStack:**
  - Currently ~500 cores, growing to ~3500 end may and ~5000 by end 2018.

• **Added**
  - **Hypervisors:** 108 Dell 6420 sleds (27 x 4-up 2U 6400 chassis)
    - 16 physical cores each, 6GB RAM/core, 25G NIC
      - Testing complete, about enter production
  - **Cloud Storage:** 12 Dell R730xd (12 bay 2U)
    - 12 x 4TB each, total 576TB, 25G NICs
    - Replicated CEPH (3x), VM image storage
      - Testing complete, about enter production
  - **Data Storage:** 21 x (Dell R630 + 2 x MD1400) sets
    - Added to ECHO for UKT0 project
    - 4PB raw /~2.9TB configured
      - Just finished testing
  - **Network:**
    - 7 x Mellanox SN2100 16x100Gb port - Cumulus

• **Typical load**
  - ~130 VMs, currently quotas due to resource limits
    - Expect number increase sharply when new resources in production

15/05/2018 HEPiX Spring 2018 - RAL Site Report
11000 HPC cores for STFC staff, collaborators and users of STFC Facilities

- Added 3552 cores this year - Dual Intel Xeon 6126s, 192GB RAM, 10GbE, 4X EDR
- Added 5 new shelves of Panasas ActiveStor 20 parallel storage connected via ethernet
- Migrating applications serving from SPOF nfs server to Quobyte 4 node cluster (possibly integrated into Jasmin)
- Network Overhaul - moving network core from pair of Dell S4810s to pair of Mellanox 2100s and (hopefully) pair of Mellanox 2700s to enable link into SCD Bridge network

- Investigating options to change batch system from LSF
- Interest in deploying storage attached to IB fabric, investigating options for linking multiple IB fabrics together
<table>
<thead>
<tr>
<th></th>
<th>E5-2660</th>
<th>E5-2650v2</th>
<th>E5-2650v2</th>
<th>E5-2640v3</th>
<th>E5-2640v3</th>
<th>E5-2630v3</th>
<th>E5-2630v4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microcode original</td>
<td>lcg1555</td>
<td>lcg1611</td>
<td>lcg1675</td>
<td>lcg1803</td>
<td>lcg1863</td>
<td>lcg1999</td>
<td>lcg2151</td>
</tr>
<tr>
<td></td>
<td>0x710</td>
<td>0x428</td>
<td>0x428</td>
<td>0x38</td>
<td>0x38</td>
<td>0x38</td>
<td>0xb000025</td>
</tr>
<tr>
<td>Kernel (3.10.0-)</td>
<td>693.1.1</td>
<td>693.1.1</td>
<td>693.1.1</td>
<td>693.1.1</td>
<td>693.2.2</td>
<td>693.2.2</td>
<td>693.11.6</td>
</tr>
<tr>
<td>Geo Mean</td>
<td>332.84</td>
<td>369.02</td>
<td>368.86</td>
<td>378.83</td>
<td>382.78</td>
<td>349.89</td>
<td>412.57</td>
</tr>
<tr>
<td>Microcode 20180312</td>
<td>0x715</td>
<td>0x42c</td>
<td>0x42c</td>
<td>0x3c</td>
<td>0x3c</td>
<td>0x3c</td>
<td>0xb000025</td>
</tr>
<tr>
<td>Kernel (3.10.0-)</td>
<td>693.11.6</td>
<td>693.11.6</td>
<td>693.11.6</td>
<td>693.11.6</td>
<td>693.11.6</td>
<td>693.11.6</td>
<td>693.11.6</td>
</tr>
<tr>
<td>Geo Mean</td>
<td>331.11</td>
<td>362.56</td>
<td>363.97</td>
<td>378.31</td>
<td>382.04</td>
<td>350.24</td>
<td>414.65</td>
</tr>
<tr>
<td>Difference (means)</td>
<td>-1.73</td>
<td>-6.46</td>
<td>-4.90</td>
<td>-0.51</td>
<td>-0.74</td>
<td>0.35</td>
<td>2.09</td>
</tr>
<tr>
<td>Difference New/Old</td>
<td>99.48%</td>
<td>98.25%</td>
<td>98.67%</td>
<td>99.86%</td>
<td>99.81%</td>
<td>100.10%</td>
<td>100.51%</td>
</tr>
<tr>
<td></td>
<td>0.52%</td>
<td>1.75%</td>
<td>1.33%</td>
<td>0.14%</td>
<td>0.19%</td>
<td>-0.10%</td>
<td>-0.51%</td>
</tr>
</tbody>
</table>

Normal expected variation
## Storage

### Read heavy

<table>
<thead>
<tr>
<th></th>
<th>fdsstoraged19</th>
<th>fdsstoraged20</th>
<th>Geomean</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel</td>
<td>2.6.32-696.18.7.el6</td>
<td>2.6.32-696.13.2.el6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run1</td>
<td>91239.64</td>
<td>60612.30</td>
<td>89046.92</td>
<td>-30660.73</td>
</tr>
<tr>
<td>Run2</td>
<td>87037.08</td>
<td>55897.16</td>
<td>79133.00</td>
<td>152.51%</td>
</tr>
<tr>
<td>Run3</td>
<td>89080.56</td>
<td>59659.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run4</td>
<td>90038.61</td>
<td>56287.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run5</td>
<td>87901.20</td>
<td>59635.84</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Write heavy

<table>
<thead>
<tr>
<th></th>
<th>fdsstoraged19</th>
<th>fdsstoraged20</th>
<th>Geomean</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel</td>
<td>2.6.32-696.18.7.el6</td>
<td>2.6.32-696.13.2.el6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run1</td>
<td>21959.44</td>
<td>18486.96</td>
<td>22626.91</td>
<td>-2318.44</td>
</tr>
<tr>
<td>Run2</td>
<td>23705.30</td>
<td>18870.39</td>
<td></td>
<td>111.42%</td>
</tr>
<tr>
<td>Run3</td>
<td>22729.15</td>
<td>19431.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run4</td>
<td>21425.90</td>
<td>21572.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run5</td>
<td>23395.78</td>
<td>23621.89</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Opportunistic test - patched was faster!**
• Area-wide power outage @ lunchtime
  - Controlled panic

• Generator UPS didn’t start
  - Power comes back ~ 8 minutes - phew!
  - All of Ceph on UPS so it was unaffected

• Why didn’t the generator start?
  - BMS noted power down, asserted generator start signal
  - In-cabin generator controller received signal, but...
  - Faulty EPO button in-cabin asserting ‘off’ so it didn’t start
  - Red light on control box

• Lesson:
  - Expose the generator control system status where it can bee seen
• Tier1 move from Hyper-V to VMware for core infrastructure
  - Cluster testing complete, migration starting

• Oracle Databases
  - Plan to migrate to RH7 before 4 April 2018

• Patching for Spectre/Meltdown
  - And other fubars in the kernel...

• Mobile Device Management to be rolled out
  - WiFi infrastructure at RAL will refuse access to core services such as email from mobile devices (phones, tablets, laptops) not running vendor-supported OS versions.
    • Registration (enrolment) of devices will be required
    • Rollout starts 28\textsuperscript{th} May
We are interested solutions for system inventory / asset management for Linux (desktop) estate, to provide ‘numbers’ for audits, FOI requests et al.