RAL Tier 1 Site Report

James Adams
Scientific Computing Department

HEPSysMan @ RAL
2017-06-15

ROW 6
TIER 1
ROW 7

Tier 1 “Capacity” Hardware

- CPU: ~240k HS06 (~24k cores)
  - FY16/17: Additional ~19.6kHS06, 1920 cores (E5-2630-v4)
    - Dell CloudEdge
- Storage:
  - ~16.5 PB usable in Castor
  - ~13.3PB raw for Ceph
  - FY 16/17: Additional 6720TB raw (~4.9TB configured) for Ceph
    - 35 x (Dell R630 + 2 x MD1400) units, SAS interconnect
- Tape: 10k slot SL8500 (one of two in system)
  - 50PB (T10KD)
  - Migrations to D-only completed
    - LHCb: 600 tapes, ~3PB: no errors at all

Networking

- Tier1 WAN
  - OPN link to increased to 30Gb/s
    - 2 x 10Gb/s over same path to CERN
    - 1 x 10Gb/s over alternative path
    - Operated as parallel routes (BGP)
- LAN
  - New! Mellanox SN2100 and SN2700 switches
    - Switches used for the SCD Private Cloud running Cumulus Linux

IPv6

IPv6

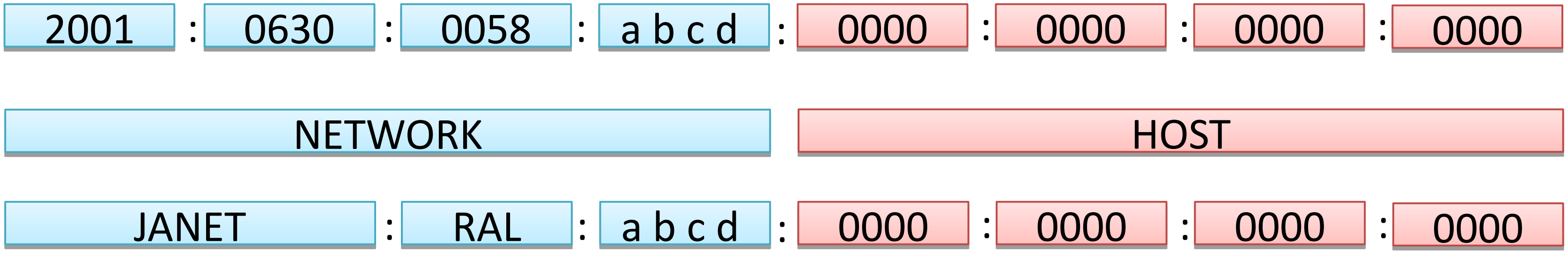
- IPv6 now available on Tier1 network
- Global addressing scheme agreed
- See my HEPiX talk:
  - https://indico.cern.ch/event/595396/contributions/2558578/

Science & Technology
Facilities Council

STFC Addressing Scheme

Each project allocated one or more IPv6 /64

- 16 bits available to describe subnet



2001 : 0630 : 0058 : a b c d : 0000 : 0000 : 0000 : 0000

NETWORK HOST

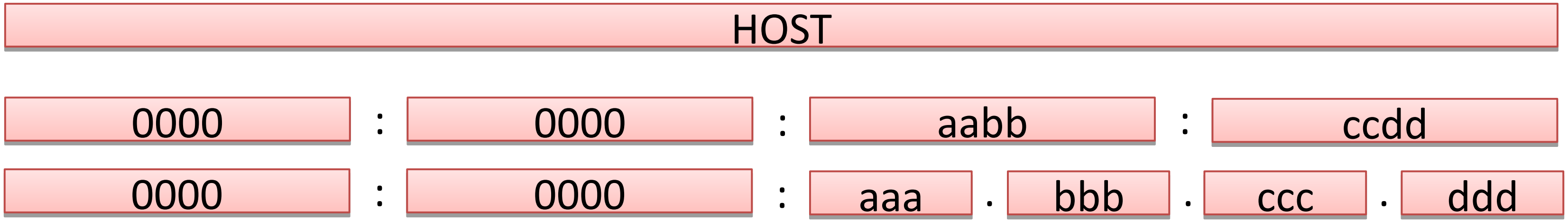
JANET : RAL : a b c d : 0000 : 0000 : 0000 : 0000

a = STFC Address plan version (0-15)
b = Network Type
c = Network Subtype
d = Assigned by subnet owner (Tier 1 addressing scheme version)

Science & Technology
Facilities Council

Tier 1 Addressing Scheme (v0)

- Assumption: All hosts will be dual-stack
- Map all existing IPv4 address (RFC2374 style)
  - Allocate addresses automatically with Quattor
- DNS entries just a sed script away…



HOST

0000 : 0000 : aabb : ccdd

0000 : 0000 : aaa . bbb . ccc . ddd

In hex notation:

`::82F6:B43C`

Or mixed notation:

`::130.246.180.60`

Services

- Batch farm
  - ~24000 job slots
  - Completed migration to SL7
  - HTCondor Docker universe running jobs in SL6 containers
  - Experiment: On node xrootd caches and ECHO gateways
- Container Orchestration
  - Investigating Kubernetes as a means of providing portability between on-premises resources and multiple public clouds
  - See Andrew's HEPiX talk:
    - https://indico.cern.ch/event/595396/contributions/2556631

More Services

- Load balancers
  - Pair of VMs running HAProxy and Keepalived as a highly-available load balancer (see previous HEPiX reports)
  - Used in front of FTS3 for over a year
  - Top BDII, Site BDII, Dynafed, Argus
- Monitoring
  - Ganglia still exists, usage is slowly fading
  - Telegraf → InfluxDB → Grafana
    - Grid services, batch system, Ceph, Windows HyperV
- Turned off old (c.2008) ElasticSearch cluster.

Even More Services

- CVMFS
  - New HW for CVMFS Stratum-0
  - See Catalin’s HEPiX talk
    - https://indico.cern.ch/event/595396/contributions/2532590/
- Planning move from Hyper-V to VMware
  - Consolidation of resources across department
  - A couple of instances of hypervisors crashing in Hyper-V 2012.....
- Still dealing with retirement of (S|RHE)L5 systems
  - Oracle DBs still on RHEL5, purchased extended lifetime support
- Windows administration privileges removed from all accounts
  - Consequence of Cyber Essentials activity at BEIS
  - Separate logon accounts with admin privileges

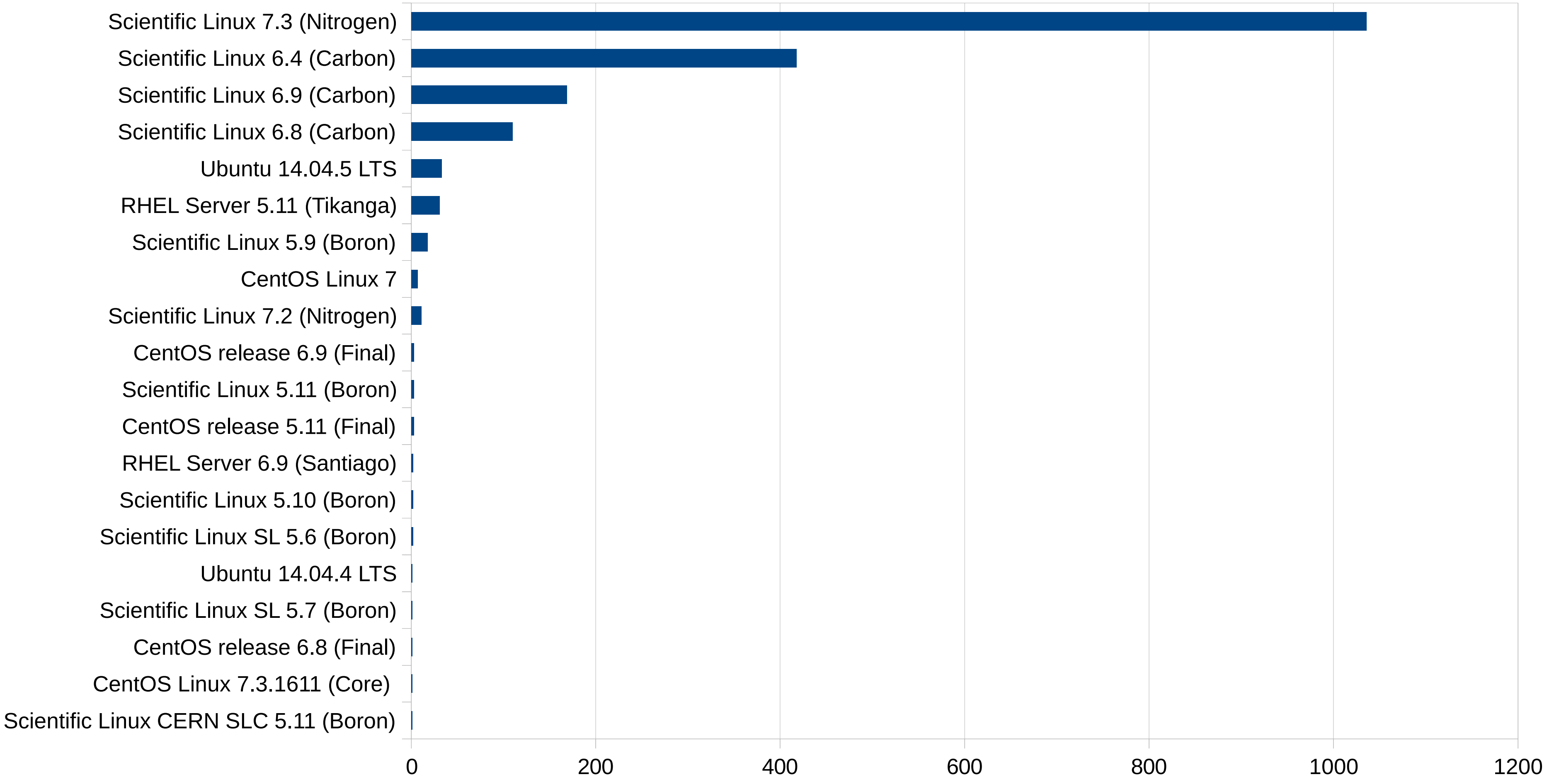
Configuration Management

- Finally got rid of Puppet!

  "Decommissioned the same way the Titanic was" ~ RA

- Declared SCDB end-of-life
  - Everything except CASTOR moving to Aquilon quickly
  - CASTOR requested SCDB remain until end-of 2018.
- Started to introduce support for RHEL7 and Debian-based (Mint, Ubuntu, Cumulus) distros
- Very heavily invested in Quattor
  - Infrastructure more and more shared across STFC
  - Next workshop at RAL in October!

OS long tail



Science & Technology Facilities Council

Storage - CASTOR

- Storing ~130TB/day of data for WLCG
- Upgraded to 2.1.15-20 in January
- SRM updated to 2.1.16-10, then rolled back for LHCb following performance problems
- Upgraded to 2.1.16-13 in May
- Fixed `transfermanager` memory leak
  - Being triggered by our “bananas” monitoring

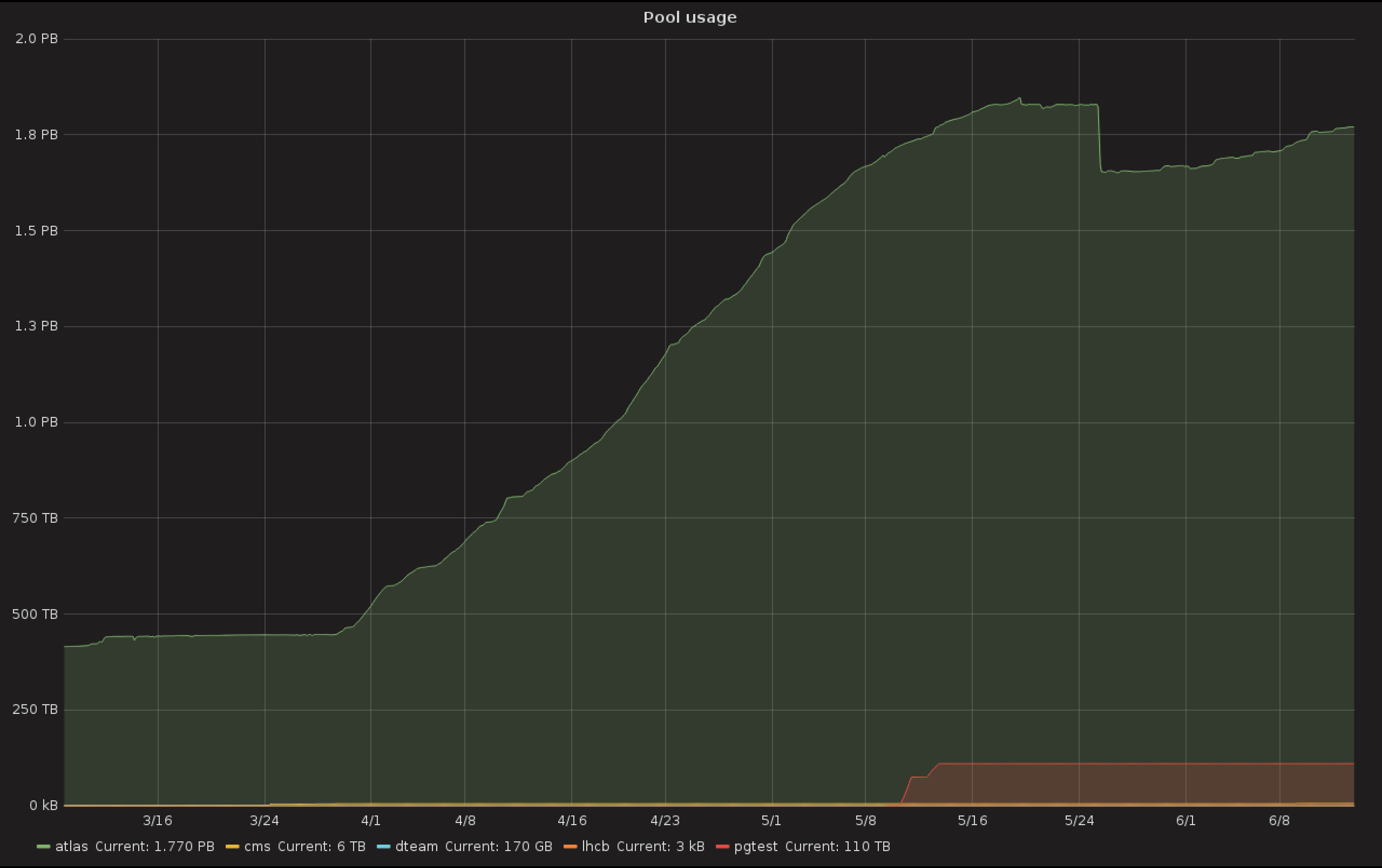
- Some 2014 generation disk servers (used for ECHO testing) moved to CASTOR

Science & Technology
Facilities Council

Storage - ECHO

- Underlying Ceph cluster upgraded to Kraken
- Accepting production data from LHC VOs
  - GridFTP and XrootD supported as production protocols
  - VO pools can be accessed via either protocol.
- Storing ~2PB of data for ATLAS
  - Input via GridFTP
  - Batch farm talking to ECHO via XrootD
- Will provide 7.1PB of wLCG pledge this year
- See Tom's HEPiX talk:
  - https://indico.cern.ch/event/595396/contributions/2553417/

Science & Technology
Facilities Council


Pool usage
2.0 PB
1.8 PB
1.5 PB
1.3 PB
1.0 PB
750 TB
500 TB
250 TB
0 kB
3/16
3/24
4/1
4/8
4/16
4/23
5/1
5/8
5/16
5/24
6/1
6/8
atlas Current: 1.770 PB
cms Current: 6 TB
dteam Current: 170 GB
lhcb Current: 3 kB
pgtest Current: 110 TB

Tape Now

- Two StorageTek (~~Sun~~ Oracle) SL8500 Libraries
- Standardised on T10K tape media
  - Tier 1
    - 5167 T10KD (8.5TB)
  - STFC Facilities
    - 2606 T10KD (8.5TB)
    - 3525 T10KC (5.0TB)
    - 2435 T10KB (1.0TB)

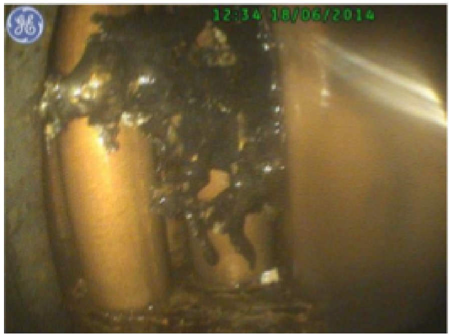
Tape Future

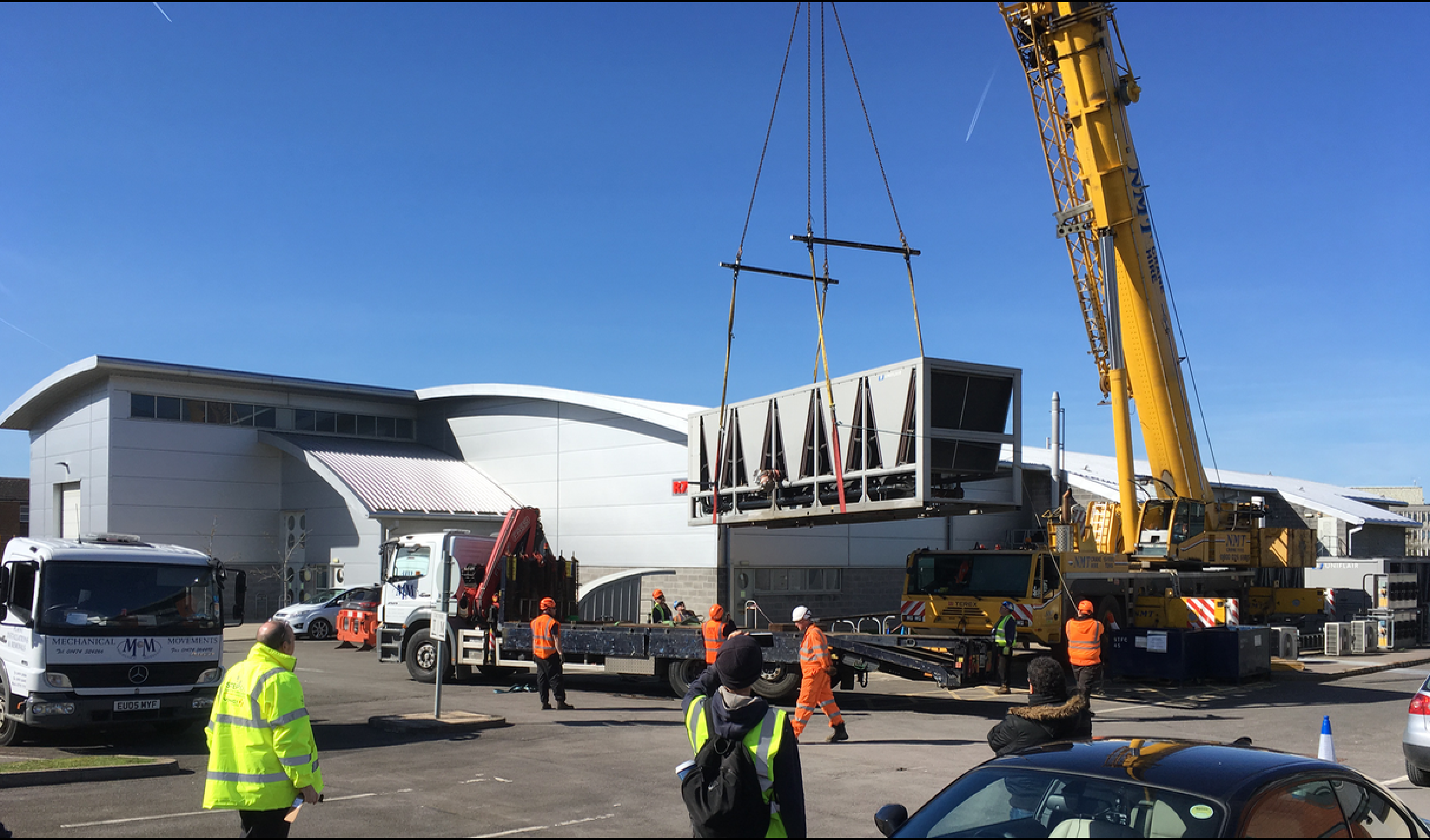
- Plan ***was*** to move to T10KE with T2 and eventually T3 media
- However… Long term support for both drives and Libraries is now unclear
- No official announcements
- Can operate on existing tech until 2020 if supported
- Waiting to see what Oracle does next

Science & Technology
Facilities Council

Infrastructure - Chillers

- Moved into building in 2009
  - 2 x 750kW chillers with free cooling
  - Later added another 2 x 750kW without free cooling
  - Original pair end-of-life
    - One lost half capacity due to component failure
- Replaced under Laboratory spend-to-save initiative
  - 2 x 1MW chillers with free cooling
    - More efficient
    - Higher capacity
  - Commissioned March-April
  - Reduced PUE from ~1.64 to ~1.35
- New sequencer
  - More intelligent control system for all four chillers
  - Reduced power consumption by ~60kW
  - Reduced PUE to ~1.31

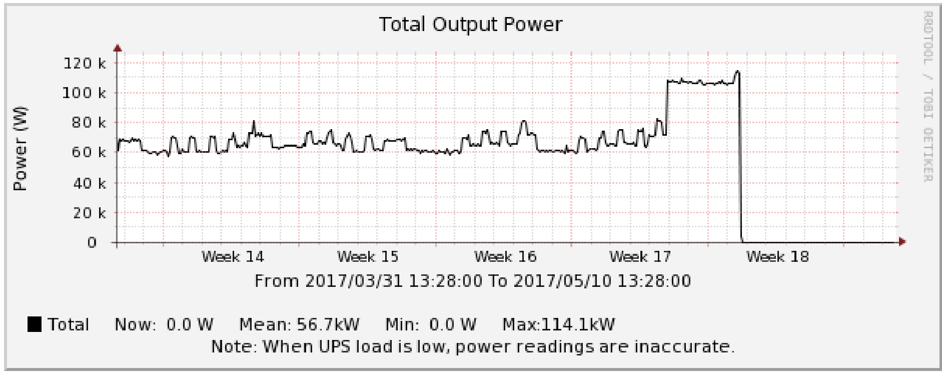


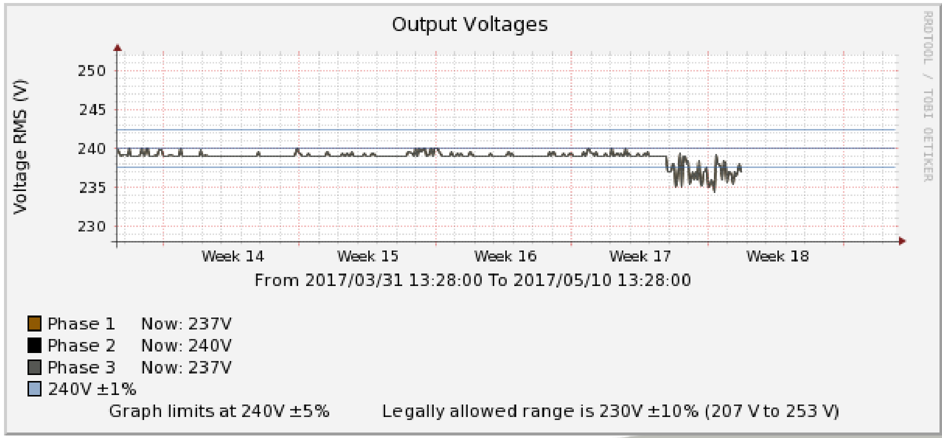

MECHANICAL
M&M
MOVEMENTS
EU05 WYF
NMT
CRANE HIRE
TEREX
UNIFLAIR
R7

NMT

R89 UPS

- 28th April
  - UPS detects serious fault, switches to bypass
- 2nd May
  - Engineers arrive, UPS shutdown
  - Options discussed...
- 11th May
  - Replacement of UPS approved
- 12th-14th May (weekend)
  - Faulty UPS removed
  - Replacement installed and tested
- 15th May
  - Replacement UPS commissioned
- 16th May
  - All UPS feeds restored

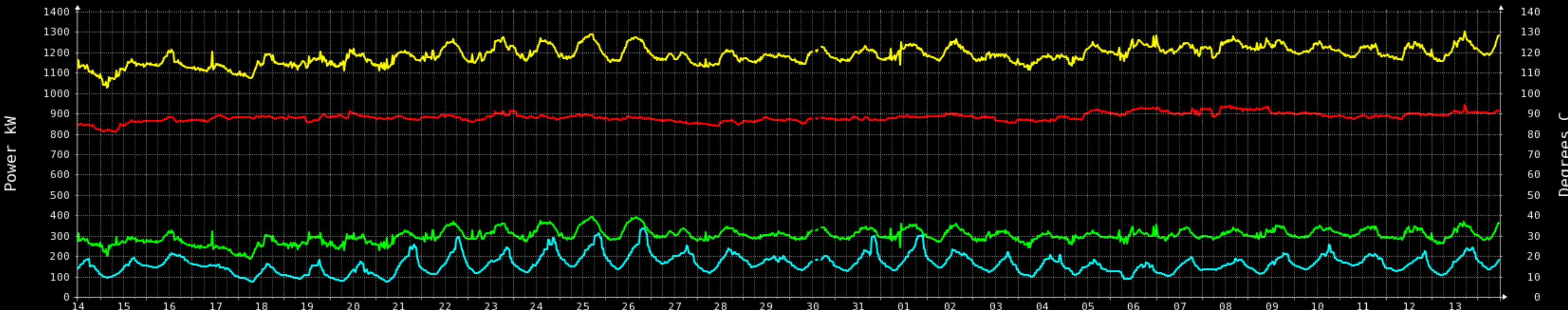




Science & Technology Facilities Council

Power Consumption

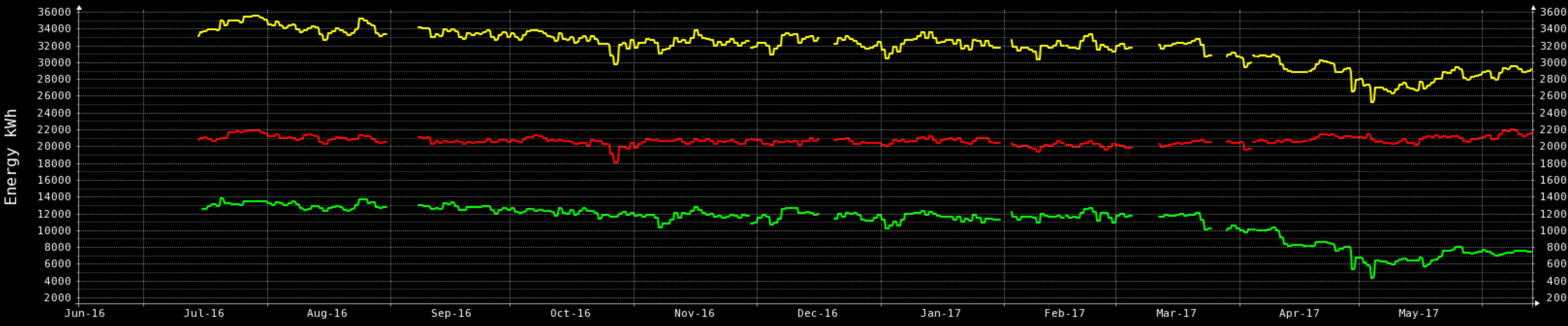
14 May 2017 to 14 June 2017



```
🟨 TOTAL Power         cur: 1286 kW      avg: 1191 kW      max: 1340 kW
🟥 IT Power            cur:  915 kW      avg:  886 kW      max:  948 kW
🟩 HVAC Power          cur:  371 kW      avg:  304 kW      max:  409 kW

🟦 Outside Temperature cur:  18.8 °C      avg:  16.7 °C      max:  34.1 °C
```

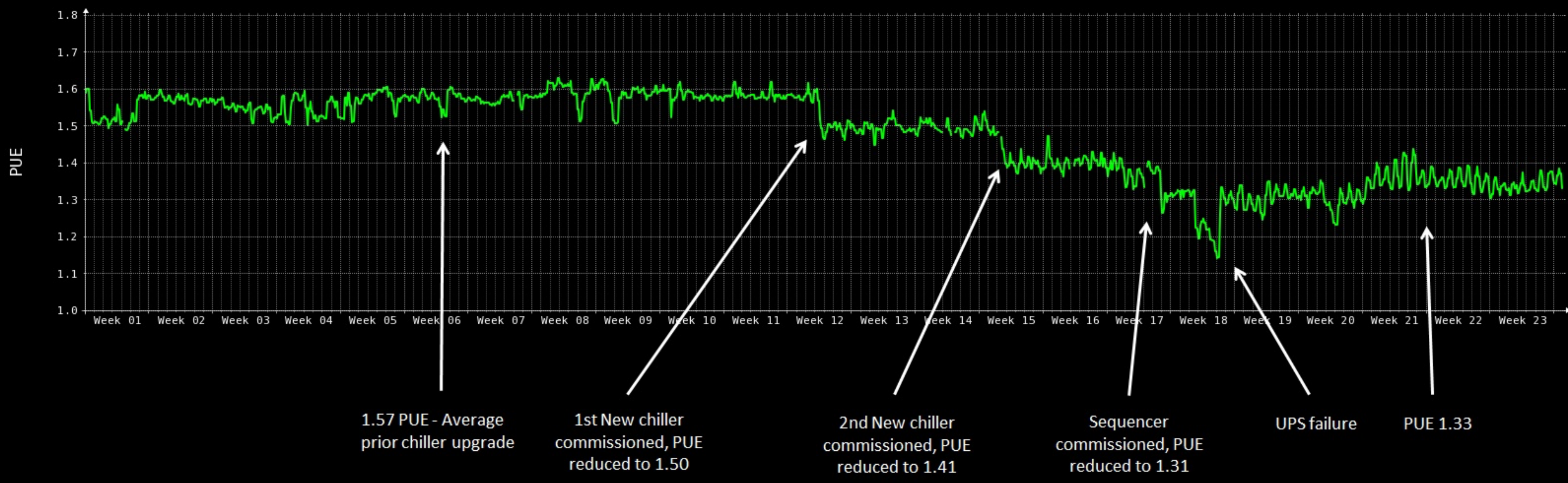
Daily Energy Consumption - 14 June 2016 to 14 June 2017



```
🟨 TOTAL kWh    cur:   29294 kWh     avg:   31908 kWh     max:   35619 kWh

🟥 IT kWh       cur:   21761 kWh     avg:   20784 kWh     max:   22178 kWh
🟩 HVAC kWh     cur:    7533 kWh     avg:   11119 kWh     max:   13937 kWh
```

PUE - January to June



1.57 PUE - Average prior chiller upgrade

1st New chiller commissioned, PUE reduced to 1.50

2nd New chiller commissioned, PUE reduced to 1.41

Sequencer commissioned, PUE reduced to 1.31

UPS failure

PUE 1.33

- New UPS saving ~20kWh
  - Doesn’t change the PUE change by much
  - But still £12k+ per year saving
- Summer has only increased PUE from 1.31 to 1.33
  - Even without free cooling the new chillers are proving to be very efficient
- Winter & fine tuning should see PUE drop below 1.3

Questions?