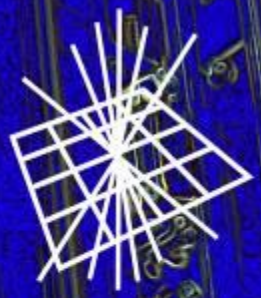


GridPP

UK Computing for Particle Physics

RAL Site Report

HEPiX Autumn 2021

25-28 October 2021

Martin Bly,

STFC UK Research and Innovation

Overview

- DC
- Networks
- Procurements
- Ceph
- Tape
- Cloud

Thanks to colleagues for contributions

DC Operations

- UPS
  - Additional capacity on the R89 diesel-backed UPS, allows more circuit capacity
    - Additional distribution board(s) to be installed
  - UPS in older R26 DC commissioned, not diesel-backed

- Cooling capacity
  - Remaining two older chillers replaced with new units, now fully commissioned

- Project underway to study options for future additional DC provision, examining various scenarios ahead of submission of a business plan to funding bodies

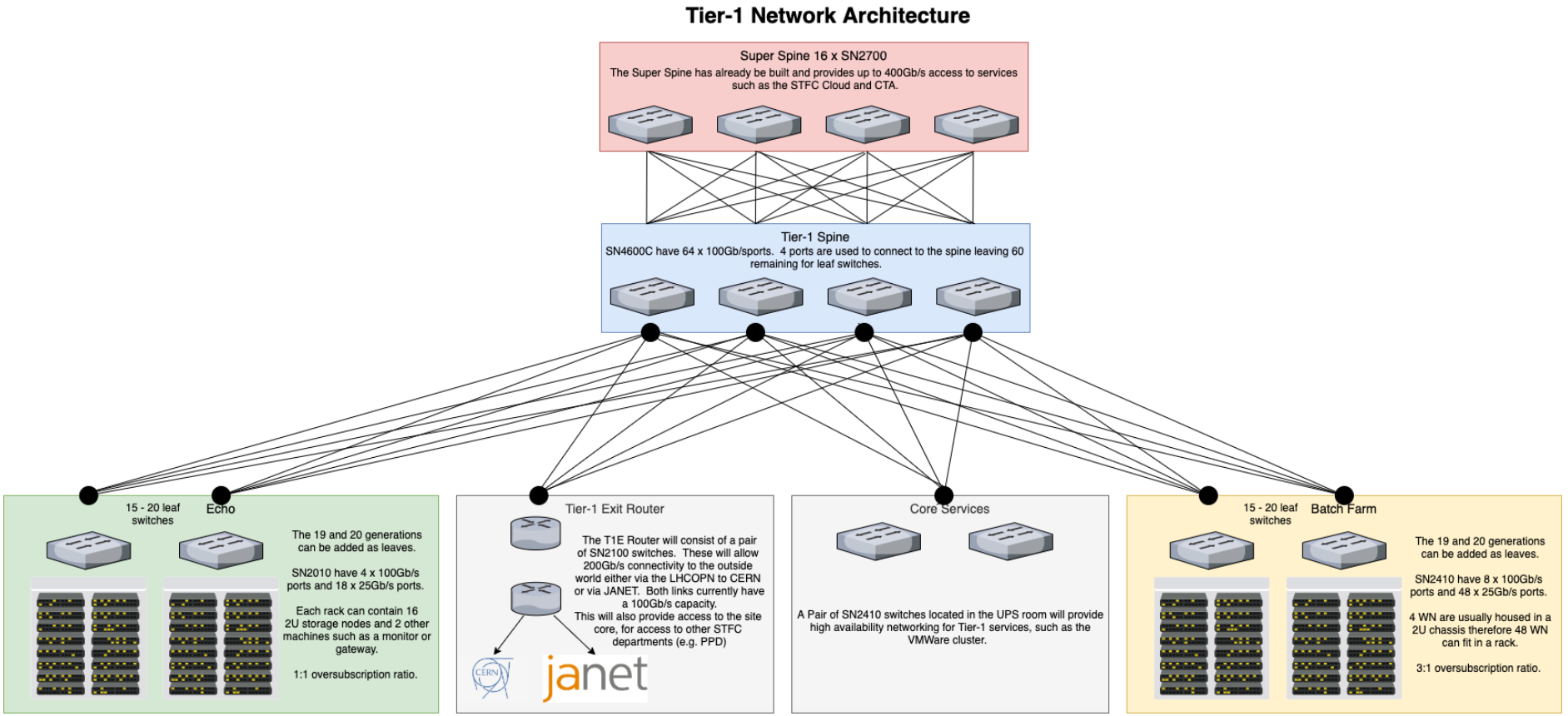
Networking - RAL Site

- Firewall
  - New Fortinet 4201F firewalls commissioned
    - Capable of 800Gbps total (8 x 100Gbps)
- Site WAN connection
  - Connection to JANET network now redundant 200Gbps
- Upgrade of site switch core
  - New switch core Summer 2021 @ 100Gbps per circuit
    - Also @40Gbps and 10GBps for legacy links
    - Increased intra-site connection bandwidth
    - Extreme: x870, x695 in two stacks of 4 x x670 + 2 x x656
- New point-to-point routing regime to be implemented in the site core shortly
- RAL now uses IP telephony via Zoom
  - Old numbers transferred
- Following recent public incidents, UKRI and STFC have been addressing security through auditing, with other projects in process

Networking - Tier1

- LHCOPN RAL to CERN on 100Gbps circuit, no ‘standby’
  - Failover is via public Internet
- Plan to replace Tier1 network with Leaf/Spine topology
  - Progressing, concentrating on newest generation of storage and compute, and core/WAN connectivity
    - Mellanox/Cumulus, routed network
    - Legacy and New network connected via SCD SuperSpine
    - 4 spines, 2 exit routers and various leaf ToR switches connected
    - Connections to WAN and Site core in place, operational soon
      - Will provide single (resilient) exit point to WAN and core for both legacy and new network, old routers will be decommissioned
      - Provides 100Gbps bandwidth for traffic to WAN from Legacy network
  - Will be fully operational for Run 3

Networking - Tier1



Procurements

- Compute for FY21/22, joint procurement for Tier1, Cloud and JASMIN: 92 for JASMIN, 48 for Tier1, 30 for Cloud
  - 1U SuperMicro servers
    - 2 x AMD EPIC CPUs: Tier1 and cloud: 7763, JASMIN: 7643
  - 1TB RAM (Tier1), 2TB RAM (cloud), 512GB (JASMIN)
  - 7.68 TB Enterprise SSD (Tier1, Cloud), 480GB Enterprise SSD (JASMIN)
  - Mellanox ConnectX-5 10/25GB DP NIC
- GPUs (cloud)
  - 15 x 4U SM servers each with 4 x Nvidia Tesla 80GB A100 GPUs
  - 15 x 1U SM servers each with 4 x Nvidia Quadro RTX4000 GPUs
- Switches
  - Mellanox SN2410, SN2100
- Storage
  - Tier1 ECHO: 80 x Dell R740XD2 w/24 x 18TB SAS HDDs
  - Cloud: 15 x Dell R7525 w/24 x 3.84TB SATA RI SSDs
  - CTA EOS: 3 x Dell R7525 w/24 x 3.84TB SATA RI SSDs

Echo: new Failure Domains

- Echo is a large Ceph cluster, 229 nodes, erasure coding
  - 51PB raw capacity, 40PB raw used, 29PB data stored
  - 8+3 erasure coding with the failure domains set to node-level
    - each of the 11 stripes is guaranteed to be on a different server
  - Plan to migrate to rack-level failure domains
    - each stripe is guaranteed to be on a different rack of servers
- Advantages:
  - Resilience to rack-level failures
  - Matches latest storage 'Unit' with 16 servers and a ToR switch
  - When hosts need new kernels, can reboot an entire rack concurrently rather than going through the whole cluster 2 hosts at a time ☺

Tape

- Data on tape
  - Tier1: 75PB, all on Tfinity
  - Facilities: 85PB, mostly on Tfinity
  - Migration of science data in Castor almost complete
    - ~400 DLS tapes to migrate
- Castor
  - Stable operation, tape only
  - Final upgrade to v2.1.19 complete
- CTA
  - Testing new tape system for operation soon
- DMF
  - Direct access to Tfinity libraries, pending upgrade to DMS software
- Exploring future solutions for system backup technology

SCD Cloud

- ~60k vCPU cores, ~500 GPUs, 260TB RAM, 768TB raw/256TB useable NVMe Ceph, 1.4PB local SSDs connected to HVs
- VM flavours available up to 124 VCPUs with 1.9TB RAM and 3.6TB local SSD

Other

- Talks – David Crooks
  - Resource Trust Evolution: Host Certificates in a Dynamic Landscape
    - https://indico.cern.ch/event/1078853/contributions/4580733/
  - Threat Intelligence and Security Operations Centres: Collaborative Security
    - https://indico.cern.ch/event/1078853/contributions/4580732/
  - Both in the late session on Thursday

Questions?