Science and Technology Facilities Council

Tier-1 Network Upgrade

Alastair Dewhurst

Agenda

1 Current Setup

2 Requirements

3 Design

4 Migration

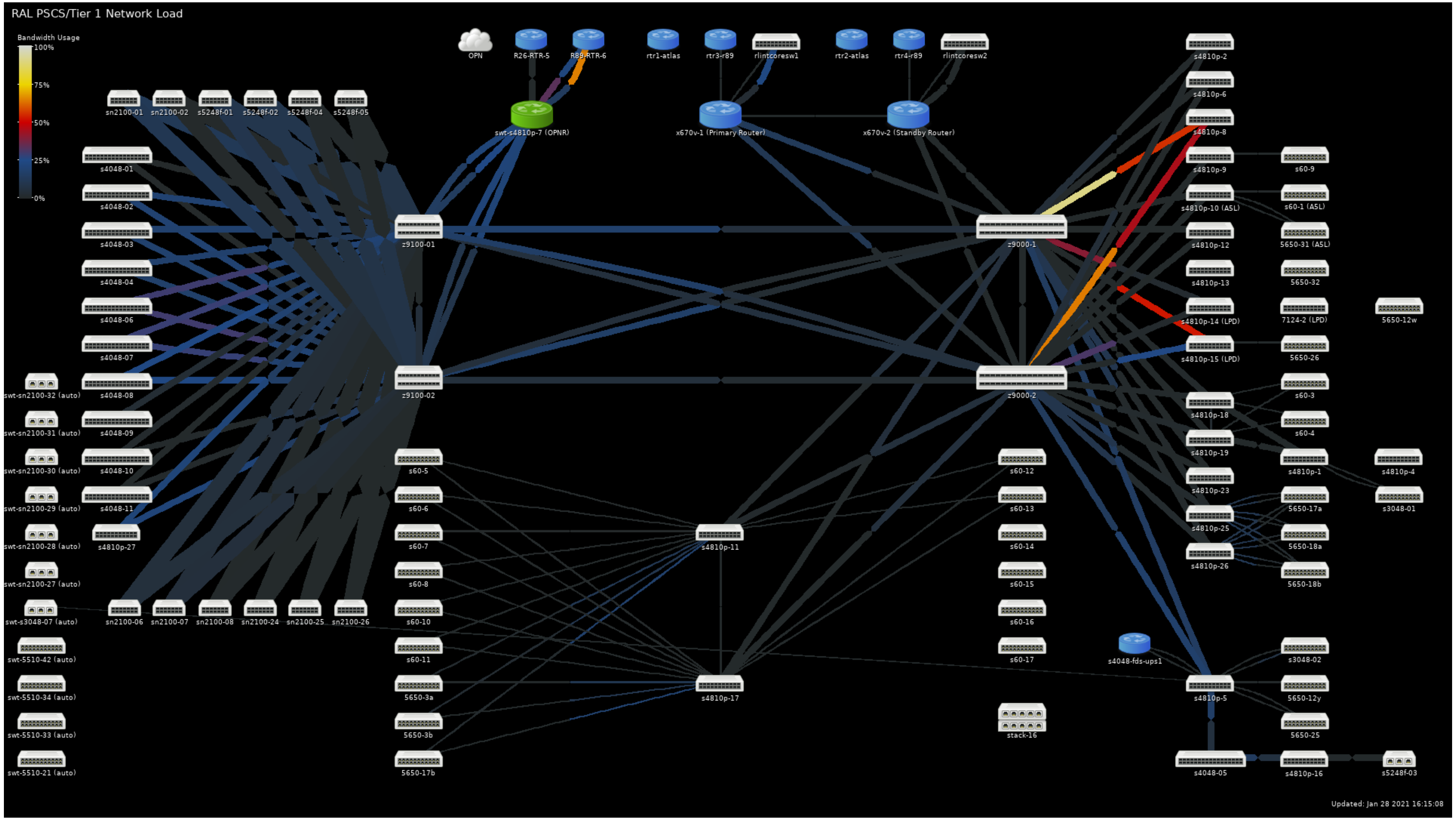
Science and
Technology
Facilities Council

Current Setup

Tier-1 Network

- The RAL Tier 1 currently connects to:
  - The world via Janet via the RAL Campus Network.
  - Tier 0s and other Tier 1s via the LHCOPN via a private router (OPNR).
  - Tier 2s via Janet via a private router (OPNR).
- Currently the Tier 1 provides compute, storage and services over both IPv4 and IPv6 on a single L2 segment.
  - 3× IPv4 subnets (one for LHCOPN)
  - 2× IPv6 subnets (one for LHCOPN)
- Routing to deal with this is a little arcane…
  - 3+1 physical routers
  - ~7 virtual routers
  - Nodes have ~16 IPv4 and ~8 IPv6 routing table entries
  - More default route (gateway) options than subnets

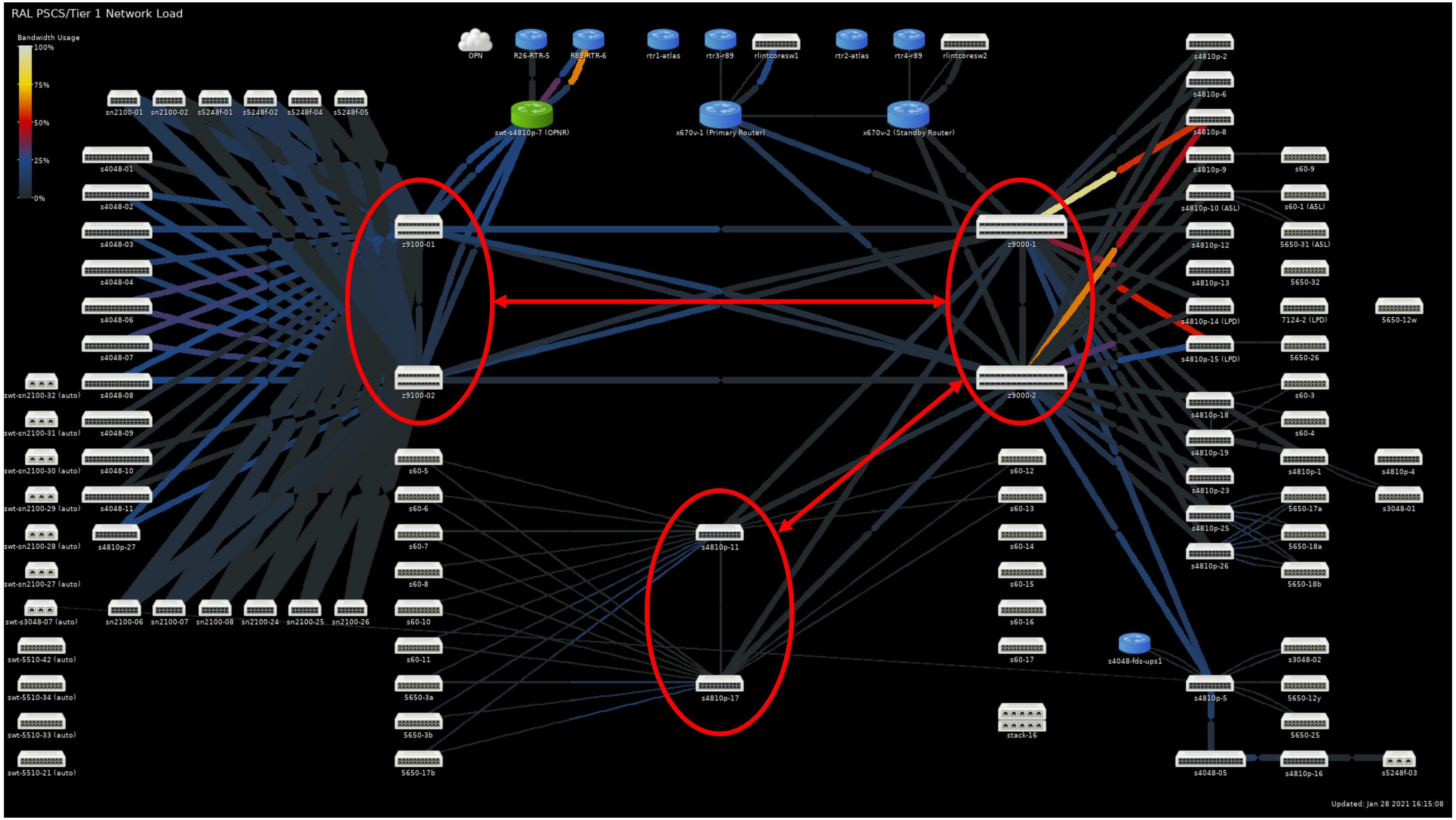
Tier-1 Network



RAL PSCS/Tier 1 Network Load

Updated: Jan 28 2021 16:15:08

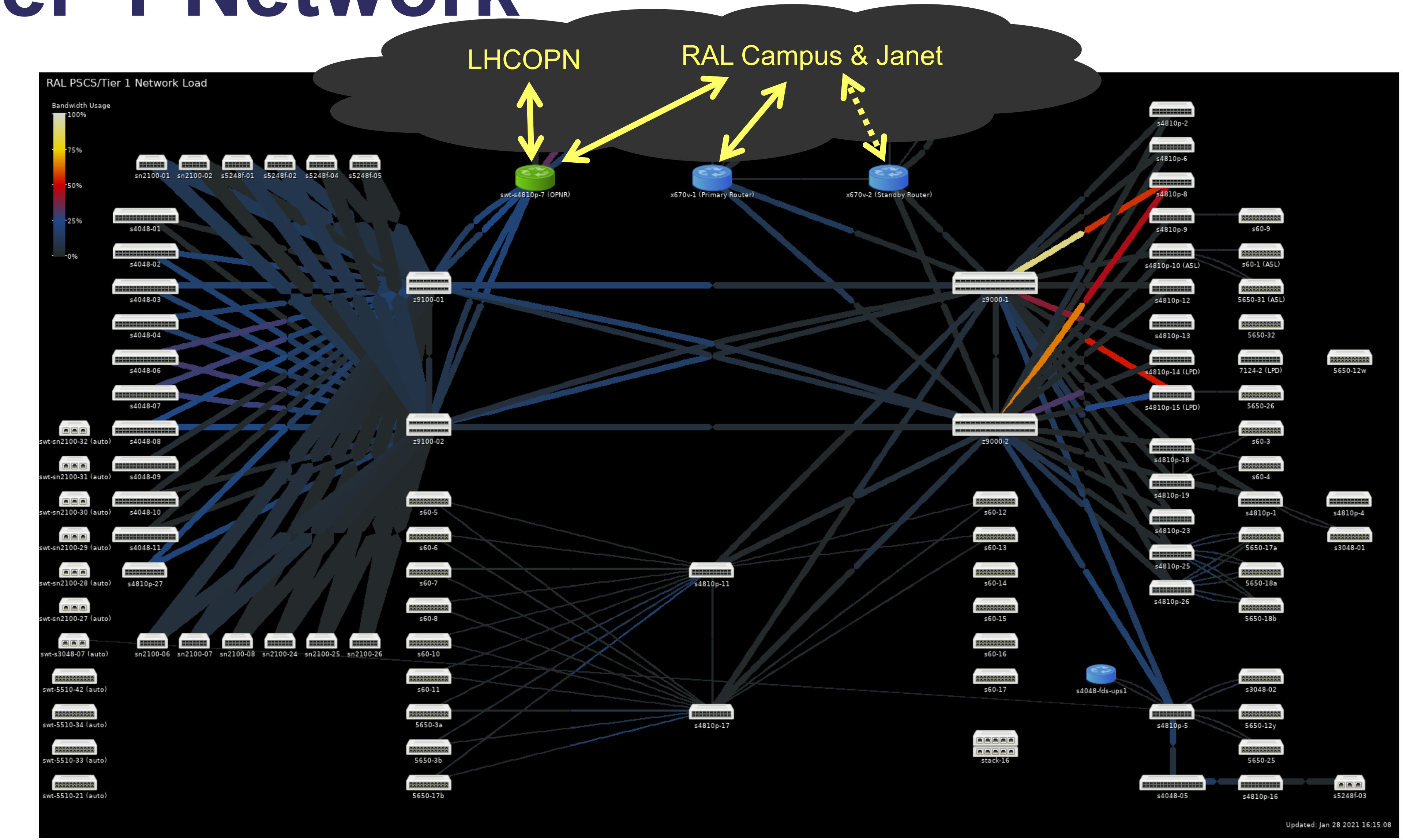
Tier-1 Network



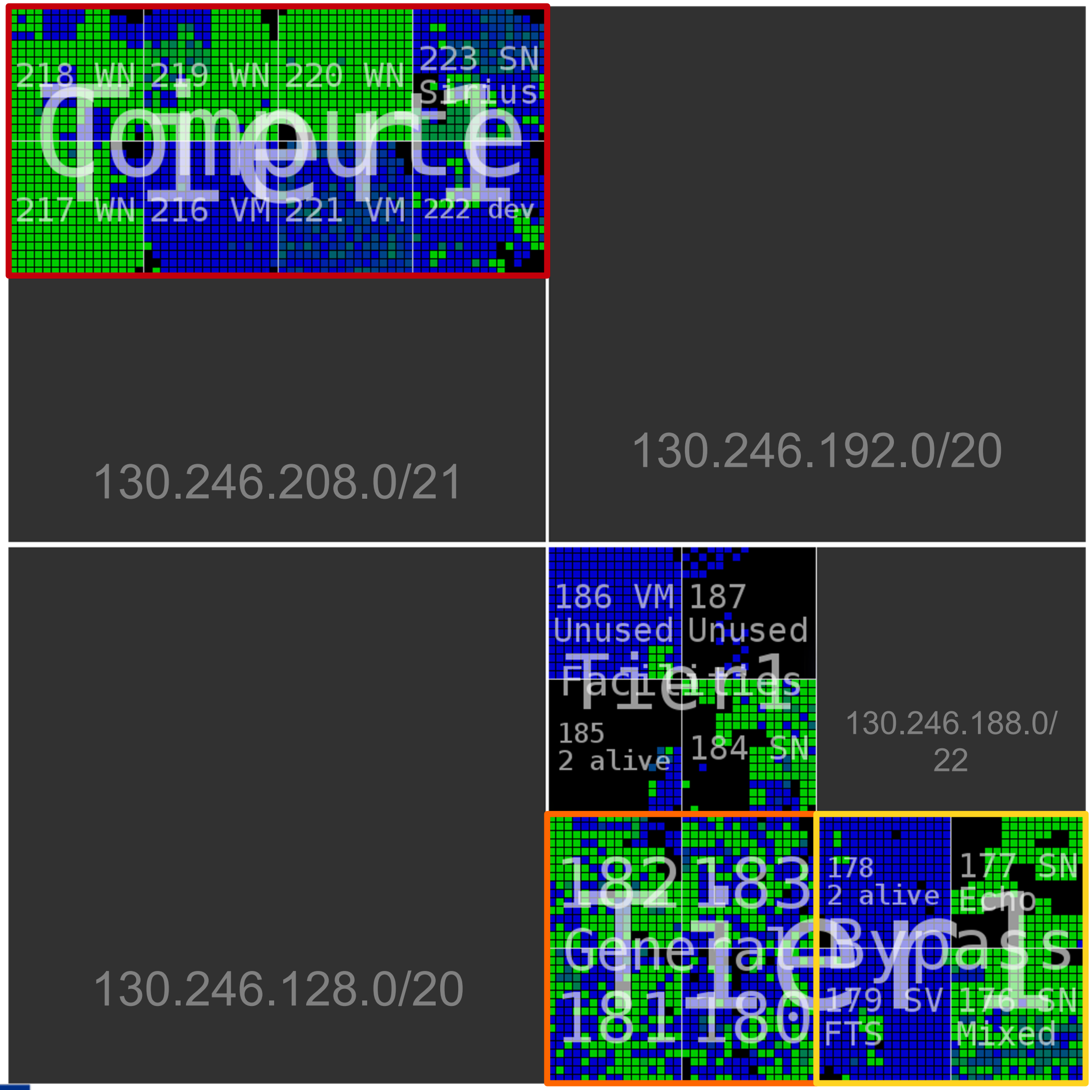
RAL PSCS/Tier 1 Network Load

Updated: Jan 28 2021 16:15:08

Tier-1 Network



Tier-1 Subnets



3 x Tier 1 subnets:

OPN: 130.246.176.0/22

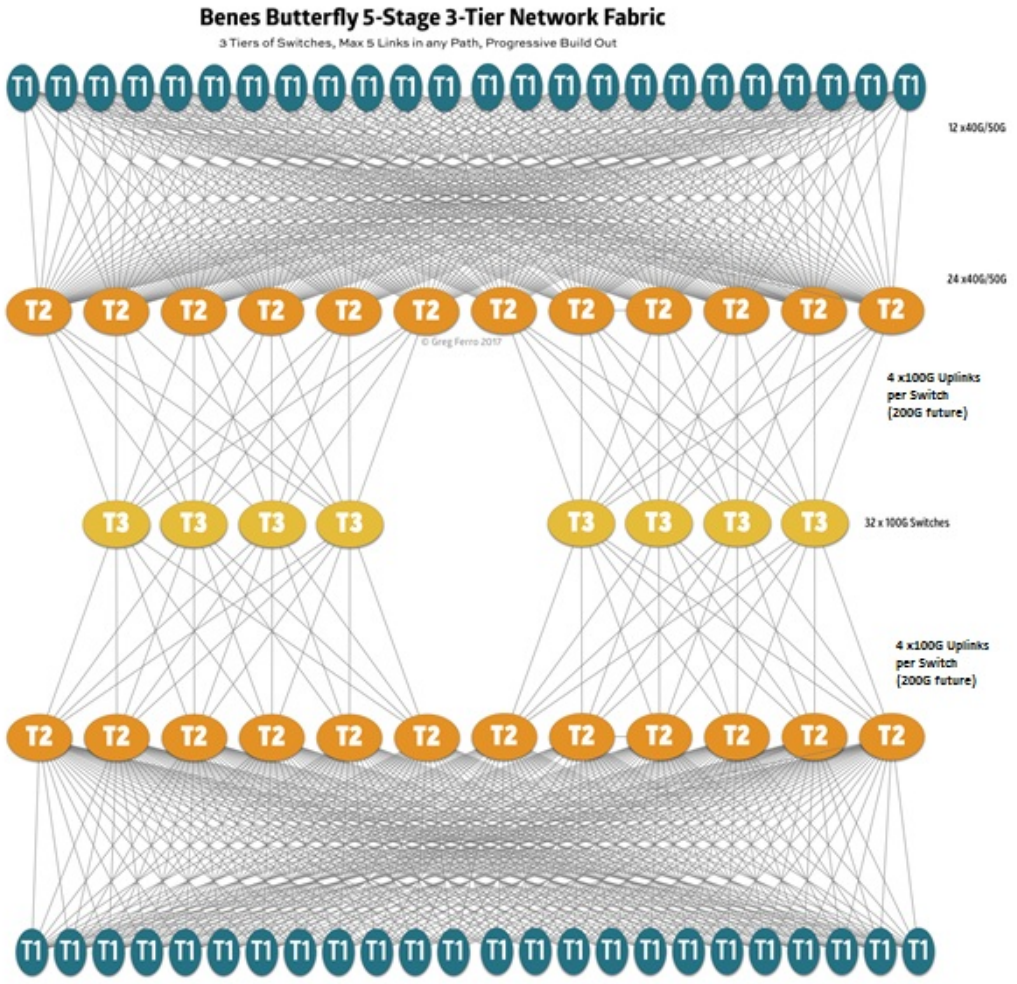
Services: 130.246.180.0/22

Compute: 130.246.216.0/21

Subnet design was done a decade before services like CMS AAA were thought of.

SCD Super Spine

- In 2017 - 18, Jonathan Churchill designed an built an SCD Super Spine.
  - Originally to move data between Jasmin projects.
  - Currently high-bandwidth bypass of site core.
- 3 Tier, Spine/leaf architecture following data centre best practises.
- 16 x SN2700 switches in 4 blocks.
  - 32 x 100Gb/s each
- Cabling has been laid to connect up the Tier-1.



Benes Butterfly 5-Stage 3-Tier Network Fabric

3 Tiers of Switches, Max 5 Links in any Path, Progressive Build Out

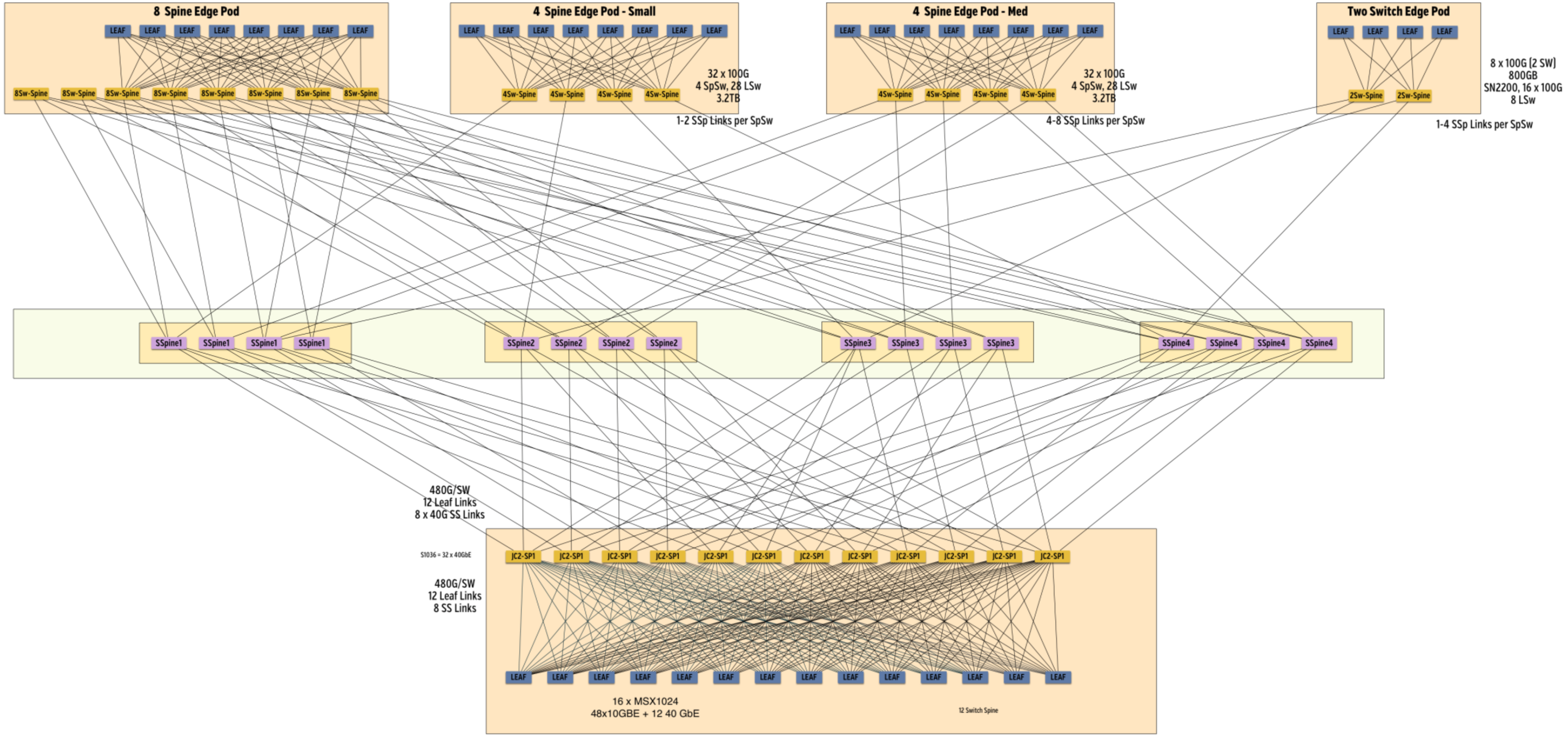
© Greg Ferro 2017

T1
T2
T3
12 x40G/50G
24 x40G/50G
4 x100G Uplinks per Switch (200G future)
32 x 100G Switches
4 x100G Uplinks per Switch (200G future)

SCD Super Spine

5 Stage Benes Fabric, Core/Pod Architecture, Variable Pod Size

Sweet Sans Pro - Modular Build Out strategy for SDN Networking



Science and
Technology
Facilities Council

Requirements

Requirements

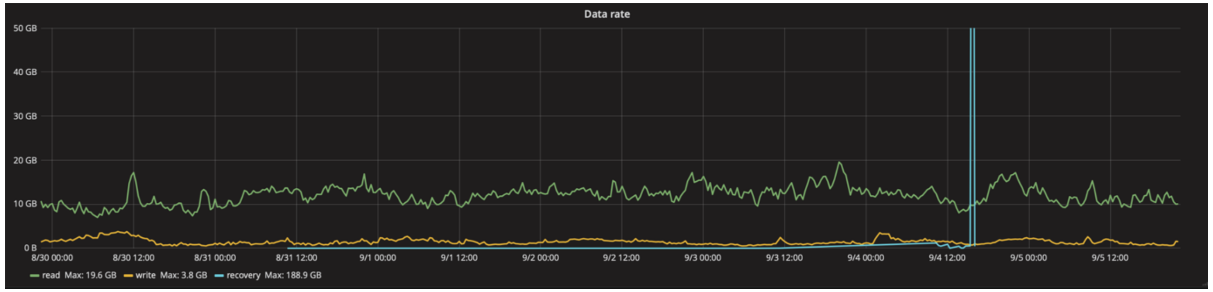
- The WLCG has estimated some throughput requirements for the different Tier-1s.
- Everything must support IPv6.
- Must join the LHCONE.
  - With a single 100Gb/s OPN link, will use LHCONE as failover.
- Must be future proof.
  - Easily scale up bandwidth.
  - Be able to take part in future network activities (e.g. SKAONE)

RAL-LCG2	2021 Target (Gb/s)	2023 Target (Gb/s)	2025 Target (Gb/s)	2027 Target (Gb/s)
OPN to CERN	50	116	198	331
Over JANET	50	116	198	331

Storage Requirements - Ceph

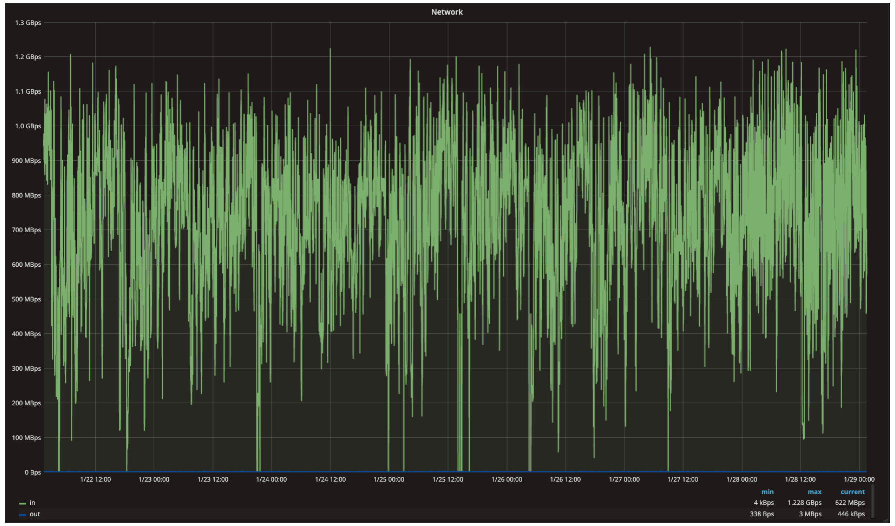
- Most production load on Echo is internal to RAL.
- Ceph cluster network:
  - Doubles bandwidth due to Erasure Coding across nodes.
  - Can saturate disk I/O during rebalancing (e.g. after a node failure).
- **We aim to provide ~1Gb/s per HDD**

The plot shows the throughput in and out of Echo during the first week in September 2020. This is a relatively quiet time as the LHC is not taking data. During previous data taking periods average read rates were 20 – 30GB/s with peaks as high as 50GB/s.



Storage Requirements - Tape

- The Tape Robotics are expensive and should therefore be the bottleneck in the system.
- Each Tape Server can saturate a 10Gb/s link.
- We have 20 Tape Servers currently although will probably need more.
- **Aim to provide 200Gb/s capacity.**
- CTA, uses SSD buffers to maintain this kind of performance.



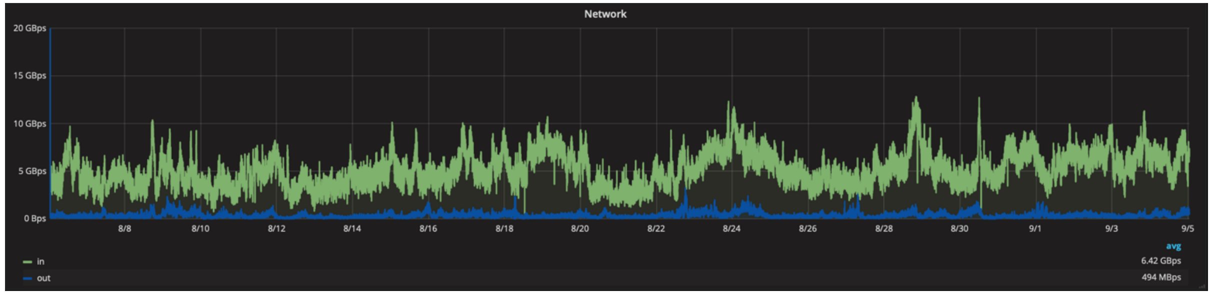
Plot shows the network throughput of a tape server that is migrating data to the new Robot.

Castor performance isn't quite as good as CTA.

CPU Requirements

- Measuring the average throughput of the most recent generation of CPUs provides 6.42GB/s reads.
- 12288 jobs slots with 95% average occupancy.
- **Require ~0.5MB/s per job slot.**
- We have also measured the write rate to SSD and get a similar number.

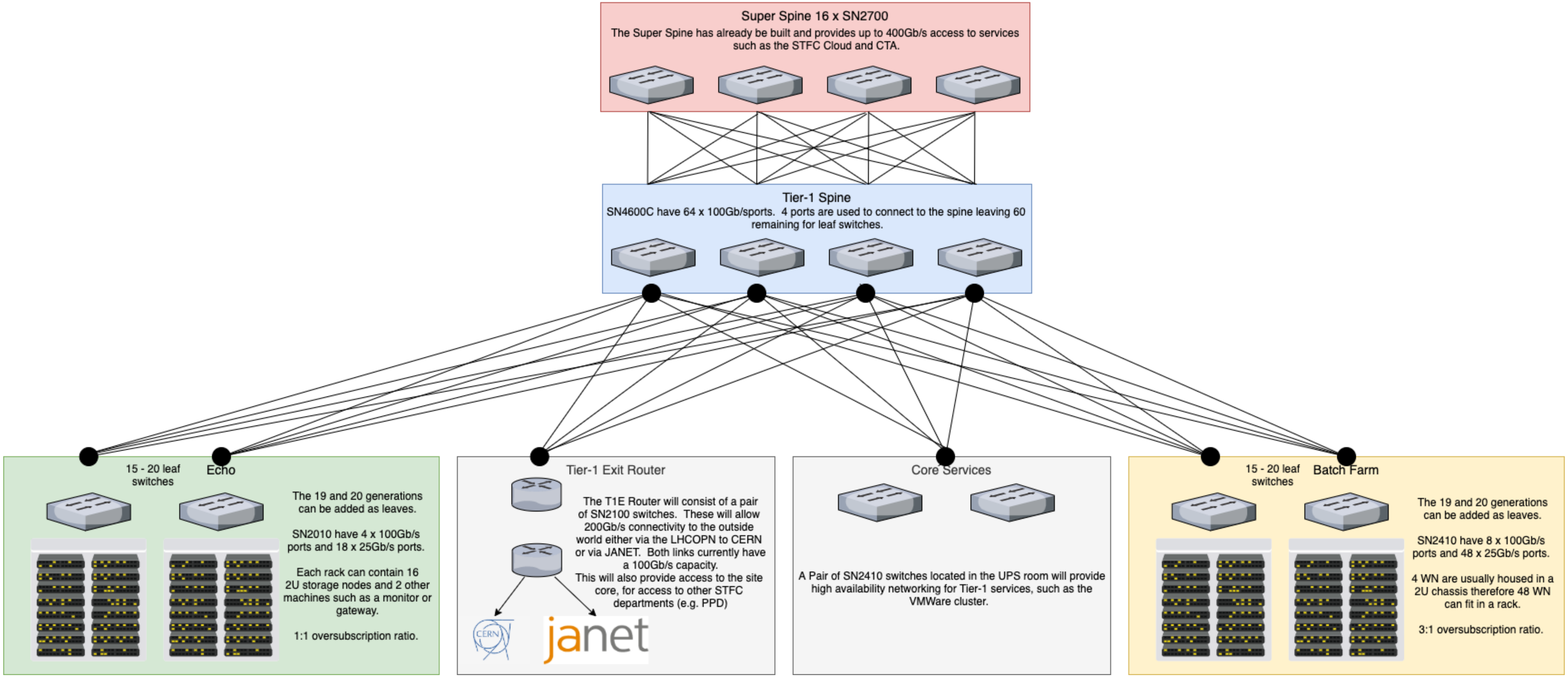
Figure shows the total network throughput for the Dell19 servers during August and the start of September 2020.



Science and Technology Facilities Council

Design

Tier-1 Network Architecture



Super Spine 16 x SN2700

The Super Spine has already be built and provides up to 400Gb/s access to services such as the STFC Cloud and CTA.

Tier-1 Spine

SN4600C have 64 x 100Gb/sports. 4 ports are used to connect to the spine leaving 60 remaining for leaf switches.

The 19 and 20 generations can be added as leaves.

SN2010 have 4 x 100Gb/s ports and 18 x 25Gb/s ports.

Each rack can contain 16 2U storage nodes and 2 other machines such as a monitor or gateway.

1:1 oversubscription ratio.

Tier-1 Exit Router

The T1E Router will consist of a pair of SN2100 switches. These will allow 200Gb/s connectivity to the outside world either via the LHCOPN to CERN or via JANET. Both links currently have a 100Gb/s capacity.

This will also provide access to the site core, for access to other STFC departments (e.g. PPD)

Core Services

A Pair of SN2410 switches located in the UPS room will provide high availability networking for Tier-1 services, such as the VMWare cluster.

The 19 and 20 generations can be added as leaves.

SN2410 have 8 x 100Gb/s ports and 48 x 25Gb/s ports.

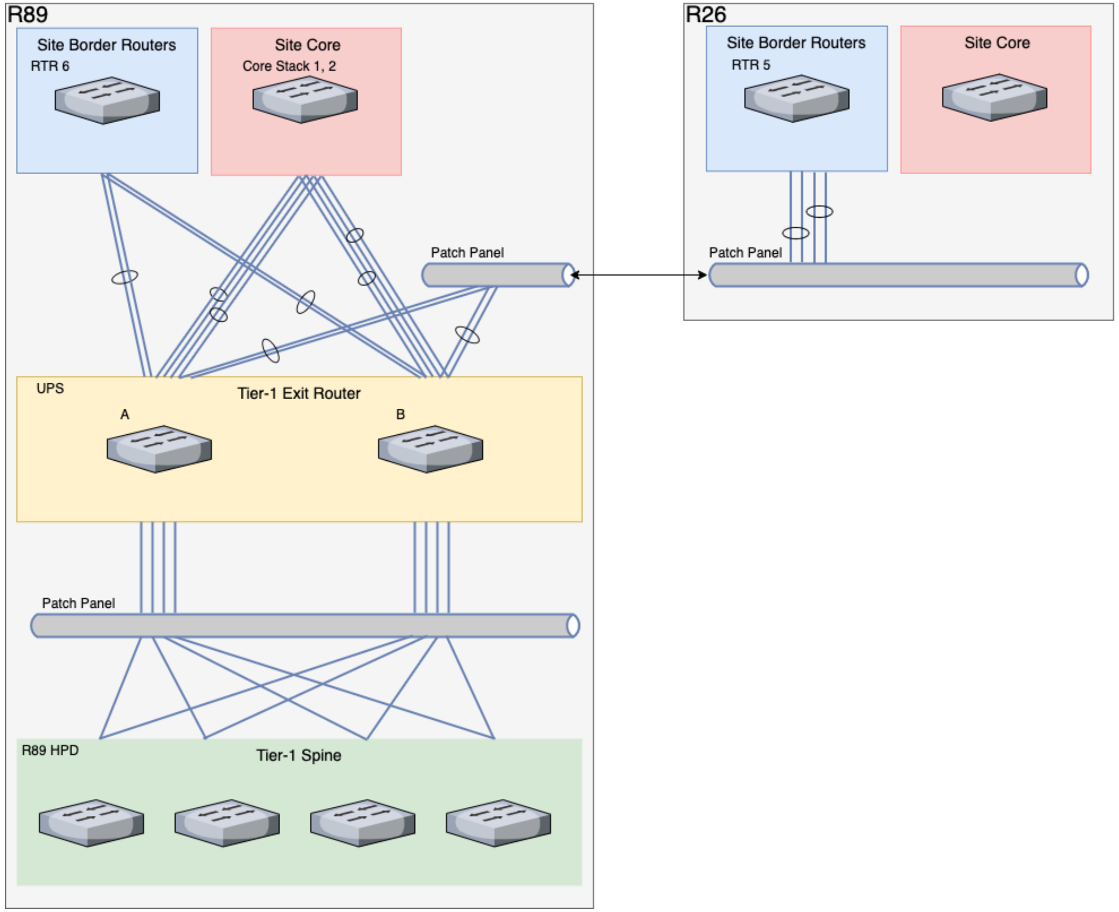
4 WN are usually housed in a 2U chassis therefore 48 WN can fit in a rack.

3:1 oversubscription ratio.

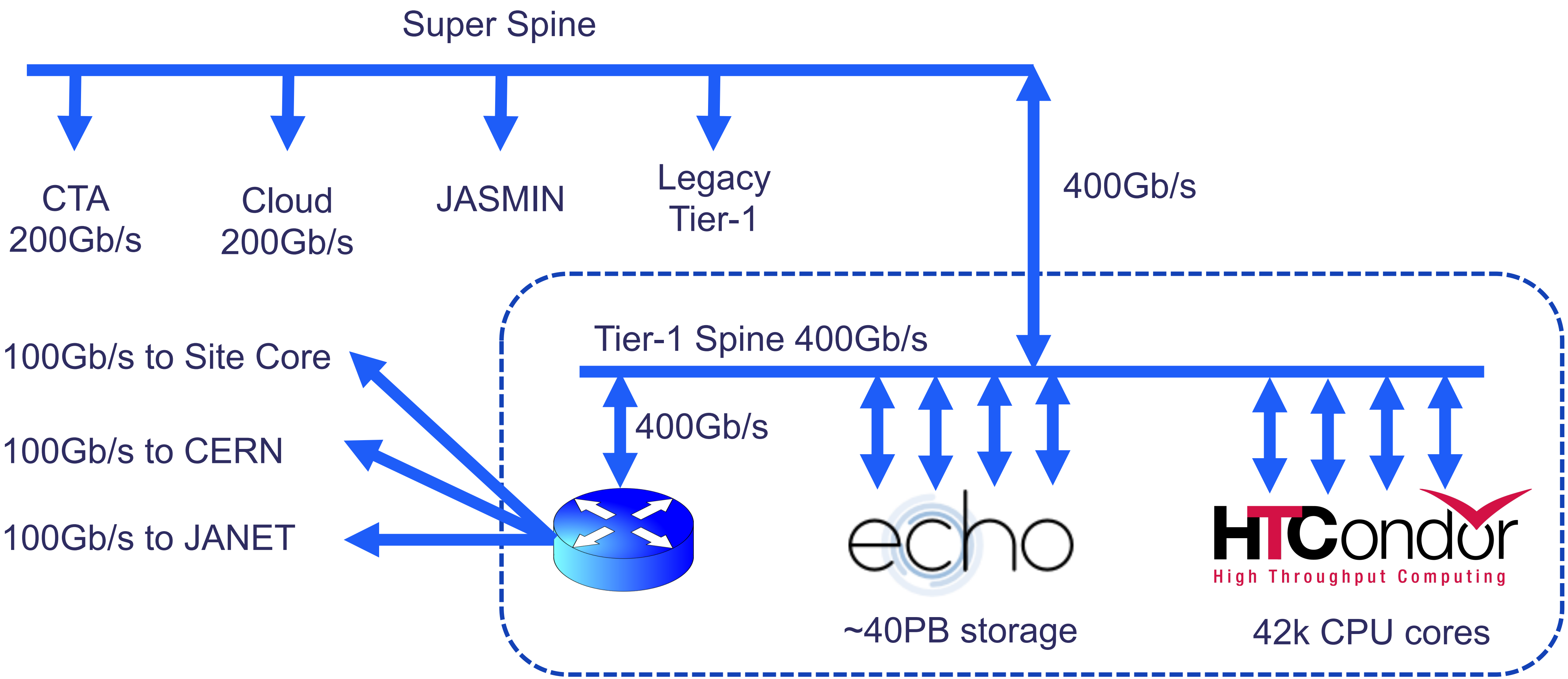
Tier-1 Exit Router

- T1E Router will be connected to both Border routers.
- OPN link currently lands on Border Router 6.
- Initially we will connect to site core network in R89 only.
  - Site core is being upgraded soon.

Tier-1 Exit Router



Core network infrastructure

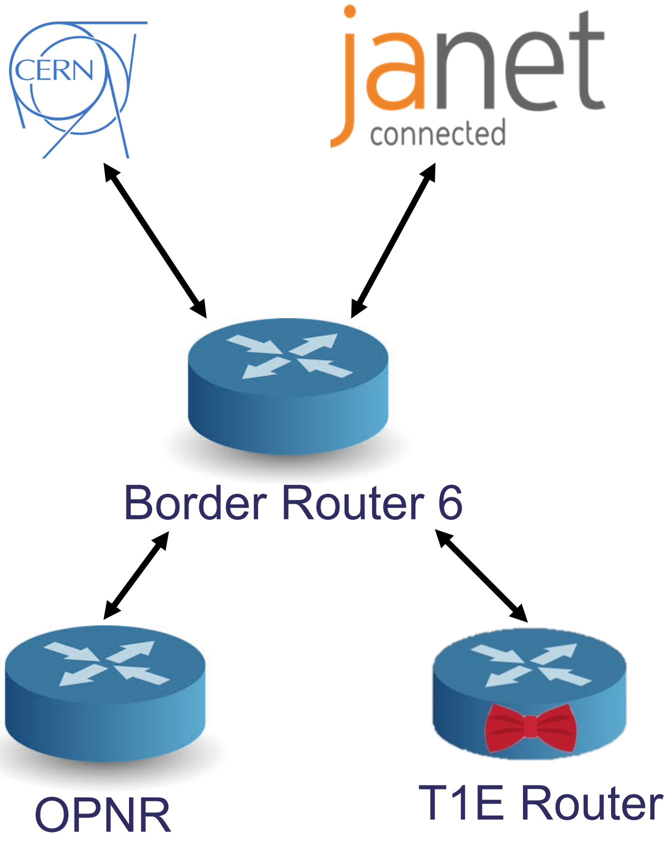


Science and
Technology
Facilities Council

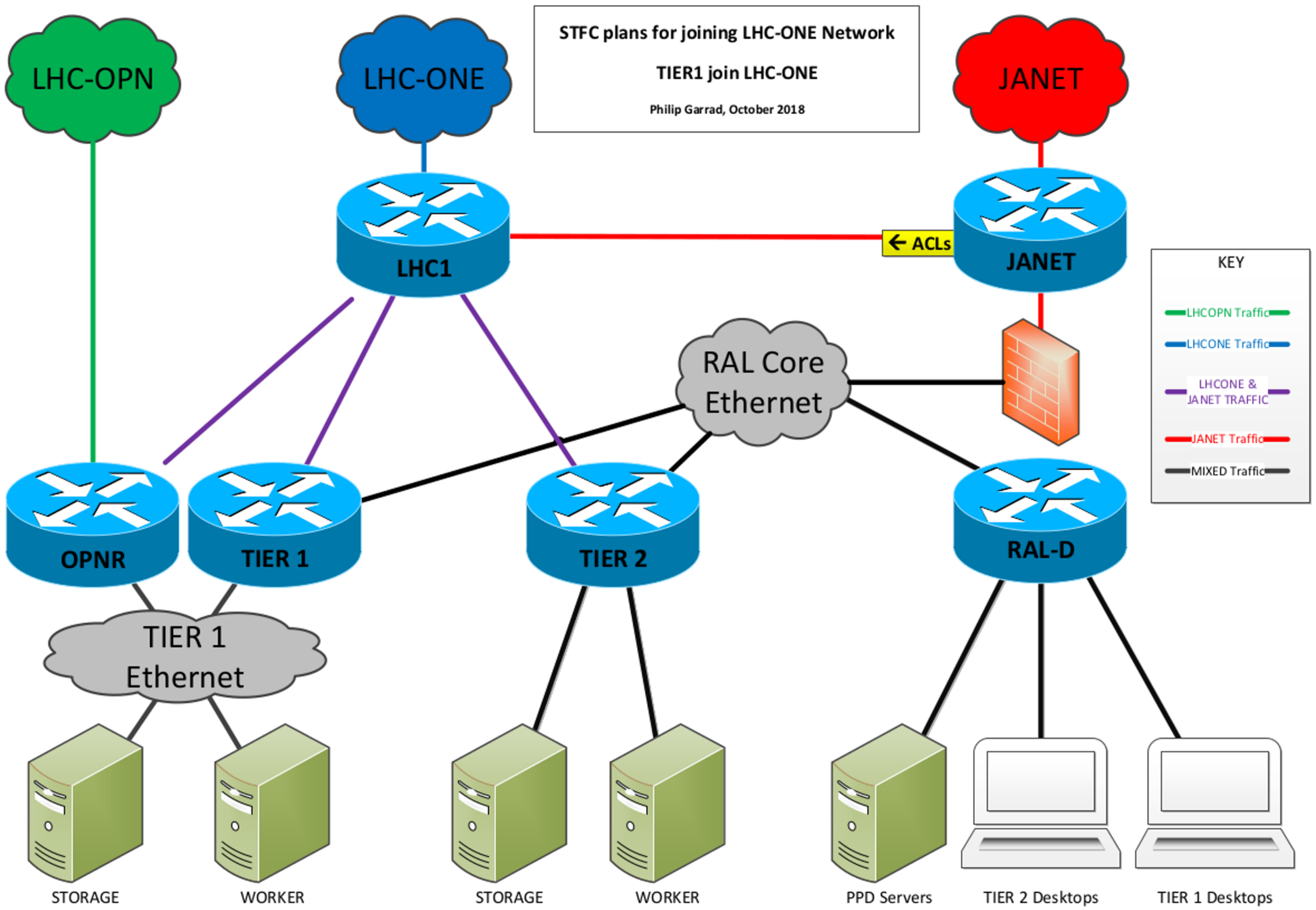
Migration

Peering

- Currently the OPNR peers with CERN.
- We believe only one of the OPNR and T1E Router can do all the peering.



Old LHCONE plan



STFC plans for joining LHC-ONE Network

TIER1 join LHC-ONE

Philip Garrad, October 2018

Time Line

1)Build new Tier-1 Network

- XMA are doing all installation and cabling inside the data centre. Hopefully allowed in from start of March. Target completion April 1st.
- Dedicated contractor effort (Anil) to configure setup. Target completion May 1st.

2)Connect Network pods to Super Spine

- Can happen in parallel to 1). Target completion May 1st.

3)Switch Peering from OPNR to TIE Router.

4)Announce 130.246.216.0/21 and 2001:630:58:1820/64 to LHCOPN.

5)Announce 130.246.216.0/21 and 2001:630:58:1820/64 to LHCONE.

6)Migrate older hardware to new network. Q3 2021

Science and
Technology
Facilities Council

Questions?