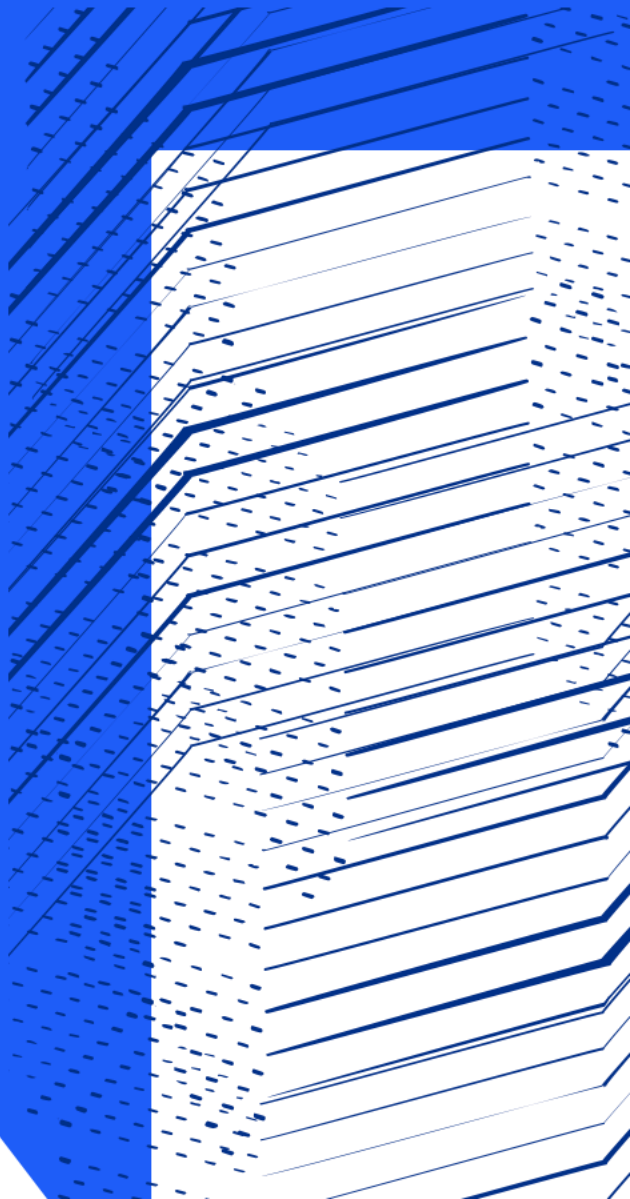
Science and Technology Facilities Council

RAL-LCG2 Network Outage October 2022

Alastair Dewhurst



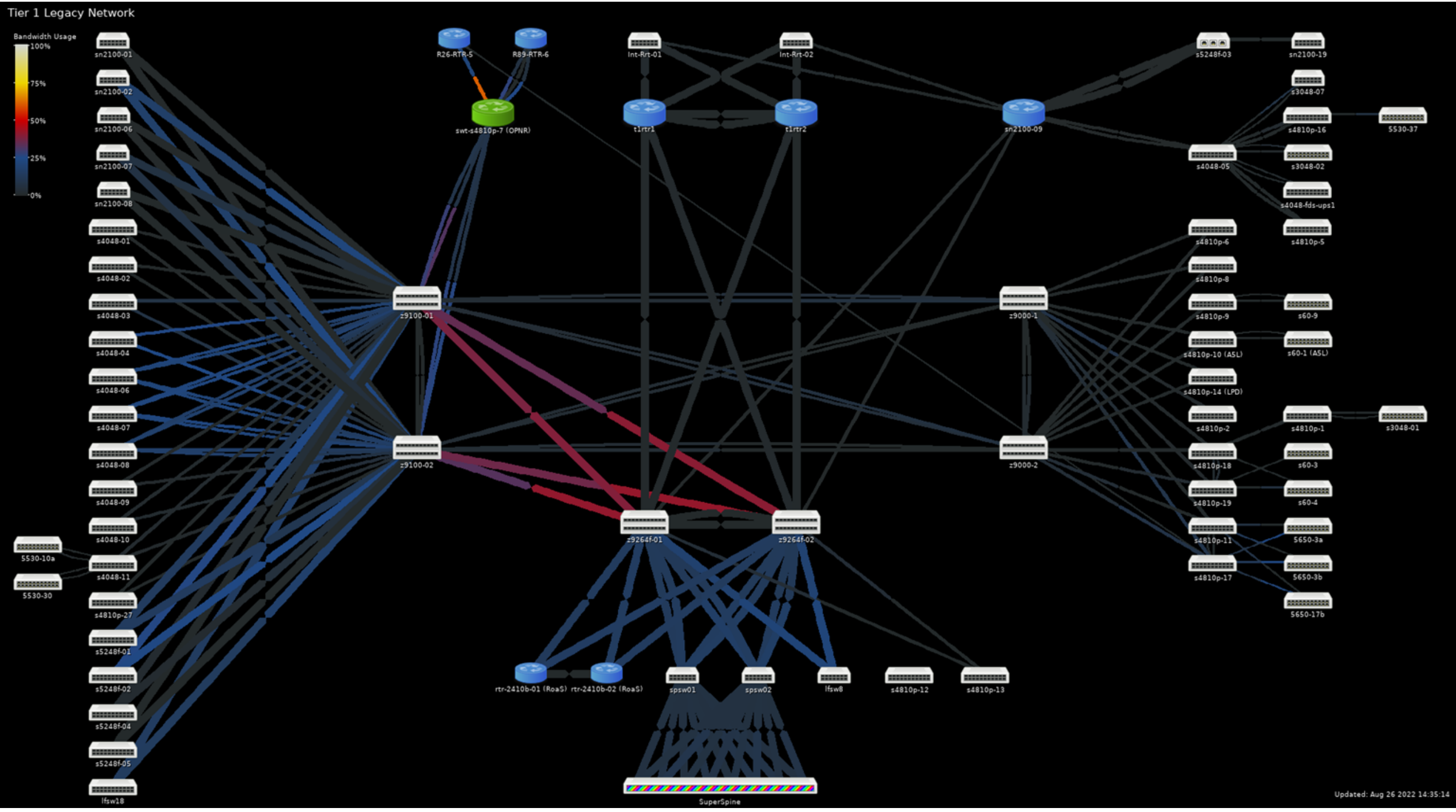
Introduction

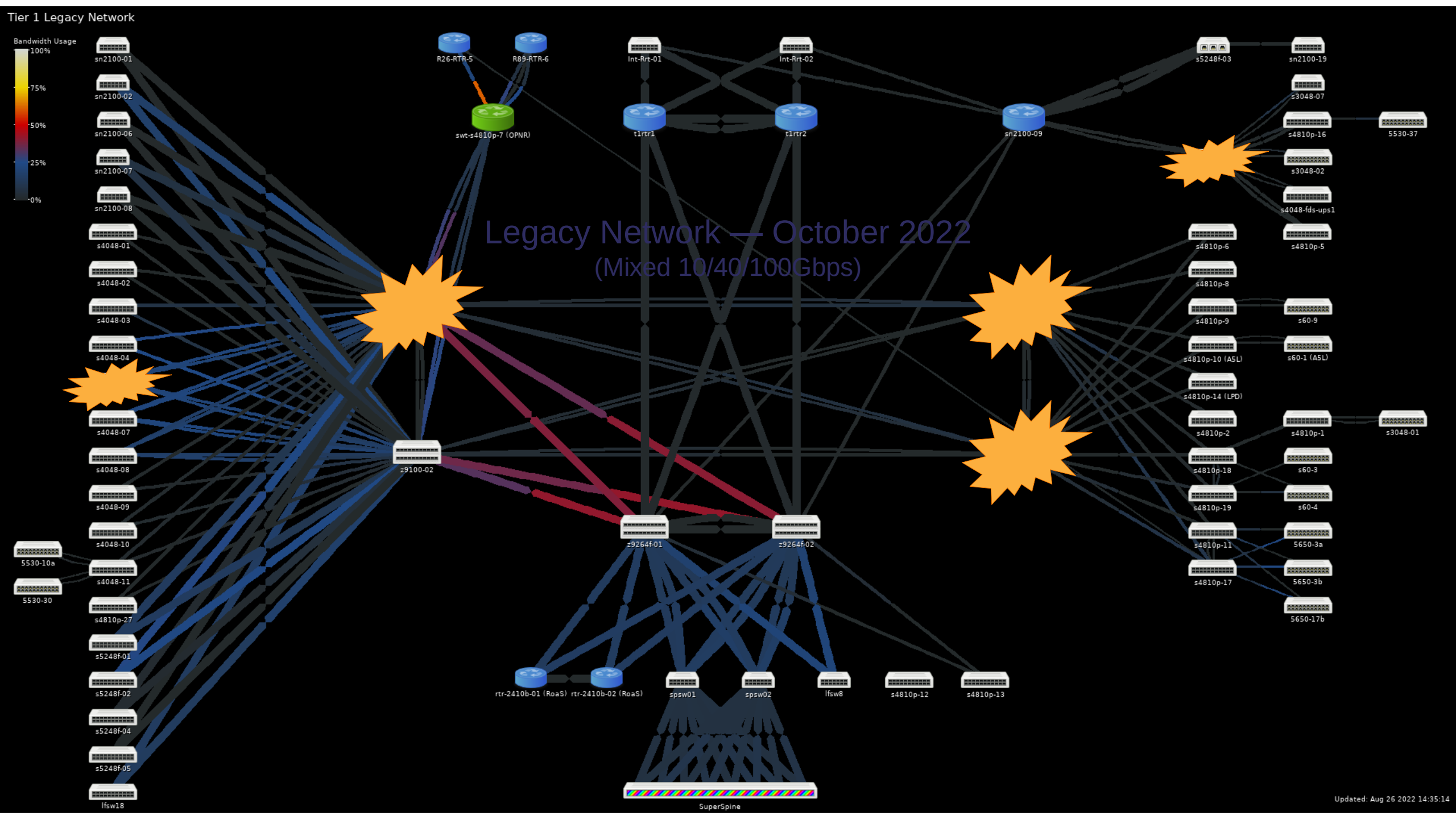
- This report summarises the RAL-LCG2 Network outage which involved two outages and caused significant degradation to services for ~10 days.
- Full SIR can be found here.

Legacy Network Meltdown

On Monday 17th October 2022 ~15:00 local time:

- One of the two core Z9100 failed
- One of the Facilities database server switches suffered a hardware failure
- Took out RT ticketing system, FTS and GOCDB
- One of the echo storage switches crashed
  - Took out 25 storage nodes
- Lots of spanning tree changes on management network
- Rapid toplogy changes caused instability of MLAG & VRRP in core
- Remaining core switches failed over to broadcast (i.e. hub) mode


Tier 1 Legacy Network
Bandwidth Usage
100%
75%
50%
25%
0%
sn2100-01
sn2100-02
sn2100-06
sn2100-07
sn2100-08
R26-RTR-5
R89-RTR-6
swt-s4810p-7 (OPNR)
Int-Rrt-01
Int-Rrt-02
t1rtr1
t1rtr2
sn2100-09
s5248f-03
sn2100-19
s3048-07
s4810p-16
5530-37
s4048-05
s3048-02
s4048-fds-ups1
s4048-01
s4048-02
s4048-03
s4048-04
s4048-06
s4048-07
s4048-08
s4048-09
s4048-10
s4048-11
z9100-01
z9100-02
z9000-1
z9000-2
s4810p-6
s4810p-8
s4810p-9
s4810p-10 (A5L)
s4810p-14 (LPD)
s4810p-2
s4810p-18
s4810p-19
s4810p-11
s4810p-17
s4810p-5
s60-9
s60-1 (A5L)
s4810p-1
s3048-01
s60-3
s60-4
5650-3a
5650-3b
5650-17b
5530-10a
5530-30
z9264f-01
z9264f-02
rtr-2410b-01 (RoaS)
rtr-2410b-02 (RoaS)
spsw01
spsw02
lfsw8
s4810p-12
s4810p-13
s4810p-27
s5248f-01
s5248f-02
s5248f-04
s5248f-05
lfsw18
SuperSpine
Updated: Aug 26 2022 14:35:14

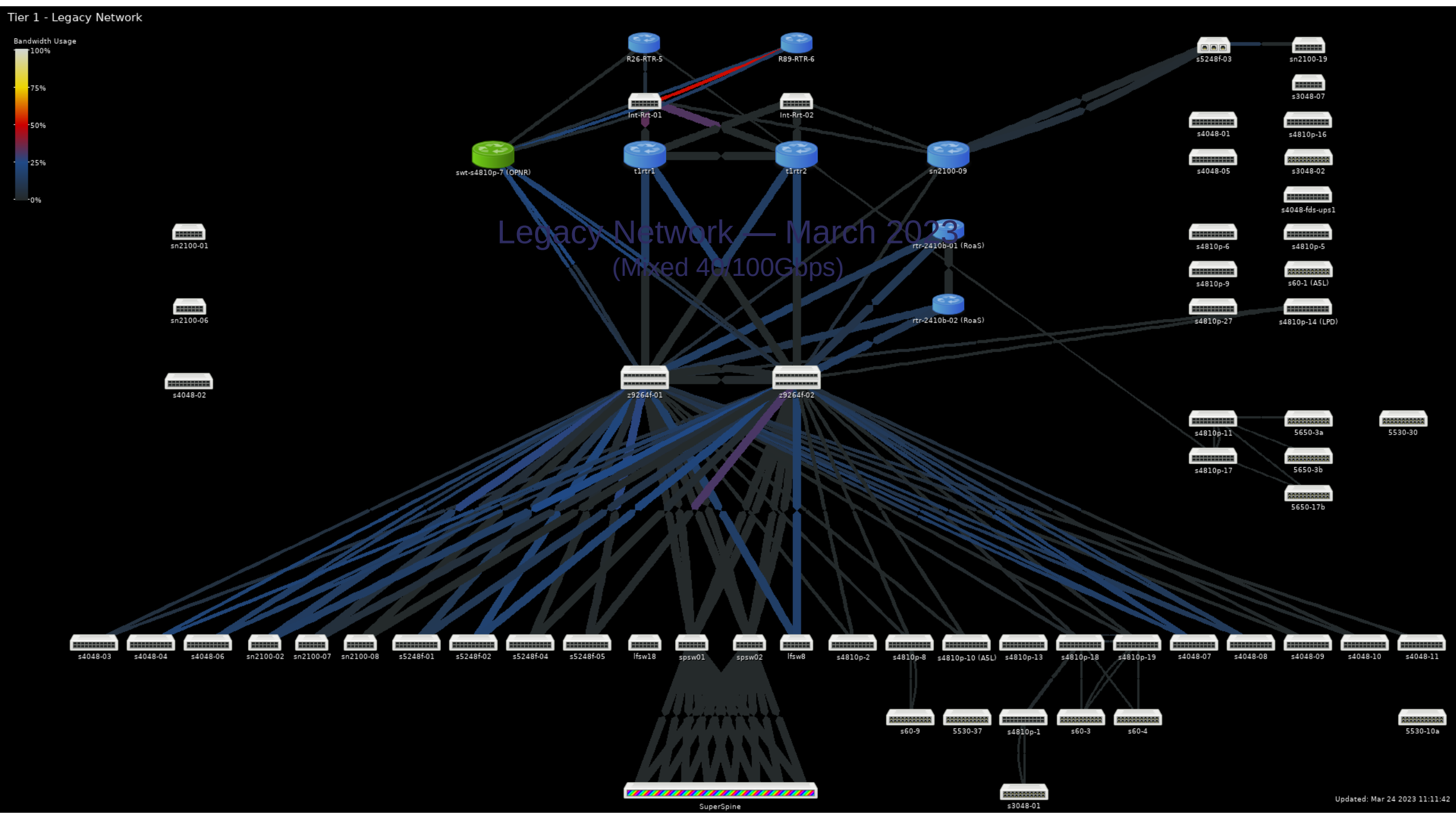

Tier 1 Legacy Network
Bandwidth Usage
100%
75%
50%
25%
0%
Legacy Network — October 2022
(Mixed 10/40/100Gbps)
sn2100-01
sn2100-02
sn2100-06
sn2100-07
sn2100-08
s4048-01
s4048-02
s4048-03
s4048-04
s4048-07
s4048-08
s4048-09
s4048-10
s4048-11
5530-10a
5530-30
s4810p-27
s5248f-01
s5248f-02
s5248f-04
s5248f-05
lfsw18
R26-RTR-5
R89-RTR-6
swt-s4810p-7 (OPNR)
Int-Rrt-01
Int-Rrt-02
t1rtr1
t1rtr2
z9100-02
z9264f-01
z9264f-02
rtr-2410b-01 (RoaS)
rtr-2410b-02 (RoaS)
spsw01
spsw02
lfsw8
s4810p-12
s4810p-13
SuperSpine
sn2100-09
s5248f-03
sn2100-19
s3048-07
s4810p-16
5530-37
s3048-02
s4048-fds-ups1
s4810p-5
s4810p-6
s4810p-8
s4810p-9
s60-9
s4810p-10 (A5L)
s60-1 (A5L)
s4810p-14 (LPD)
s4810p-2
s4810p-1
s3048-01
s4810p-18
s60-3
s4810p-19
s60-4
s4810p-11
5650-3a
s4810p-17
5650-3b
5650-17b
Updated: Aug 26 2022 14:35:14

Initial Response

- Key staff away at HEPiX
- Additional faults found when trying to restart
- Restoring VLT crashed switches immediately
- Fan and power supply failed
- Z9000s & Z9100s out-of-maintenance
  - No spare hardware
- We were able to get the network functional by the end of the day (17th) and many services returned to production the following day.
- However we had no resilience and batch and Echo needed to be limited.
- One good thing: Plans had been made for migrating off them

Follow up response

- Decided to make 12 months of planned changes in a single day
  - Scheduled an intervention for 26th October.
  - Moved 56 cables and config from four core switches
- Not everything came back correctly so downtime had to be extended.
  - It wasn't bad for the complexity of the change and the time pressure.
- Disabled management network entirely
  - Designed an interim replacement architecture
  - Replacement has been slow, but steady progress.


Tier 1 - Legacy Network
Bandwidth Usage
100%
75%
50%
25%
0%
Legacy Network — March 2023
(Mixed 40/100Gbps)
R26-RTR-5
R89-RTR-6
Int-Rrt-01
Int-Rrt-02
swt-s4810p-7 (OPNR)
t1rtr1
t1rtr2
sn2100-09
rtr-2410b-01 (RoaS)
rtr-2410b-02 (RoaS)
s5248f-03
sn2100-19
s3048-07
s4048-01
s4810p-16
s4048-05
s3048-02
s4048-fds-ups1
sn2100-01
s4810p-6
s4810p-5
s4810p-9
s60-1 (A5L)
sn2100-06
s4810p-27
s4810p-14 (LPD)
s4048-02
z9264f-01
z9264f-02
s4810p-11
5650-3a
5530-30
s4810p-17
5650-3b
5650-17b
s4048-03
s4048-04
s4048-06
sn2100-02
sn2100-07
sn2100-08
s5248f-01
s5248f-02
s5248f-04
s5248f-05
lfsw18
spsw01
spsw02
lfsw8
s4810p-2
s4810p-8
s4810p-10 (A5L)
s4810p-13
s4810p-18
s4810p-19
s4048-07
s4048-08
s4048-09
s4048-10
s4048-11
s60-9
5530-37
s4810p-1
s60-3
s60-4
5530-10a
SuperSpine
s3048-01
Updated: Mar 24 2023 11:11:42

Science and Technology Facilities Council

Lessons learnt

Lesson learnt – legacy network

- We weren't surprised that the Legacy network failed.
  - We have been working to replace it since 2020.
- Between 2015 and 2018 there was no permanent RAL Tier-1 Manager
  - Strategic decisions about the network weren't made which delayed necessary upgrades and replacements.
- The legacy network has been rebuilt simpler and with less ancient hardware and will be completely phased out by end of 2024.
- The new network is starting to deliver significant benefits (e.g. LHCONE, multiple 100Gb/s OPN links)
- The criticality of the network has been recognized and effort to support it is a core part of the GridPP7 funding bid.

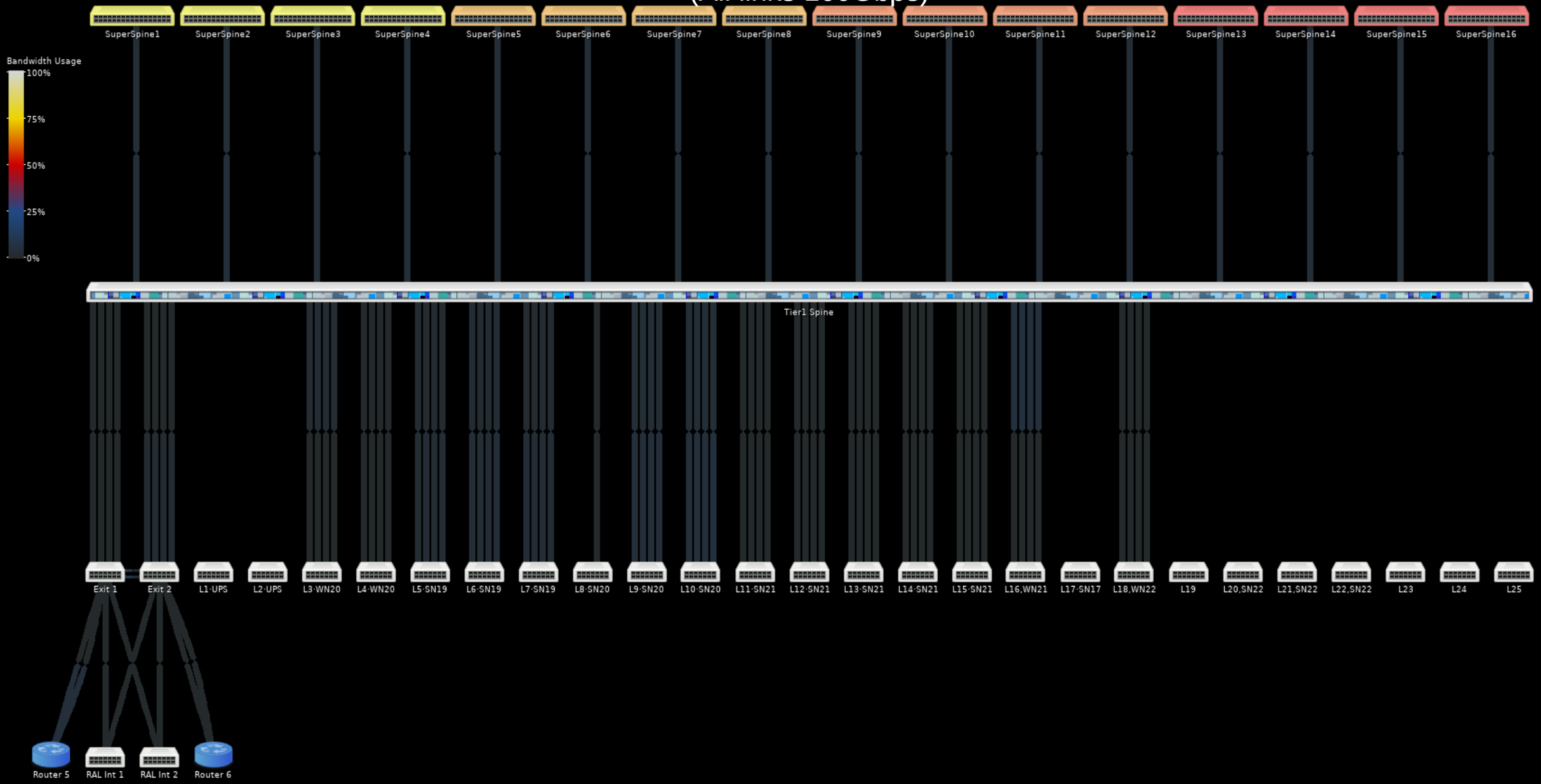
Lesson learnt – operational

- When RAL has a major outage it is likely to affect both the standard Tier-1 services and GOCDB.
  - Better documentation has been put in place to ensure that more people know how to announce a downtime when you can't use the GOCDB.
- One Antares server (CTA-Frontend) was setup on the legacy Tier-1 network.
  - This has now been moved to the correct location on the network.
- The criticality of the management network for disaster recovery had been underestimated.
  - We have invested £350k to purchase 110 new switches to modernize the hardware.
- We were already in the process of moving our documentation and ticket system to Jira and Confluence in the Cloud.
  - This should allow for more effective Disaster Recovery as it won't be affected by an outage.

Tier 1 - New Network

New Network — March 2023

(All links 100Gbps)



Science and
Technology
Facilities Council

Questions?