The RAL Tier-1 Network

James Adams
RAL, STFC, UKRI
GridPP49, Abingdon

2023-03-30

Overview

- Background
- What’s Changed?
- What’s Next?

Background

- Building out a new network for the Tier-1 alongside the “legacy” network
  - Fully-routed eBGP ECMP architecture
  - Mellanox switches running Cumulus Linux
  - Joined to legacy network by SCD SuperSpine
- Started work July 2021
- Connected to SCD SuperSpine October 2021
- Connected to RAL site November 2021
- First worker nodes live by December 2021

What’s Changed?

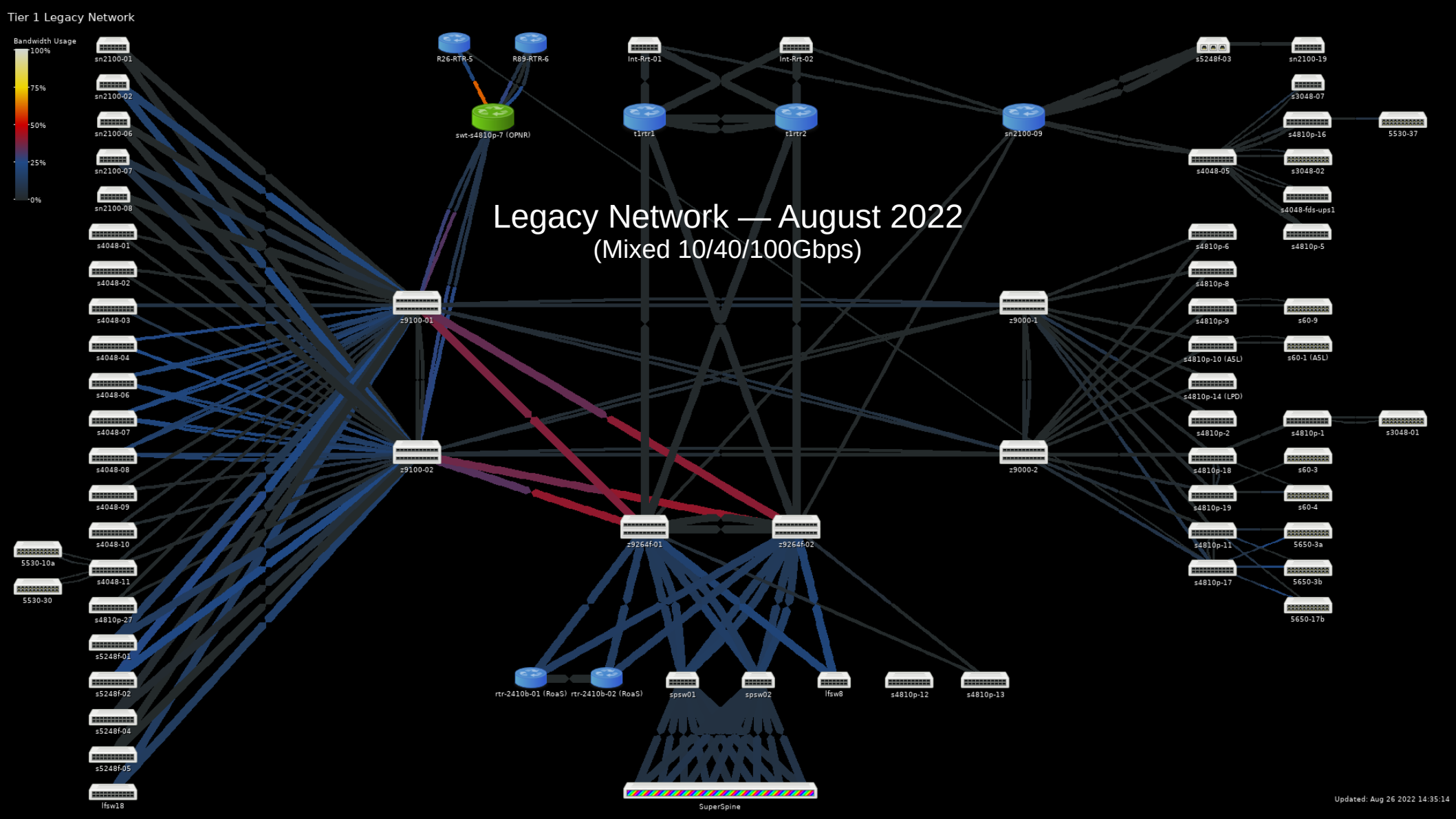
- Legacy network melt-down
- More hardware on new network
- Peering with LHCONE
- IPv6

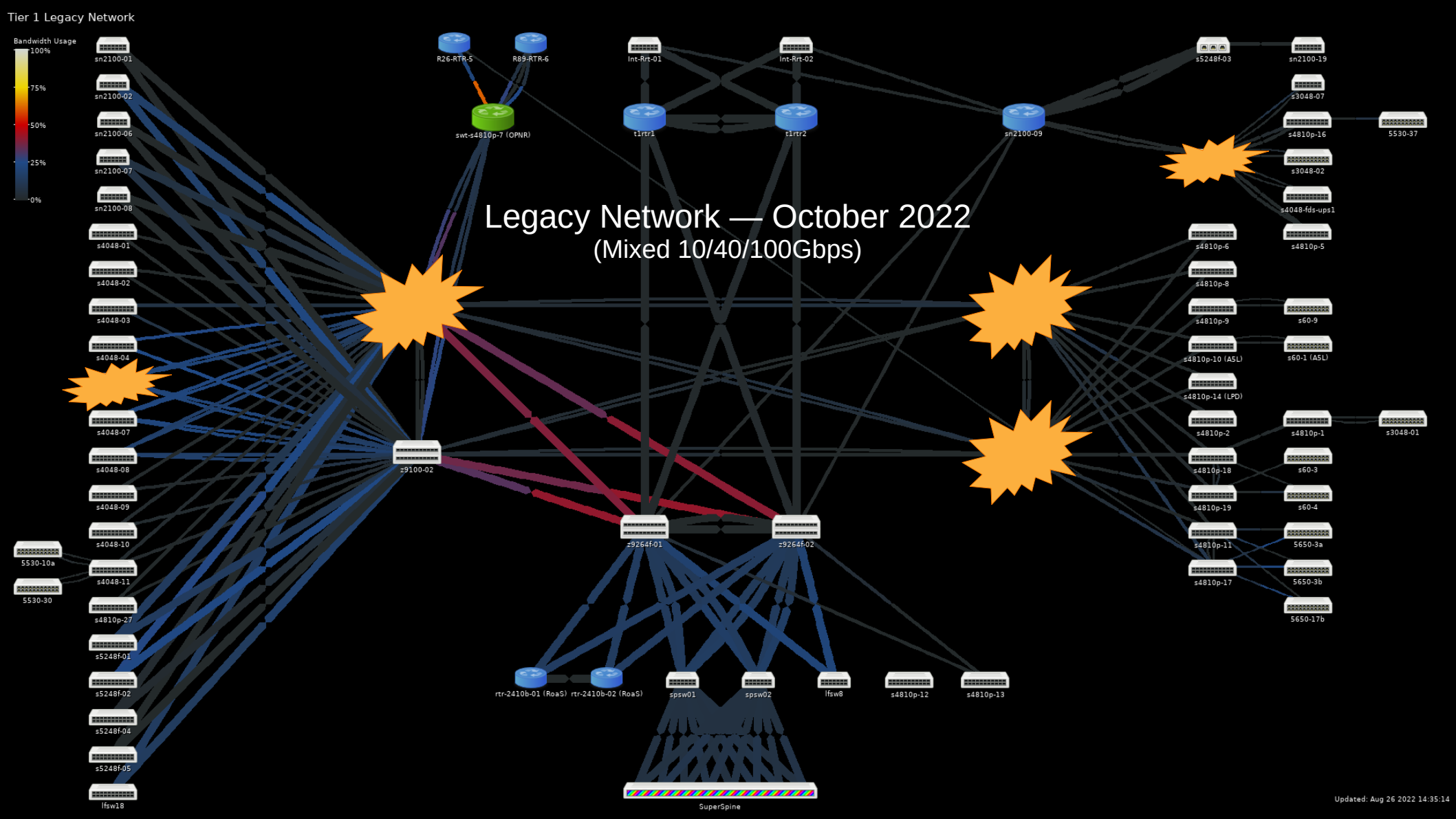
Legacy Network Meltdown

- 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

Legacy Network Meltdown

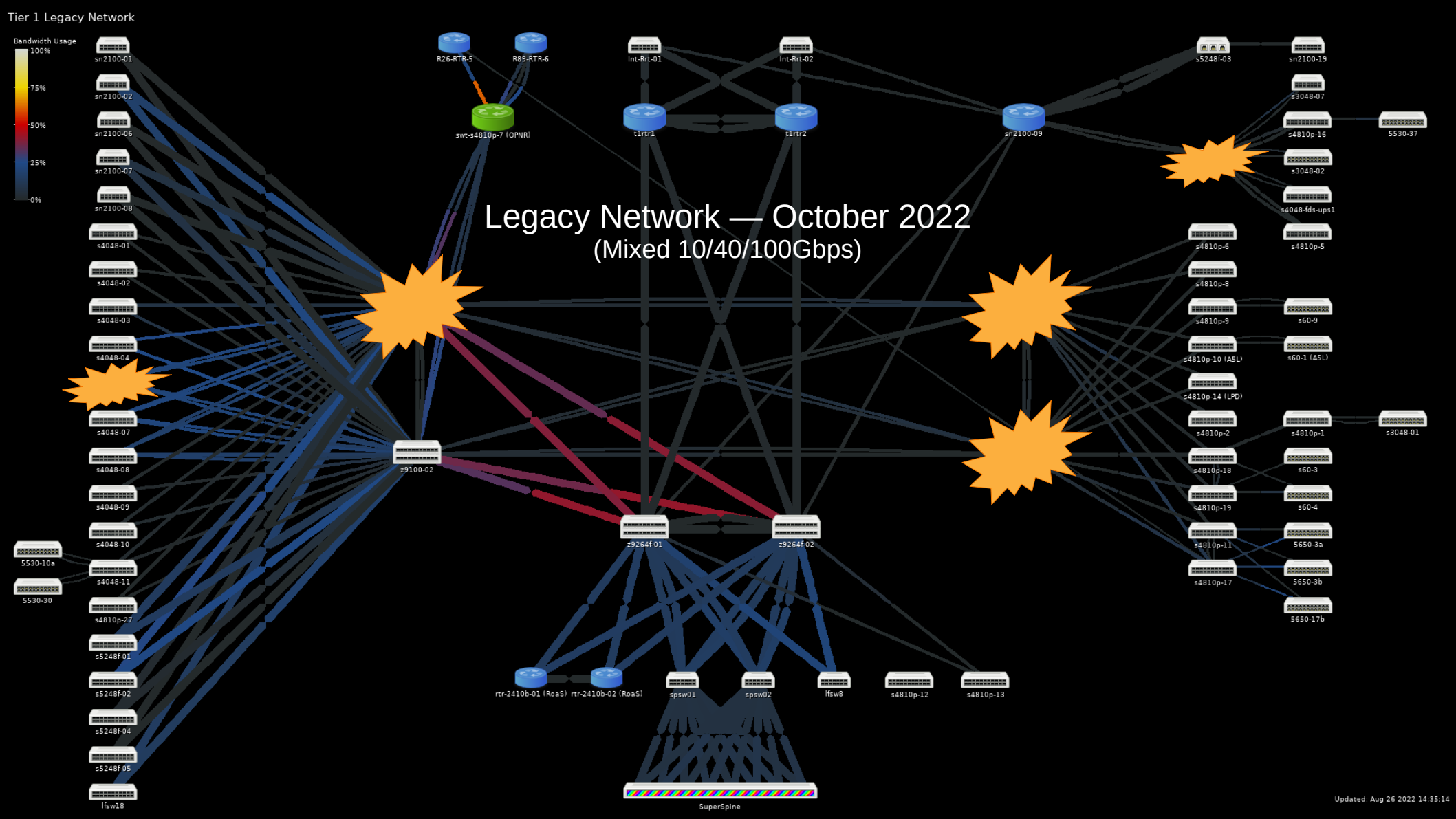
- Key staff away at HEPiX
- Additional faults found when trying to restart
  - Restoring VLT crashed switches immediately
  - Fan and power supply failed
- Z9000s & Z9100s out-of-maintenance
  - No spare hardware
  - Planning had already started for migrating off them

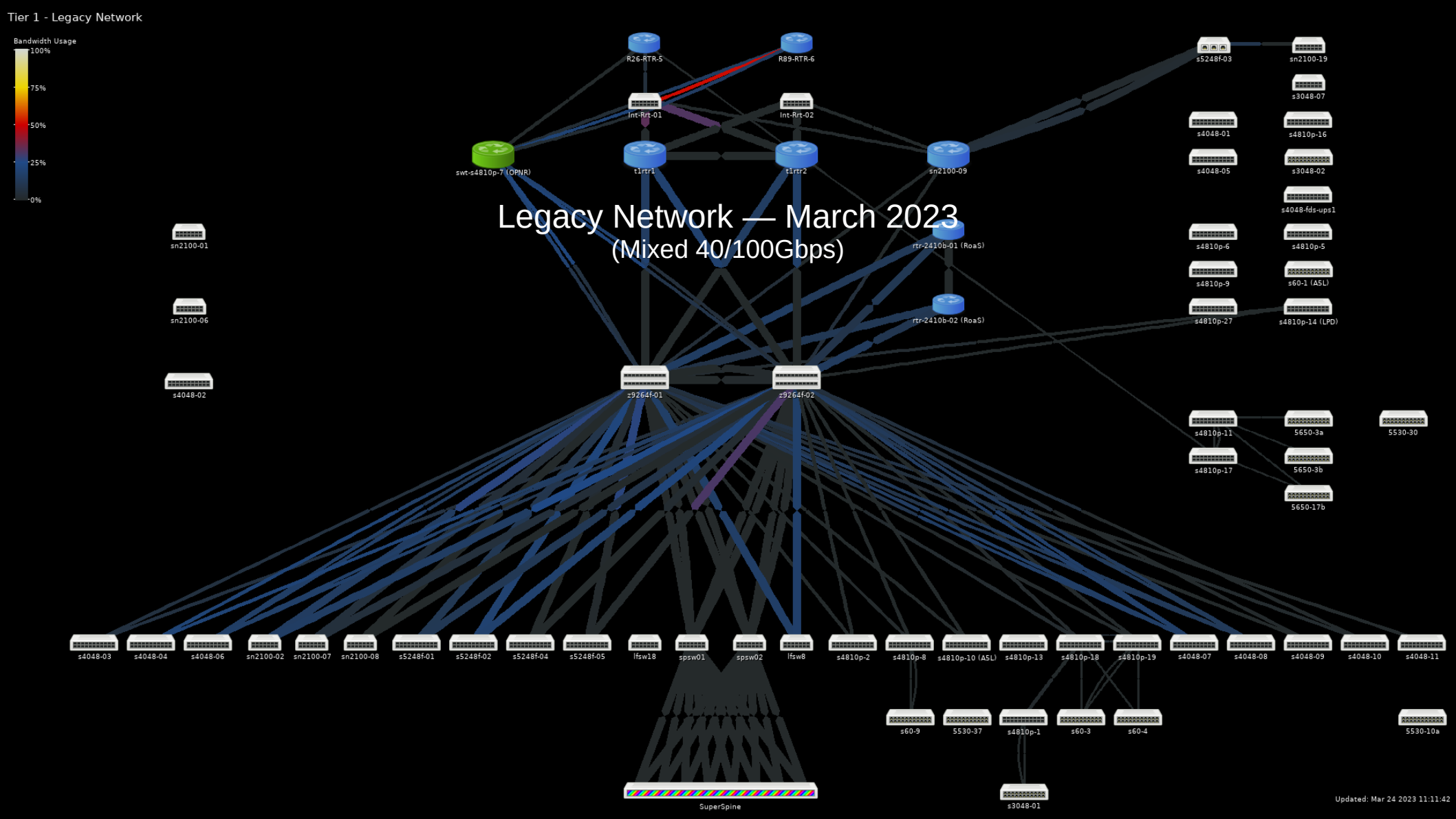

Tier 1 Legacy Network
Bandwidth Usage
100%
75%
50%
25%
0%
Legacy Network — August 2022
(Mixed 10/40/100Gbps)
sn2100-01
sn2100-02
sn2100-06
sn2100-07
sn2100-08
R26-RTR-5
R89-RTR-6
swt-s4810p-7 (OPNR)
Int-Rrt-01
Int-Rrt-02
t1rtr1
t1rtr2
sn2100-09
s5248f-03
sn2100-19
s3048-07
s4810-16
5530-37
s4048-05
s3048-02
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 (ASL)
s4810p-14 (LPD)
s4810p-2
s4810p-18
s4810p-19
s4810p-11
s4810p-17
s4810p-5
s60-9
s60-1 (ASL)
s4810p-1
s3048-01
s60-3
s60-4
5650-3a
5650-3b
5650-17b
5530-10a
5530-30
s4810p-27
s5248f-01
s5248f-02
s5248f-04
s5248f-05
lfsw18
z9264f-01
z9264f-02
rtr-2410b-01 (RoaS)
rtr-2410b-02 (RoaS)
spsw01
spsw02
lfsw8
s4810p-12
s4810p-13
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
R26-RTR-5
R89-RTR-6
swt-s4810p-7 (OPNR)
lnt-Rrt-01
lnt-Rrt-02
t1rtr1
t1rtr2
sn2100-09
s5248f-03
sn2100-19
s3048-07
s4810p-16
5530-37
s3048-02
s4048-fds-ups1
s4810p-5
s4048-01
s4048-02
s4048-03
s4048-04
s4048-07
s4048-08
s4048-09
s4048-10
s4048-11
s4810p-27
s5248f-01
s5248f-02
s5248f-04
s5248f-05
lfsw18
5530-10a
5530-30
z9100-02
z9264f-01
z9264f-02
rtr-2410b-01 (RoaS)
rtr-2410b-02 (RoaS)
spsw01
spsw02
lfsw8
s4810p-12
s4810p-13
SuperSpine
s4810p-6
s4810p-8
s4810p-9
s4810p-10 (ASL)
s4810p-14 (LPD)
s4810p-2
s4810p-18
s4810p-19
s4810p-11
s4810p-17
s60-9
s60-1 (ASL)
s4810p-1
s3048-01
s60-3
s60-4
5650-3a
5650-3b
5650-17b
Updated: Aug 26 2022 14:35:14

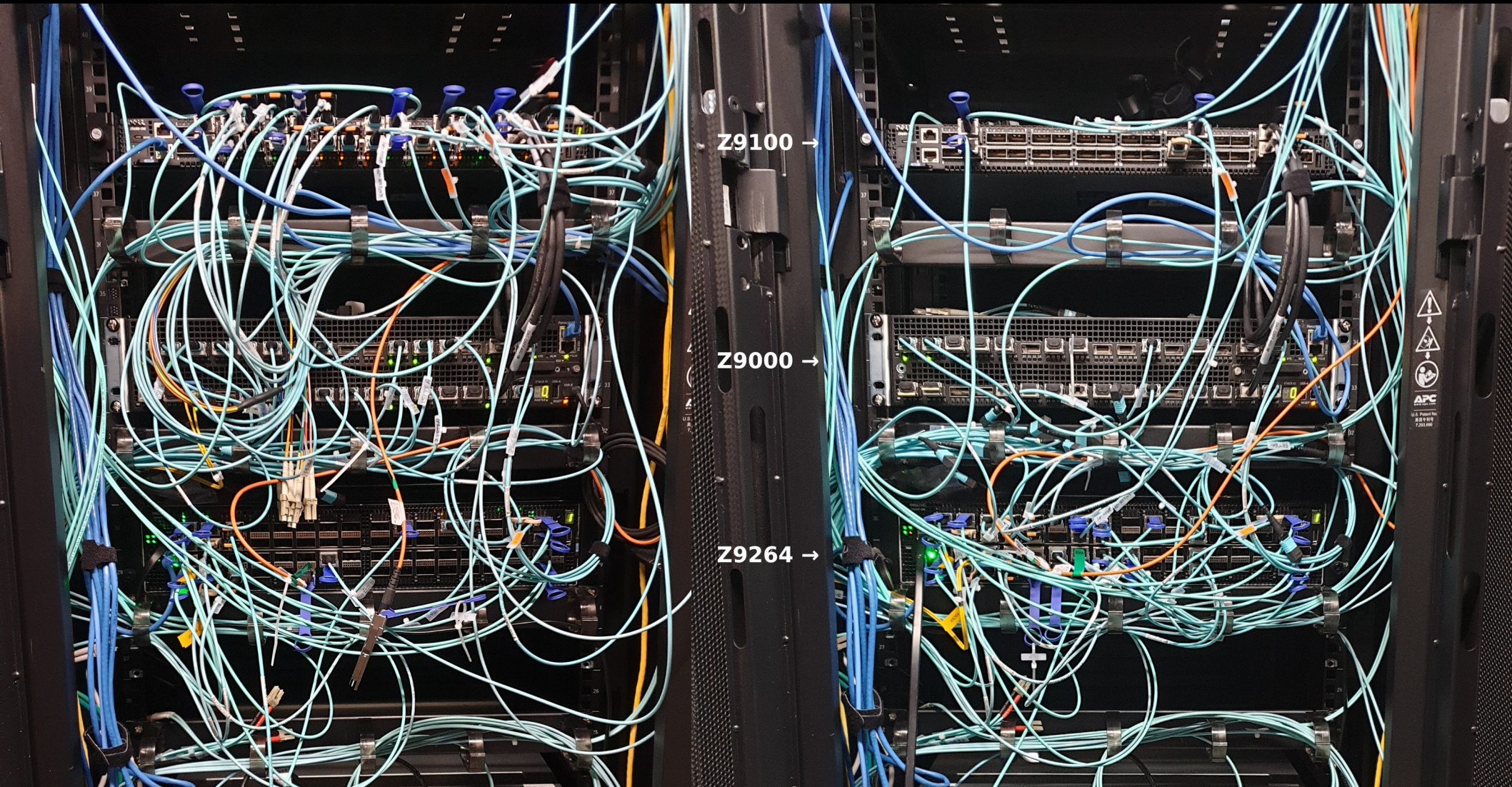
Legacy Network Meltdown

- Decided to make 12 months of planned changes in a single day
  - Moved 56 cables and config from four core switches
- 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 — October 2022
(Mixed 10/40/100Gbps)
sn2100-01
sn2100-02
sn2100-06
sn2100-07
sn2100-08
R26-RTR-5
R89-RTR-6
swt-s4810p-7 (OPNR)
lnt-Rrt-01
lnt-Rrt-02
t1rtr1
t1rtr2
sn2100-09
s5248f-03
sn2100-19
s3048-07
s4810p-16
5530-37
s3048-02
s4048-fds-ups1
s4810p-5
s4048-01
s4048-02
s4048-03
s4048-04
s4048-07
s4048-08
s4048-09
s4048-10
s4048-11
s4810p-27
s5248f-01
s5248f-02
s5248f-04
s5248f-05
lfsw18
5530-10a
5530-30
z9100-02
z9264f-01
z9264f-02
rtr-2410b-01 (RoaS)
rtr-2410b-02 (RoaS)
spsw01
spsw02
lfsw8
s4810p-12
s4810p-13
SuperSpine
s4810p-6
s4810p-8
s4810p-9
s4810p-10 (ASL)
s4810p-14 (LPD)
s4810p-2
s4810p-18
s4810p-19
s4810p-11
s4810p-17
s60-9
s60-1 (ASL)
s4810p-1
s3048-01
s60-3
s60-4
5650-3a
5650-3b
5650-17b
Updated: Aug 26 2022 14:35:14


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
s5248f-03
sn2100-19
s3048-07
s4048-01
s4810p-16
s4048-05
s3048-02
s4048-fds-ups1
sn2100-01
rtr-2410b-01 (RoaS)
s4810p-6
s4810p-5
s4810p-9
s60-1 (ASL)
sn2100-06
rtr-2410b-02 (RoaS)
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 (ASL)
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

Before *(all of this, twice)* After



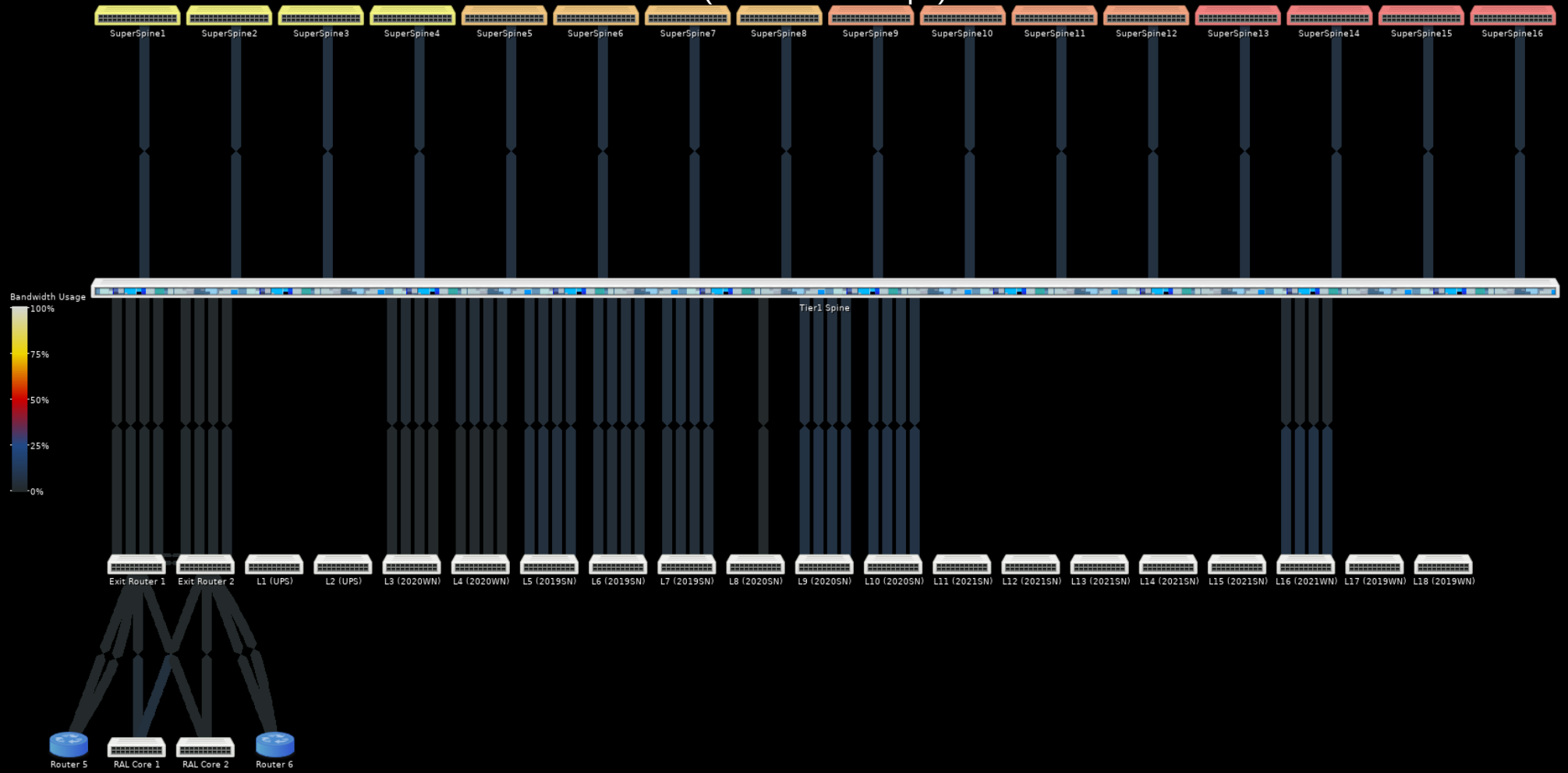
New Network

- Four leaves of workers (almost) +1
  - 104% of 2023 CPU pledge (146% soon)
- Eleven leaves of storage in production +6
  - 33% of Echo capacity
  - FY2022 nodes currently being deployed

New Tier1 Network

New Network — August 2022

(All links 100Gbps)

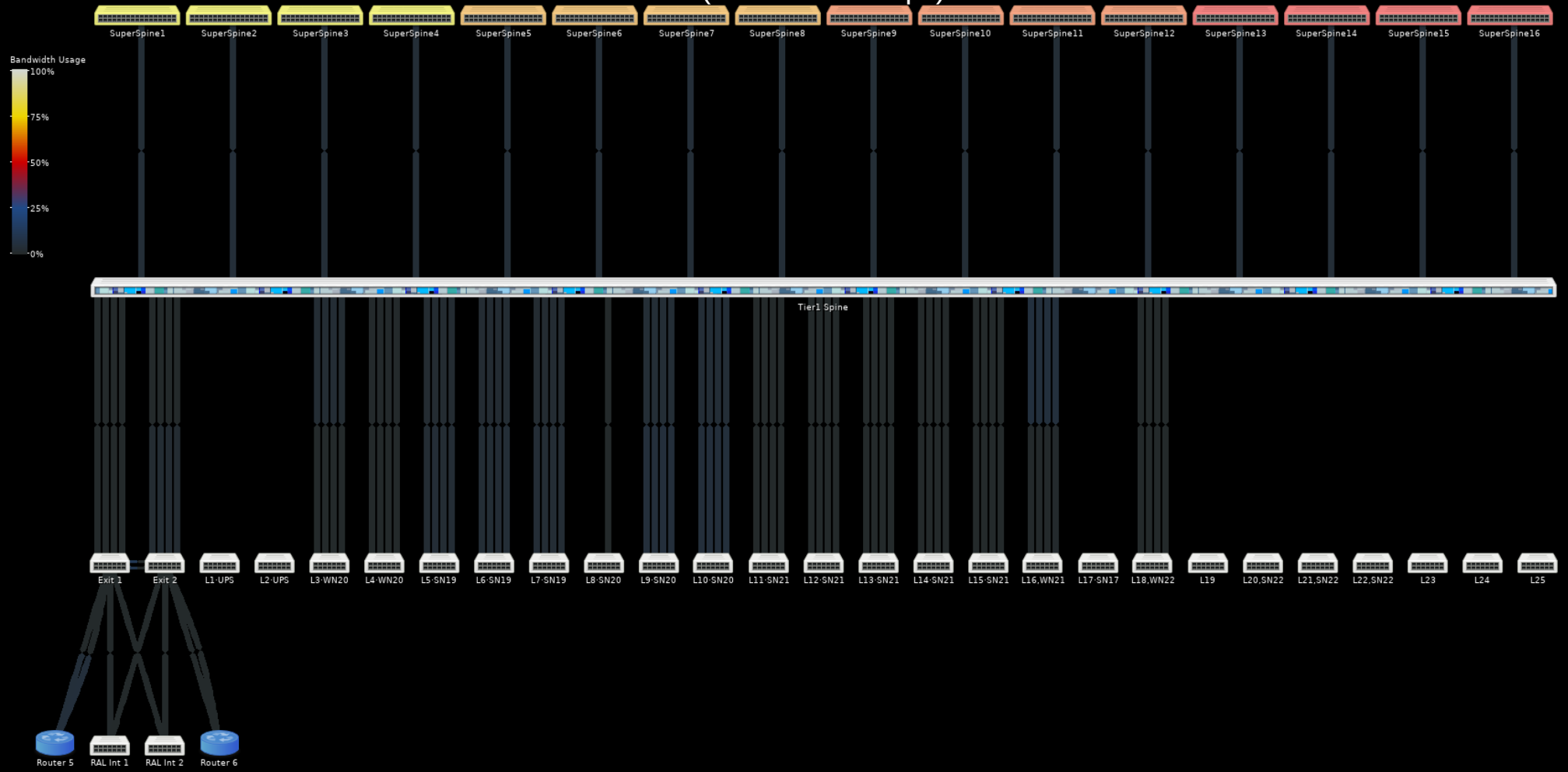


Updated: Aug 26 2022 14:40:10

Tier 1 - New Network

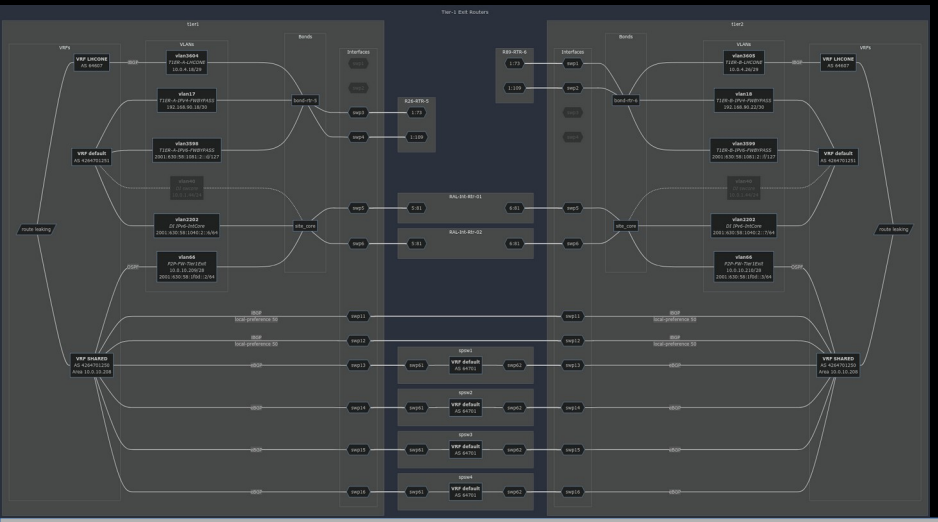
New Network — March 2023

(All links 100Gbps)

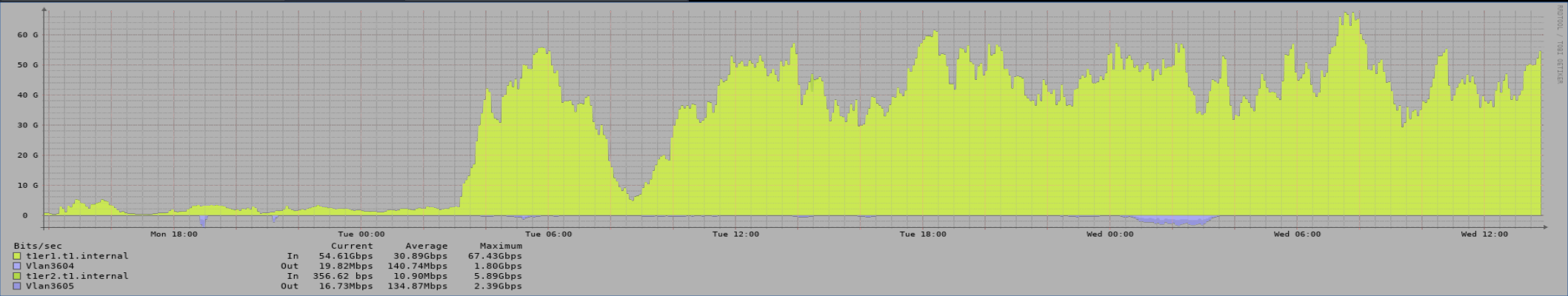


Updated: Mar 24 2023 12:15:44

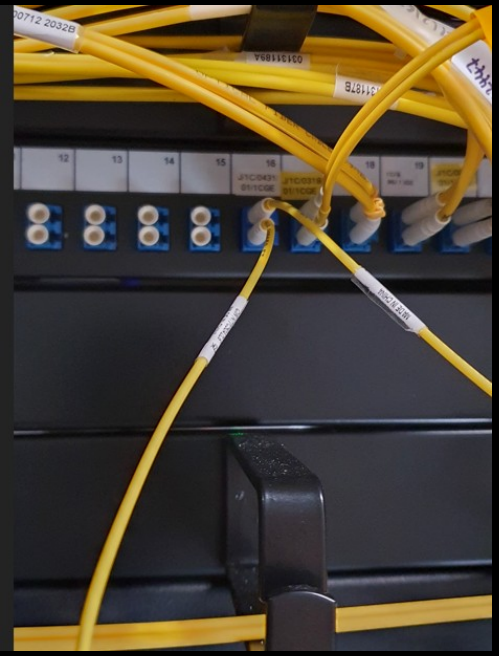
LHCONE



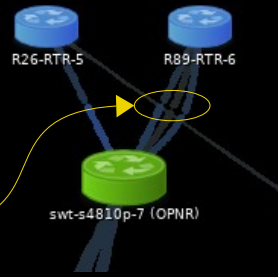
- Peering with both Exit Routers
- Everything on new network now advertised
  - 2020+ Worker nodes
  - 2019+ Storage
    - No echo gateways (yet)

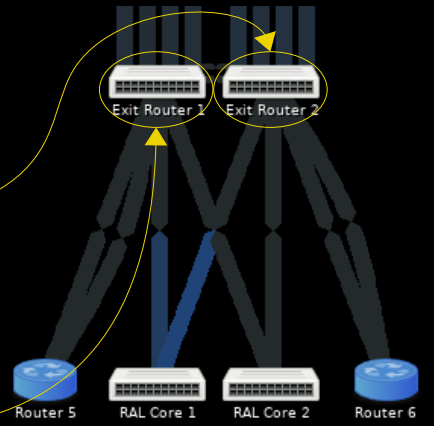


LHCOPN



- Still peering with legacy OPNR via RAL Border Router 6.
- Second 100Gbps link now provisioned!
  - Will connect directly to Exit Router 2.
- Then move existing link to Exit Router 1.





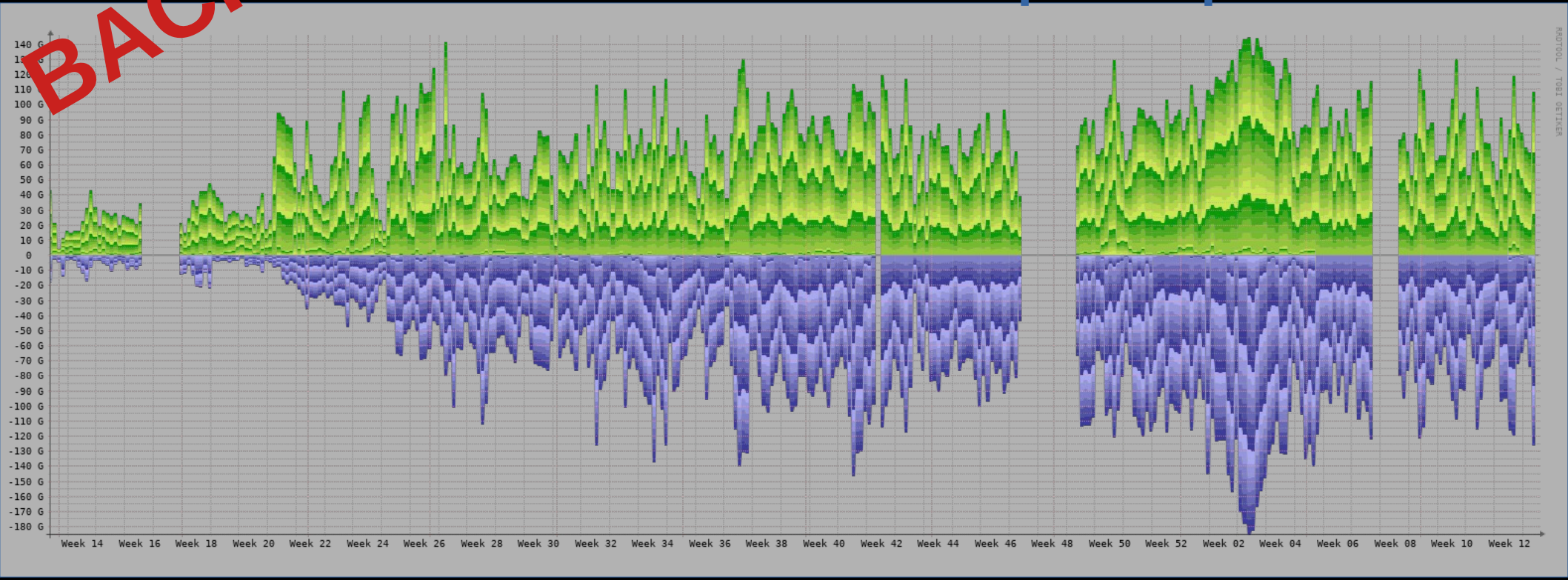
IPv6

- In production on legacy network
  - Many services dual-stack for many years
- Rolled out to all leaves on new network
  - Started thinking about dual-stack worker nodes
  - Decisions to make around Docker integration
- Preparing leaves on CTA network (for Antares)
  - Required by ALICE

Questions?

Last Year – Tier-1/SuperSpine

BACKUP



BACKUP

Last Year – LHCOPN