The RAL Tier-1 Network

James Adams, RAL
GridPP48, Ambleside

2022-09-01

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?

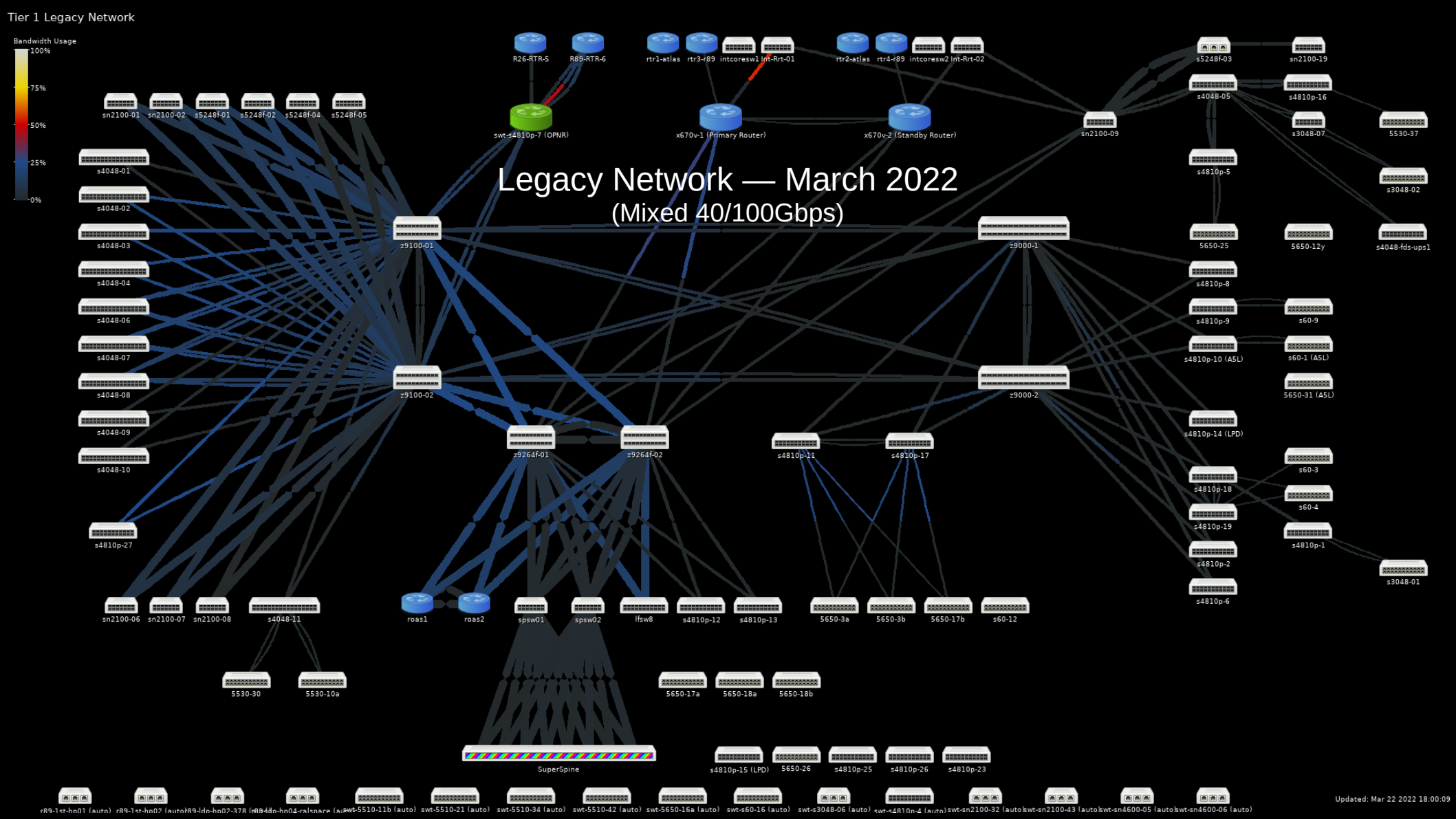
- RAL Macro-Segmentation
- Legacy network clean-up
- More hardware on new network
- More projects on SuperSpine
- Peering with LHCONE

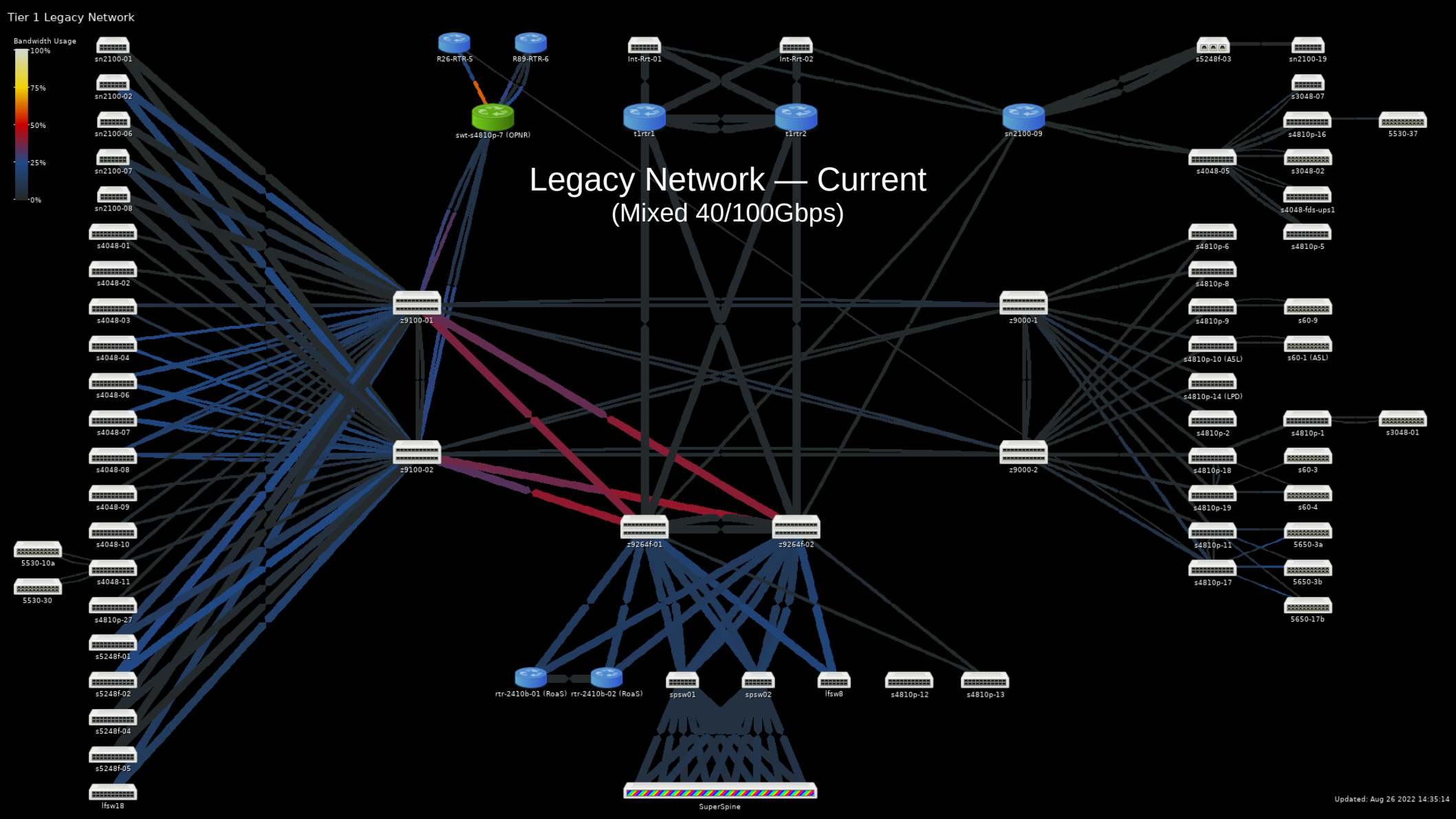
RAL Macro-Segmentation

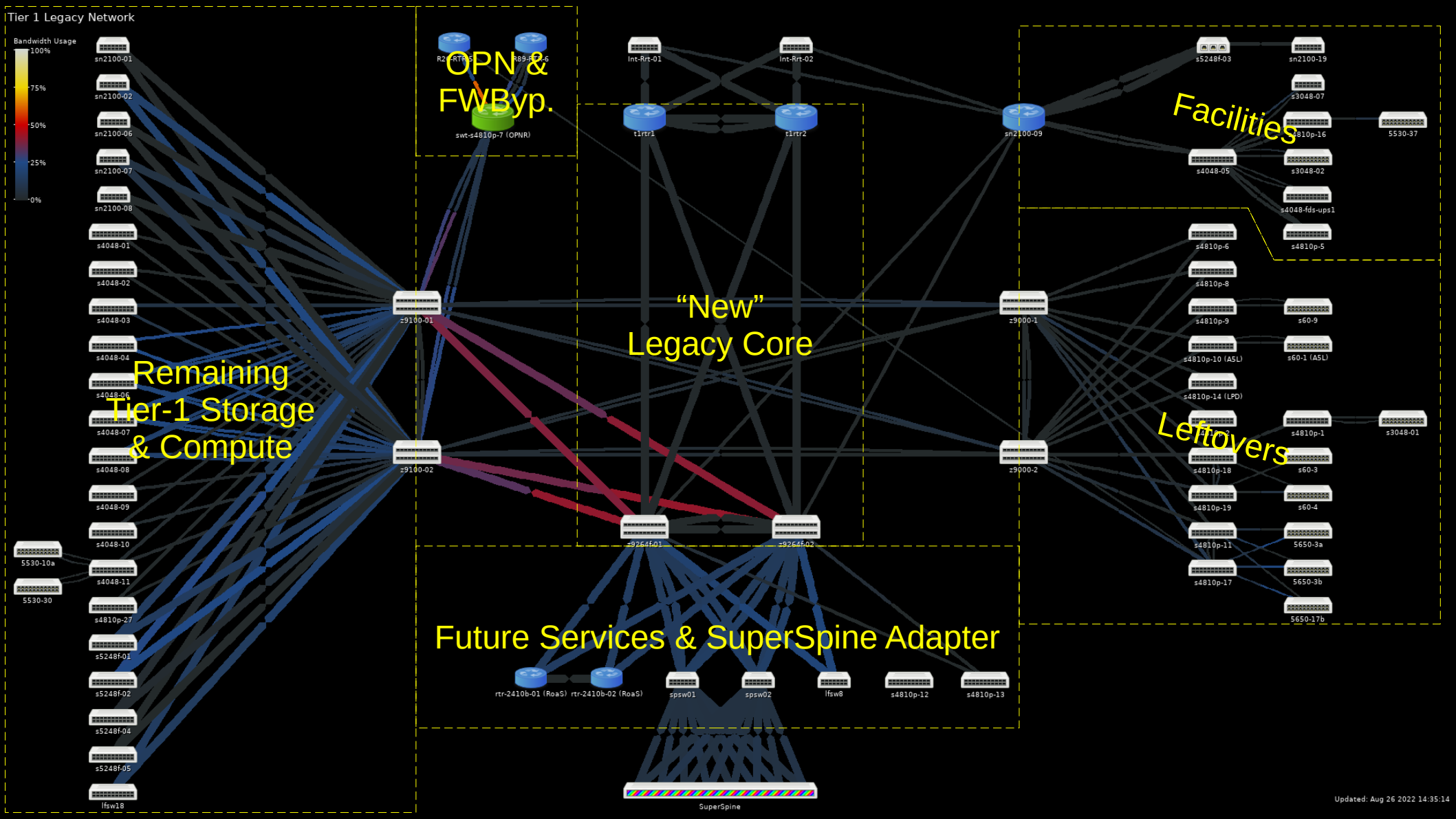
- Essentially puts a firewall into the RAL Core
  - Designed to limit blast-radius of incidents
  - Breaks decades of connectivity assumptions
- Still discovering side effects
  - Active-Active router pairs no longer fail-out correctly
  - Long running SSH sessions dropped
- Somewhat mitigated by SuperSpine

Legacy Network

- Removed lots of old equipment
- Legacy path to LHCOPN increased to 80Gbps
  - Bottleneck to CERN improved
- Replaced Tier-1 Legacy Routers
  - Routes to RAL Campus and Janet
  - Removed 40Gbps bottleneck (now 200Gbps)
  - Will become the routers for facilities network later


Tier 1 Legacy Network
Bandwidth Usage
100%
75%
50%
25%
0%
Legacy Network — March 2022
(Mixed 40/100Gbps)
R26-RTR-5
R89-RTR-6
rtr1-atlas
rtr3-r89
intcoresw1
Int-Rrt-01
rtr2-atlas
rtr4-r89
intcoresw2
Int-Rrt-02
s5248f-03
sn2100-19
s4048-05
s4810p-16
s3048-07
5530-37
s3048-02
s4048-fds-ups1
sn2100-01
sn2100-02
s5248f-01
s5248f-02
s5248f-04
s5248f-05
swt-s4810p-7 (OPNR)
x670v-1 (Primary Router)
x670v-2 (Standby Router)
sn2100-09
s4810p-5
s4048-01
s4048-02
s4048-03
s4048-04
s4048-06
s4048-07
s4048-08
s4048-09
s4048-10
z9100-01
z9100-02
z9000-1
z9000-2
5650-25
5650-12y
s4810p-8
s4810p-9
s4810p-10 (ASL)
s60-9
s60-1 (ASL)
5650-31 (ASL)
s4810p-14 (LPD)
s60-3
s60-4
s4810p-1
s3048-01
s4810p-18
s4810p-19
s4810p-2
s4810p-6
z9264f-01
z9264f-02
s4810p-11
s4810p-17
s4810p-27
sn2100-06
sn2100-07
sn2100-08
s4048-11
roas1
roas2
spsw01
spsw02
lfsw8
s4810p-12
s4810p-13
5650-3a
5650-3b
5650-17b
s60-12
5530-30
5530-10a
5650-17a
5650-18a
5650-18b
SuperSpine
s4810p-15 (LPD)
5650-26
s4810p-25
s4810p-26
s4810p-23
r89-1st-bn01 (auto)
r89-1st-bn02 (auto)
r89-ldp-bn02-378 (auto)
r89-ldp-bn04-ralspace (auto)
swt-5510-11b (auto)
swt-5510-21 (auto)
swt-5510-34 (auto)
swt-5510-42 (auto)
swt-5650-16a (auto)
swt-s60-16 (auto)
swt-s3048-06 (auto)
swt-s4810p-4 (auto)
swt-sn2100-32 (auto)
swt-sn2100-43 (auto)
swt-sn4600-05 (auto)
swt-sn4600-06 (auto)
Updated: Mar 22 2022 18:00:09


Tier 1 Legacy Network
Bandwidth Usage
100%
75%
50%
25%
0%
Legacy Network — Current
(Mixed 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
s4048-05
s3048-02
s4048-fds-ups1
s4810p-5
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
s60-9
s4810p-10 (ASL)
s60-1 (ASL)
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
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%
OPN &
FWByp.
“New”
Legacy Core
Remaining
Tier-1 Storage
& Compute
Facilities
Leftovers
Future Services & SuperSpine Adapter
sn2100-01
sn2100-02
sn2100-06
sn2100-07
sn2100-08
s4048-01
s4048-02
s4048-03
s4048-04
s4048-06
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
swt-s4810p-7 (OPNR)
int-rrt-01
int-rrt-02
t1rtr1
t1rtr2
z9100-01
z9100-02
z9264f-01
z9264f-02
sn2100-09
z9000-1
z9000-2
s5248f-03
sn2100-19
s3048-07
s4810p-16
5530-37
s4048-05
s3048-02
s4048-fds-ups1
s4810p-5
s4810p-6
s4810p-8
s4810p-9
s60-9
s4810p-10 (ASL)
s60-1 (ASL)
s4810p-14 (LPD)
s4810p-1
s3048-01
s4810p-18
s60-3
s4810p-19
s60-4
s4810p-11
5650-3a
s4810p-17
5650-3b
5650-17b
rtr-2410b-01 (RoaS)
rtr-2410b-02 (RoaS)
spsw01
spsw02
lfsw8
s4810p-12
s4810p-13
SuperSpine
Updated: Aug 26 2022 14:35:14

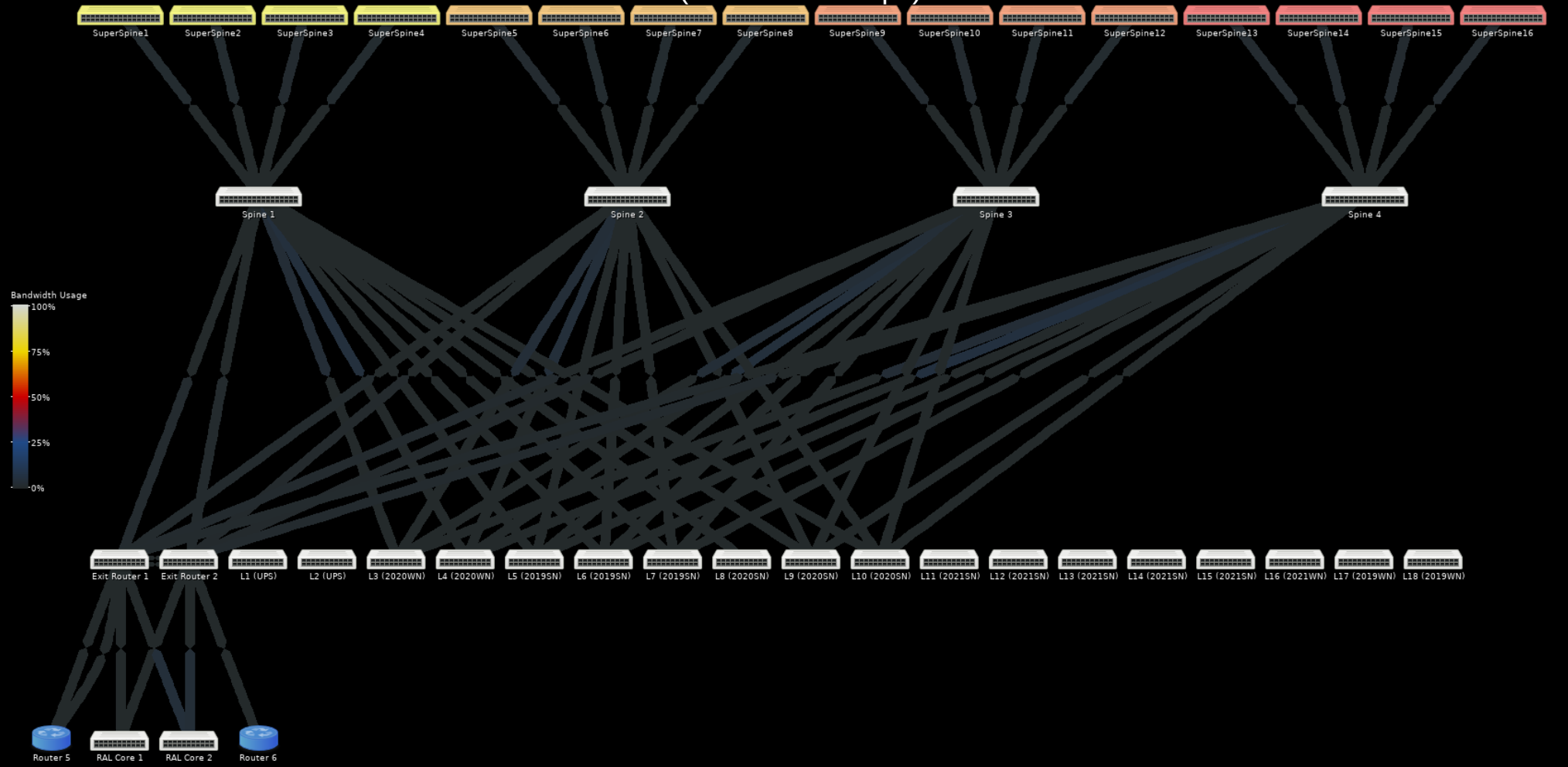
New Network

- Three leaves of workers in production (+1)
  - 85% of pledge
- Five leaves of storage in production (+5)
  - 33% of Echo capacity
  - Data being migrated carefully
- Joined LHCONE
  - Initially with a very small prefix for PerfSonar

New Tier1 Network

New Network — March 2022

(All links 100Gbps)

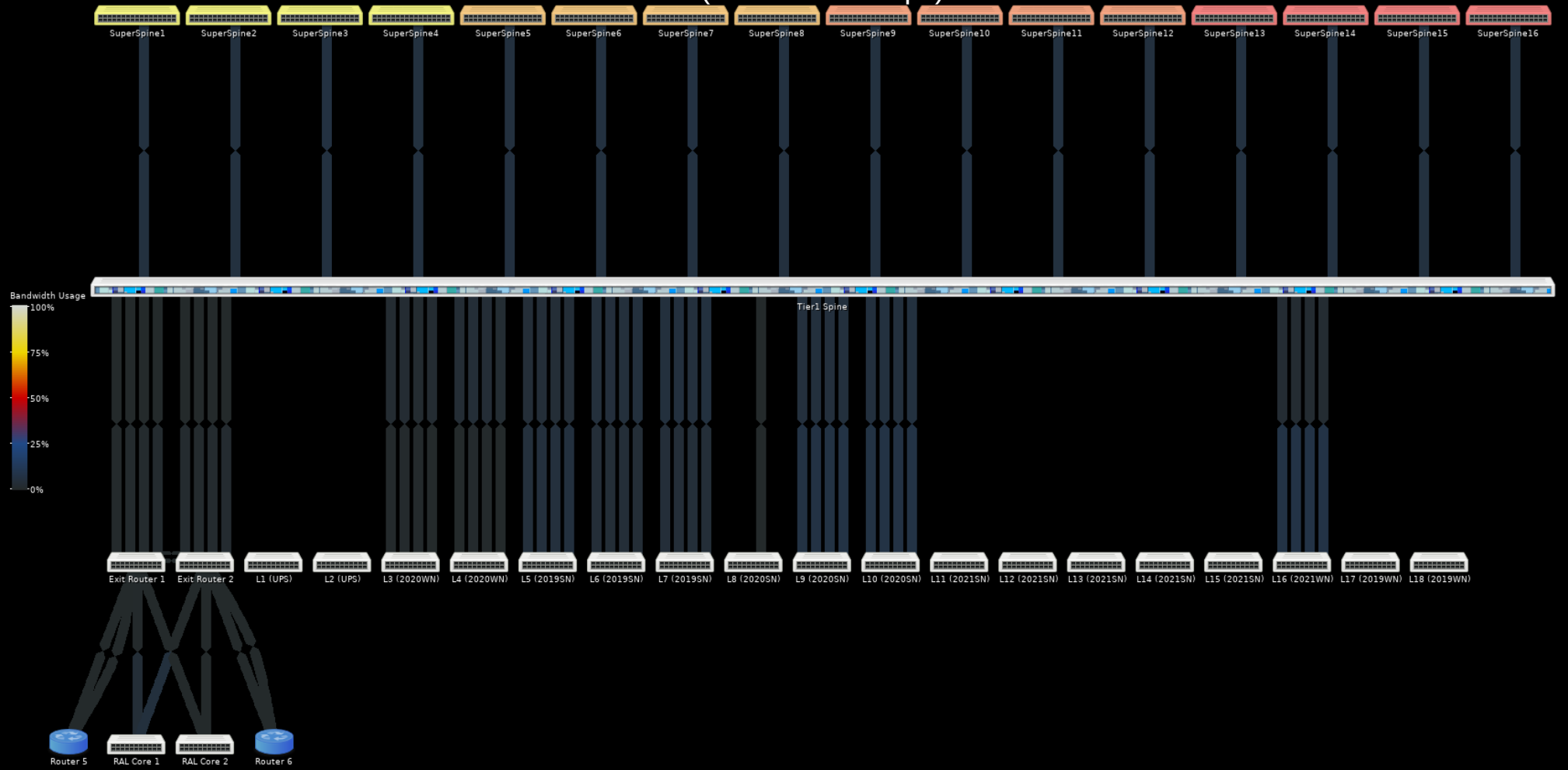


Updated: Mar 22 2022 18:00:10

New Tier1 Network

New Network — Current

(All links 100Gbps)



Updated: Aug 26 2022 14:40:10

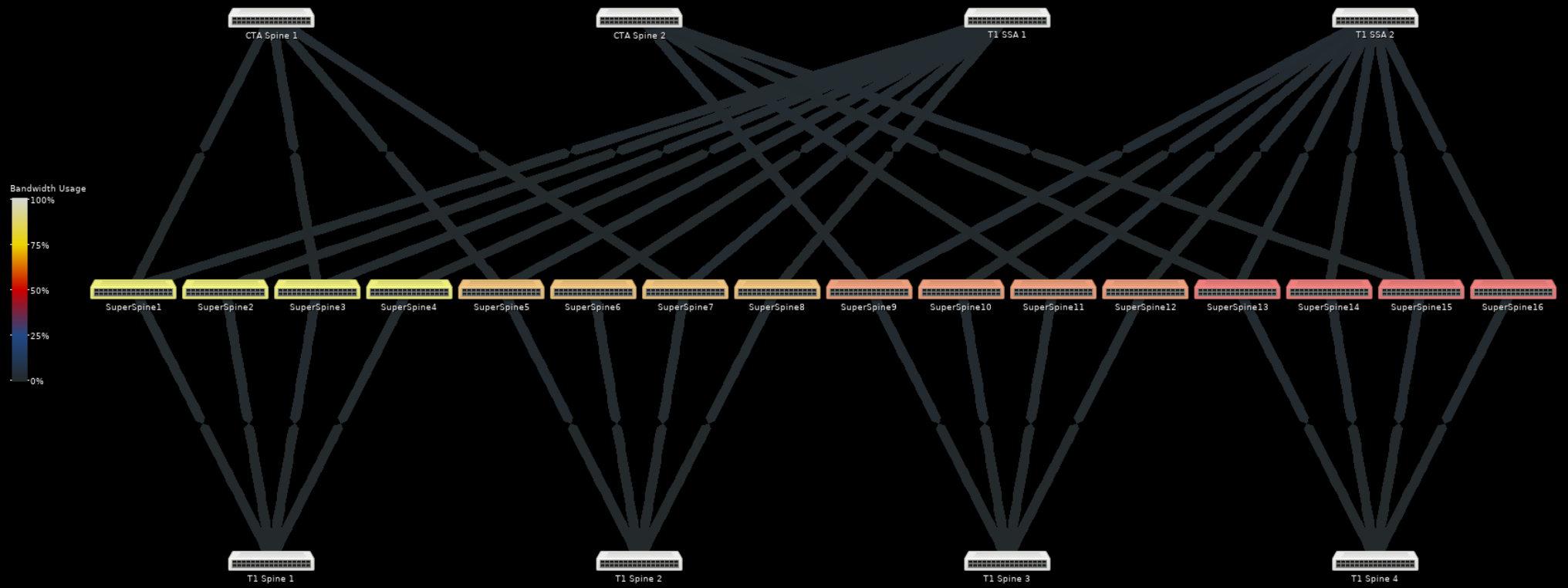
SuperSpine

- Joins SCD networks at up to 1.6Tbps (5-Stage Clos)
  - JASMIN
  - DAFNI
  - Tier1 legacy and new networks
  - Antares (CTA)
- SCD Cloud (R89 Pod) recently joined
  - R26 Pod joining soon

Superspine

SuperSpine — March 2022

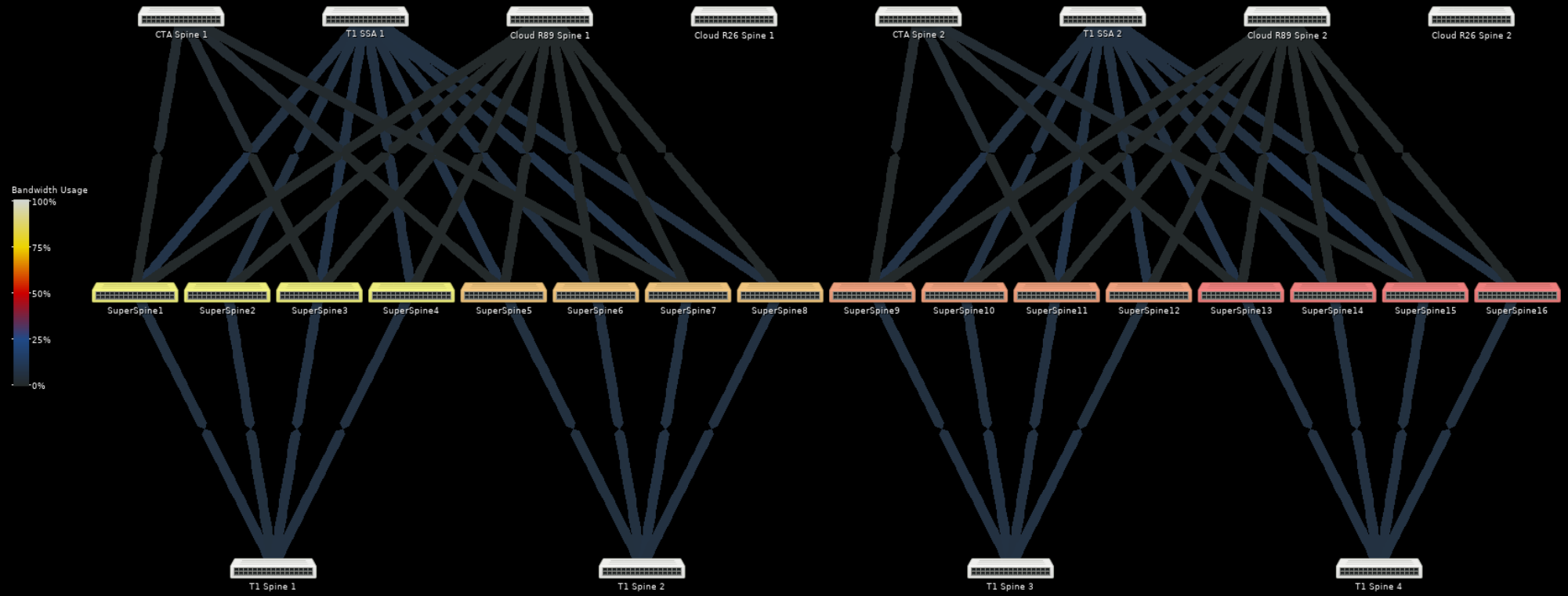
(All links 100Gbps, JASMIN & DAFNI not shown)



Updated: Mar 22 2022 18:00:11

SuperSpine — Current

(All links 100Gbps, JASMIN & DAFNI not shown)



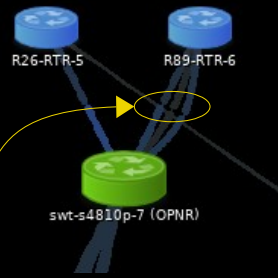
Updated: Aug 26 2022 14:35:11

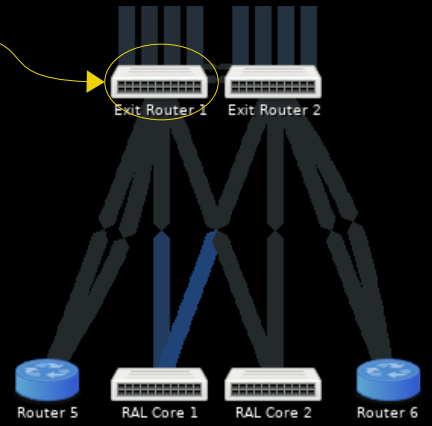
IPv6

- Needed more IPv6 space for new network
  - Legacy network has two /64
  - New network needs a /64 per leaf switch (currently twenty)
  - STFC addressing plan too restrictive (8-bits per department)
- A new /48 has been allocated by JISC for the Tier-1
  - Discussion happening internally to agree routing over SuperSpine
- New network will become “IPv6 first”

LHCOPN

- Currently peers with legacy OPNR via RAL Border Router 6.
- Move directly to Tier-1 Exit Router 1
  - Legacy network access over SuperSpine
  - Reverses existing traffic flow
- Pending IPv6 routing review
  - 60% OPN traffic is IPv6





LHCONE

- Currently peers with Tier-1 Exit Routers
- No services other than PerfSonar advertised yet
  - Reminder: LHCONE bypasses site firewalls
- Worker nodes will join after Condor 9 upgrade
  - Still need to roll out host firewalls
- Pending IPv6 routing review
  - Legacy network will access over SuperSpine

PerfSonar

- New nodes deployed on Tier-1 and Antares pods
- Problem with routes for the additional interfaces
  - Not being installed in the switching silicon
  - Being discussed with vendor
  - Alternative plan to move back to traditional method
    - Discouraged by PerfSonar developers
    - Remains the fall-back plan if not resolved by October

Questions?

BACKUP

STFC Addressing Scheme

Each project allocated one or more IPv6 /64

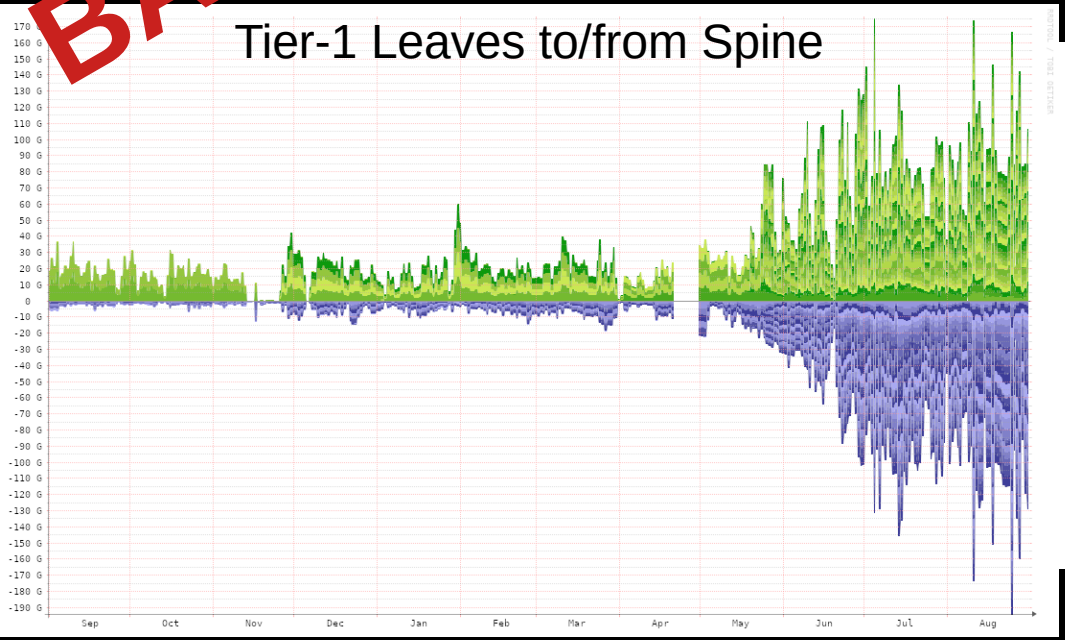
- 16 bits available to describe subnet

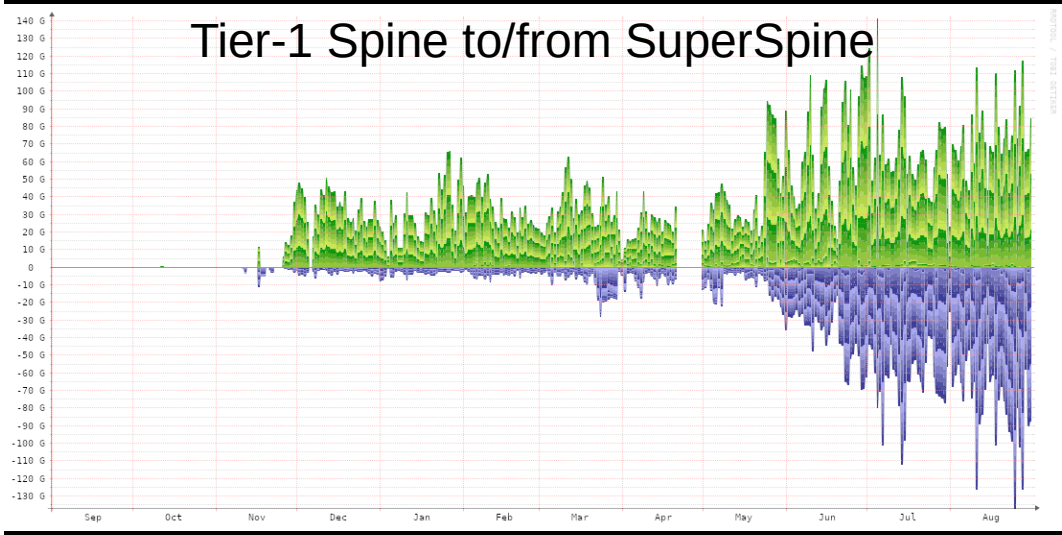
2001	:	0630	:	0058	:	abcd	:	0000	:	0000	:	0000	:	0000
NETWORK								HOST						
JANET	:	RAL	:	abcd	:	0000	:	0000	:	0000	:	0000		

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

BACKUP

One Year of Traffic





BACKUP

One Year of Traffic