The RAL Tier-1 Network

James Adams, RAL
GridPP51, Sheffield

2024-03-27

Overview

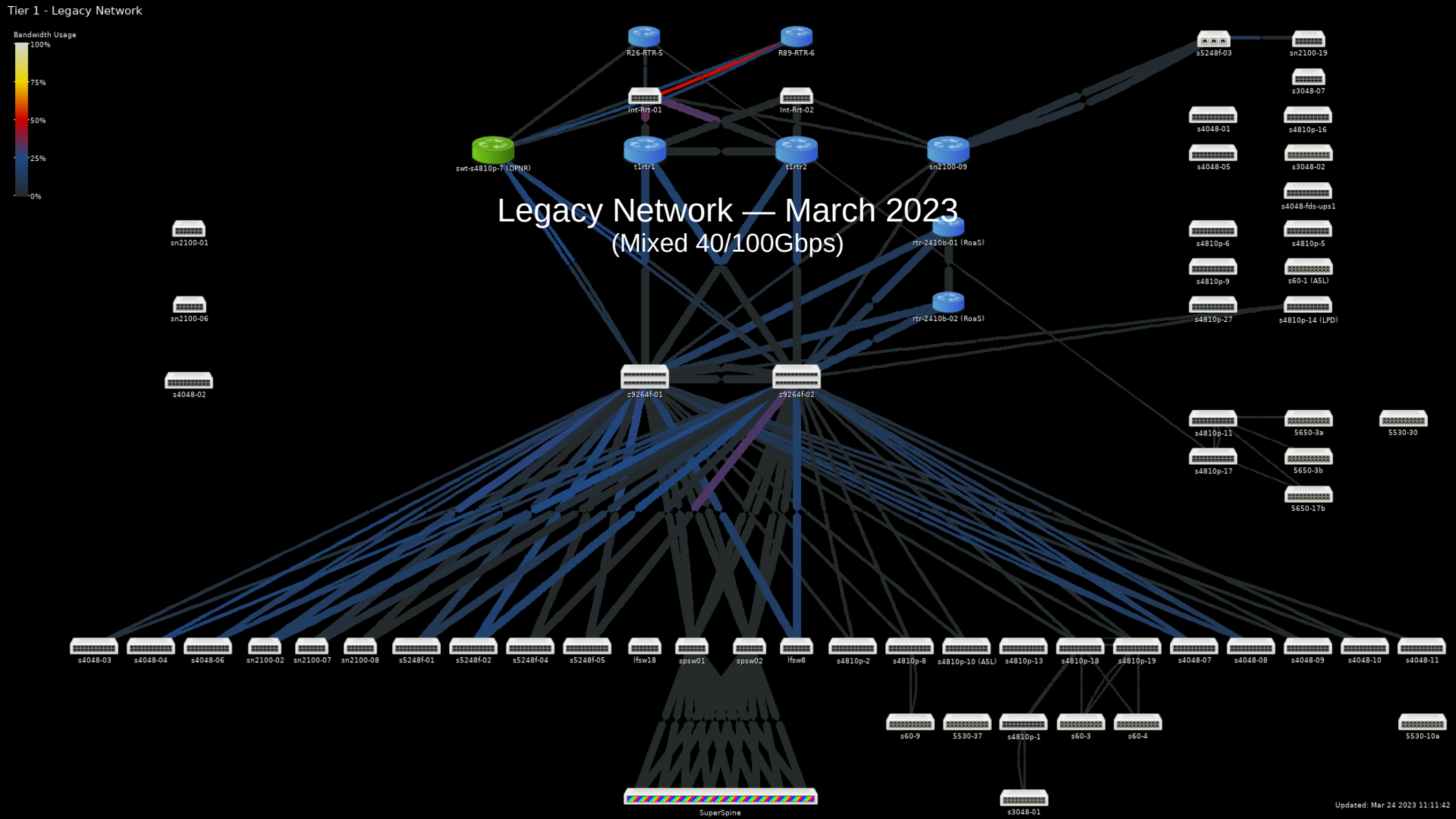
- Background
- What’s Changed?
- DC24
- What’s Next?

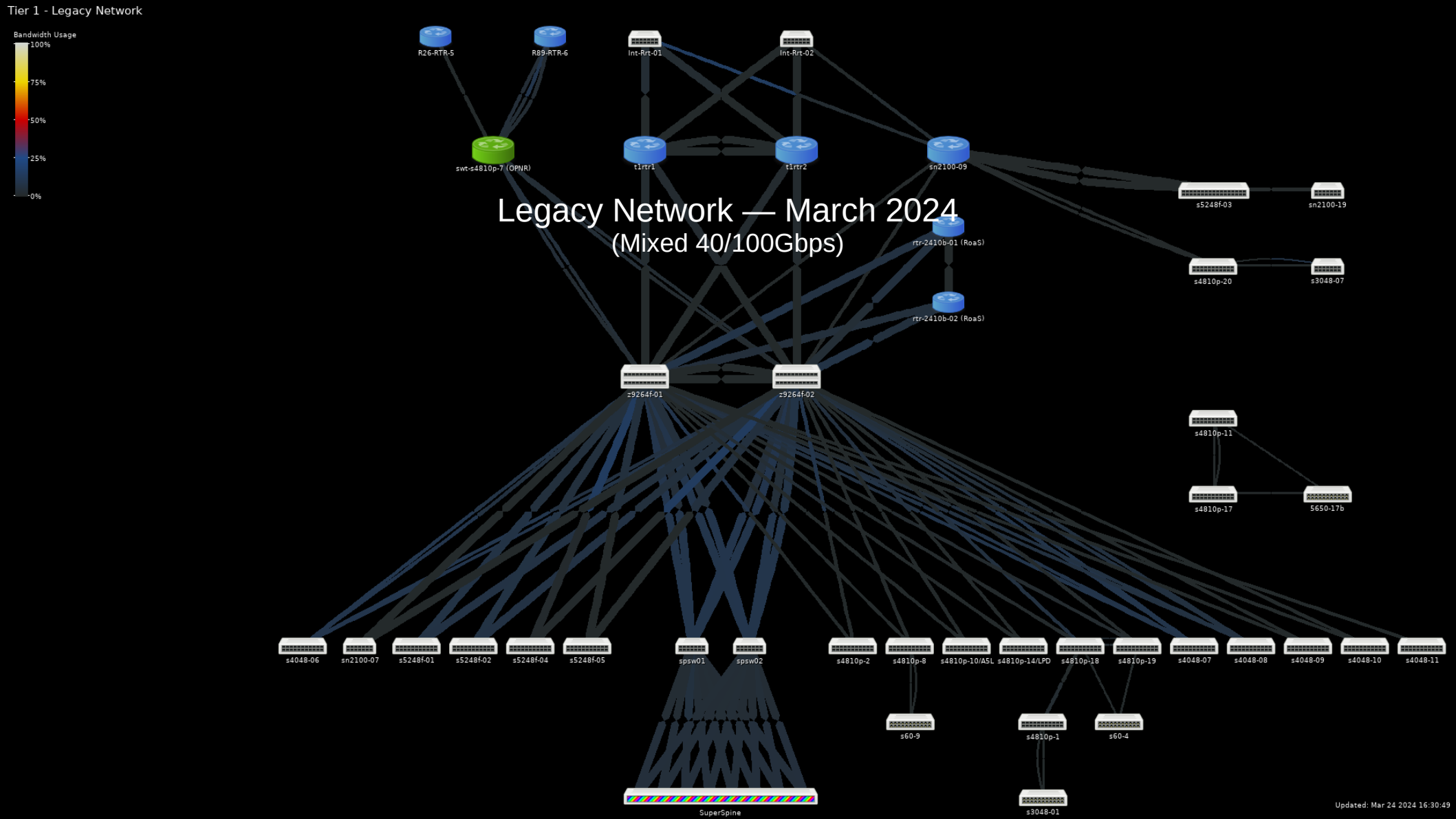
Background

- Built out a replacement network for the Tier-1 alongside the “legacy” network
  - Fully-routed eBGP ECMP architecture
  - nVidia/Mellanox switches running Cumulus Linux
  - Joined to legacy network and other SCD projects by SuperSpine
- First went live with worker nodes December 2021
  - Now hosting majority of Tier-1 Storage and Compute

What’s Changed?

- Less hardware on legacy network
- More hardware on ~~new~~ main network
- Second LHCOPN link in production
  - August 2023
- IPv6 rollout
  - Temporary split of Exit Routers


Tier 1 - Legacy Network
Bandwidth Usage
100%
75%
50%
25%
0%
Legacy Network — March 2023
(Mixed 40/100Gbps)
R26-RTR-5
R89-RTR-6
lnt-Rrt-01
lnt-Rrt-02
swt-s4810p-7 (OPNR)
tlrtr1
tlrtr2
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
sn2100-06
rtr-2410b-02 (RoaS)
s4810p-9
s60-1 (ASL)
s4048-02
s4810p-27
s4810p-14 (LPD)
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


Tier 1 - Legacy Network
Bandwidth Usage
100%
75%
50%
25%
0%
R26-RTR-5
R89-RTR-6
Int-Rrt-01
Int-Rrt-02
swt-s4810p-7 (OPNR)
t1rtr1
t1rtr2
sn2100-09
s5248f-03
sn2100-19
Legacy Network — March 2024
(Mixed 40/100Gbps)
rtr-2410b-01 (RoaS)
s4810p-20
s3048-07
rtr-2410b-02 (RoaS)
z9264f-01
z9264f-02
s4810p-11
s4810p-17
5650-17b
s4048-06
sn2100-07
s5248f-01
s5248f-02
s5248f-04
s5248f-05
spsw01
spsw02
s4810p-2
s4810p-8
s4810p-10/A5L
s4810p-14/LPD
s4810p-18
s4810p-19
s4048-07
s4048-08
s4048-09
s4048-10
s4048-11
s60-9
s4810p-1
s60-4
SuperSpine
s3048-01
Updated: Mar 24 2024 16:30:49

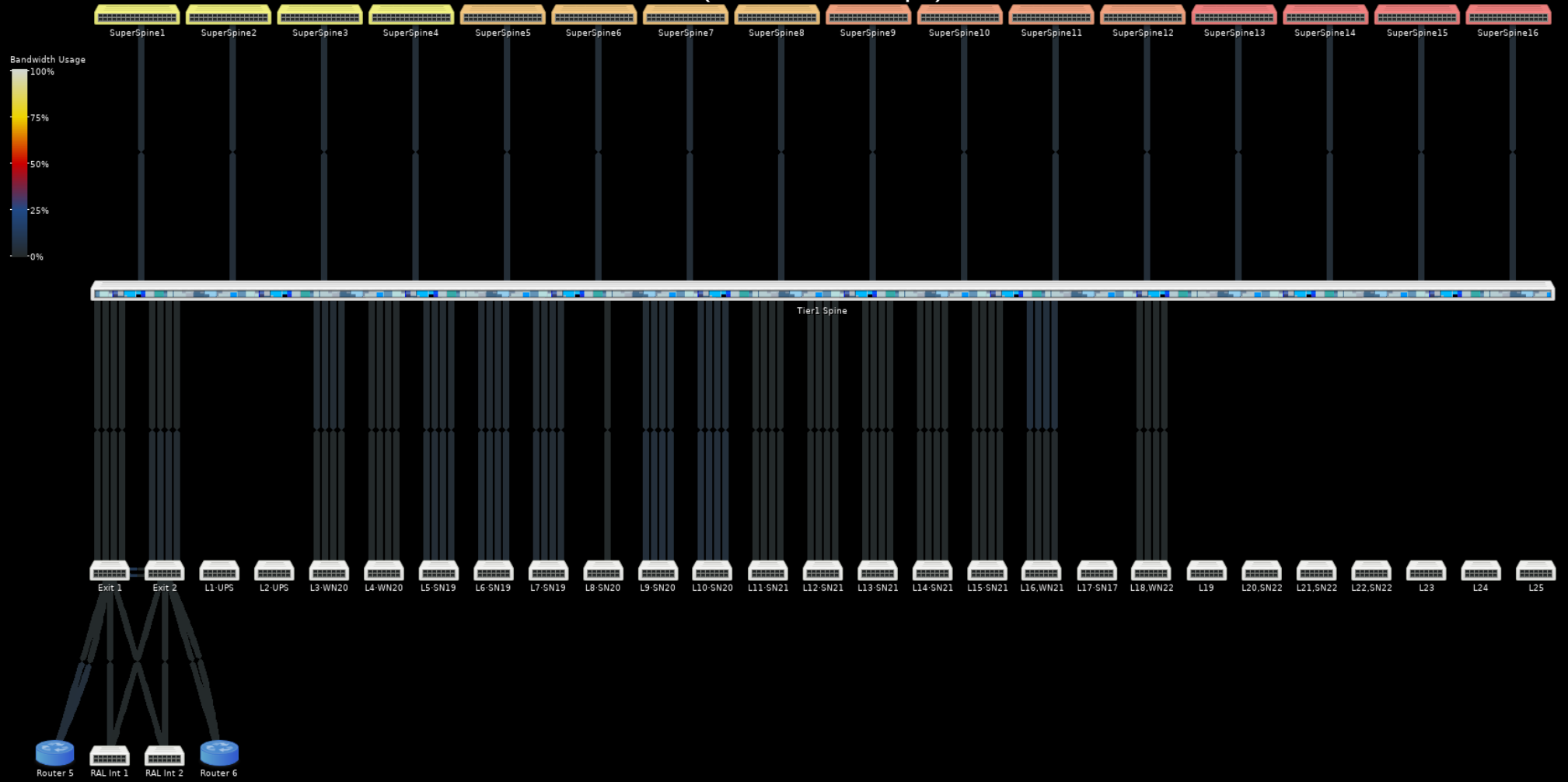
~~New~~ Main Network

- Four leaves of workers in production (±0)
  - 85% of pledge
- 21 leaves of storage in production (+10)
  - 90% of Echo capacity
  - Almost all Echo gateways

Tier 1 - New Network

Main Network — March 2023

(All links 100Gbps)

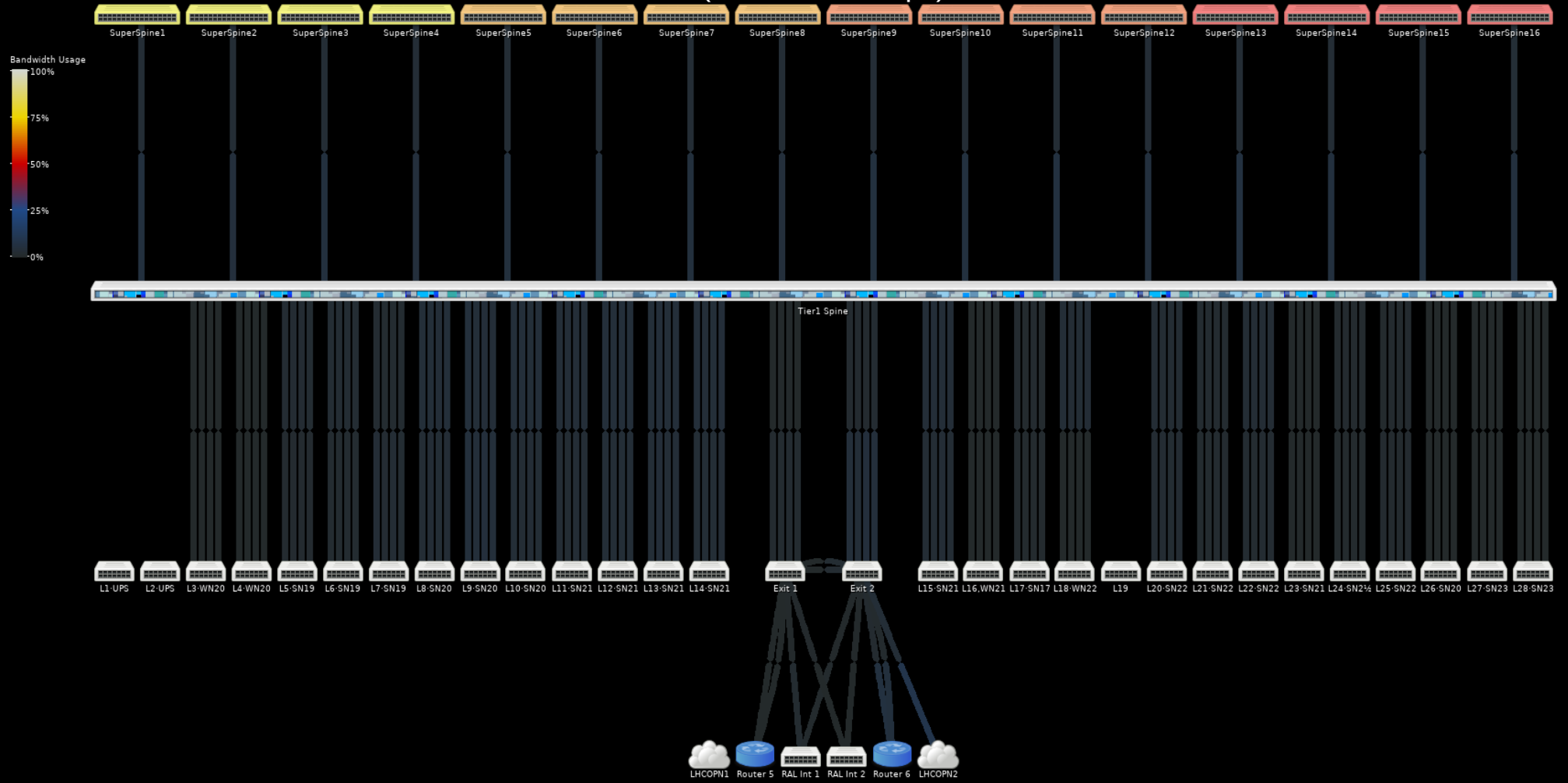


Updated: Mar 24 2023 12:15:44

Tier 1 - New Network

Main Network — March 2024

(All links 100Gbps)

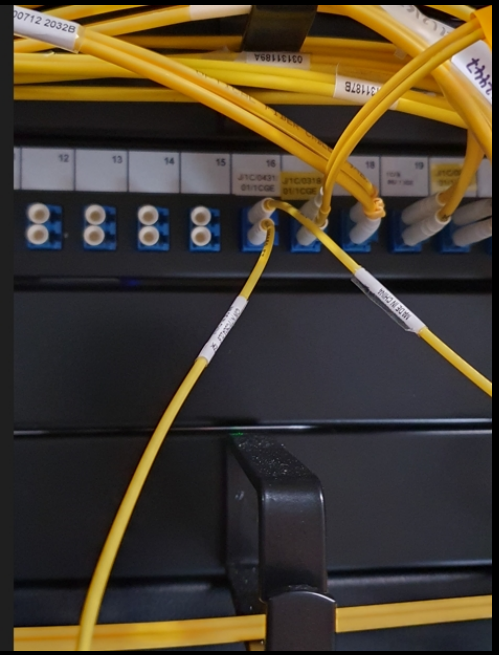


Updated: Mar 24 2024 23:05:27

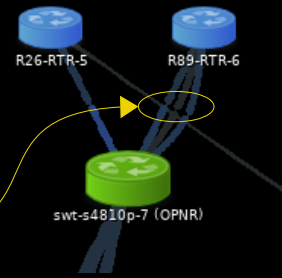
IPv6

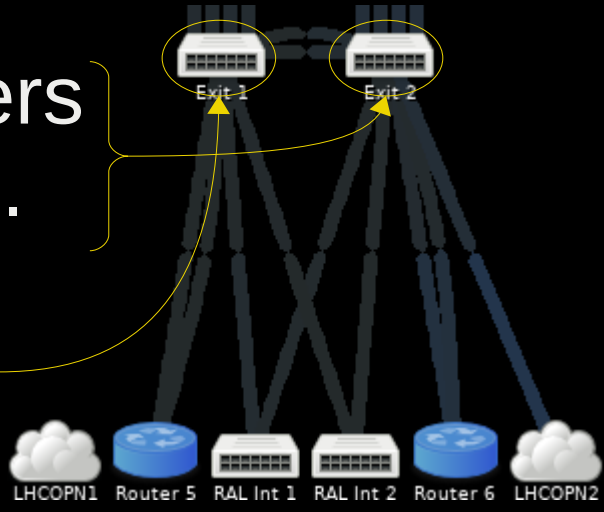
- New allocation `2001:630:54::/48` rolled out
  - `2001:630:54:0000::/52` → Tier-1 in R89
  - `2001:630:54:2000::/52` → CTA/Antares in R89
- Out-of-date FRR package in Cumulus Linux
  - Missing VRF awareness for OSPFv3
  - Split functionality of Exit Routers (for now)
    - #1 dedicated to routing to RAL site without VRFs (OSPF)
    - #2 dedicated to LHCONE & LHCOPN (BGP)

LHCOPN



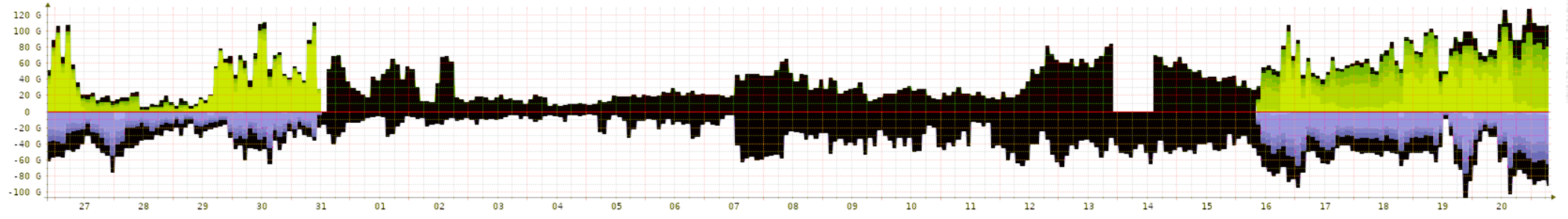
- Original link peers with legacy OPNR via Border Router 6
  - Limited to 80Gbps
- Second 100Gbps link peers directly with Exit Router 2.
- Will move original link to Exit Router 1 this year





LHCOPN

- Undersea cable carrying both links damaged
  - OPN went down 31st January
  - Break found and cable lifted 12th February
  - First repair (14th February) unsuccessful
  - Final repaired completed 16th January
- LHCONE picked up *most* of the traffic



DC24

- Extended failure of LHCOPN was definitely educational
  - Some sites on LHCOPN had different routes on LHCONE
    - Largely resolved, but needs ongoing monitoring
  - Conflicts emerged between the split Exit Routers
    - Asymmetric routing depending on source IP
- Explicit Congestion Notification
  - Unclear whether a bug in implementation or genuine congestion at NIKHEF
- Successfully moved the bottleneck to echo gateways
  - Moved on to tuning of network stack on hosts
  - Lots of hands on work from many people

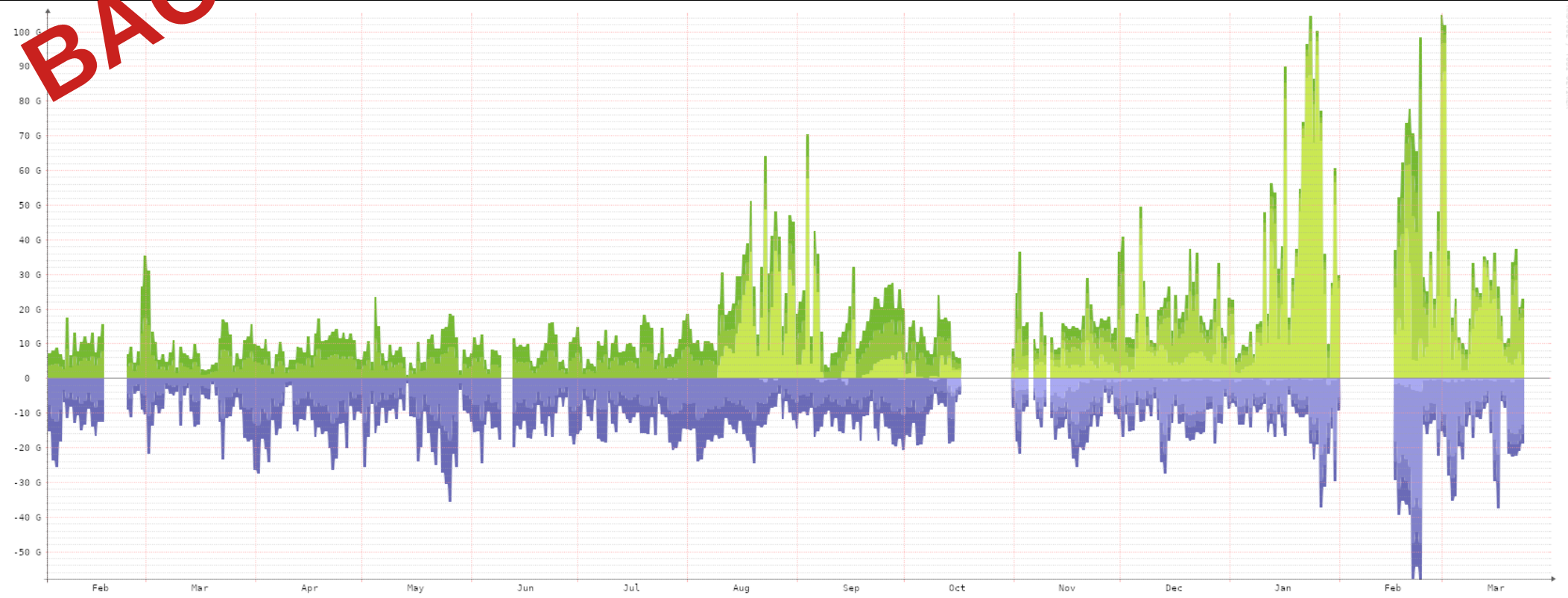
What’s Next?

- Move or decommission remaining legacy network hosts
  - Compute and Storage
- Migrate default route to SCD Exit Pod
  - Shutdown local RAL site links
- Move original LHCOPN link to Exit Router 1
  - After all Echo Gateways are moved
- Retire legacy network from Tier-1 use

Questions?

BACKUP

One Year of LHCOPN



Bits/sec		Current	Average	Maximum
tier2.tl.internal	In	4.12Gbps	9.22Gbps	93.20Gbps
Vlan3561	Out	505.48Mbps	601.25Mbps	5.81Gbps
tier2.tl.internal	In	13.45Gbps	4.76Gbps	57.29Gbps
Vlan3562	Out	13.00Gbps	3.78Gbps	33.82Gbps
swt-s4810p-7.pscs.internal	In	2.66Gbps	4.22Gbps	17.74Gbps
fortygige 0/52	Out	2.51Gbps	5.14Gbps	17.98Gbps
swt-s4810p-7.pscs.internal	In	2.67Gbps	4.22Gbps	17.78Gbps
fortygige 1/52	Out	2.58Gbps	5.17Gbps	17.50Gbps

BACKUP One Year of LHCONE

Bits/sec		Current	Average	Maximum
tler1.tl.internal	In	25.81 bps	9.56Gbps	84.88Gbps
Vlan3604	Out	27.35 bps	75.11Mbps	1.06Gbps
tler2.tl.internal	In	16.02Gbps	9.77Gbps	64.76Gbps
Vlan3605	Out	5.49Gbps	10.90Gbps	50.31Gbps