Science and Technology Facilities Council

Scientific Computing

The UK RAL Data Centre Network

Jonathan Churchill

SCD Network Architect

Scientific Computing Department

STFC – Rutherford Appleton Labs. UK

Agenda

1 Network Expansion Experiences

JASMIN rapid expansion history

2 Two layer L3 CLOS to Five layer

What to do when a 2 layer CLOS cant expand.

3 Generalising to a Data Centre Network

Solving the interoperation bandwidth problem

4 Flexibility

In toplogy, requirements, operating systems, routing protocols.

Image © STFC Alan Ford

Science and Technology Facilities Council

Scientific Computing

SCD Networking Context at STFC

- Tier1
  (52k Cores, ~800 nodes, 45PB, 400+ VMs)
- JASMIN
  (20k Core HPC, 300 Nodes, 37PB Storage, 3 Clouds, 1,100+ VMs …..)
- STFC Cloud (~3000 VMs, 700 Nodes)
- Facilities (4 10k slot tape robots, 150+PB, 40 Tape servers)
- SCARF (20k Core HPC, 4PB Storage)
- DAFNI (1k Core Compute, 1PB Storage)
- IRIS etc

- Separate projects and teams
- Different functional requirements
- Different network requirements
- Different network topologies
- Different histories
- …No network team

Science and Technology Facilities Council

Scientific Computing

JASMIN 2011 .. 2014

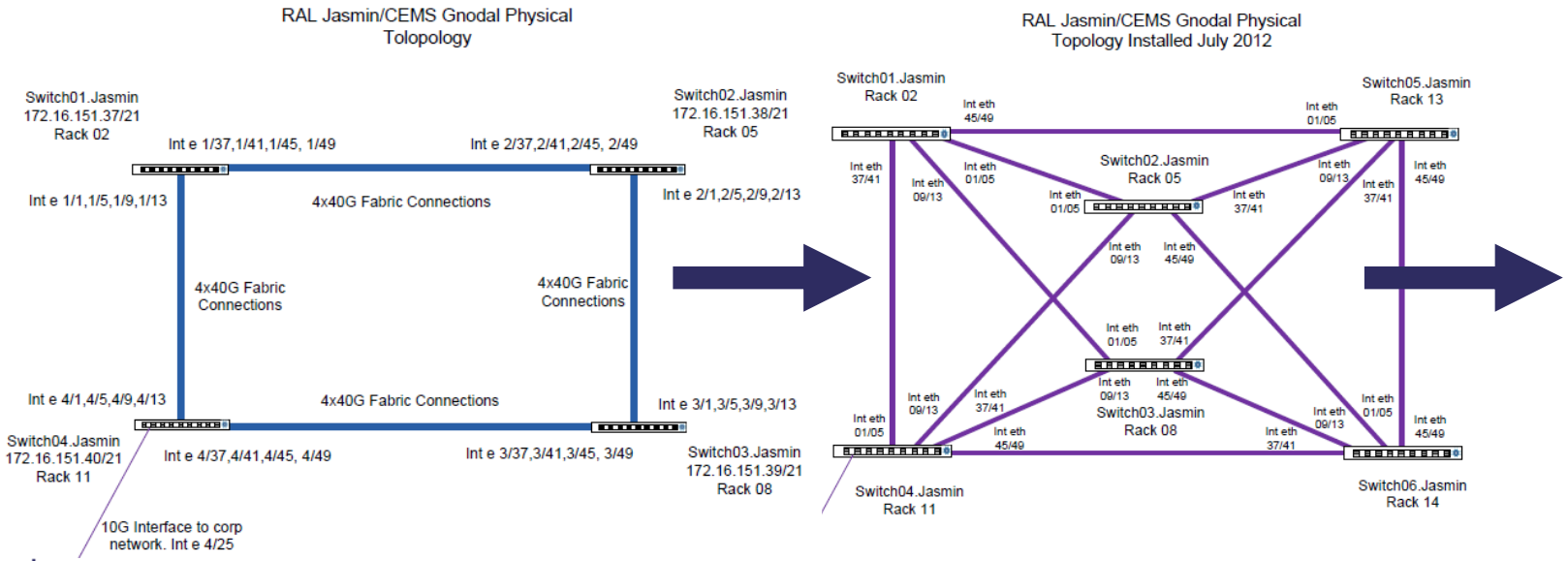
4.5PB Panasas + 240 Cores

/21

/21

Flat Overlaid L2

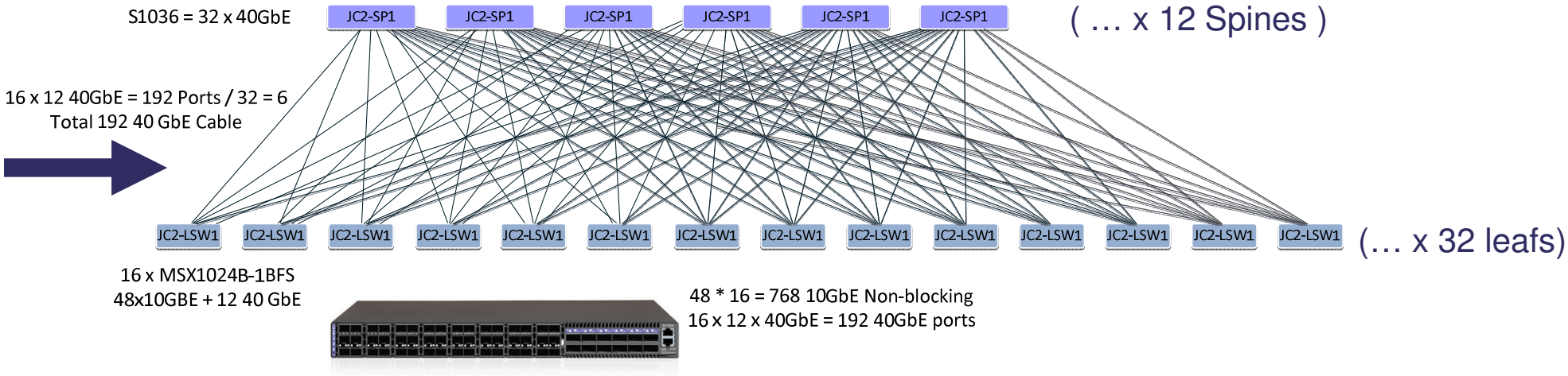
160->240 Ports @ 10Gb



Scie
Technology
Facilities Council

Scientific Computing

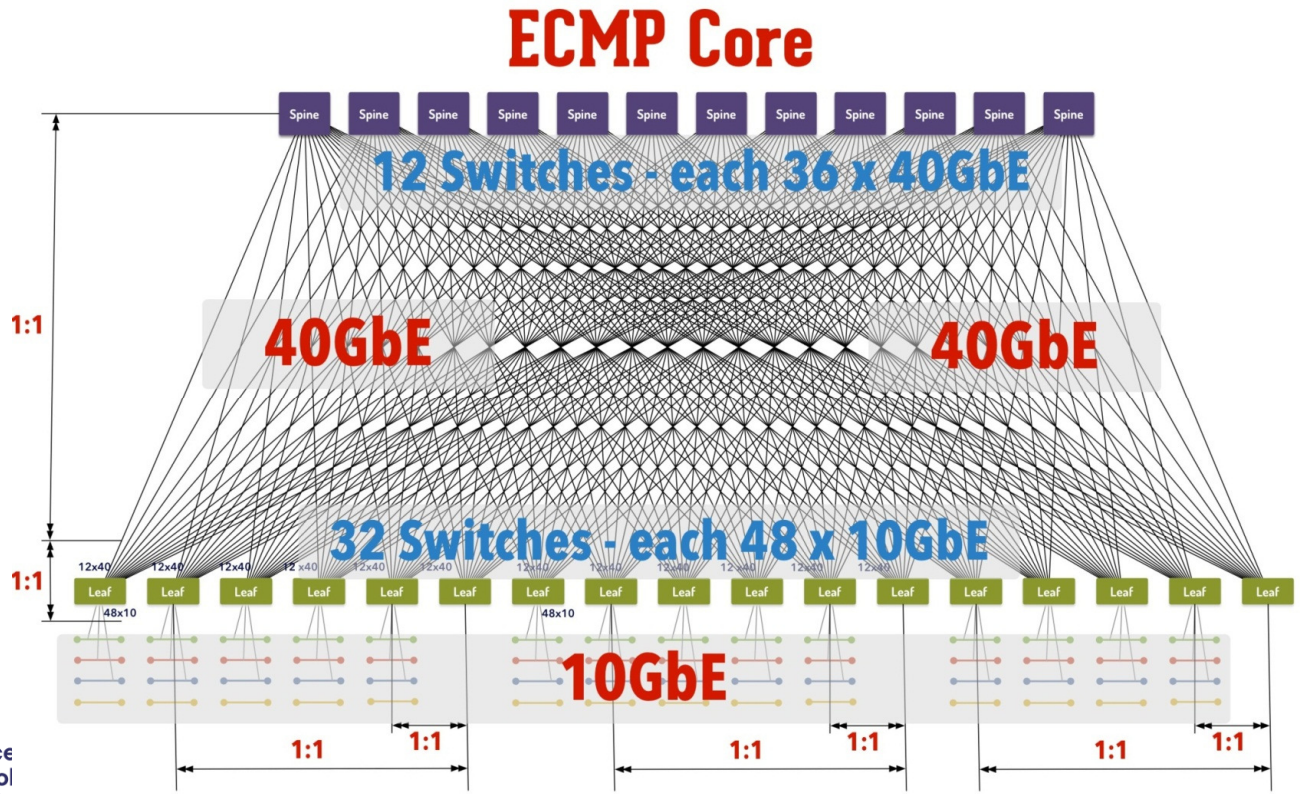
1,104 x 10GbE Ports CLOS L3 ECMP OSPF

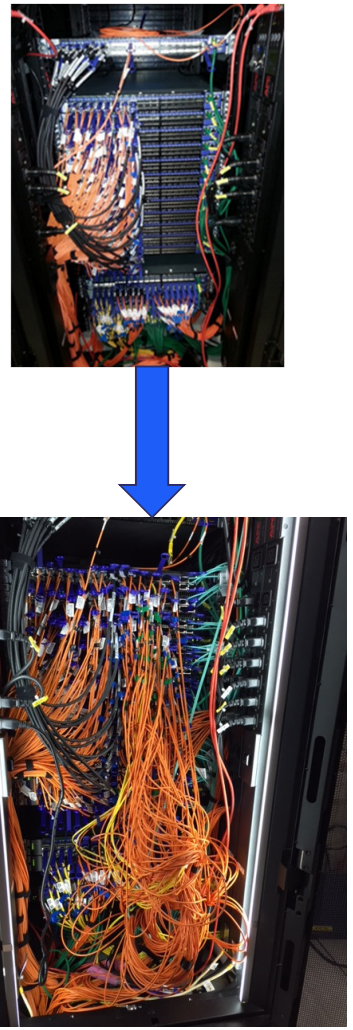


- 768 Ports max. (no expansion) … so 12 spines
- Max 36 leaf switches :1,728 Ports @ 10GbE
- Non-Blocking. Zero Contention (48x10Gb = 12x 40Gb uplinks)
- Low Latency (250nS L3 / per switch/router). 7-10uS MPI
- Cheap(ish) < £400k(inc Cables) vs > £1.5M chassis spine

JASMIN 2/3 Expansion (2014)

- 30 of 36 Leaf Ports in use per Spine.
- Only ≤ 6x 40G per spine for uplinks
- JASMIN4: 22-25 Racks, ≥ 18 Leaf Switches

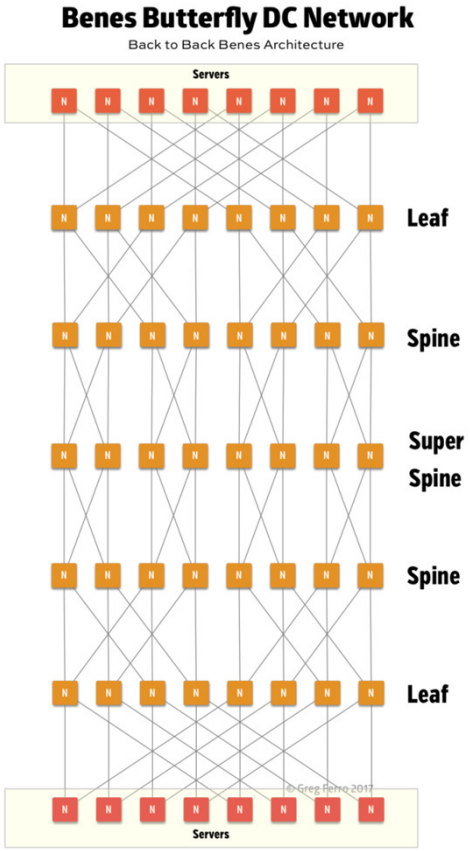




Science Technol Facilities Council

Scientific Computing

A Data Centre Network

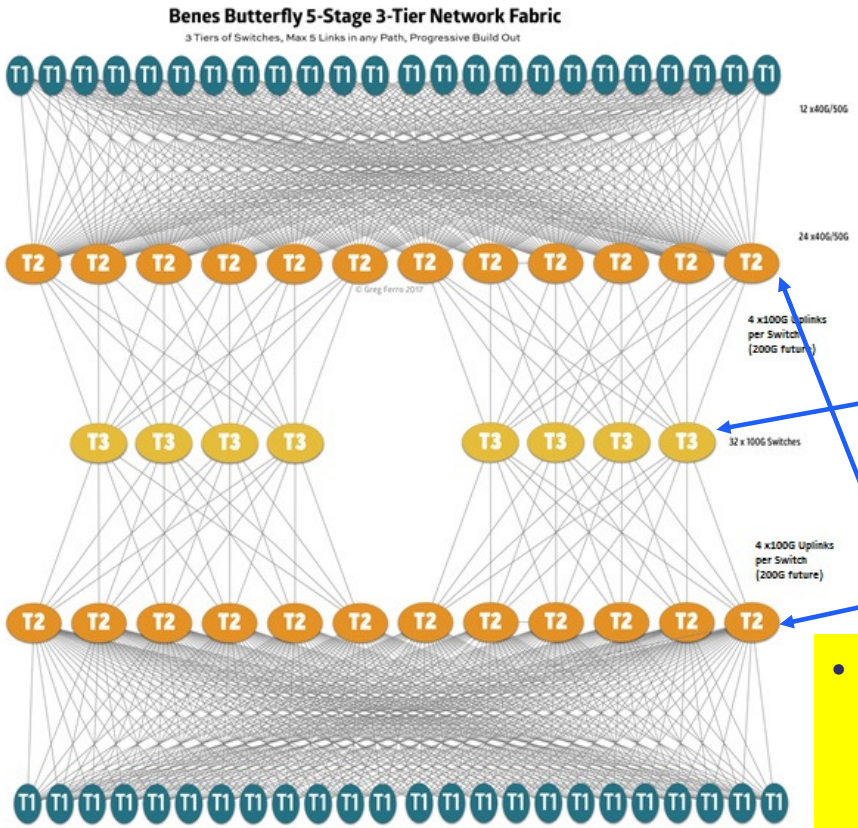


What do Google and Facebook do ?

Science and Technology Facilities Council
Scientific Computing

Facebook has published this article about its three tier architecture:
https://code.facebook.com/posts/360346274145943/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/

A “Data Centre” Network for JASMIN and then ….



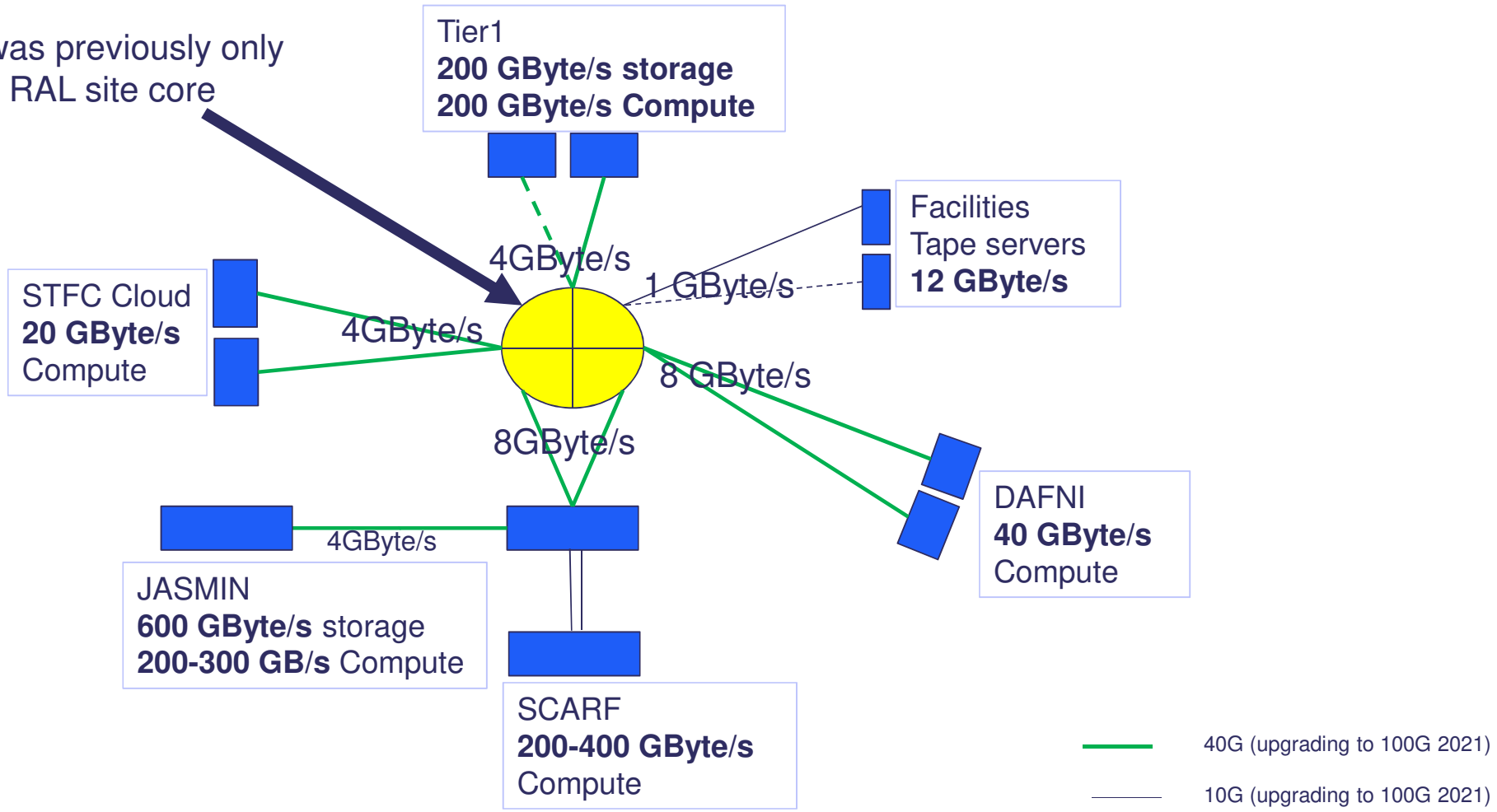
16 ‘T3’ routers in JASMIN implementation in 4 groups of 4

JASMIN has a mix of 2,8,12 ‘T2’ pods

- But ….
  Tier1, STFC Cloud , Facilities, DAFNI networks look similar to JASMIN Pods, or can be made to look similar.
  - Tier1, STFC Cloud, etc use 2x or 4x ‘T2’

Trapped Bandwidth

- Interoperation was previously only possible via the RAL site core network.

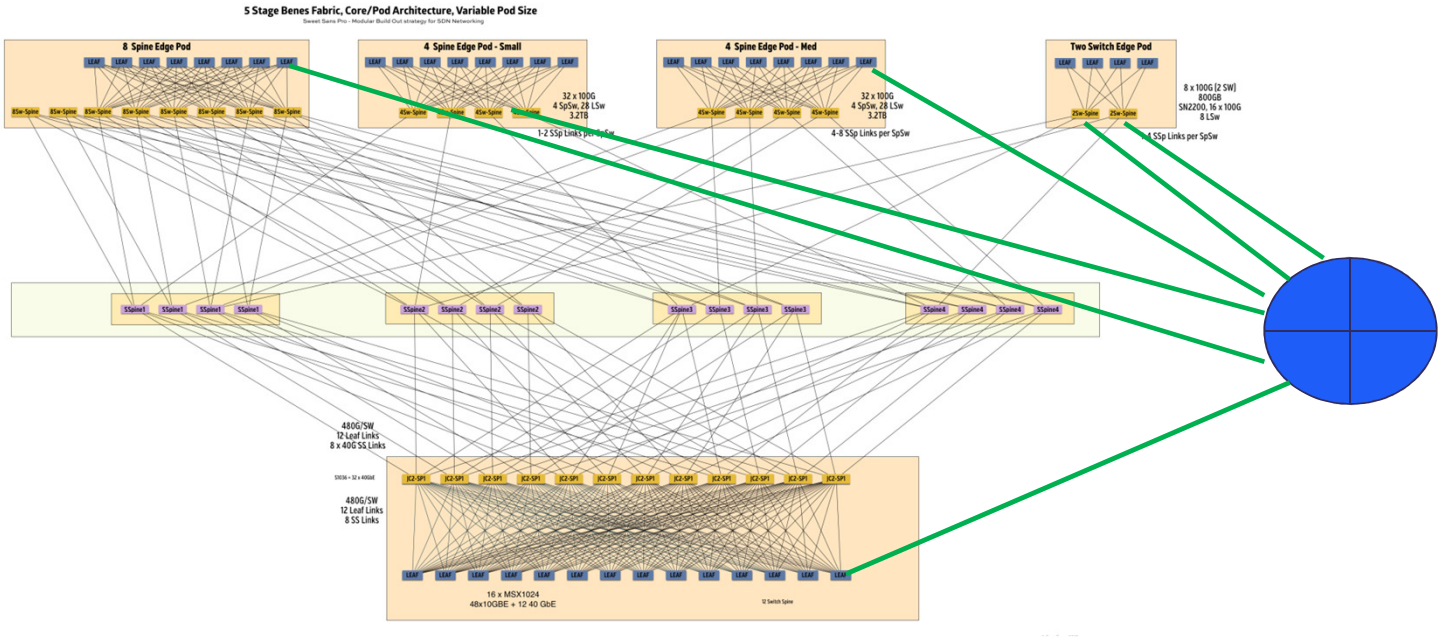


Science and Technology Facilities Council

Scientific Computing

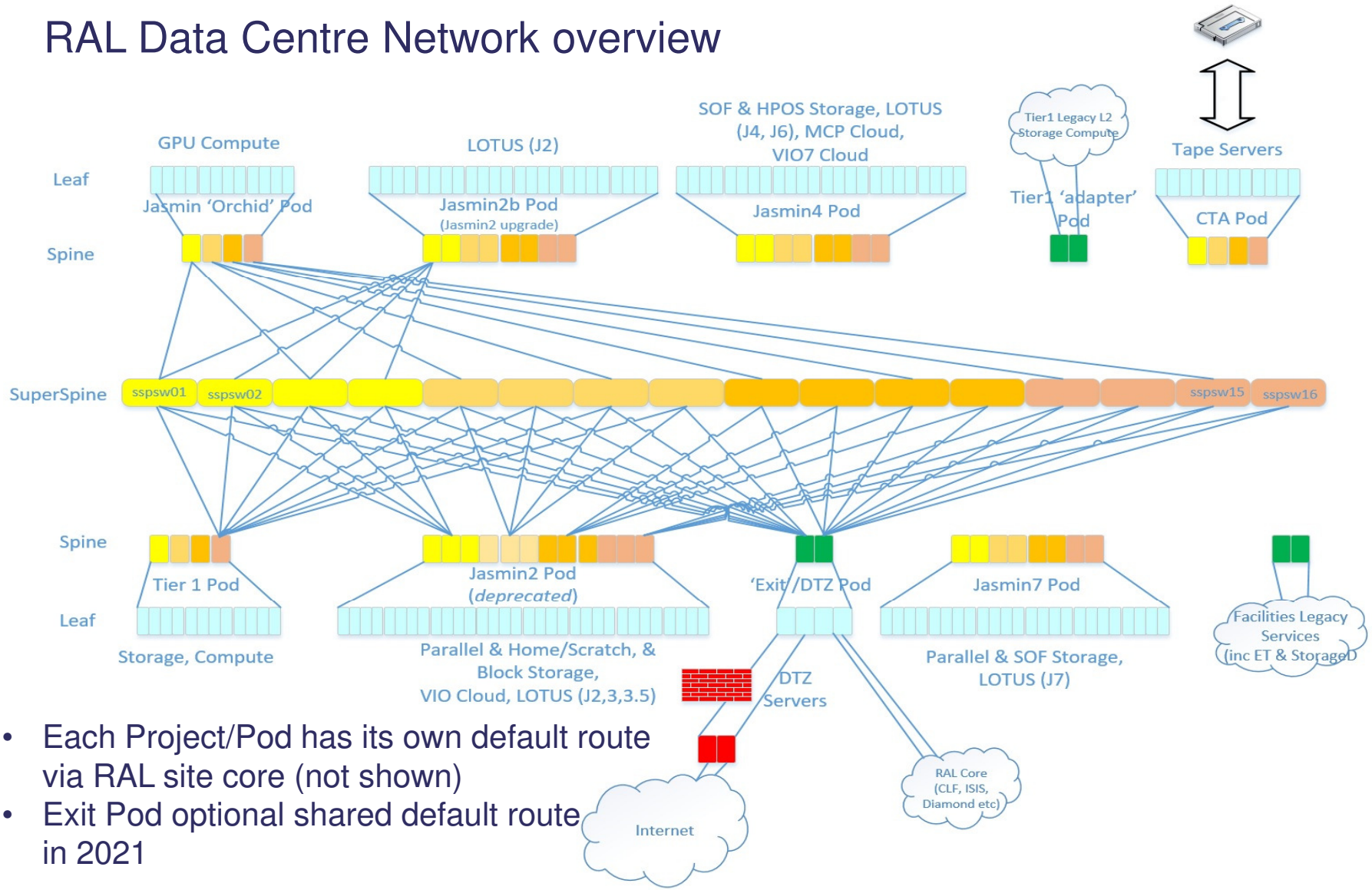
- *>1000 Gbytes/sec trapped behind 25 Gbytes/sec site core links*
- *Can't use the RAL site core network to interoperate even at 100Gb…..*

Site Core Default Routes Remain



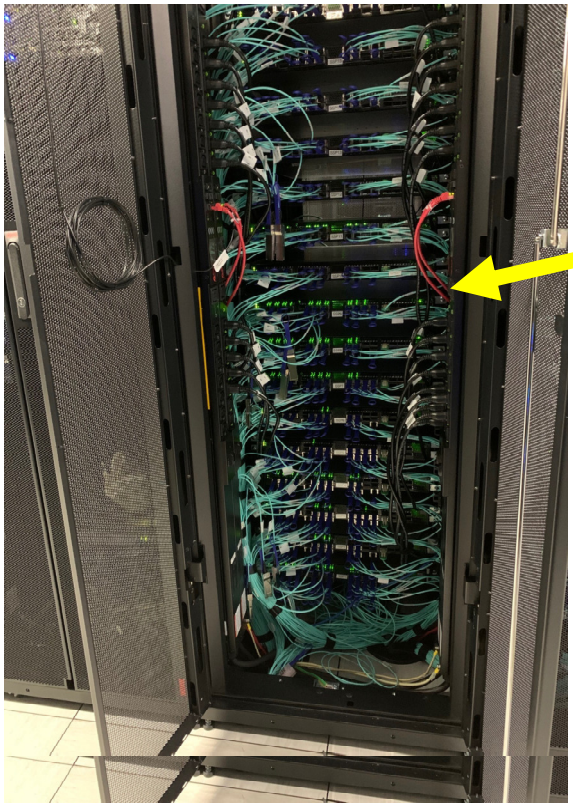
- Data Centre network is effectively a high bandwidth bypass of the site core
- But Pods cant utilise the single 0.0.0.0/0 default route for now.
  - Default route in the superspine is configured to be excluded (or lowest priority)
- Shared “Exit” Pod to site front door later in 2021

RAL Data Centre Network overview

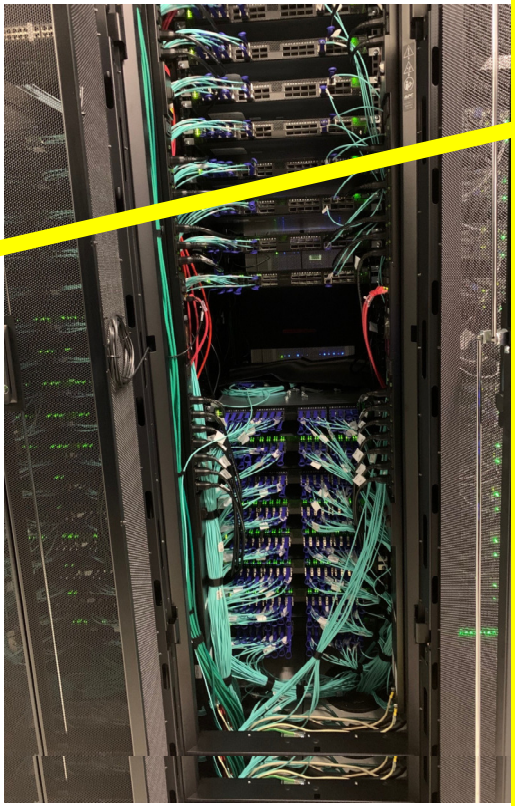


- Each Project/Pod has its own default route via RAL site core (not shown)
- Exit Pod optional shared default route in 2021

What it Looks like….



Superspine/T3 Rack



JASMIN 4&7 Spine Rack
8 Spine Pods

SuperSpine Switch Port/Bandwidth Allocation

Ports 1-4 : JASMIN4 6.4Tb
Ports 5-7 : JASMIN2 (40Gb) 1.92Tb
Ports 11 : Tier1 L2 'Adapter' 1.6Tb
Ports 12-15: JASMIN7 6.4Tb
Ports 16-19: JASMIN2B 6.4Tb
Ports 21: Facilities 'CTA' 0.8+Tb
Ports 8: Facilities 'Adapter' 0.4+Tb
Port 23: JASMIN Orchid 0.8+Tb
Port 24: STFC Cloud (R89) 1.6Tb
Port 25: STFC Cloud (R26) 1.6Tb
Port 9: DAFNI 0.4+Tb
Port 27: Exit Pod
Port 26: SCARF

64x100Gb Patch Panel to R26
Rack layout for future 200/400Gb Upgrade

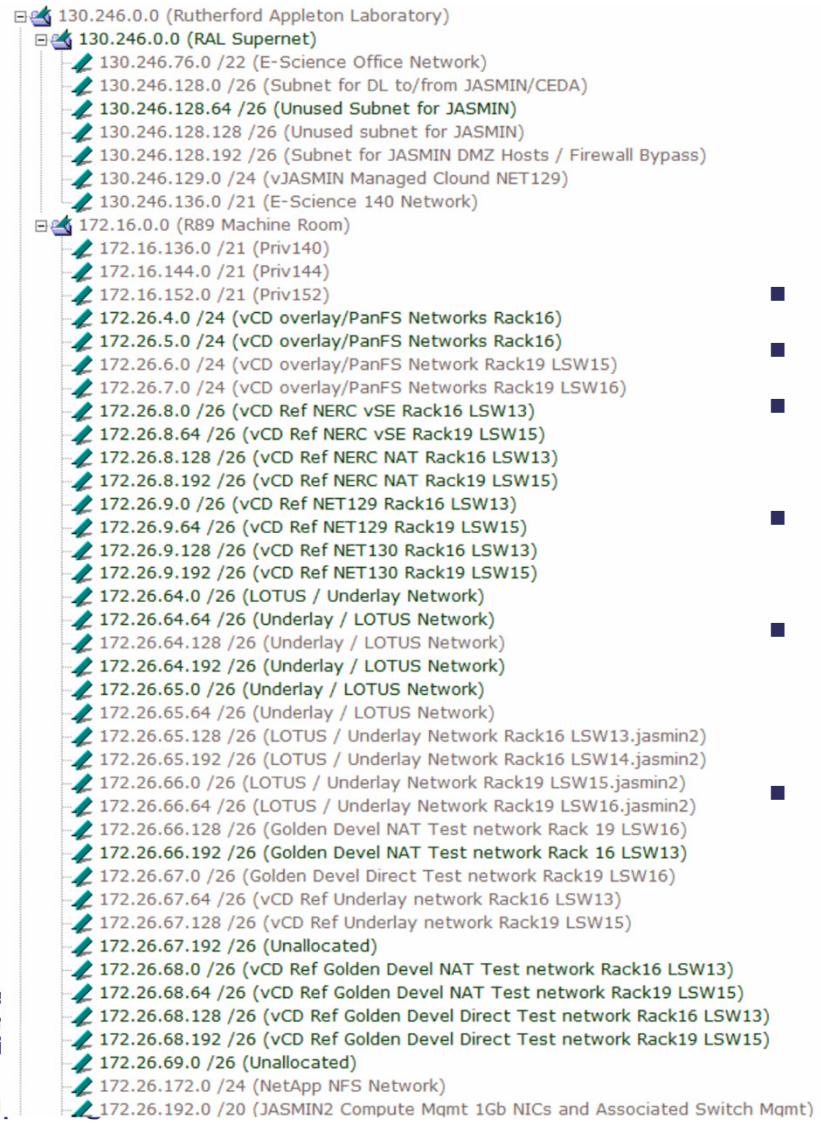
All AOC Cabling (except to R26 – 2km 100G optics)

Pod Comparison & Routing Protocols

Pod	Routing Protocols	NOS	Pod Spines	Superspine Connections	Notes
Superspine	eBGP, iBGP	Cumulus		(16x 32 Ports)	
JASMIN	eBGP, iBGP, OSPF	Cumulus, Onyx	12,8,2	36x40G 64x100G 8x100G	Cumulus Spines, Onyx Leafs OSPF Leaf to Spine BGP Spine to Superspine
Tier1 adapter	eBGP	Cumulus	2	16x100G	L2 Leaf to Spine, eBGP to SSpine
Tier1	iBGP	Cumulus	4	16x100G	BGP all OSPF to site core
STFC Cloud	iBGP , eVPN (OSPF)	Cumulus	2	16x100G	BGP all, OSPF to site core
Facilities CTA	iBGP (OSPF)	Cumulus	2	8x100G	BGP all, OSPF site core
'Exit' Pod	eBGP, eVPN, static	Cumulus, Onyx, FTOS	2	8x 100G	FTOS to the border routers BGP all. eVPN to client Pods.

- Wide Mix of topologies, protocols, NOS, compute and storage
- Tuneable Pod to superspine bandwidth
- Mellanox SwitchX, Spectrum1, Spectrum2, Dell Force10 (~250 IP Fabric Switch/Routers Total)
- Config mgmt.: Mix of Ansible, NEO and manual/vendor
- Monitoring: LibreNMS, Icinga2, NetQ

Subnet & IP Management headache



- ~1,000 Underlay Routes (+ >1,000 p2p links)
- 6,500+ IPs (in JASMIN alone)
- > 300 L2 VLAN IDs

- Managing Subnets and IPs is a headache we really haven't solved.
- IPplan but only beta support IPv6 and no longer maintained
  http://iptrack.sourceforge.net/
- Migrating to NetBox (since 2018 !)
  https://netbox.readthedocs.io
  and/or RAL site IPAM.

Summary

- RAL Data Centre Network :
  - Supports high bandwidth (multi Terabit/s), low latency, interoperability of STFC's hosted infrastructures, such as Tier1, JASMIN, STFC Cloud, SCARF, IRIS, DAFNI
  - Allows a wide range of infrastructure CLOS and L2 legacy network topologies, to inter communicate at tunable bandwidths.
  - Supports a mix of Cumulus, Onyx (and FTOS) NOS using a mix of routing protocols.
- It's currently a high bandwidth bypass of the RAL site core network
  - Care how the default route is treated.
  - A shared 'Exit' Pod is planned for 2021 to bypass the RAL site core for science data traffic.

Science and Technology Facilities Council

Scientific Computing

Questions?

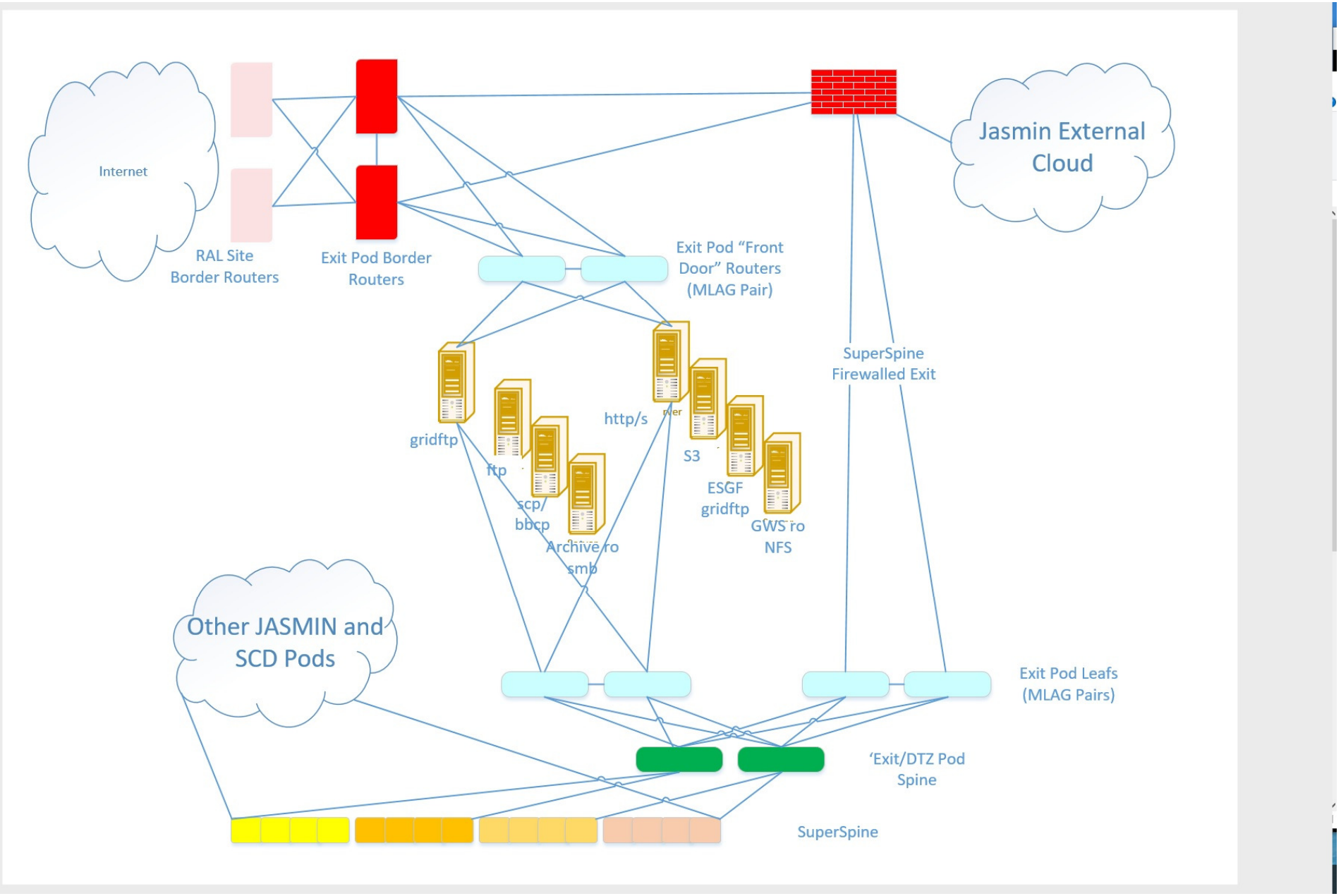
Science and Technology Facilities Council

Scientific Computing

Thank you

scd.stfc.ac.uk

@SciComp_STFC


Internet
RAL Site
Border Routers
Exit Pod Border
Routers
Jasmin External
Cloud
Exit Pod "Front
Door" Routers
(MLAG Pair)
SuperSpine
Firewalled Exit
gridftp
ftp
scp/
bbcp
Archive ro
smb
http/s
S3
ESGF
gridftp
GWS ro
NFS
Other JASMIN and
SCD Pods
Exit Pod Leafs
(MLAG Pairs)
'Exit/DTZ Pod
Spine
SuperSpine