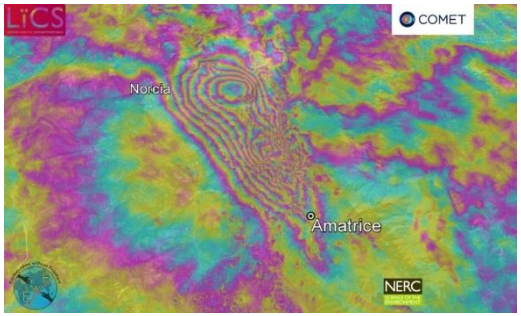
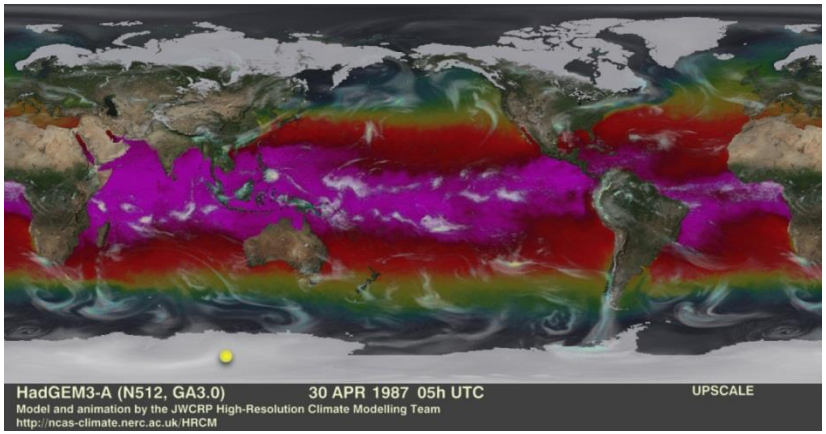
On the challenges of deploying an unusual high performance hybrid object/file parallel storage system in JASMIN

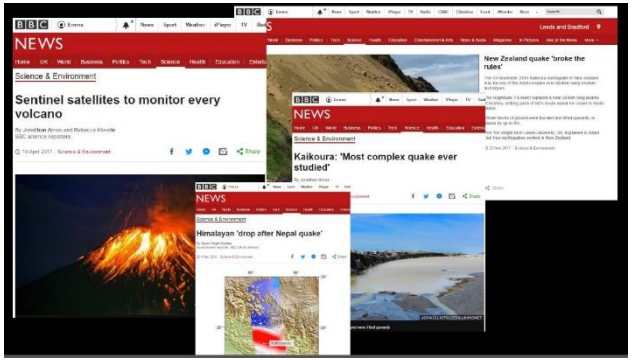
Cristina del Cano Novales[1], Jonathan Churchill[1] , Athanasios Kanaris[1], Robert Döbbelin[2], Felix Hupfeld[2], Aleksander Trofimowicz[2]

1 Scientific Computing Department, Science and Technology Facilities Council, RAL, Didcot OX11 0QX, UK
2 Quobyte GmbH, Berlin, AG Charlottenburg HRB 149012 B, Germany

Environmental Data Analysis







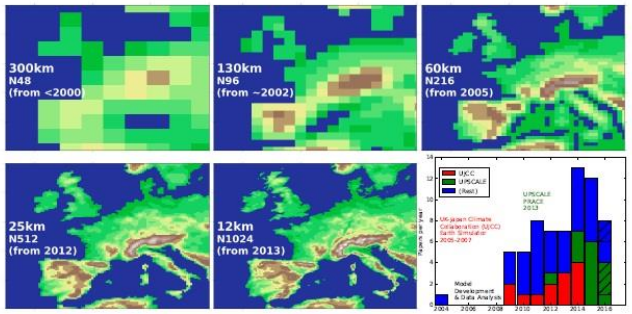
- Centre for Environment and Hydrology
- Trends for 1000's of species
- Analysis unprecedented in complexity and scope within the UK.

- COMET-CPOM UoLeeds
- Near real time monitoring of all active earthquake and volcanos.
- Relies on full ESA Sentinel data, Managed and unmanaged tenancies, LOTUS batch

Science & Technology Facilities Council

JASMIN: the missing piece

Growing Need - High Resolution Climate Programme!

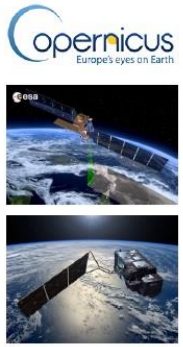


Just one example, of the *many* axes of growing scientific demand in simulations and observation:

- From 7K to 3.1M points (0.05 MB to 25MB) for a single timestep of a single level of a global field.
- Multi-year data management campaigns support the data analysis (which needs to include similarly high-resolution observations).

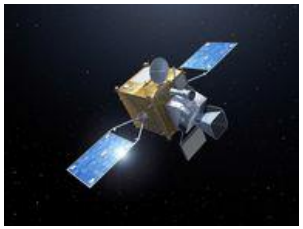
The Organised Data Deluge

CMIP6 data volumes and data rates not yet known, but the European contribution to HiresMIP alone is expected to exceed 2 PB.



Sentinel 1A (2014), 1B (2016)
Sentinel 2A (2015) 2B (2017?)
Sentinel 3A (2016) 3B (2018?)

Data rate: o(6) PB/year



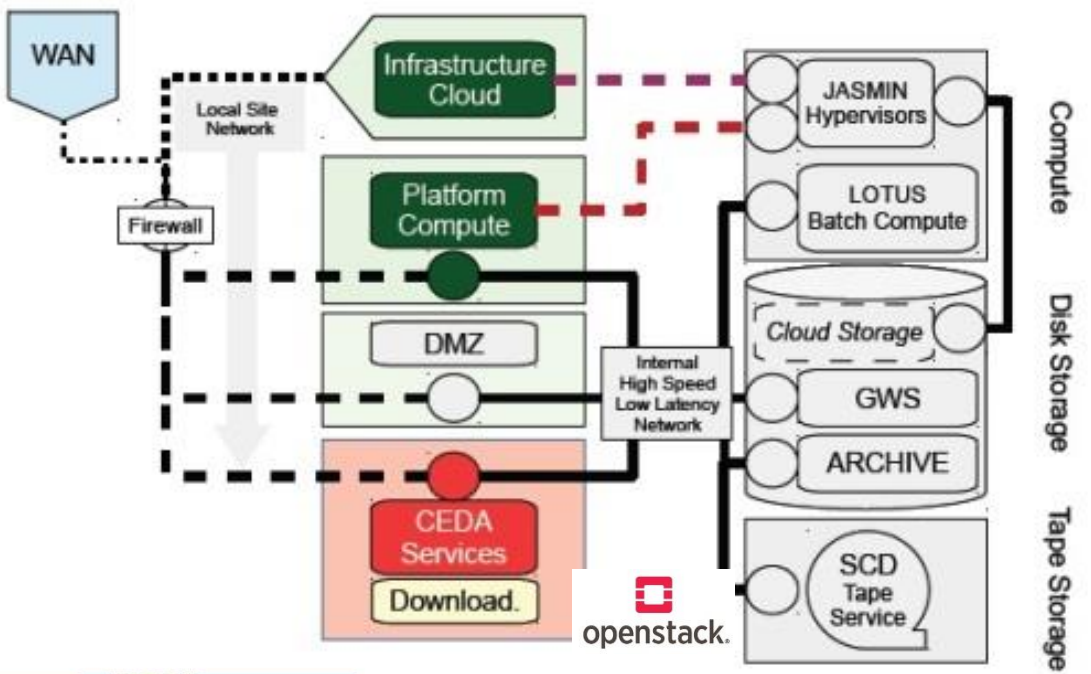
MetOffice supercomputer

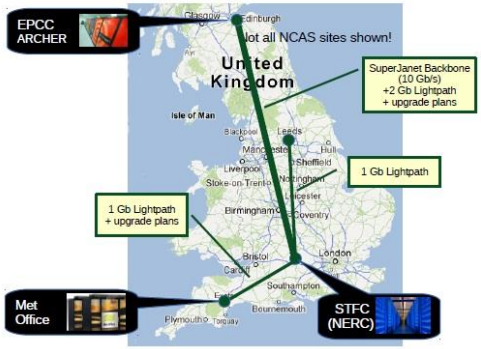
ARCHER supercomputer (EPSRC/NERC)

JASMIN (STFC/Stephen Kill)

Science & Technology Facilities Council

Blending PB's of data, 1000's of Cloud VM's, Batch Computing & WAN Data transfer

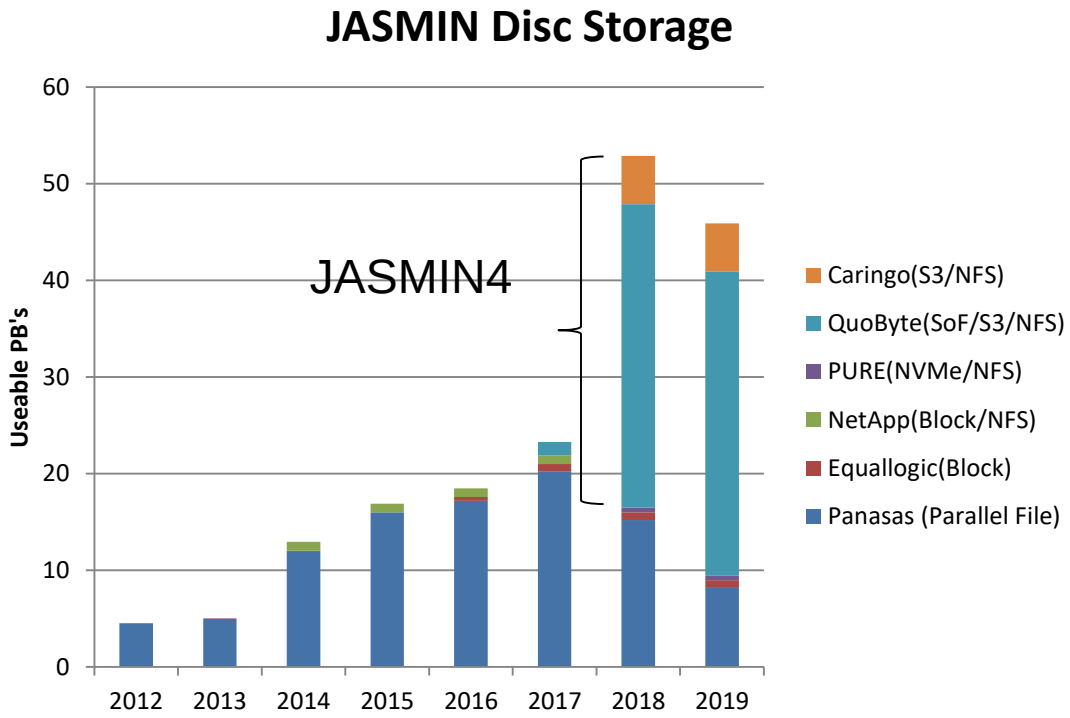




- ✓ 24.5 PB Panasas ~ 250GByte/s
- ✓ 44 PB Quobyte SDS ~ 220GBytes/s
- ✓ 5PB Caringo Object Store
- ✓ 80PB Tape
- ✓ Batch HPC 6-10k cores
- ✓ Optical Private WAN + Science DMZ
- ✓ "Managed" VMware Cloud
- ✓ OpenStack "Community" Cloud
- ✓ Pure FlashBlade scratch
- ✓ Non-blocking ethernet 12-20Tbit/sec

Science & Technology Facilities Council

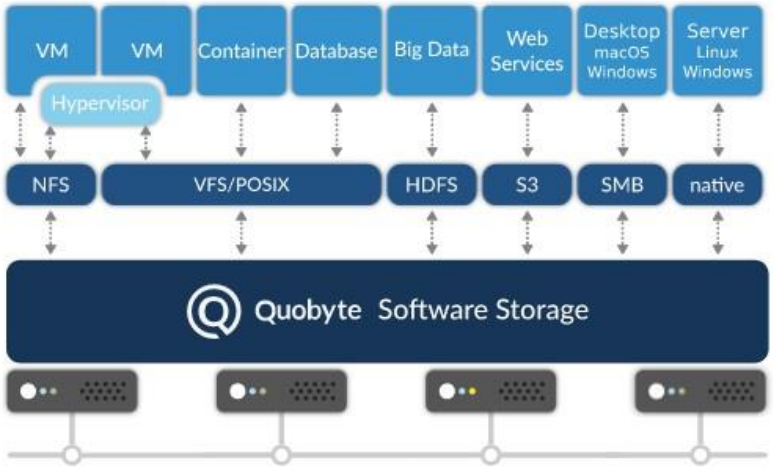
JASMIN4 Disc Storage



- No boundaries on data growth (or network topology)
- S3 interface to file and object system. RW Both sides.
- Performance similar to Panasas PFS
- Online upgrades. Redundant networking.
- No client “call back” port.
  - Previous root /network and UMC restrictions

Science & Technology Facilities Council

Quobyte SDS



Parallel File System
HPC

Distributed File System
Video, CGI, EDA

Storage for containers
Kubernetes, Mesos, Docker

Scale-out NAS
Enterprise applications

Hadoop File System
Big Data

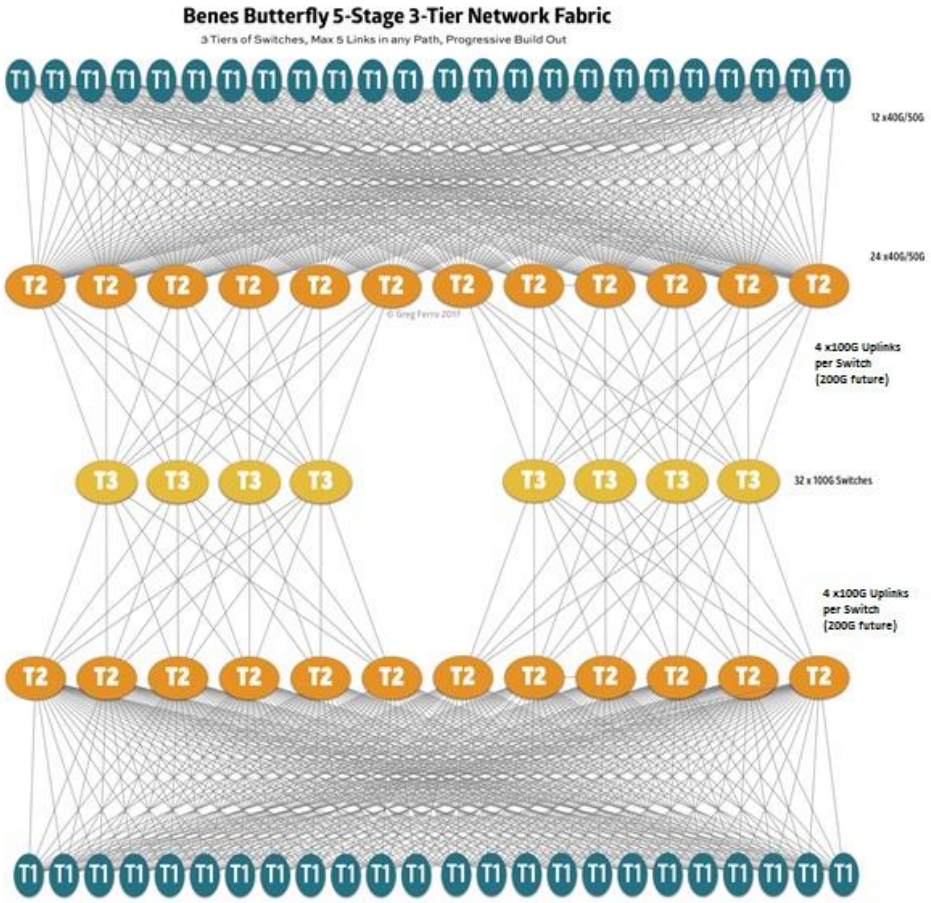
Archival storage
HPC, backup, e-Science

Block storage for VMs
OpenStack, hyperconverged

Object storage
Service provider-grade S3 storage

- 45PB raw, ~30PB usable (EC 8+3)
- Hardware split 50:50 Dell / Supermicro
- 47x R730xd's + MD3060 arrays (1 / server pair) - 40Gb NICs
- 40x Supermicro 4U "Top loader" servers – 50Gb NICs
- Target > 50MB/sec/HDD. Ideally 70-100MB/sec/HDD

“5 Tier” CLOS Network



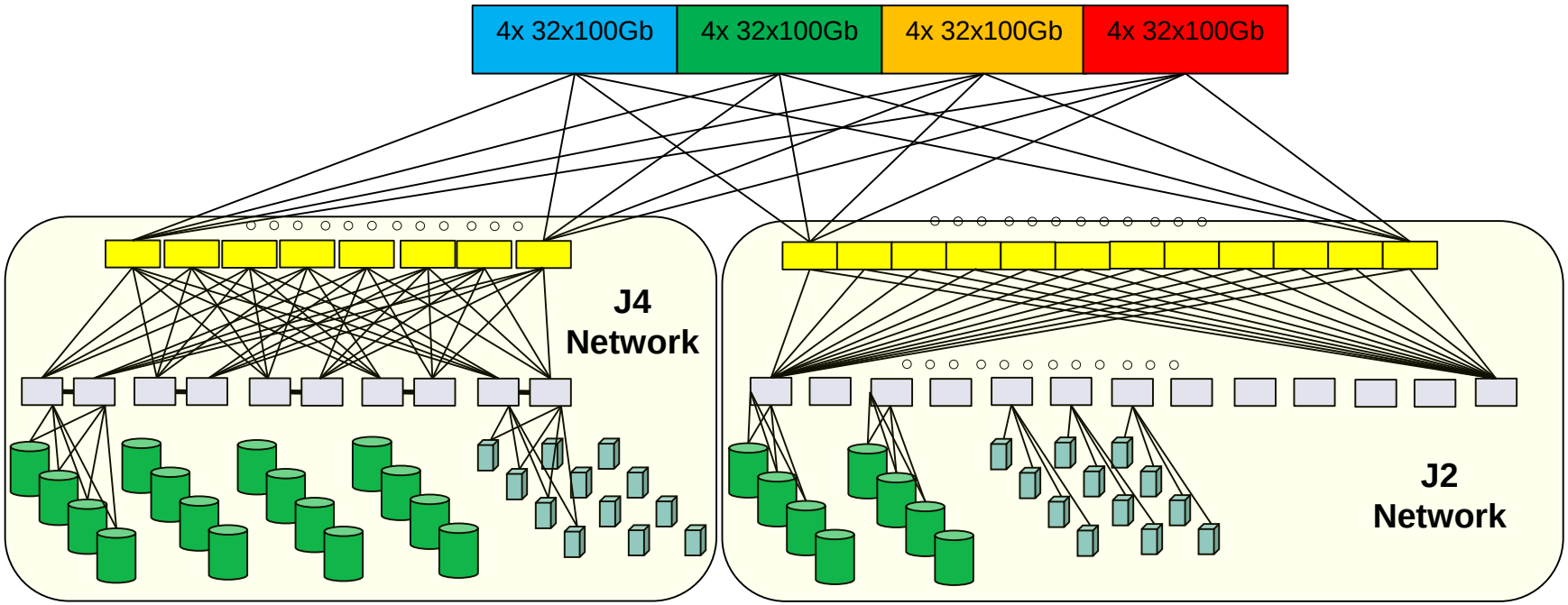
- Traditional for BGP throughout
- JASMIN2/3 all OSPF
- OSPF Lower complexity cf BGP
- Keep OSPF Leaf-Spine for JASMIN4
  - Ease of use at the edges.
- BGP only in Spine to SuperSpine
  - For the core network specialists
  - But stops EVPN leaf use for now

Science & Technology Facilities Council

Connecting JASMIN2 to JASMIN4

Superspine: 16 Spines (32x 100Gb)

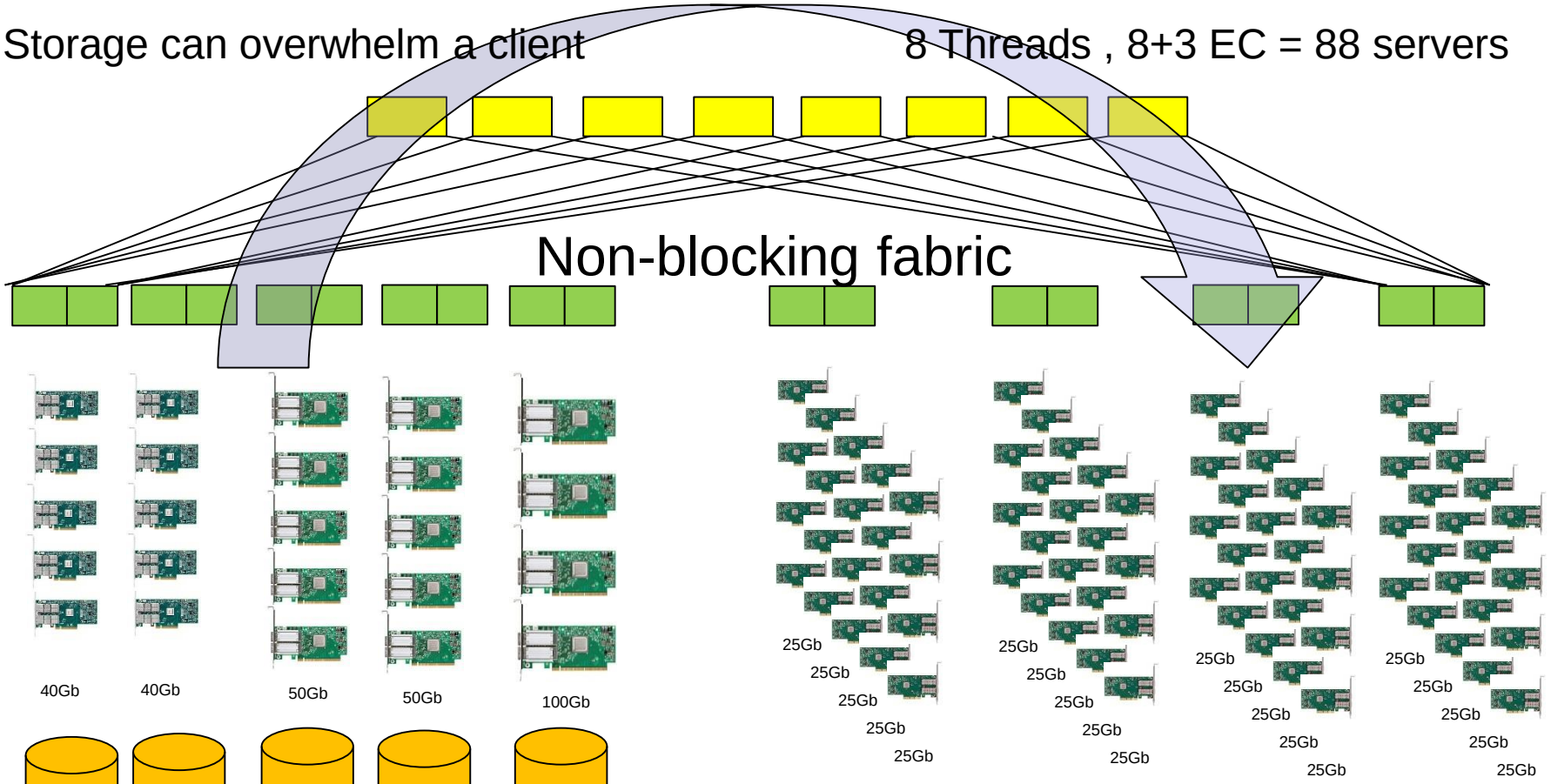
- 4 Cluster/groups of 4 routers



- 8 Spines (32x 100Gb)
  - 4x 100Gb to Super-Spine
- 17 Leaf pairs (2 of 16x 100Gb)
  - 8x 100Gb uplinks. 1 per spine
- Storage/Compute
  - 1x 25/40/50Gb to ‘A’ and ‘B’ le

- 12 Spines (36x 40Gb)
  - 4x 40Gb to Super-Spine
- 30 Leafs (48x10Gb+12x40Gb)
  - 12x 40Gb uplinks. 1 per spine
- Storage/Compute
  - 2x 10Gb to local leaf

Congestion in a “non-blocking” network



But 180x25Gb > 4Tbits/s

3090 HDD’s x 70MB/s > 250GBytes/sec > 2Tbits/sec

~200GB/s for a few minutes

Science & Technology Facilities Council

Thank you!