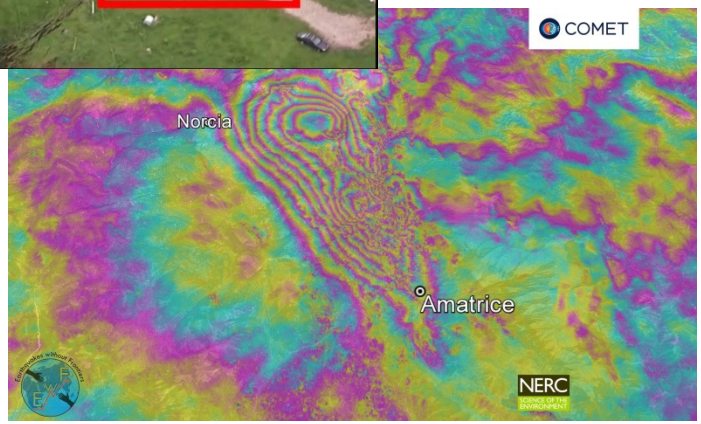
JASMIN - On the road to High Performance Object Stores

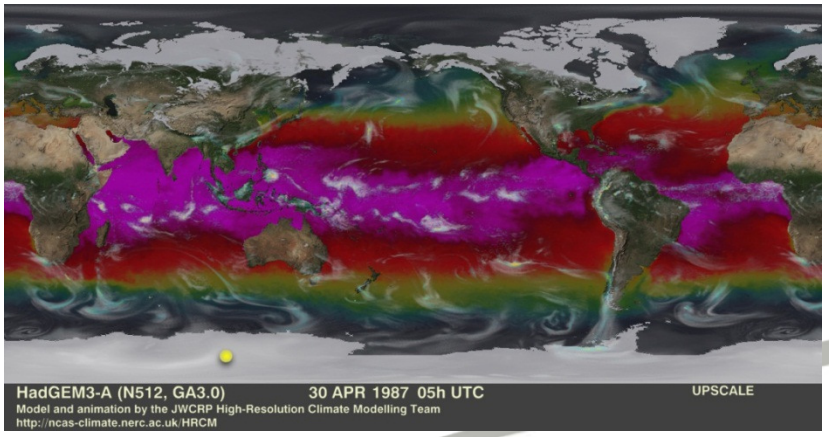
Jonathan Churchill
JASMIN Architect & Infrastructure Manager

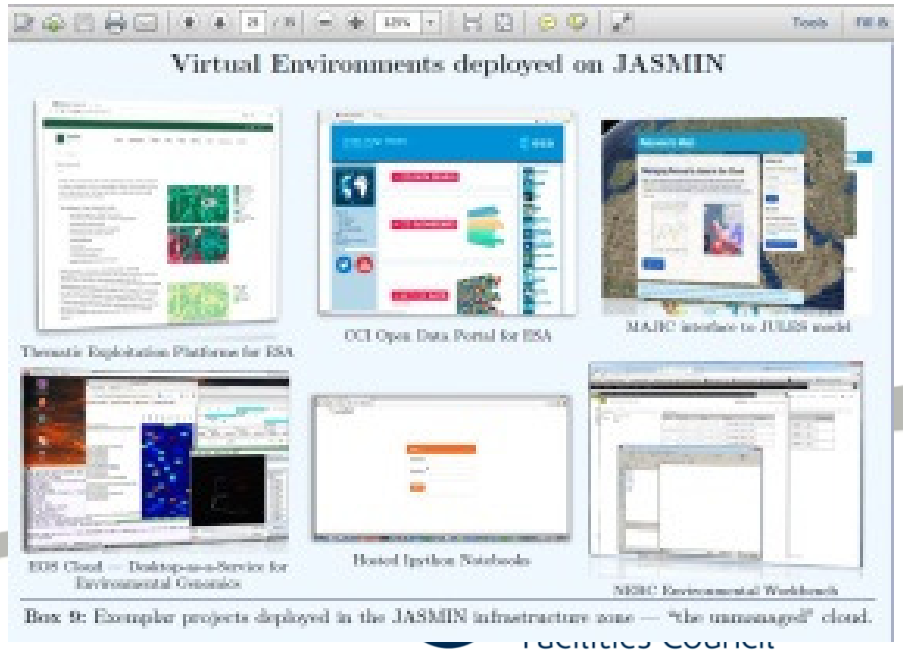
Research Infrastructure Group
Scientific Computing Department
STFC – Rutherford Appleton Labs. UK

Environmental Data Analysis

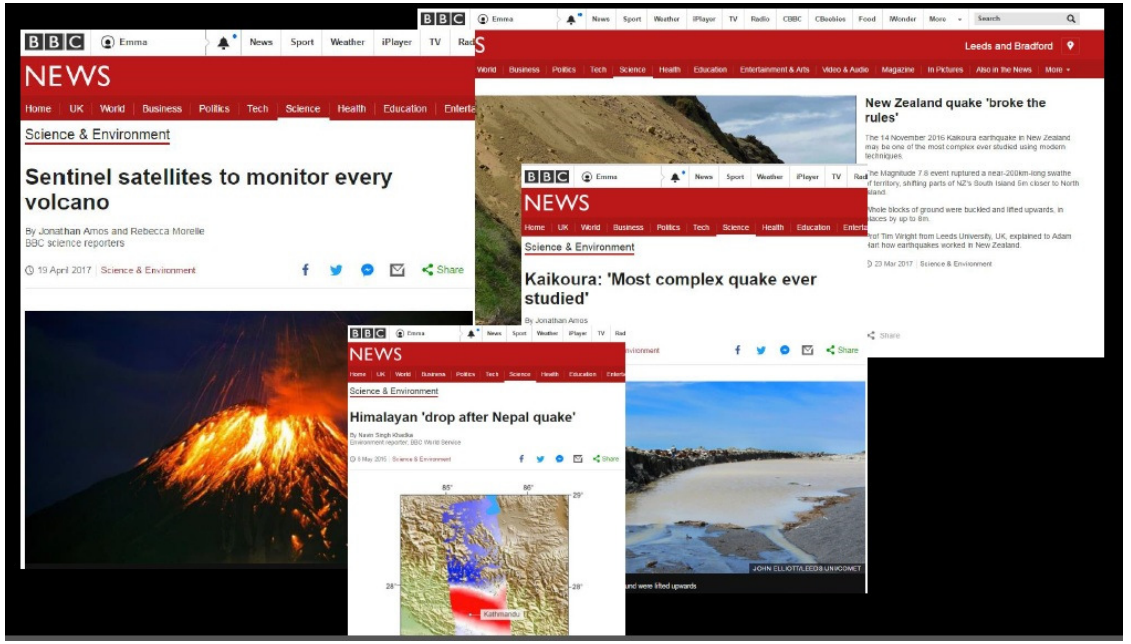


Biases in *ad-hoc* data





Jonathan.Churchill@stfc.ac.uk STFC RAL



Department for Environment Food & Rural Affairs | Agriculture, Environment and Rural Affairs | The Scottish Government

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

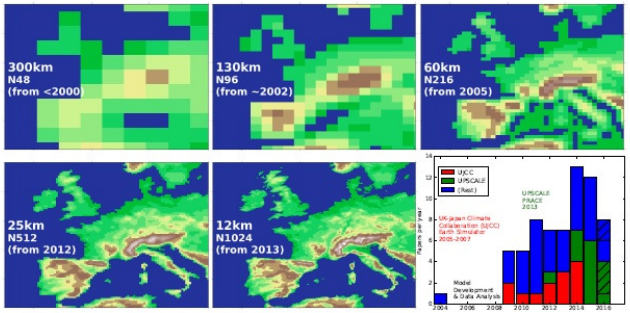
- COMET-CPOM UoLeeds
- Near real time monitoring of all active earthquake and volcanos.
- Relies on full ESA Sentinel data, Managed and unmanaged tenancies, LOTUS batch

Science & Technology Facilities Council

Jonathan.Churchill@stfc.ac.uk STFC RAL

JASMIN: the missing piece

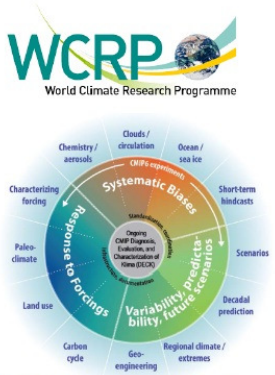
Growing Need - High Resolution Climate Programme!



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



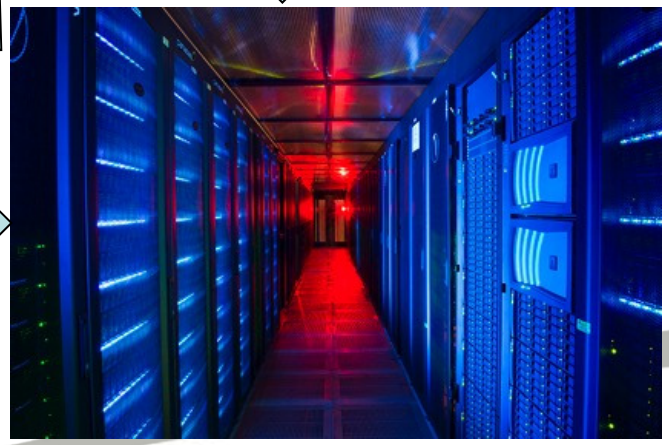
CMIP6 data volumes and data rates not yet known, but the European contribution to HiresMIP alone is expected to exceed 2 PB.

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)



JAMSIN (STFC/Stephen Kill)

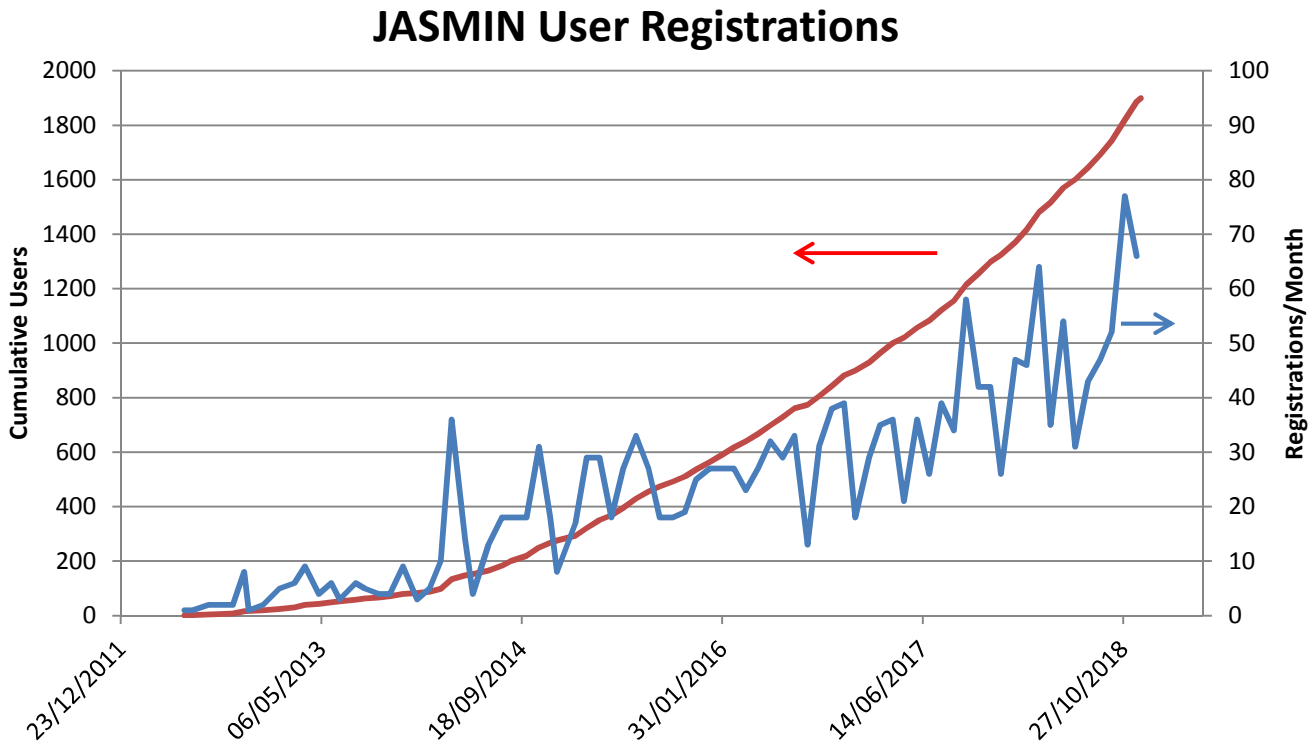
Science & Technology Facilities Council

chill@stfc.ac.uk STFC RAL

Outline

- Overview of JASMIN
- Why Object Stores
- File/Object Storage PoC.
- Quobyte @ JASMIN
- Caringo @ JASMIN
- Implementing a Data Centre Network
- Summary

Why Object Stores ?
- UID / GID Numbers



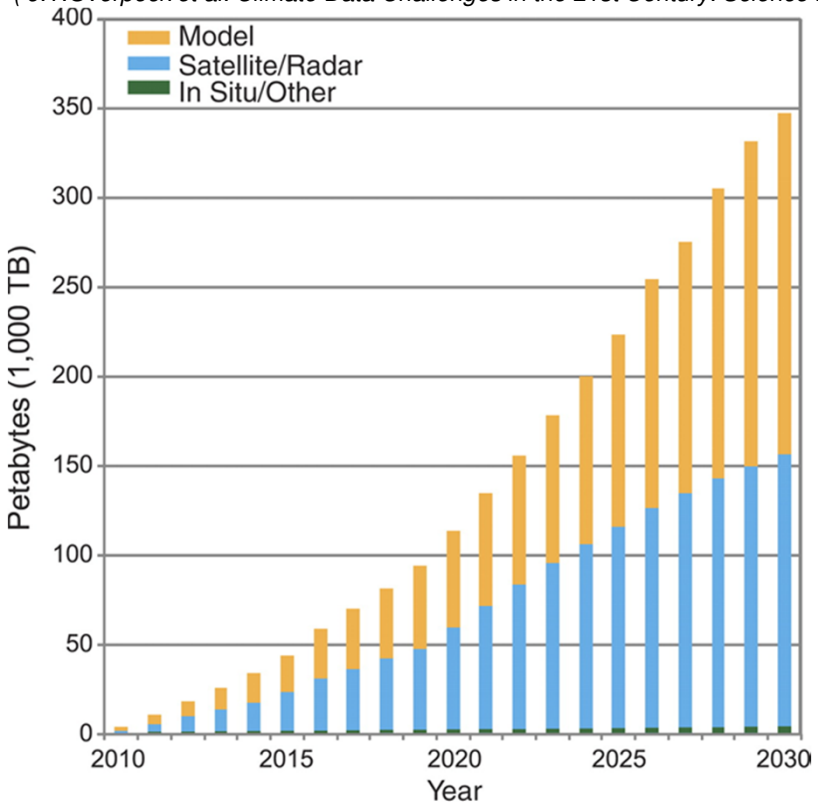
- Excludes 11,000+ CEDA Download users
- Cloud admin `useradd`'s clashing UID/GIDs

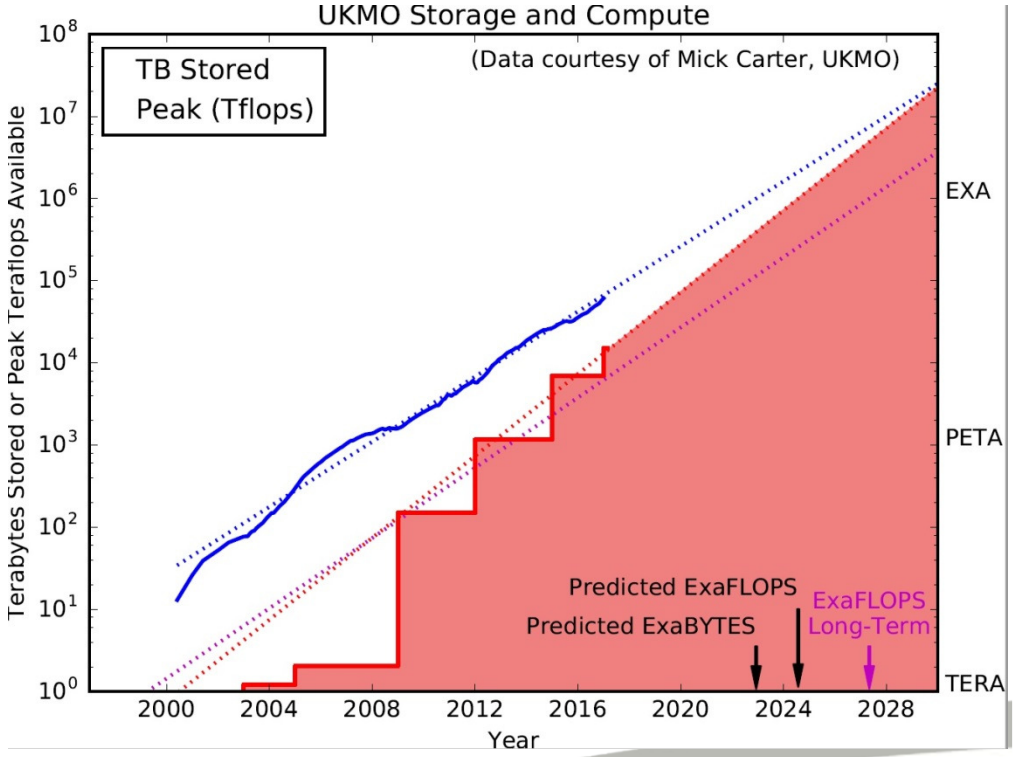
Jonathan.Churchill@stfc.ac.uk STFC RAL

Science & Technology Facilities Council

Why Object Stores ?
- Data Volume & # Projects

(J.T.Overpeck et al: Climate Data Challenges in the 21st Century. Science Feb 2011)



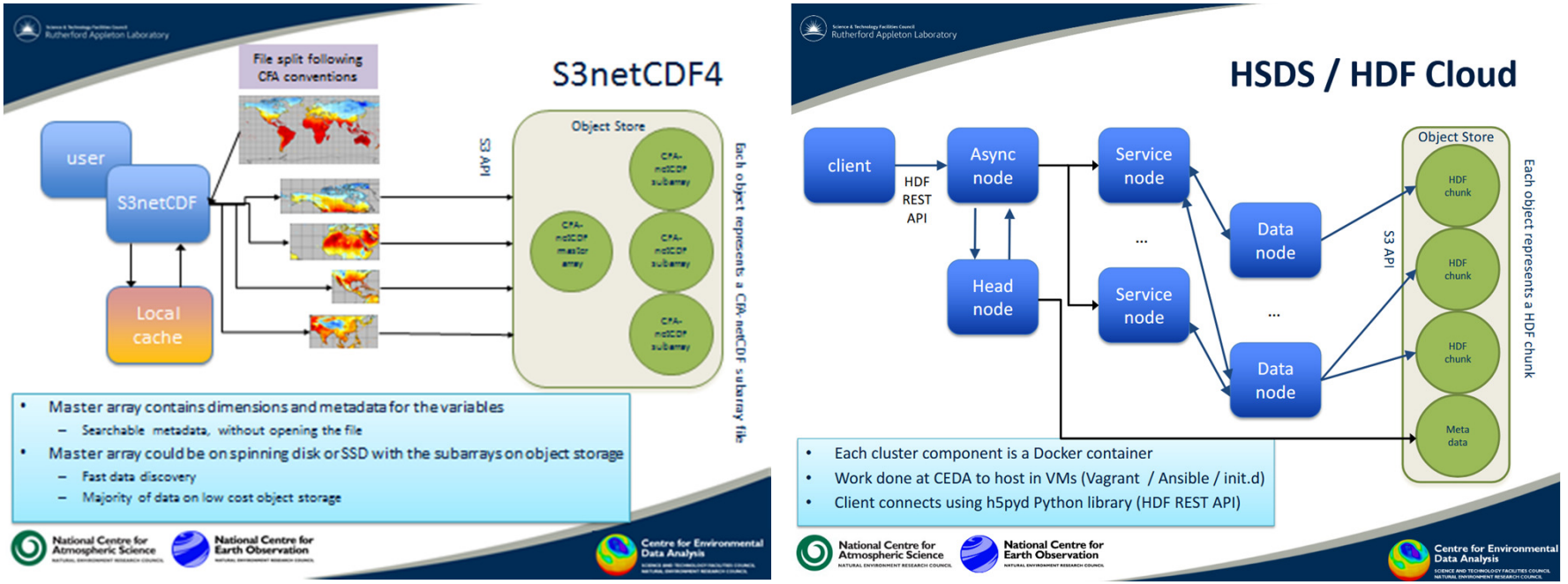


- Matching rise in analysis project / Tenant numbers
- Matching rise in (400++) storage volumes

Science & Technology Facilities Council

Jonathan.Churchill@stfc.ac.uk STFC RAL

Why Object Stores ?
- File formats



Slides courtesy Neil Massey STFC RAL

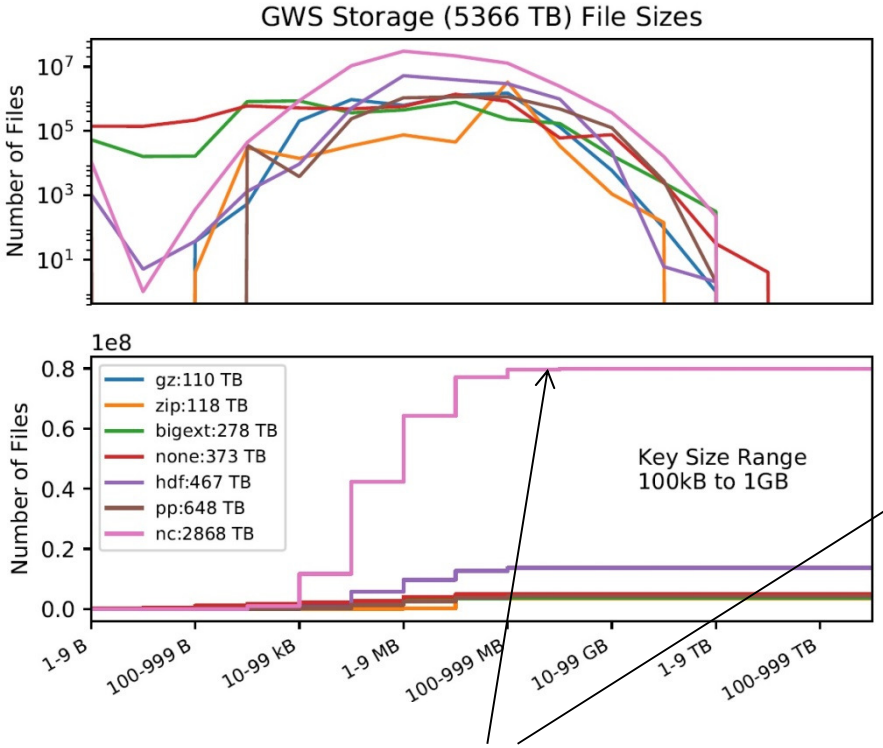
- NetCDF / HDF5:
  Object: Chunk access vs File: serial ‘skipping’

Science & Technology Facilities Council

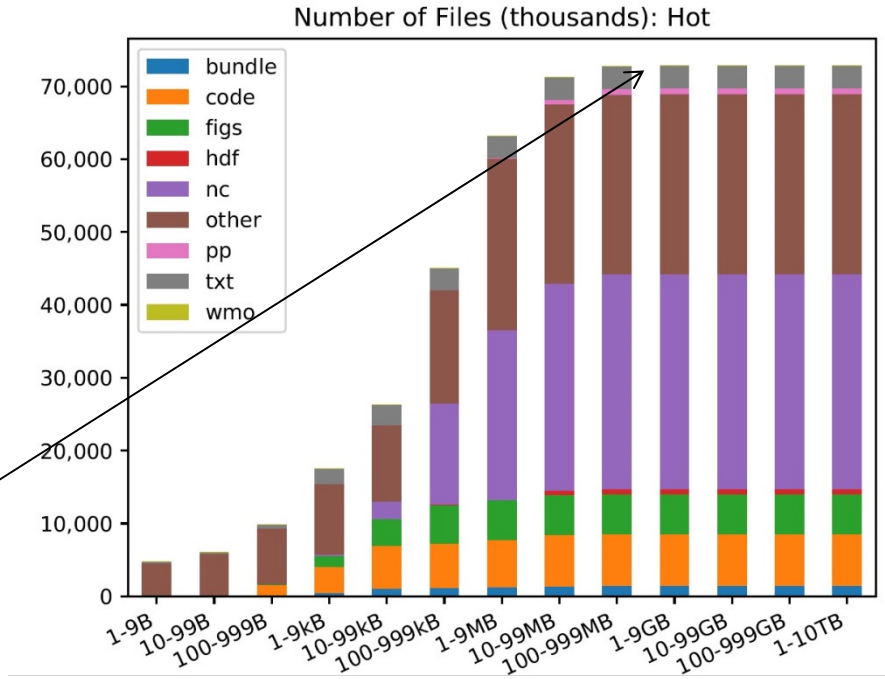
Jonathan.Churchill@stfc.ac.uk STFC RAL

Object PoC Object Size Choice.
JASMIN File Size and Type

All files



'Hot' files

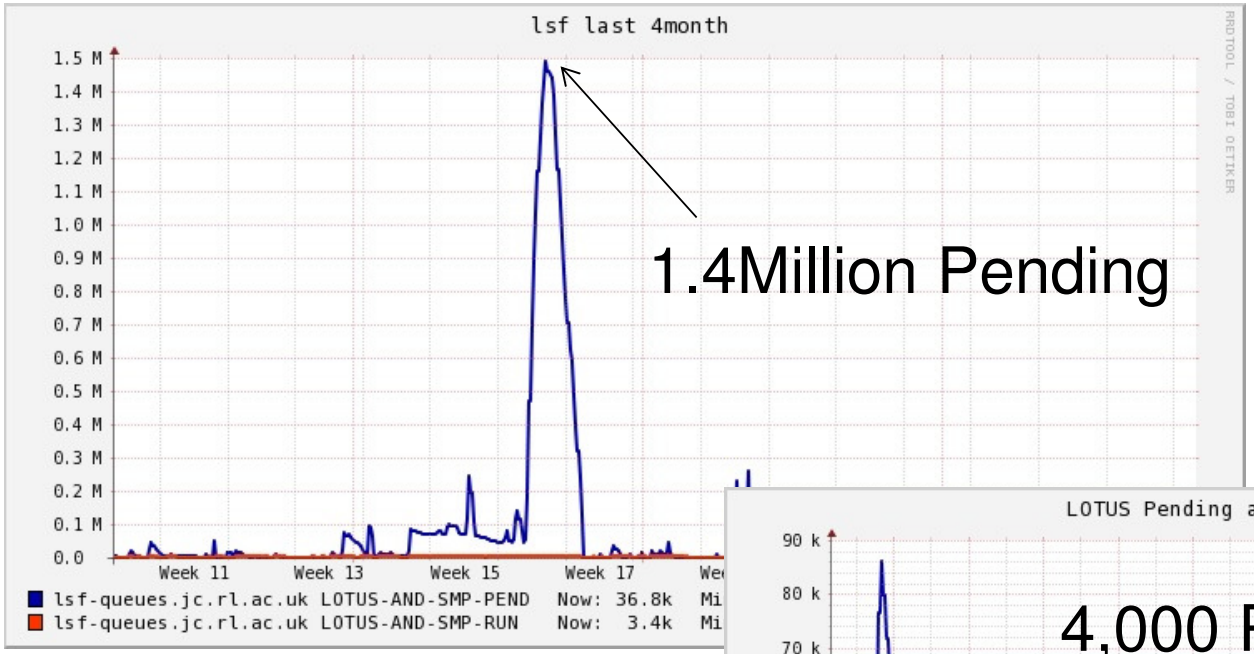


- Choose 2GB File/Objects for the PoC

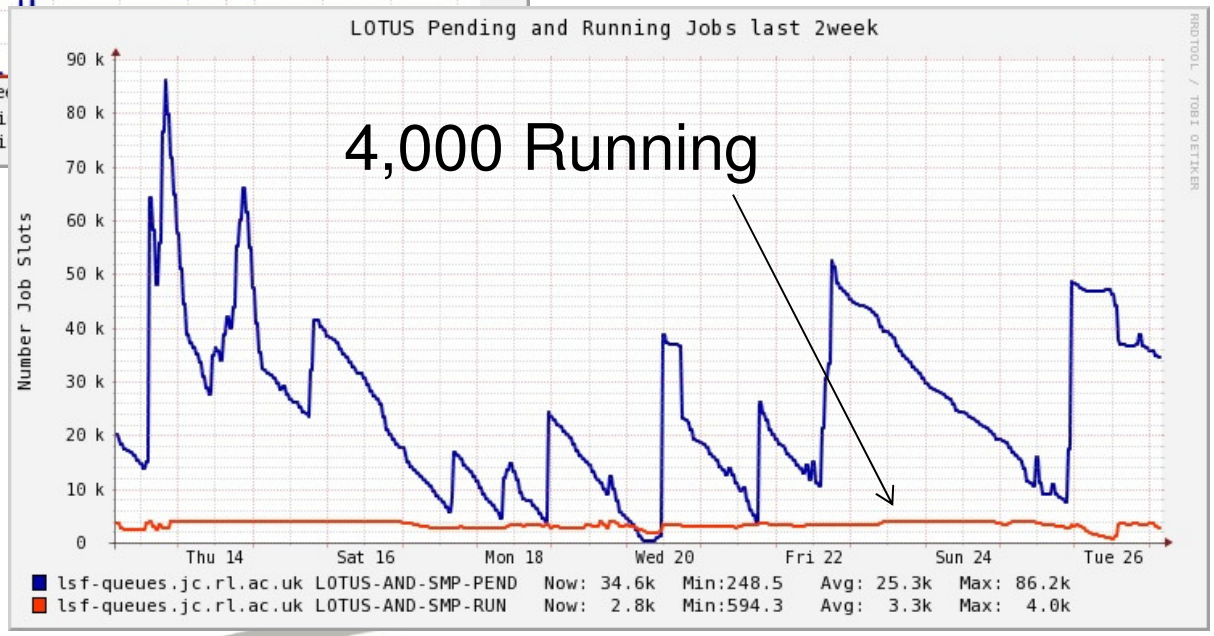
Jonathan.Churchill@stfc.ac.uk STFC RAL

Science & Technology Facilities Council

Why *High Performance* Object Stores ?



Running vs Pending LOTUS Jobs

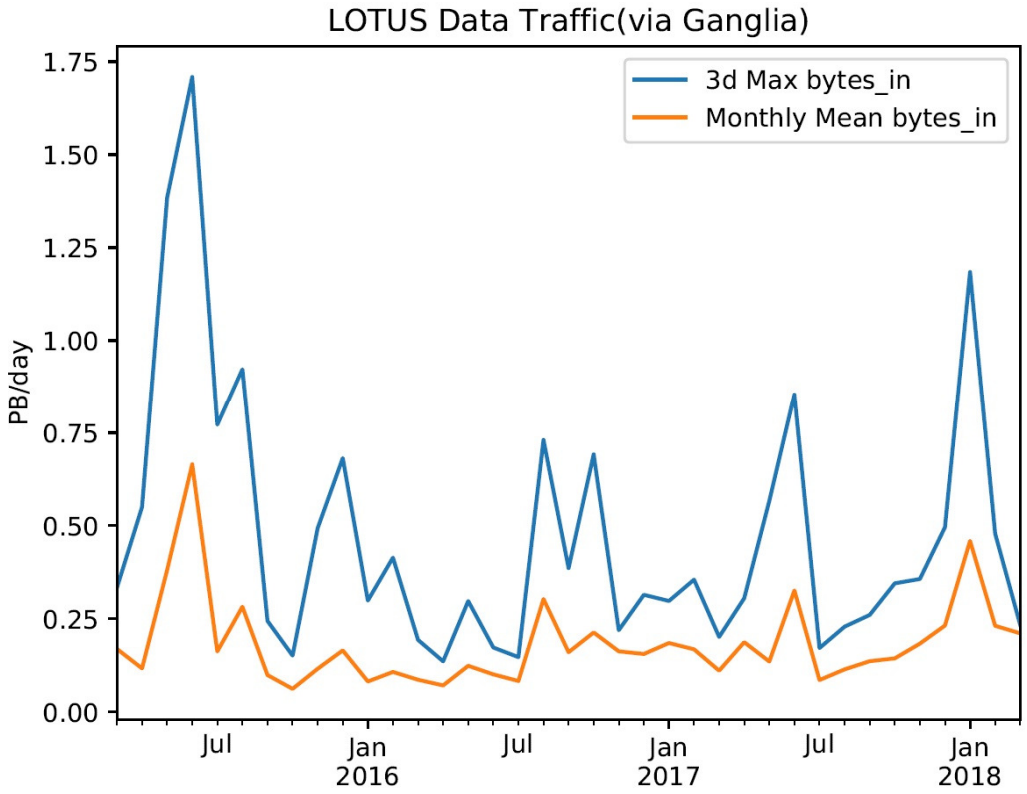


Read performance is key.

Science & Technology Facilities Council

Why *High Performance* Object Stores ?

JASMIN Data Rates



- 1Pbyte/day analysis reads from ~250-300x 10Gb nodes
- …it's not all the same data or project

Science & Technology Facilities Council

Jonathan.Churchill@stfc.ac.uk STFC RAL

JASMIN PoC Objectives

- Easily manage and scale a "large" single resilient pool (10-50Pbytes) using mixed vendor hardware.
  - Comprehends multiple annual rounds of procurement purchasing.
  - Ideally minimum increment 1 HDD.
  - Must be Erasure coded (£££).
- Accessible via S3/REST API's and NFS and/or custom file client.
  - Ideally S3 and NFS/client can read/write the same objects/files.
- Benchmark Objectives for S3/REST and NFS/custom interfaces
  - How much of the native HDD performance can the s/w extract.
  - Expectations :
    <= 40MB/s/HDD SoSo, 50MB/s/HDD Good, > 60MB/s/HDD Awesome.
- Reduce total CAPEX cost per usable TByte cf current PFS by 2-3x.
  - Without increasing human FTE support costs….
- Investigate capabilities of inbuilt metadata/search engines
- Investigate capabilities for multi-tenancy and object access
  eg multi-user/multi-admin and delegated admin

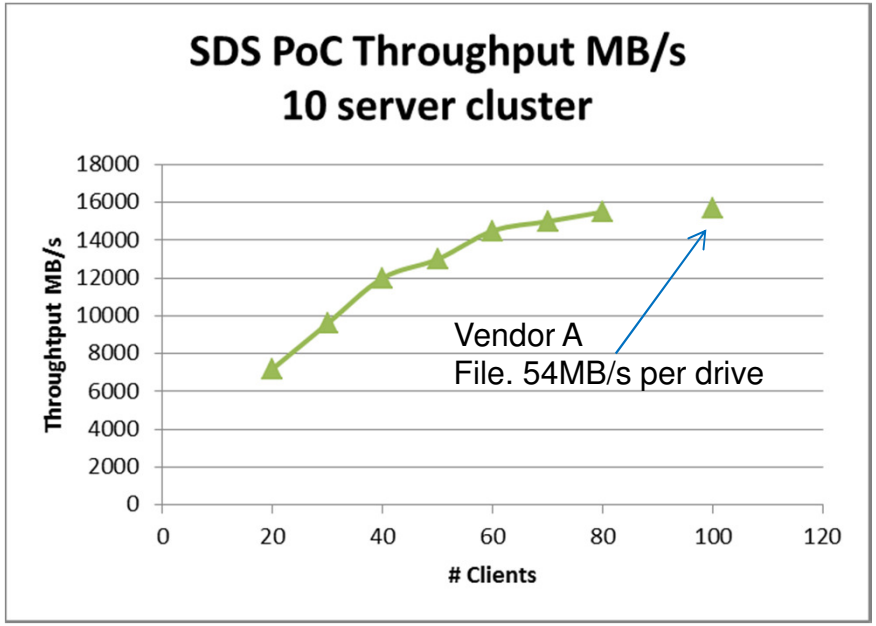
Science & Technology Facilities Council

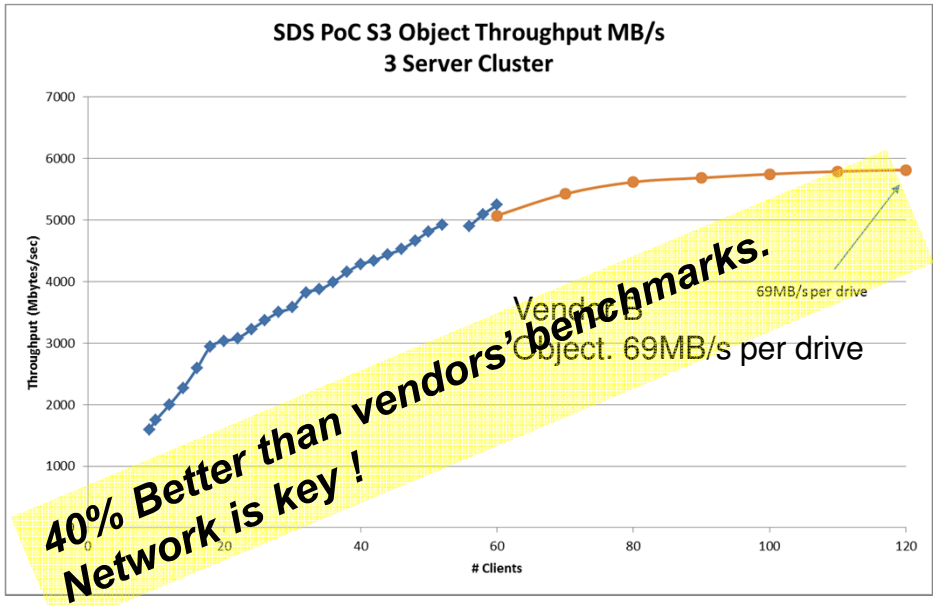
PoC Setup

- 8x Dell R730XD/MD1400
  - 30 drives / 4U (28x 8TB +/-SSD)
  - 64GB RAM, 20 Core, E5-2630v4 2.2GHz
  - Dual 40Gb Mellanox Connect-X3Pro NIC
- 2x HX645S-36i Supermicro
  - 36 drives (34x 6TB)
  - 128GB RAM, 12 Core, E5-2603v4 1.7GHz
  - Dual 40Gb Mellanox Connect-X3Pro NIC
- 2x R730xd (2x 100Gbit port networking)
  - “Connector” nodes if required.
- ~ 2Pbyte Raw Total over 10 nodes.

Science & Technology Facilities Council

Software Defined Storage PoC



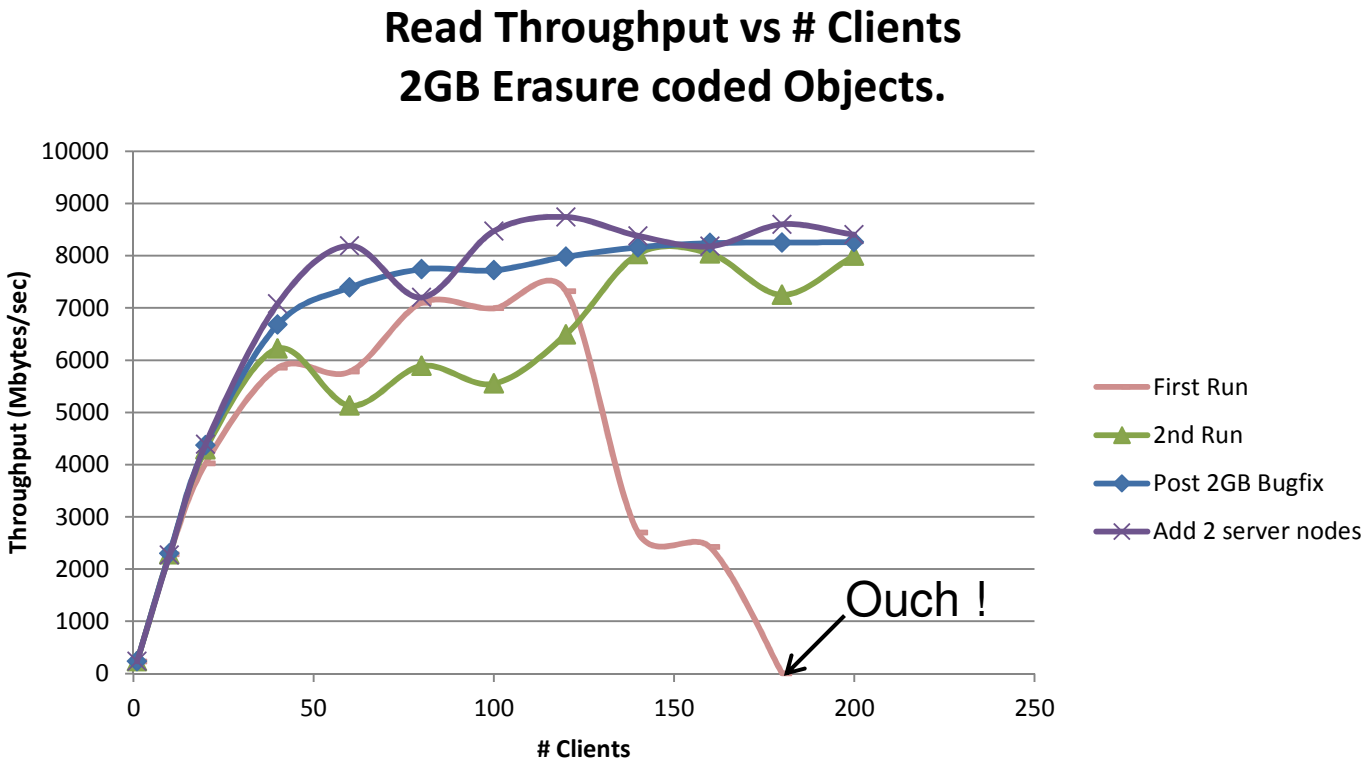


- Can we find file and object SDS comparable with PFS?
- 40Gb NIC / Chassis. No network bottlenecks.
- File 50-60MByte/sec/HDD
- S3 Object 60-70MByte/sec/HDD
- 27 HDD's / chassis = >>15Gbps
  - But 20-30Gbps with reconstruction/scrubs

Jonathan.Churchill@stfc.ac.uk STFC RAL

Science & Technology Facilities Council

Software Defined PoC – Is 2GB big or small Object ?

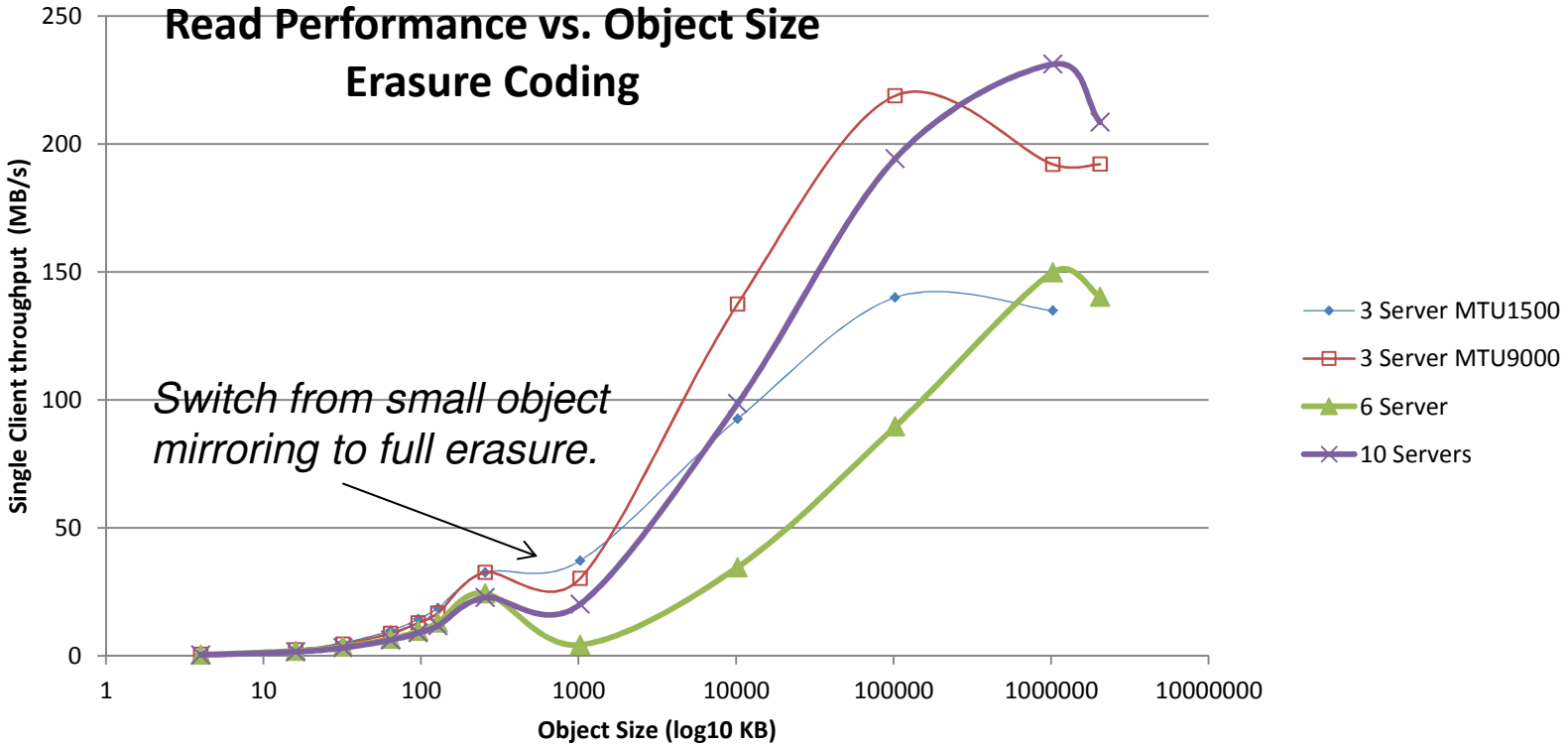


- Mismatched understanding of a “small file/object”
  - 2GByte vs 10-100MBytes

Jonathan.Churchill@stfc.ac.uk STFC RAL

Science & Technology Facilities Council

Software Defined PoC - Impact of Object size

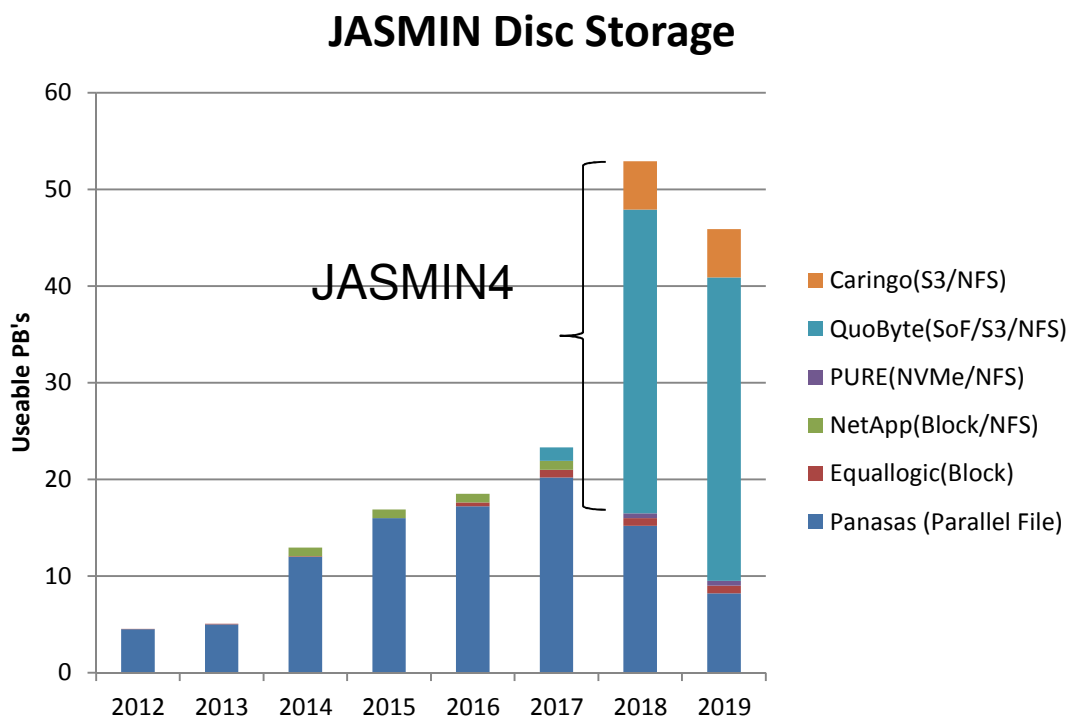


- ✓ Increases with MTU.
- ✓ Increases with Storage Cluster Size

Science & Technology Facilities Council

Jonathan.Churchill@stfc.ac.uk STFC RAL

JASMIN4 Disc Storage

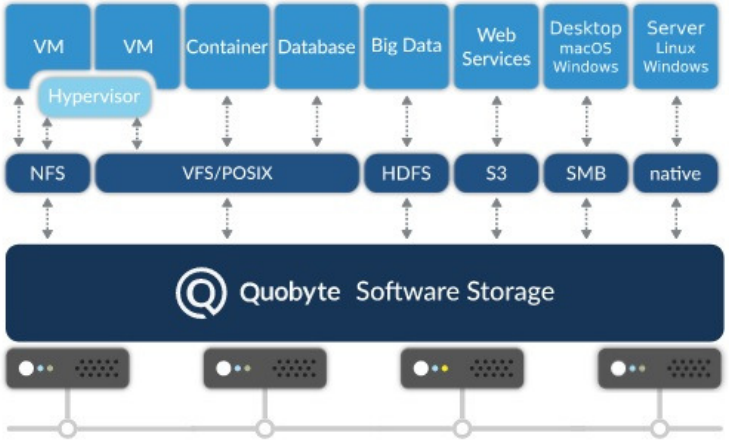


- No boundaries on data growth (or network topology)
- S3 interface to file and object system. RW Both sides.
- Performance comparable with 2014 Gen. Panasas PFS
- Online upgrades. Redundant networking.
- No client “call back” port.
  - Previous root /network and UMC restrictions

Jonathan.Churchill@stfc.ac.uk STFC RAL

Science & Technology Facilities Council

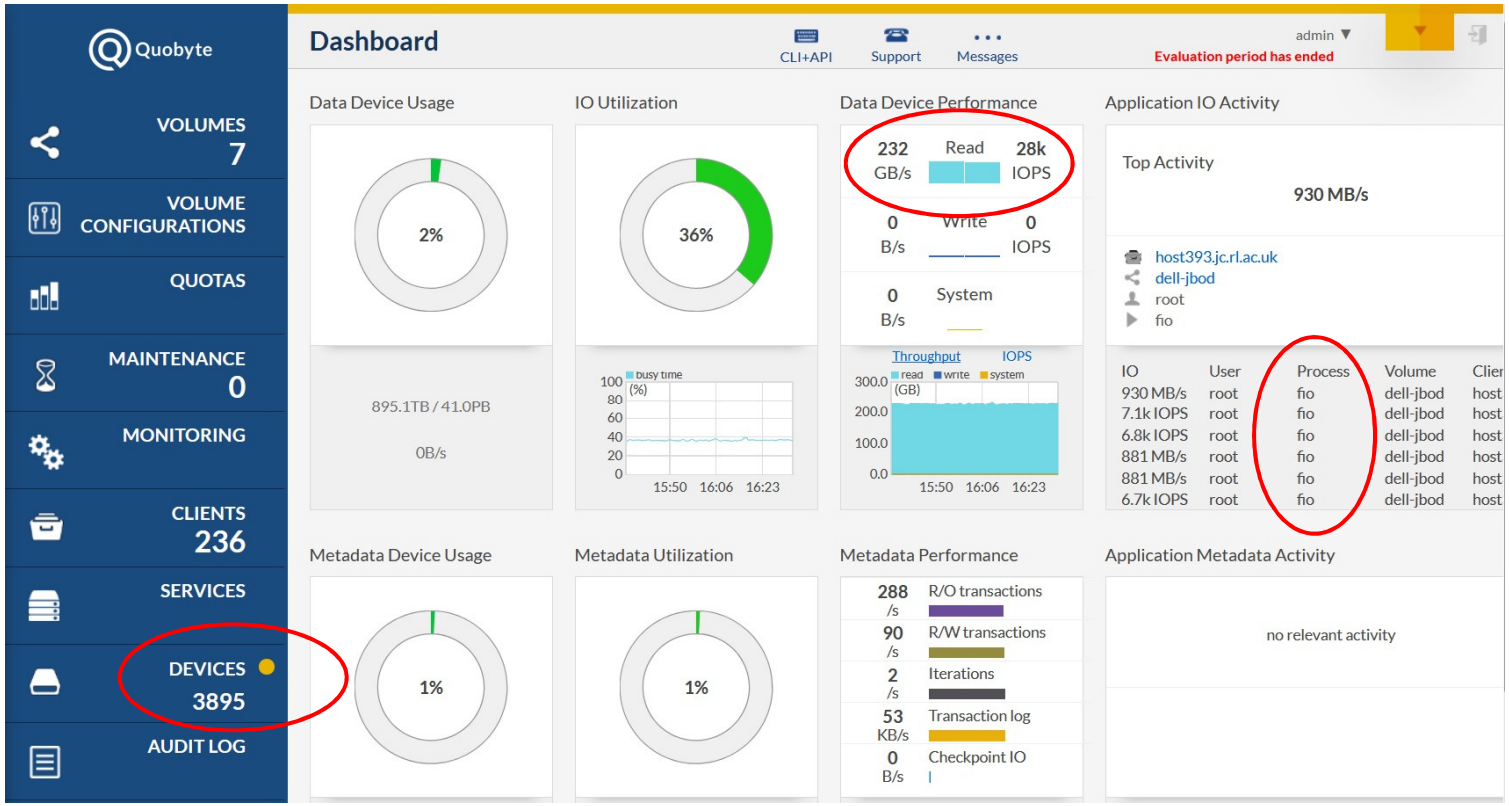
Quobyte @ JASMIN



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

- 42PB raw, 30PB usable (EC 8+3)
  - Largest world-wide capacity deployment
- Capacity split 50:50 Dell / Supermicro
- 47x R730xd's + MD3060 arrays (1 / server pair)
  - 32 10TB drives/ server, Dual 40Gb NICs
- 40x Supermicro 4U "Top loader" servers
  - 44 x 12TB drives, Dual 50Gb NICs
- Measured Read 220Gbytes/sec (fio) File access.
  - Theoretical of ~270Gbyes/sec ➔ 72MB/sec/drive

Quobyte @ JASMIN

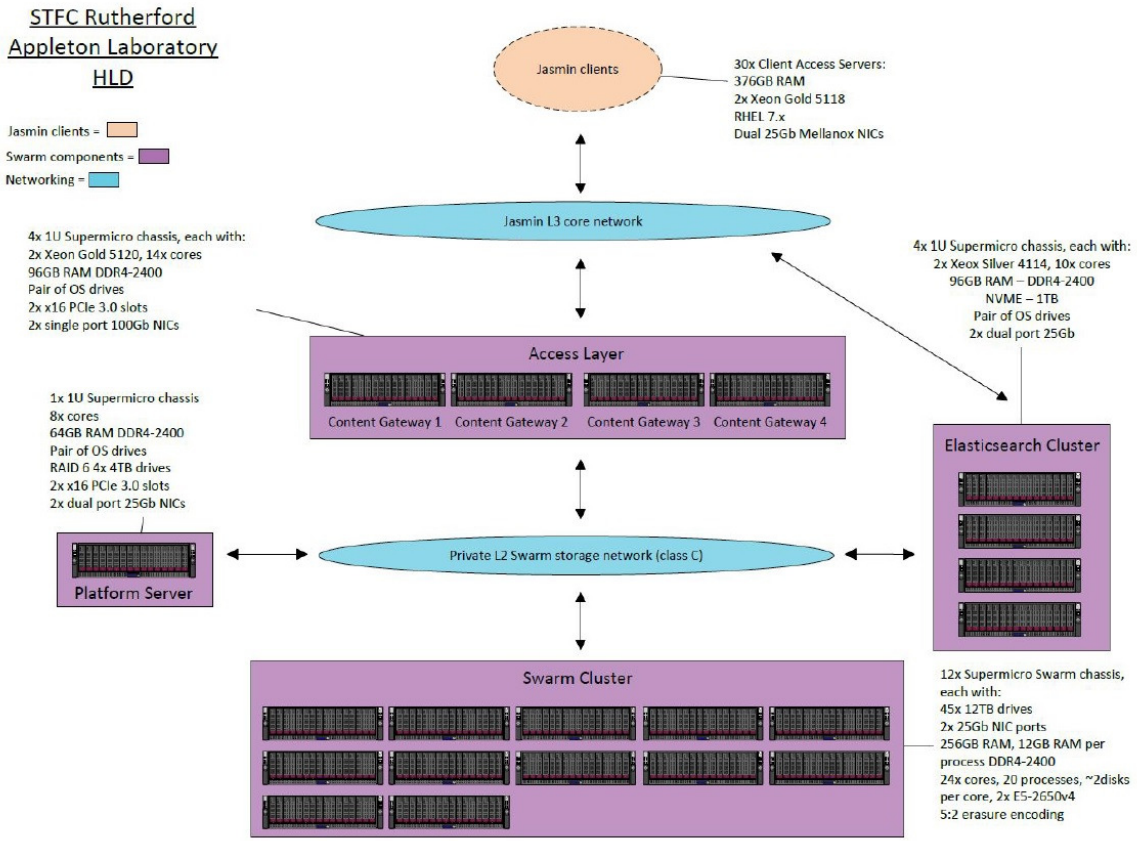


- 232GB/s ➔ >1.9 Terabit/sec network traffic.

Jonathan.Churchill@stfc.ac.uk STFC RAL

Science & Technology Facilities Council

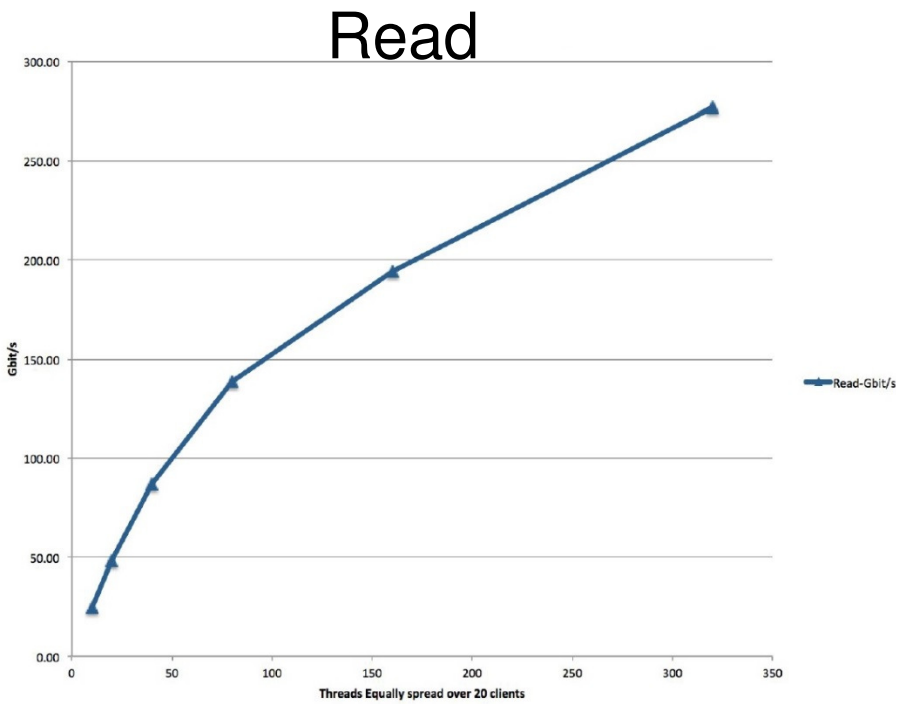
Caringo @ JASMIN

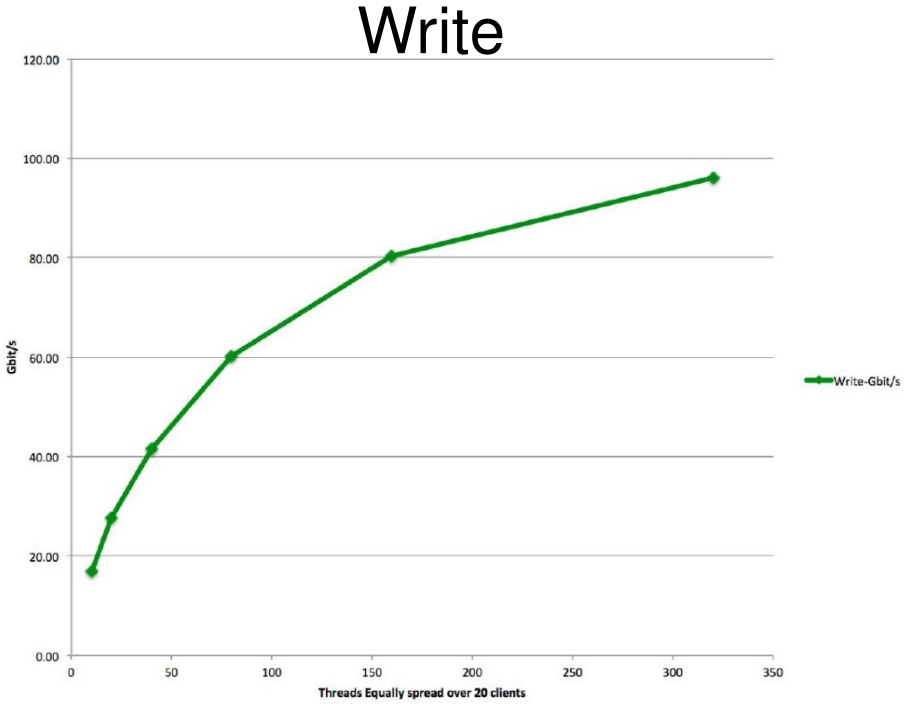


Science & Technology Facilities Council

Jonathan.Churchill@stfc.ac.uk STFC RAL

Caringo @ JASMIN



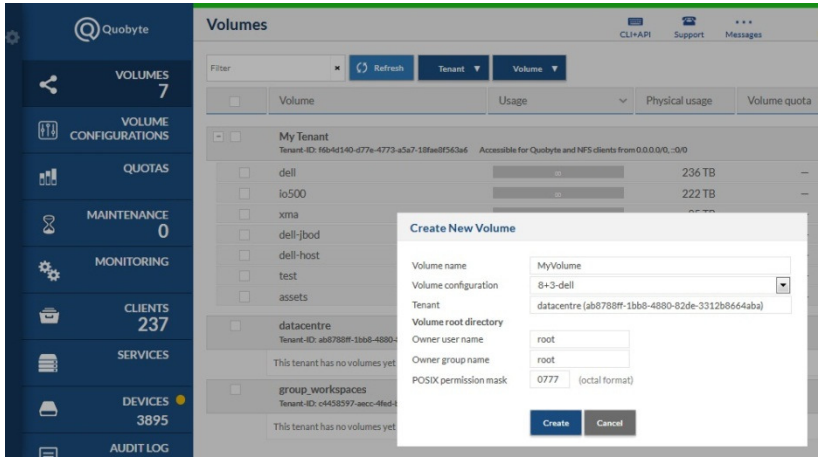


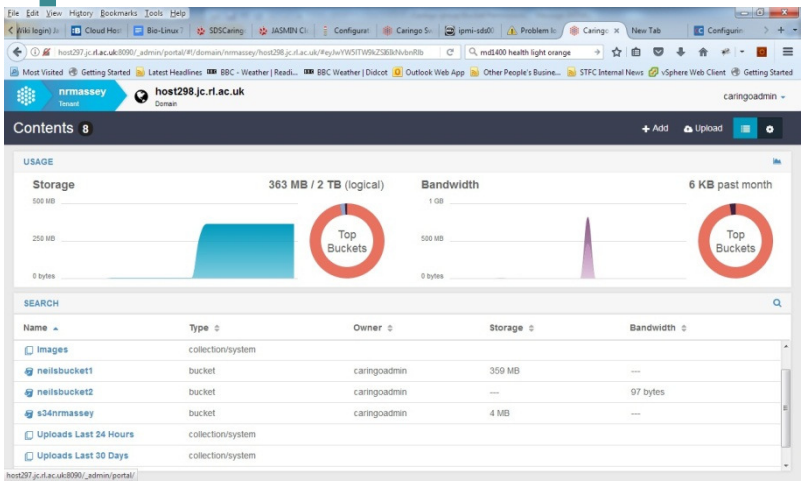
- Read 35Gbytes/sec ➔ ~64MB/s per drive
- Write 12.3Gbytes/sec
  - With more software optimisation ongoing.

Jonathan.Churchill@stfc.ac.uk STFC RAL

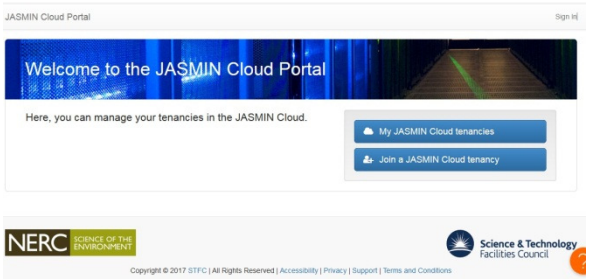
Science & Technology Facilities Council

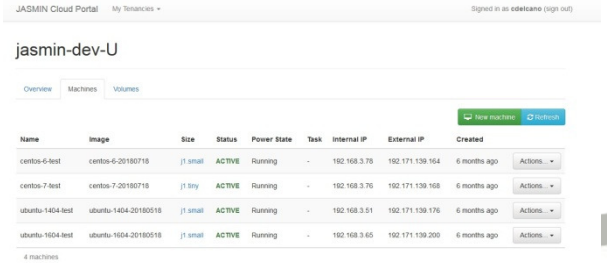
Storage and Compute Tenancies





- QuoByte File/S3 and Caringo S3/File User tenancies
  - Self provisioned, quota'ed storage.
- JASMIN OpenStack Cloud portal uses the same tenancies.

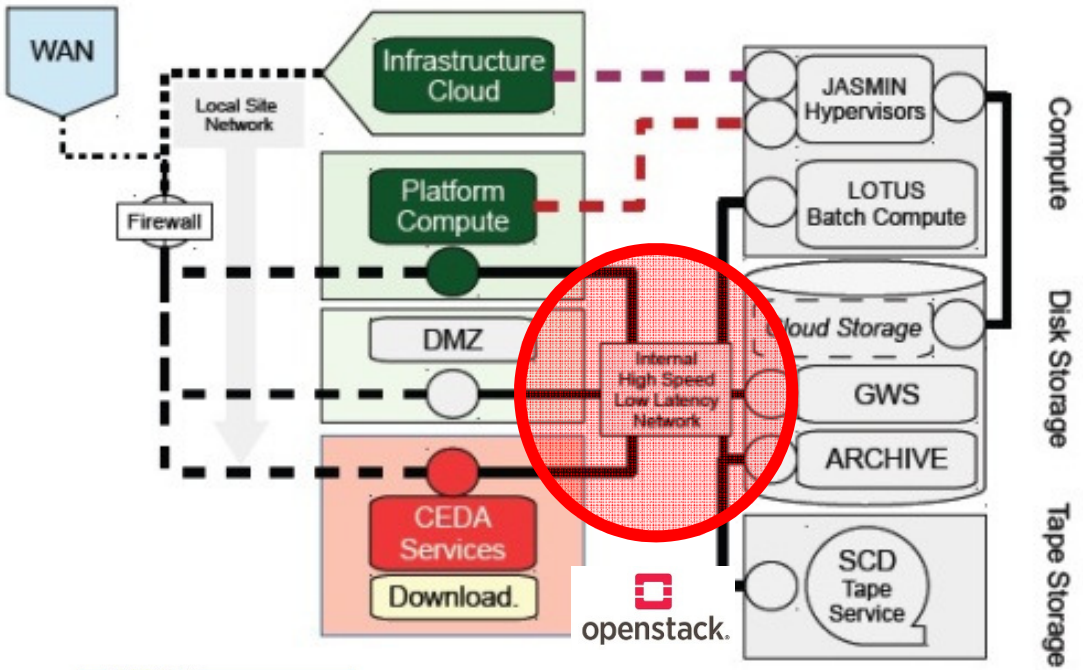


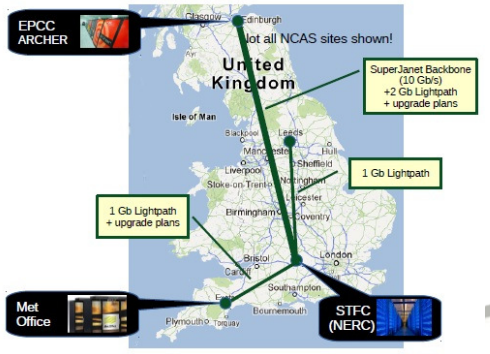


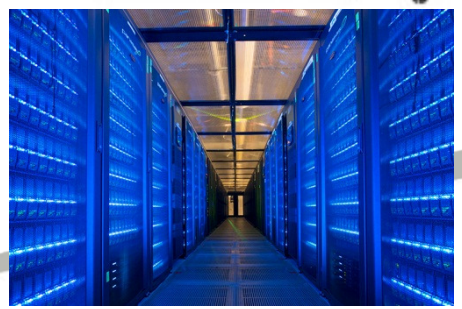
Jonathan.Churchill@stfc.ac.uk STFC RAL

Science & Technology Facilities Council

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







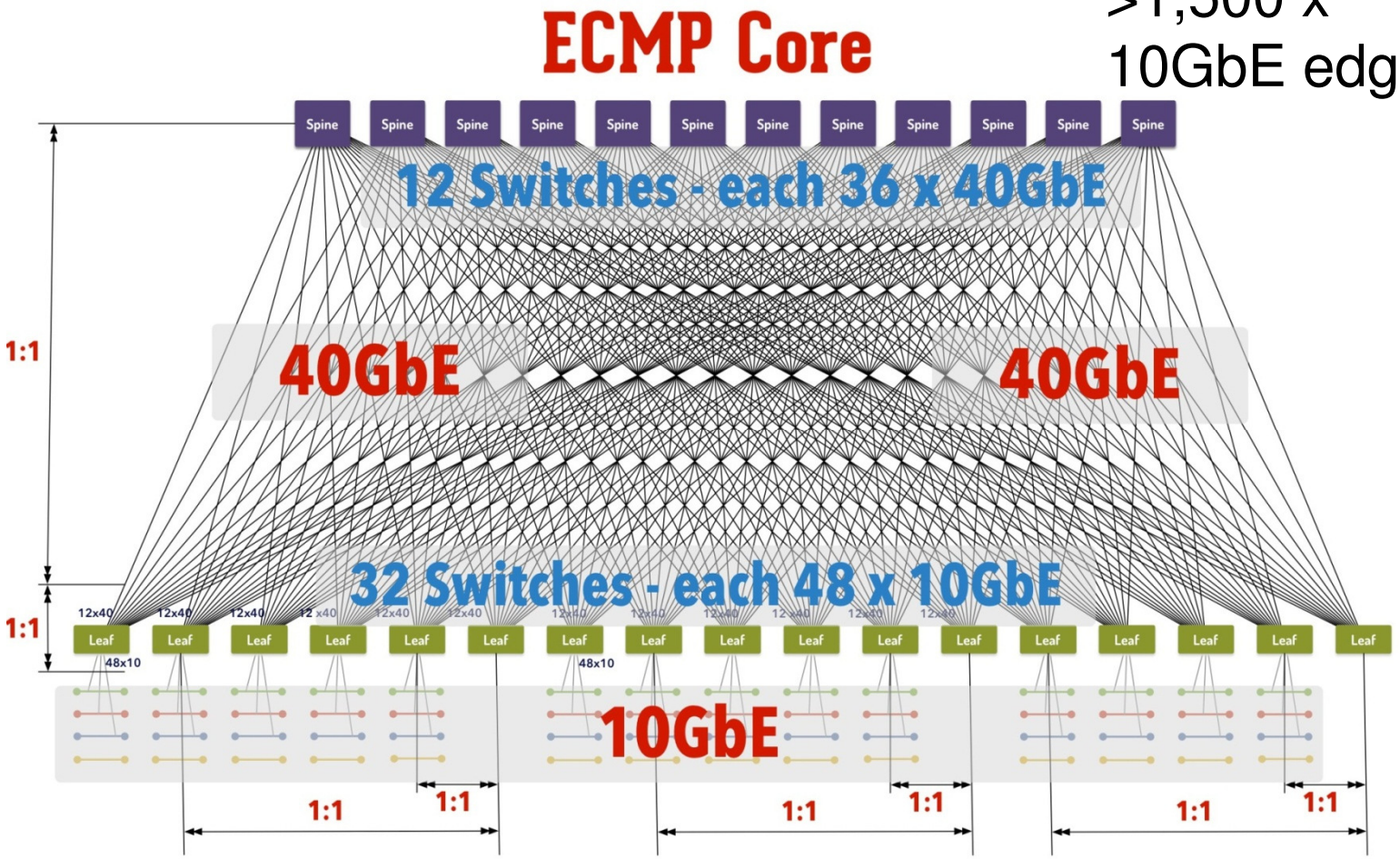
- 24.5 PB Panasas ~ 250GByte/s
- 44 PB Quobyte SDS ~ 220GBytes/s
- 5PB Caringo SDS Object
- 80PB Tape
- Batch HPC 6-10k cores
- Optical Private WAN + Science DMZ
- "Managed" VMware Cloud
- OpenStack Cloud
- PURE FlashBlade scratch/home
- Non-blocking ethernet > 12Tbit/sec

Science & Technology Facilities Council

Jonathan.Churchill@stfc.ac.uk STFC RAL

JASMIN 2,3 “Fabric” Network

>1,500 x
10GbE edge ports

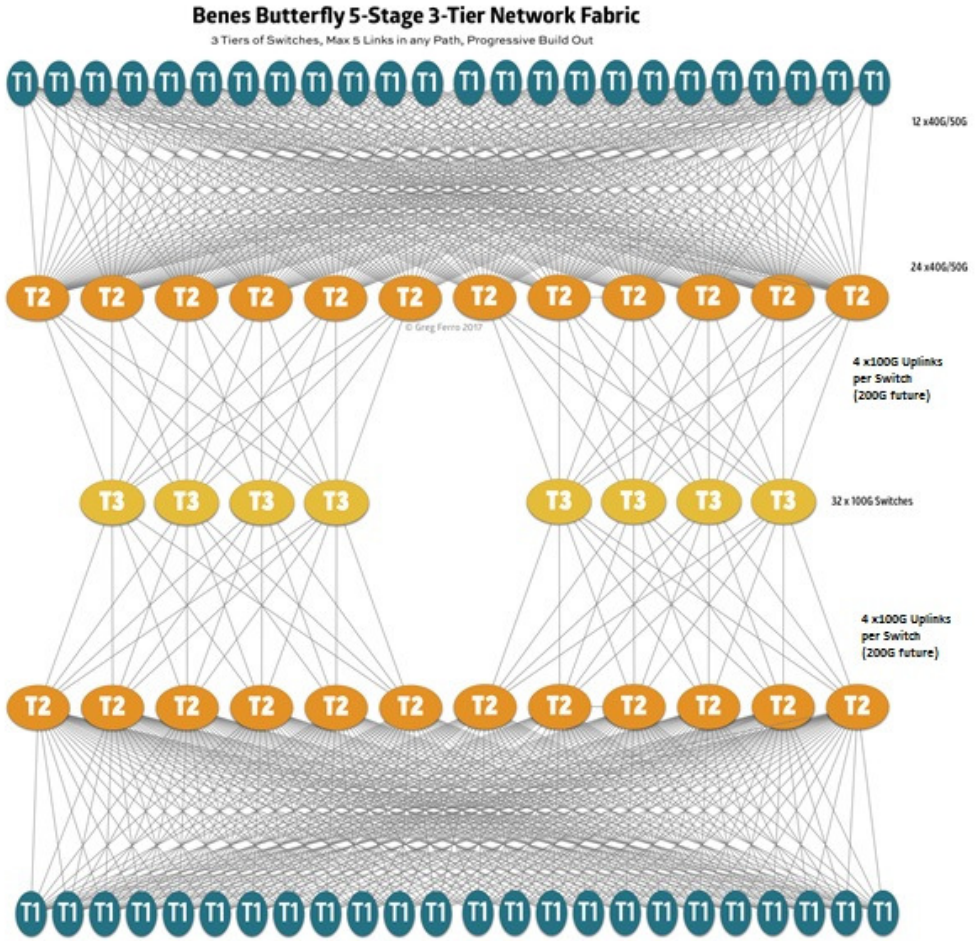


Any-any Low Latency:
< 7uS MPI ping-pong
(< 3uS with RoCEv3)

>12Terabits/sec
Non-blocking 1:1

Science & Technology Facilities Council

A Data Centre Network

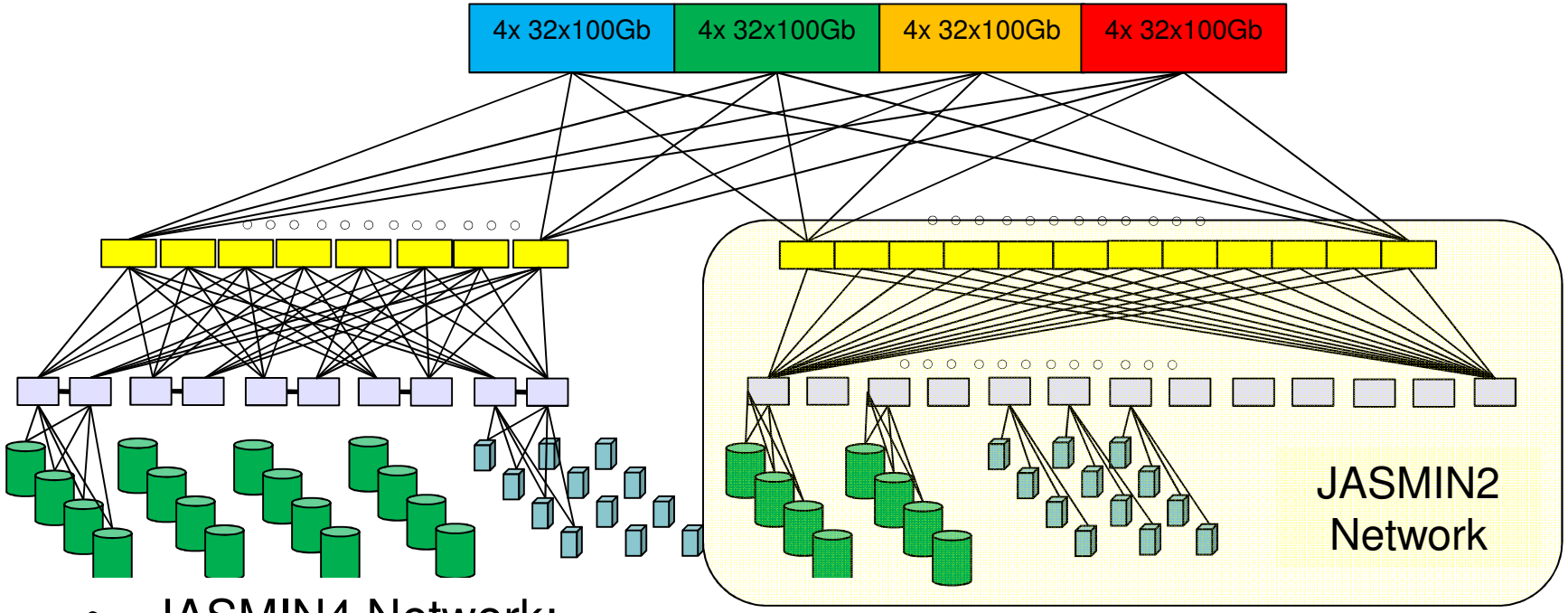


Facebook article for its three tier architecture:
https://code.facebook.com/posts/360346274145943/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/

Science & Technology Facilities Council

Connecting JASMIN2 to JASMIN4

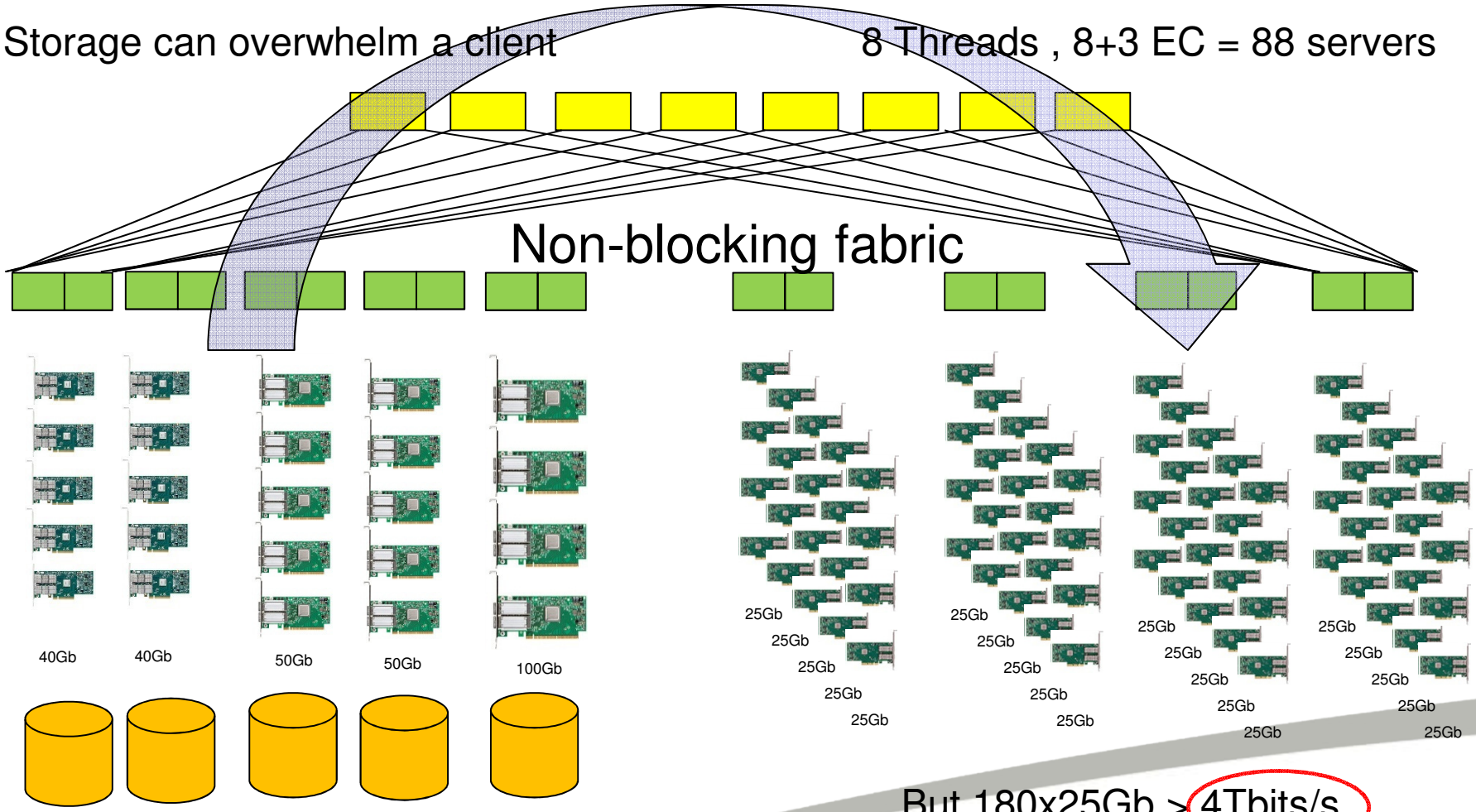
- Superspine: 16 Spines (32x 100Gb)
  - 4 Cluster/groups of 4 routers



- JASMIN4 Network:
  - 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’ leaf

  - 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



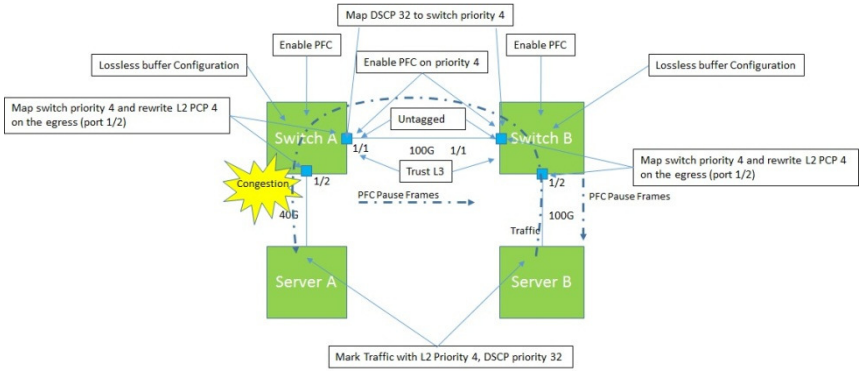
3090 HDD’s x 70MB/s > 250GBytes/sec
> 2Tbits/sec

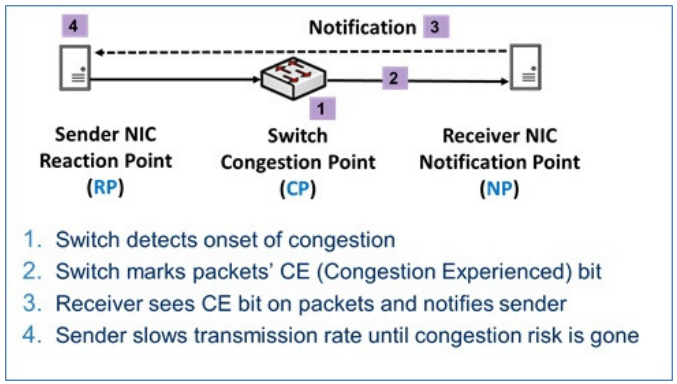
But 180x25Gb > 4Tbits/s

~200GB/s for a few minutes

Science & Technology Facilities Council

Controlling Packet Loss





DSCP + PFC

Distributed Services Code Point +Priority Flow Control

ECN

Explicit Congestion Notification

- DSCP/PFC Recommended for lossless networks eg RCoE/FCoE
  - But no RDMA / RCoE to prioritise for PFC
  - Complex to configure
- ECN proposed over a decade ago. Simpler to configure. Slow uptake.
  - Concerns expressed for high speed networks re:
  - HoL blocking, ECN propagation times/reaction times, buffer sizes
- ECN chosen
  - Simpler to configure, applies to all traffic
- **220GB/s sustained. Virtually no packet loss**

Images: http://www.mellanox.com/blog/2016/07/resilient-roce-relaxes-rdma-requirements/

Science & Technology Facilities Council

Summary

- Single H/W environment PoC @ PB scale key to procurement tech spec and expectation setting.
  - There be dragons !
- Some SDS (Object & File) systems approach PFS speed
  - With caveats !
- The network needs to match with bandwidth and latencies.
  - 5 (3) Tier Low latency Routed CLOS with > 12 (20) Tbps bw
  - Expect to implement some form of PFC.

- ***JASMIN :***
  On the road to high performance Object stores

Jonathan.Churchill@stfc.ac.uk STFC RAL

Thank you !

Science & Technology
Facilities Council