Implementing a scratch filesystem on E8 Storage NVMe

SSUG London 2019

Peter Childs

Update and Technical Deep Dive from talk originally given at SSUG@SC18 By Tom King

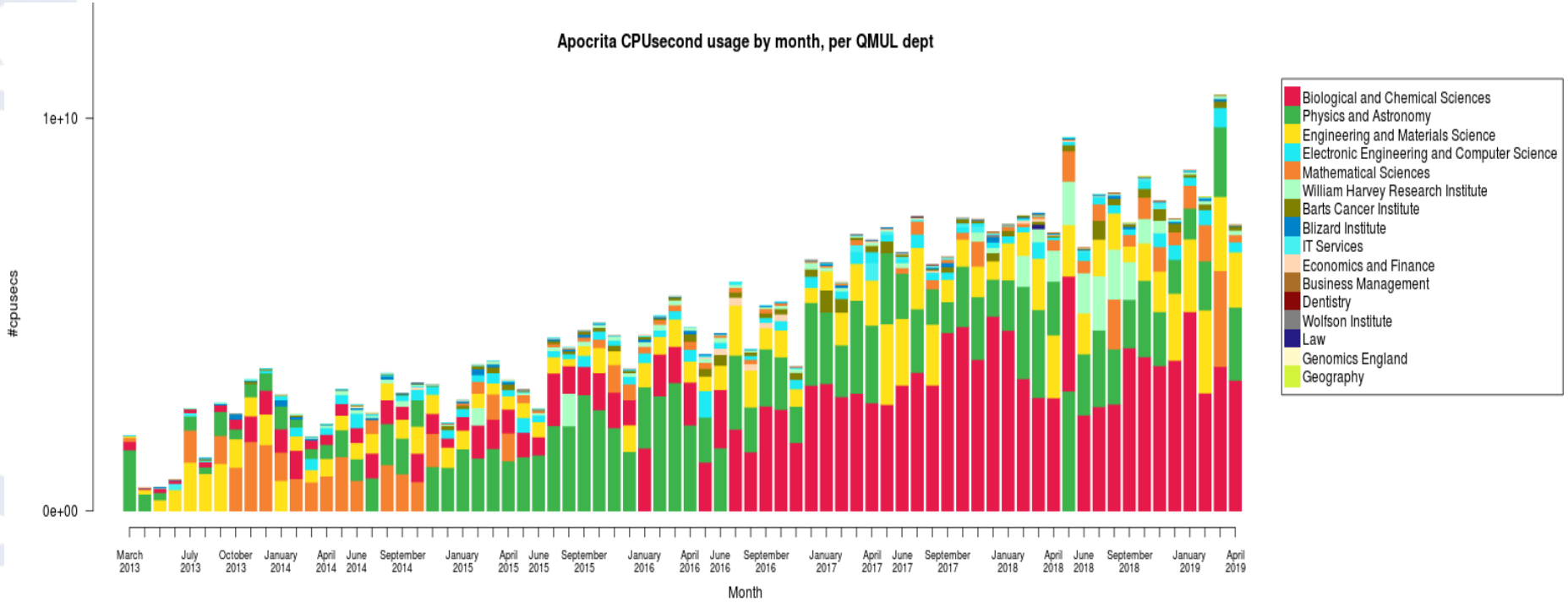
QMUL in numbers

- 25,000 students
- 4,500 staff
- £428m annual income (£144m research)
- 4 campuses in East and Central London
- Russell group member

- Science & Engineering
- Human & Social Sciences
- Medicine & Dentistry

- QMUL physicists discovered Proxima Centauri B
- Research into Tamoxifen breast cancer treatment
- Hosts Genomics England

HPC for all


Apocrita CPUsecond usage by month, per QMUL dept
#cpusecs
1e+10
0e+00
March 2013
July 2013
October 2013
January 2014
April 2014
June 2014
September 2014
January 2015
April 2015
June 2015
September 2015
January 2016
April 2016
June 2016
September 2016
January 2017
April 2017
June 2017
September 2017
January 2018
April 2018
June 2018
September 2018
January 2019
April 2019
Month
Biological and Chemical Sciences
Physics and Astronomy
Engineering and Materials Science
Electronic Engineering and Computer Science
Mathematical Sciences
William Harvey Research Institute
Barts Cancer Institute
Blizard Institute
IT Services
Economics and Finance
Business Management
Dentistry
Wolfson Institute
Law
Genomics England
Geography

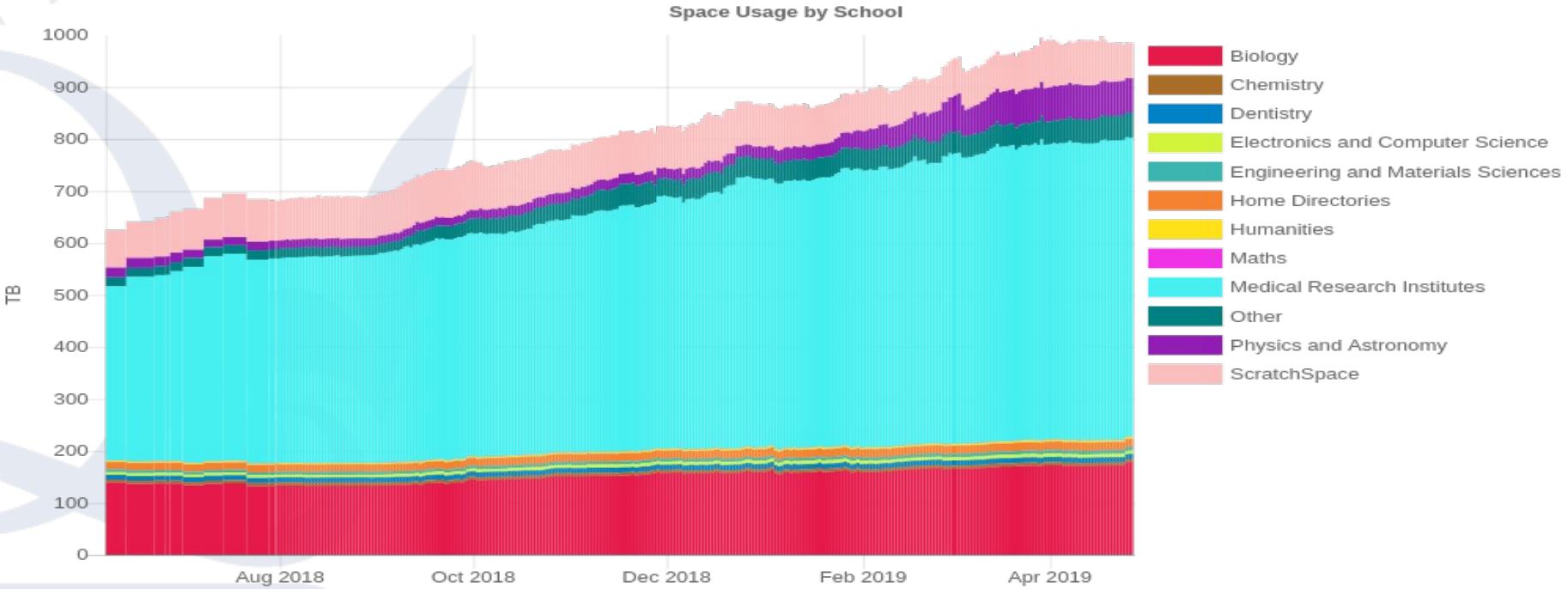
Old Spectrum Scale deployment

300 node Ethernet cluster running 4.2.3

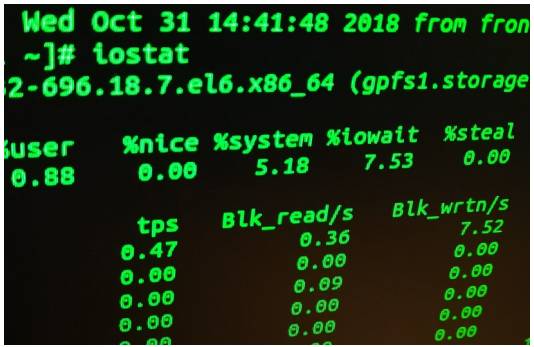
Scratch file system
0.5 PB over 180 spindles
Free but quota'd
Weekly/Monthly

Home file system
1.2 PB over 130 spindles
And 8 SSDs for MD
Chargeable

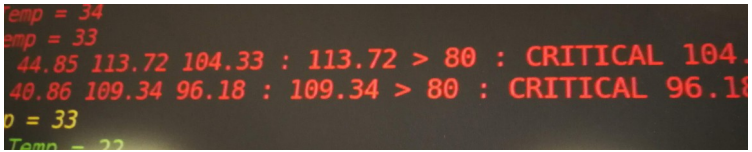
Storage Usage



Storage I/O intensive workloads



- Maker – genome annotation
- Canu – short sequence alignment tool
  FAQ includes “My run of Canu was killed by the sysadmin?”
- Guidance
- OpenMolcas

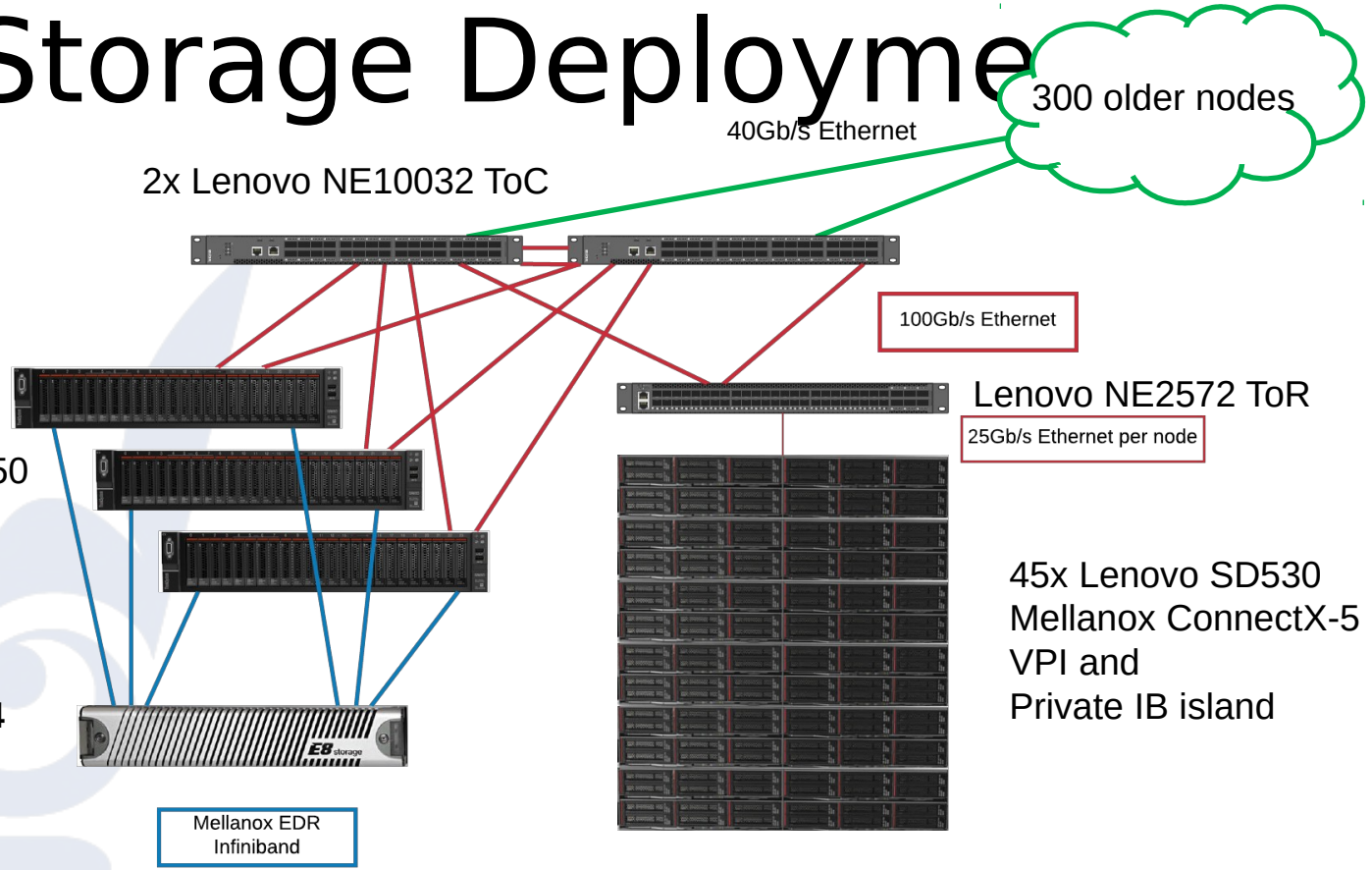


E8 Storage D24

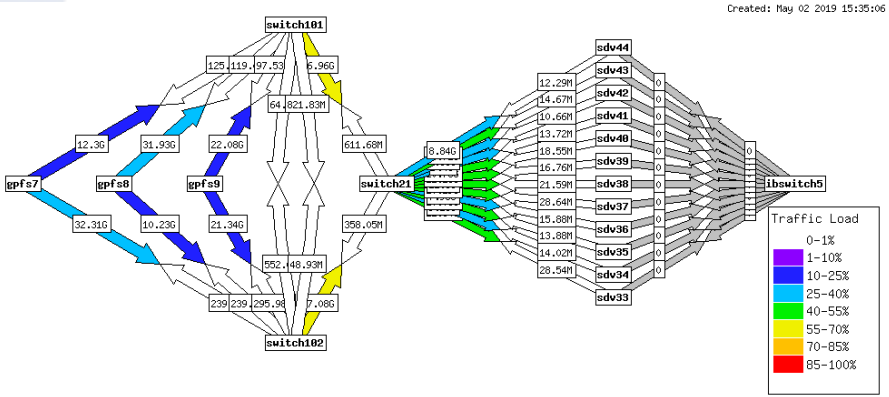
- 24x 6.4TB NVMe drives (3 dwpd)
- 8x 100Gb/s IB or Ethernet
- Dual controllers
- 40GB/s read maximum
- 20GB/s write maximum
- IPoIB for connection management
- Tracing kernel modules + E8 Storage daemon in user-space
- RAID calculations carried out on compute

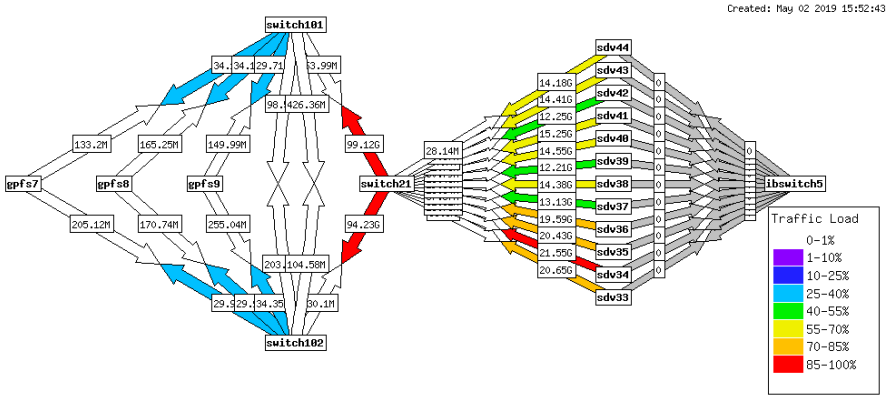
Gulp. We’re going to need a new core network!

E8 Storage Deployme



Weathermap during nsdperf – 12 node

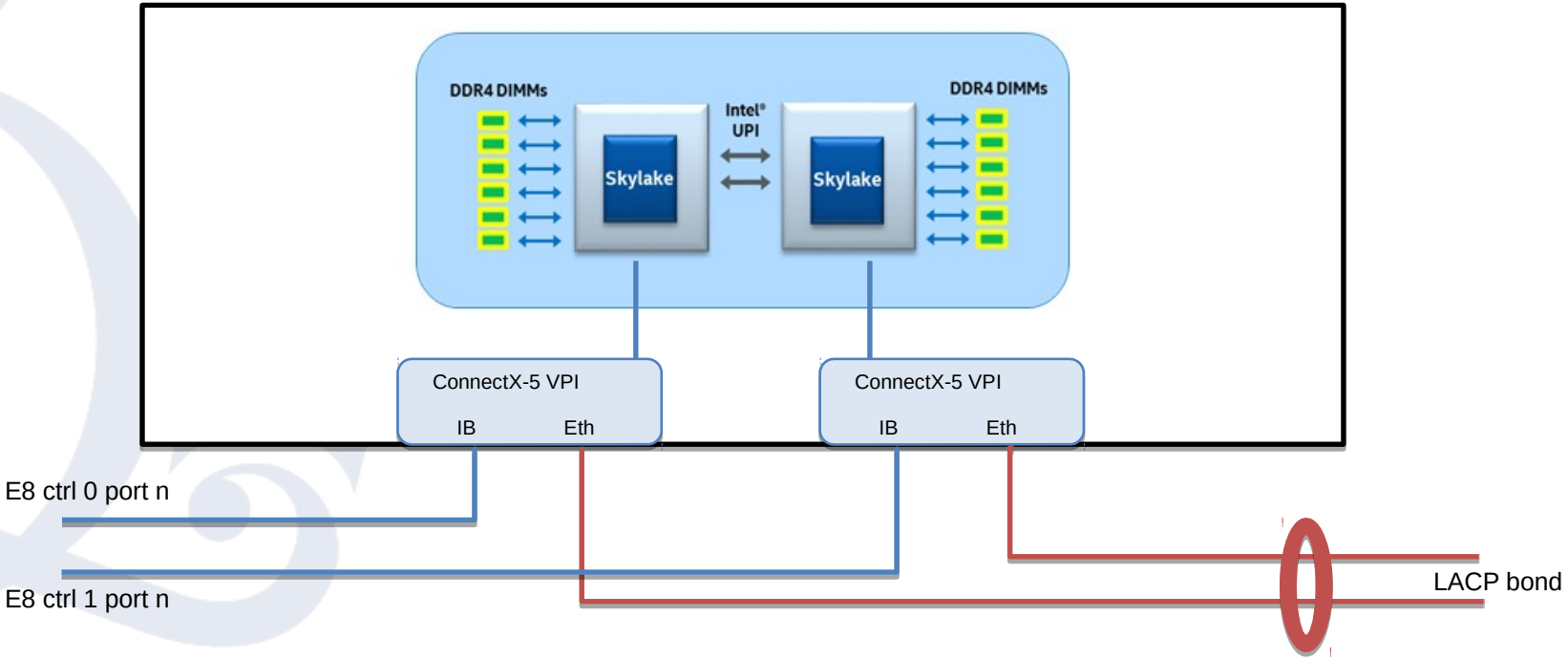




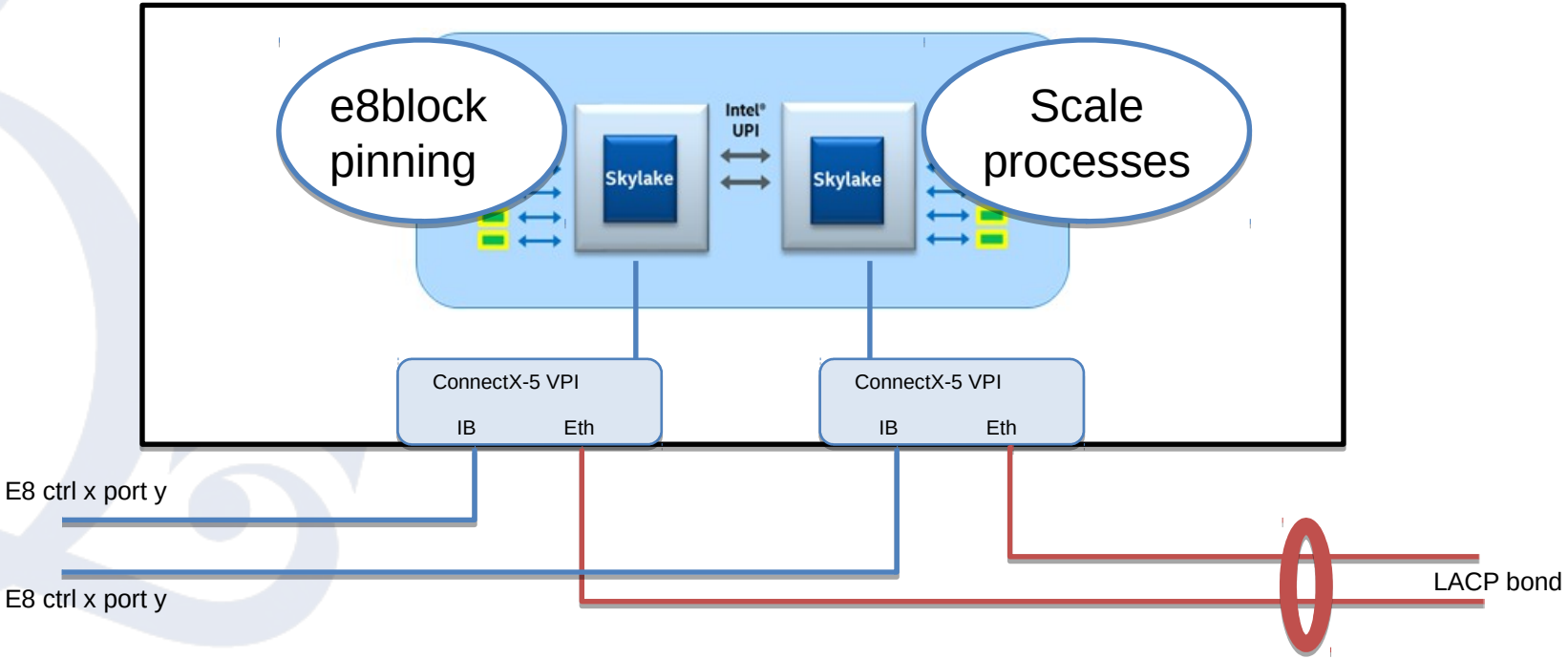
Spectrum Scale view

- 6 striped volumes presented as NSDs from the D24’s 24 devices
- Separate filesystem, not tiered with existing GS7K
- No separation of metadata
- Tests with block-size of 256KB up to 16MB
- Need to identify sweet spot trading off block-size against wastage in SS4

NSD server setup



NSD server setup



IO500

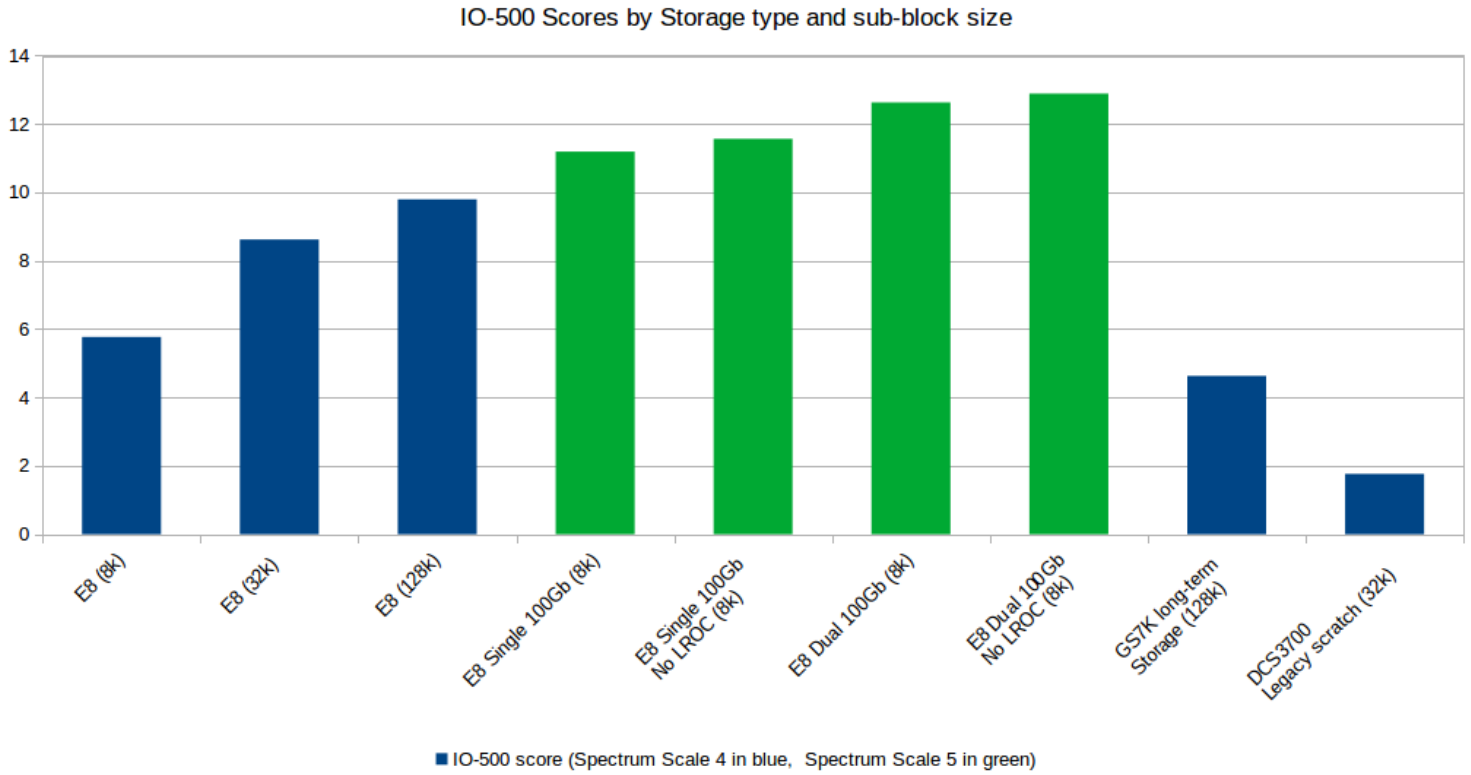
- https://www.vi4io.org/io500/start
- Standard Test of Storage.
  - Weighted Average of different tasks storage does.
- 10 Node Challenge
  - QMUL E8 Storage is 8th
  - Latest results using Spectrum Scale 5 would put us in 6th

Results

	IO-500 b/w 10 nodes	IO-500 iops 10 nodes	IO-500 Score 10 nodes
Spectrum Scale 4.2.3.0 with LROC 40GB on SSD			
E8 Storage (256KB)	1.6 GB/s	20.0 kiops	5.69
E8 Storage (1MB)	3.5 GB/s	20.8 kiops	8.62
E8 Storage (4MB)	4.3 GB/s	22.2 kiops	9.79
Spectrum Scale 5.0.2-1			
E8 Storage (SS5)	5.1 GB/s	28.3 kiops	12.63
E8 Storage (SS5 No LROC)	5.4 GB/s	30.7 kiops	12.89
cf.			
IBM DCS 3700	0.4 GB/s	7.8 kiops	1.76

Peak Results IO-500
10 node challenge

Test	DCS3700 (SS4)	GS7K (SS4)	E8 (SS5)
IOR Easy Write (GB/s)	0.568	4.344	10.817
IOR Easy Read (GB/s)	1.560	5.008	12.580
Mdtest Easy Stat (kiops)	8.276	72.315	62.094
Mdtest easy Delete (kiops)	5.639	18.572	40.308


IO-500 Scores by Storage type and sub-block size
14
12
10
8
6
4
2
0
E8 (8k)
E8 (32k)
E8 (128k)
E8 Single 100Gb (8k)
E8 Single 100Gb No LROC (8k)
E8 Dual 100Gb (8k)
E8 Dual 100Gb No LROC (8k)
GS7K long-term Storage (128k)
DCS3700 Legacy scratch (32k)
IO-500 score (Spectrum Scale 4 in blue, Spectrum Scale 5 in green)

Further work done

- Upgrade to Spectrum Scale 5.0.2-1 (Now nearly complete)
  - This has removed the need to worry about block size
  - Numerous code changes makes the storage work faster even on our long term disk based storage.
- Addition of correct network cables, our initial results were limited by cables originally shipped not working.
- Network Bond balancing (using of Layer 3+4, rather than layer 2)
- Need for monitoring writes per day – We've been stung in the past and we're now doing this

Bio tool results

Spectrum Scale 5.0.2-1 – single node – run time (sec)

	DCS 3700	GS7K	E8 D24	Local SAS SSD ext4
Maker	4335	2374	2136	1989
GuidanceIO	171.37	153.51	147.98	132.36

Further Gains have been small, as the pressure moves from the io system onto the CPU.
Biggest gains seen by just having metadata on SSD.

Auto-deleting Scratch

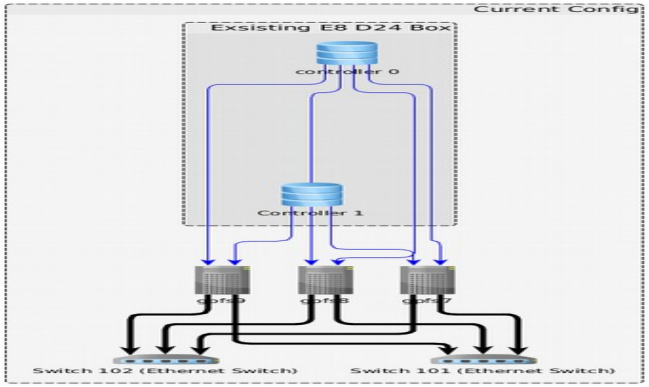
- Users don’t delete data unless forced
- Need to ensure a fair share usage policy while allowing people to handle large amounts of data
- Standard Quotas don’t work as while users will tidy up their own space to reuse it they won’t once they leave
- Encourage users to ensure backups of important data are done, rather than leaving it scratch

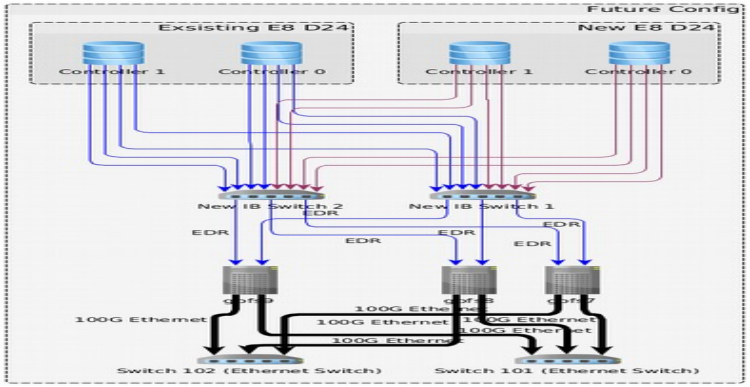
- Use of Spectrum Policies to check age of files and delete accordingly using time stamps
- Need to delete old directories (Which isn’t done by default with a delete policy)
- Sending email from policies before the files are actually deleted

- If people are interested, I can open our code.

Current Plans

- Doubling capacity,
- add pair of IB switches
  - Aid Future Expansion
  - Standard Solution





Conclusions

- Good Performance
- Network Limited
- Now in service
- Tested and Benchmarked
- Spectrum Scale 5 is significantly faster than previous versions
- Planning to expand it shortly

Thanks

- Tom King, QMUL
- Peter Childs, QMUL
- Dr. Chris Walker, QMUL
- Sammie Frisch, E8 Storage
- Stuart Campbell, E8 Storage
- OCF team