

Implementing a scratch filesystem on E8 Storage NVMe

SSUG@SC18

Tom King

QMUL in numbers

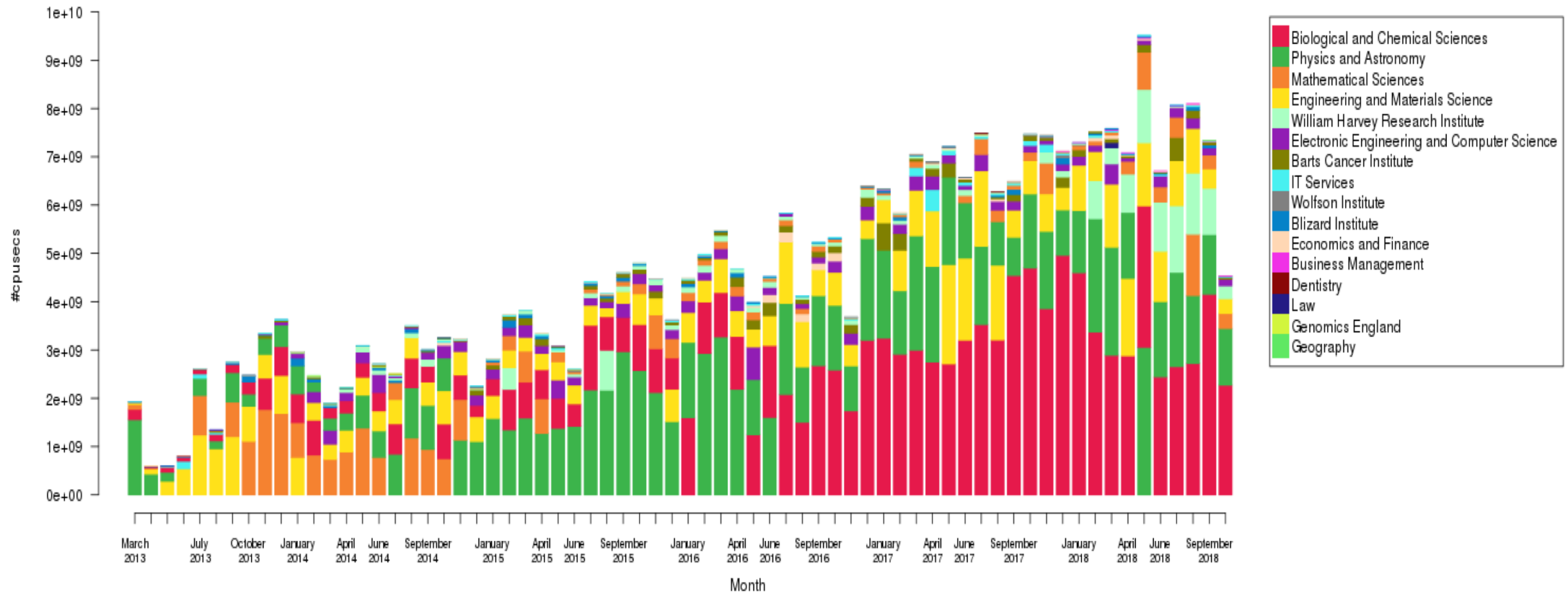
- 25,000 students
- 4,500 staff
- £428m annual income (£144m research)
- 4 campuses in East and Central London
- Russell group member



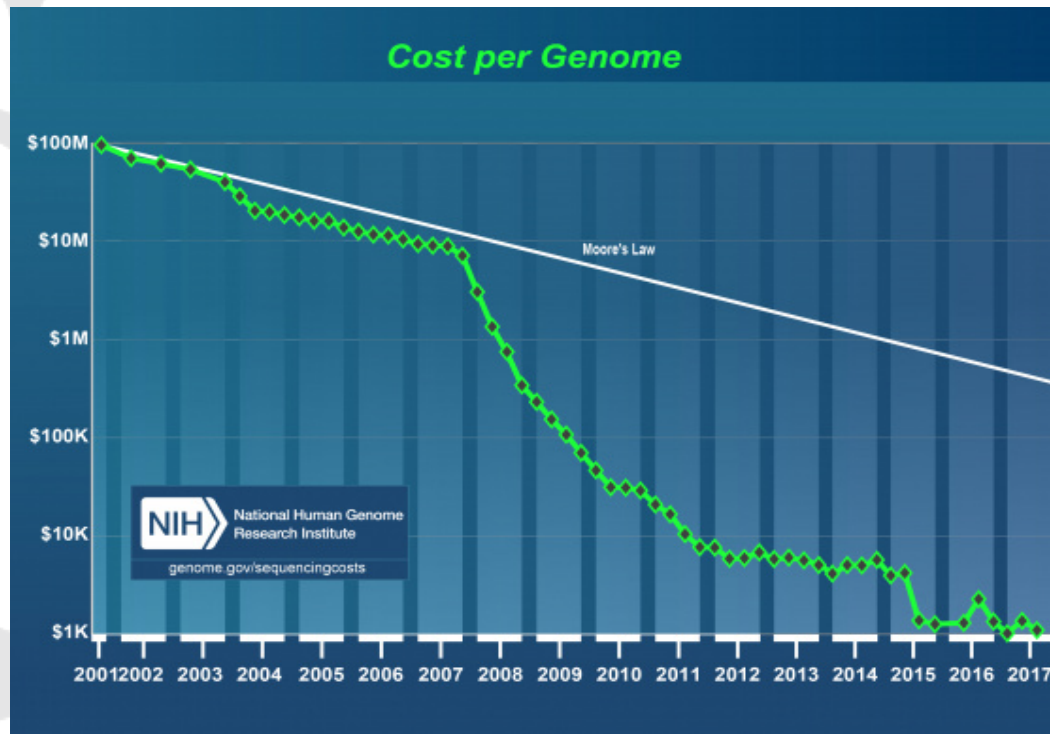
- Science & Engineering
- Human & Social Sciences
- Medicine & Dentistry
- QMUL physicists discovered Proxima Centauri B
- Research into Tamoxifen breast cancer treatment
- Hosts Genomics England

HPC for all

Apocrita CPUsecond usage by month, per QMUL dept



Sequencing for all



Backblaze Average Cost per Drive Size

By Quarter: Q1 2009 - Q2 2017



BACKBLAZE

Current Spectrum Scale deployment

300 node Ethernet cluster

IBM DCS3700

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

DDN GS7K

Home file system
1.2 PB over 190 spindles
Chargeable

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

```
Wed Oct 31 14:41:48 2018 from from
~]# iostat
2-696.18.7.el6.x86_64 (gpfs1.storage

user      %nice    %system  %iowait  %steal
0.88      0.00     5.18     7.53     0.00

          tps      blk_read/s  blk_wrtn/s
0.47      0.36      7.52
0.00      0.00      0.00
0.00      0.09      0.00
0.00      0.00      0.00
0.00      0.00      0.00
0.00      0.00      0.00
```

```
emp = 34
emp = 33
44.85 113.72 104.33 : 113.72 > 80 : CRITICAL 104.33
40.86 109.34 96.18 : 109.34 > 80 : CRITICAL 96.18
p = 33
Temp = 22
```

E8 Storage D24

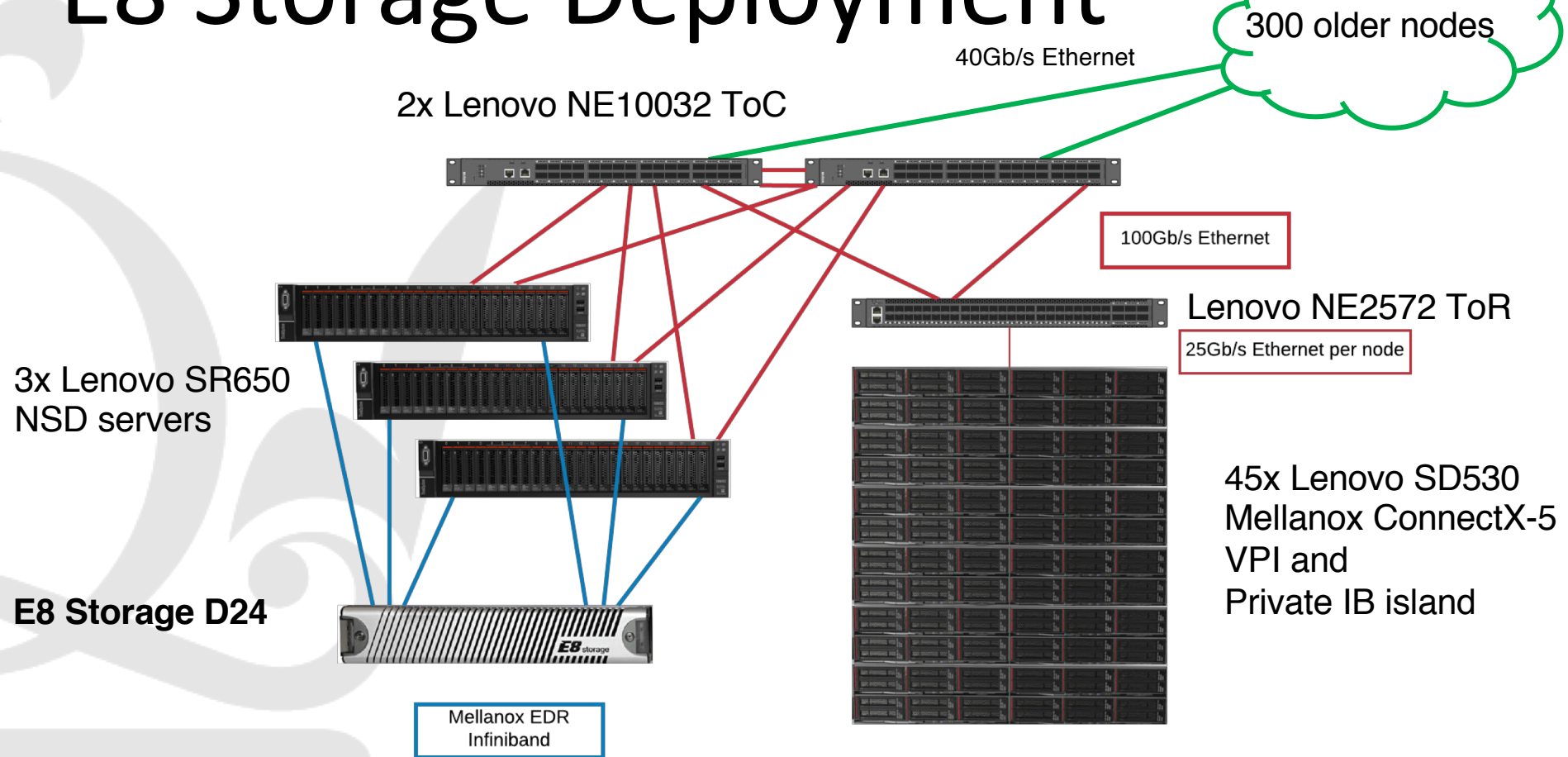
- 24x 6.4TB NVMe drives (3 dwpd)
- 8x 100Gb/s IB or Ethernet
- Dual controllers
- 40GB/s read maximum
- 20GB/s write maximum
- IPoIB for connection management
- Tracing kernel modules + E8 Storage daemon in user-space
- RAID calculations carried out on compute

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

NVMe

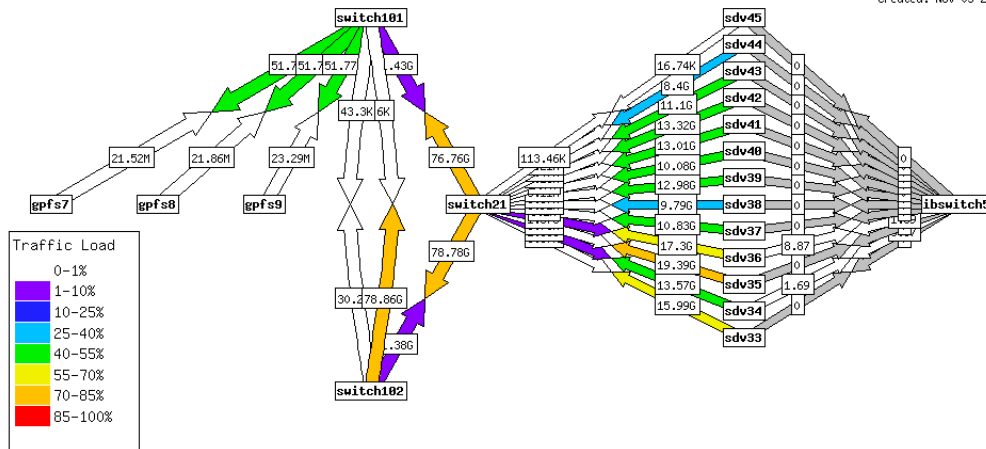
- Commercially available since 2014
- Straight on to PCIe bus avoiding the overhead of “legacy” storage buses
- ~x10 IOPS increase compared with SATA
- Still reliant S/M/TLC NAND
- Limited lifetime
- Need to assess and monitor writes per day
- NVMeOF – fabric
- Mellanox IB providing lowest latency fabric

E8 Storage Deployment

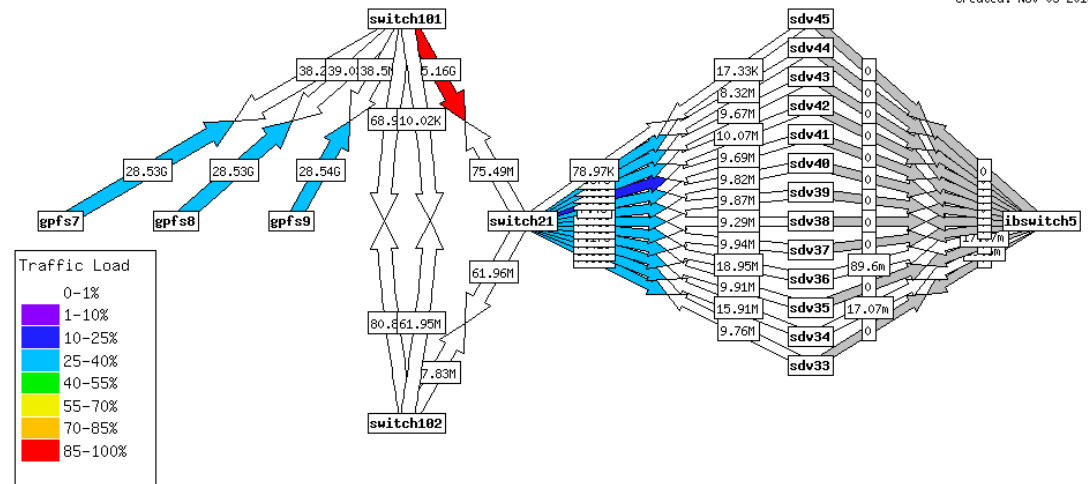


Weathermap during nsdperf – 12 node

Created: Nov 05 2018 10:15:11



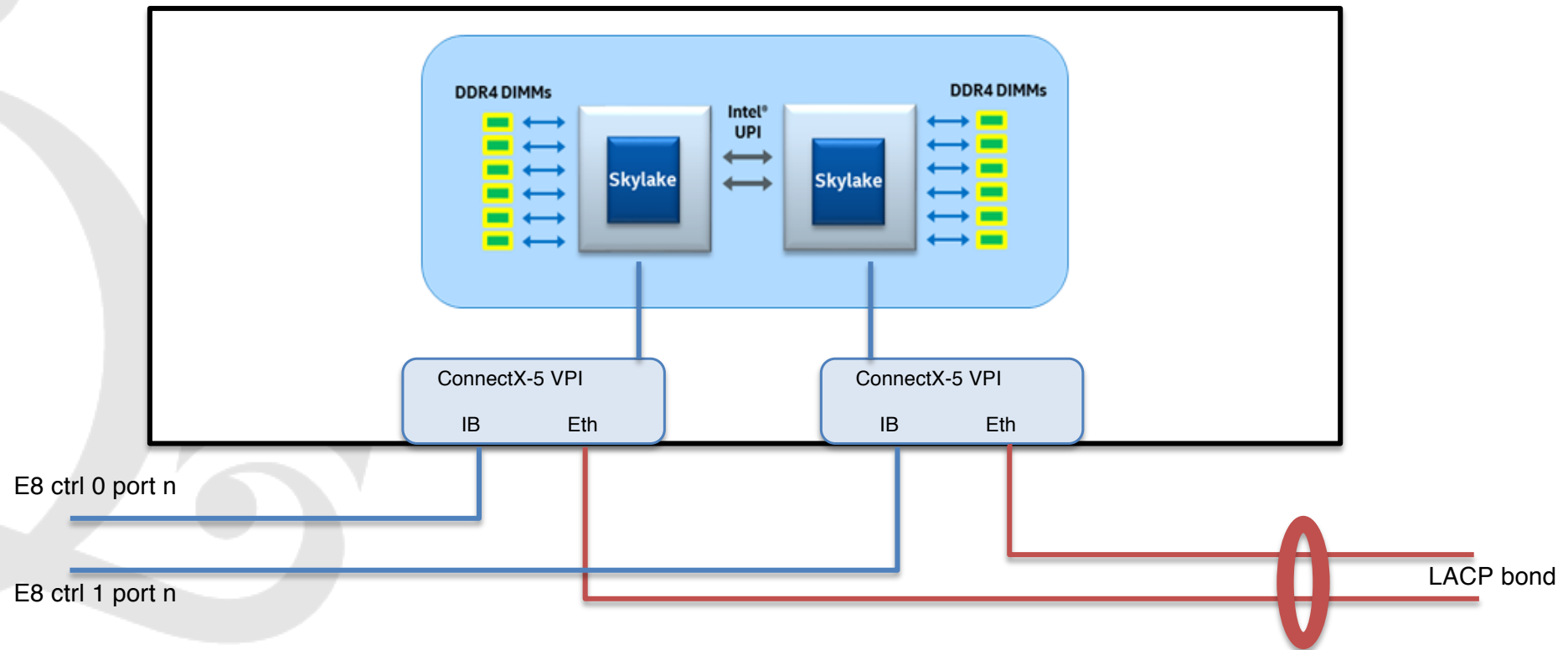
Created: Nov 05 2018 10:30:11



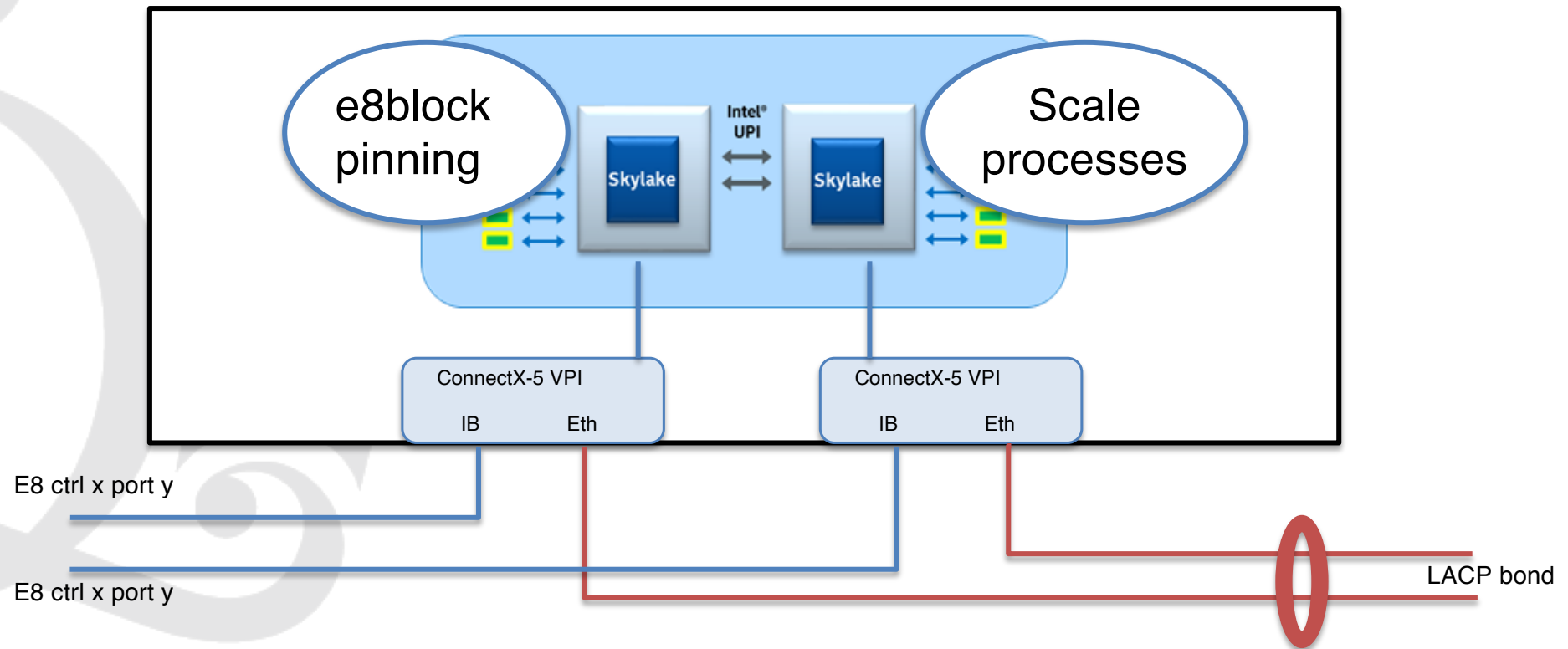
Spectrum Scale view

- 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



First results

Spectrum Scale 4.2.3.0 with LROC 40GB on SSD

	IO-500 b/w 10 nodes	IO-500 iops 10 nodes	fio b/w read 1n/16t	fio b/w write 1n/16t
E8 Storage (256KB)	1.6 GB/s	20.0 kiops	- GB/s	- GB/s
E8 Storage (1MB)	3.5 GB/s	20.8 kiops	9.9 GB/s	4.4 GB/s
E8 Storage (4MB)	4.3 GB/s	22.2 kiops	11.4 GB/s	5.7 GB/s
E8 Storage (16MB)*	-	-	-	-
cf.				
IBM DCS 3700	0.4 GB/s	7.8 kiops	1.1 GB/s	0.3 GB/s

Bio tools results

Spectrum Scale 4.2.3.0 – single node – run time (sec)

	DCS 3700	GS7K	E8 D24	Local SAS SSD
Maker (LROC)	4335	2374	2407	1989
Maker (no LROC)	4239	2363	2113	1935

Conclusions

- Good performance observed so far
- Expect that further tweaks are possible before go-live to users

Further developments

- Re-run with Scale 5.0.2-1 and with HAWC
- Test with the full network capacity/cables
- Need for monitoring writes per day – I've been stung in the past
- Doubling of capacity through additional VPI cards in NSD servers
- Further expansion - IB switches and blocking topology or more NSD servers
- Offer Scale over IB to reduce latency to newer nodes
- Evaluate NSD servers with Cascade Lake and Apache Pass for meta-data

Thanks

- Peter Childs, QMUL
- Dr. Chris Walker, QMUL
- Sammie Frisch, E8 Storage
- OCF team