UNIVERSITY OF
BIRMINGHAM

Scale out storage systems to support research and cloud

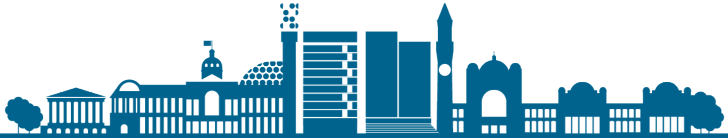
Simon Thompson, Research Computing Infrastructure Architect
IT Services

BEAR

BIRMINGHAM ENVIRONMENT
FOR ACADEMIC RESEARCH

□ Services free at point of use

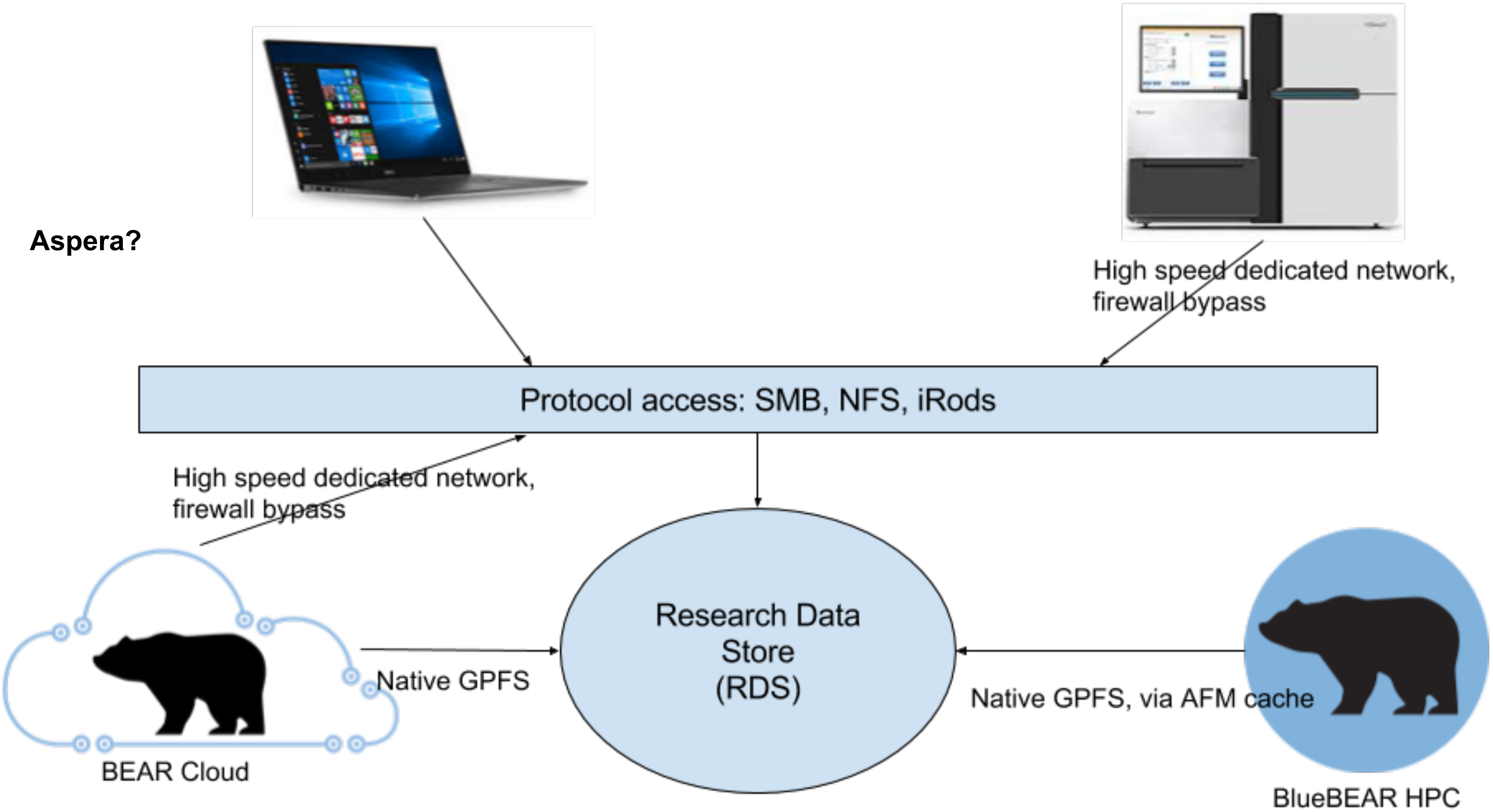
□ Across all research disciplines



BEAR Services

- HPC (BlueBEAR)
- Private cloud (BEAR Cloud, CLIMB)
- Research Data Storage and Archive
- High speed research networking
- Data Visualisation

Aspera?



High speed dedicated network, firewall bypass

Protocol access: SMB, NFS, iRods

High speed dedicated network, firewall bypass

Research Data Store (RDS)

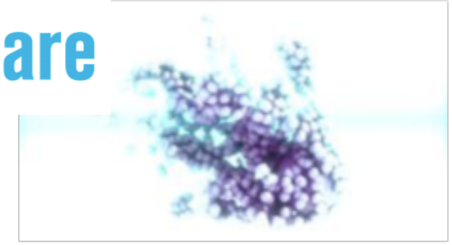
Native GPFS

Native GPFS, via AFM cache

BEAR Cloud

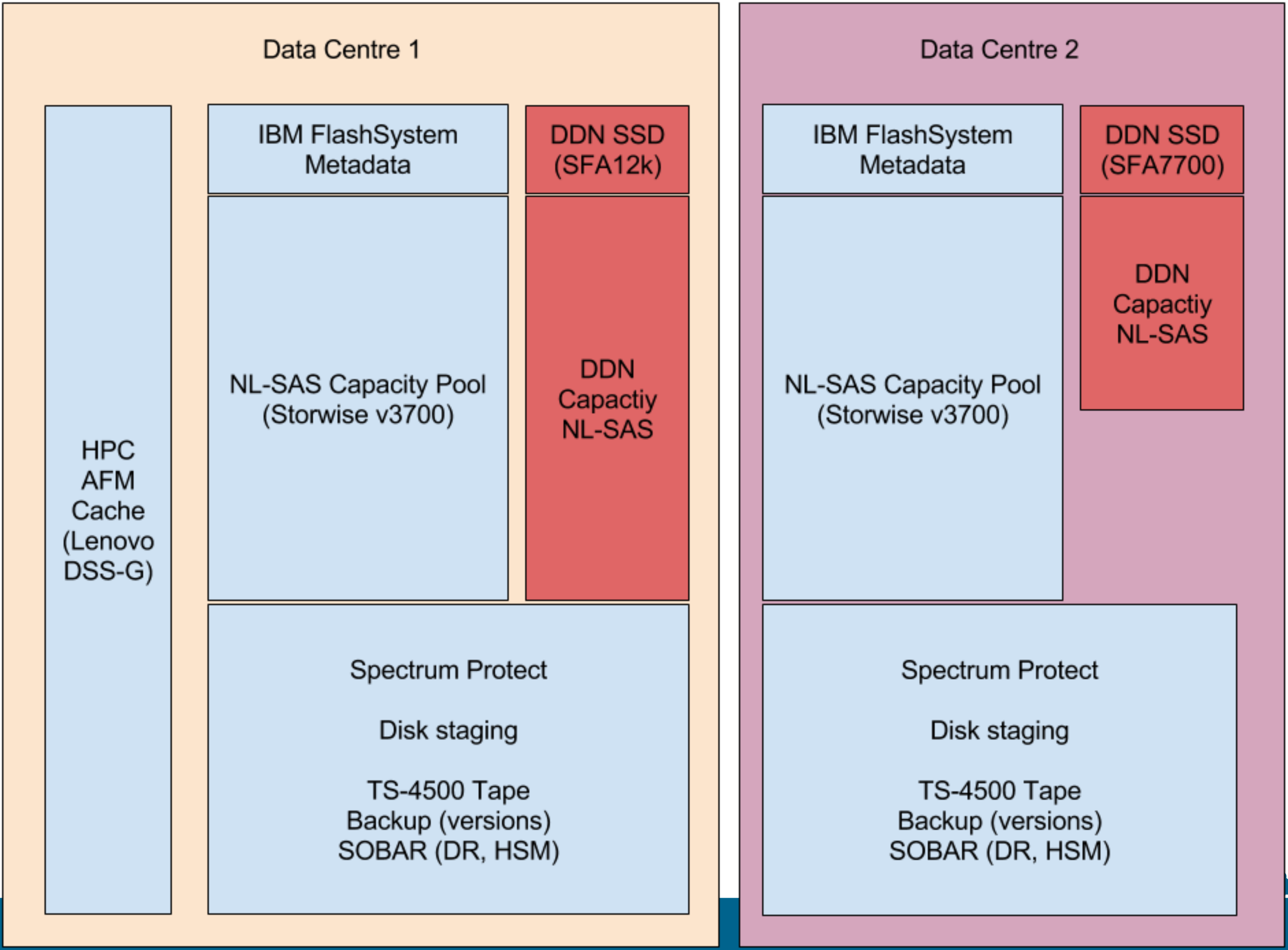
BlueBEAR HPC

BEAR DataShare



Data visualisation

- BEAR Cloud
- DCV
- Visualisation Centre



Data Centre 1

HPC AFM Cache (Lenovo DSS-G)

IBM FlashSystem Metadata

DDN SSD (SFA12k)

NL-SAS Capacity Pool (Storwise v3700)

DDN Capactiy NL-SAS

Spectrum Protect

Disk staging

TS-4500 Tape
Backup (versions)
SOBAR (DR, HSM)

Data Centre 2

IBM FlashSystem Metadata

DDN SSD (SFA7700)

DDN Capactiy NL-SAS

NL-SAS Capacity Pool (Storwise v3700)

Spectrum Protect

Disk staging

TS-4500 Tape
Backup (versions)
SOBAR (DR, HSM)

Scale under your cloud …

- □ Cinder/Glance integration
  - – Volumes/Images
- □ Manilla integration
  - – NFS “as a service”
- □ Single data management platform
  - – We already run for data services
  - – Standard placement rules for optimisation
  - – Integrate into existing backup as required

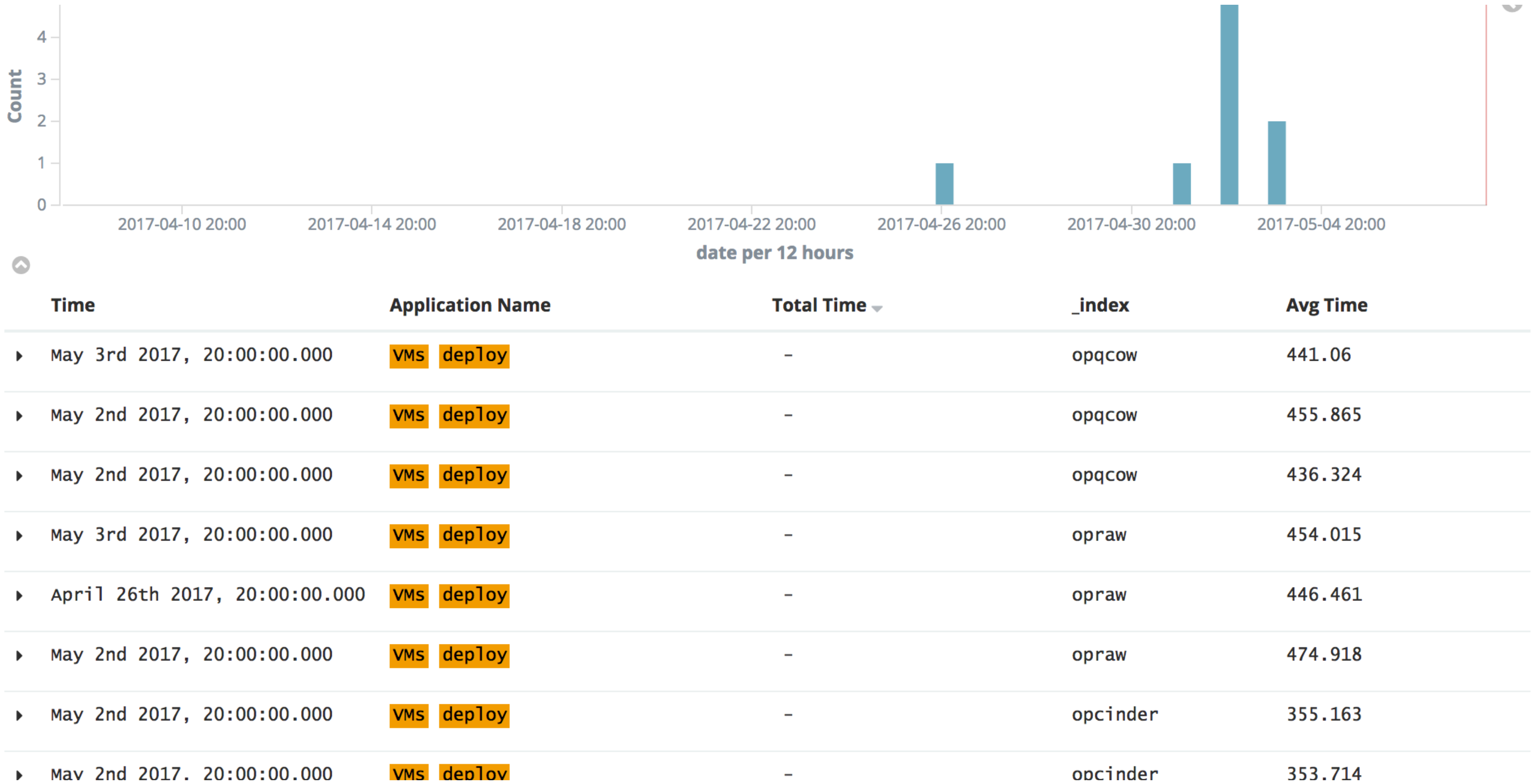
Optimising for OpenStack VMs

- □ What effect do various factors have on VMs?
  - – Disk format (raw/qcow2/cinder volume)?
  - – HAWC?
  - – LROC?
  - – Blocksize?
  - – Is it workload dependent?
  - – Can SFX Cache help?

DDN STORAGE

Working with DDN on this

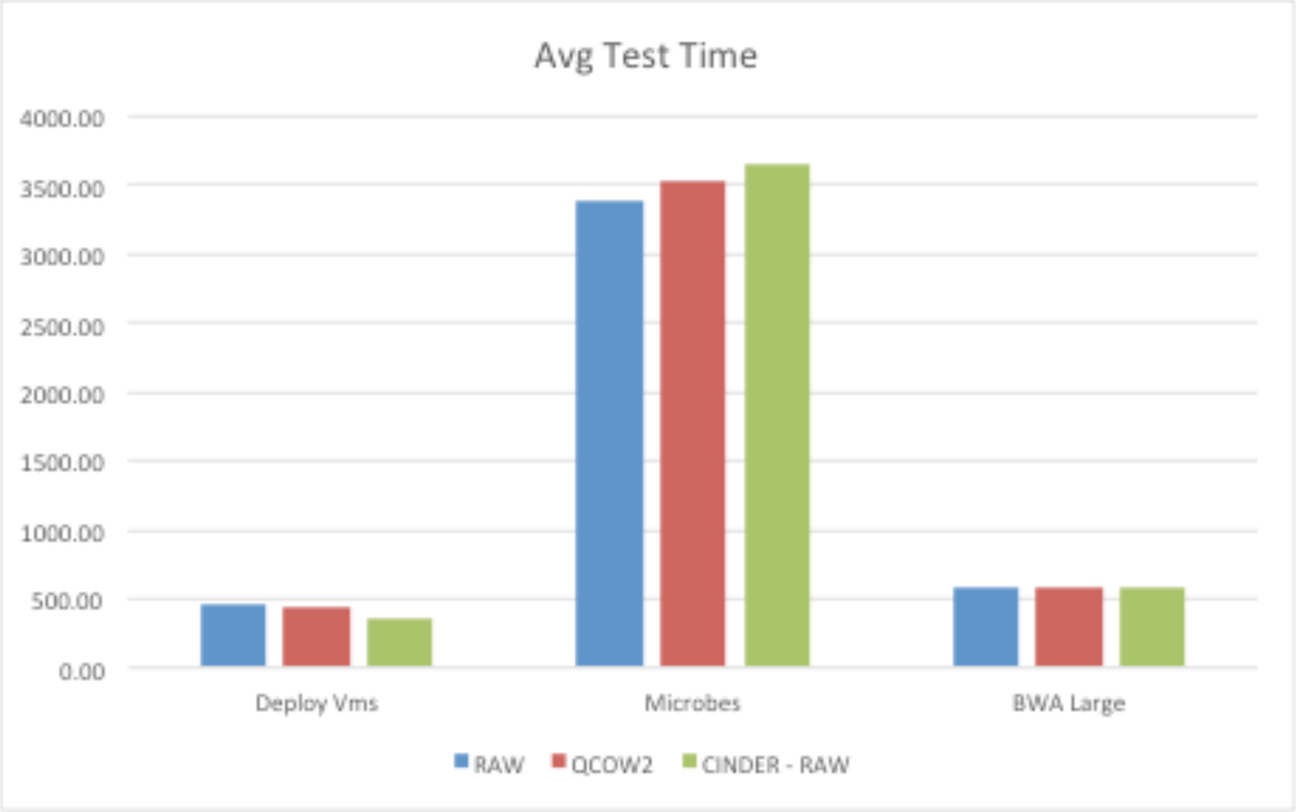
Early results – image format



Time	Application Name	Total Time	_index	Avg Time
May 3rd 2017, 20:00:00.000	VMs deploy	-	opqcow	441.06
May 2nd 2017, 20:00:00.000	VMs deploy	-	opqcow	455.865
May 2nd 2017, 20:00:00.000	VMs deploy	-	opqcow	436.324
May 3rd 2017, 20:00:00.000	VMs deploy	-	opraw	454.015
April 26th 2017, 20:00:00.000	VMs deploy	-	opraw	446.461
May 2nd 2017, 20:00:00.000	VMs deploy	-	opraw	474.918
May 2nd 2017, 20:00:00.000	VMs deploy	-	opcinder	355.163
May 2nd 2017, 20:00:00.000	VMs deploy	-	opcinder	353.714

Cinder volume fastest to boot…

Early results – image format



Credit: Maria Gutierrez, Abdul Alkhamees - DDN

Poking into HAWC

- Log is per client, but in system pool
- SSD metadata on FS already
  - `mmchfs bearcloud -L 128M`
  - `mmchfs bearcloud --write-cache-threshold 32K`
- Move fs manager and restart hypervisor GPFS
- `mmfsadm saferdump log| grep minNumFreeBytes`

  -
    ```
    nBytesFree 129990972 nBytesReserved 0
    maxNumFreeBytes 129991680 minNumFreeBytes
    129987260
    ```

Scale data into your cloud …

- How do we get integrated access?
- Manilla doesn’t work for us
- NFS with VXLAN to network nodes
  - Slow!
- NFS to existing protocol nodes
  - Pagepool and understanding ganesha
- We need to work on NFSv4+sec=krb

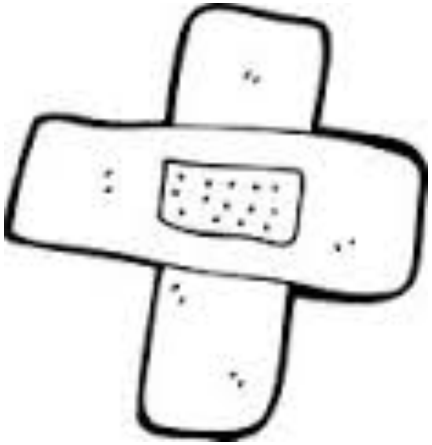
Scale data into your cloud …

- Native Spectrum Scale client
  - Works
    - Optimal networking needs tuning
    - SR-IOV IB
  - “Elastic” scaling is difficult!
  - Bulk destroy requires recovery

    (this is expected, but more likely to occur than with traditional HPC nodes)

Growing pains!

- SMB encryption performance issues
- Rapid expansion of services in last 18 months
- Storage instability
  - Pinch points in network
  - /rds use case change
  - HPC clients hanging
- mmnetverify helped with finding some issues
- Multi-homed boxes & rp_filter



Growing pains!

- mmnetverify helped with finding some issues
- Multi-cluster is great
  - Track back over 5 systems to find cause of issues
  - “Reverse” node expels
- NFS instability (some was our fault!)
- Taken time to implement (disruptive) changes

Challenges

- □ Remote data collection
  - – In the forest
  - – In the field

Current developments

- Just arrived, new Lenovo DSS-G system
  - First customer unit into Europe
  - Replace current multi-cluster for RDS with AFM cache
  - Decommission existing HPC storage
    - Legacy project storage solution
    - All projects will move to RDS storage

Lenovo
CAB 09
CAUTION PRUDENCE

Current developments

- Build some new data centres
- We just upgraded to Data Management Edition
  - Encryption
    - Securing research data
    - SMB3 end to end?

TCT capacity tier

- Researchers say we are too expensive
  - They can buy NAS for £40-50/TB
- Currently copies=2 (+ RAID6 overheads)
- TCT may help us here
  - Erasure code over 3 sites for 1.5x overhead