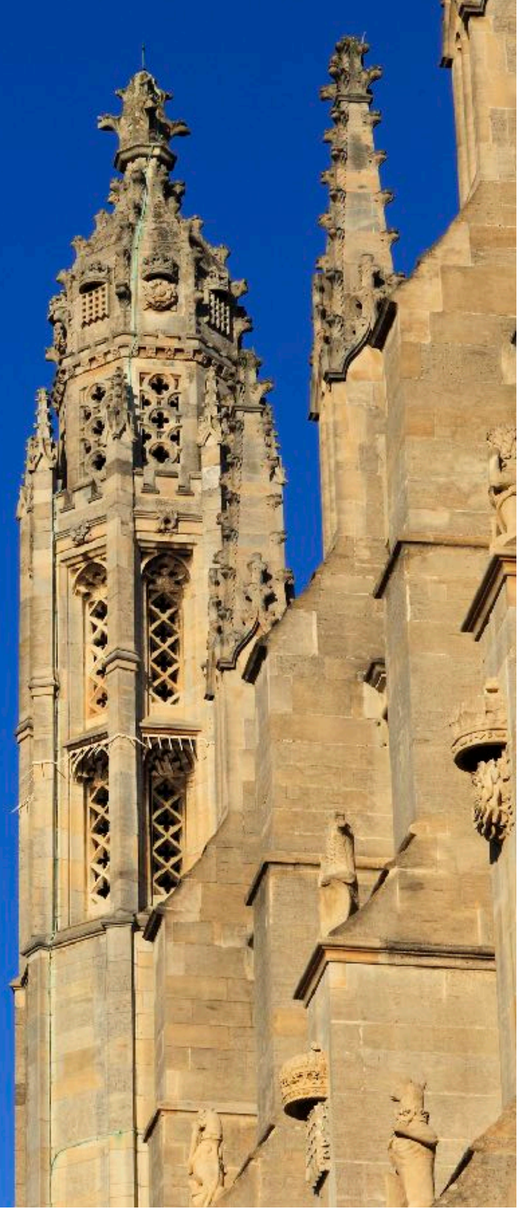
The Data Accelerator

PDSW-DISCS'18 WIP

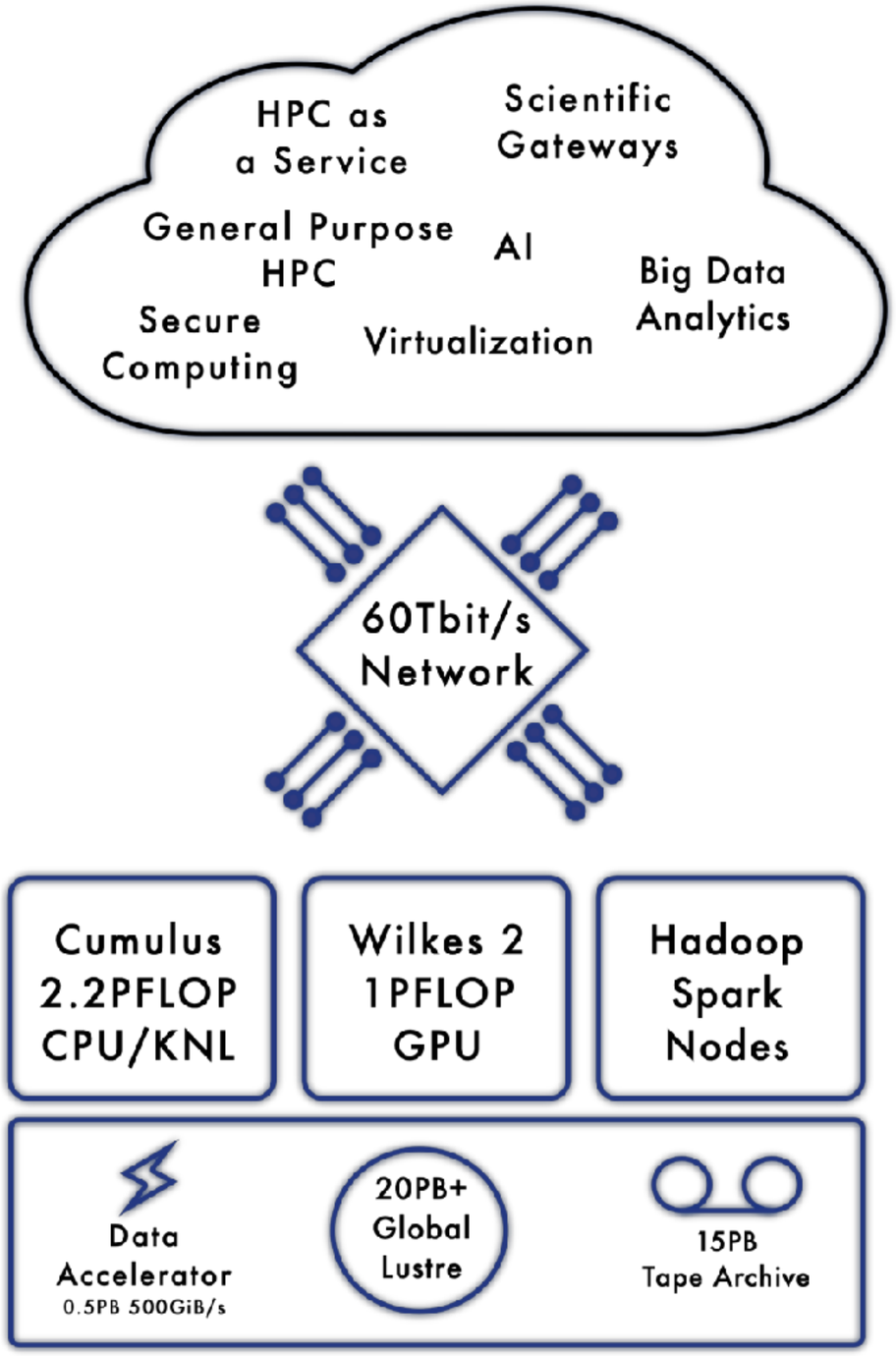
Alasdair King SC2018



UNIVERSITY OF CAMBRIDGE
Research Computing Services

CUMULUS
UK SCIENCE CLOUD

UNIVERSITY OF CAMBRIDGE
Research Computing Services



Data Accelerators Workflows and Features

- **Stage in/Stage out**
- Transparent Cashing
- **Checkpoint**
- Background data movement
- Journaling
- **Swap memory**

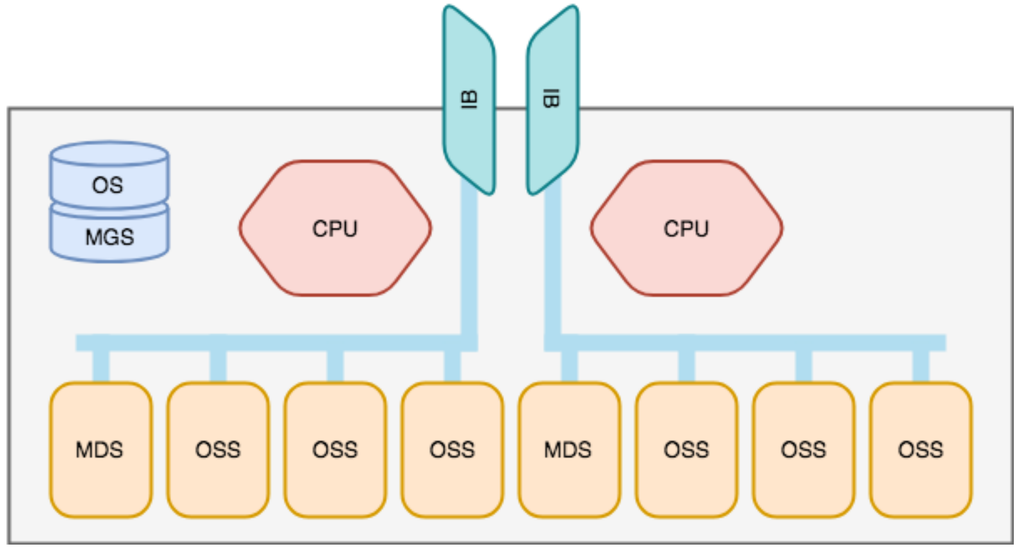
Storage volumes - namespaces - can persist longer than the jobs and shared with multiple users, or private and ephemeral.

POSIX or Object (this can also be at a flash block load/store interface)

Use cases in Cosmology, Life Sciences - Genomics, Machine learning workloads, Big Data analysis.

UNIVERSITY OF CAMBRIDGE
Research Computing Services

The Data Accelerator Platform



- Each DAC uses an internal SSD for the MGS should it be elected to run a file system.
- NVMeS then have an MDS or OSS applied. This arrangement can be changed as required.

24 Dell EMC PowerEdge R740xd

2 Intel Xeon Scalable Processors

2 Intel Omni-Path Adaptors

Each with 12 Intel SSD P4600

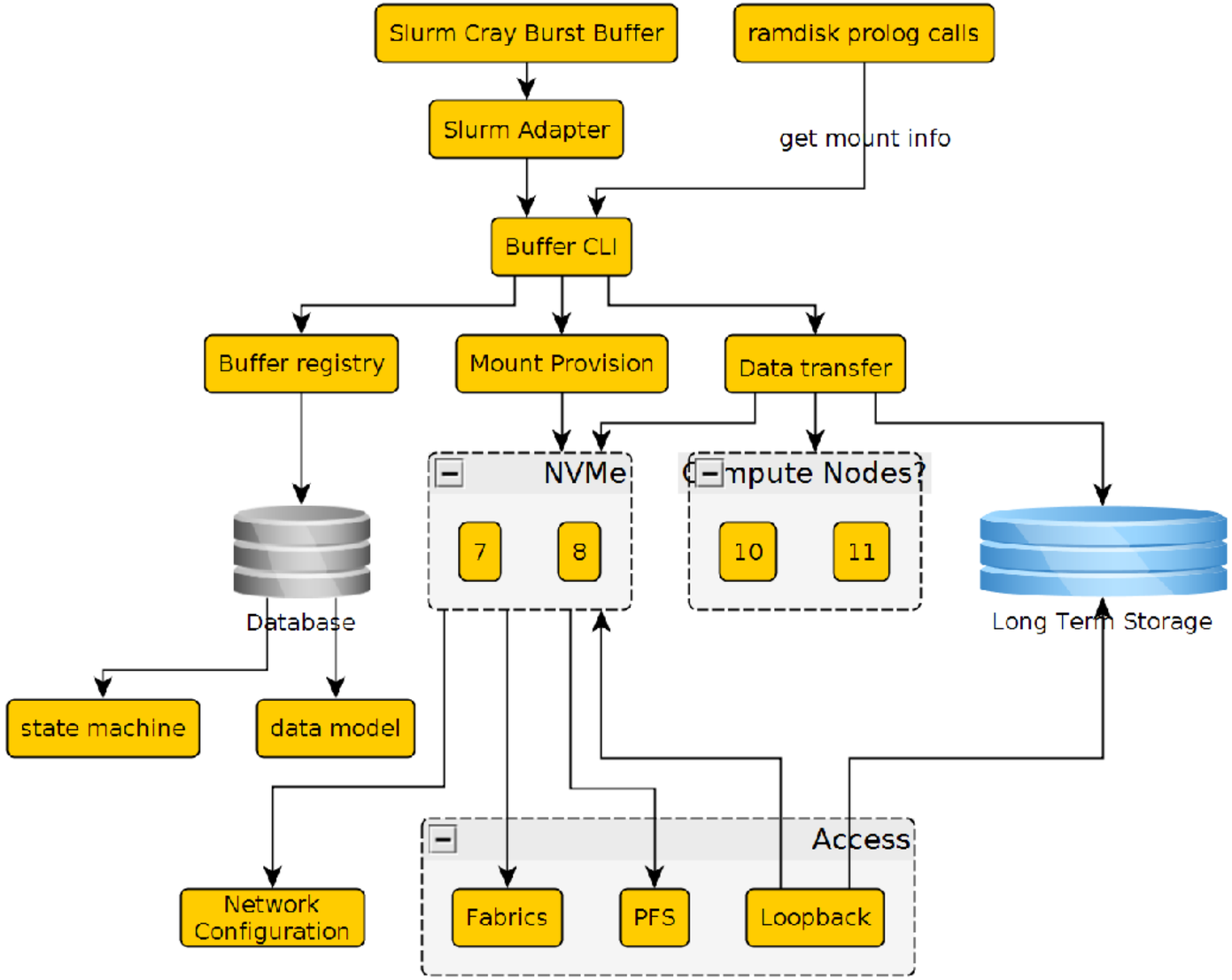
½PB of Total Available Space

UNIVERSITY OF CAMBRIDGE
Research Computing Services

Integration with SLURM via flexible storage orchestrator

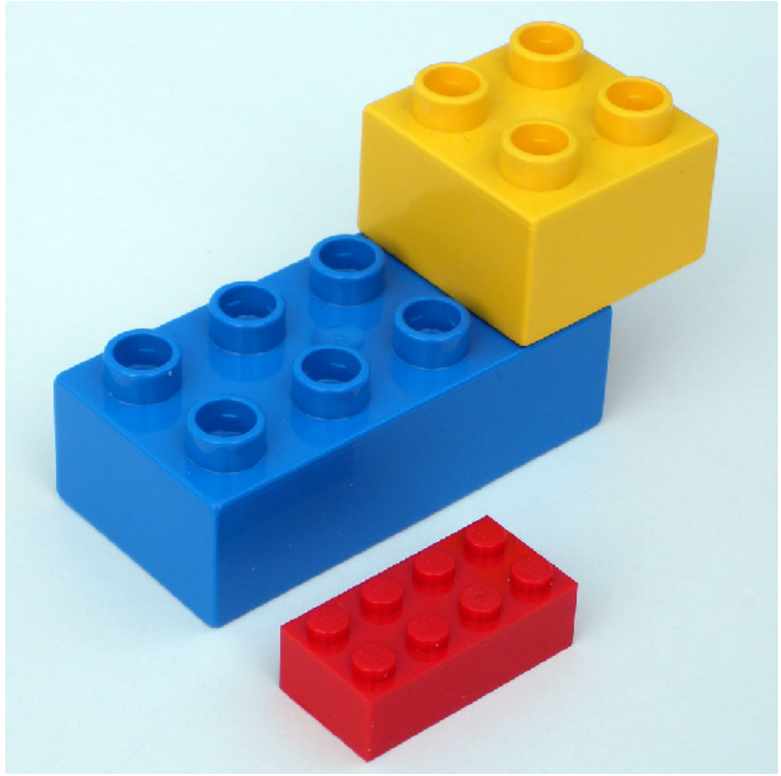
SLURM DAC Plugin

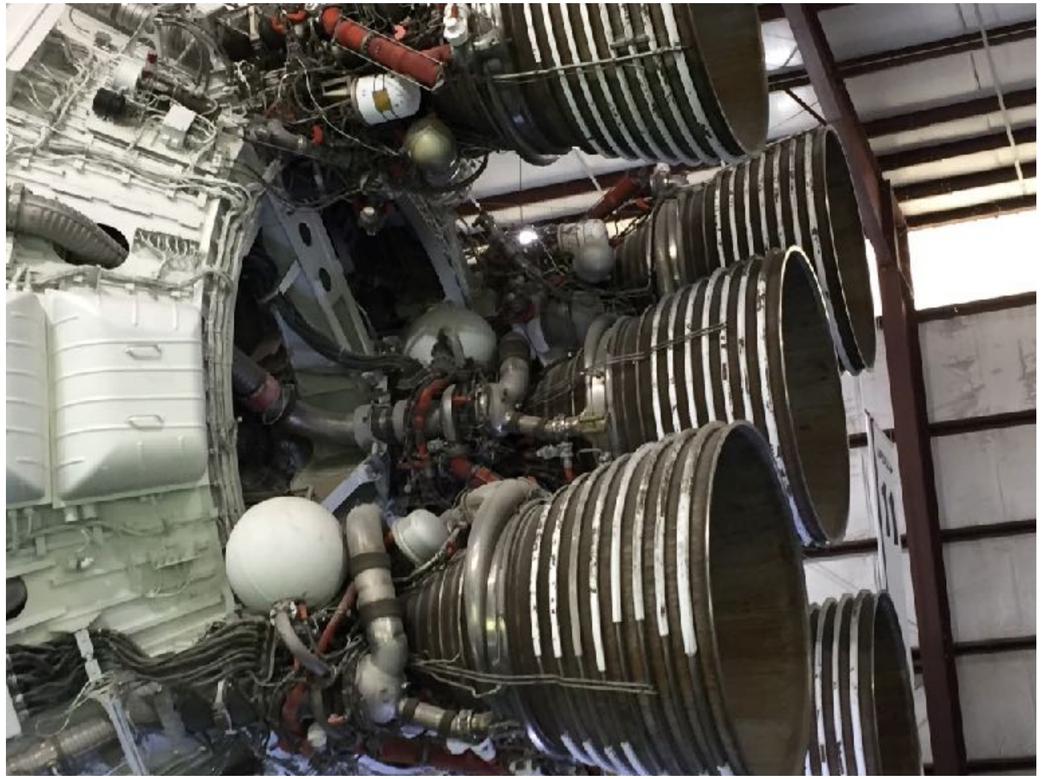
- Reuses the existing Cray plugin.
- Cambridge has implemented an orchestrator to manage the DAC nodes.
- Go project utilising ETCd and Ansible for dynamic automated creation of filesystems
- To be released as an OpenSource project.



UNIVERSITY OF CAMBRIDGE
Research Computing Services

Technical challenges





UNIVERSITY OF CAMBRIDGE
Research Computing Services

Problems Discovered

- ARP Flux in Multi-rail networks
- Multicast and Static Routing
- Lustre patches to bypass page cache on SSD
- BeeGFS multipal filesytem organisation
- Omni-Path errors and original system topology design

*Please email if you're interested in the writeup of solving some of these problems.

UNIVERSITY OF CAMBRIDGE
Research Computing Services

ARP Flux

Compute Nodes

Who has the MAC Address of 10.47.18.1?

Compute node A

10.47.18.1 its at 00:00:FA:12

Who has the MAC Address of 10.47.18.1?

Compute node B

10.47.18.1 its at 00:00:FB:16

Storage Multi-Rail Nodes

I have 10.47.18.1 Its at 00:00:FA:12

IB0 10.47.18.1

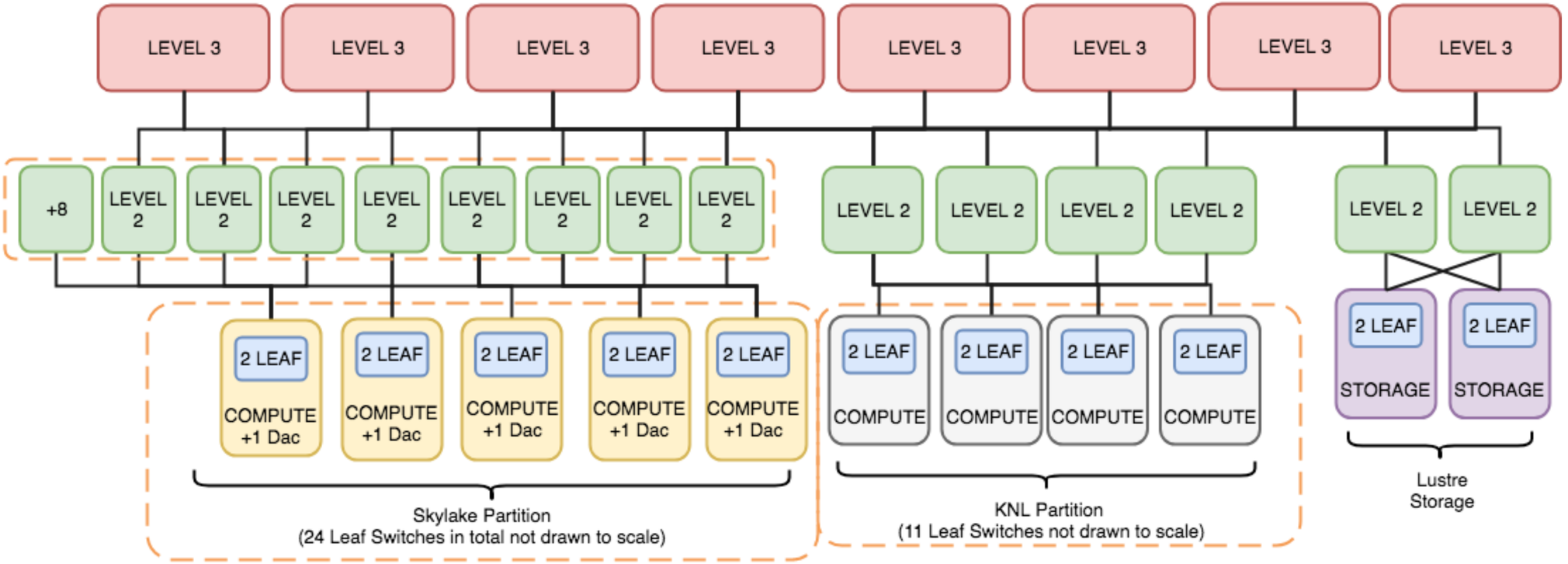
I have 10.47.18.1 Its at 00:00:FB:16

IB1 10.47.18.25

Multi-Rail node A

UNIVERSITY OF CAMBRIDGE
Research Computing Services

Cumulus OPA Interconnect Topology



UNIVERSITY OF CAMBRIDGE
Research Computing Services

* **Each Level is 2:1 Blocking with the exception of the DAC (1:1)**

* **Wilkes II (Not shown) Connects via LNET routers to access storage only**

Performance on Cumulus

- Can reach 500GiB/s Read and 300GiB/s Write on Synthetic IOR for 184 Nodes 32 ranks per node (5888 MPI Ranks)
- x25 faster than Cumulus's existing 20GiB/s Lustre scratch
- Cambridge would have to spend over x10 to reach the same performance target without considering space and power implications.

UNIVERSITY OF CAMBRIDGE
Research Computing Services

IO500 and some Numbers

Sneak Peek Lustre Numbers

mdtest_hard_stat	**2112.230**	**kiops**	**(2.1 Million iops)**
mdtest_hard_read	**1618.130**	**kiops**	**(1.6 Million iops)**

UNIVERSITY OF CAMBRIDGE
Research Computing Services

*Tested with both BeeGFS and Lustre

Further work

- Integration and testing on the live system
- Testing UK Science. Working with DiRAC to evaluate the impact on their workloads.
- Filesystem tuning and I/O Job monitoring
- General Release for all as a resource on Cumulus and as an Open Source solution.

Questions and Comments?

UNIVERSITY OF CAMBRIDGE
Research Computing Services

Alasdair King ajk203@cam.ac.uk

Thanks for the Continued Support of :