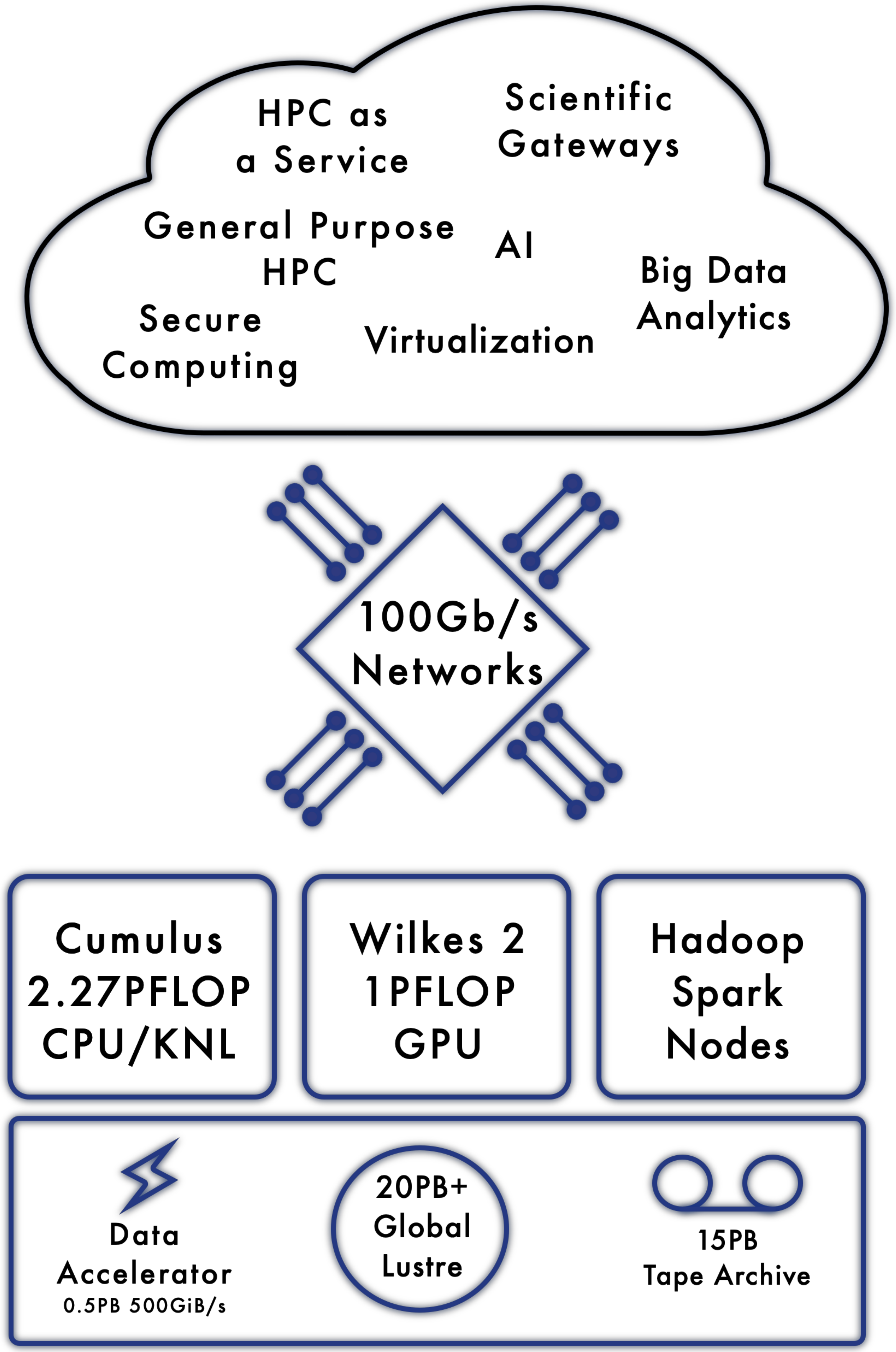
The Data Accelerator

University of Cambridge IO500

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.

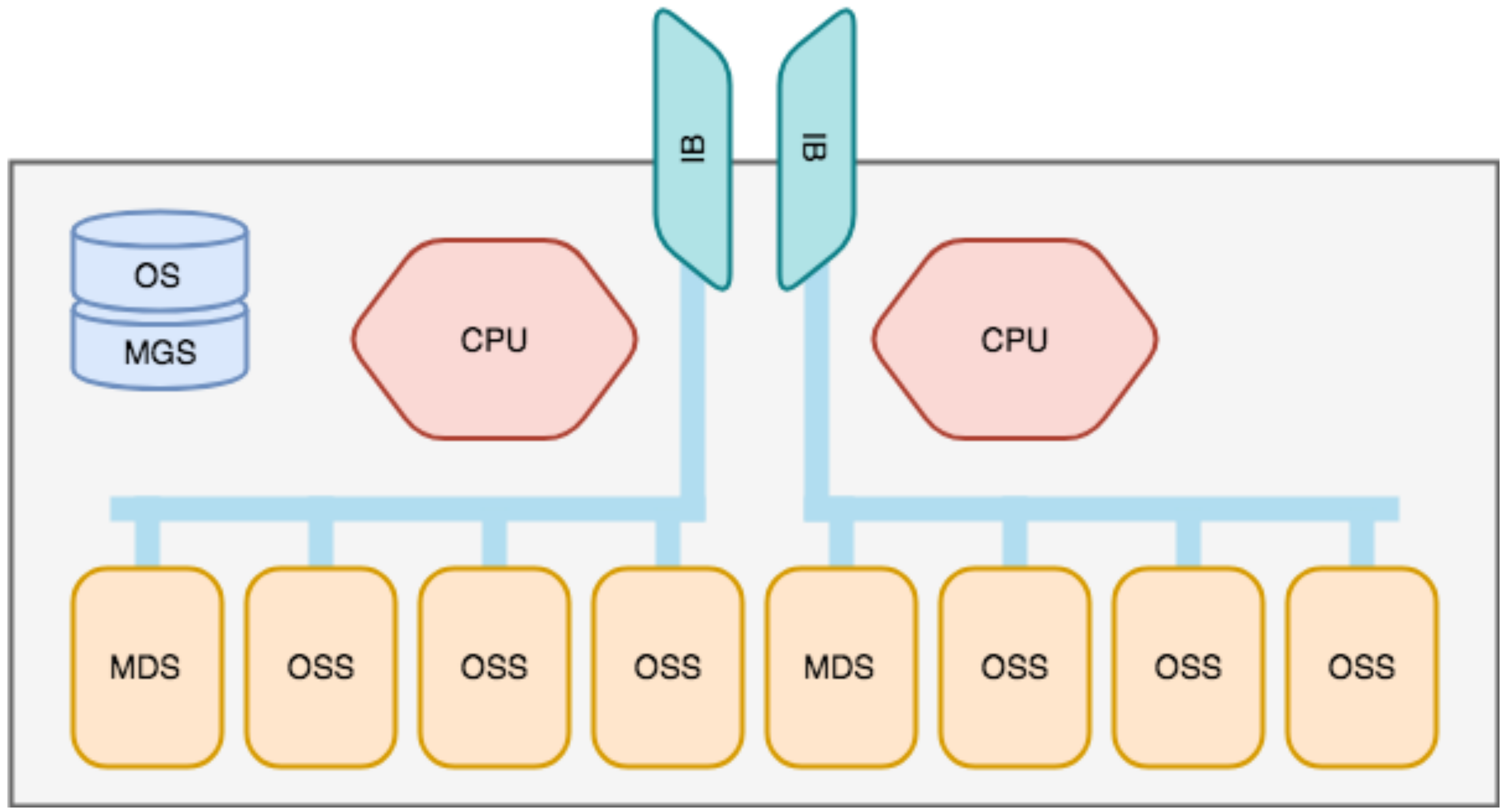
POSIX or Object (this can also be at a flash block load/store interface)

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

UNIVERSITY OF CAMBRIDGE
Research Computing Services

(Ref. https://glennklockwood.blogspot.com/2017/03/reviewing-state-of-art-of-burst-buffers.html)

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

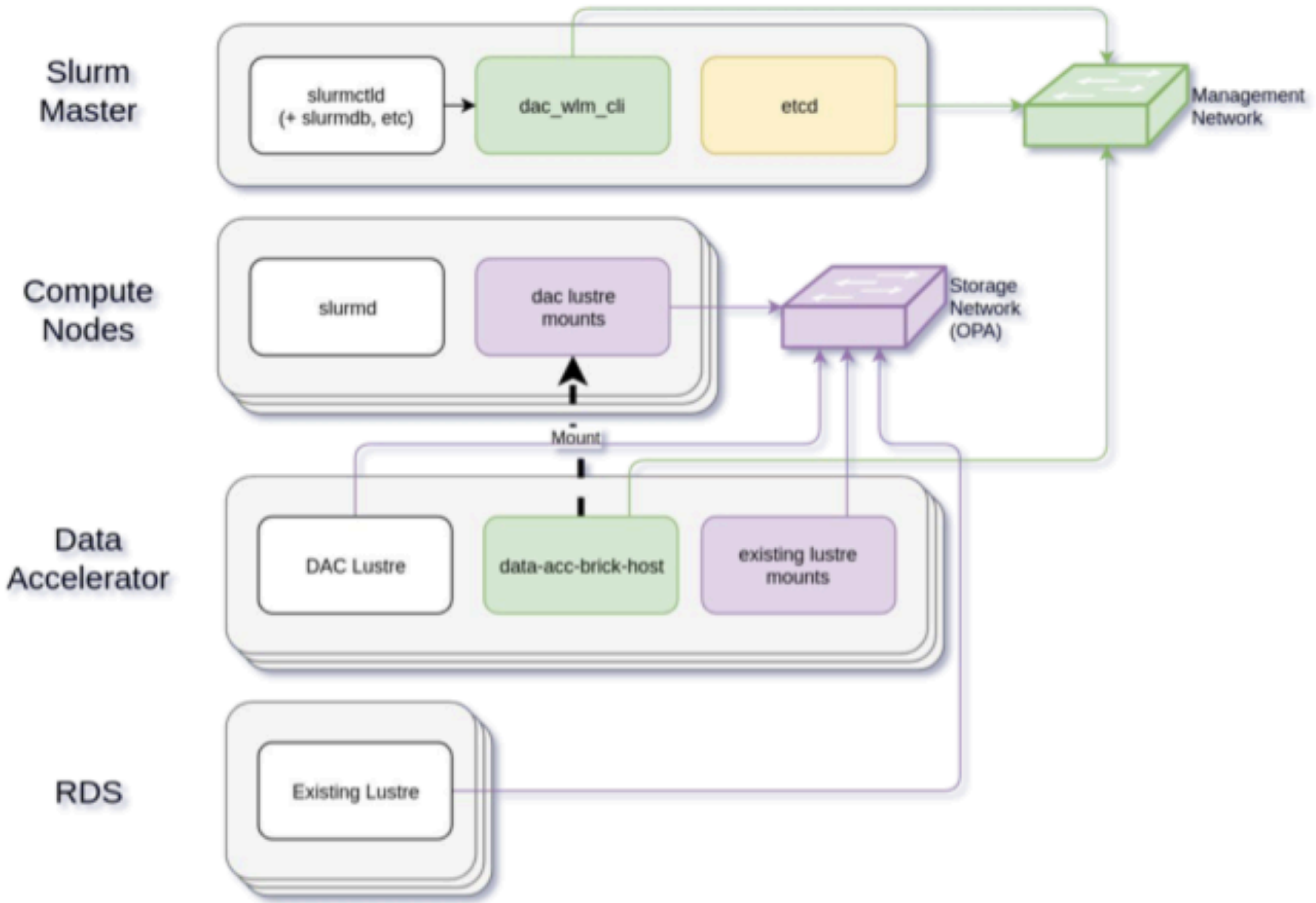
½PB of Total Available Space

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

Integrating Lustre for the Data Accelerator

UNIVERSITY OF CAMBRIDGE
Research Computing Services

Ansible Enabled Lustre Install

```
ansible-playbook test-dac-lustre.yml -i test-inventory-lustre --tag format --tag reformat_mgs
ansible-playbook test-dac-lustre.yml -i test-inventory-lustre --tag mount,create_mdt,create_mgs,create_osts,client_mount
ansible-playbook test-dac-lustre.yml -i test-inventory-lustre --tag stop_all,unmount,client_unmount
ansible-playbook test-dac-lustre.yml -i test-inventory-lustre --tag format

ansible-playbook test-dac-lustre.yml -i test-inventory-lustre --tag stop_mgs
ansible-playbook test-dac-lustre.yml -i test-inventory-lustre --tag reformat_mgs
```

*test-inventory-lustre

```
dac:
  children:
    fs1:
      hosts:
        dac1:
          fs1_mgs: nvme0n1
          fs1_mdt: nvme1n1
          fs1_osts: {nvme2n1: 2}
        dac2:
          fs1_osts: {nvme3n1: 1}
      vars:
        fs1_mgsnode: dac1
```

*test-dac-lustre.yml

```
---
- name: Setup buffer for fs1
  hosts: fs1
  become: yes
  roles:
    - role: lustre
      vars:
        fs_name: fs1
```

UNIVERSITY OF CAMBRIDGE
Research Computing Services

Multirail Lustre

- Set up the ARP and Linux Kernel Routing before enabling multirail

#Setting ARP so it doesn't broadcast (Do this for every IB interface)

```
sysctl -w net.ipv4.conf.all.rp_filter=0
sysctl -w net.ipv4.conf.ib0.arp_ignore=1
sysctl -w net.ipv4.conf.ib0.arp_filter=0
sysctl -w net.ipv4.conf.ib0.arp_announce=2
sysctl -w net.ipv4.conf.ib0.rp_filter=0
```

Multirail Lustre

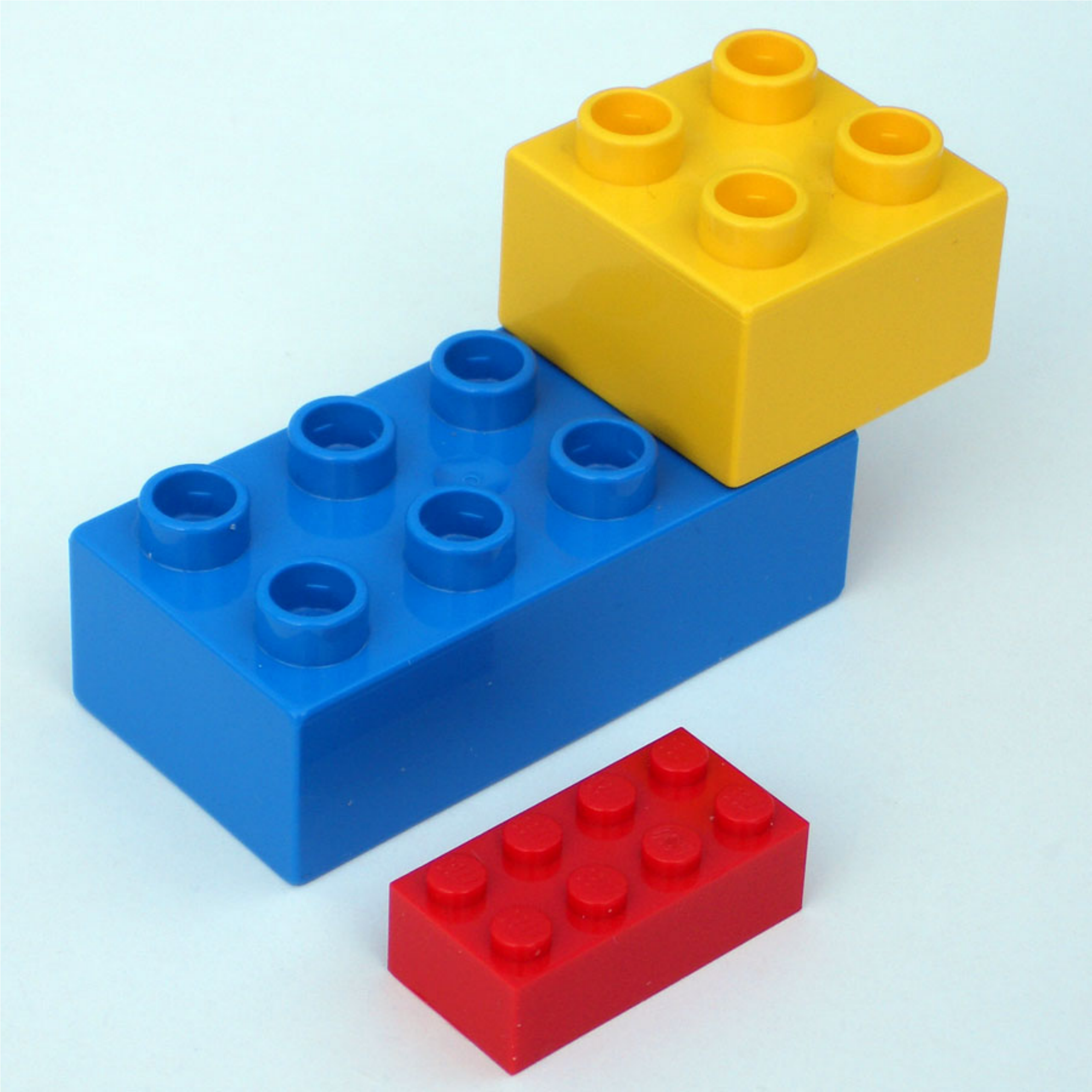
- Set up the ARP and Linux Kernel Routing before enabling multirail

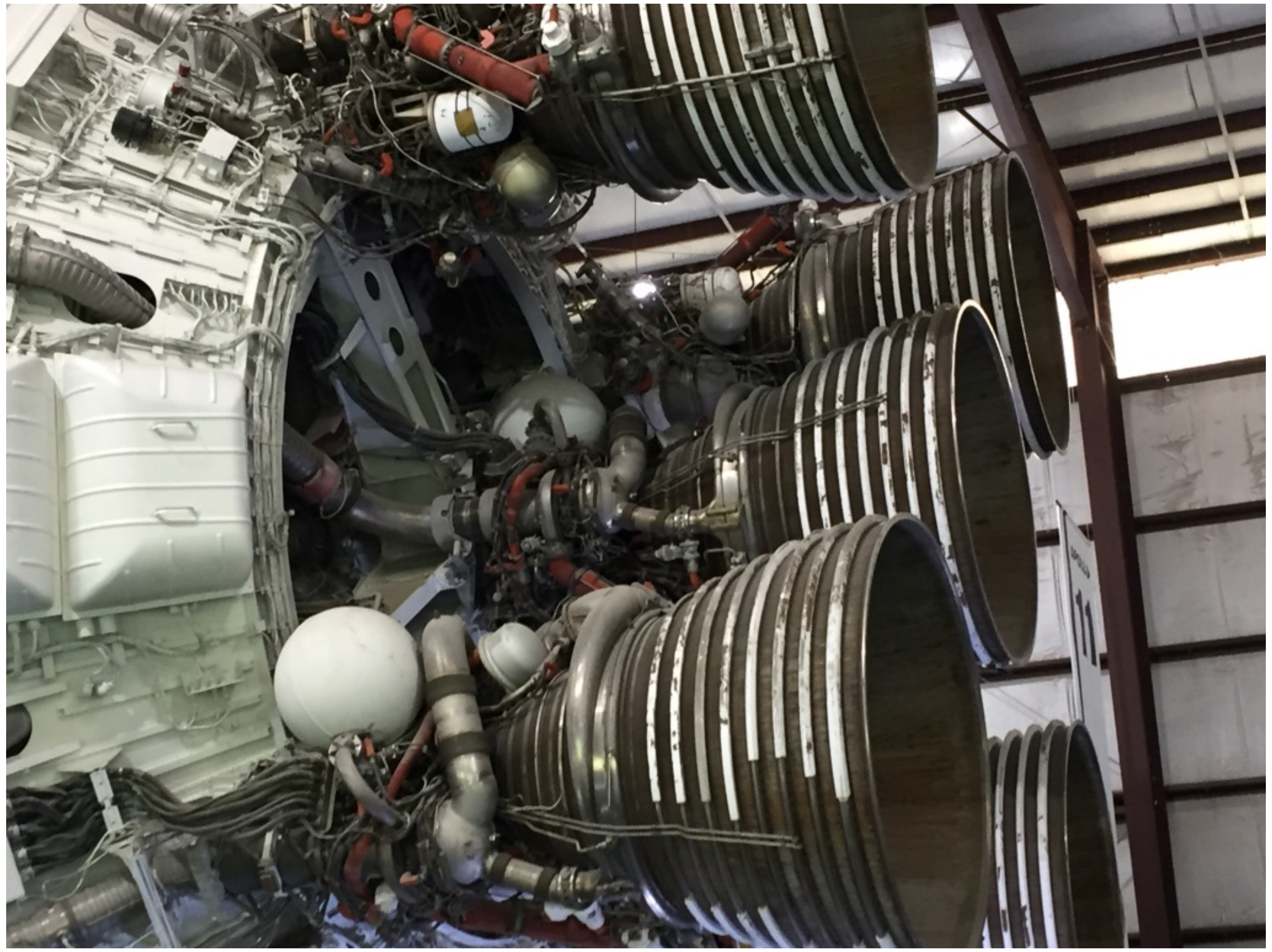
```
ip neigh flush dev ib0
ip neigh flush dev ib1

echo 200 ib0 >> /etc/iproute2/rt_tables
echo 201 ib1 >> /etc/iproute2/rt_tables

ip route add 192.168.0.0/16 dev ib0 proto kernel scope link src 192.168.1.1 table ib0
ip route add 192.168.0.0/16 dev ib1 proto kernel scope link src 192.168.2.1 table ib1

ip rule add from 192.168.1.1 table ib0
ip rule add from 192.168.2.1 table ib1

ip route flush cache
```

MDT Large_dir for DNE 2

- Default ext4/e2fsprocs is a 2 Level htree for 10M files
- Can be increased to 3 levels with large_dir option in e2fsprogs 2.14
- add this to mkfsoptions or tune2fs to enable

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 (If using SSD for lustre use 2.12)
- BeeGFS multipal filesytem organisation
- Omni-Path errors and original system topology design

UNIVERSITY OF CAMBRIDGE
Research Computing Services

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

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

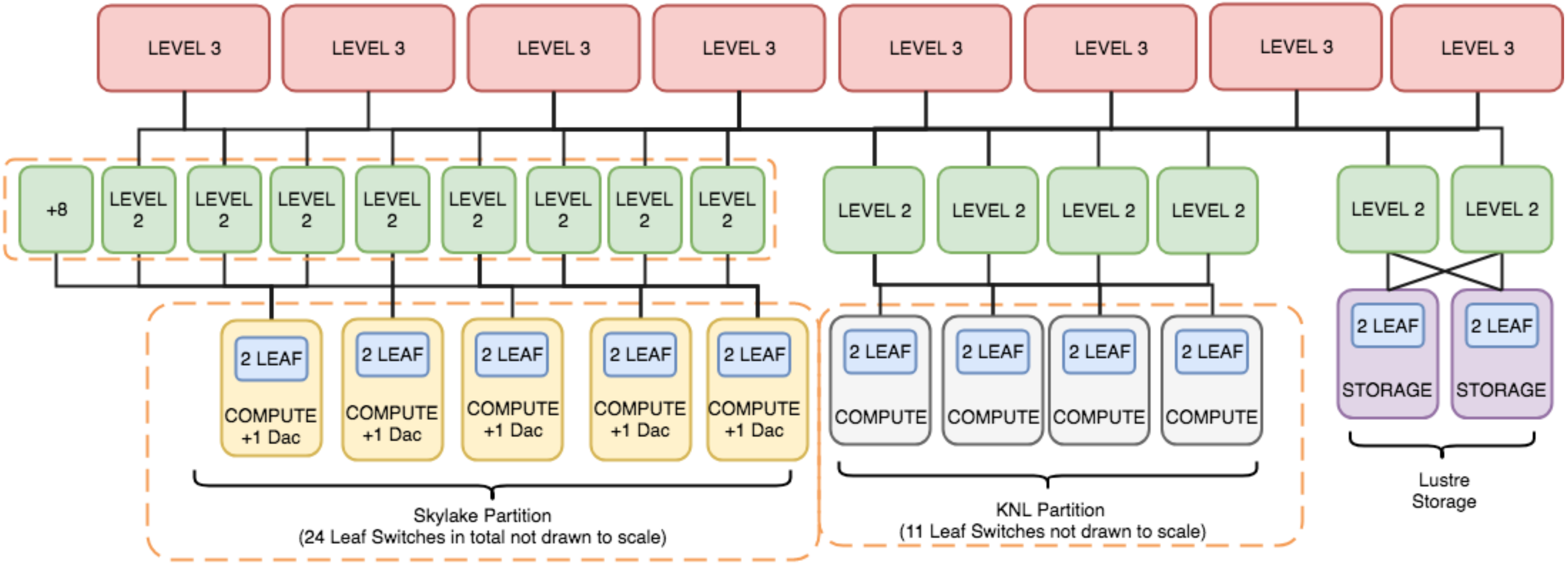
IB1 10.47.18.25

Multi-Rail node A

UNIVERSITY OF CAMBRIDGE
Research Computing Services

Fixed if setting up Multi-rail as per previous slide

Cumulus OPA Interconnect Topology

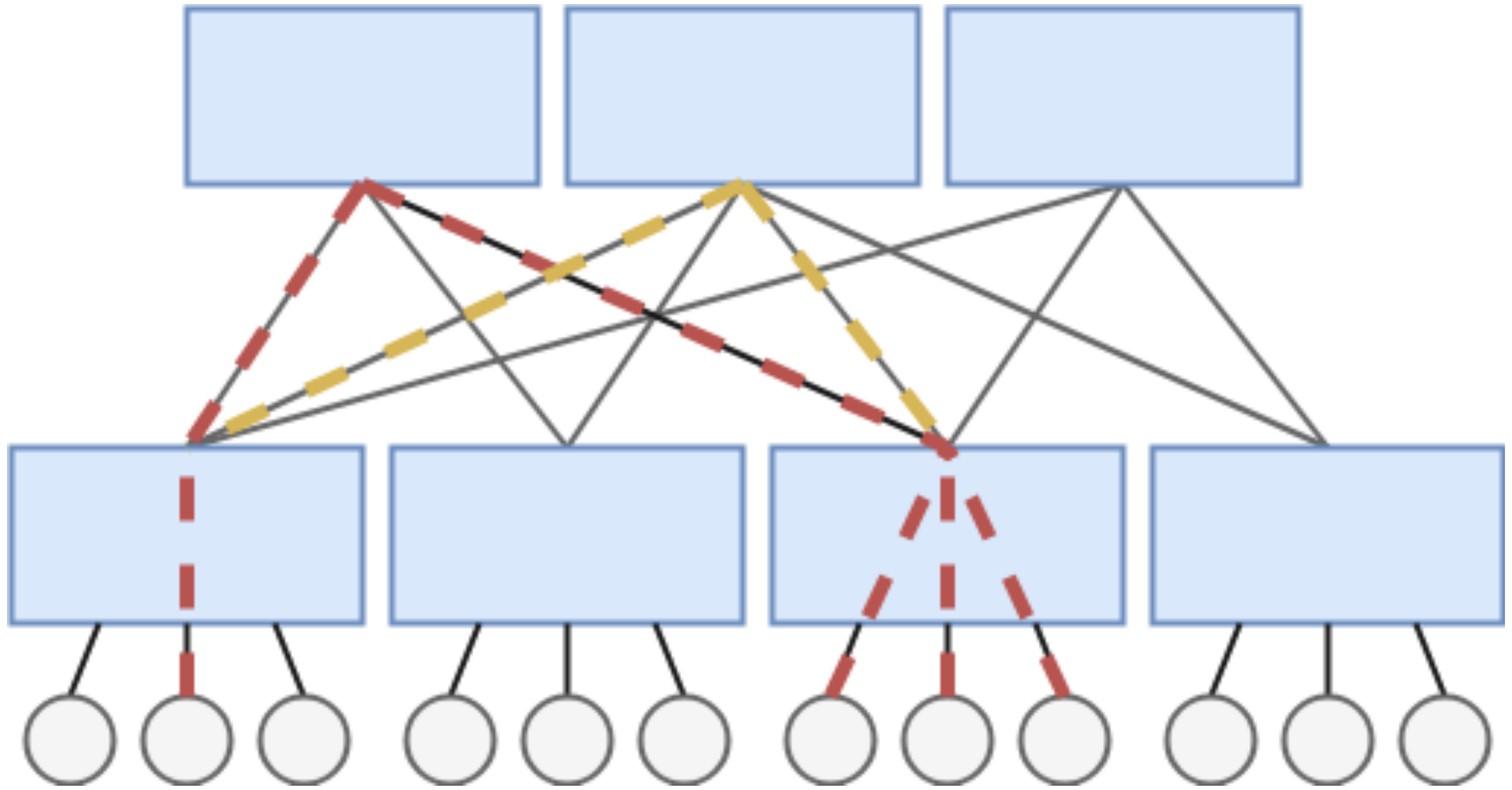


* Each Level is 2:1 Blocking with the exception of the DAC (1:1)

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

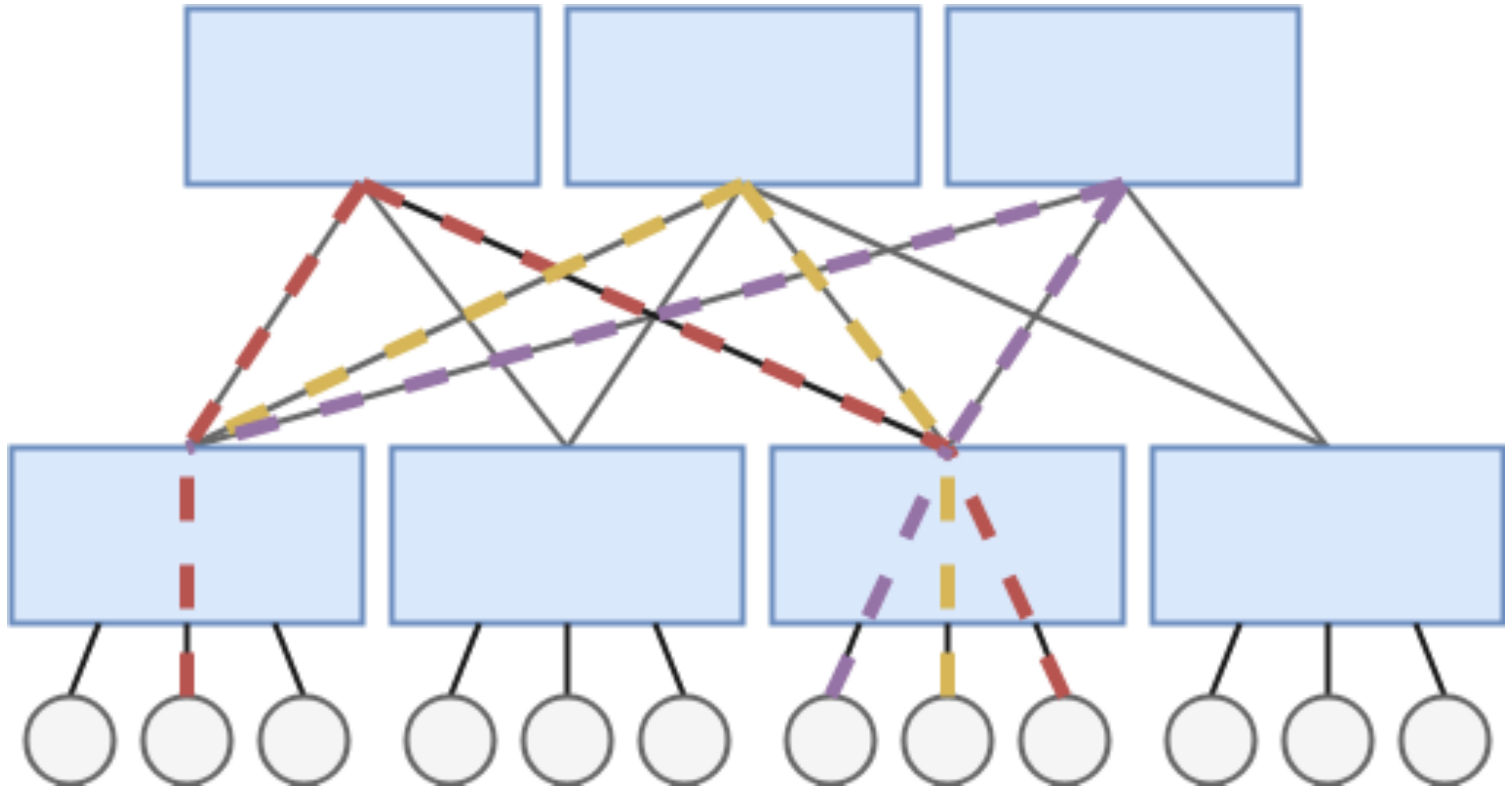
UNIVERSITY OF CAMBRIDGE
Research Computing Services

Fat Tree Static Routing



- All nodes take the same Inter Switch Links(Red)
- Other Links are Posible(Gold)

Adaptive Routing



- Nodes can now take alternate routes (Gold,Purple)
- Utilisation of Inter switch links improved

Diagnosing in Intel Omni-Path

```
Group Focus: All   GrpNumPorts: 4015  NumPorts: 10  Number: 10
  Ix  Util-High   LIDx   Port   Node GUID 0x    NodeDesc
   0        0.0 0113  46  00117501020D8FAA opasw-fr16-u40
  <->      84.6 04B3   3  00117501020F29B1 opasw-dr20-u35
   1        0.0 0113  15  00117501020D8FAA opasw-fr16-u40
  <->      74.8 0190   3  00117501020D8F8D opasw-dr20-u30
   2        0.0 04B3  25  00117501020F29B1 opasw-dr20-u35
  <->      70.0 04D2  19  00117501020D805E opasw-dr19-u42
   3        0.1 0113  11  00117501020D8FAA opasw-fr16-u40
  <->      67.4 04AE   3  00117501020F4147 opasw-dr20-u42
   4       57.4 005A   3  00117501020C57AF opasw-dr20-u33
  <->       0.1 0113  30  00117501020D8FAA opasw-fr16-u40
   5        0.0 04D2  14  00117501020D805E opasw-dr19-u42
  <->      51.9 04FB   1  00117501010DBA30 dac-e-13 hfi1_1
   6        0.0 0190  41  00117501020D8F8D opasw-dr20-u30
  <->      51.4 01A9  37  0011750102702B0F opasw-dr19-u41
   7        0.0 04AE   9  00117501020F4147 opasw-dr20-u42
  <->      49.9 04B6  33  00117501020C47F7 opasw-dr19-u42
   8        0.0 018B  11  0011750102702978 opasw-fr16-u38
  <->      49.8 04AE  36  00117501020F4147 opasw-dr20-u42

Quit up Live/rRev/fFwd/bookmrked Bookmrk Unbookmrk ?help | sS cC N0-n P0-n:
```

- Example of *opatop* during a test. Can highlight oversubscribed links based on the percentage utilised.

Topology Problems

- The speed at which the SSDs can achieve forces changes away from placement of traditional disk systems.

- DAC nodes are now in place with compute nodes.

- If out on an island, static routing hurts performance, and can be relieved with adaptive routing.

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.

IO500

Lustre 2.11 on 24 DAC - 8NVMe - with 20 MDT

[RESULT]	BW	phase	1	ior_easy_write	208.252	GB/s	:	time	365.1	seconds
[RESULT]	IOPS	phase	1	mdtest_easy_write	53.451	kiops	:	time	352.43	seconds
[RESULT]	BW	phase	2	ior_hard_write	7.441	GB/s	:	time	509.42	seconds
[RESULT]	IOPS	phase	2	mdtest_hard_write	366.946	kiops	:	time	349.35	seconds
[RESULT]	IOPS	phase	3	find	729.39	kiops	:	time	192.27	seconds
[RESULT]	BW	phase	3	ior_easy_read	358.561	GB/s	:	time	212.05	seconds
[RESULT]	IOPS	phase	4	mdtest_easy_stat	247.4	kiops	:	time	91.97	seconds
[RESULT]	BW	phase	4	ior_hard_read	46.78	GB/s	:	time	81.04	seconds
[RESULT]	IOPS	phase	5	**mdtest_hard_stat**	**2112.23**	**kiops**	:	time	72.16	seconds
[RESULT]	IOPS	phase	6	mdtest_easy_delete	50.864	kiops	:	time	365.44	seconds
[RESULT]	IOPS	phase	7	**mdtest_hard_read**	**1618.13**	**kiops**	:	time	96.21	seconds
[RESULT]	IOPS	phase	8	mdtest_hard_delete	389.67	kiops	:	time	333.57	seconds
[SCORE]	**Bandwidth**	**71.4032**	**GB/s**	**:**	**IOPS**	**352.754**	**kiops**	**:**	**TOTAL**	**158.707**

UNIVERSITY OF CAMBRIDGE
Research Computing Services

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.

THANKS TO EVERYONE!

Paul Calleja
Wojciech Turek
Stuart Rankin
John Garbutt
Jeffery Salmond
Dominic Friend
Joe Stankiewicz
Matt Raso-Barnett
Paul Brown
John Taylor
Sean McGuire

UNIVERSITY OF CAMBRIDGE
Research Computing Services

StackHPC

whamCloud

DELL EMC

BeeGFS®

intel®

Questions and Comments?

UNIVERSITY OF CAMBRIDGE
Research Computing Services

Alasdair King ajk203@cam.ac.uk