Research Storage 2019 Update

Matt Rásó-Barnett
Research Computing Platforms
October 2019

What do we offer? (HPC Cluster Customer View)

- High Performance
- Large
- Scalable under parallel IO
- 1 copy on disk
- No Backup

- ***Active* data here**

Data Accelerator (DAC)

~500TiB NVMe Lustre

Research Data Store (RDS)

~ 12PiB Lustre

Performance Tier

- Reasonably Fast
- Small
- Non-scalable under parallel IO
- 2 copies on disk
- Backed up

- **All *Code* here**

? /

NFS /home

~1PiB ZFS

Resilience Tier

- Slow**, Very High Latency
- Large
- Non-scalable under parallel IO
- 2 copies on tape
- Not backed up

- ***Inactive* data here**

Research Cold Store (RCS)

~10PiB Tape

Archive Tier
Two Copies on Tape

Data Accelerator (DAC)

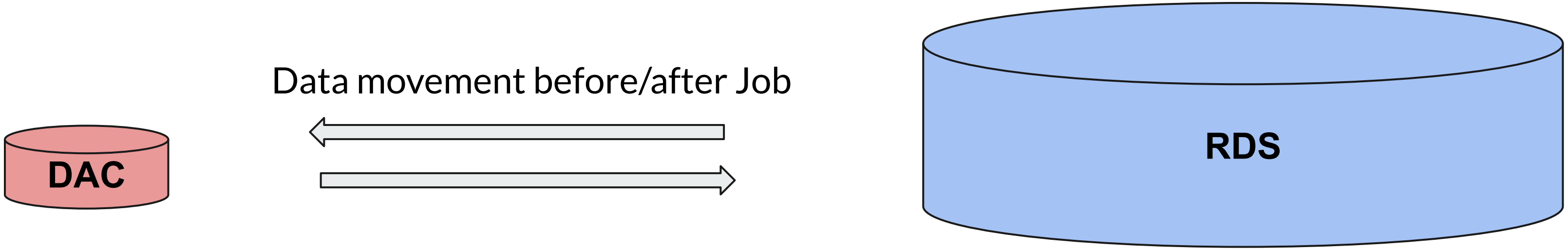
What is it?

- All-Flash Storage service, designed for maximum performance for HPC jobs

- Presented as ‘Burst-Buffer’-style storage area for CSD3 Cluster

    - **500 TiB Flash**
    - **Schedulable**
    - **Exclusive access to the storage during your job**
        - **Closer to providing guaranteed level of performance**
    - **No persistency beyond length of job* (or beyond scheduled time-period)**

    - **Stage-in and stage-out from larger bulk parallel filesystem (RDS)**



Why is this useful?

- Economics of flash - can only buy *so* much!
- Where to put it? How to best use it?
  - Salami-slicing it in the same shared model of RDS would not provide enough
  - Shared-scratch with purging still suffers from neighbor effects
    - Need to trim those SSDs sometime too
- DAC attempts to take the ‘buffer’ approach as pioneered in Cray DataWarp™, DDN IME™
- Ideal for:
  - Applications with large checkpointing bandwidth needs
  - Applications with heavily random IO patterns, or IOPS intensive
  - Metadata-intensive applications

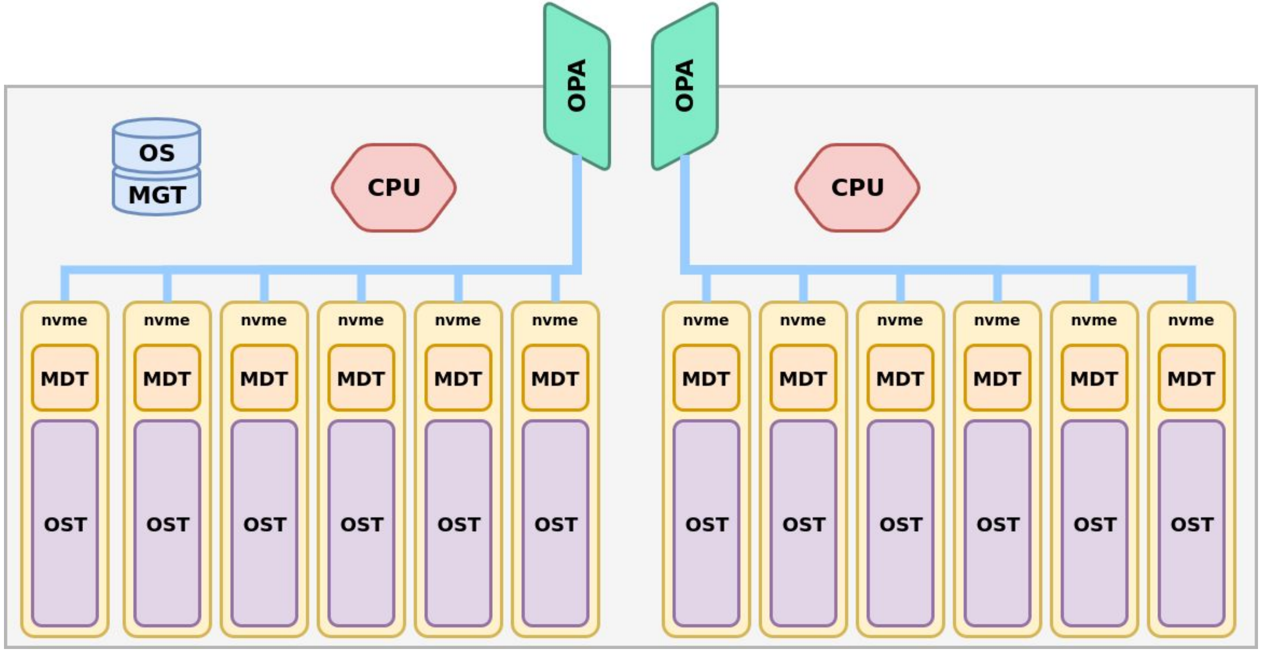
Hardware Platform

- ❖ **~500 TiB of NVMe Flash**

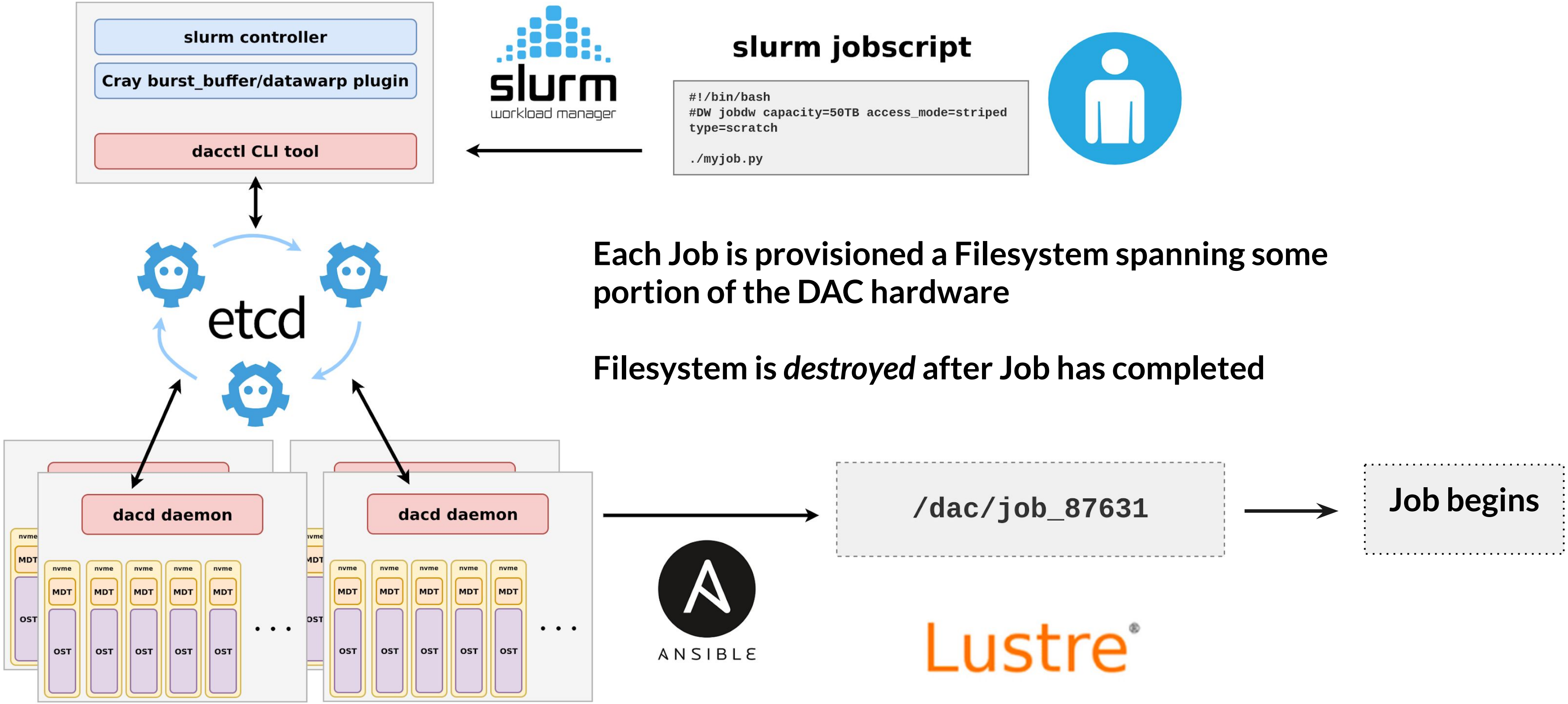
- ❖ **24x Dell R740xd Servers**

Each Server contains:

- ➢ 12x Intel SSD P4600 1.4TiB NVMe per server
- ➢ 2x Intel Omnipath HFIs @100Gbps per server
- ➢ 2x Intel Xeon Gold 6142 CPU 32C @2.60GHz
- ➢ 192GiB DDR4



How we use it - Lustre Filesystems-on-demand



slurm controller

Cray burst_buffer/datawarp plugin

dacctl CLI tool

slurm workload manager

slurm jobscript

```
#!/bin/bash
#DW jobdw capacity=50TB access_mode=striped
type=scratch

./myjob.py
```

etcd

dacd daemon

dacd daemon

nvme

MDT

OST

. . .

ANSIBLE

Lustre®

/dac/job_87631

Job begins

Each Job is provisioned a Filesystem spanning some portion of the DAC hardware

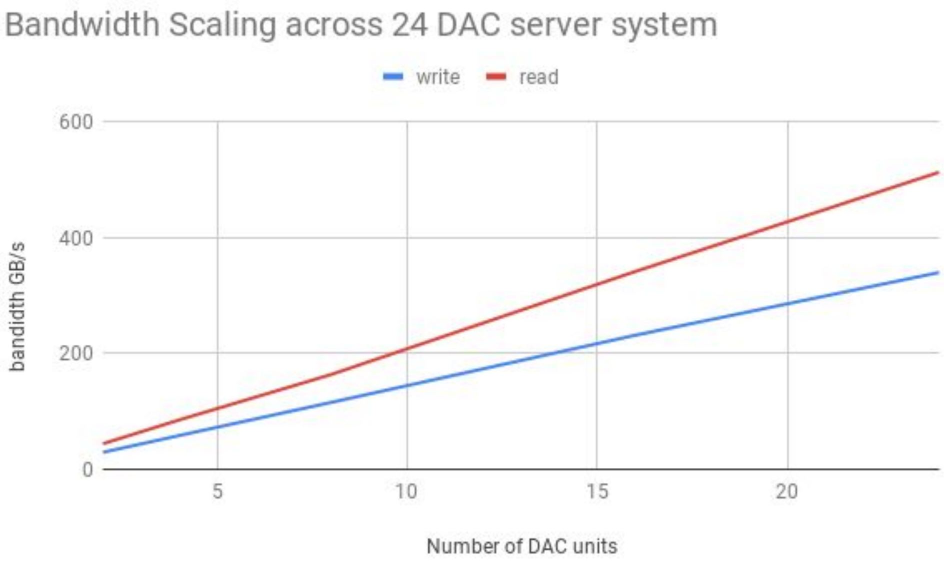
Filesystem is *destroyed* after Job has completed

DAC Software Project

https://github.com/RSE-Cambridge/data-acc

- Open-source. Developed in-house in collaboration with StackHPC (stackhpc.com)
- Repo contains installation instructions, as well as quickstart demo environments deployable with Docker or Openstack
- Core code written in Golang, along with Ansible to do Lustre filesystem creation/deletion
- Contributions/Feedback welcome!

Peak Performance - Flash is Fast

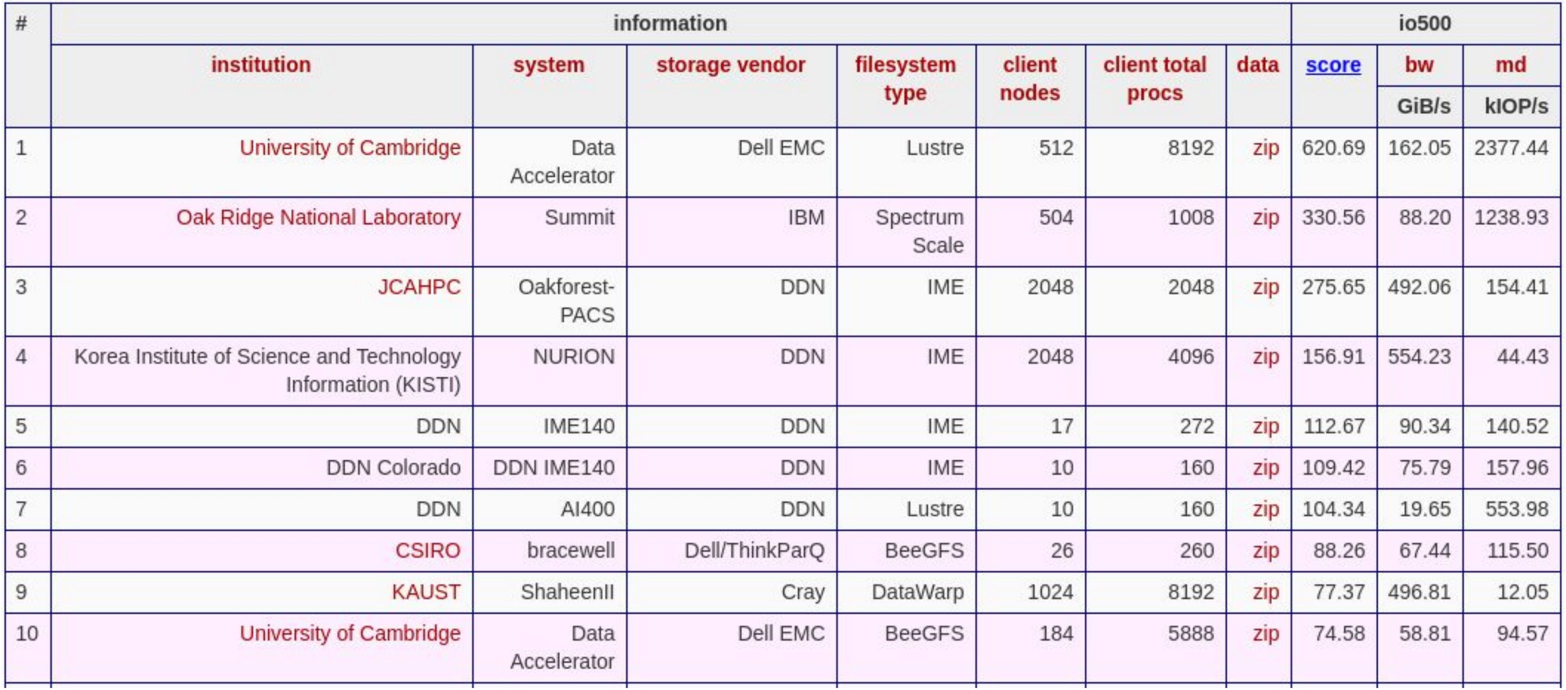


Some Headline Numbers:

- Best-case bandwidth
  (aligned, large streaming, file-per-process):
  - ➢ 530 GiB/s Read
  - ➢ 350 GiB/s Write
- Best-case metadata:
  - ➢ 1.9M file creates per second
  - ➢ 1.2M file deletes per second
- ‘find’ lookups (stat)
  - ➢ 2.2M IOPS
- Worst-case bandwidth
  (unaligned, small IO, single shared-file)
  - ➢ 80 GiB/s Read
  - ➢ 50 GiB/s Write

ISC’19 IO500 Results

Cambridge DAC took #1 Position in this HPC IO Benchmark Competition
https://www.vi4io.org/io500/list/19-06/start



#	information							io500		
	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
									GiB/s	kIOP/s
1	University of Cambridge	Data Accelerator	Dell EMC	Lustre	512	8192	zip	620.69	162.05	2377.44
2	Oak Ridge National Laboratory	Summit	IBM	Spectrum Scale	504	1008	zip	330.56	88.20	1238.93
3	JCAHPC	Oakforest-PACS	DDN	IME	2048	2048	zip	275.65	492.06	154.41
4	Korea Institute of Science and Technology Information (KISTI)	NURION	DDN	IME	2048	4096	zip	156.91	554.23	44.43
5	DDN	IME140	DDN	IME	17	272	zip	112.67	90.34	140.52
6	DDN Colorado	DDN IME140	DDN	IME	10	160	zip	109.42	75.79	157.96
7	DDN	AI400	DDN	Lustre	10	160	zip	104.34	19.65	553.98
8	CSIRO	bracewell	Dell/ThinkParQ	BeeGFS	26	260	zip	88.26	67.44	115.50
9	KAUST	ShaheenII	Cray	DataWarp	1024	8192	zip	77.37	496.81	12.05
10	University of Cambridge	Data Accelerator	Dell EMC	BeeGFS	184	5888	zip	74.58	58.81	94.57

In development - early-access users before Christmas

- Aim to start getting early users on the platform before Christmas
- More details to be announced through HPC users mailing list (hpc-user@lists.cam.ac.uk)
- If you have researchers with use-cases that could benefit, or would like to be considered for early-access trial, get in touch at support@hpc.cam.ac.uk