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

- Reasonably Fast
- Small
- Non-scalable under parallel IO
- 2 copies on disk
- Backed up
- All *Code* here

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

Data Accelerator (DAC)
~500TiB NVMe
~12PiB Lustre

Research Data Store (RDS)
~ 12PiB Lustre

?/
NFS /home
~1PiB ZFS

Research Cold Store (RCS)
~10PiB Tape

Performance Tier

Resilience Tier

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

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

Filesystem is destroyed after Job has completed

Job begins
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

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