Improved Solutions for I/O Provisioning and Application Acceleration

August 11, 2015

Jeff Sisilli
Sr. Director Product Marketing
jsisilli@ddn.com
Why Burst Buffer?  
The Supercomputing Tug-of-War

"A supercomputer is a device for turning compute-bound problems into I/O bound problems”
Ken Batcher, Emeritus Professor of Computer Science, Kent State University

"DDN’s storage mission is to eliminate I/O-bound problems and revert them back to compute-bound ones”
Alex Bouzari, CEO & Founder, DDN

The Divide Driving Exascale Innovation

© 2015 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others. Any statements or representations around future events are subject to change.
PAIN: “Problem” I/O Bound Applications
Longstanding PFS I/O Bottlenecks Must Be Eliminated

Current I/O Challenges at HPC Sites

75% of surveyed IT organizations face “problem applications” that are I/O bound in their environment.

Source: TechValidate survey of 118 users of current I/O challenges and solution requirements in HPC

TAKEAWAY
1. “Problem applications” are a huge source of known pain in HPC

HOW A BURST BUFFER HELPS
1. Accelerates applications and returns time available for computation by orders of magnitude
Polling HPC TOP500 Sites

RFP Mindshare of Various Flash-based I/O Acceleration Technologies

Which of the following are you including in upcoming RFP's to speed-up I/O and applications? Check all that Apply.

<table>
<thead>
<tr>
<th>Choice</th>
<th>Responses</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burst Buffer</td>
<td>40</td>
<td>44%</td>
</tr>
<tr>
<td>Flash and/or Edge Appliances</td>
<td>27</td>
<td>30%</td>
</tr>
<tr>
<td>All Flash Storage Arrays</td>
<td>23</td>
<td>25%</td>
</tr>
<tr>
<td>Hybrid (SSD + SAS) Arrays</td>
<td>63</td>
<td>69%</td>
</tr>
</tbody>
</table>
Current Motivation for I/O Accelerators
Faster Time to Results vs. Gaining New Efficiencies

For which purpose are you primarily considering these technologies?

<table>
<thead>
<tr>
<th>Choice</th>
<th>Responses</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster time to discovery, insight or results</td>
<td>57</td>
<td>58%</td>
</tr>
<tr>
<td>Reducing hardware, cost and footprint of provisioning bandwidth in traditional spinning disk approaches</td>
<td>60</td>
<td>61%</td>
</tr>
</tbody>
</table>
The Next I/O Provisioning Revolution:
Decoupling Physical Storage from Compute Resources!

BEFORE

Too Many:
- COMPUTE NODES
- DISKS, NETWORKING NODES, ARRAYS, ADMIN, H/W

EVERYTHING IS OVERPROVISIONED

AFTER

Much Fewer:
- COMPUTE NODES
- DISKS, NETWORKING NODES, ARRAYS, ADMIN, H/W

A LOT MORE SPEED TO THE APPLICATION & A LOT LESS COMPONENTS
Even Building the World’s Fastest PFS . . .
Will NOT Fix These I/O Challenges

PFS Locking

Storage Latency

Fragmented I/O Patterns

Out of Core Data

Limited Power

PFS are not designed for today’s mixed I/O & ensembles

HDD seek times & network traversing add latency

Mal-aligned apps slow down the PFS & entire cluster

Many datasets are too big for expensive DRAM

Exascale or next scale needs more space & power

No matter how many HDDs you add to a PFS, you can’t break I/O bottlenecks without a burst buffer
Burst Buffer & Beyond: IME®
The New I/O Acceleration Architecture

**AIM™ - Acceleration IN Memory**

Introducing AIM, An *Active I/O Tier*, inserted right between compute and your PFS

Intelligent IME software virtualizes disparate NVMe SSDs into a single pool of shared memory that accelerates I/O, PFS & Applications
Take AIM & Target I/O Bottlenecks With IME
Eliminate Overprovisioning & Storage Sprawl

- Slow PFS
- I/O Bound Applications
- Long Checkpoints
- POSIX Locking
- HDD Latency

- Big Footprints
- Large Datasets
- Low Capacity HDDs
- Latency
- High Power
Introducing IME®

Key Components and Operations

A thin IME Client resides on compute nodes or I/O nodes within the cluster. Client overloads (intercepts) I/O calls.

Traditional Lustre® (or GPFS™) Client model with MDS (or NSD) interface.

The Parallel File System (PFS) is unaware that IME may or may not be an intermediary.

Application interfaces can include: POSIX, MPI-IO, ROMIO, HDF5, NetCDF, etc. Application is unaware that it is accessing PFS or burst buffer. Application does not require modification of any type.

Low-level communications key value pair based protocol; is not POSIX. This allows for breakthrough performance and scalability.

IME Server software does the heavy lifting of managing the state.
IME: A Burst Buffer & Way, Way, Way Beyond
Game Changing, Enabling Technology

Cache is only the beginning. Right out of the box, IME does so much more . . .

- **BURST BUFFER**
  - Most cost & space efficient way to provision peak performance

- **PFS ACCELERATOR**
  - Finally breaks POSIX locking bottleneck with instant open/close

- **APP OPTIMIZER**
  - Dynamically aligns mal-formed I/O into striped writes without code mods

- **DRAM EXTENDER**
  - No dataset is too big with TBs or even PBs of fast, cost efficient NVMe
New Considerations for Architecting I/O Performance
IME introduces a more efficient way to provision performance than just storage arrays alone.

**BEFORE IME:**

PFS Systems were designed to handle the entire performance load. This required lots of storage controllers, enclosures and drives to deliver full bandwidth –

**STORAGE BANDWIDTH UTILIZATION OF A MAJOR HPC PRODUCTION STORAGE SYSTEM**
- 99% of the time < 33% of max
- 70% of the time < 5% of max

**TODAY, WITH IME:**

IME’s BURST BUFFER Absorbs the Peak Load
PARALLEL FILE SYSTEM Handles the Sustained Load

IME enables peak performance to be provisioned with much less hardware, power, space.
IME Accelerates I/O in Several Ways

“Problem Application” Case Study: S3D

1) MITIGATES POOR PFS PERFORMANCE caused by PFS locking, small I/O, and mal-aligned, fragmented I/O patterns.

IME “makes bad apps run well” and also prevents a poor-behaving app from impacting the entire supercomputer.

This is especially valuable to diverse workload environments and ISV applications.

At SC14, we demonstrated 1000x speed-up on mal-formed I/O when using non-POSIX low-level communications.

2) PROVIDES HIGHER PERFORMANCE I/O (bandwidth and latency) to the application.

Providing additional bandwidth here is relatively inexpensive. Configuring 10x more bandwidth compared to PFS is typical.

3) IME DRIVES I/O MORE EFFICIENTLY TO THE PFS by re-aligning and coalescing data within the non-volatile storage.

At SC14, we demonstrated 100x speed-up due to this efficiency. IOR benchmarks show a 3x – 20x speedup on I/Os <32KB.
1. IME enables organizations to separate the provisioning of peak & sustained performance requirements with greater operational efficiency and cost savings than utilizing exclusively disk-based parallel file systems.

IME Reduces Storage Hardware up to 70%:
- Fewer systems to buy, power manage, maintain

Storage Bandwidth Utilization of a Major HPC Production Storage System:
- 99% of the time < 33% of max
- 70% of the time < 5% of max
How Does IME Help?
Increases I/O & Application Performance

2. IME Accelerates applications, especially those with small or mal-aligned I/O for faster time to results & insight.
Thank You!

Keep in touch with us

sales@ddn.com
@ddn_limitless
company/datadirect-networks

2929 Patrick Henry Drive
Santa Clara, CA 95054
1.800.837.2298
1.818.700.4000

© 2015 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others. Any statements or representations around future events are subject to change.