Using Native NVMe-oF SSDs to Advance Computational Storage

Shahar Noy
Product Marketing, Marvell
# Introduction to Native NVMe-oF SSDs

<table>
<thead>
<tr>
<th>NVMe-oF SSD</th>
<th>NVMe SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface</td>
<td>PCIe and Ethernet</td>
</tr>
<tr>
<td><strong>Performance (Single port)</strong></td>
<td>50Gb/s</td>
</tr>
<tr>
<td></td>
<td>100Gb/s</td>
</tr>
<tr>
<td><strong>Performance (Dual Port)</strong></td>
<td>2x 25Gb/s</td>
</tr>
<tr>
<td></td>
<td>2x 50Gb/s</td>
</tr>
<tr>
<td><strong>System Power</strong></td>
<td>X</td>
</tr>
<tr>
<td><strong>System Cost</strong></td>
<td>X</td>
</tr>
</tbody>
</table>

(higher when factoring system reliability)

* 2U/24 System; Ethernet Switch; High Availability; 200Gb to 600Gb throughput; vs PCIe based fabric with X86 and NICs. The higher the throughput the greater are the savings.
The Challenges of Storage Controllers

Centralized System
- One component with non-autonomous parts
- Compute is shared
- Software runs in a single process
- Single point of control
- Single point of failure

Network
- Oversubscribed
Computational Storage To The Rescue

**Distributed System**
- Multiple autonomous components
- Compute is dedicated
- Software runs in concurrent processes
- Multiple points of control
- Multiple points of failure

**Network**
- Subscribed/Oversubscribed
Distributed Computational NVMe-oF SSDs

More Efficient Data Center
- Improved Economics
- Full Throughput
- Shared Resources
- Point-to-Point Communication
- Lower Up Front Cost
- High Utilization
- Better Reliability

What Causes Outages?
Percentage estimate of root causes by type

- Hardware Failure: 55.0%
- Human Error: 22.0%
- Software Failure: 18.0%
- Natural Disaster: 5.0%

Data source: Quorum via Seagate Report 2017
A Word on Datacenter Outages

Statistics

• 33% of Datacenters had at least 1 outage
• 80% of outages were preventable
• HA (2N Architecture) suffered 22% outage

• 1 Hour downtime = $260,000
Additional Benefits of Distributed Computational NVMe-oF SSD

- Object Store: CEPH Daemon
- Database: KV Interface
- Virtualization: Live Migration
Object Store: CEPH Daemon

Object Store Daemon (OSD) Read and Write Flow

1. Client app issues read request, RADOS sends request to primary OSD
2. Primary OSD reads data from local disk and completes read request

Source: https://blog.cypressxt.net/osd-performances-scaling/
Database: KV Interface

Protocol Optimization

Previous Drive Protocol Stack

New Drive Protocol Stack

Source: https://www.samsung.com/semiconductor/global.semi.static/Samsung_Key_Value_SSD_enables_High_Performance_Scaling0.pdf
Virtualization: Live Migration
Summary

Native Computational NVMe-oF SSD

Best Thing Since Sliced Bread!
Thank You!