Enabling Scalable Ethernet JBOF with a Native NVMe-oF™ SSD

Shingo Tanaka
Toshiba Memory Corporation
NVMe™ over Fabrics

Conventional system

• Direct Attached Storage (DAS)

NVMe-oF JBOF based system

• Disaggregated shared storage with DAS-like performance
• Better overall storage utilization

Drive-less Computing nodes

NVMe-oF JBOF
Challenge:
Break the Performance Bottleneck

768Gbps ➔ 1.5Tbps ➔ 3.0Tbps
Storage “Drive” Disaggregation

Solution: Native NVMe-oF SSD and Ethernet JBOF
Native NVMe-oF SSD

- NVMe-oF operation is guaranteed at SSD level
  - Overall cost reduction for end customers
- HW optimization with eliminating NVMe-oF bridge in the middle
Ethernet JBOF

**Existing JBOF**

- Remote Host
- Rack Switch
- Ethernet
- NIC
- CPU
- DRAM
- Sw
- PCIe
- SSD

**Bottleneck**

- NVMe-oF is realized by SW on CPU

**Limited performance**

- High CPU & NIC power consumption

**Ethernet JBOF**

- Remote Host
- Rack Switch
- Ethernet Switch
- Native NVMe-oF SSD

**Native performance**

- Extremely lower power consumption

**Simple & Scalable performance**
# Ethernet JBOF – Frontend Options –

<table>
<thead>
<tr>
<th></th>
<th>Ethernet Switch</th>
<th>Ethernet PHY</th>
<th>Ethernet Port Extender</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cost (BOM &amp; Power)</strong></td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Port Flexibility</strong></td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td><strong>Remarks</strong></td>
<td>Could omit rack switch in small deployments</td>
<td>Port/cable can be physically aggregated (c.f. 4x25G in a single QSFP28)</td>
<td>Best option for large deployments</td>
</tr>
</tbody>
</table>
Proof-of-Concept prototype

- Best-in-class performance w/ 24x SSDs
- Low cost & power for NVMe-oF bridging
- Scalable system performance

2.5” U.2 Native NVMe-oF SSD w/ NVMe-oF interposer
Dual 25GbE I/F

2RU 24xSSD HA Ethernet JBOF
Dual 6x 100GbE I/F

Demo at Toshiba Booth #307, Hall A
Preliminary TCO analysis

TCO comparison of NVMe-oF bridging & aggregation overhead (excluding SSDs)

TCO = CAPEX + OPEX

CAPEX: based on market prices in small quantity
OPEX: based on running cost estimate with power cost 0.1 $/KWh, system power efficiency 0.8, Lifetime 5 yrs
System level POC concept

- High performance storage system based on Ethernet JBOF with data services

**Data Services**
- Multi-tenancy
- High-availability
- Inline Deduplication
- Inline Compression
- Snapshot
- Thin-provisioning

**NVMe over Fabrics Solution**
Best-in-class performance

w/ Data Services
Thank You

Contact: shingo3.tanaka@toshiba.co.jp