NVMe-oF JBOF:
An ideal solution to integrate PCIe/NVMe SSDs in storage systems

At the heart of a new generation of data center infrastructures and appliances

Sept 2017
VIRTUALIZED DATACENTER: THE BLENDER EFFECT FOR STORAGE I/O OPERATIONS

10,000 VMs

Aggregate switch

10,000s of VMs generate millions of random IOPs on the storage side.

MIOPs in random

Centralized storage
NVMe SSDs: 
THE ANSWER TO THE RANDOM MIOPs DEMAND

- NVMe SSDs deliver 4000x better performances than traditional SAS HDDs.

<table>
<thead>
<tr>
<th>SAS HDD</th>
<th>SAS SSD</th>
<th>NVMe SSD</th>
</tr>
</thead>
</table>
| • 150 IOPs  
  • 6ms-200ms latency | • 200 KIOPs  
  • 140µs latency | • 800 KIOPs  
  • 115µs latency |

X 1000 more compute  
X4 more compute
FROM SAS TO NVMe
PCIe: The issues

- Primary designed for CPU-to-peripheral communication
- Not designed for Rack communication

Diagram showing:
- Compute Nodes
- Storage Servers
- Top of Rack Switch (ToR)
- Ethernet Switch
- FE Fabric
PCle : The issues

Adding JBOFs means adding head nodes

- Primary designed for CPU-to-peripheral
- Scale storage capacity (pay as you grow)
PCle : The issues

- Primary designed for CPU-to-peripheral
- Scale storage capacity (pay as you grow)
- Scale storage head nodes based on services
PCle : The issues

- Primary designed for CPU-to-peripheral
- Scale storage capacity (pay as you grow)
- Scale storage head nodes based on services
- Limited compute-to-storage ratio and flexibility

1 to 6 Head Nodes per JBOF
NVMe JBOF:
THE 2 OTHER SOLUTIONS EXISTING TODAY

X86 JBOF

- Lower Density
- High Cost/High Power

HYPERCONVERGED

- Compute/storage ratio is fixed
THE IDEAL SOLUTION: NVMe-oF JBOF

- Scale head nodes based on services
- Scale storage as needed
- Leverage existing PCIe JBOF designs
- High Density
- Cost/power optimized

Density of PCIe JBOF with the flexibility of x86 JBOF
KALRAY

NVMe-oF

Solution
NVMe-oF STORAGE SOLUTION: KALRAY TARGET CONTROLLER (KTC40/KTC80)

KALRAY TARGET CONTROLLER FUNCTION

TARGET CONTROLLER FEATURE

PCIe RC MODE FOR DIRECT SSD CONTROL
- Standard Linux with NVMe Driver
- Control up to 255 PCIe endpoints
- Any NVMe SSD supported – no need for CMB
- SSD Hot Plug Support

NVMe-oF PROTOCOL OVER RoCEv1/v2
- 4x + performant than SAS (IOPs & throughput)
- Scalability: Connect up to 2048 initiator cores
- standard ethernet connectivity

LOW ADDITIONAL LATENCY
- 15 µs for 4KB block transfer

BOARD MANAGEMENT CONTROL (BMC)
- Supervise enclosure

HIGH AVAILABILITY ARCHITECTURE
- End-to-end Multipath architecture

END USER INLINE PROCESSING
- Compression, Encryption, …

Manages all the storage functions of the new generation storage JBOF.
KTC40 & KTC80 HARDWARE SPECIFICATION

**KTC80**
- MPPA®2.2-256 (Bostan2 processor)
- 80 GbE of sustained throughput
- 2 x QSFP+ ports
- 16-lane PCIe Gen3
- 2 x DDR3-1866 with ECC (4GB)
- FHHL (Full-Height, Half-Length)
- Embedded switch with bifurcation up to 4 x 4-lane

**KTC40**
- MPPA®2.2-256 (Bostan2 processor)
- 40GbE
- 2 x QSFP+ ports
- 8-lane PCIe Gen3
- 2 x DDR3-1866 with ECC (2GB)
- LP (Low-profile)
KALRAY LEADS THE INDUSTRY IN NVMe-oF COMPATIBILITY

<table>
<thead>
<tr>
<th>OPERATION</th>
<th>KTC40</th>
<th>KTC80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet ↔ SSD (NVMe Direct/Root Complex)</td>
<td>1.6 MIOPs</td>
<td>3.2 MIOPs</td>
</tr>
<tr>
<td>67%RD / 33% WR @4KB</td>
<td>15 µs latency</td>
<td>15 µs latency</td>
</tr>
</tbody>
</table>

Highest possible throughput.

A whole family of products.
KALRAY I/O BOSTAN PROCESSOR OVERVIEW

HIGH-SPEED INTERFACES:
• 2x 40GbE
• 2x PCIe Gen3 8-lanes (EP/RC)

CONNECTED TO A LARGE ARRAY OF PROCESSING
• Full C/C++ Programmable
• Dataplane execution

VIA A HIGH BANDWIDTH LOW LATENCY NETWORK ON CHIP
• Direct packet-to-core delivery
• Direct core-to-core transfers
• Direct connect between multiple MPPAs

AND I/O Quad CORES
• Runs Linux
• Runs control plane
KTC NVMe-oF SOFTWARE STACK
END USER CUSTOMIZABLE SOLUTION

CUSTOMIZABLE FUNCTIONS

INLINE PROCESSING
- Compression
- Encryption
- Deduplication
- Erasure Coding

BOARD MANAGEMENT CONTROL (BMC)
- REDFISH/SWORDFISH
- SES
- openBMC

END USER READ/WRITE OPERATION POLICY
- Implement optimized Read/write scheduling to eliminate outliers on critical streams
- Achieve a low latency for 99.9999%
YOUR PCIe JBOF EASILY BECOMES AN ETHERNET JBOF WITH KALRAY TARGET CONTROLLER

PCIe JBOF  \[\text{No Modifications}\]  NVMe-oF JBOF

KTC ENABLES A FAST TIME-TO-MARKET TO BUILD NVMe-oF JBOF
INITIATOR BOTTLENECK SOLUTION

PCle JBOF

<table>
<thead>
<tr>
<th>MIOPs</th>
<th>Initiator cores</th>
<th>KIOPS/Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>432</td>
<td>28</td>
</tr>
</tbody>
</table>

NVMe-oF JBOF

<table>
<thead>
<tr>
<th>MIOPs</th>
<th>Initiator cores</th>
<th>KIOPS/Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>24,576</td>
<td>1</td>
</tr>
</tbody>
</table>

Up to 6 x 72 Initiator cores
Up to 6 x 255 PCle End points

NVMe-oF KTC connects to 28x more initiator cores than PCle adapters. This solves the initiator bottleneck issue!
NVMe-oF JBOF: Scales the storage capacity
KTC ENABLES HIGH AVAILABILITY ARCHITECTURE

END-TO-END REDUNDANT PATH

- Dual port U.2 NVME SSD
- Dual PCIe Trees
- Dual KTC40/80 connectivity

MULTIPATH HANDLED AT THE INITIATOR SIDE

- Standard feature available in Linux Kernel
- Support Active-Active or Active-Standby modes
SCALE PERFORMANCE UP TO SSD PEAK CAPABILITIES

Global performances

- 6.4 MIOPs
- 17.6 GB/s

JBOF

- 6.4 MIOPs
- 17.6 GB/s

- 15 MIOPs
- 48 GB/s

Scale up to SSD peak performances

Global performances

- 19.2 MIOPs
- 52.8 GB/s

- 15 MIOPs
- 48 GB/s
x86-based JBOF Versus KTC-based JBOF: performance optimized

**X86 JBOF architecture**

- **DENSITY:** 24 SSDs in 2U (77TB)
- **CPU + NIC FUNCTION**
  - 2 x XEON E5-2667v4
  - 8 x 16GB DDR4
  - 3 x 100G NIC
- **POWER:** 309 W
- **PERFORMANCE:** 9.4 MIOPs

**KTC-based architecture**

- **DENSITY:** 24 SSDs in 2U (77TB)
- **CPU + NIC FUNCTION**
  - 6 x KTC80
- **POWER:** 210 W
- **PERFORMANCE:** 15 MIOPS

**Same density.**

**More power efficient.**

- **32%**

**Higher performance**

- **60%**

ELIMINATE THE HIGH COST/ HIGH POWER x86 SYSTEM (CPU, MEMORY, …) WHILE INCREASING THE PERFORMANCES BY 60%
x86-based JBOF Versus KTC-based JBOF: density optimized

**DENSITY:** 154TB in 2U

**SPECIFICATION**
- 2 x XEON E5-2667v4
- 8 x 16GB DDR4
- 2 x 100G NIC

**POWER:** 294 W

**PERFORMANCE:** 6.25 MIOPs

**Better performance.**

20%

**More Power effective.**

64%

**Greater density.**

58%

**DENSITY:** 240 TB in 2OU

**SPECIFICATION**
- CHASSIS WITH 250 M.2 SSD in 2OU
- 3 x KTC80-LP

**POWER:** 105 W

**PERFORMANCE:** 7.5 MIOPS

**ELIMINATE THE HIGH COST/ HIGH POWER x86 SYSTEM (CPU, MEMORY, …) WHILE INCREASING DENSITY AND OPTIMIZING COST AND POWER**
**STORAGE: PAY AS YOU GROW WITH KALRAY TARGET CONTROLLER**

**KALRAY UNIQUE ADVANTAGE:**

- **CHAIN NVMe-oF JBOFs**

**KEEP THE SAME INFRASTRUCTURE**
- ToR switch
- Number of storage servers

**PAY AS YOU GROW!**
- Pay only for additional storage capacity
- Not for additional storage servers or Top of Rack Switch

The chaining equivalent to SAS protocol.
NVMe-oF JBOF ENABLES DISAGREGATED HYPERCONVERGED ARCHITECTURE

Hyperconverged / SDS

- Hyperconverged/SDS scales naturally
- Compute/storage ratio is fixed
- DAS is expansive

Disaggregated Hyperconverged/SDS

- Scale compute & storage independently
- Leverage existing PCIe JBOF designs
Conclusion

How Kalray’s NUMe-oF can benefit you?
NVMe-oF : THE SOLUTION FOR NEW GENERATION OF STORAGE SYSTEMS

**NVMe-oF TARGET**

**THE SOLUTION FOR NVMe-oF JBOF**

Eliminate the need of X86 and associated system memory

**4X HIGHER IOPS THAN SAS SSD**

End-to-end NVMe/NVMe-oF capabilities ensure 4X more IOPS

**SCALABLE & FLEXIBLE**

Scale the Head Nodes and Storage capacity independently

**$$\text{FAST TIME TO MARKET}$$**

Plug NVMe-oF Target controller in your standard PCIe JBOF
MPPA, ACCESSCORE and the Kalray logo are trademarks or registered trademarks of Kalray in various countries. All trademarks, service marks, and trade names are the marks of the respective owner(s), and any unauthorized use thereof is strictly prohibited. All terms and prices are indicative and subject to any modification without notice.