Challenges and Effects of EDSFF-based NVMe-oF Storage Solution
Challenges and Opportunities of EDSFF based Storage Solution

Jungsoo Kim, System Architect, Samsung Electronics
Duckho Bae, Software Architect, Samsung Electronics
Agenda

- EDSFF Benefits
- Poseidon V1 – E1.S Reference System
- Poseidon V2 – E3.X Reference System
- Poseidon Storage OS – NVMe-oF Solution
EDSFF – Benefits

- Designed to overcome conventional device limitations
- Improves thermals, power, and scalability
- Various power options - 25W, 40W, 70W
- Various PCIe lane options - x2, x4, x8, x16
- Advantageous for high-speed interface (< PCIe Gen5)
- Built in LEDs, carrier-less design

Source: https://www.snia.org/forums/cmsi/knowledge/formfactors
EDSFF Reference System

- E1 and E3 based system can increase the performance and density
- Have more flexibilities than a traditional system

Against comparison of different SSD configurations:
- U.2 SSD x 10ea vs. E1.S SSD x 32ea
- U.2 SSD x 24ea vs. E3.S SSD x 40ea

Diagram showing components:
- SmartSSD
- E3.S SSD
- CXL Memory
- NIC

Open Possibilities.
Poseidon Project

- OCP based industrial collaboration b/w “Component Vendor↔System Vendor ↔Data Center”
- Open-source HW & SW project to expand NVMe eco-system
E1.S Reference System – PSD V1

**Front View**

1. E1.S SSD (5.9/8.0/9.5) 32ea
2. 32 E1.S BP 1
3. PSU 2ea
4. IO Module 1ea
5. FAN 8ea
6. MB 1ea
7. FHHL Card 2ea
8. OCP NIC V3 1ea
9. NVDIMM Power Module 2ea

**Motherboard**

- Support 1 x16 lanes PCIe Card
- Genz X16

**Backplane**

- Support 1 x16 lanes PCIe Card
- Genz X16

**Support 1 x16 lanes PCIe Card**

- Slimline X8
- Slimline X8
- Slimline X8
- Slimline X8
- Slimline X8
- Slimline X8
- Slimline X8
- Slimline X8
- Slimline X8

- PCIe Cables
- PCIe Cables
- PCIe Cables
- PCIe Cables
- PCIe Cables

- SLV 32 x32 lanes
- SLV 32 x32 lanes
- SLV 32 x32 lanes
- SLV 32 x32 lanes
- SLV 32 x32 lanes
- SLV 32 x32 lanes

- CPU X16
- CPU X16
- CPU X16

- Genz X16
- Genz X16
- Genz X16

- CPU1 X16
- CPU1 X16
- CPU1 X16

- OCP NIC
- OCP NIC
- OCP NIC

- 4pin I2C Comm
- 4pin I2C Comm
- 4pin I2C Comm

- BMC Ast25000
- BMC Ast25000
- BMC Ast25000

- 1 PC9548 1
- 2 PC9548 2
- 1 PC9548 3

- JTAG TCK
- JTAG TCK
- JTAG TCK

- Thermal Sensor
- Thermal Sensor
- Thermal Sensor

- IFX_PCIE x8
- IFX_PCIE x8
- IFX_PCIE x8

- IFX_PCIE x8
- IFX_PCIE x8
- IFX_PCIE x8

- I2C_SCL SDA
- I2C_SCL SDA
- I2C_SCL SDA

- I2C Switch
- I2C Switch
- I2C Switch

- Amber LED on every E1.S SSD is used
- Locate Error
- JTAG Header

- 2 I2C Switch
- 2 I2C Switch
- 2 I2C Switch

- Mos
- Mos
- Mos

- CPA9548 0 E1.S 0-7
- CPA9548 3 E1.S 24-31
- CPA9548 3 E1.S 24-31

- FRU
- FRU
- FRU

- Downstream x64 lanes
- Downstream x64 lanes
- Downstream x64 lanes

- UPC_PCIE x8
- UPC_PCIE x8
- UPC_PCIE x8

- E1.S Connector 0
- E1.S Connector 15
- E1.S Connector 15

- E1.S Connector 16
- E1.S Connector 31
- E1.S Connector 31

- E1.S Connector 31
IO Performance – Single vs Aggregated

- Each 32 SSDs shows stable IO performance around 6.4GB/s
- Aggregated IO performance achieved > 102GB/s
- Lanes b/w CPU↔PCIe Switch (x16) are saturated with 4 SSDs

*Theoretical B/W limit: 128GB/s

---

**Samsung PM9A3 Specification**

- Form factor: E1.5
- Capacity: 960 GB, 1.92 TB, 3.82 TB, 7.68 TB
- Sequential read: Up to 6,500 MB/s
- Sequential write: Up to 3,200 MB/s
- Random read: Up to 900,000 IOPS
- Random write: Up to 150,000 IOPS
- Physical Dimensions: 31.5 x 111.49 x 5.9 mm
- Power consumption: Read: <= 9.7W, Write: <= 11.7W
- Host interface: PCIe Gen 4 x4
IO Performance – Raid 0&5

- Compared Raid 0 & 5 performance of 4 x NVMe SSDs
- Observed significant performance degradation on Raid 5 write operation
- Further R&D would be need for S/W Raid on high density system

**OS:** Ubuntu 20.04.2 LTS  
**Kernel:** Linux 5.4.0-65-generc  
**Num. of threads:** 112 (56 * 2)  
**Model name:** Intel Xeon Gold 6330N CPU @ 2.20GHz  
**fio:** fio=3.16
DC customers want to improve serviceability in their datacenter by removing the screws. E1.S + extension kit with screws are the only option in the market, and we developed the innovative new tool-less ext. kit design to satisfy the requirements.

Current Design
- Screw type

Tool-less Design
- Clip type

Compatible w/ a screw type extension kit!

Screw Type (40sec) Tool-less Type (4sec)
E3 Reference System – PSD V2

- Designed to maximize the benefits of E3.x form factor
- Can configure the system according to application’s needs

**[High Density Storage]**
- E3.S SSD
- Type1: 24x E3.S 1T (PCIe x4)

**[Memory Cache / Cloud]**
- CXL DRAM
- SSD
- CXL DRAM
- Type2: 8x E3.S 1T + 8x E3.S 2T (PCIe x8)

**[AI / Video Transcoding]**
- SmartSSD
- SSD
- Accelerator

**[E3.S Gen5 SSD]**

**[CXL Memory Expander]**

**[Smart SSD]**

*Above E3 SmartSSD picture is for illustration purpose only*
Poseidon V2 – PCIe/CXL Topology

- PCIe & CXL topology varies depending on the system mode
PoseidonOS

- Open-Source NVMe-oF Solution for Disaggregated Datacenter Storage Solution
  - Develop NVMe Technology and Eco-system
  - Improve Datacenter Storage Efficiency and Performance
  - Leverage the potential of NAND Flash

※ Up to few hundreds
PoseidonOS Features

• User-space storage OS for NVMe-oF
• Provide PCIe Gen4 < performance via network
  – Up to 200GbE
• Support valuable storage features
  – NUMA-Aware, Volume Mgmt, Perf Throttling, SW RAID, ⋯
• Easily integrate with upper orchestration layers
  – RESTful, CSI, ⋯
Performance Numbers

- Achieved up to 200GbE Performance via NVMe/TCP
- Random Write has room for improvement

<table>
<thead>
<tr>
<th></th>
<th>Seq. Read</th>
<th>Seq. Write</th>
<th>Rnd. Read</th>
<th>Rnd. Write</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq. Read</td>
<td>24.5GB/s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seq. Write</td>
<td>18.6GB/s</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rnd. Read</td>
<td></td>
<td>19.4GB/s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rnd. Write</td>
<td></td>
<td></td>
<td>8.7GB/s</td>
<td></td>
</tr>
</tbody>
</table>
Poseidon in Datacenter

Compute Pool

Storage Pool

V2 (Type 2) / SSD
V2 (Type 2) / CXL
V2 (Type 2) / Smart

V1 / SSD
V2 (Type1) / SSD
V1 / ZNS
V2 (Type1) / ZNS

PoseidonOS

OPEN POSSIBILITIES.
Trident : Test Tool for Poseidon OS

- Helps open source users quickly setup and explore all features of Poseidon OS (POS)
- Planned to be open sourced to compliment the user documentation provided with POS
- Enables community users to quickly validate their lab setup
- Along with POS, can also be deployed on Qemu/KVM VMs and emulated NVMe devices
- Includes Test cases that are light-weight and intuitive as they exploit features of pytest framework like fixtures and parametrization.

Test Framework and Lab
Future Work

• Poseidon V2 contribution in 2022
• Tool-less design upgrade and datacenter adoption
• Support innovative devices (ex. ZNS, QLC)
• Support more features/provide developers toolkits
• Enable PCIe Gen5 performance
• Available at Github

https://github.com/poseidonos/poseidonos
Call to Action

• Timeline for Contribution Availability
  • PM9A3 OCP SSD (‘21 Q1) – OCP Inspired
  • Poseidon V1 (‘21 Q3) – OCP Inspired
  • Tool-less SSD Design (‘21 Q3) – OCP Inspired
  • Poseidon V2 (Target in ‘22 Q3)
• Timeline for Product/Facility Availability
• Link to Contribution DB/OCP Marketplace
• Samsung PM9A3 SSD complies with the Open Compute Project (OCP) NVMe Cloud SSD (https://www.samsung.com/semiconductor/ssd/pm9a3/)
• Poseidon OS github: https://github.com/poseidonos/poseidonos
Thank you!