OPEN. FOR BUSINESS.
CINABRO: a Software Driven, Open Flash Array Architecture for Scalable Cloud Storage Services

Sungjoon Ahn, VP of Engineering, Circuit Blvd., Inc.
Motivation

1. Serves diverse cloud storage requirements
   - Data center workloads are dynamic, diverse and constantly evolving
   - Data center SSDs typically run 3 to 5 years after rigorous qualification process
   - SSD FW update is expensive and usually limited to critical bug fixing

2. Streamlines flash memory deployments
   - SSD designs optimized for single self contained units
   - Data center SSDs often have old generation NANDs
   - Need for deploying latest NANDs in scale

Solution

- Lower Total BOM and Simpler SSD Device Architecture
- Large portions of SSD intelligence run on host server CPU

<table>
<thead>
<tr>
<th>Category</th>
<th>Conventional SSD</th>
<th>Cinabro SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1GB DRAM</td>
<td>$$</td>
<td>X</td>
</tr>
<tr>
<td>1TB NAND</td>
<td>$$$</td>
<td>$$$</td>
</tr>
<tr>
<td>SoC Controller</td>
<td>$$</td>
<td>$</td>
</tr>
<tr>
<td>Capacitors (x20) – Power Loss Protection</td>
<td>$</td>
<td>X</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Development Complexity</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>
Cinabro™ System Architecture

Disaggregated and composable All Flash Array based on COTS server
Leverage OCSSD Standard to Provide Optimized Solution with Simpler ASIC

- **Open-Channel SSD (OCSSD)**
  - Standard NVMe based protocol
  - Facilitates host FTLs and good fit for cloud providers

- **Optimized protocol translation between host and NAND interface**

- **Performance acceleration and reliability enhancement features for 3D NAND TLC/QLC**

- **Cost and power efficient ASIC design**
CBBridge™ OCSSD Controller

Simple but robust SoC handling essential NAND media functions

- Open-Channel SSD spec and additional features for cross layer optimizations
- 28nm process technology accommodates 96+ layer 3D toggle3 TLC/QLC NAND with LDPC

System Performance: G3 2-L
Network Interface

Leverage New Standard for Networked Storage Interfaces

- NVMe-over-Fabrics (NVMe-oF)
  - Faster access between hosts and storage systems
  - Much lower latency than iSCSI

- Flexible system design to support various fabrics of NVMe-oF standard (Ethernet, Infiniband, etc.)

- Open architecture allows incorporating new system technologies (e.g. SmartNIC, FPGA acceleration, SDN)

- Seamless integration with Open-Channel SSDs
Software Design

Advanced Open Source Software Optimized for All Flash Array

- Host-managed array FTLs
- User-level device driver configurable/adapting to various workload
- Scalable design to manage array of NAND modules
- Leverage multi-core/multi-processor CPU to maximize parallelism
- Data center friendly orchestration utilizing Linux Containers and Kubernetes Ready design
CBOS™ Software Architecture

Container based storage and application software modules
• Host based Flash Array FTLs come with libraries that apps can pick and choose
CBOS™ Container Example

1. NVMe-oF Target
2. High Performance Computing
3. In Storage Processing

- 2 Socket, 4 Core System
- Each CPU core handles 8 OCSSD
- Per different application needs, matching FTL container is deployed
- CBOS storage containers run on available CPU cores executing managerial tasks
## Development Milestones

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CINABRO™ Appliance</strong></td>
<td><strong>First Working Prototype</strong>&lt;br&gt;• Commodity x86 server&lt;br&gt;• x4 Cosmos OpenSSD Mini PCIe cards&lt;br&gt;• Mellanox RDMA network cards</td>
<td><strong>Alpha</strong>&lt;br&gt;• Commodity x86 server&lt;br&gt;• x8 Cosmos OpenSSD Ultra PCIe cards&lt;br&gt;• NVMe-oF network cards: TBD</td>
<td><strong>Beta</strong>&lt;br&gt;• Customized PCIe fabrics&lt;br&gt;• Commodity CPU, DRAM, PCIe&lt;br&gt;• BMC: ready; Backup power: TBD</td>
</tr>
<tr>
<td><strong>CBBridge™</strong></td>
<td><strong>FPGA1</strong>&lt;br&gt;• FPGA code with OCSSD compliant FW&lt;br&gt;• 16nm 2D MLC NAND w/ BCH</td>
<td><strong>FPGA2</strong>&lt;br&gt;• RTL release: CBBridge™ SoC ready. Runs on FPGA&lt;br&gt;• 64L 3D TLC NAND w/ LDPC</td>
<td><strong>ASIC</strong>&lt;br&gt;• SoC tape-out (mid ’19)&lt;br&gt;• 96L 3D TLC/QLC NAND w/ LDPC</td>
</tr>
<tr>
<td><strong>CBOS™</strong></td>
<td><strong>Prototype release</strong>&lt;br&gt;• NVMe-oF drive interface&lt;br&gt;• Baseline data path working:&lt;br&gt;• OCSSD pblk / lightnvm, NVMe-oF, SPDK/DPDK</td>
<td><strong>Alpha release</strong>&lt;br&gt;• Host-based flash array FTLs&lt;br&gt;• Storage management layer&lt;br&gt;• Application plugins&lt;br&gt;• System resource &amp; performance manager design complete</td>
<td><strong>Beta release</strong>&lt;br&gt;• Core feature complete&lt;br&gt;• OpenStack compliant&lt;br&gt;• Data management beta&lt;br&gt;• System resource &amp; performance Manager beta&lt;br&gt;• Out-of-Band management beta</td>
</tr>
<tr>
<td><strong>Open Source</strong></td>
<td><strong>SPDK contribution</strong>&lt;br&gt;• Functions to help writing OCSSD access from user level&lt;br&gt;• Included in SPDK v17.10, v18.01</td>
<td><strong>R&amp;D version alpha</strong>&lt;br&gt;• OpenSSD FPGA RTL codes v1.2&lt;br&gt;• Developer edition CBOS™ alpha: includes device drivers, user level libraries, and pilot apps</td>
<td><strong>R&amp;D version beta</strong>&lt;br&gt;• OpenSSD FPGA RTL codes v1.3&lt;br&gt;• Developer edition CBOS™ beta</td>
</tr>
</tbody>
</table>
Current Prototype

Multiple FPGA based OCSSDs running in our lab

- 2 Socket Intel Xeon Server (CINABRO node running CBOS™)
- Cosmos OpenSSD Mini FPGA Boards (Open-Channel SSD)
- Mellanox CX5 40/100Gb RoCE
- Mellanox CX4 40/100Gb RoCE
- 1 Socket Intel Xeon Server (Admin/App node)
- Mellanox CX3 40Gb RoCE
- 1 Socket Intel Xeon Server (App node running Video Streaming Servers on Containers)
- 40/100Gb Network
- Client PC (Browsers)
- Dashboards

Current Prototype
Multiple FPGA based OCSSDs running in our lab

- 2 Socket Intel Xeon Server (CINABRO node running CBOS™)
- Cosmos OpenSSD Mini FPGA Boards (Open-Channel SSD)
- Mellanox CX5 40/100Gb RoCE
- Mellanox CX4 40/100Gb RoCE
- 1 Socket Intel Xeon Server (Admin/App node)
- Mellanox CX3 40Gb RoCE
- 1 Socket Intel Xeon Server (App node running Video Streaming Servers on Containers)
- 40/100Gb Network
- Client PC (Browsers)
- Dashboards
Prototype FTL Evaluations

Our 1st host FTLs, UBM and GFTL*, have been implemented in SPDK.

1. Array with up to 4 units of FPGA OCSSDs:
   - Showing reasonable performance for FPGA based SSDs

2. Array with up to 24 units of OCSSD Qemu-nvme emulators:
   - Used for qualitative test of multiple applications
   - 24 copies of GFTL working correctly over same number of emulated OCSSDs

* GFTL is based on Hanyang University’s Greedy FTL.
## Prototype Applications

### End-to-end integration tests with multiple applications

<table>
<thead>
<tr>
<th>Containerized Video Server</th>
<th>Containerized RocksDB</th>
<th>SK telecom’s AF Ceph</th>
</tr>
</thead>
<tbody>
<tr>
<td>• OCSSD arrays host movie files and are exposed via SPDK nvmf_tgt containers</td>
<td>• Containerized RocksDB, both local and remote (over NVMe-oF)</td>
<td>• All Flash Ceph is flash optimized version of Ceph.</td>
</tr>
<tr>
<td>• Video servers run inside containers, made of Nginx web server with RTMP module</td>
<td>• Local: SPDK’s RocksDB plugin over CBOS™ GFTL</td>
<td>• Initial data verification test over 4 FPGA OCSSDs ran successfully.</td>
</tr>
<tr>
<td></td>
<td>• Remote: SPDK’s nvmf-tgt over CBOS™ GFTL</td>
<td></td>
</tr>
</tbody>
</table>

- [8 concurrent video server example]
- [24 RocksDB plugin per CPU core example]
- [AF Ceph over 4 OCSSDs example]
Summary

Solution Benefits
• Flexibility to accommodate NAND generations from various vendors
• Adaptable to various Cloud Data Center network infrastructure
• Customizable SW architecture to meet ever-evolving cloud data center requirements

Communities
• OpenSSD, OCSSD, SPDK: Our work has been integrated
• OCP: open to collaboration about making our hardware design available to the community

Resources
• OpenSSD FPGA SSD available at: http://openssd.io
• SPDK OCSSD contributed codes: https://github.com/spdk/spdk.git (SPDK v17.10, v18.01)
• CBOSTM development edition codes: TBD