OCP SUMMIT

March 20-21
2018
San Jose, CA

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

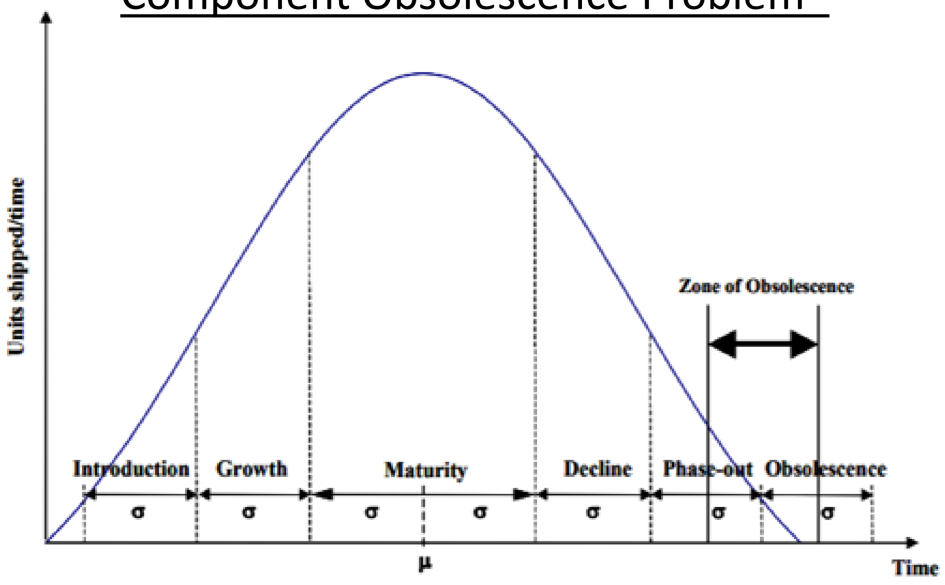
1. Serves diverse cloud storage requirements

- 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

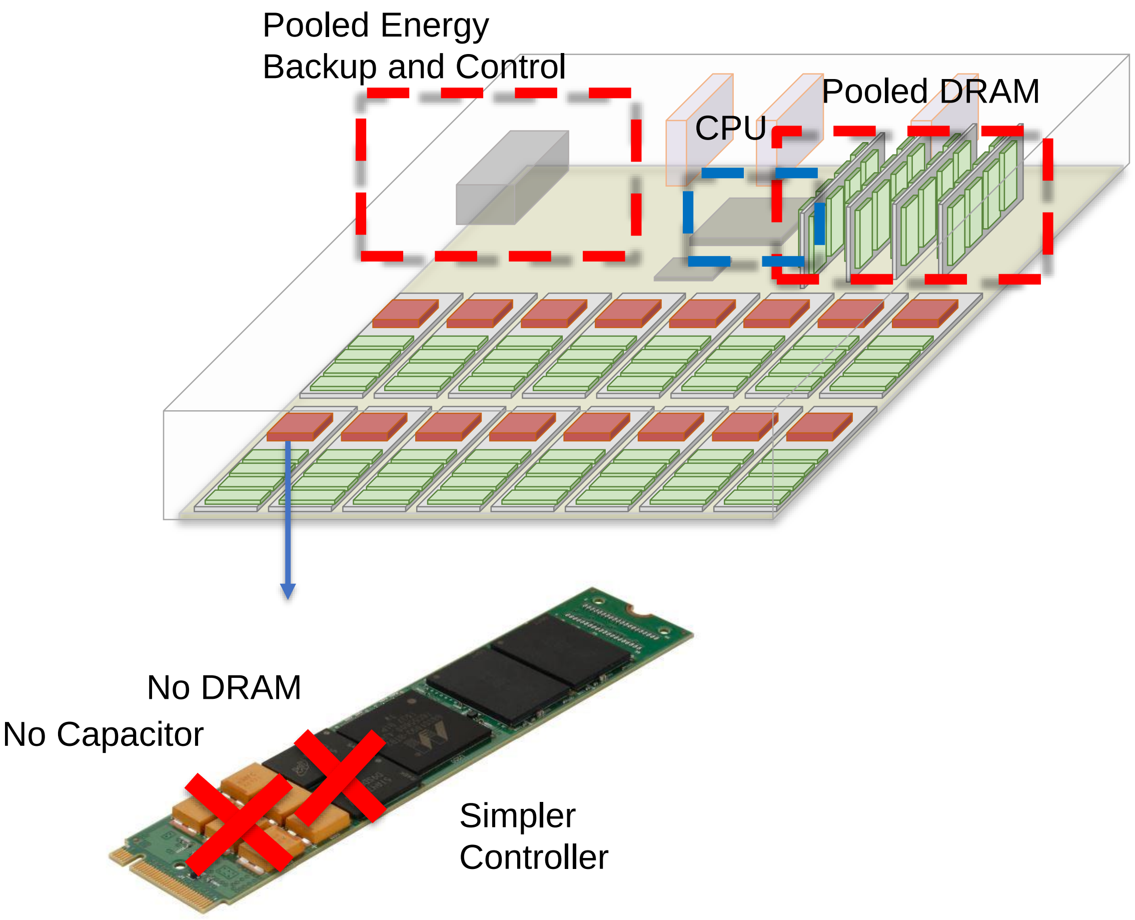
Component Obsolescence Problem*



* Source: Electronic Part Life Cycle Concepts and Obsolescence Forecasting, Solomon, Sandborn, and Pecht

Solution

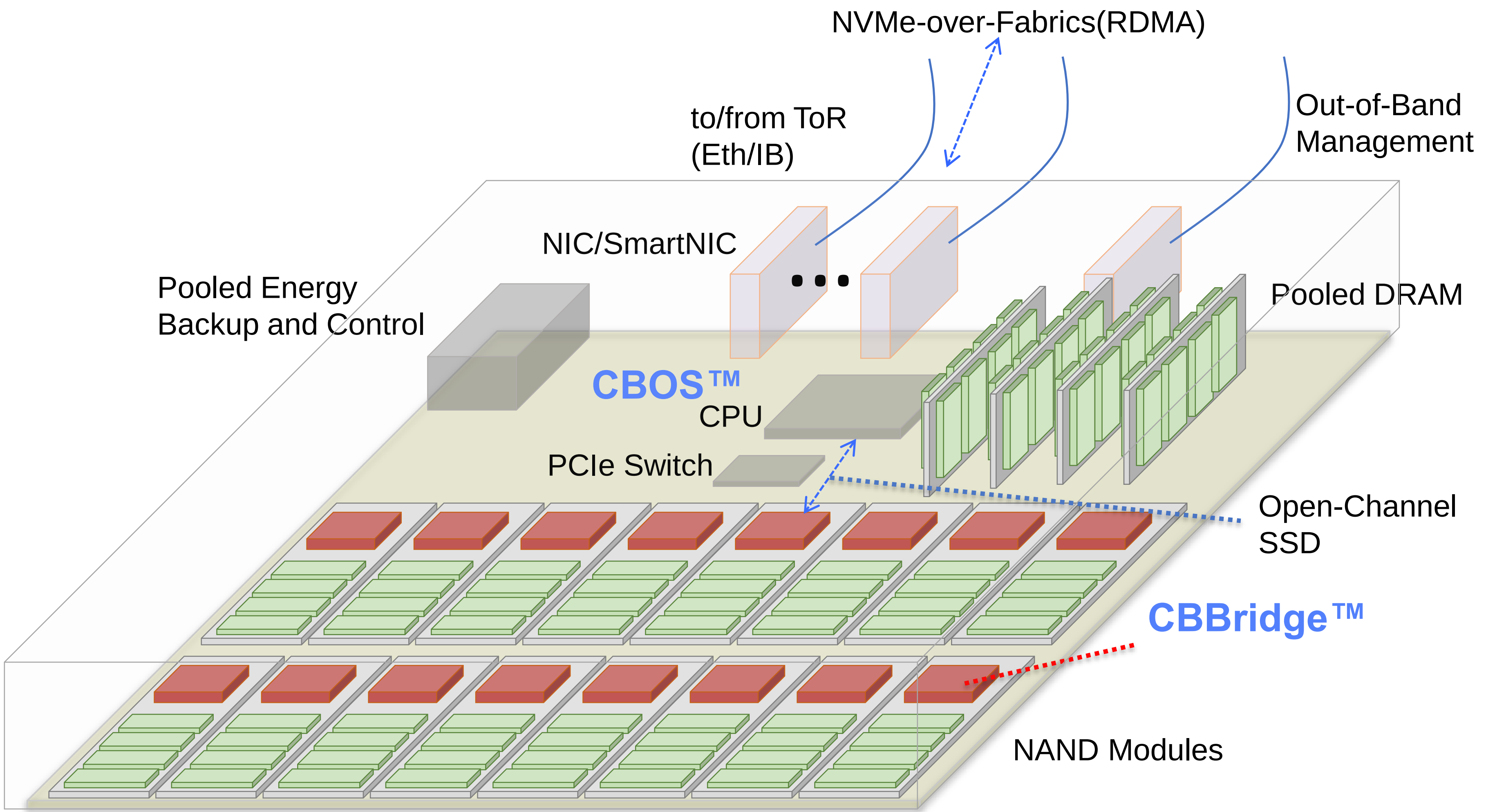
- Lower Total BOM and Simpler SSD Device Architecture
- Large portions of SSD intelligence run on host server CPU



Category	Conventional SSD	Cinabro SSD
1GB DRAM	$$	X
1TB NAND	$$$	**$$$**
SoC Controller	$$	**$**
Capacitors (x20) – Power Loss Protection	$	X
Power Consumption	High	**Low**
Development Complexity	High	**Low**

Cinabro™ System Architecture

Disaggregated and composable All Flash Array based on COTS server

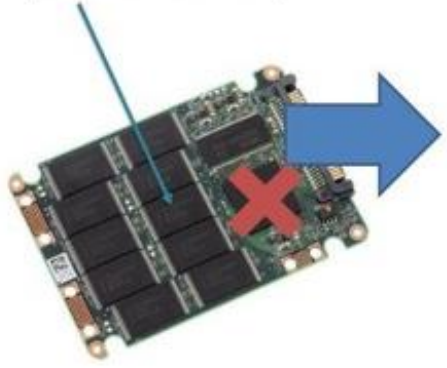


CIRCUITBLVD
Change your memories. Change your life.

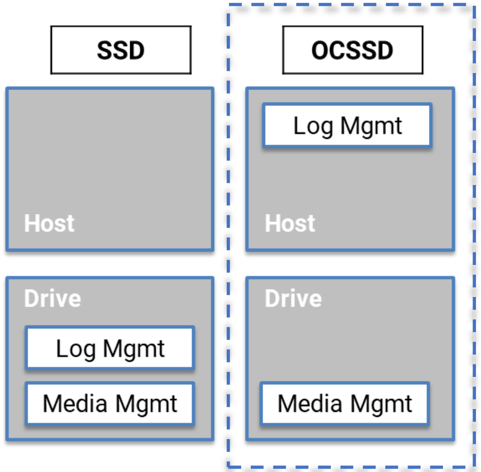
NAND Flash Interface

Leverage OCSSD Standard to Provide Optimized Solution with Simpler ASIC

Physical flash exposed to the host (Read, write, erase)



Host
- Data placement
- IO Scheduling
- Over-provisioning
- Garbage collection
- Wear levelling



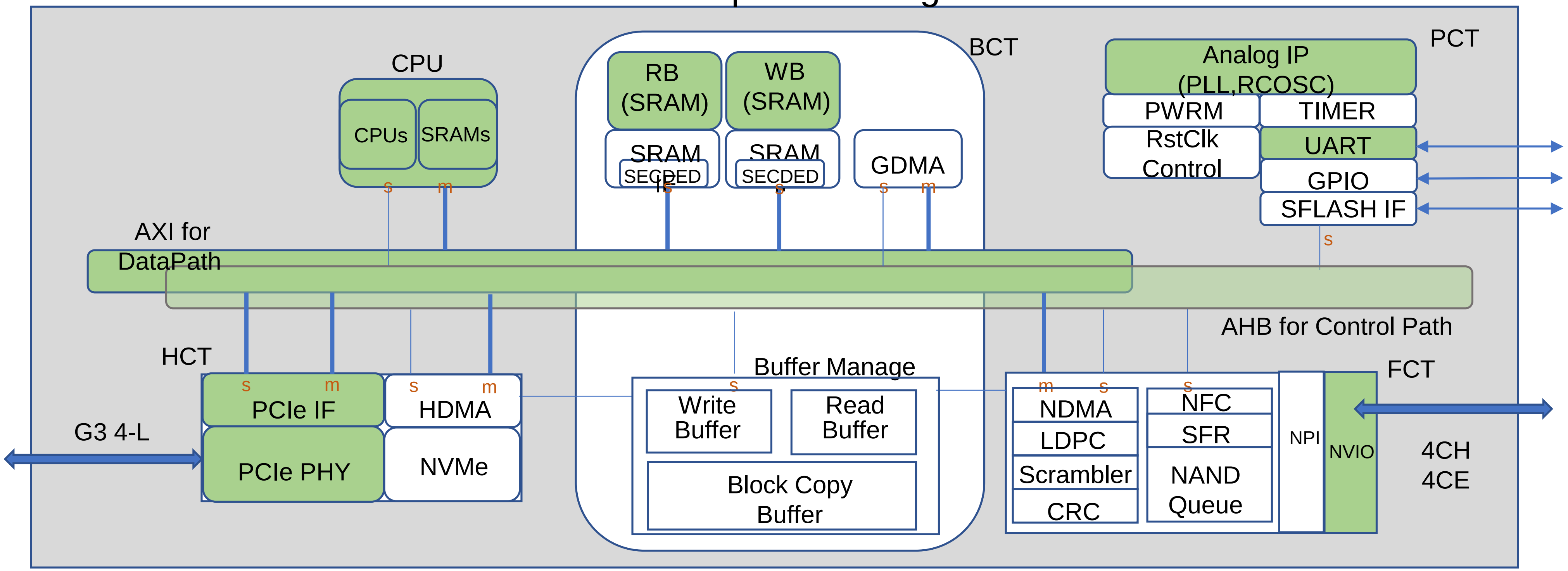
- Open-Channel SSD (OCSSD)
  - Standard NVMe based protocol
  - Facilitates host FTLs and good fit for cloud providers
- Optimized protocol translation between host and NAND interface
- Performance acceleration and reliability enhancement features for 3D NAND TLC/QLC
- Cost and power efficient ASIC design

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

SoC Chip Block Diagram



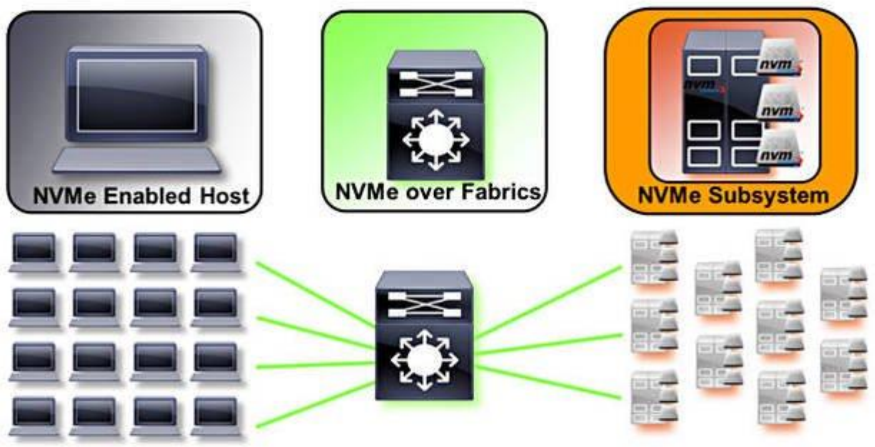
System Performance : G3 2-L

🟩 3rd Party IPs

CIRCUITBLVD
Change your memories. Change your life.

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

Latest Open Source SW Solution

Linux LightNVM
SPDK/DPDK
RocksDB
OpenStack
Ceph
Docker
Kubernetes

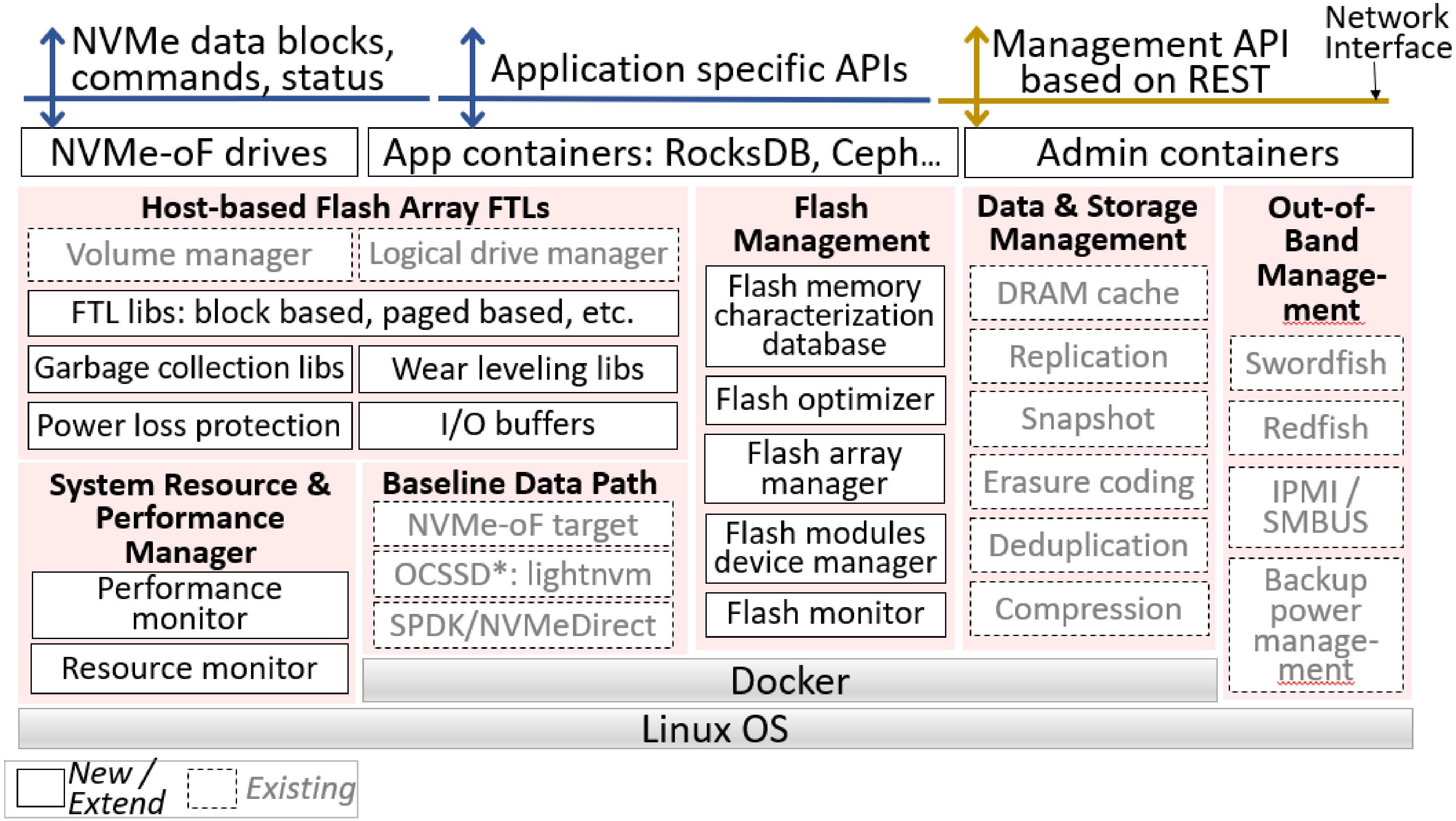
SPDK

- 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



NVMe data blocks, commands, status

Application specific APIs

Management API based on REST

Network Interface

NVMe-oF drives

App containers: RocksDB, Ceph…

Admin containers

Host-based Flash Array FTLs

Volume manager

Logical drive manager

FTL libs: block based, paged based, etc.

Garbage collection libs

Wear leveling libs

Power loss protection

I/O buffers

Flash Management

Flash memory characterization database

Flash optimizer

Flash array manager

Flash modules device manager

Flash monitor

Data & Storage Management

DRAM cache

Replication

Snapshot

Erasure coding

Deduplication

Compression

Out-of-Band Manage-ment

Swordfish

Redfish

IPMI / SMBUS

Backup power manage-ment

System Resource & Performance Manager

Performance monitor

Resource monitor

Baseline Data Path

NVMe-oF target

OCSSD*: lightnvm

SPDK/NVMeDirect

Docker

Linux OS

□ *New / Extend*

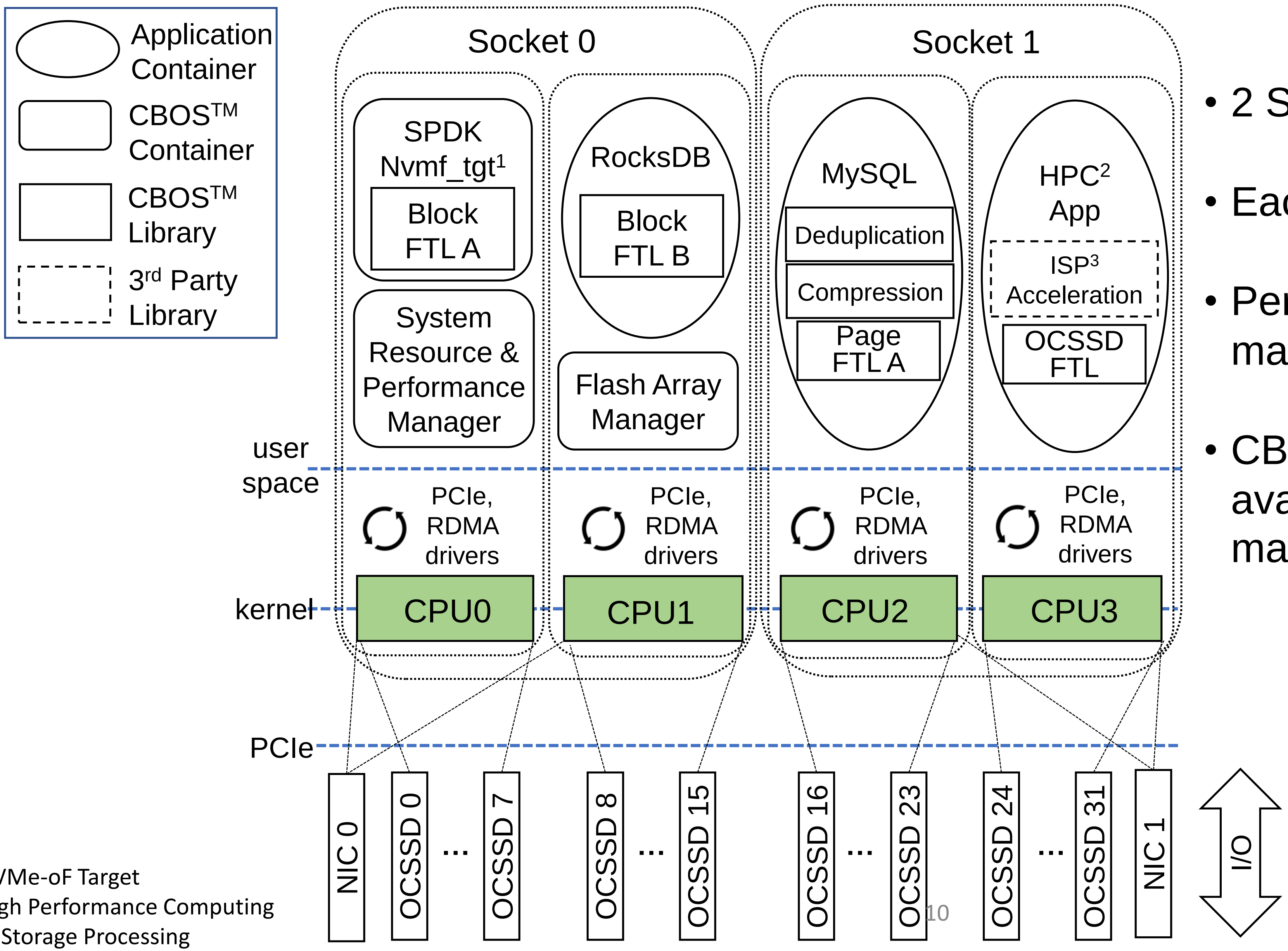
⬚ *Existing*

CIRCUITBLVD

Change your memories. Change your life.

* OCSSD: Open-Channel SSD

CBOS™ Container Example

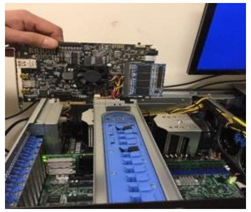
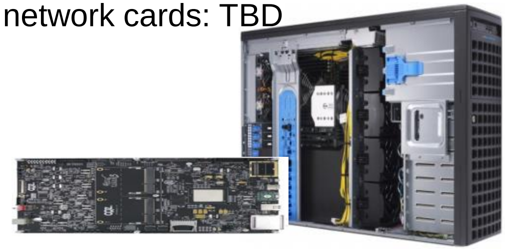
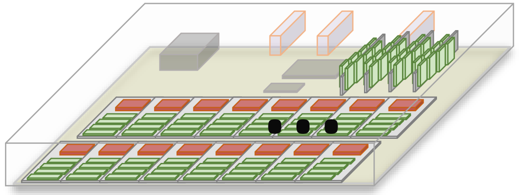


- 2 Socket, 4 Core System
- Each CPU core handles 8 OCSSD
- Per different application needs, matching FTL container is deployed
- CBOS storage containers run on available CPU cores executing managerial tasks

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

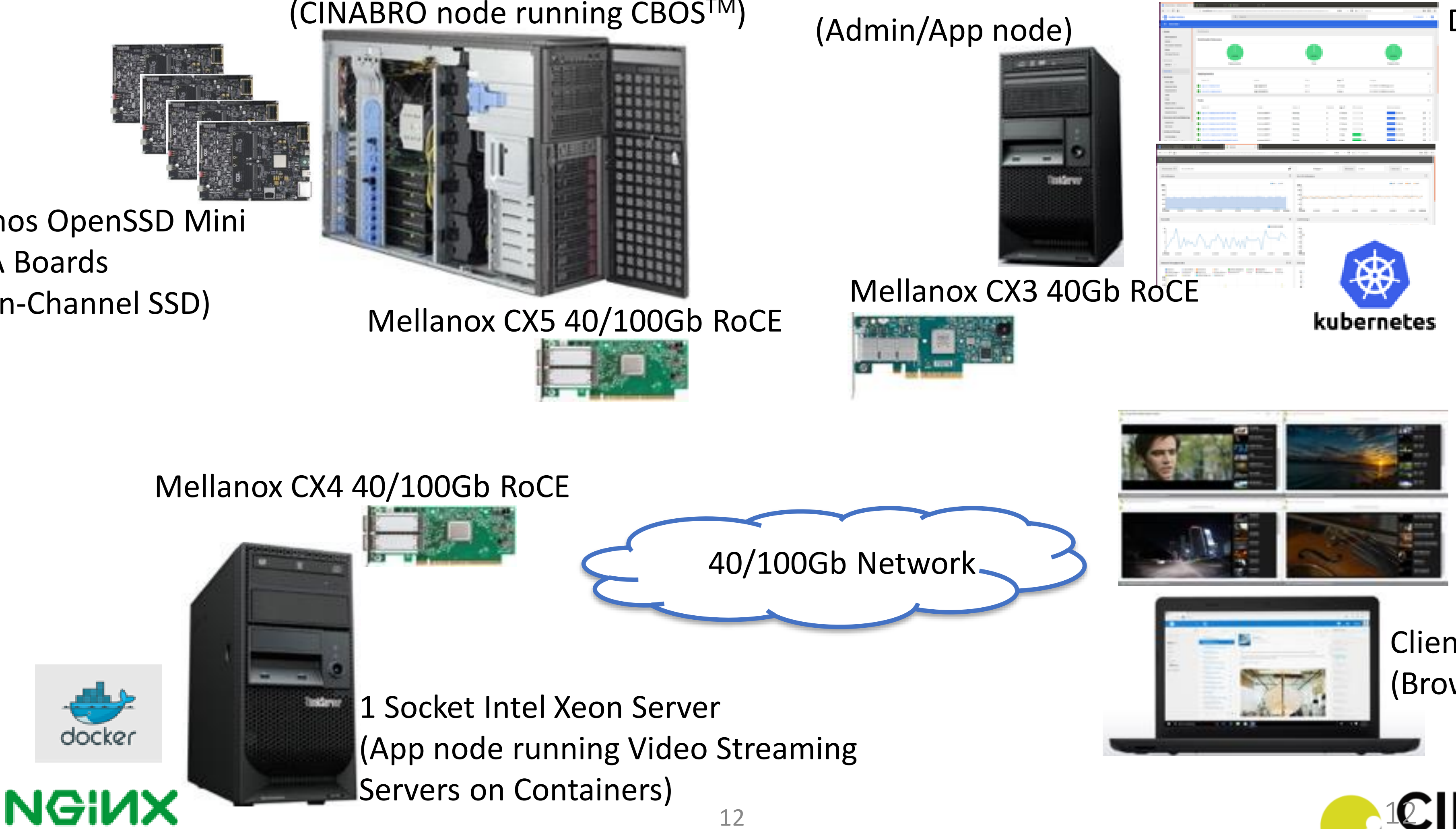
CIRCUITBLVD
Change your memories. Change your life.

Development Milestones

	Phase I (-Dec. '17)	Phase II (-Dec. '18)	Phase III (-Dec. '19)
CINABRO™ Appliance	**First Working Prototype** • Commodity x86 server • x4 Cosmos OpenSSD Mini PCIe cards • Mellanox RDMA network cards 	**Alpha** • Commodity x86 server • x8 Cosmos OpenSSD Ultra PCIe cards • NVMe-oF network cards: TBD 	**Beta** • Customized PCIe fabrics • Commodity CPU, DRAM, PCIe • BMC: ready; Backup power: TBD 
CBBridge™	**FPGA1** • FPGA code with OCSSD compliant FW • 16nm 2D MLC NAND w/ BCH	**FPGA2** • RTL release: CBBridge™ SoC ready. Runs on FPGA • 64L 3D TLC NAND w/ LDPC	**ASIC** • SoC tape-out (mid '19) • 96L 3D TLC/QLC NAND w/ LDPC
CBOS™	**Prototype release** • NVMe-oF drive interface • Baseline data path working: • OCSSD pblk / lightnvm, NVMe-oF, SPDK/DPDK	**Alpha release** • Host-based flash array FTLs • Storage management layer • Application plugins • System resource & performance manager design complete	**Beta release** • Core feature complete • OpenStack compliant • Data management beta • System resource & performance Manager beta • Out-of-Band management beta
Open Source	**SPDK contribution** • Functions to help writing OCSSD access from user level • Included in SPDK v17.10, v18.01	**R&D version alpha** • OpenSSD FPGA RTL codes v1.2 • Developer edition CBOS™ alpha: includes device drivers, user level libraries, and pilot apps	**R&D version beta** • OpenSSD FPGA RTL codes v1.3 • Developer edition CBOS™ beta

Current Prototype

Multiple FPGA based OCSSDs running in our lab



2 Socket Intel Xeon Server
(CINABRO node running CBOS™)

1 Socket Intel Xeon Server
(Admin/App node)

Dashboards

Cosmos OpenSSD Mini
FPGA Boards
(Open-Channel SSD)

Mellanox CX5 40/100Gb RoCE

Mellanox CX3 40Gb RoCE

Mellanox CX4 40/100Gb RoCE

40/100Gb Network

Client PC
(Browsers)

1 Socket Intel Xeon Server
(App node running Video Streaming
Servers on Containers)

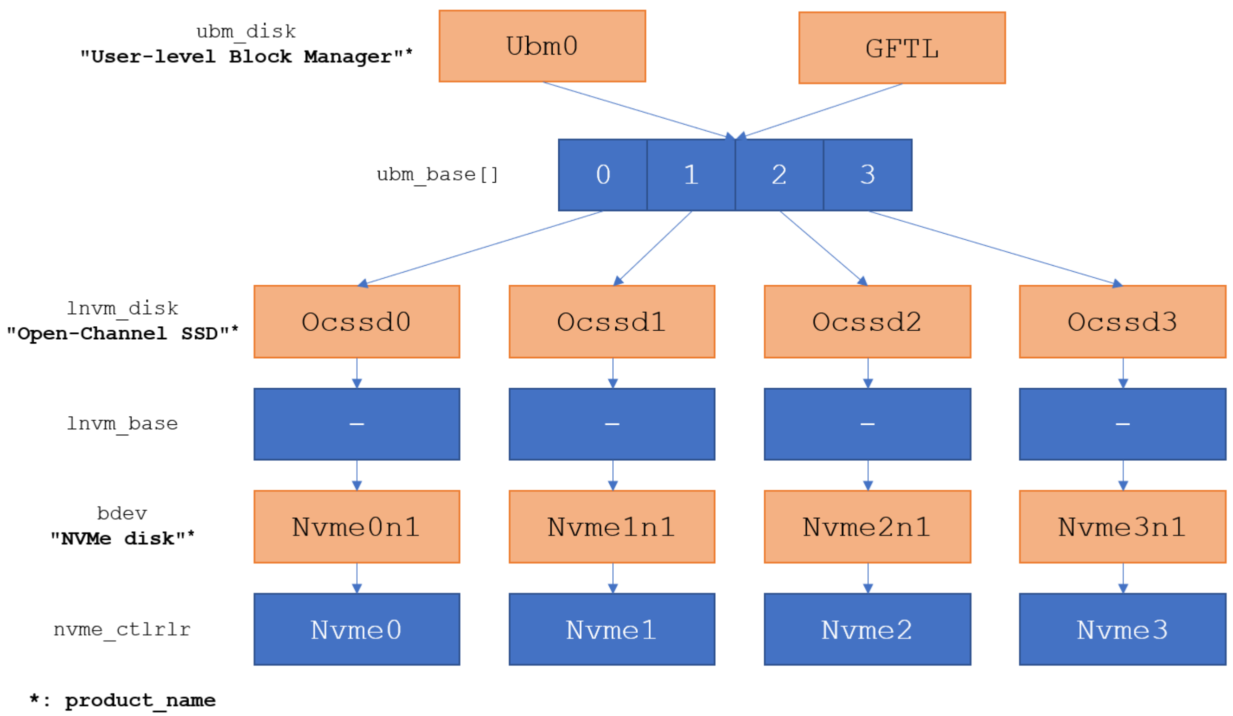
Prototype FTL Evaluations

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

Access

Local (SPDK API)

Remote (NVMe-oF)



ubm_disk
"User-level Block Manager"*
Ubm0
GFTL
ubm_base[]
0
1
2
3
lnvm_disk
"Open-Channel SSD"*
Ocssd0
Ocssd1
Ocssd2
Ocssd3
lnvm_base
–
–
–
–
bdev
"NVMe disk"*
Nvme0n1
Nvme1n1
Nvme2n1
Nvme3n1
nvme_ctlrlr
Nvme0
Nvme1
Nvme2
Nvme3
*: product_name

(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

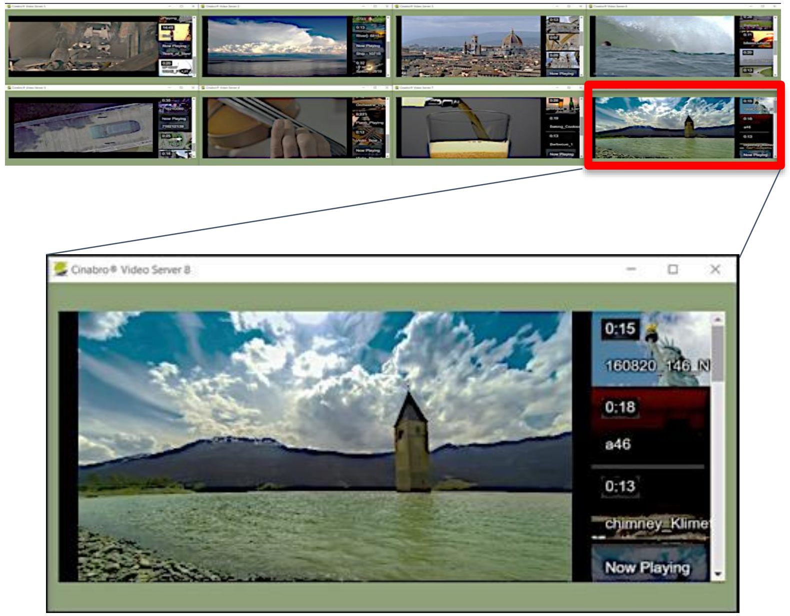
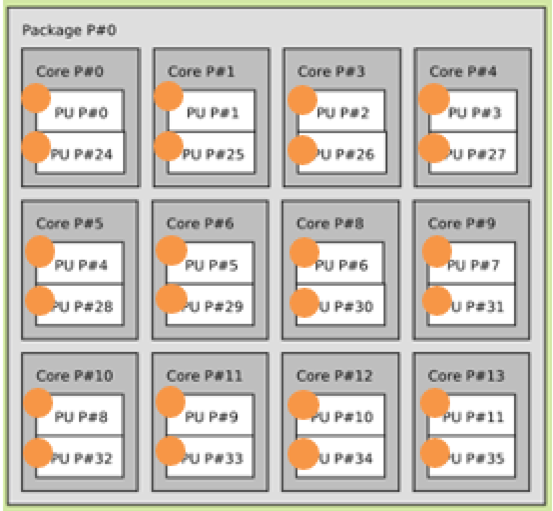
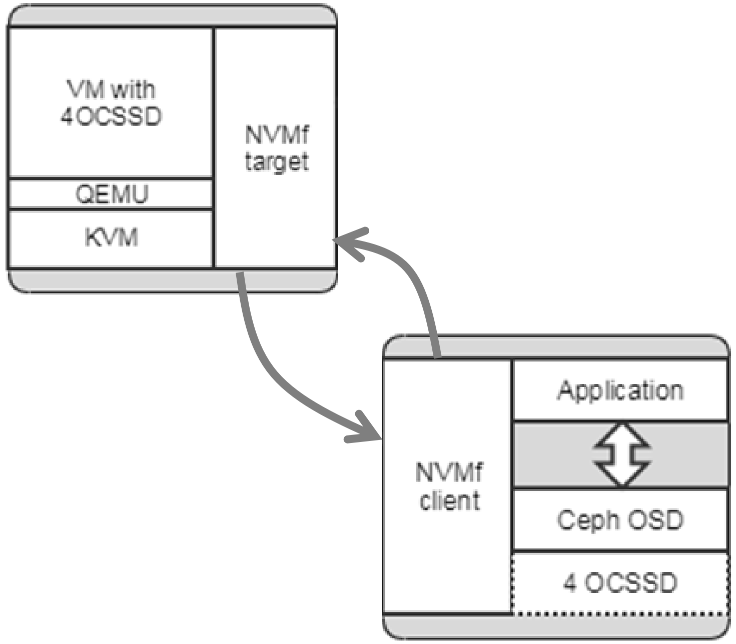
: used for qualitative test of multiple applications

: 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

Containerized Video Server	Containerized RocksDB	SK telecom's AF Ceph
- OCSSD arrays host movie files and are exposed via SPDK nvmf_tgt containers - Video servers run inside containers, made of Nginx web server with RTMP module	- Containerized RocksDB, both local and remote (over NVMe-oF) - Local: SPDK's RocksDB plugin over CBOS^{TM} GFTL - Remote: SPDK's nvmf-tgt over CBOS^{TM} GFTL	- All Flash Ceph is flash optimized version of Ceph. - Initial data verification test over 4 FPGA OCSSDs ran successfully.
 Cinabro® Video Server 8 0:15 160820_146_N 0:18 a46 0:13 chimney_Klime Now Playing [8 concurrent video server example]	SPDK: STORAGE PERFORMANCE DEVELOPMENT KIT RocksDB  Package P#0 Core P#0 PU P#0 PU P#24 Core P#1 PU P#1 PU P#25 Core P#3 PU P#2 PU P#26 Core P#4 PU P#3 PU P#27 Core P#5 PU P#4 PU P#28 Core P#6 PU P#5 PU P#29 Core P#8 PU P#6 PU P#30 Core P#9 PU P#7 PU P#31 Core P#10 PU P#8 PU P#32 Core P#11 PU P#9 PU P#33 Core P#12 PU P#10 PU P#34 Core P#13 PU P#11 PU P#35 [24 RocksDB plugin per CPU core example]	 VM with 4OCSSD QEMU KVM NVMf target NVMf client Application Ceph OSD 4 OCSSD [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)
- CBOS^{TM} development edition codes: TBD

OCP
SUMMIT