Emerging Data-centric Storage Architectures

Pankaj Mehra, Ph.D.
VP Storage Pathfinding
Samsung Semiconductor Inc.
Challenges of Data@Scale

Bottlenecks
- Processing power and processing bandwidth
- Metadata inefficiency of object storage & retrieval
- Wire protocol termination for disaggregated flash

Inefficiencies
- Inability to deliver both performance and scale
- Wasted endurance
- Wasted memory BW
- CPU overhead of I/O
- CPU overhead of I/O virtualization
Good Ideas, Already In-Play

- Virtualization offload
- SMRDB (since HDD days)
- DB filtering acceleration
- Storage NWconv (since FC)
- Active Disk (since HDD days)
- OSD (since HDD days)

Why Revisit?

Because in 2020, three distinct 25-y.o. ideas meet the SSD!
SmartSSD® CSD Scales to Accelerate Data-Rich Workloads

**Computational Storage**
- **3 & 6 GBps internal BW per device:** Minimize external data movement
- **FPGA:** Each device has 3x~10x core equivalents for offload/acceleration
- **4TB storage, 4 GB FPGA DRAM:** For Inline and Data@Rest processing

**Scalable Performance**
- **Near Data Processing:** Data format conversion, Filtering, Metadata management, DB Analytics, Video processing
- **New Services:** Secure content, Edge acceleration

---

SmartSSD U.2 Platform

**Acceleration Concept**

**Partner Solutions**

**SmartSSD**

- SSD Controller
- FPGA

**4TB V-NAND**

- FPGA DRAM

---

SparkSQL with Parquet Data

- Transfer Full Data
- Filter and Decompress
- SQL Process

H.264 Video Transcoding

- P2P Compression and Decompression
Computational Storage Use Cases Examples

- 3rd party and proprietary acceleration stacks run on Computational Storage to accelerate real-time analytics and regex searches for cybersecurity.

### Analytics Cache Node

- **New:** Accelerated Cache Nodes
- **SmartSSD**
- **Computational Storage Processor (CSP)**
- **Computational Storage Drive (CSD)**

### RegEx Appliance

- **Throughput scales to large datasets and complex searches**
- **>10x throughput improvement compared to x86**

---

**RE2 vs SmartSSD Throughput**

(As function of expression complexity, set size, match density, near-match density, etc.)

---

2020 Storage Developer Conference. © Samsung Electronics. All Rights Reserved.
Samsung SmartSSD® Technology Roadmap

- Samples, development tools, partners solutions available for immediate PoC
- Customer PoC Test&Dev systems/support available from Samsung and partners

**v1.0 SmartSSD® U.2 CSD**

- U.2 FF: Scale Processing to 24 ~ 48 devices
- 4TB, PCIe Gen3x4 External, ~530K LUTs,

**Next Gen SmartSSD® CSD**

- Customers requirements: Integration, Interfaces, FF, workloads

**Best Practices**

SNIA API and NVMe Protocol for Computational Storage

**IH’20**

- Partner PoCs
- U.2 ES
- Partner Solution

- Customer PoCs

**2nd Generation**
Ethernet SSD targets IO bottleneck in Storage Chasses

2019
3 X 100GbE ➔ ~36 GB/s
24x NVMe SSDs ➔ ~48 GB/s

Future
3 X 200GbE ➔ ~75 GB/s
200GbE
200GbE
200GbE

CPU

CPU

CPU

CPU

CPU

NVMe SSD

NVMe SSD

24x NVMe SSDs ➔ ~150+ GB/s

CPU and IO bottleneck for storage throughput performance
NVMe-oF SSD based EBOF

**Pros**
- Enables disaggregation of NVMe SSDs
- Management & Storage Services
- Utilizing existing storage & server architectures

**Cons**
- Non-scalable Storage Controller - PCIe single root constraint
- Bandwidth Limitation
  - CPU, PCIe, NW Constraints
- Power and Thermals

**Pros**
- High Bandwidth
- Scaled Linearly (Ethernet)
- Sharable via NVMe-oF
- Less power
- Lower latency

**Cons**
- New platform architecture
- Management of Storage Services & Network Devices

NVMe-oF EBOF can address bandwidth, scalability, and flexibility
Samsung Ethernet SSD Technology Roadmap

- Samples, development tools, partners solutions available for immediate PoC
- Customer PoC Test&Dev systems/support available from Samsung and partners

Samsung U.2 eSSD POC

- U.2, 2x25 GbE, RDMA/RoCeV2, iWARP NVMe-oF
- 2020

Next Gen Ethernet SSD

- Customers requirements: NVMe-oF/TCP, HW Offloads
- FF, Congestion Control (ECN/PFC), Security, Management

PoCs

- U.2 ES POC

2nd Generation
KV SSD is about Efficient Block Operations

Block operations on CPUs ⇒ Bottlenecks, Scaling Inefficiency.
KV SSD offloads Block operations from CPUs.
KV-optimized SW shows Multiple Efficiency Benefits

**Average Latency (us)**

- RocksDB
- KVRocks

- FillSeq: 1.854, 0.638
- ReadRandom: 1.033, 1.014
- ReadSeqUni: 0.446, 0.492
- UpdateBath: 38.697, 28.78
- FillRandom: 43.203, 42.214
- ReadRand: 30.264, 4.472

Average Latency is up to 60% better on KVRocks

**Insertion operations/sec**

- KVRocks: 400, 419194.76
- RocksDB: 110, 149194.76

SSD lifetime is more than 11x for KVRocks compared to RocksDB ⇒ 11x less CapEx
Samsung KV SSD Technology Roadmap

- Stack open sourced at https://github.com/OpenMPDK/KVSSD

U.2 KV SSD
- U.2 FF: Scale Processing to 24 devices
- 4TB, PCIe Gen3x4 External

Next Gen KV SSD
- Your requirements? (Integration, Interfaces, FF, Workloads)

IH'20

Software dev
- Rocks DB repl.
- KVCeph
- Minio

Partner Solution

Partner Testing

2nd Generation
## Benefits Summary

<table>
<thead>
<tr>
<th></th>
<th>SmartSSD</th>
<th>Ethernet SSD</th>
<th>Key-Value SSD</th>
<th>Zoned Name Spaces</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application Awareness</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Acceleration</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduce data-related CPU load</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Improved Write Endurance</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Fewer protocol terminations</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Min device virtualization o/h</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Fewer stack translations</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Metadata Optimization</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Scaling Data Bandwidth</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saving L2-to-Memory BW</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Control@Scale (IODT, QoS)</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Maximize #SSDs/chassis</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

2020 Storage Developer Conference. © Samsung Electronics. All Rights Reserved.
## Possible Convergence

<table>
<thead>
<tr>
<th>Host Interface</th>
<th>Addressing</th>
<th>Accelerator</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCIe</td>
<td>Block</td>
<td>None</td>
</tr>
<tr>
<td>Ethernet</td>
<td>ZNS</td>
<td>FPGA</td>
</tr>
<tr>
<td></td>
<td>Key-Value</td>
<td></td>
</tr>
</tbody>
</table>

### Diagram

- **Object**
  - e-KV
  - KV Smart
  - eSmart
- **Block**
  - OLTP
  - OLAP
  - KV SSD
  - SmartSSD
  - eSSD

**Key-Value**
- Beyond block CPU util 10PB+
- Near-data proc Performance Scalability (100TB+)
- Disagg. Block TCO IOPS

**OLTP**
- Blob
- Media
- Dense VMs

**OLAP**
- Serverless
- Data Lake

---

2020 Storage Developer Conference. © Samsung Electronics. All Rights Reserved.
Please take a moment to rate this session.

Your feedback matters to us.
- Bullet one
  - Bullet two
    - Bullet 3
    - Bullet 4
Dr. Pankaj Mehra

Vice President,
Storage Pathfinding

Samsung Semiconductor Inc.
- Subhead
  - Example 1
  - Example 2
- Subhead