Zoned Flash SSDs in Advanced Storage Systems

Robert Lercari
Radian Memory Systems, Inc.
This presentation and/or accompanying oral statements by Radian Memory Systems, Inc. ("RMS") representatives collectively, the “Presentation”, is intended to provide information concerning Flash memory technology. While RMS strives to provide information that is accurate and up-to-date, this Presentation may nonetheless contain inaccuracies or omissions. As a consequence, RMS does not in any way guarantee the accuracy or completeness of the information provided in this Presentation.

This Presentation may include forward-looking statements, including, but not limited to, statements about any matter that is not a historical fact; statements regarding RMS’ intentions, beliefs or current expectations concerning, among other things, market prospects, technological developments, growth, strategies, and the industry in which RMS operates; and statements regarding products or features that are still in development. By their nature, forward-looking statements involve risks and uncertainties, because they relate to events and depend on circumstances that may or may not occur in the future. RMS cautions you that forward looking statements are not guarantees of future performance and that the actual developments of RMS, the market, or industry in which RMS operates may differ materially from those made or suggested by the forward-looking statements in this Presentation.

In addition, even if such forward-looking statements are shown to be accurate, those developments may not be indicative of developments in future periods.
Agenda

• Cooperative Flash Management
• ASL Configurator
• Idealized Flash
• Delegated Copy-Move
• Back Channel
• Performance Benchmarks
Cooperative Flash Management (CFM)

Redistribution of Flash Management between Host/Device

Host System Software

- Data Placement
- Leverages host segment cleaning for Garbage Collection
- Scheduling

Symphonic Host Libraries and API

- ASL Configurator

Symphonic Firmware

- Garbage Collection
- Space Mgt., Command Interface
- Address Configurator
- Wear Leveling
- Configuration Mgt.
- Geometry Emulation
- Error Handling
- State Mgt.
- Flush-to-Flash

- Wear Leveling
- NAND Maintenance
- Maintains device state
- Idealized Flash
- Configurable Addressing
- Offload process execution

Standard NVMe API plus vendor specific extensions
Cooperative Flash Management (CFM)

2014
Symphonic™ v1
Cooperative Flash Management
Demo’d CFM to leading Flash Fabs

2015
Symphonic™ v2
Cooperative Flash Management
Won FMS
CFM and SMR

2017
SMErr Drive
SMR-F SSD
Zoned Flash

2018
Implementation for OC2

2018
Symphonic™ v3
Cooperative Flash Management
‘All Firmware’ implementation

2019
Zoned Flash
Zoned Flash

Zoned Flash ver. 3.0; RMS-350

- Idealized Flash
- ASL Configurator
- Decoupled Wear Leveling and NAND Maintenance
- Back Channel*
- Delegated Copy-Move offload*
- Zone Append*
- Relaxed Write Pointer

*Optional feature
Smerf Bridge *(optional)*

- Zone Block Device (ZBD) to NVMe protocol translation bridge
- Provides support for Zone Report and Zone Reset commands between:
  - Hosts using the ZBD interface
  - Zoned devices using NVMe vendor extensions for Zone Report and Zone Reset
  - Support for multi-drive volumes
Address Space Layout (ASL) Configurator

Iso-Region
Dies form discrete, physically isolated regions

Iso-Box
One or more iso-regions that can be associated with a namespace

- Performance
- Endurance
- Capacity
Address Space Layout (ASL) Configurator

- Zones and Application Segments: Write Amp
- Write Stripes: Bandwidth/Latency

NAND Erase Units (blocks) from dies from within the same iso-region

NAND pages from within Erase Units (blocks) within a zone.

- Zone Report command
Address Space Layout (ASL) Configurator

Configurable Parameters
Idealized Flash

- Geometry Emulation
  - Presents idealized NAND
  - Hierarchal Address Virtualization
- Abstracts vendor-specific attributes
- Maps Bad Blocks
- Not a FTL
  - No 0.1% mapping storage requirements
  - Deterministic

Contiguous Addressing

Zone 1 | Zone 2 | Zone 3 | Zone N
---|---|---|---
bad | bad | bad | bad

Zoned Namespaces SSD Firmware
- Space Mgt. Command Interface
- Decoupled Wear Leveling
- Error Handling
- SSD Maintenance
- ASL Configurator
- State Management
- Flash-to-Flash

Geometry Emulation
Strict Write Pointers

- NAND requires sequential programming
- Tangled Ordering
- Performance Impact

Tangled Ordering Write Errors

<table>
<thead>
<tr>
<th>QD/Thread</th>
<th>% Write Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>98.87%</td>
</tr>
</tbody>
</table>

- Without Idealized Flash (Geometry Emulation turned off)
- Closed system
Zone Append

**Pros**
- No Strict Write Pointer requirement
- Overcomes NAND addressing anomalies, geometry or vendor specific attributes
- No FTL L2P storage requirements, .1%, 1GB mapping space for 1TB capacity

**Cons**
- Modifications to host system software
- New consistency models
- Potential latency impact

---

**Radian’s Zone Append can support multiple, concurrent append request/completions**

1. **Host sends data and specifies zone**
2. **SSD determines LBA in designated zone and provides it to host**
3. **Host updates mapping table**

---

2019 Storage Developer Conference. © Radian Memory Systems, Inc. All Rights Reserved.
Relaxed Write Pointer

- Idealized Flash enables Relaxed Write Pointer
- Overcomes Tangled Ordering if host attempts to write sequentially
- No modifications to host software, no new consistency models or additional latency
- Minimal SSD memory (not 0.1% like L2P tables)
Variable Capacity & Zone Excursions

- No addressing gaps
- Zones can change capacity due to bad blocks
- Host software requires modifications

Idealized Flash

- No addressing gaps
- Automatic bad block replacement
- Zones do not change size/capacity
- No modifications to host software
Idealized Flash

- Idealized Flash overcomes the limitations or need for...
  - Strict Write Pointers
  - Zone Append
  - Variable Zone Capacities and Excursions
- Simplifies integration into existing storage systems
- Minimizes host software modifications when transitioning to different NAND
Idealized Flash

- Idealized Flash overcomes the limitations or need for...
  - Strict Write Pointers
  - Zone Bound
  - Variable Zone Capacities and Excursions

- Simplifies integration into existing storage systems
- Minimizes host software modifications when transitioning to different NAND
Host Storage Stack

Garbage Collection

1. Host determines what zone to clean
2. Copies valid data from that zone
3. Writes valid data to new zone
4. Issues ‘Zone Reset’ command to device for original zone

ZNS SSD

Zone 1

Zone 2

Zone 1

Device erases original zone

Zone 2

Zone 2
Delegated Copy-Move Commands

- NVMe vendor specific extension
  - NV-RAM → Flash
  - Flash → NV-RAM
  - Flash → Flash

User NV-RAM

- Multiple use cases and advantages
- Can appear as distinct...
  - Zone(s)
  - Namespaces and/or unique block device
Host Storage Stack

1. Host determines what zone to clean

2. Host commands device to copy valid data and move it to new zone

3. Device copies and moves specified data to specified zone

4. Issues ‘Zone Reset’ command to device for original zone

5. Device erases original zone
Host Storage Stack

Garbage Collection

1. Host determines what zone to clean
2. Copies valid data from that zone
3. Writes valid data to new zone
4. Issues ‘Zone Reset’ command to device for original zone

Host sequentially writes new data to specific LBA and zone

ZNS SSD

2019 Storage Developer Conference. © Radian Memory Systems, Inc. All Rights Reserved.
Delegated Copy-Move

100% of Host Writes are new Host Writes

1.5 Host WA

Final Result: 67% of Media Writes are new Host Writes

1.5 Total System WA
(Host WA performed internally in SSD)

67% of Host Writes are new Host Writes

1.5 Host WA

Final Result: 34% of Media Writes are new Host Writes

3.0 Total System WA
(1.5 x 2.0 performed internally in SSD)

Assumes 1.5 Host WA factor
Assumes 2.0 FTL SSD internal WA factor

= Copied Data
Disaggregated Storage

DCM avoids host/device copying over the wire
‘Back Channel’ (optional)
Wear Leveling & NAND Maint.

- ‘Back Channel’ – optional out-of-band communication path
- Supports various cooperative host/device management activities
- Decoupled Wear leveling and NAND Maintenance managed transparently by device
- Routine management handled transparently in coherently aligned manner
- Additional wear and maintenance signaled by device through Back Channel
- Scheduled by host to avoid unpredictable spikes
- Ultimately enforced by device to maintain warranty
‘Back Channel’ *(optional)*

Wear Leveling & NAND Maint.

Lifecycle QoS

Device initiated requests

Escalating Severity

**Host Storage Stack**

- **NAND Maintenance**
  - Host includes data in GC relocation process, device performs relocation as instructed
  - Device requests data at specified locations to be moved

**ZNS SSD**

- Requests Data Movement
- Performs NAND Maintenance
Block Translation Layer

Enables testing random, overwriting workload

- Log Structured design serializes random writes
- Performs segment cleaning (garbage collection) with Zone Reset
- Can enable Conventional Zones

SDC ‘19 Demo
Apples to Apples Comparison

Radian Zoned SSD

Identical Silicon

- Same SSD Processor
- Same Flash Array
  - 3D TLC NAND Dies/Package
  - # of Channels
  - # of Packages/Channel
  - 4.6TB Raw capacity
- Same DDR4 array
  - DDR4
  - # of Devices

FTL SSD

Zoned Flash
U.2 NVMe SSD
RMS-350

FTL
U.2 NVMe SSD
Measuring at the system level

- Log Structured design serializes random writes
- Performs segment cleaning (garbage collection) with Zone Reset
- Can enable Conventional Zones
- Enables testing random, overwriting workload
- Creates Log-on-Log

Emulates log structured storage management layer:
- All-Flash Storage Array
- Hyperconverged appliance
- SDS frameworks
- KVS and other storage engines
### Overprovisioning (OP)

#### 30% Example

<table>
<thead>
<tr>
<th></th>
<th>FTL SSD</th>
<th>Zoned SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advertised User Capacity</td>
<td>3.23TB</td>
<td>3.23TB</td>
</tr>
<tr>
<td>Total OP</td>
<td>30%</td>
<td>30%</td>
</tr>
<tr>
<td>LS Host Free Space</td>
<td>13%</td>
<td>27%</td>
</tr>
<tr>
<td>Advertised Device Capacity</td>
<td>3.84TB</td>
<td>4.49TB</td>
</tr>
<tr>
<td>SSD Internal OP</td>
<td>17%</td>
<td>3%</td>
</tr>
<tr>
<td>Raw Capacity</td>
<td>4.62TB</td>
<td>4.62TB</td>
</tr>
</tbody>
</table>

- **Advertised User Capacity**
- **Total Overprovisioning**
- **LS Host Free Space**
- **Advertised Device Capacity**
- **SSD Internal Overprovisioning**
- **Raw Device Capacity**
• 70/30 Mix
• 4K Random Read
• 4K Random Write
• SSD Queue Depth = 32
  4 worker threads
  IOD = 8/thread
• Total Overprovisioning = 30%
• Single Namespace

99.99% Latency Over Time

99.99% Latency @ IOPS

Latency (ms)

Latency Time (s)

IOPS (K)

FTL SSD

Radian Zoned SSD

>75ms Delta @ 250K IOPS
- 70/30 Mix
- 4K Random Read
- 4K Random Write

- SSD Queue Depth = 32

99.99% Latency @ IOPS
Single Namespace, 25% OP and 30% OP

Latency (ms)

FTL SSD

Radian Zoned SSD

100
80
60
40
20
0

IOPS (K)

25% OP

30% OP

70ms Delta @ 250K IOPS

99.99% Latency @ IOPS
Sixteen Namespaces, 30% OP

Latency (ms)

FTL SSD

Radian Zoned SSD

100
50
10
5

IOPS (K)

>185ms Delta @ 240K IOPS

Log-on-Log + Noisy Neighbors