Raising QLC Reliability in All-Flash Arrays

Jeff Yang
Principal Engineer
Storage Research Dept.
Silicon Motion, Inc.
QLC Characteristics

- QLC Endurance: 1~3K P/E. (limited DWDP)
- Enterprise TLC: 10K P/E.
- Average tPro per page is 3~4X to TLC.
- QLC Cost reduction ~25~30%.
- Ungraceful shutdown failure range on QLC is a conflict to the atomic write.
- SLC block from TLC or QLC flash are almost the same.

QLC is good enough for the SMB (Small and Medium-sized Business) AFA applications.
Vertical integration for storage technology

Now: iSCSI, future: NVMe-of
1. Volume control/QOS/snapshot/Clone/thin-Pro.
3. High ability, healthy monitoring.

SMI SSD controller
key advantage for AFA. (unique technology)
1. Lowest cost SSD for highest reliability
2. Lowest cost SSD for shortest latency
3. Cross boundary optimization
4. Extend QLC reliability.

SMI controller Key advantage.
Flexible ECC engine and RAID design to meet NAND’s characteristics. AES/cypher IP.
NAND analysis and tuning, understand all kind of NAND flash characteristics.
Flash Array

**Storage work**

**Buffer management**

**Caching algorithm**

**Flash Array (FA)**

**Group of disk**

<table>
<thead>
<tr>
<th>SSD0</th>
<th>SSD1</th>
<th>SSD2</th>
<th>SSD3</th>
<th>SSD4</th>
</tr>
</thead>
</table>

**SLBA:** System Logical block address

**DLBA:** Disk Logical block address

**FPPA:** Flash physical page address

**RD**

**WR**

**XOR**

**U**

**P**

**Disk 0**

**Disk 1**

**Disk 2**

**Disk 3**

**Disk 4**

**U0**

**U1**

**U2**

**U3**

**P0**

**U4**

**U5**

**U6**

**U7**

**P1**

**P1_TMP**

**RD U4, P1**

**WR U4', P1'**

**XOR P1 U4 → P1_TMP**

**XOR P1_TMP U4' → P1'**
N SSDs with parity P, PQ, and PQR

- N + P: 1Write $\rightarrow$ 2Read + 2Write
  $\rightarrow$ Single parity: WAI = 2
- N + PQ: 1Write $\rightarrow$ 3 Read + 3Write
  $\rightarrow$ Double parity: WAI = 3
- N + PQR: 1Write $\rightarrow$ 4 Read + 4Write
  $\rightarrow$ Triple parity: WAI = 4

- The traditional method is not suitable for SSDs, because of the high WAI factor and consume the SSD’s endurance faster.
Lower WAI RAID method on SSD

- Map the SLBA to the DLBA.
- Generate two parity on the same DLBA cross different SSDs to provide the protection.
- Flash array software layer maintain a lookup table.
- When writing the existed data into the Flash array:
  - Existed RAID link will not be changed.
  - Write the data to the new location with new RAID link.
Lower WAI RAID method on SSD

- Write the data to the new location with new RAID link.
- Invalidate the old location and update the L2P table.
- The GC work will be applied.
- The WAI will be related to the Overprevision (OP) ratio.

![Diagram showing L2P table and SSD mapping]

OP ratio to WAI

![Graph showing OP ratio against WAI]

Assume: FIO with norandommap
The SSD awareness RAID will construct a remapping table in host side.

For the single SSD point of view, the access behavior become the sequential write. (DLBA is a sequential write.)

When issuing a read CMD on SLBA, host will redirect to DLBA by using the mapping table.
Multi-thread operation scenario

- Read/write disk using single thread will cause context switch problems.
- Single disk will need to handle the requests from different threads.
Access mixing from thread to SSD channel

- All the threads issue the access to the same SSD.
- It becomes the random access behavior.
- Huge DRAM for SSD device mapping table and another OP-ratio are required.
- Powerful SSD device CPU for GC work.
- Larger Capacitance for ungraceful shutdown handling.
- Enterprise SSD: OP=20%, WAI = ~3.
- Much more expensive enterprise SSD.
A flexible controller to solve problem. Service oriented SSD

- Flexible switching provides several different types of NAND groups for applications.
- The service oriented RAID is configurable for different types of NAND groups and different types of NAND failure behavior.
- Each Channel becomes sequential program and erase.
- Use SLC as caching buffer on ungraceful shutdown handling.
Every thread gets its own NAND Flash.

- Dedicated NANDs for dedicated threads respectively.
- The simple mapping removes DRAM requirement in SSD device.
- Remove GC work from device SSD.
- Data will write into SLC first, remove Capacitance in ungraceful shutdown handling flow.
- Cost efficient SSD.
- Sequential Write/Erase behavior is the perfect match for QLC.
## Comparison on N + PQR (triple RAID-parity)

<table>
<thead>
<tr>
<th></th>
<th>Traditional Flash array RAID</th>
<th>SSD awareness RAID</th>
<th>Vertical integration RAID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host Flash array (SLBA to DLBA)</td>
<td>1W = 4R + 4W. OP = ~0% WAI = 4</td>
<td>OP = 20% WAI = 3</td>
<td>OP = 20% WAI = 3</td>
</tr>
<tr>
<td>SSD device (DLBA to FPPA)</td>
<td>Enterprise SSD. OP = 20%. WAI = 3</td>
<td>Enterprise SSD. OP = 20%. WAI = 3</td>
<td>Service oriented SSD. OP = ~0%. WAI = ~1</td>
</tr>
<tr>
<td>Overall (SLBA to FPPA)</td>
<td>WAI = 12 OP = 20% (additional read latency)</td>
<td>WAI = 9 OP = ~36%</td>
<td>WAI = 3. OP = 20%</td>
</tr>
</tbody>
</table>

- Reduce the overall WAI and Over Provisioning will increase the life time of QLC
Conclusion

• SSD controller is a key to connect the NAND to applications.
• Both reliability and efficiency will be improved by controller.