Raising QLC Reliability in All-Flash Arrays

Jeff Yang

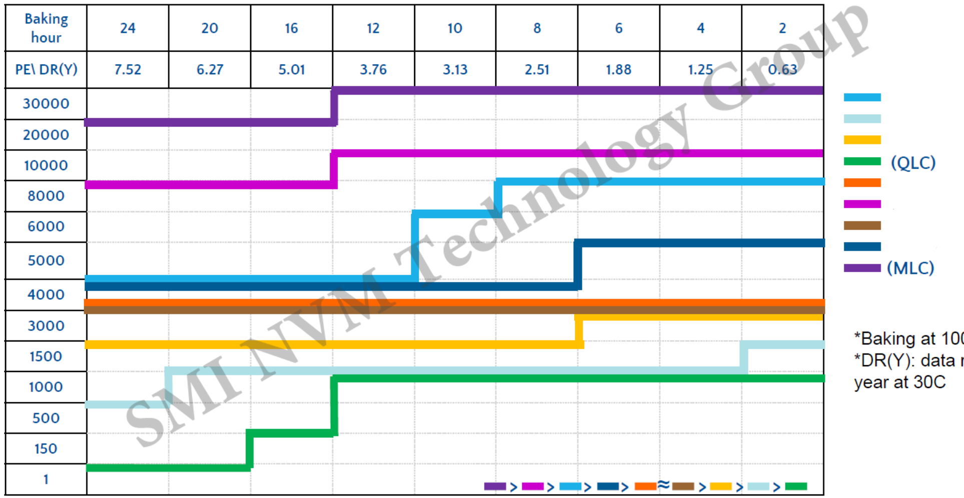
Principal Engineer

Storage Research Dept.

Silicon Motion, Inc.

QLC Characteristics

Endurance Rating (Estimation)



*Baking at 100C
*DR(Y): data retention year at 30C

- 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

- NVMe-OF/networking
- Failover technology
- Storage work
- Buffer management Caching algorithm
- Flash Array(FA) Group of disk
- NVME-driver/OCSSD
- NVME-PCIE/SOC architecture
- ECC and error recovery
- NAND study NVM-TG

Now:ISCSI, future: NVMe-of

1. Volume control/QOS/snapshot/Clone/thin-Pro.
2. Data-reduction: compression/deduplication.
3. High ability, healthy monitoring.

SMI SSD controller
key advantage for AFA. (unique technology)

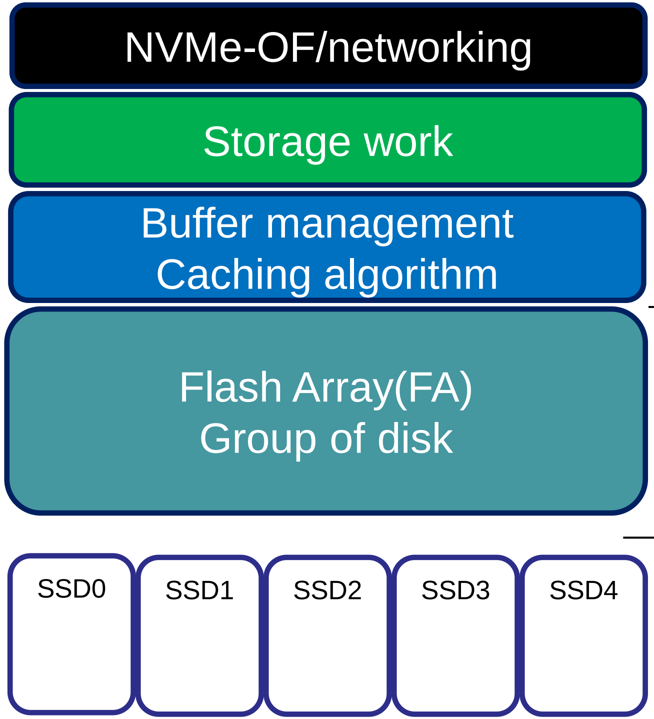
1. Lowest cost SSD for highest reliability
2. Lowest cost SSD for shortest latency
3. Cross boundary optimization
4. Extend QLC reliability.

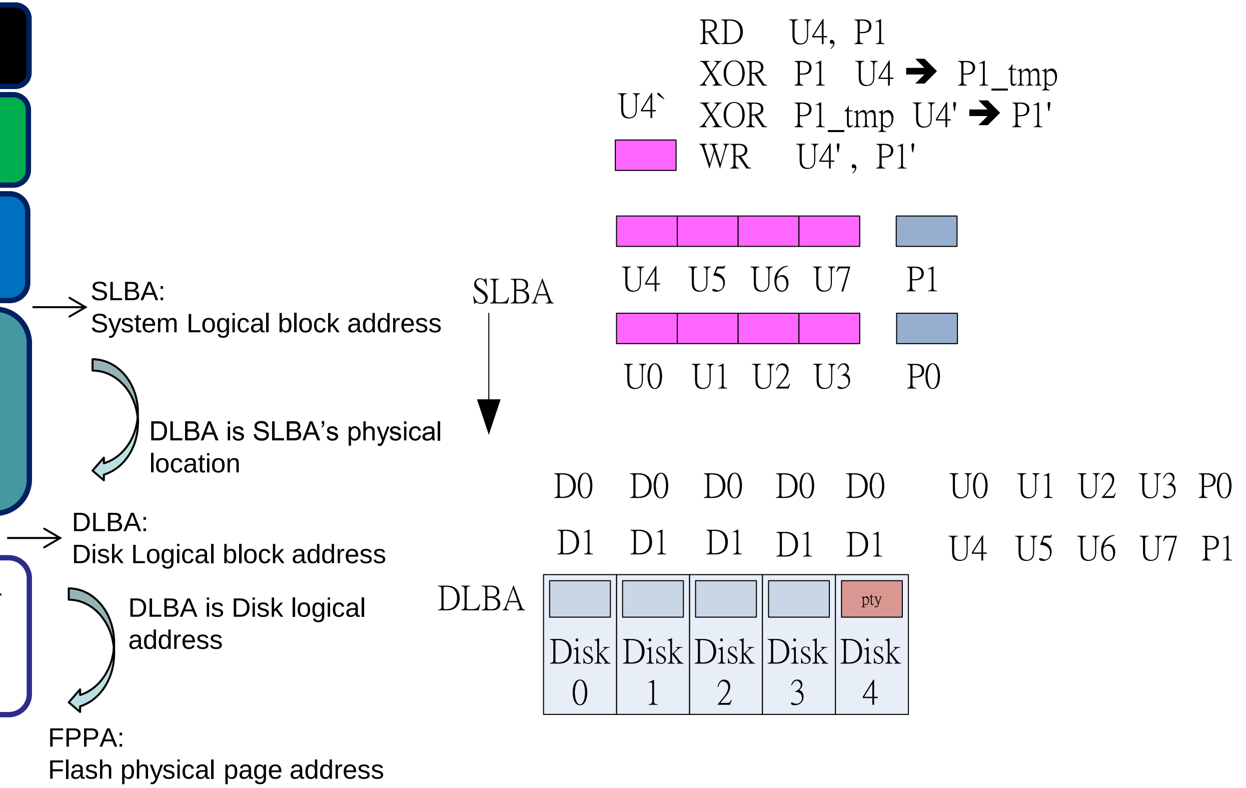
SMI controller Key advantage.

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


NVMe-OF/networking
Storage work
Buffer management
Caching algorithm
Flash Array(FA)
Group of disk
SSD0
SSD1
SSD2
SSD3
SSD4


SLBA:
System Logical block address
DLBA is SLBA’s physical location
DLBA:
Disk Logical block address
DLBA is Disk logical address
FPPA:
Flash physical page address
U4`
RD U4, P1
XOR P1 U4 ➔ P1_tmp
XOR P1_tmp U4' ➔ P1'
WR U4' , P1'
SLBA
U4 U5 U6 U7 P1
U0 U1 U2 U3 P0
D0 D0 D0 D0 D0
D1 D1 D1 D1 D1
U0 U1 U2 U3 P0
U4 U5 U6 U7 P1
DLBA
pty
Disk 0
Disk 1
Disk 2
Disk 3
Disk 4

N SSDs with parity P, PQ, and PQR

- N + P: 1Write → 2Read + 2Write

  → Single parity: WAI = 2
- N + PQ: 1Write → 3 Read + 3Write

  → Double parity: WAI = 3
- N + PQR: 1Write → 4 Read + 4Write

  → 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.

L2P table

U1	D_{00}
U2	D_{10}
U3	D_{20}
U4	D_{01}
U5	D_{11}
U6	D_{21}
U7	D_{02}
U8	D_{12}
U9	D_{22}

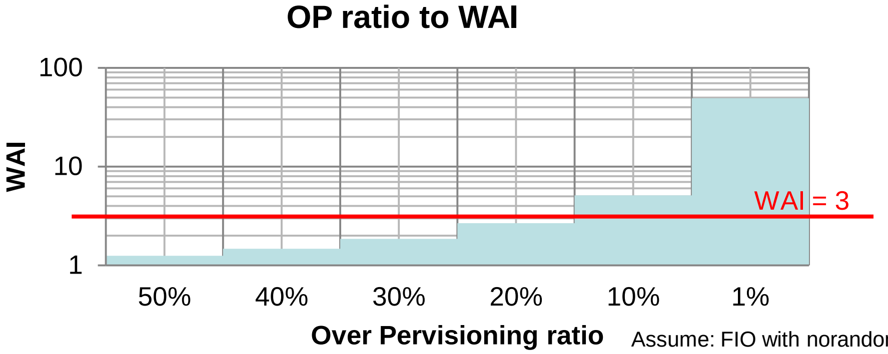
SSD0	SSD1	SSD2	SSD3	SSD4
U1 D_{00}	U2 D_{10}	U3 D_{20}	P0 D_{30}	Q0 D_{40}
U4 D_{01}	U5 D_{11}	U6 D_{21}	P1 D_{31}	Q1 D_{41}
U7 D_{02}	U8 D_{12}	U9 D_{22}	P2 D_{32}	Q2 D_{42}
D_{03}	D_{13}	D_{23}	D_{33}	D_{43}
D_{04}	D_{14}	D_{24}	D_{34}	D_{44}
D_{05}	D_{15}	D_{25}	D_{35}	D_{45}

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.

L2P table

U1	D_{00}
U2	$D_{10} \rightarrow D_{05}$
U3	$D_{20} \rightarrow D_{03}$
U4	$D_{01} \rightarrow D_{04}$
U5	$D_{11} \rightarrow D_{13}$
U6	D_{21}
U7	D_{02}
U8	$D_{12} \rightarrow D_{14}$
U9	$D_{22} \rightarrow D_{23}$
U10	D_{24}



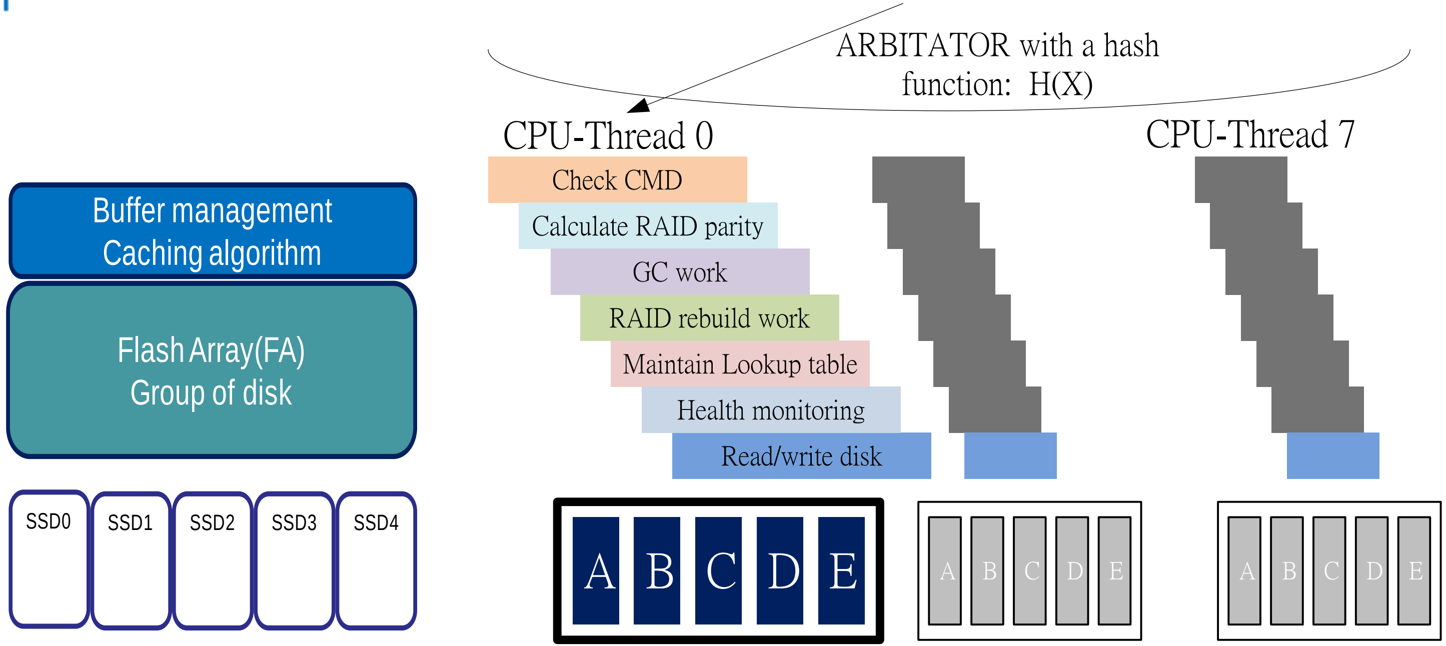
Assume: FIO with norandommap

Single SSD Read/Write behavior

- 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.

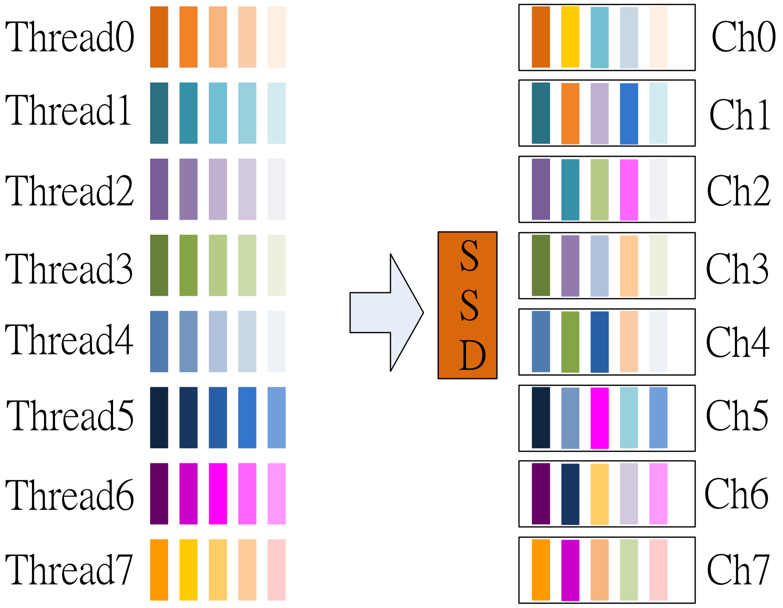
SSD0	SSD1	SSD2	SSD3	SSD4
D_{00}	D_{10}	D_{20}	D_{30}	D_{40}
D_{01}	D_{11}	D_{21}	D_{31}	D_{41}
D_{02}	D_{12}	D_{22}	D_{32}	D_{42}
D_{03}	D_{13}	D_{23}	D_{33}	D_{43}
D_{04}	D_{14}	D_{24}	D_{34}	D_{44}
D_{05}	D_{15}	D_{25}	D_{35}	D_{45}
D_{06}	D_{16}	D_{26}	D_{36}	D_{46}
D_{07}	D_{17}	D_{27}	D_{37}	D_{47}
D_{08}	D_{18}	D_{28}	D_{38}	D_{48}

Multi-thread operation scenario



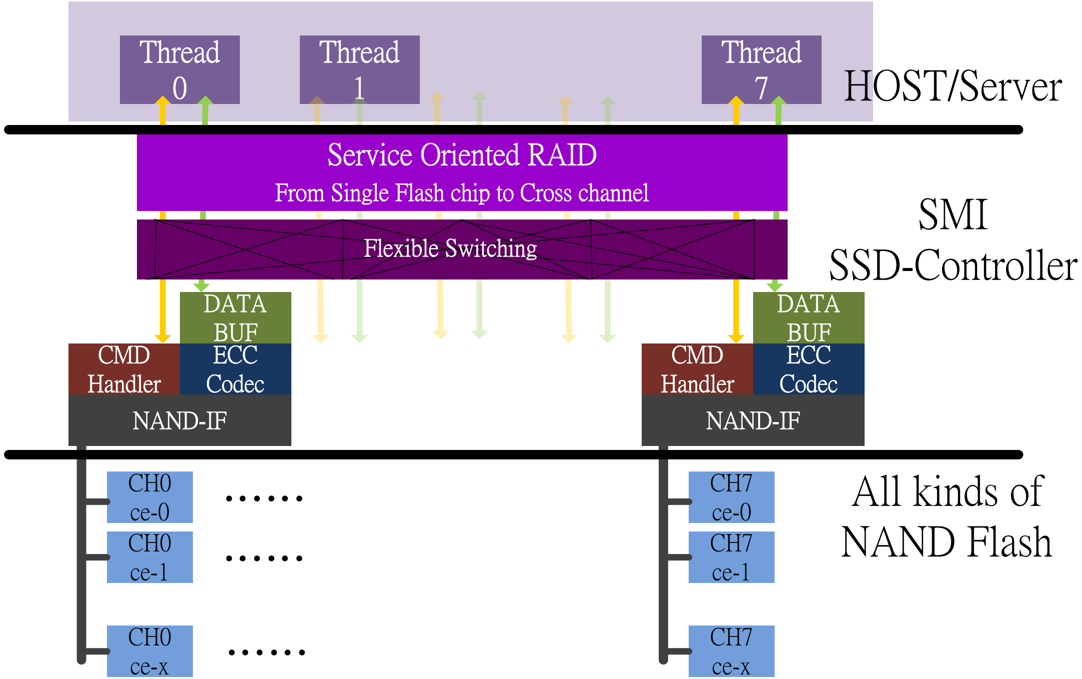
- Read/write disk using single thread will cause context switch problems.
- Single disk will need to handle the requests from different threads.

Access mixing from thread to SSD channel



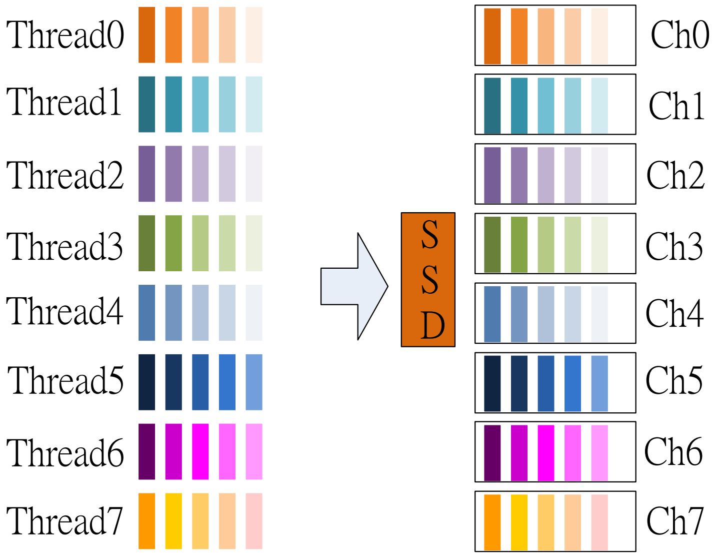
- All the threads issue the access to the same SSD.
- It becomes the random access behavior.
- Huge DRAM for SSD device mapping table and another OP-ratio are required.
- Powerful SSD device CPU for GC work.
- Larger Capacitance for ungraceful shutdown handling.
- Enterprise SSD: OP=20%, WAI = ~3.
-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

Flash Memory Summit

Comparison on N + PQR (triple RAID-parity)

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

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

Flash Memory Summit

Conclusion

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