An Advanced Error Recovery Scheme for Open-Channel SSDs.

Jeff Yang
Storage research team

Silicon Motion
Storage request and environment limitation

- CPU: 112~128 threads.
- PCIE port: 96~128 lanes
- Ethernet: 16~32 lanes
- Limited Lanes for PCIE-SSD.
- Single application occupies single thread and requires 600G~1TB data access.
- 30~60iops per GB.
- Access latency is very important.
- Data integrity is very important.
- Host want to controller everything on data storage
- Single NAND chip will have 128GB.
- If 1TB (8 NAND chip) consume 4-lanes to provide best latency…..
The Benefit of OCSSD

- Macro control
  - Open channel SSD allows the host to perform data placement and I/O scheduling, defined storage interface (e.g., block device or object store) to higher level applications.

- Cut through complexity
  - By managing data placement, the host is able to aggregate data with a similar life time within the same chunk thus achieving lower write amplification.

- Multi-pronged
  - Expose their internal parallelism
  - The I/O isolation between the parallel units provides achieving predictable latency,
  - Logical I/O traffic isolation is provided by partitioning the physical media of the Open channel SSD.

- Fast harvest
  - A new idea can be implemented by the software engineer directly. Cheaper but faster.

EXCEPT

Host want to controller everything on data storage…….

- Error correction code
- NAND bad block management
- Low level wear leveling
- Error recovery latency from........ Read disturbance, data-retention....
- Ungraceful shut down.
- RAID for NAND failure
- And Many.. Many.. Dirty Jobs.
IOPS vs. Latency vs. Reliability vs. Cost

The BEST SSD design imagination

128 NAND-chips SSD: 16TB. RW: 1M-iops, RR: ~2M-iops, Latency: 70usec

- Is it possible to keep the latency? ➔ NO WAY
- 16TB will request for 0.5~1M iops under gen3x4, but 1M~2M iops for gen4x4.
- The more NAND chips will introduce higher failure dppm.
- It needs RAID protection scheme.
- RAID will introduce other write-amplifier and capacity lost.
RAID overhead and benefit

8 Dies

| Die 0 | Die 1 | ... | parity Die 7 |

1 - (1 - 50 ppm)^8 \approx \binom{8}{1} \times 50 ppm = 400 ppm

1 - (1 - 50 ppm)^8 - \binom{8}{1} \times (1 - 50 ppm)^7 \times 50 ppm \approx 0.07 ppm

RAID capacity overhead

12.5%

RAID WAI overhead

8/7 = 1.14

128 Dies

| Die 0 | Die 1 | Die 2 | ... | Die 125 | Die 126 | Die 127 |

Sequential write the data cross the die, put the parity on one of them.

1 - (1 - 50 ppm)^{128} \approx \binom{128}{1} \times 50 ppm = 6400 ppm

1 - (1 - 50 ppm)^{128} - \binom{128}{1} \times (1 - 50 ppm)^{127} \times 50 ppm \approx 20 ppm

RAID capacity overhead

0.8%

RAID WAI overhead

128/127 = 1.007

In real case, the parity will be put on different die with a rotation scheme.
Under the OCSSD scheme, after the RAID protection:

- Each thread will have the sequential write behavior, but different timing and frequency.
- Because of the RAID protection. Even if the host is sequential write, it still need a mapping table for the host-address to physical address.
- Different application’s data will be mixed and located on every NAND chips.
RAID protection cause worse latency.

Each read physical address will randomly located on a certain die. Assume the uniform distribution

TWO WAYS to reduce the latency.
1. Reduce the read-busy time.
2. Separate the data into several different physical groups.
Separate the different application’s data.

- Dedicated SOC-Buffer (BF), and NVME Queue-Pair (QP) for each NAND group belong to single applications.
- Reduce the latency, but also limited the RAID protection in single group.
- After reducing the RAID group size, it will consume more capacity for RAID overhead.
- Each group still can reserve the benefit of the sequential write only on OCSSD/ZNS.
Group RAID protect region

Widest failure range, single die failure. Largest RAID overhead under few chip number But strongest protection

1die/128dies = 1/128

Single plan failure. Smaller RAID overhead. (planar RAID) 1plan/8diesx4plan Cannot protect the single die failure = 1/32

TWO WL(layer) protection Smallest RAID overhead. Less than 1%

Plan 0 Plan 1 Plan 2 Plan 3

Block 0 Block 1 Block 2 Block 3

b0 b1 b2 b3

CE-0 CE-1 CE-2 CE-3

Block N-1

Ch0 Ch1 Ch15
NAND chip NAND chip NAND chip

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

CE-0 CE-1 CE-2 CE-3

b0 b1 b2 b3

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short

Plan 0 Plan 1 Plan 2 Plan 3

b0 b1 b2 b3

WL0 WL1 WL2 WL3

Damage range of Program fail / Word-line Open

Damage range of Word-line short
Service Oriented data protection

- A configurable data protection engine will provide the flexibility.
- Reserve the good property from host sequential write.
- Keep the shorter latency under the multi-tenancy access scenario.

DIE protection or plan protection or two-WL protection.

Reserve a Channel for data protection. But each will have different endurance. The RAID parity will be updated frequently.

Still provide the option on the limited raid overhead.

Capacity overhead: 1/8 = 12.5%
Write Amplifier: 8/7 = 1.14x

Capacity overhead: 1/128 = 0.8%
Write Amplifier: 8/7 = 1.14x

Keep good latency property, and save the capacity.

Provide the DIE protection.
We can do more…

- The server host software engineer can rely on these technology and focus on higher level applications.
- Silicon Motion help to take care the NAND’s physical issues.
- Advanced Error correcting code.
- Data-retention immunity.
- Auto dynamic SLC/TLC/QLC swapping for performance acceleration.
- Flexible Bad block management.
- Read-disturbance auto detection.
- Combo GC command set.
- Support DRAM and NonDRAM cases operation.
- Beyond-SMART technology for better life perdition.

Contact Silicon Motion to enhance your genius.
Thanks for your attention!
Visit our booth #413 for more information