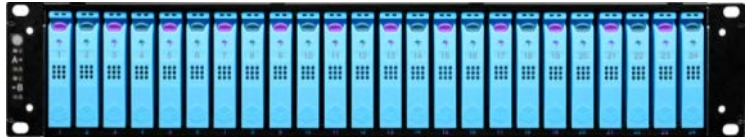
An Advanced Error Recovery Scheme for Open-Channel SSDs.

Jeff Yang

Storage research team

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

Host want to controller everything on data storage……

- 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

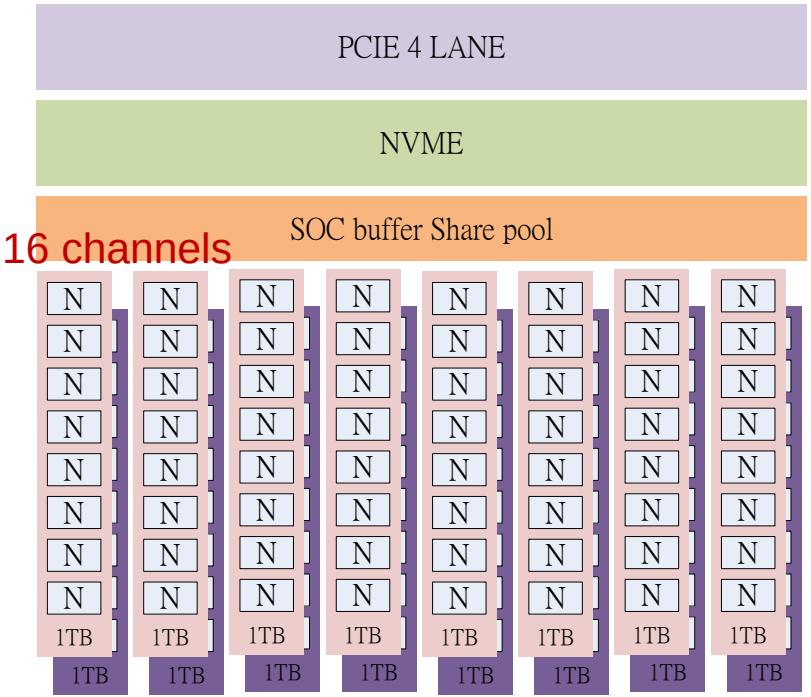
Error correction code **NAND bad block management**

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

Single NAND-chip SSD: 128GB. RW: 8K-iops, RR: 16K-iops,
Latency: 70usec(~tR only) .

X128

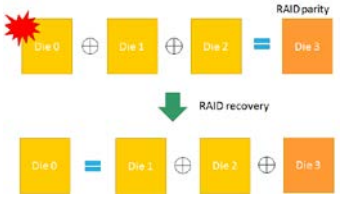
128 NAND-chips SSD: 16TB. RW: 1M-iops, RR: ~2M-iops,
Latency: 70usec

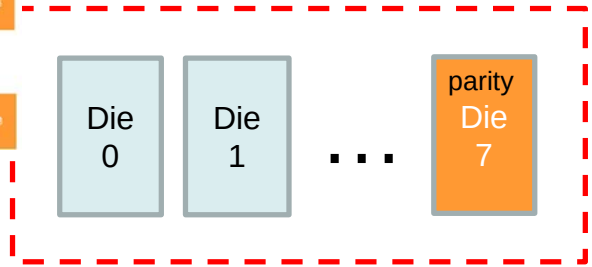
KEEP

- 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

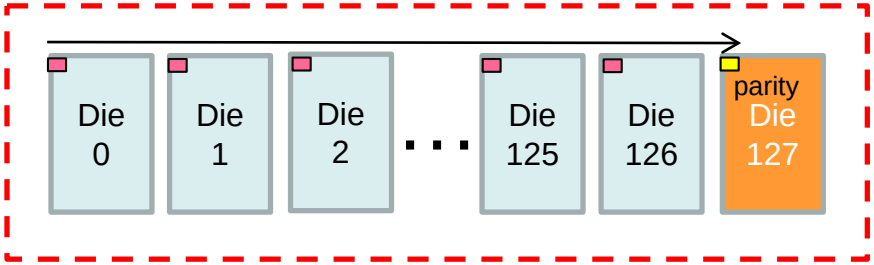
8 Dies





128 Dies

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



Broken Device without RAID

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

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

Broken Device with RAID

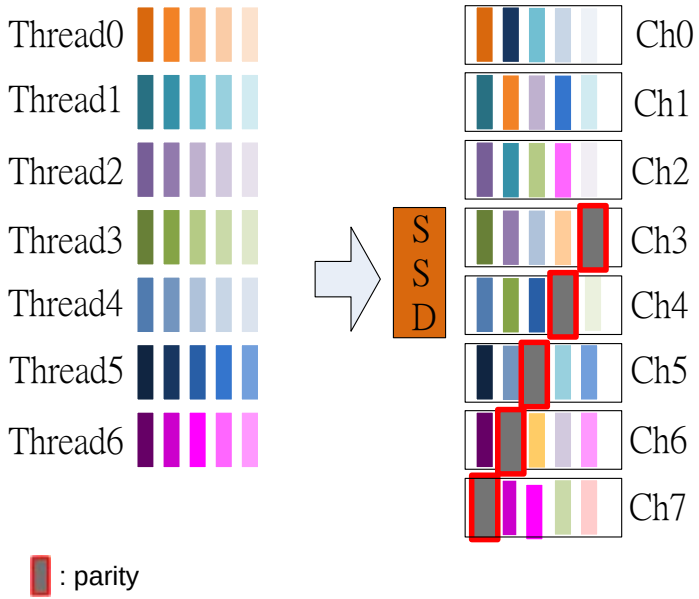
$$1-(1-50ppm)^{8} - \binom{8}{1} \times (1-50ppm)^{7} \times 50\text{ppm} \approx 0.07\text{ppm}$$

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

	8 Dies	128 Dies
RAID capacity overhead	12.5%	0.8%
RAID WAI overhead	8/7 = 1.14	128/127 = 1.007

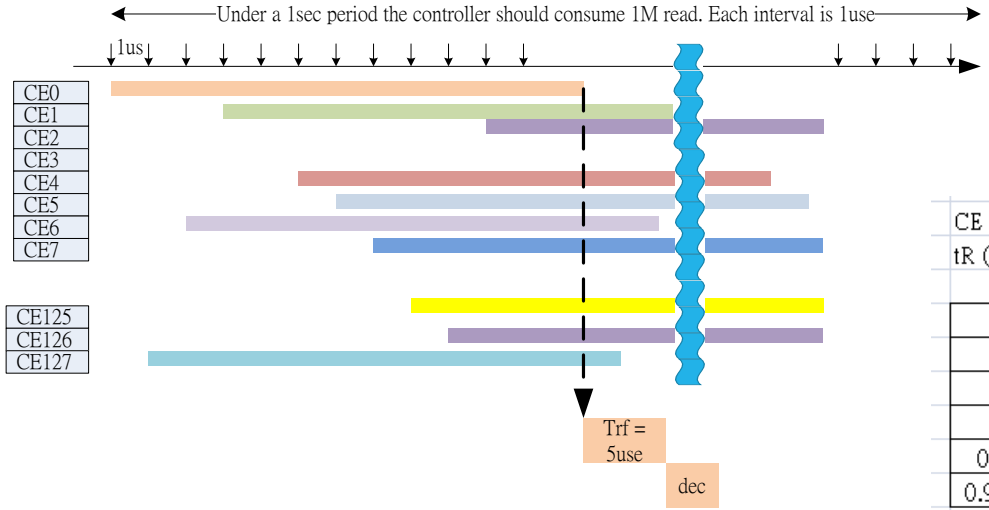
In real case, the parity will be put on different die with a rotation scheme.

Under the OCSSD scheme, after the RAID protection

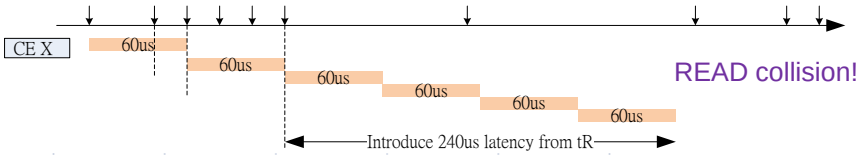


- Each thread will have the sequential write behavior, but different timing and frequency.
- Because of the RAID protection. Even if the host is sequential write, it still need a mapping table for the host-address to physical address.
- Different application's data will be mixed and located on every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

RAID protection cause worse latency.



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



READ collision!

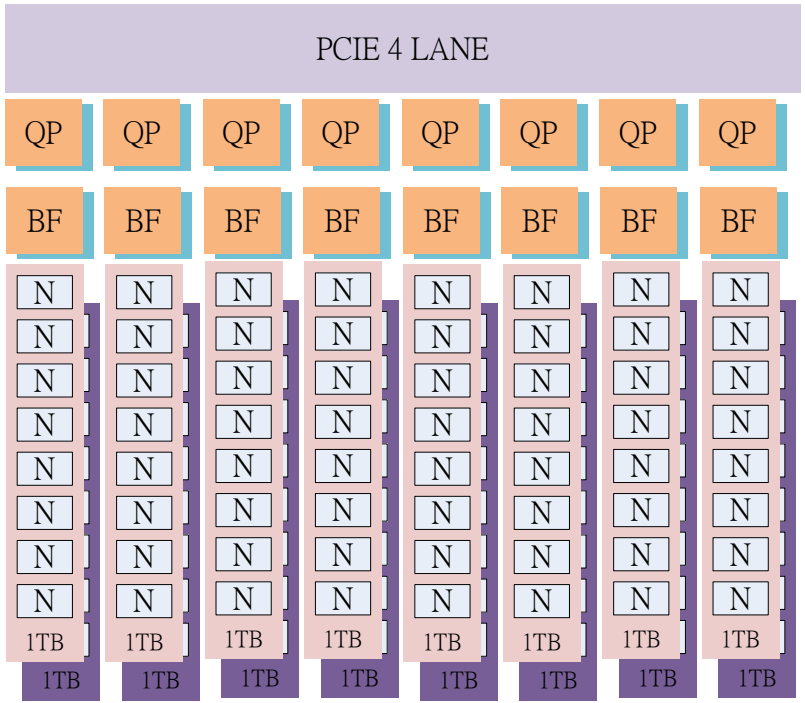
CE	128				
tR (us)	100	90	80	70	60
0.9	647	458	342	261	201
0.99	1133	767	551	406	301
0.999	1691	1119	771	549	397
0.9999	2208	1486	988	688	492
0.99999	2581	1678	1143	834	624
0.999999	2749	1812	1225	965	732

Six 9s latency will be a multiple times of Read-busy time

TWO WAYS to reduce the latency.

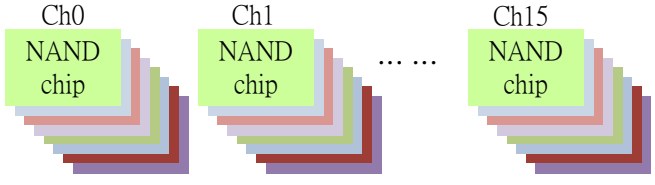
1. Reduce the read-busy time.
2. Separate the data into several different physical groups.

Separate the different application's data.



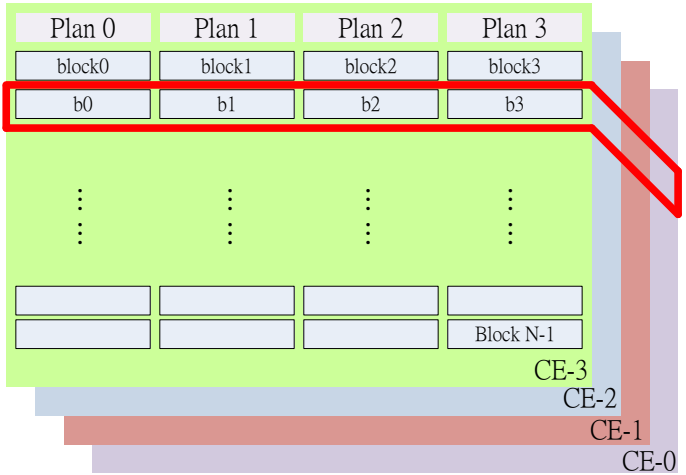
- Dedicated SOC-Buffer(BF), and NVME Queue-Pair(QP) for each NAND group belong to single applications.
- Reduce the latency, but also limited the RAID protection in single group.
- After reducing the RAID group size, it will consume more capacity for RAID overhead.
- Each group still can reserve the 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

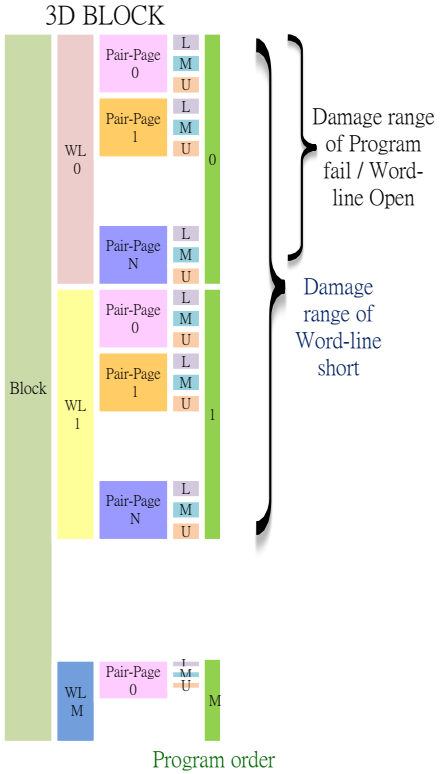


Single plan failure. Smaller RAID overhead. (planar RAID)
Cannot protect the single die failure

1plan/8diesx4plan = 1/32

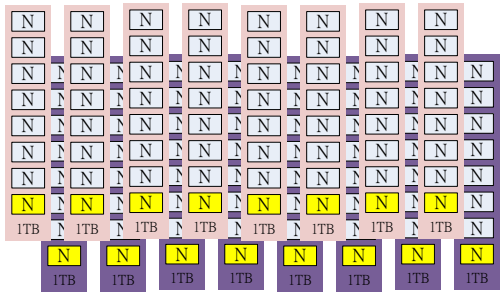
TWO WL(layer) protection
Smallest RAID overhead. Less than 1%

	CH#	0	0	1	1	
	CE#	0	1	0	1	
WL0	pair-page	P0	P0	P0	P0	S0
	pair-page	P1	P1	P1	P1	S1
	pair-page	P2	P2	P2	P2	S2
	pair-page	P3	P3	P3	P3	S3
WL1	pair-page	P4	P4	P4	P4	S4
	pair-page	P5	P5	P5	P5	S5
	pair-page	P6	P6	P6	P6	S6
	pair-page	P7	P7	P7	P7	S7
WL2	pair-page	P8	P8	P8	P8	S8
	pair-page	P9	P9	P9	P9	S9
	pair-page	P10	P10	P10	P10	S10
	pair-page	P11	P11	P11	P11	S11
WL3	pair-page	P12	P12	P12	P12	S12
	pair-page	P13	P13	P13	P13	S13
	pair-page	P14	P14	P14	P14	S14
	pair-page	P15	P15	P15	P15	S15

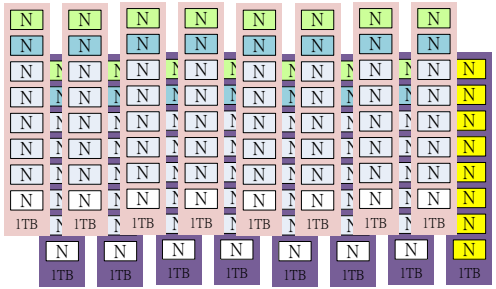


Service Oriented data protection

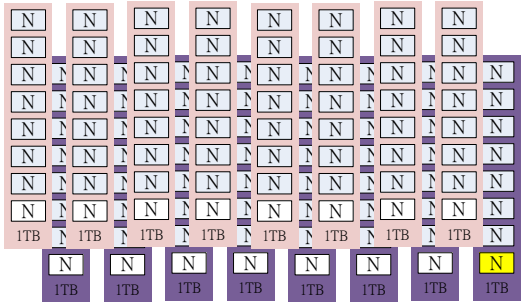
- A configurable data protection engine will provide the flexibility.
- Reserve the good property from host sequential write.
- Keep the shorter latency under the multi-tenancy access scenario.



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.

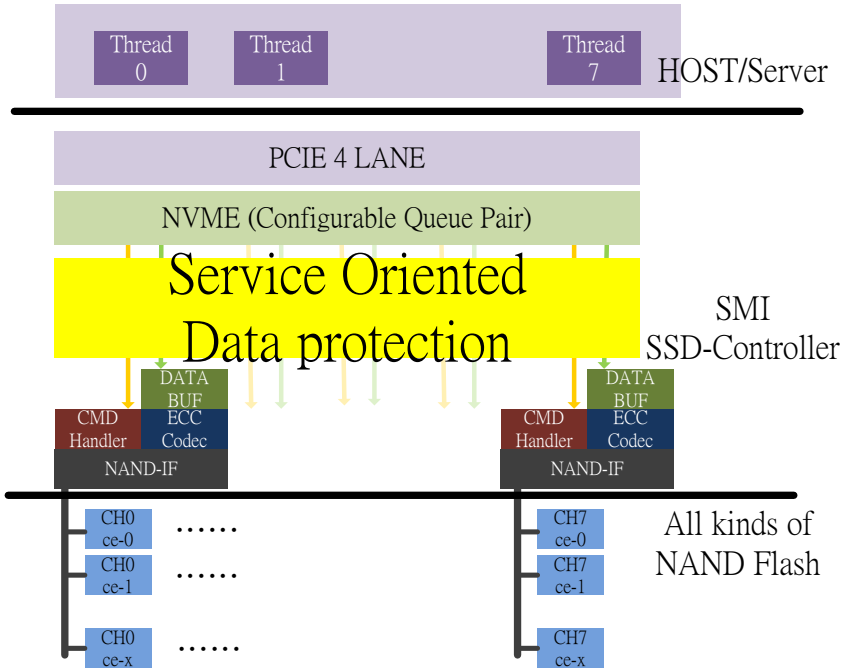
Capacity overhead: 1/8 = 12.5%
Write Amplifier: 8/7 = 1.14x

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.

SiliconMotion

www.siliconmotion.com

Thanks for your attention!

Visit our booth #413 for more information