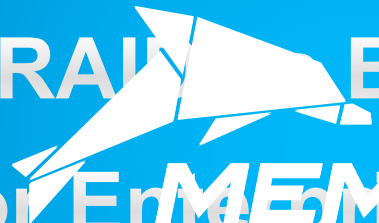

Flash Memory
SUMMIT
EMBLAZE

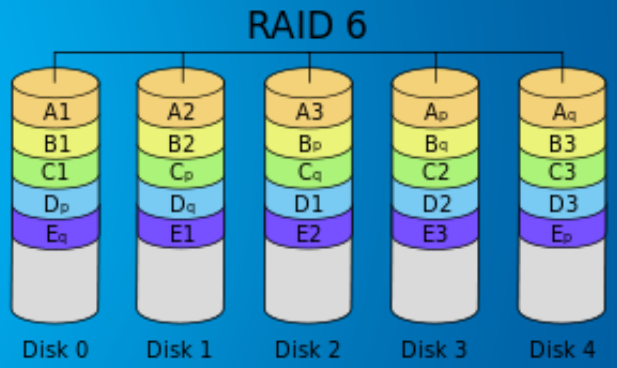
Flash RAID Building Blocks for Enterprise Storage



LU Xiangfeng CTO
Memblaze Co., Ltd

What is the status of enterprise RAID?

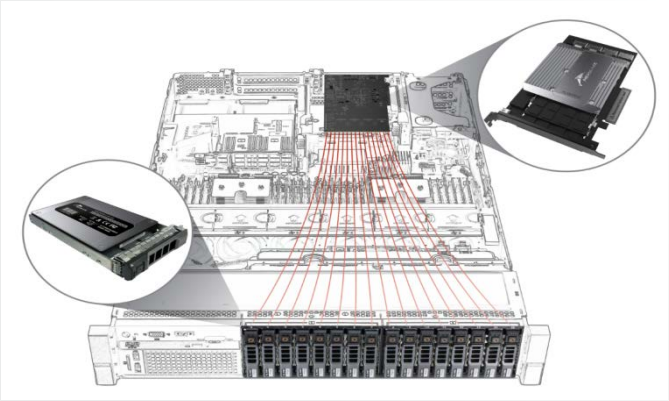
- Enterprises rely on RAID to increase system reliability, improve performance/ IOPS and acquire large capacity storage.
- Different RAID configurations are for different application scenarios. They are designed mainly for HDD, not optimized for SSD.
  - RAID 0, 1, 2, 3, 4, 5,…

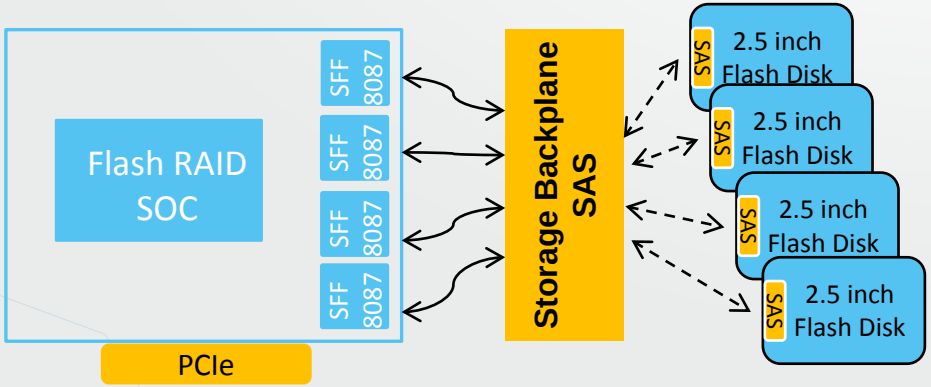


What plagues traditional RAID in flash era?

- Performance loss as result of partial stripe update.
- Increase write load and loss of efficiency due to parity update.
- Wear leveling may increase simultaneous failure rate.
- Full disk reconstruction and disk resynchronization.
- Can hardly support SSD features like trim, SMART, write atomicity, etc.

Our solution – Flash RAID storage system



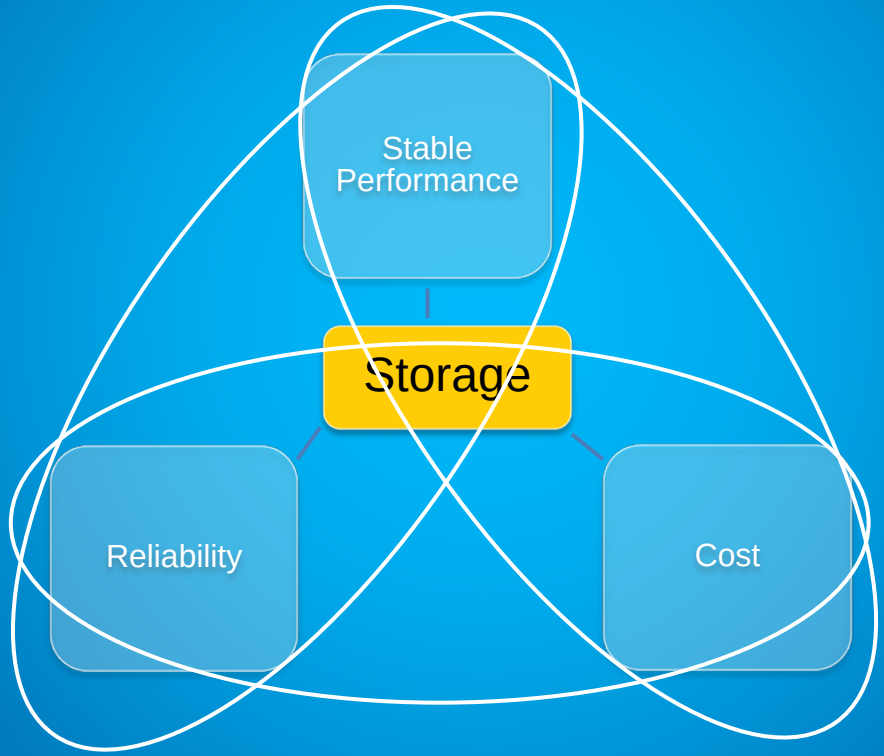


- NVMe 1.0/ PCIe 3.0 compatible host interface.
- SFF 8087 SAS cables connect one card with multiple disks to form a large storage pool. Each runs in 10Gbps bandwidth.
- A SOC card can drive up to 16 pieces of 2.5 inch flash disks with maximum capacity of 32TB.

Our solution – Flash RAID storage system (Cont'd)

- Superior IO performance and stable IO latency.
- Reliable and efficient RAID architecture supports declustered RAID which can avoid RAID6 penalty and enable fast data rebuild.
- Flexible capacity options and thin provision.
- Global wearleveling among disks and automatic load balance.

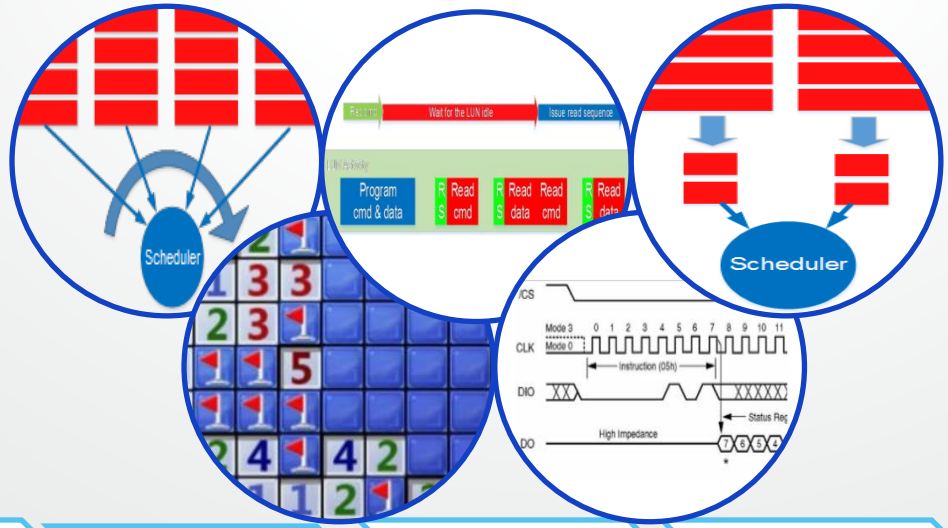
How Flash RAID trade-off these factors?



Why stable performance/ latency hard to achieve?

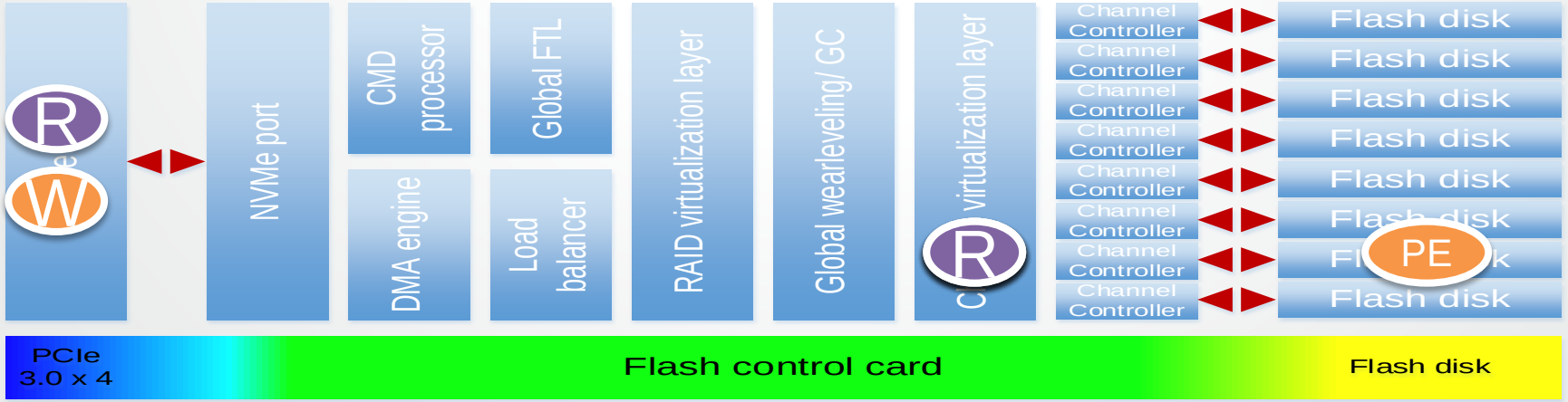
- New flash process node size and interface bring new challenges to flash controller.
  - How to leverage new flash features to compensate for the increasing latency jitter?
- Enterprise users poses critical service level requirement.
  - How to dynamically schedule IO requests so that different users are served in accordance with their demands?

Queues in typical enterprise flash drives



	IO command queuing	FTL lookup queuing	Flash lun queuing	Flash channel queuing	Host DMA queuing
Typical latency:	1us	1us	0 ~ 10ms	0 ~ 60us	~ 10us

Defend stable performance



- Aggregate flash in a server into a global storage pool. It is possible to optimize hardware resources and balance IO latency.
- New user write data and garbage collection migration data can be rearranged to minimize its competition with user data read IOs.
- Balance of different data recovery approaches to minimize user data read latency. i.e. Whether to use read retry or suspension to recover the data frame depends on the global queue condition and service quality.

How to keep consistent reliability?

- Enterprise storage has stringent requirement on data reliability. But flash cell reliability is shrinking with process node size becomes smaller.
  - How to maintain consistent data integrity with state of art flash?
- Write amplification plagues traditional RAID6. Disk replacement inures additional instability and jitter due to slow rebuild process.
  - How to keep data integrity without additional overheads like write penalty in usual RAID6?

Defend enterprise reliability

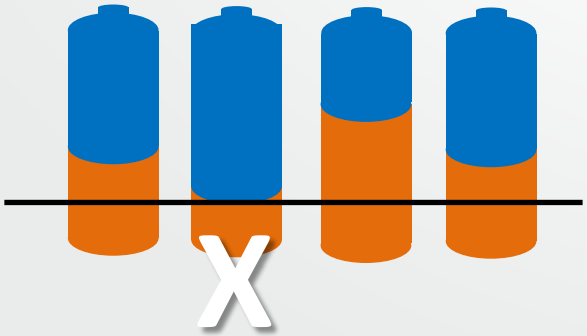
- Global wearleveling among disks delivers high endurance.
- RAID layer efficiently utilizes good blocks in each LUN and avoids RAID6 random write penalty. RAID over different disks can improve system reliability by rectifying disk level failures.
- Disk hot plug feature and fast rebuild processes improves storage data integrity and eases maintenance of fault disks.

What enterprise users care about storage features?

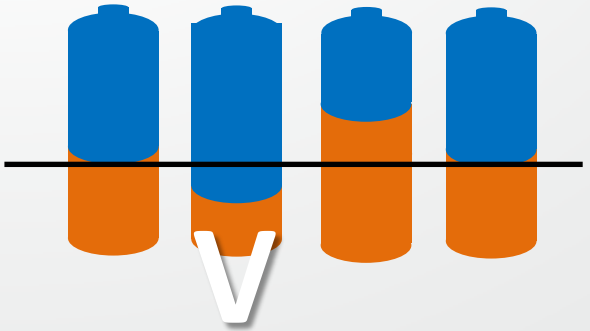
- While enterprise storage has strict requirements for performance and latency, the other factor hinders the adoption is per GB cost.
  - How to keep usable storage cheaper?
  - How to improve storage endurance?
- Enterprise usually requires dedicate policies for storage management.
  - How to keep storage management interface simple while maintain its functionality?

Global storage pool convoys cost deduction

- With global storage pool, OP can be shared globally, which can improve device utilization.
  - Suppose OP compensates for bad blocks, and the system fails when OP drops to a certain level (bad block count exceeds threshold).



Separated storages, whole system fails when bad block count exceeds threshold in any disk



Global storage pool, whole system fails when average bad block count exceeds threshold

Global storage pool convoys cost deduction (cont’d)

- Hardware resources can be shared by the global storage pool.
  - ECC engine and hardware queues can be shared globally.
  - Thin provision and deduplication can be enabled for the whole system.
- Storage pool maintenance is simplified.
  - Unique storage pool allocation and management.
  - Exhausted disk can be displaced without system service interrupt.

Conclusion

- With dedicated hardware acceleration, it is possible to deliver solutions with efficient performance, latency requirement and reliability.
- Aggregated storage pool not only offers large capacity, but also enables system level optimization for device usability, life, reliability and cost.

Flash RAID Advantage

	RAID0	RAID1	RAID5	RAID10	Flash RAID
Reliability	1	4	3	4	4
Bandwidth	5	2	2	3	4
IOPS	5	2	2	3	4
Latency	3	2	1	2	5
Cost /GB	5	1	4	1	4
Flash endurance	3	1	2	1	5
Total	19	11	12	12	22

Thank you

Please Visit us at Booth #322

MEMBLAZE