New Data Protection System for NVMe SSD

Alan Wu, Sr Director, Memblaze

zhongjie.wu@memblaze.com

2016/8/8

NVMe SSD is the Future

- Capacity of NVMe SSD keep increase, but price continue going down
  - The capacity of PblazeV is above 11TB
- NVMe SSD has been widely used in internet services and traditional industry customer start to choose NVMe SSD

nvm EXPRESS Server & SSD

Critical Issue:
Data Protection for NVMe SSD

- Hardware RAID
  - Hardware RAID supports SAS/SATA interface
    - Guarantee the data reliability of SAS/SATA SSD and HDD
  - Hardware RAID cannot support NVMe interface
    - NVMe SSD has extreme high performance
    - RAID controller becomes performance bottleneck
- Software RAID
  - Can aggregate the performance of NVMe SSD
  - Fully leverage the resource of multi-core CPU

Software RAID Concerns

- Data reliability
  - System crash, how to guarantee data reliability?
    - Considering NVMe SSD has high performance, data buffer can be removed to achieve the same data reliability as hardware RAID
- Performance
  - Does parity computing consume CPU cycles?
    - Intel CPU provides SSE/AVX acceleration instructions
- Resource utilization
  - Resource is shared by software RAID and application, how to isolate them?
    - Few source is consumed by software RAID, and CPU resource also can be controlled by Cgroup mechanism
- OS installation
  - Most of software RAID cannot support OS installation
    - Software RAID tool can be integrated into UEFI to support OS installation

Redefine NVMe software RAID

- NVMe software RAID revolution
  - Resolve all concerns about traditional software RAID
  - Address NVMe SSD new storage issues
- Key technologies to redefine NVMe software RAID
  - Leverage CPU multi-cores and improve CPU efficiency
    - Lock-free algorithm
  - Enlarge SSD lifespan and avoid simultaneous failure
    - Global wear leveling and anti-wear leveling
  - Enhance data reliability to make system more robust
    - New SSD failure model and algorithm
  - Avoid data loss in case system power failure
    - Remove data buffer and do SYNC IO handling model

FlashRAID – NVMe Software RAID

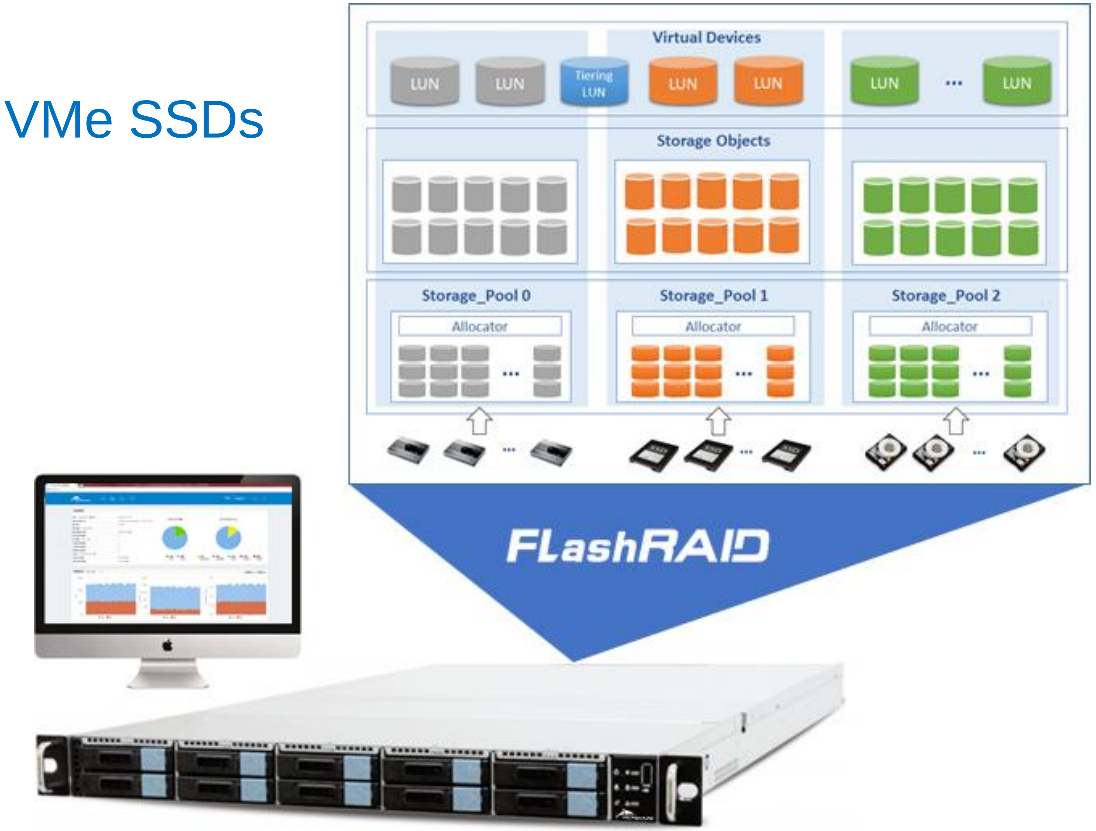
- Major features
  - Guarantee data reliability for NVMe SSD
  - Aggregate the performance of multiple NVMe SSDs
  - Fully support data automatic tiering
  - Easy storage management

1 nvm EXPRESS RAID

250 + IOPS

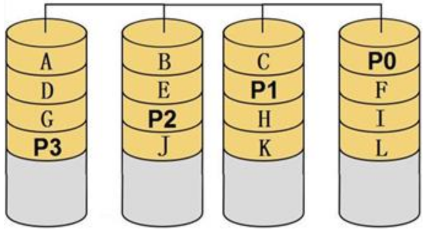
2 + Tiering

30 – Fast Rebuild

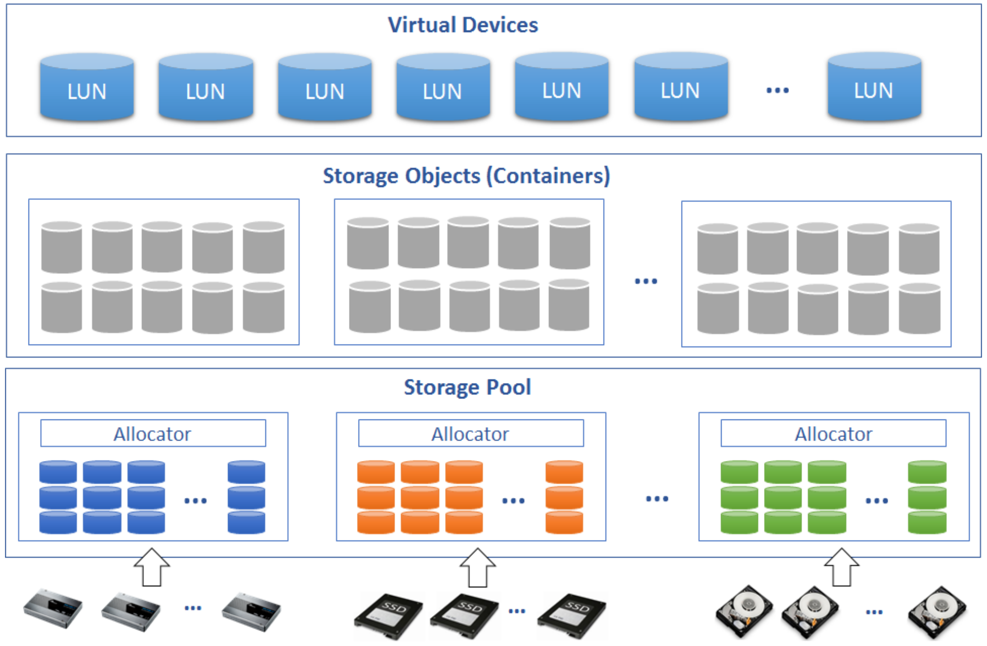


Software RAID New Architecture: Storage Virtualization based Architecture

- Storage Virtualization
  - Can improve data reconstruction performance significantly
  - Can fit SSD new failure model and the characteristics of SSD
  - Provides basic mechanism to do global wear leveling and anti-wear leveling
  - Simplify RAID to do IO QoS



VS

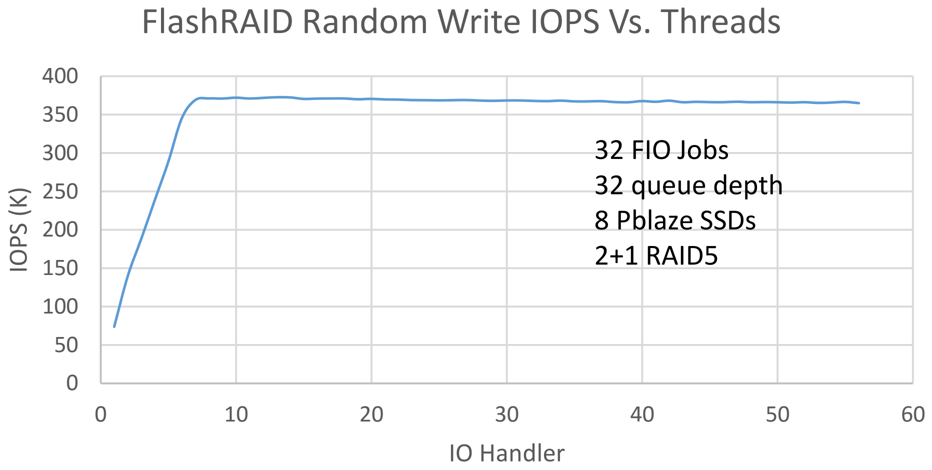


CPU becomes IO Bottleneck

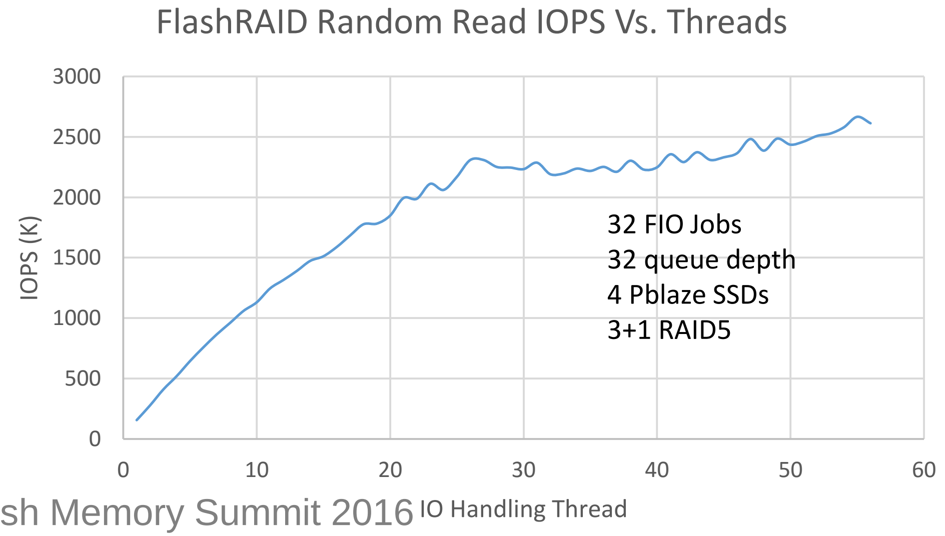
- NVMe SSD has high performance, CPU becomes bottleneck
  - Traditional storage software focus on disk issue
    - Magnetic drive has address seeking issue
  - NVMe RAID software need to take care CPU bottleneck issue, more than SSD issues
    - Memory issue / NUMA issue
    - Lock issue
    - Task scheduler issue
- Aggregate CPU multi-cores, satisfy development trends of CPU
  - Divide task and resource reasonably
  - Innovative lock-free producer-consumer algorithm
  - Patented lock-free flow control algorithm

Performance Scalability Evaluation

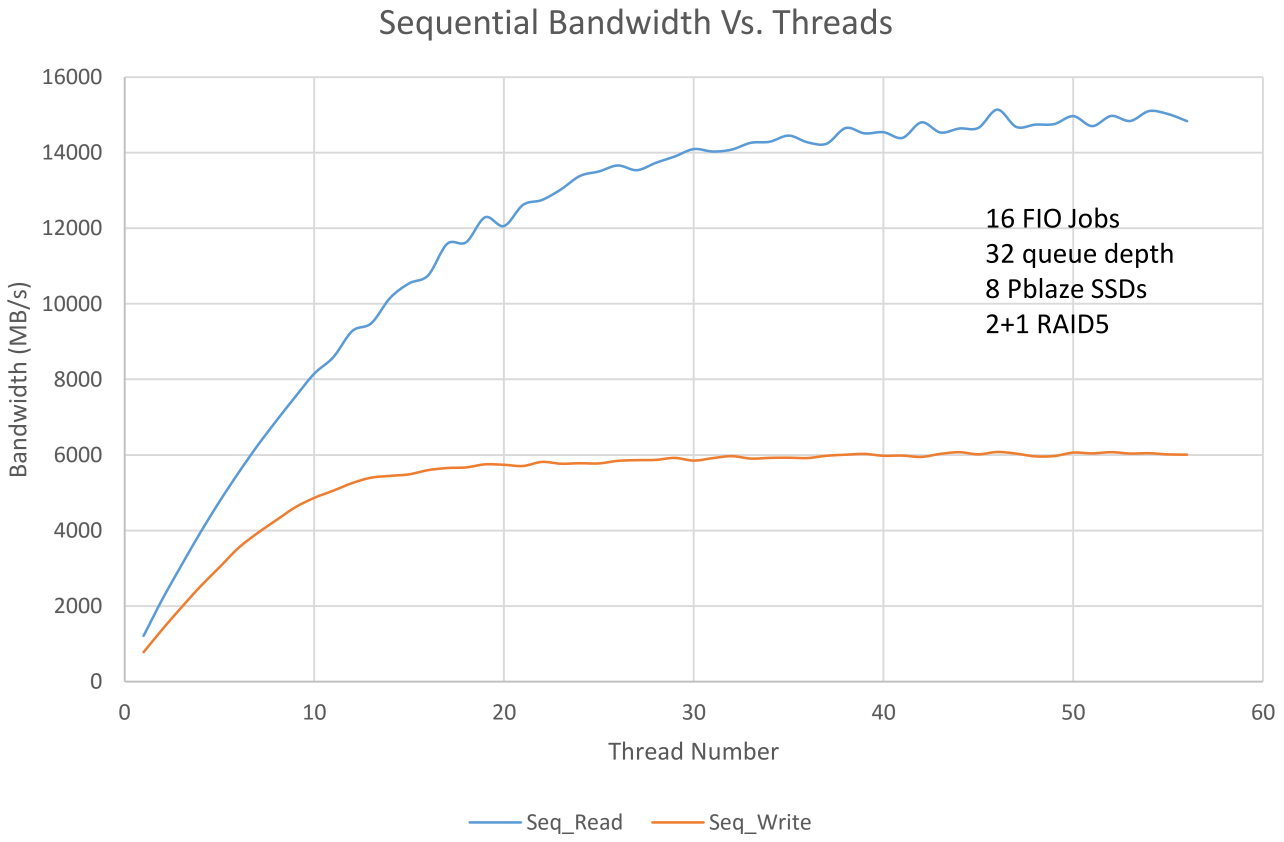
- Performance can be increased with the number of CPU cores


FlashRAID Random Write IOPS Vs. Threads

32 FIO Jobs
32 queue depth
8 Pblaze SSDs
2+1 RAID5


FlashRAID Random Read IOPS Vs. Threads

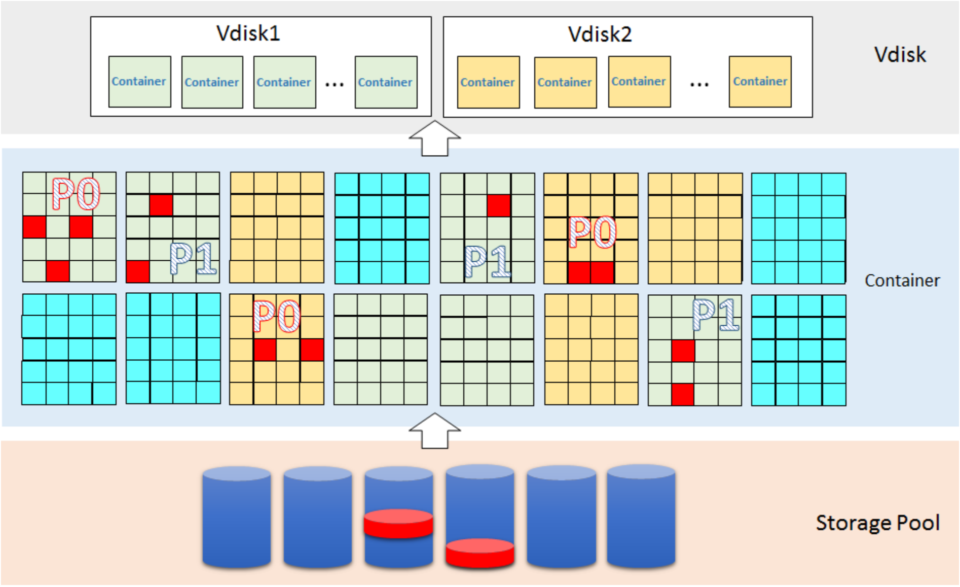
32 FIO Jobs
32 queue depth
4 Pblaze SSDs
3+1 RAID5


Sequential Bandwidth Vs. Threads

16 FIO Jobs
32 queue depth
8 Pblaze SSDs
2+1 RAID5

Innovative Smart Data Reconstruction

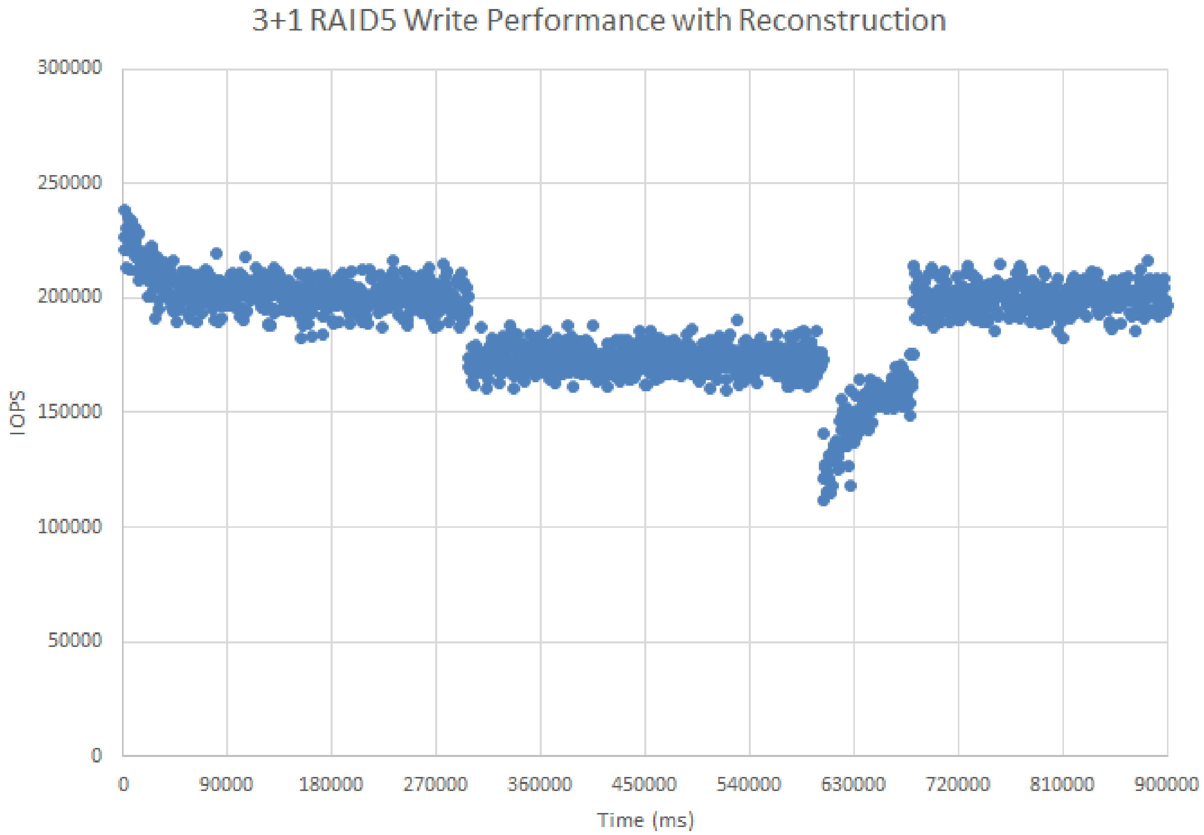
- Traditional reconstruction is running from begin to end on a hard disk drive
  - Cannot aware data and data loss risks
- Reconstruction method can be changed to enhance data reliability
  - NVMe SSD has excellent performance
- Smart data reconstruction
  - Data aware reconstruction
    - Take storage object as reconstruction unit
  - Priority based reconstruction
    - Data with high data loss risk should be reconstructed in high priority
  - Partial fast reconstruction

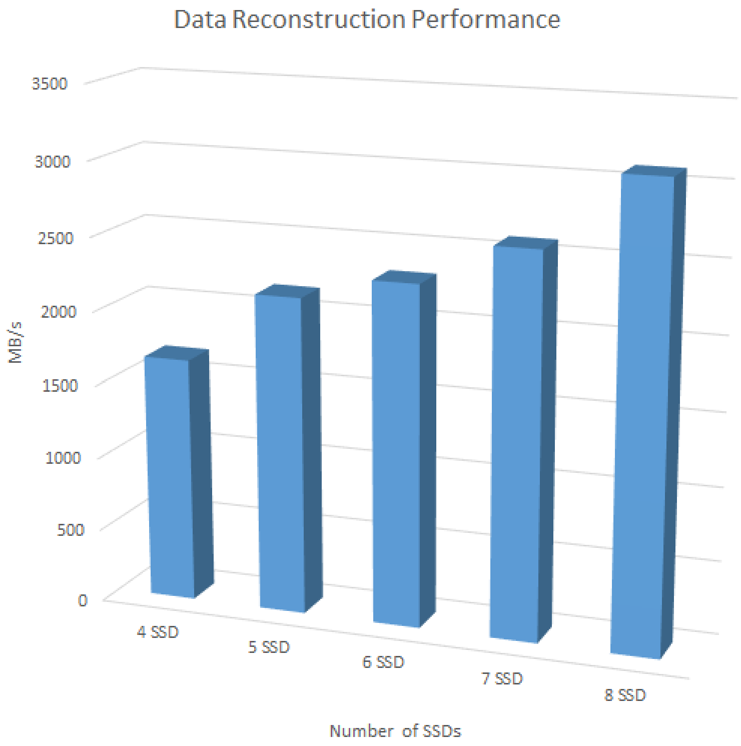


Priority based Reconstruction

FlashRAID Reconstruction Evaluation

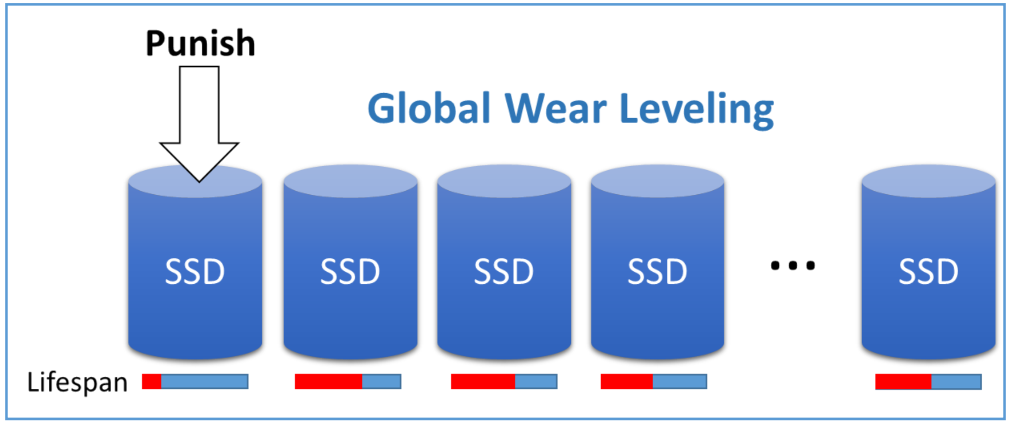
- Reconstruction performance can be scalable

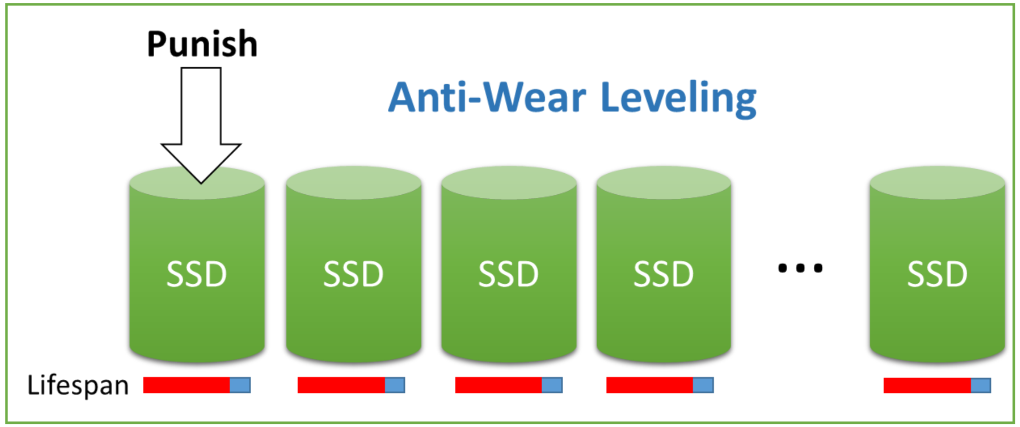




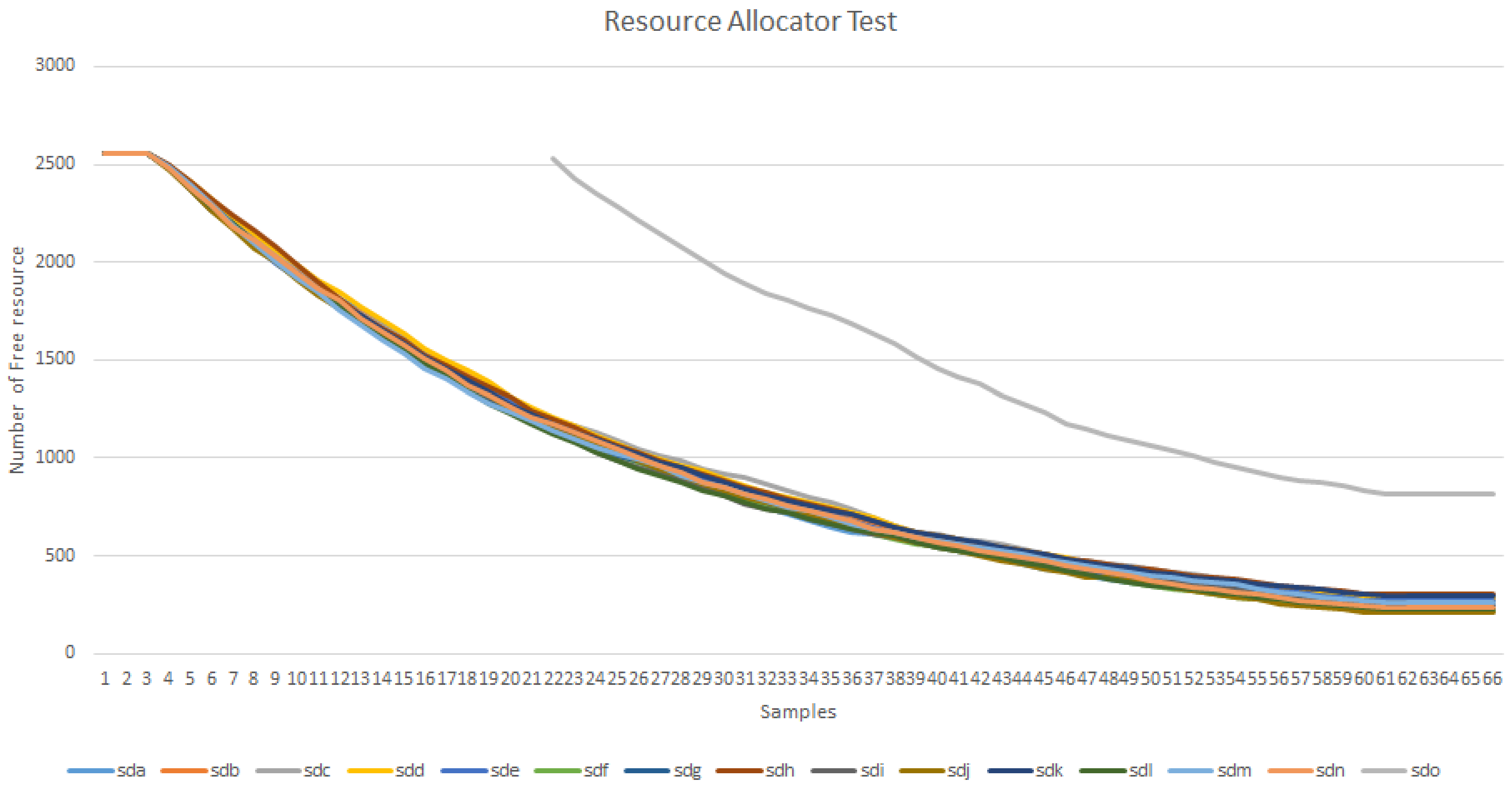
SSD Lifespan Management

- Question
  - SSD has limitation of lifespan, how to maximize lifespan in system level?
- Global wear leveling between multiple SSDs
  - Static / dynamic wear leveling algorithm
- Anti-wear leveling to avoid simultaneous failure
  - PE cycle information in each SSD is the key parameter



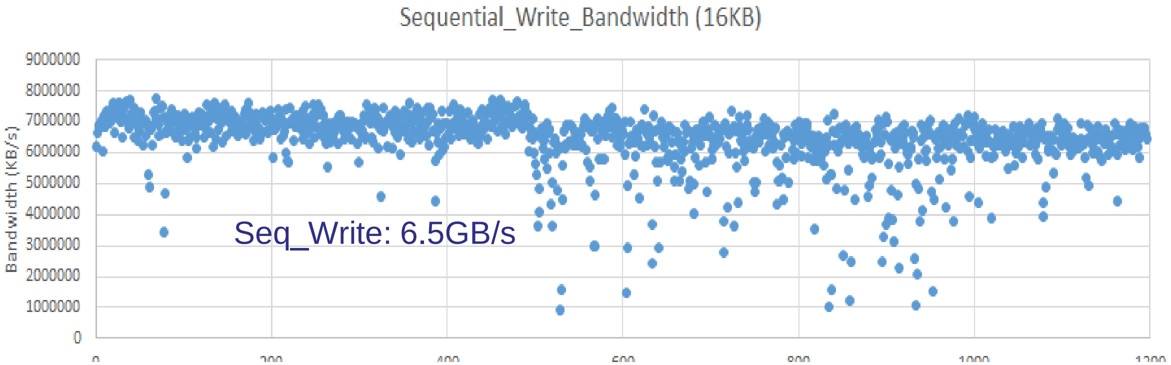


Weight based Pseudo Resource Allocator

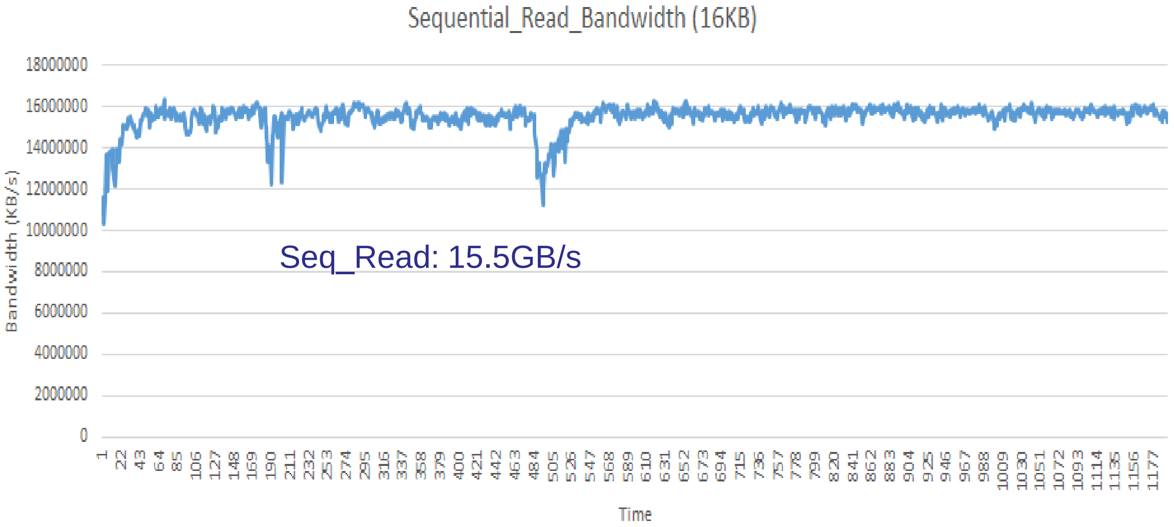


FlashRAID Performance

Test Configuration: Xeon2680 (v2), two CPU sockets, 6 Pblaze4, RAID5

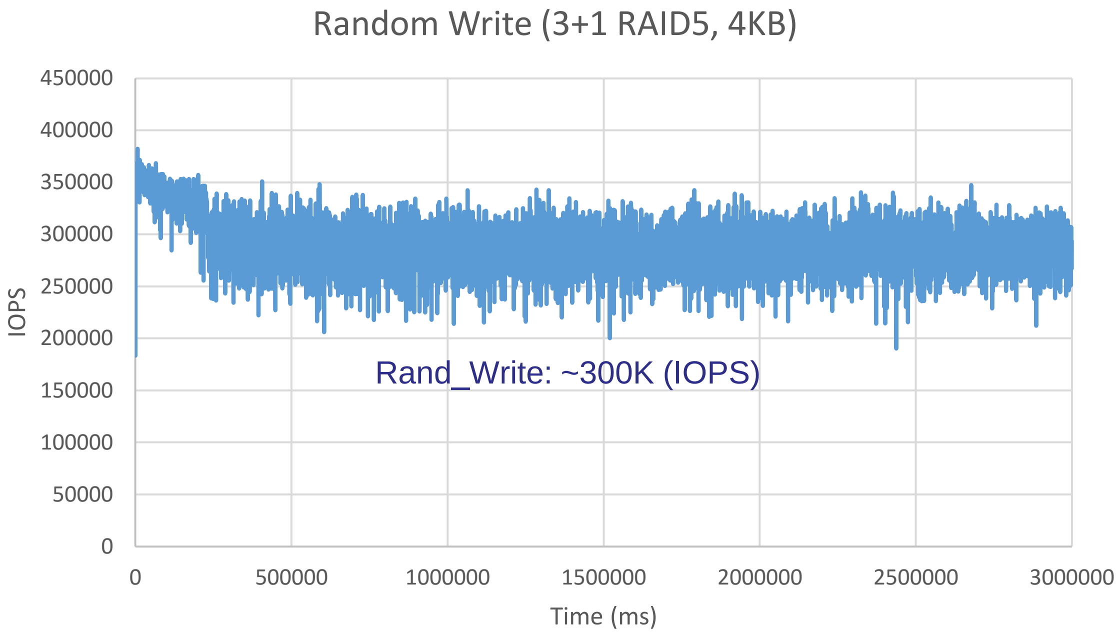

Sequential_Write_Bandwidth (16KB)

Seq_Write: 6.5GB/s

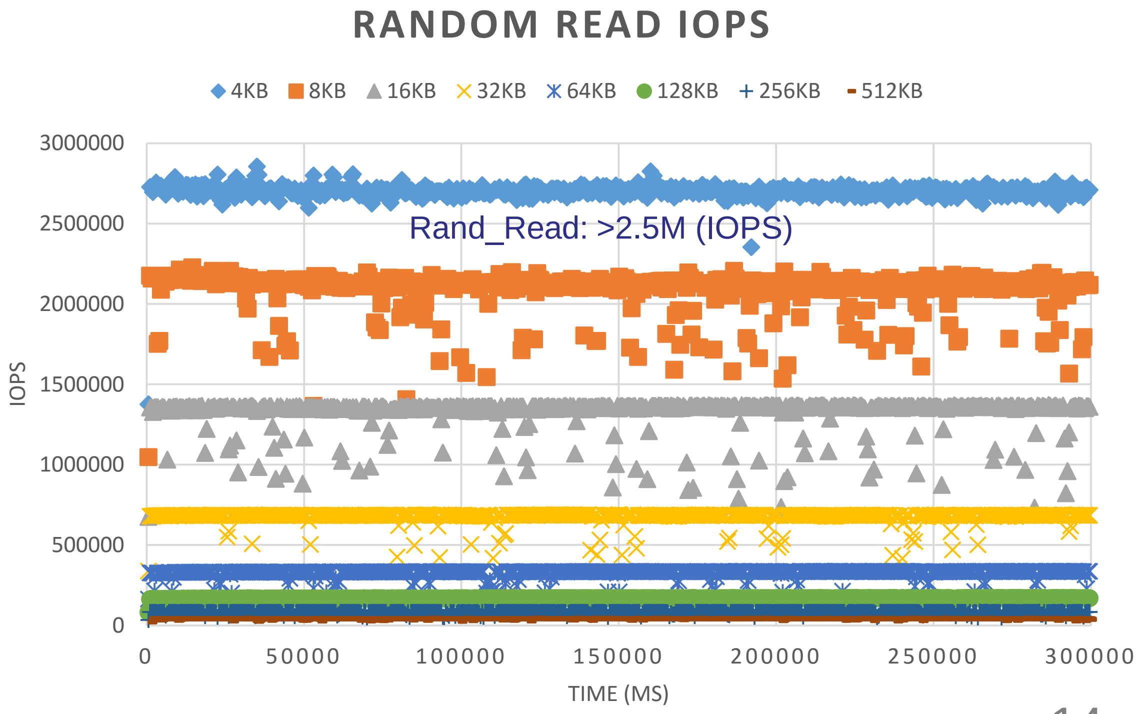

Sequential_Read_Bandwidth (16KB)

Seq_Read: 15.5GB/s

Bandwidth Test Result


Random Write (3+1 RAID5, 4KB)

Rand_Write: ~300K (IOPS)


RANDOM READ IOPS

Rand_Read: >2.5M (IOPS)

Conclusions

- Software RAID is suitable for NVMe SSD data protection
  - Aggregate performance of NVMe SSD
  - Fit the development trends of CPU multi-cores
- Redefine software RAID for NVMe SSD
  - Lock free algorithms to resolve performance bottleneck of CPU
  - NVMe oriented design methods to resolve new issues of SSD
- FlashRAID is an excellent data protection solution for NVMe SSD

MEMBLAZE

FlashRAID®

http://www.memblaze.com