Reducing Latency and Improving Performance Consistency in NVMeOF

Parag Maharana (Seagate)

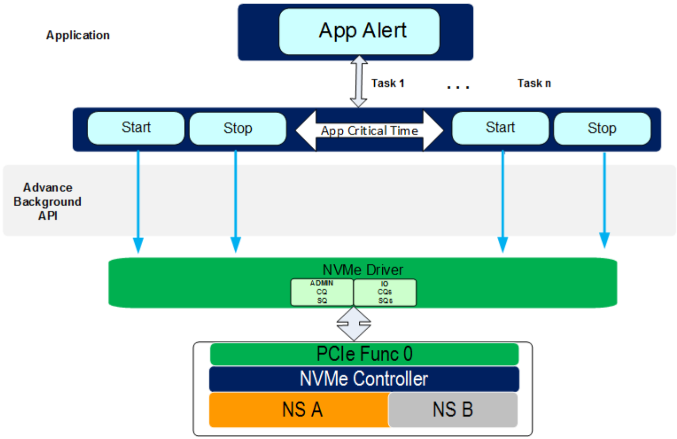
K R Kishore (Pavilion Data)

Overview

- NVMe is very well suited for All Flash Array
- NVMe provides
  - Low latency and high throughput
  - Quality of Service attributes enhance better performance
- However, NVMe has optional features that can be used to provide predictable and consistent latency

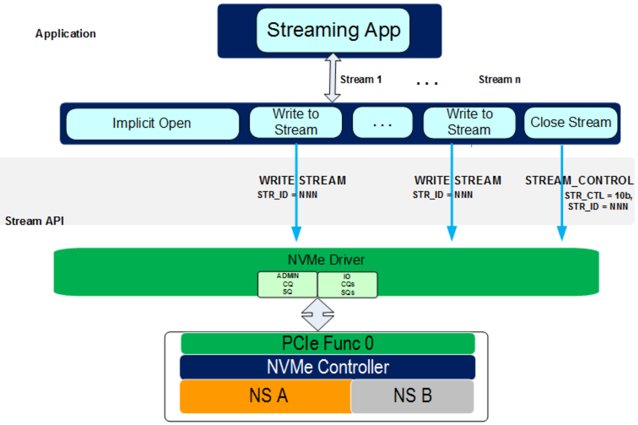
NVMe Optional Features

Advanced Background Control



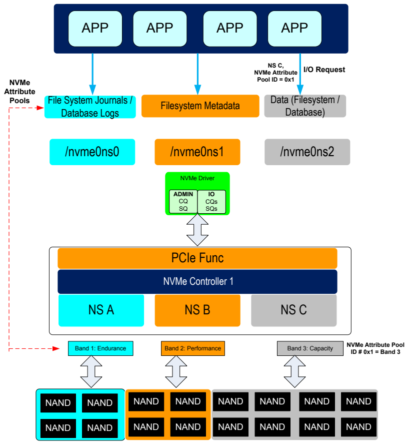
Enable Host Control base Consistency latency

Streams



Efficient Flash Utilization Performance

Attribute Pools



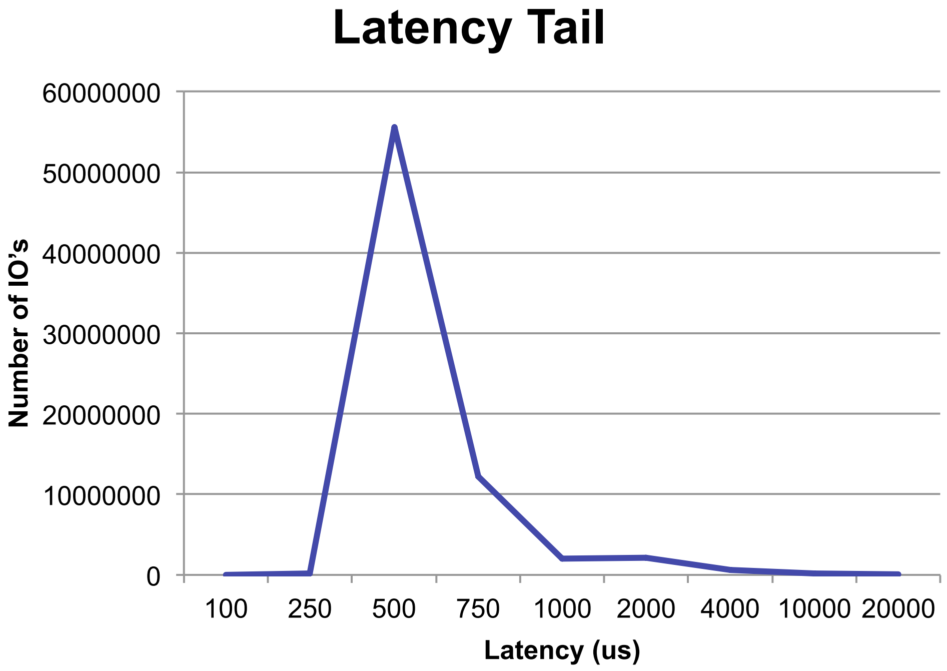
Enable multiple workloads with different characteristics on the same NAND

SEAGATE

Pavilion Data Systems

Storage System – Latency Requirements

- Storage systems desire predictably consistent latency
- Access Latency
  - Deviations from median
  - Distributed in time
- Storage Arrays can minimize the latency outliers compared to discrete direct attached storage
- Use NVMe standards framework
  - Directives implemented by Seagate
  - Directives co-ordination and data handling by Pavilion

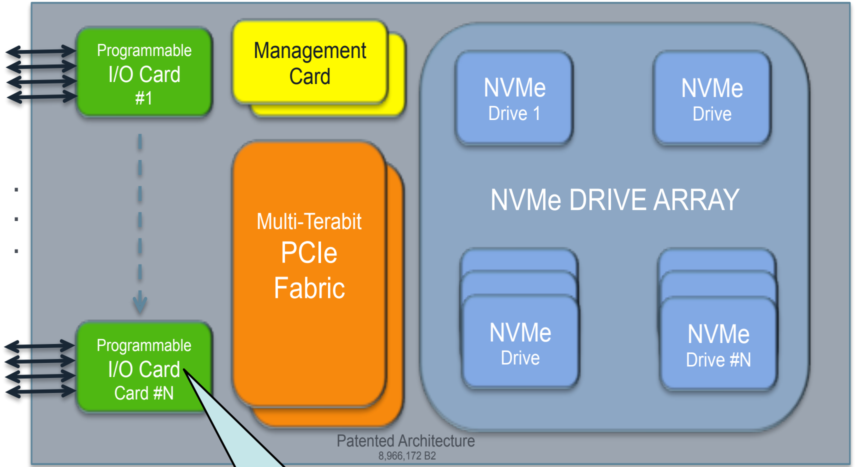


SSD Latency Management

- Read/Write latency variations occur when SSD controller is busy with media management tasks such as Garbage Collection, Wear Leveling etc.
- Standards based Advanced Background Control (ABC) primitives implemented by Seagate
  - Host controls ABC start and stop operations
  - SSD provides credit thresholds as a proxy for the amount of maintenance work that needs to be done
- Storage Array Controller takes on the responsibility of invoking ABC on all drives in the array in a round robin fashion
  - Scheduling can be time based or credit based

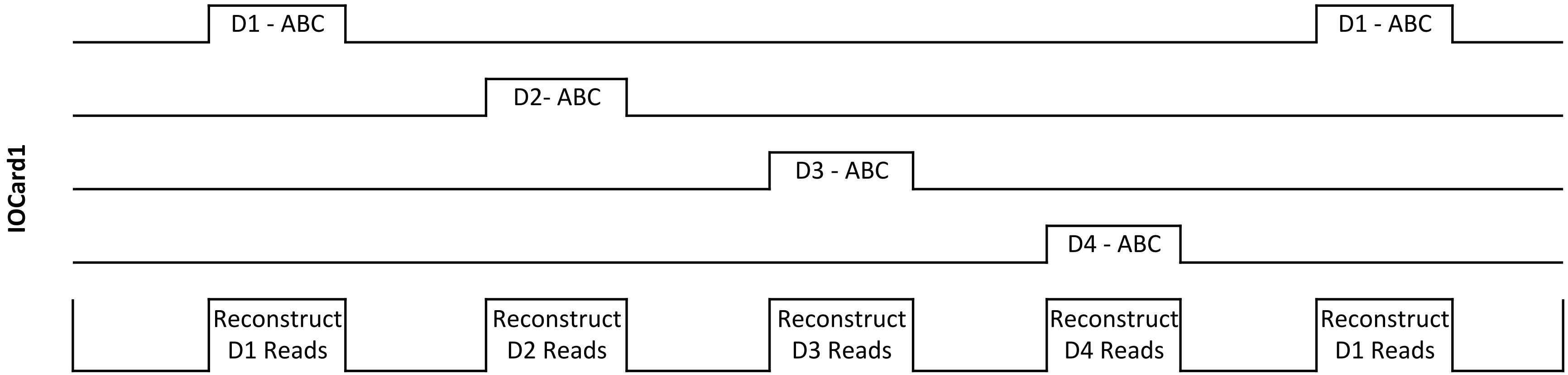
ABC in a Flash Array - mechanics

- Scheduler based ABC
  - IO controllers take on the responsibility of invoking ABCs in a round robin fashion
  - IO controllers invokes ABC on one drive at a time
  - *n* IOCs can co-ordinate scheduling across drives
  - IOCs use erasure coding to generate read responses when target SSD is under maintenance by treating them as temporary failures



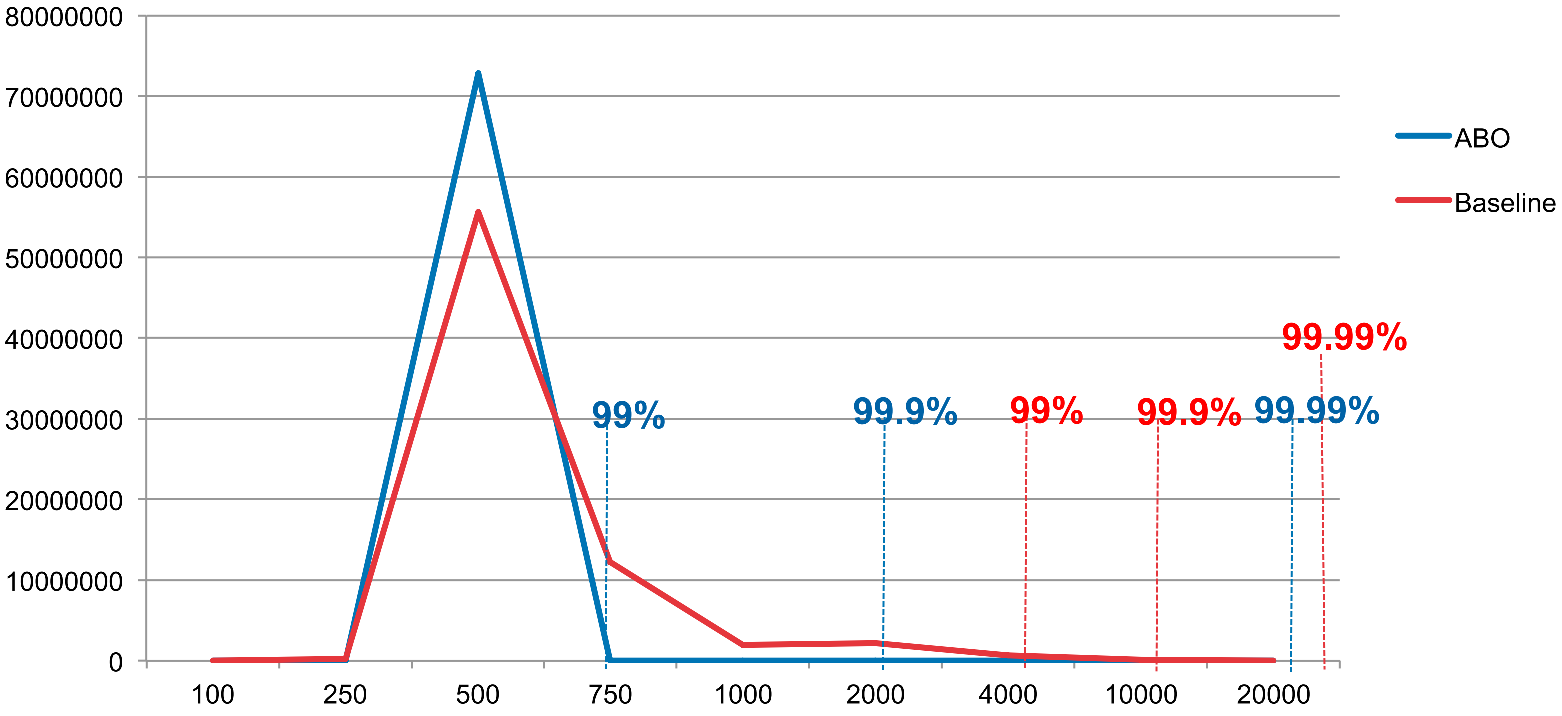
Round Robin or Credit based Scheduler managing ABC across drives

Scheduler based ABC



- IO Card1 is scheduling ABC across 4 drives
- IO Card1 will reconstuct read data when a particular drive is under maintenance

Latency distribution with ABC



Thank You! Questions?

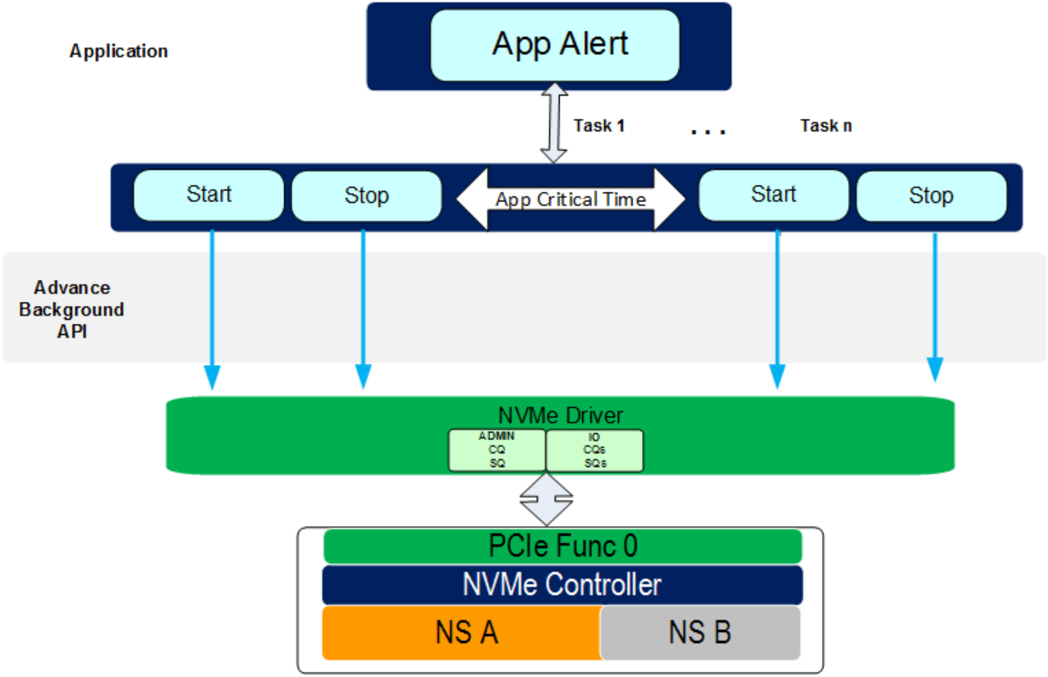
www.seagate.com/flash

Learn about Seagate's ever-expanding portfolio of SSDs,
Flash solutions and system level products for every segment

www.paviliondata.com

Advanced Background Control

- Advanced Background operation provides a method to control application critical response time
- This allows system to control Garbage collection, Wear leveling and Critical task time
- Start and Stop primitives allow application to alert SSD to start or stop background operation



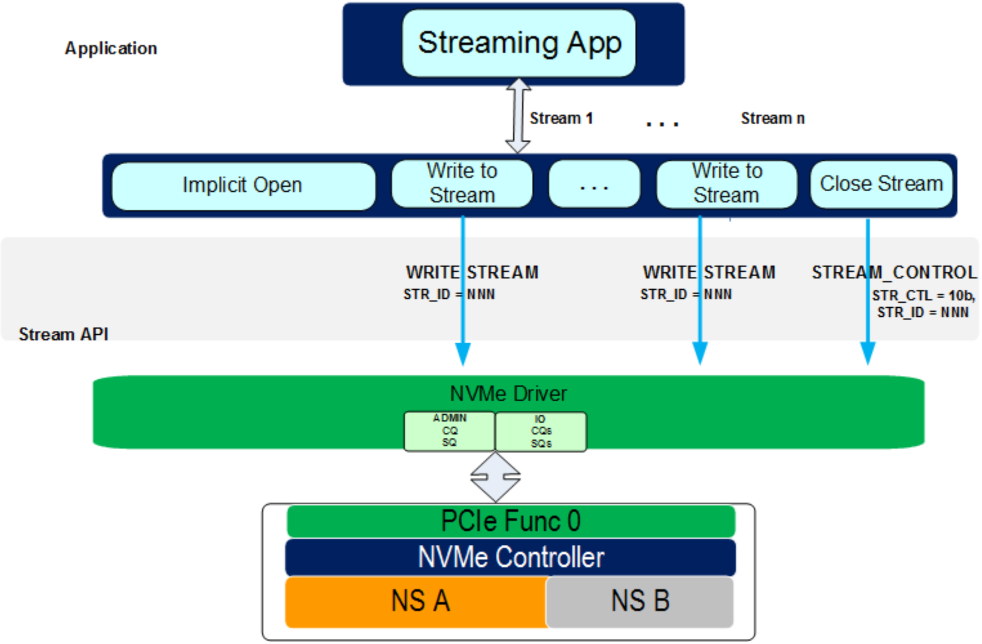
Enable Host Control base Consistency latency

SEAGATE

Pavilion Data Systems

Multi-Streaming Standard

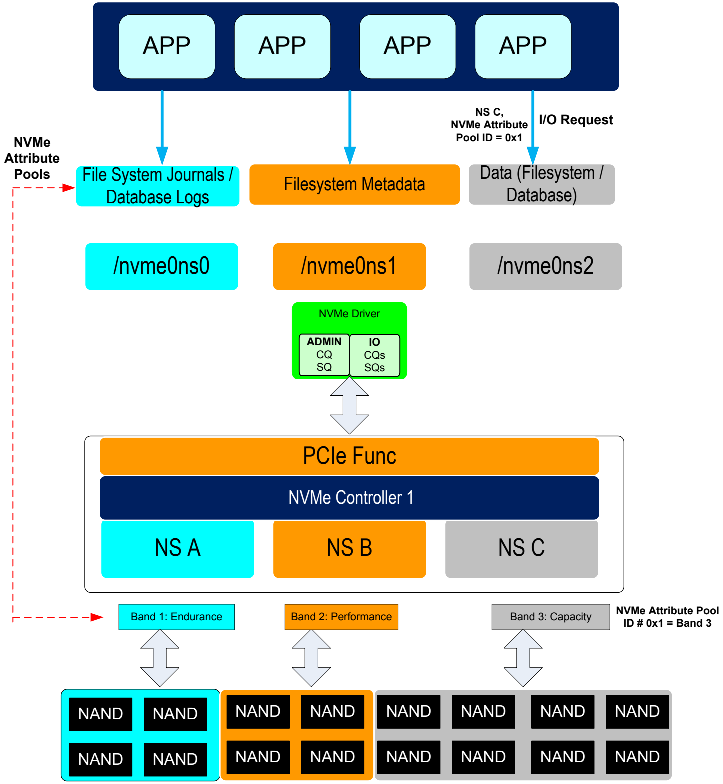
- Steams allow data written together on media so they can be erased together that will minimize garbage collection resulting in reduce write amplification.
- Stream IDs are per namespace
- Max Streams Limit (MSL) is the maximum number of concurrently opened streams in the ‘NVM subsystem’
- Optimal Stream Write Size (OSWS) and Stream Granularity Size (SGS) are per namespace



Efficient Flash Utilization Performance

Attribute Pool

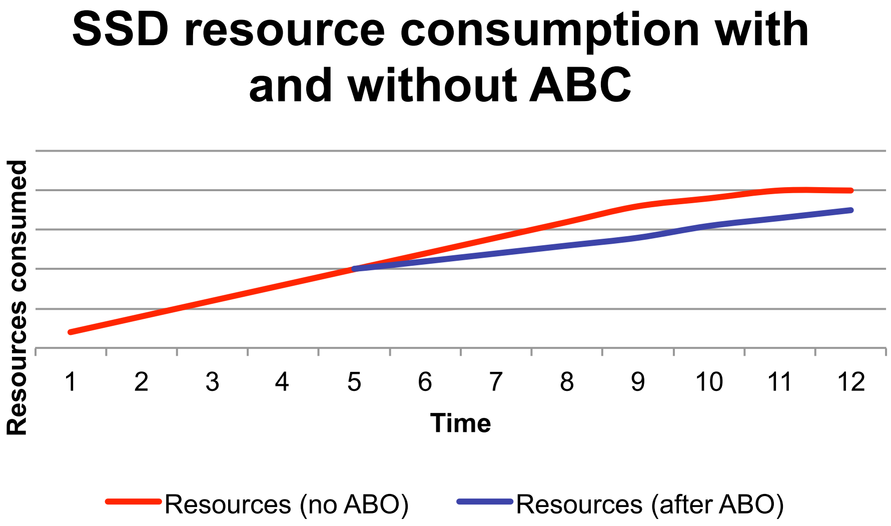
- Build bands using low cost NAND
  - Capacity band,
  - Performance Band
  - Endurance Band
- Expose Characteristics of Band/ NS
  - Latency range
  - Bandwidth and IOPs range
- Quality of Service



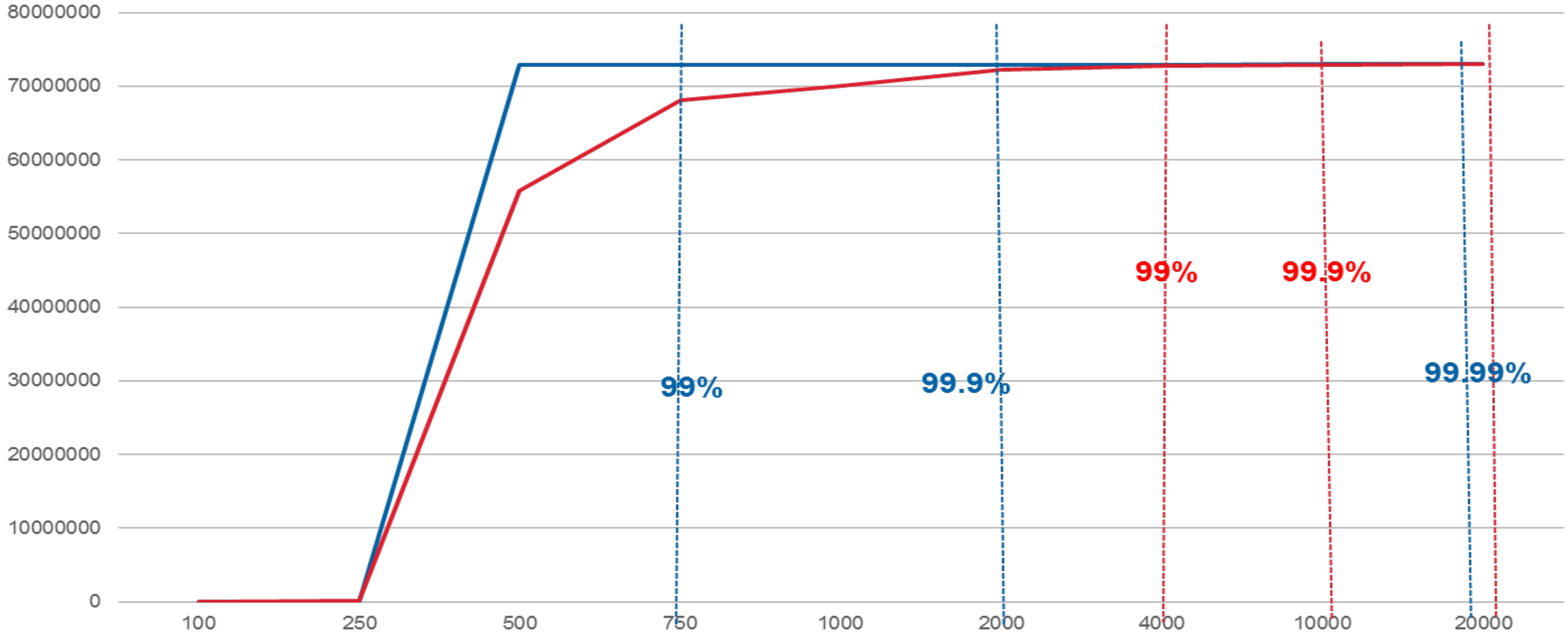
Enable multiple workloads with different characteristics on the same NAND

ABC in a Flash Array - Challenges

- Resources get consumed as the SSD fills up
- If drive becomes truly full, wiggle room is proportional to amount of Over Provisioning
- Goal is to start reclaiming space by proactively initiating ABC
- Throughput determines rate of ‘garbage’ creation
- Tradeoff cost vs benefits: (OP, max allowed throughput) vs consistent latency depending on workload



Cumulative Latency with ABC


80000000
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
100
250
500
750
1000
2000
4000
10000
20000
99%
99.9%
99%
99.9%
99.99%