SSD Architecture for Consistent Enterprise Performance

Gary Tressler and Tom Griffin
IBM Corporation

August 21, 2012

SSD Architecture for Consistent Enterprise Performance - Overview

- **Background:**
  - Client feedback indicates that traditional approach to managing SSD operations and maintenance activities concurrently is ***no longer acceptable*** (e.g., minimizing avg. maximum response per interval)
    - Enterprise users beginning to pursue 24/7/365 SSD-driven business operations – response time interruptions not tolerable throughout SSD lifetime
- **New Approach:**
  - SSD must provide consistent performance over its designated life span
  - All SSD maintenance activities must be managed in background
  - SSD performance may need to be sacrificed to a limited extent to achieve these goals

SSD Architecture for Consistent Enterprise Performance - Overview

- **Examples of Required Enterprise SSD Operation Profile**
  - **Background operations** should be performed continuously, and require a consistent level of throughput, or always done in low priority (never consuming an appreciable amount of host bandwidth)
    - No background task should take high priority if sufficient idle time not available
    - Relocation algorithms due to read disturb mitigation and wear leveling must operate consistently and constantly and should not result in large spikes or dips in host performance
  - Any **power backup circuit** check (e.g., capacitance monitoring) cannot ever stall the host
  - **Garbage collection and free space reclamation** should be managed in such a way that critical limits in free resources that will likely result in large stalls or host performance dips are not reached
  - **ECC correction** circuitry must have sufficient bandwidth to maintain performance with increased need to correct sectors as SSD ages
  - Must ensure that **mixed read and write workloads** do not dip below IOPs level that 100% reads or 100% writes can achieve
    - e.g., reads should not be gated behind large writes
  - Must be mindful of performance differences resulting from **workload changes** depending on level of preconditioning
  - All types of **software locks** should be done in such a way to minimize stalls to specific I/O

Performance Consistency Characterization Experiment #1

JEDEC Enterprise Workload

- 3 random workloads
  - Transfer size mix
    - 512B (4%)
    - 1KB (1%)
    - 1.5KB (1%)
    - 2KB (1%)
    - 2.5KB (1%)
    - 3KB (1%)
    - 3.5KB (1%)
    - 4KB (67%)
    - 8KB (10%)
    - 16KB (7%)
    - 32KB (3%)
    - 64KB (3%)
  - Max. I/O rate, QD = 32, incompressible data
  - 5s measurement intervals
  - Workload mix:
    - #1 (50% overall workload skew, 5% drive range)
    - #2 (30% overall workload skew, 15% drive range)
    - #3 (20% overall workload skew, 80% drive range)

Characterization Environment

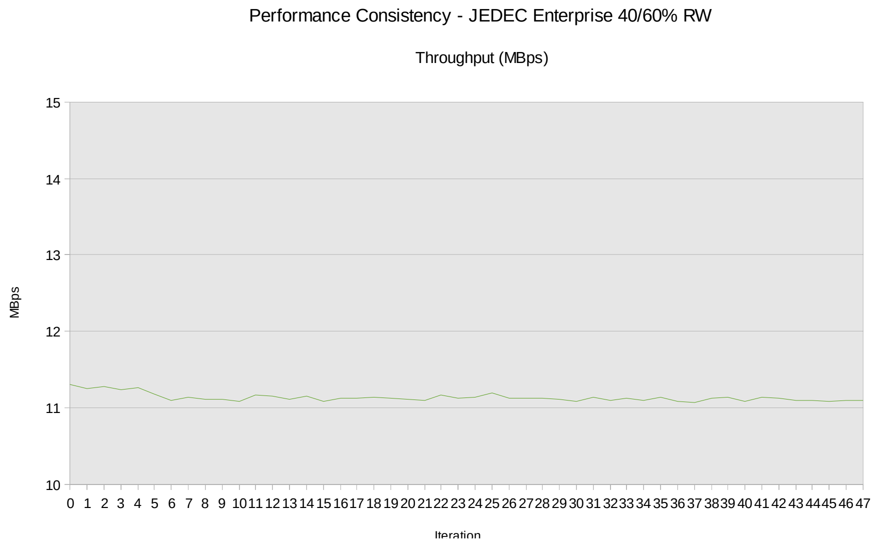
- PC-based
- Windows 7
- LSI HBA
- Various Enterprise SSDs
  - SAS, SATA
  - 2.5” SFF, 1.8” SFF
  - Different capacities

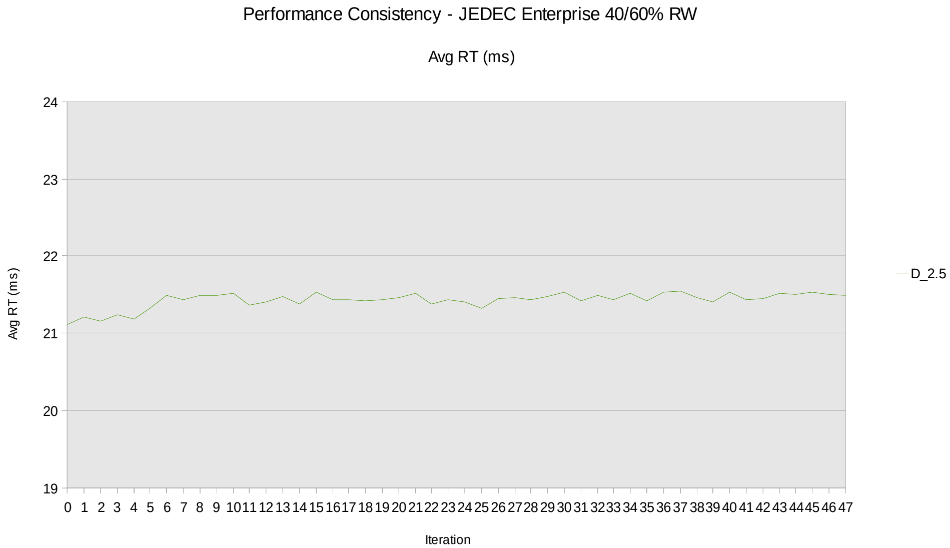
Note: Average Maximum Latency (AvgMaxRT_5sInt) = the average of the maximum latencies reported by exerciser where each maximum latency is recorded at a 5s interval

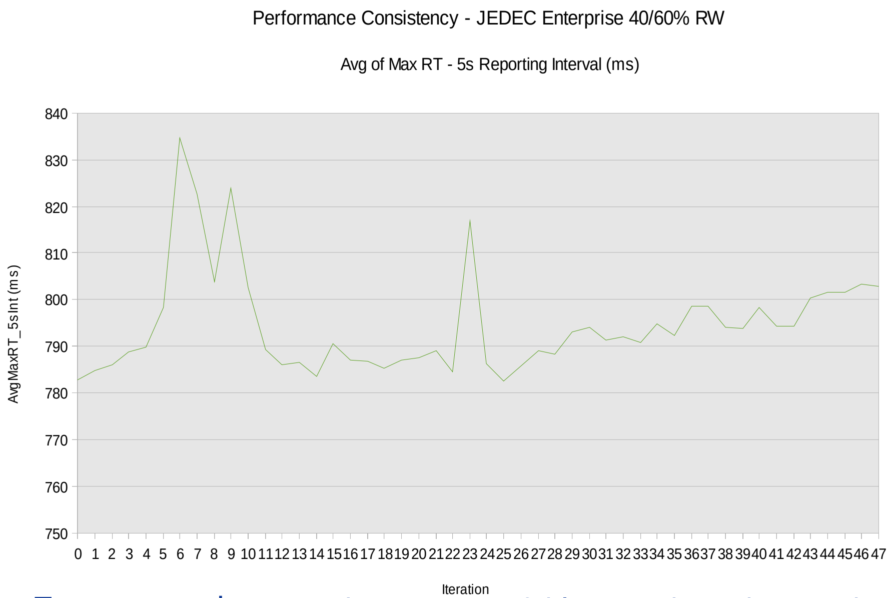
Testing

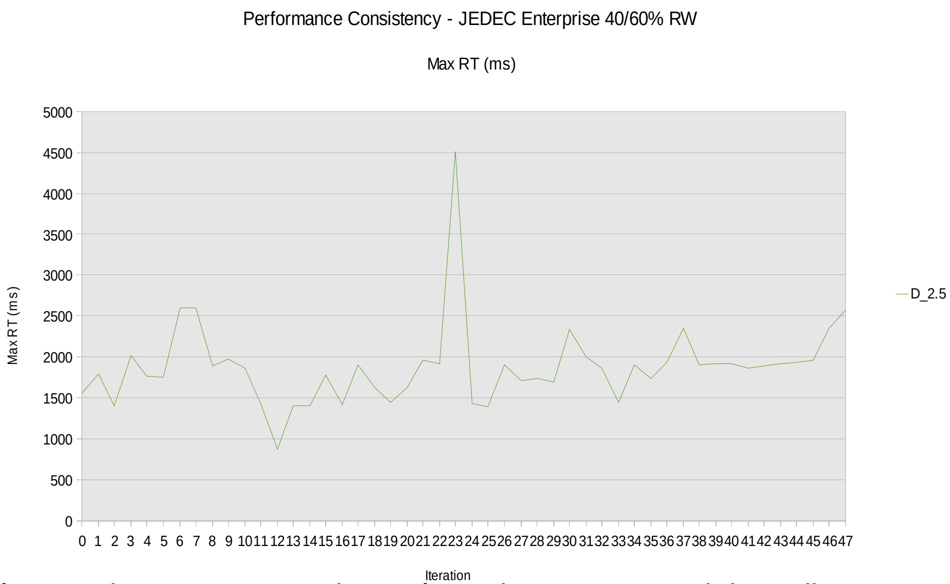
- Continuous iteration of above workload as follows:
  - 8-hour run at 100% write
  - 8-hour run at 40/60% RW mix (defined JEDEC Enterprise workload)
- Initial 24-hr. preconditioning with JEDEC Enterprise workload (100% write)

2.5” SATA - Performance Consistency Experiment #1



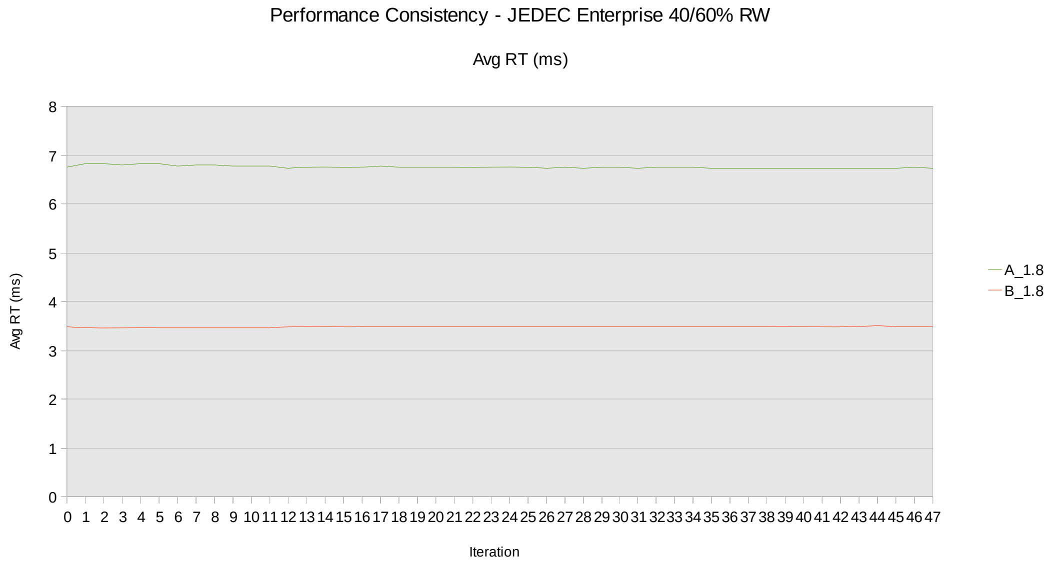




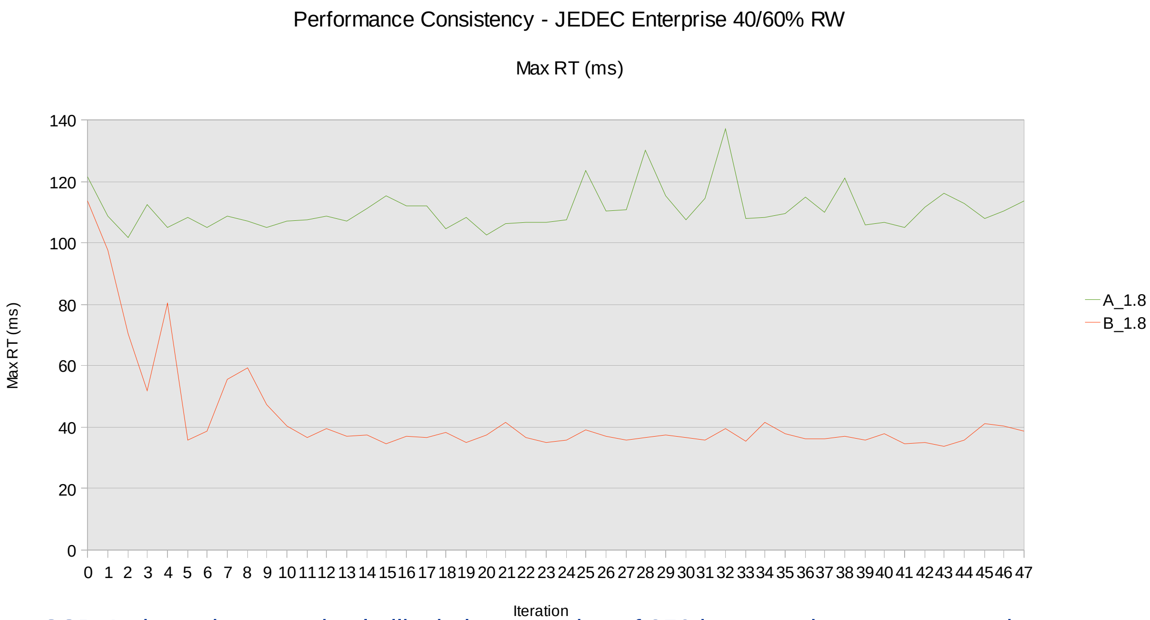


- Entry enterprise SSD demonstrates fairly even throughput and avg. latency, but avg. max. and max. latencies are poor and degrading

1.8” SATA – Performance Consistency Experiment #1

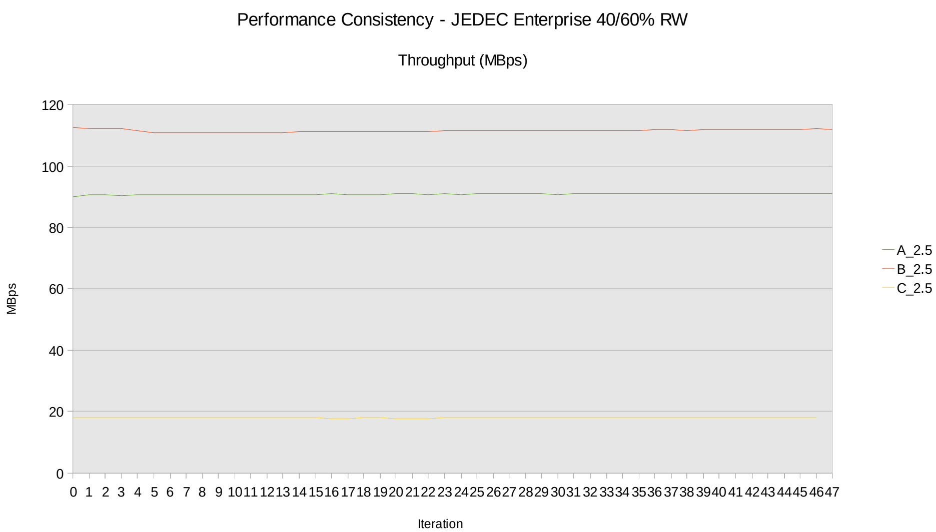


- SSDs show relatively stable average response time (and throughput) over approx. 350 hour test

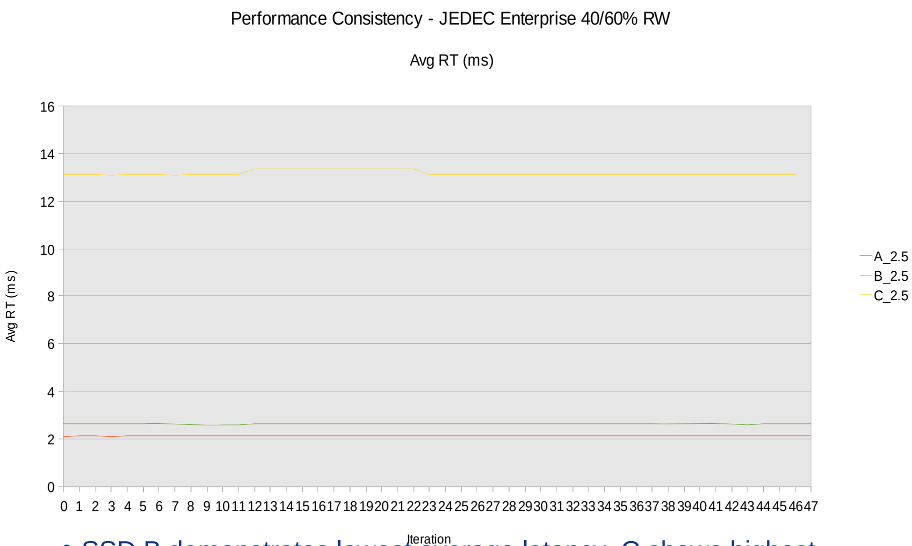


- SSD A shows increased volatility in latter portion of 350 hour maximum response time test

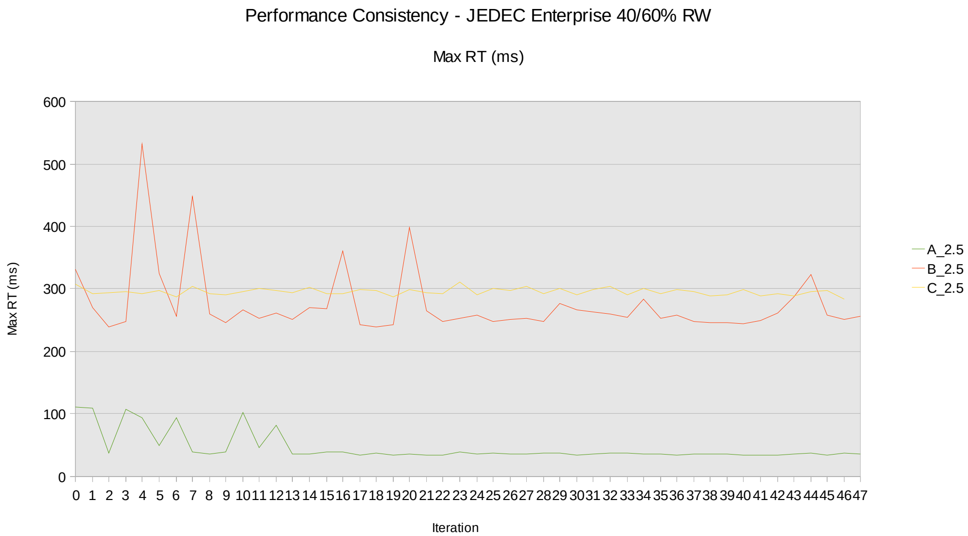
2.5” SAS – Performance Consistency Experiment #1



- SSD B demonstrates highest throughput, C shows lowest



- SSD B demonstrates lowest average latency, C shows highest



- B shows largest magnitude and deviations in maximum latency, while C demonstrates even result
  - Users may need to evaluate tradeoffs between throughput/average latency and maximum latency

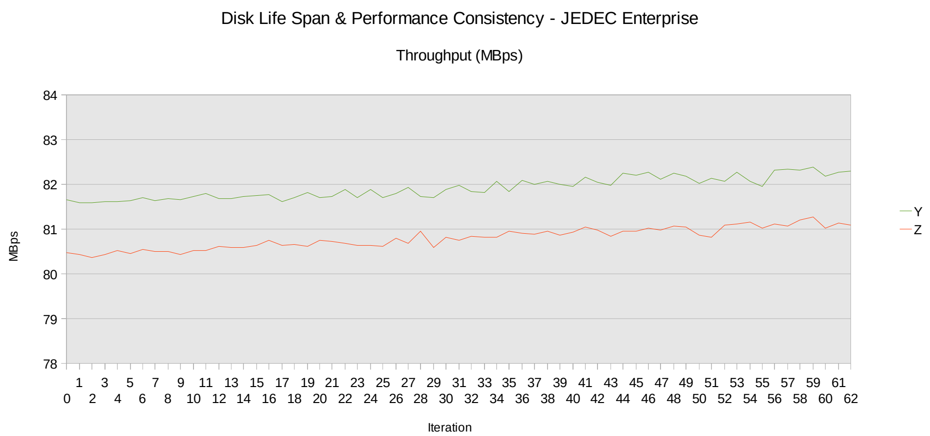
Disk Life Span / Performance Consistency Experiment #2

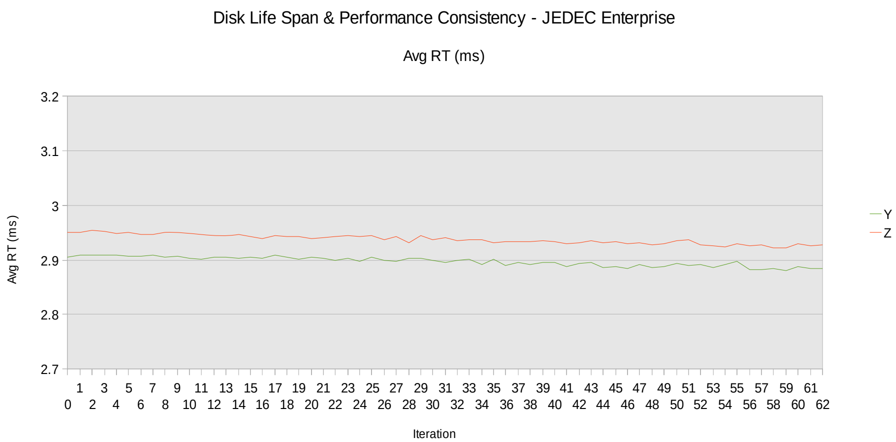
Testing Iteration

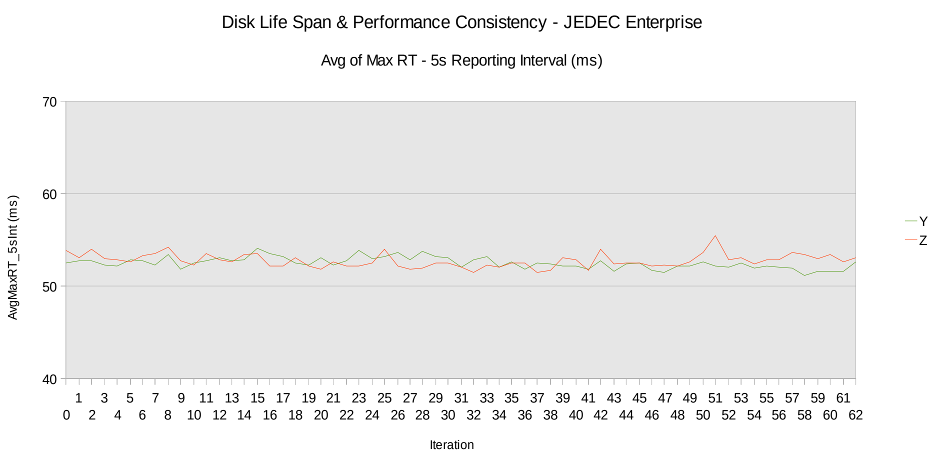
1. Sequential Write – 24 hours
    - 128K, Max IO rate, QD = 32, Incompressible data
    - 2m measurement intervals

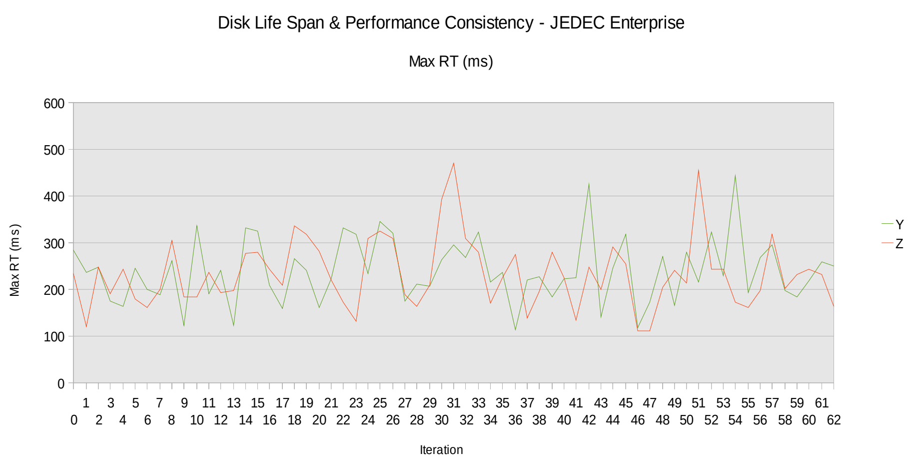
2. JEDEC Enterprise Workload – 1 hour
    - 3 Mixed RW random workloads
        - RW = 40/60%
        - Transfer size mix
            - 512B (4%)
            - 1KB (1%)
            - 1.5KB (1%)
            - 2KB (1%)
            - 2.5KB (1%)
            - 3KB (1%)
            - 3.5KB (1%)
            - 4KB (67%)
            - 8KB (10%)
            - 16KB (7%)
            - 32KB (3%)
            - 64KB (3%)
        - Max IO rate, QD = 32, Incompressible data
        - 5s measurement intervals
        - Workload mix:
            - #1 (50% overall workload skew, 5% drive range)
            - #2 (30% overall workload skew, 15% drive range)
            - #3 (20% overall workload skew, 80% drive range)

1.8" SATA – Disk Life Span / Performance Consistency Experiment #2 Results









- Although throughput and avg. response improve, max. latency peaks increasingly evident over 62 hr. test (approx. 1500 hrs. seq. write incl.)

Disk Life Span / Performance Consistency Experiment #3

<u>Testing Iteration</u>

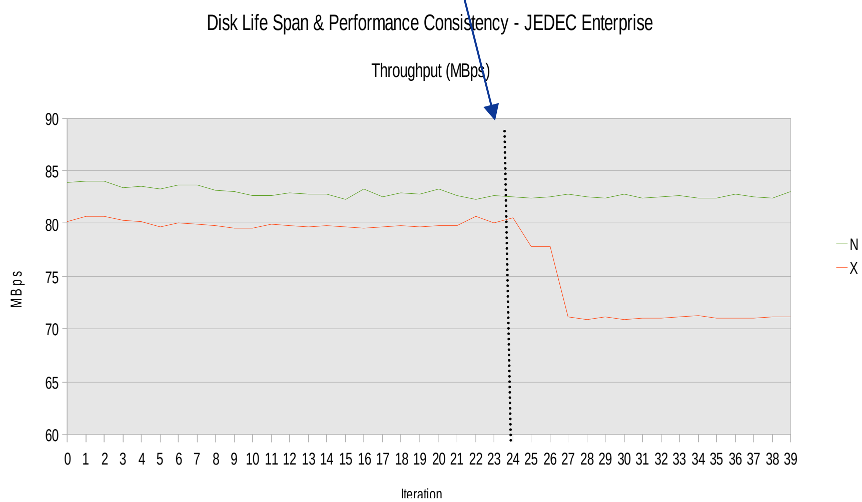
1. Sequential Write – 24 hours
    - 128K, Max IO rate, QD = 32, Incompressible data
    - 2m measurement intervals

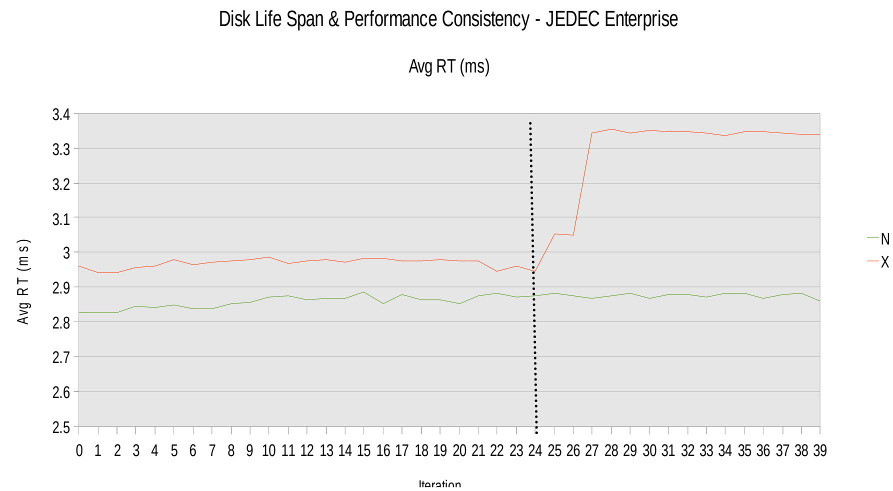
2. JEDEC Enterprise Workload – 1 hour
    - 3 Mixed RW random workloads
        - RW = 40/60%
        - Transfer size mix
            - 512B (4%)
            - 1KB (1%)
            - 1.5KB (1%)
            - 2KB (1%)
            - 2.5KB (1%)
            - 3KB (1%)
            - 3.5KB (1%)
            - 4KB (67%)
            - 8KB (10%)
            - 16KB (7%)
            - 32KB (3%)
            - 64KB (3%)
        - Max IO rate, QD = 32, Incompressible data
        - 5s measurement intervals
        - Workload mix:
            - #1 (50% overall workload skew, 5% drive range)
            - #2 (30% overall workload skew, 15% drive range)
            - #3 (20% overall workload skew, 80% drive range)

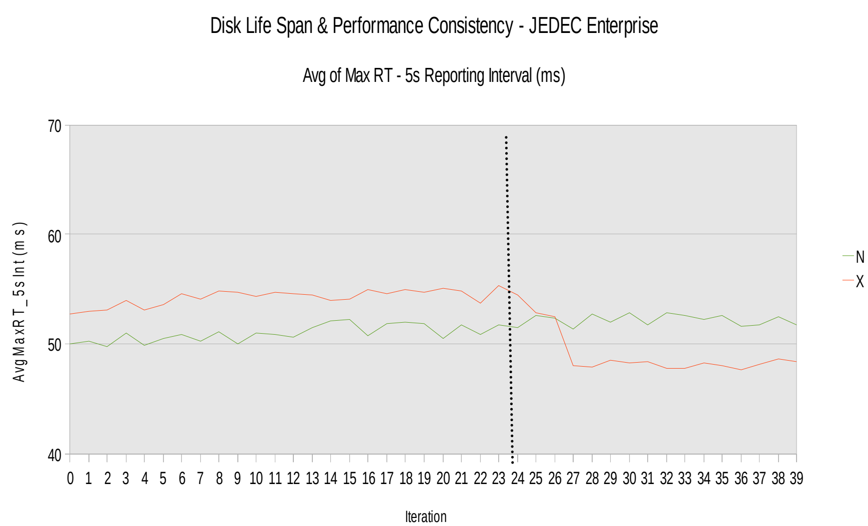
N – performance throttling disabled
X – performance throttling enabled

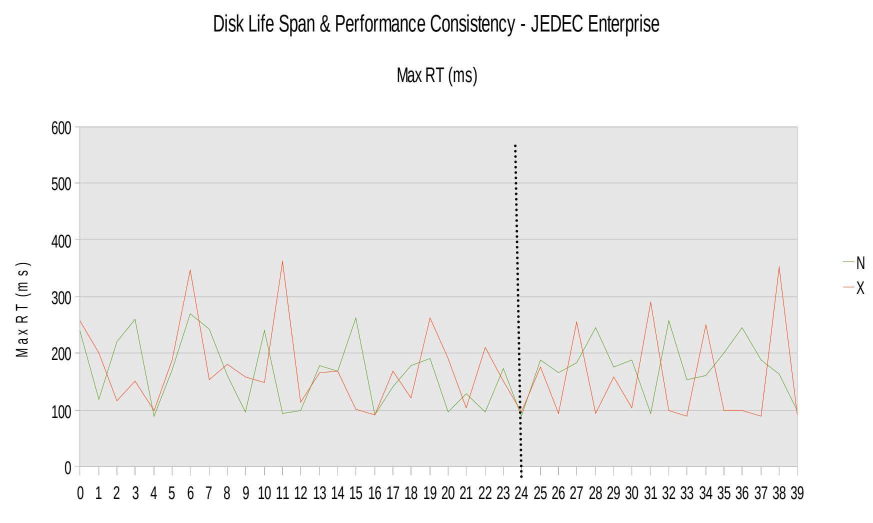
1.8” SATA – Disk Life Span / Performance Consistency Experiment #3 Results

• *Performance throttling engaged*









- User must be aware of background lifetime throttling mechanisms that can surface and impact performance
  - Although throughput/average latency degrade with throttling, avg. max. latency (and it’s standard deviation) improves

SSD Architecture for Consistent Enterprise Performance – Next Steps

• Continue to monitor ongoing experiments for inconsistent performance / long latency events and trends

• Pursue root cause investigation of long latencies to determine how these events can be better managed in SSD background operations

• Perform additional experiments to better evaluate aging SSD and end-of-life scenarios to characterize likely performance consistency impacts

• Initiate SSD performance consistency characterization within RAID configurations to better analyze read/write tradeoff behaviors that likely exist within a real system environment

SSD Architecture for Consistent Enterprise Performance – Summary

- The traditional approach for managing background operations of enterprise SSDs is no longer acceptable
  - Clients beginning to pursue 24/7/365 SSD-driven operations

- Background operations should be performed continuously, and require a consistent level of throughput, or always done in low priority (never consuming an appreciable amount of host bandwidth)
  - Key examples are – relocation algorithms due to read disturbs, garbage collection/ free space reclamation and ECC correction for aging SSDs

- Extensive characterization likely required to appropriately evaluate SSD performance consistency
  - Long duration testing and consideration of various conditions/scenarios throughout SSD life

- SSD throughput and average latency are not always good indicators of consistent SSD performance
  - Maximum and average maximum (per interval) latencies are key parameters to evaluate

- Background lifetime / performance throttling mechanisms will likely impact SSD performance consistency and must be thoroughly characterized