SNIA

Education

The Why and How of SSD Performance Benchmarking

Esther Spanjer, SMART Modular
Easen Ho, Calypso Systems

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

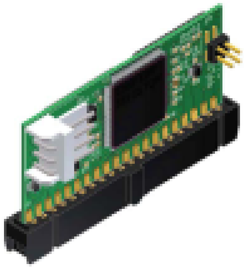
NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Abstract

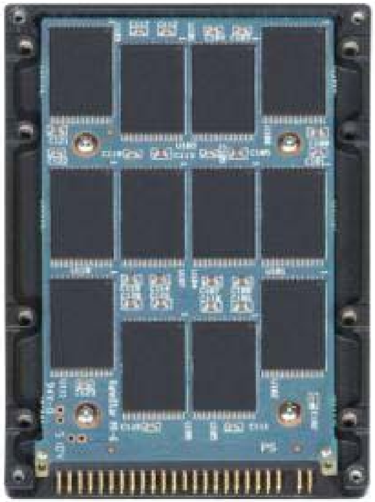
- A variety of parameters can influence the performance behavior of a solid state drive: current and previous workloads, fragmentation, block size, read/write mix, and queue depth to name a few
- SNIA's Performance Test Specification allows for performance benchmarking that result in repeatable and consistent test results
- This presentation will provide an overview of the SNIA SSD Performance Test Specification for both client and enterprise SSDs

Definition of SSS

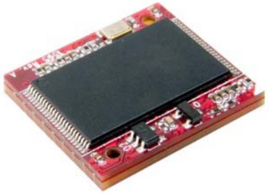
SSS = Solid State Storage



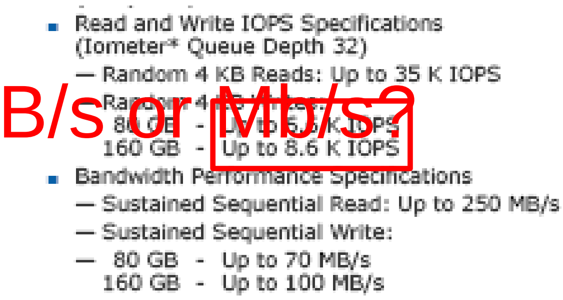
Traditional hard disk drive



Solid state hard drive

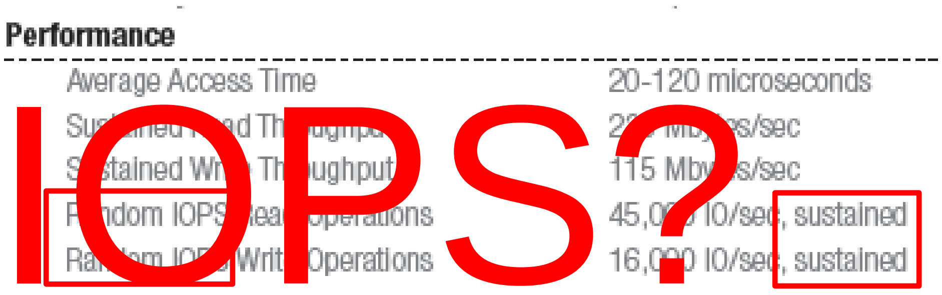


The Performance Landscape



- Read and Write IOPS Specifications (Iometer* Queue Depth 32)
  - Random 4 KB Reads: Up to 35 K IOPS
  - Random 4 KB Writes:
    80 GB - Up to 6.6 K IOPS
    160 GB - Up to 8.6 K IOPS
- Bandwidth Performance Specifications
  - Sustained Sequential Read: Up to 250 MB/s
  - Sustained Sequential Write:
  - 80 GB - Up to 70 MB/s
    160 GB - Up to 100 MB/s

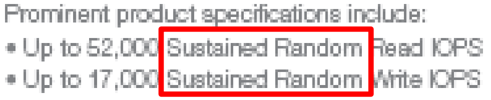
MB/s or Mb/s?



Performance

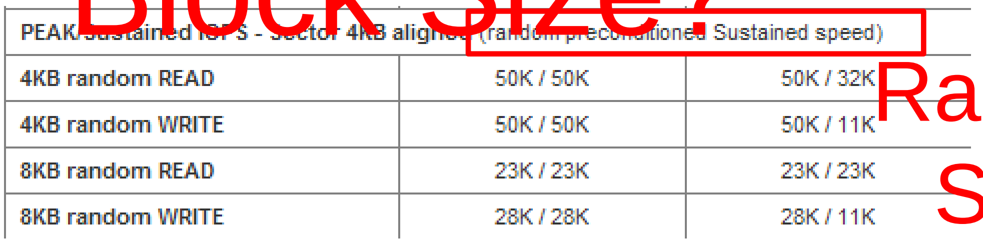
Average Access Time	20-120 microseconds
Sustained Read Throughput	2[illegible] Mbytes/sec
Sustained Write Throughput	115 Mbytes/sec
Random IOPS Read Operations	45,000 IO/sec, sustained
Random IOPS Write Operations	16,000 IO/sec, sustained

IOPS?



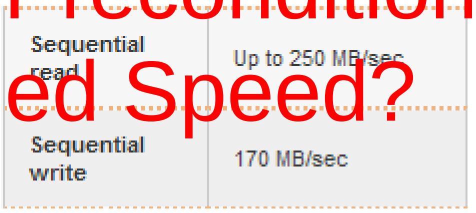
Prominent product specifications include:
- Up to 52,000 Sustained Random Read IOPS
- Up to 17,000 Sustained Random Write IOPS

Block Size?



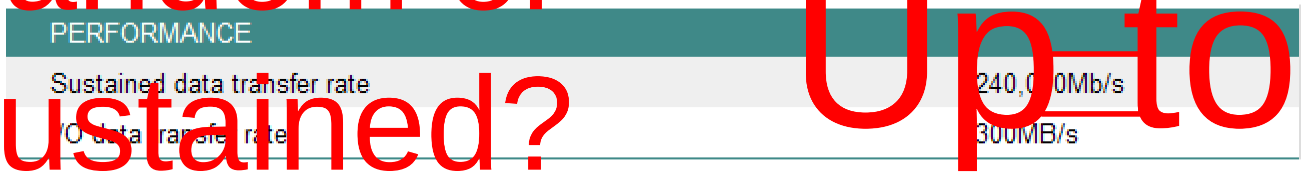
PEAK sustained IOPS - Sector 4KB aligned (random preconditioned Sustained speed)		
4KB random READ	50K / 50K	50K / 32K
4KB random WRITE	50K / 50K	50K / 11K
8KB random READ	23K / 23K	23K / 23K
8KB random WRITE	28K / 28K	28K / 11K

Random Precondition Sustained Speed?



Sequential read	Up to 250 MB/sec
Sequential write	170 MB/sec

Random or Sustained?



PERFORMANCE	
Sustained data transfer rate	240,000Mb/s
I/O data transfer rate	300MB/s

Up to?

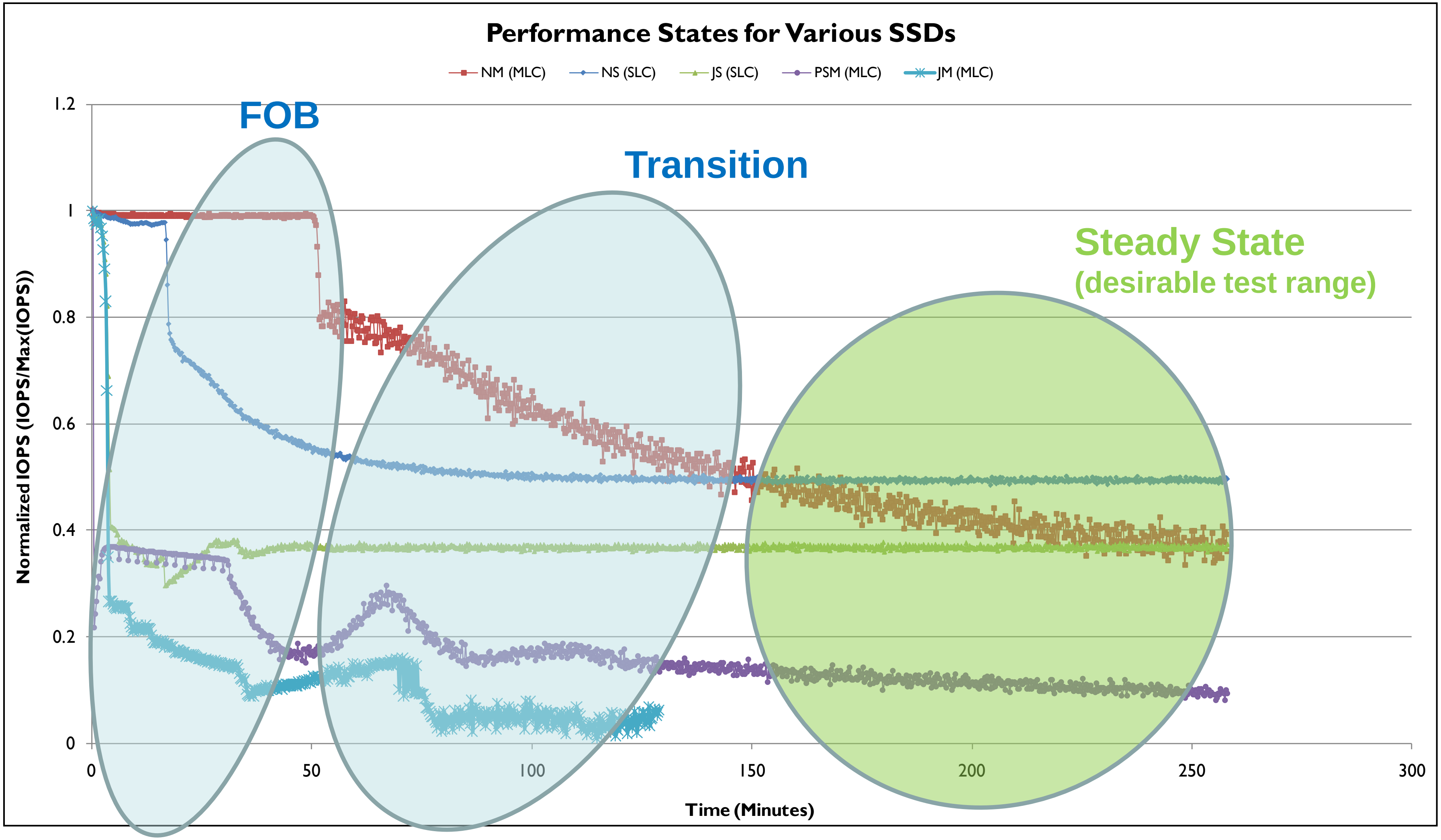
Variables influencing Performance

- Platform
  - Test Hardware (CPU, interface, chipset, etc)
  - Software (OS, drivers)
- SSS Device Architecture
  - Flash geometry, cache, flash management algorithm, etc

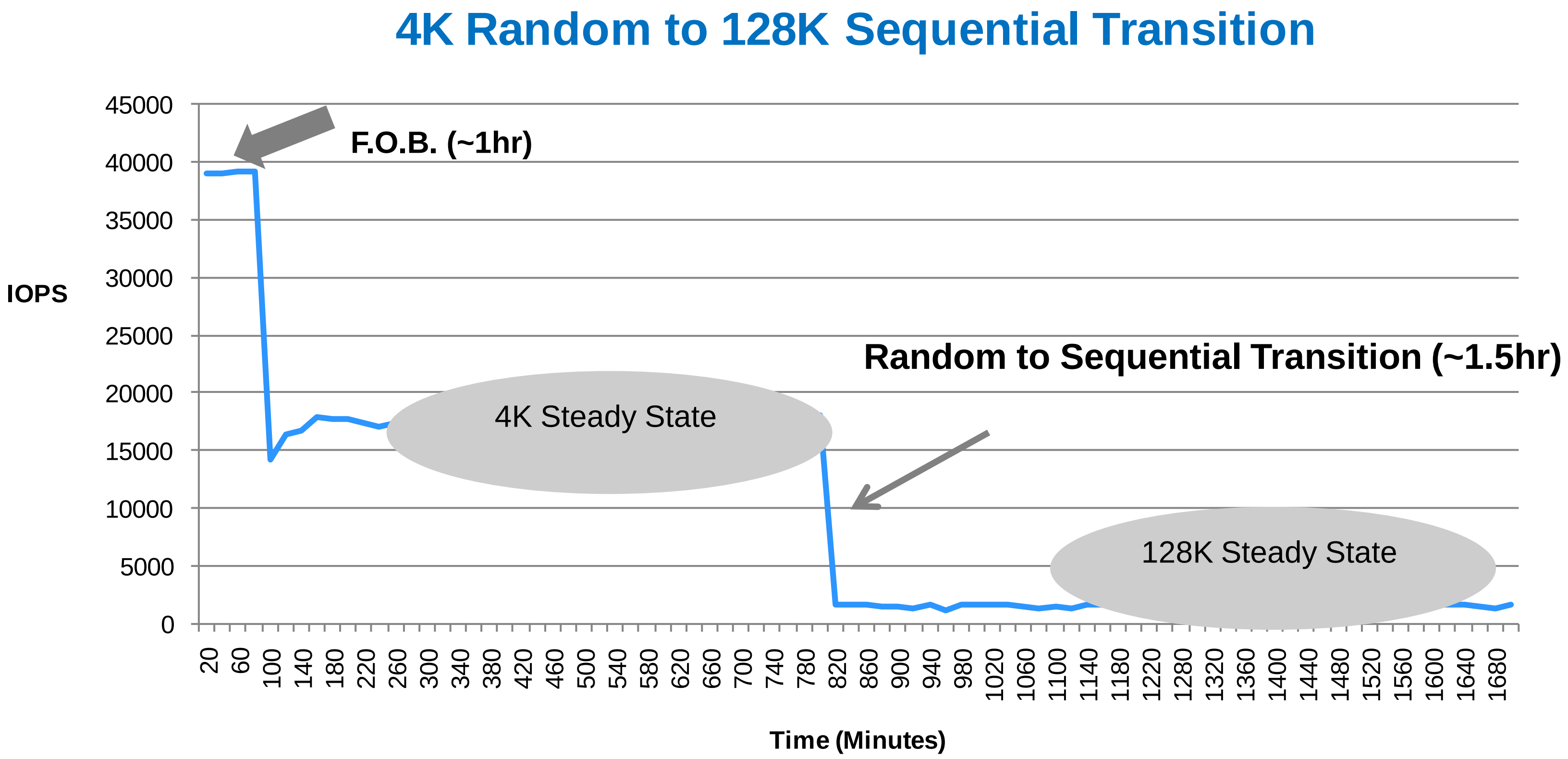
Variables influencing Performance

- Platform
  - Test Hardware (CPU, interface, chipset, etc)
  - Software (OS, drivers)
- SSS Device Architecture
  - Flash geometry, cache, flash management algorithm, etc
- Workload
  1. Write history & preconditioning: State of device before testing

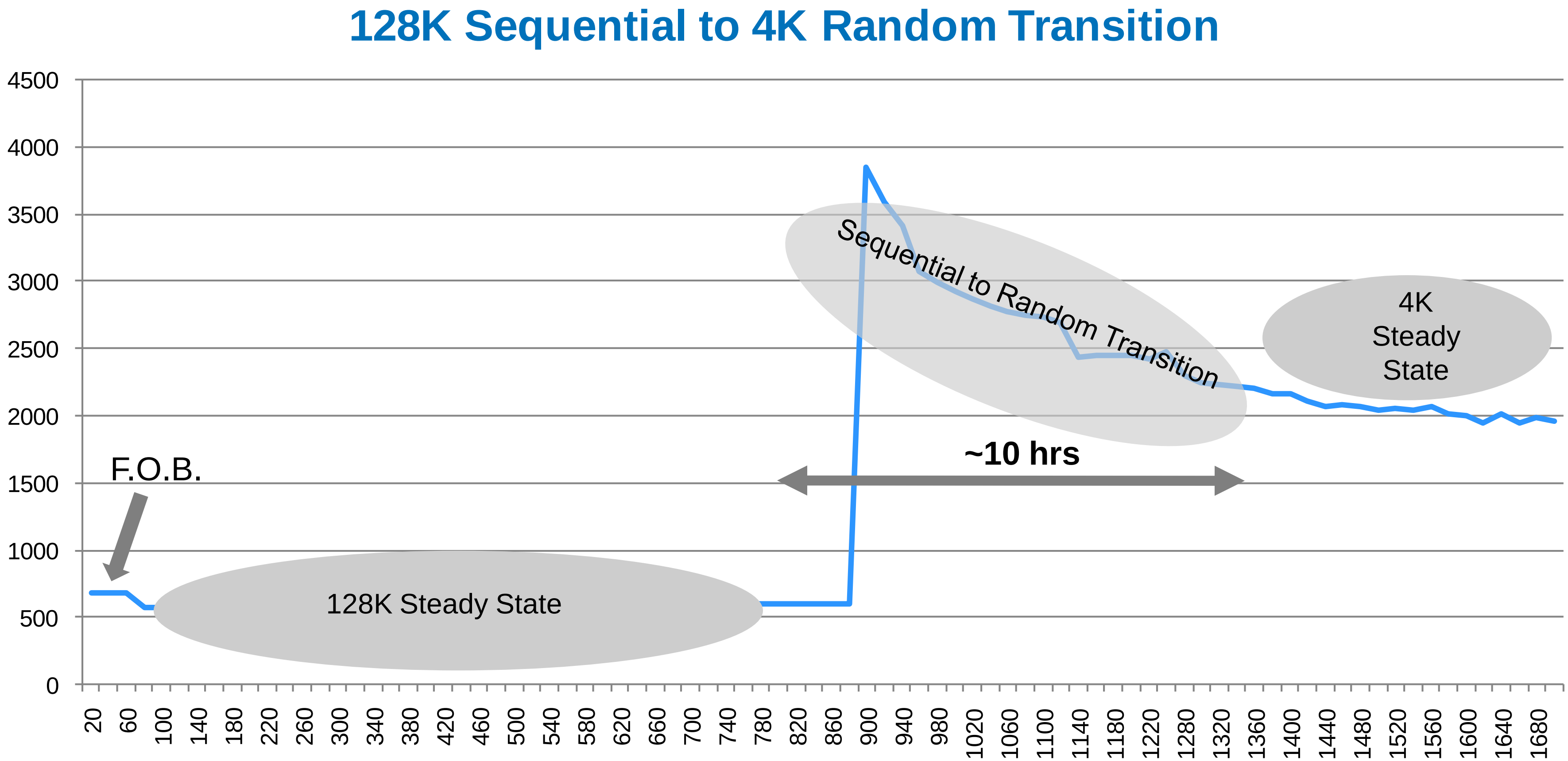
The need for Preconditioning



Write History - 1



Write History - 2

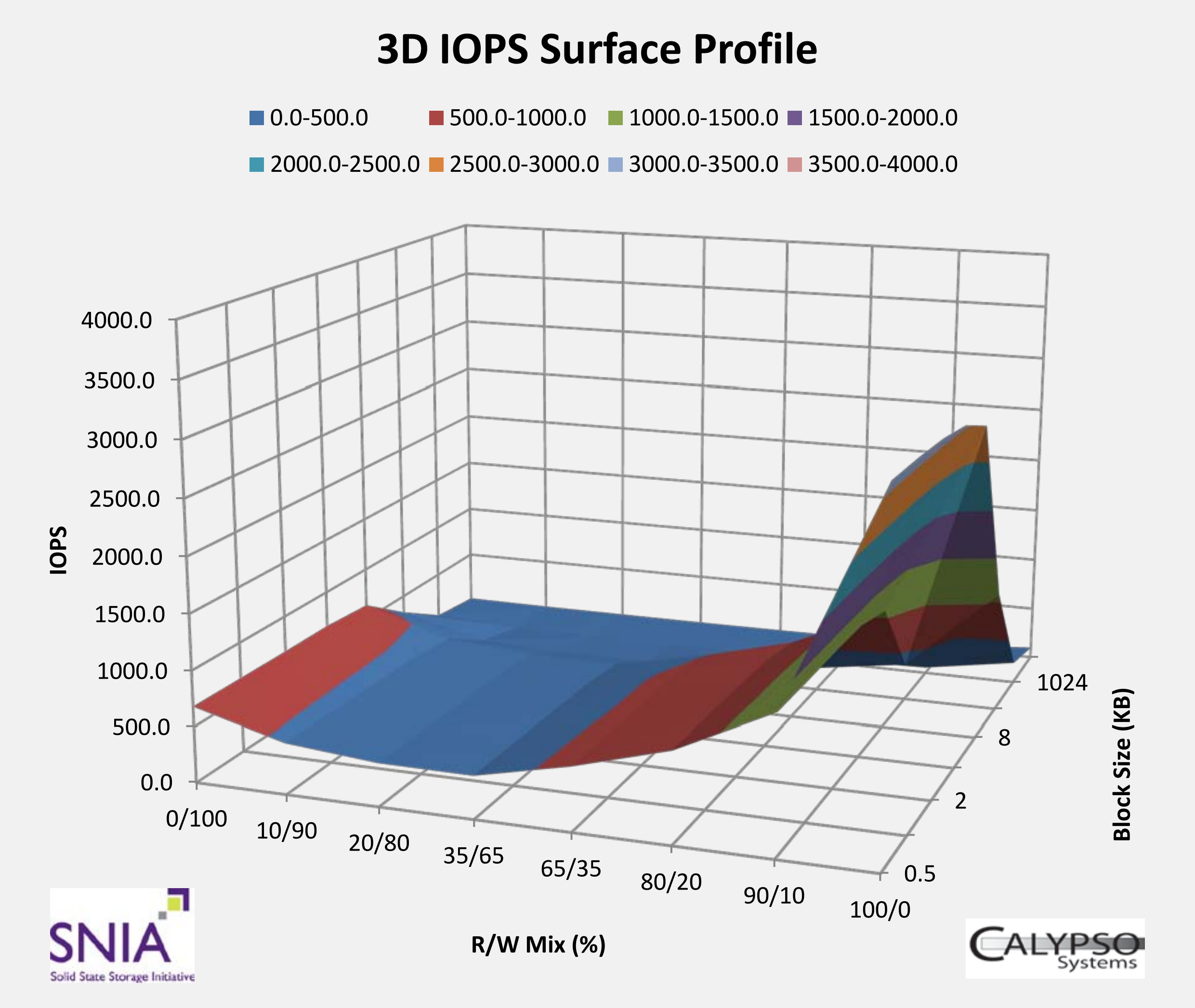


Variables influencing Performance

- Platform
  - Test Hardware (CPU, interface, chipset, etc)
  - Software (OS, drivers)
- SSS Device Architecture
  - Flash geometry, cache, flash management algorithm, etc

- Workload
  1. Write history & preconditioning: State of device before testing
  2. Workload pattern: Read/write mix, transfer size, sequential/random

Workload Pattern



Performance depends on

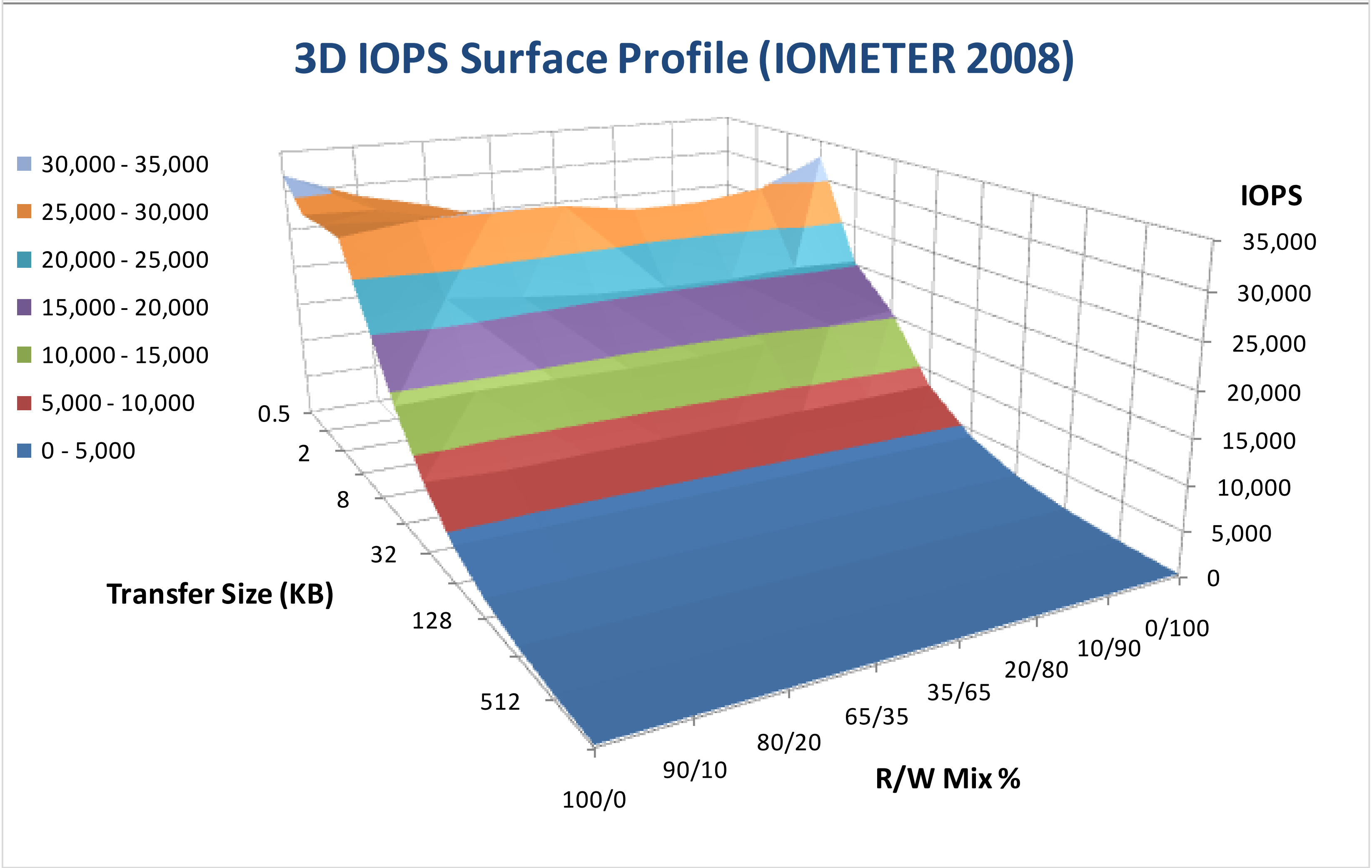
- Read/Write Mix
- Block Size
- Queue Depth (not shown)

Variables influencing Performance

- Platform
  - Test Hardware (CPU, interface, chipset, etc)
  - Software (OS, drivers)
- SSS Device Architecture
  - Flash geometry, cache, flash management algorithm, etc

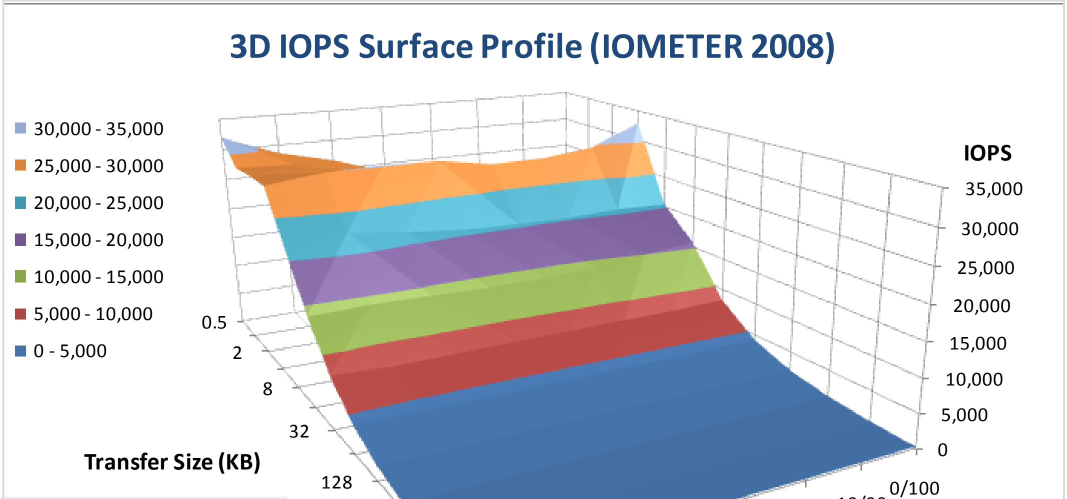
- Workload
  1. Write history & preconditioning: State of device before testing
  2. Workload pattern: Read/write mix, transfer size, sequential/random
  3. Data Pattern: The actual bits in the data payload written to the device

Dependency on data content - I

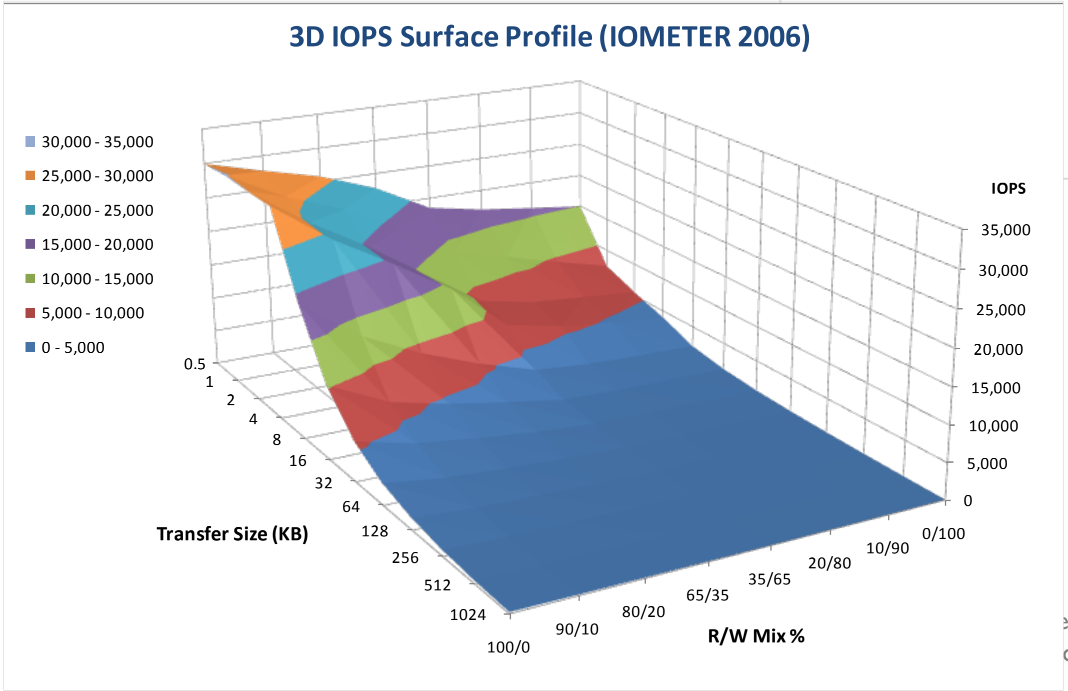


Dependency on data content - 2

IOMeter 2008
Low Entropy Data Content



3D IOPS Surface Profile (IOMETER 2008)



3D IOPS Surface Profile (IOMETER 2006)

IOMeter 2006
High Entropy Data Content

The Need for Industry Standardization!

- SNIA Technical Working Group (TWG)
  - Created in early 2009
- Specification for tests procedures to enable comparative testing of SSS performance
  - **Agnostic** – Does not favor any one technology
  - **Relevant & Repeatable** – Meaningful to end users
  - **Practical** – Complete with reasonable time and effort
- Performance Test Spec (PTS) 1.0 Client Released
- PTS 1.0 Enterprise Released
  - PTS 1.1 in progress, target release 4Q11

Benchmark Types

Synthetic

IOMeter, VDBench

- Test specific scenario (QD, block size, transfer rate)
- Good to determine corner case behavior

Application-based

SysMark, PCMark

- Test performance of specific application (ignores QD, transfer size, etc.)
- Illustrates real world differences

Trace-based

Storage Bench

- Measures performance as drive is used (traces)
- Most valid when similar applications are run (no two user workloads are the same)

SNIA PTS focuses on synthetic based benchmark tools

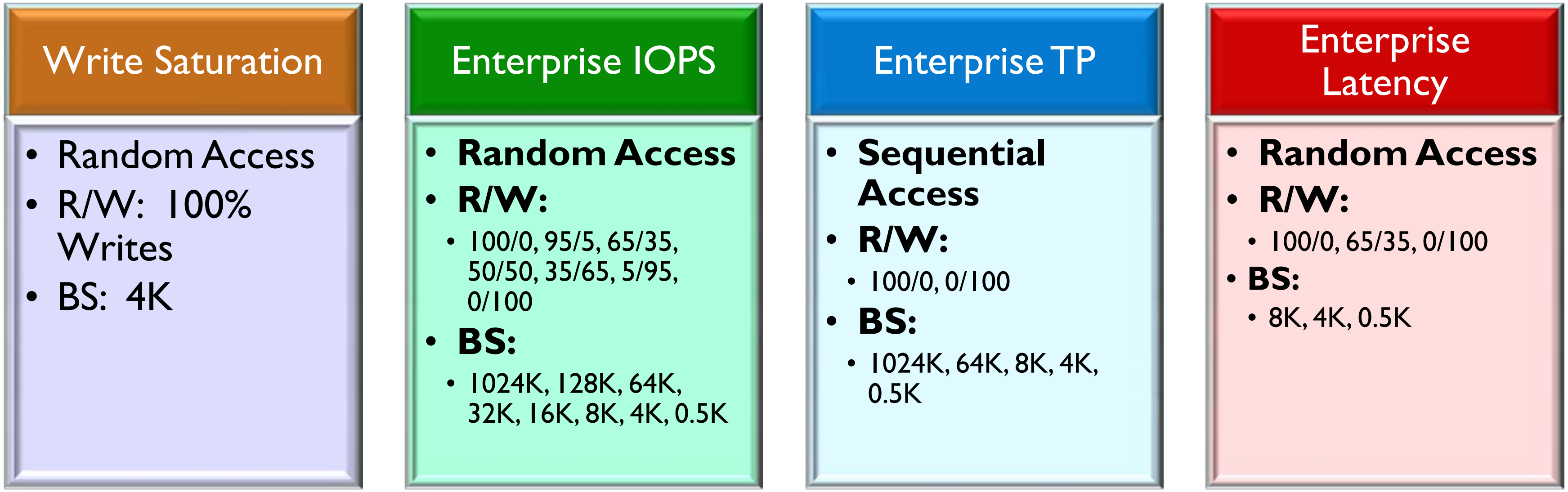
SSSI Reference Test Platform

- Intel S5520HC
- Single Intel W5580, 3.2GHz, Quad-core CPU
- 12GB, 1333MHz, ECC DDR3 RAM
- LSI 9212-4e4i 6Gb/s SAS HBA
- Intel ICH10R 3Gb/s SATA
- 8X Gen-II PCI-e

- CentOS 5.5
- Calypso RTP Backend V1.5
- Calypso Test Suite (CTS) V6.5

Tests Contained In Draft V1.0 Spec.

- The V1.0 Specification encompasses:
  - A suite of basic SSS performance tests



Write Saturation

- Random Access
- R/W: 100% Writes
- BS: 4K

Enterprise IOPS

- **Random Access**
- **R/W:**
  - 100/0, 95/5, 65/35, 50/50, 35/65, 5/95, 0/100
- **BS:**
  - 1024K, 128K, 64K, 32K, 16K, 8K, 4K, 0.5K

Enterprise TP

- **Sequential Access**
- **R/W:**
  - 100/0, 0/100
- **BS:**
  - 1024K, 64K, 8K, 4K, 0.5K

Enterprise Latency

- **Random Access**
- **R/W:**
  - 100/0, 65/35, 0/100
- **BS:**
  - 8K, 4K, 0.5K

- Preconditioning and Steady State requirements
- Standard test procedures
- Standard test reporting requirements

What Is NOT Covered In the Spec

- Application workload tests
- Matching to user workloads
- Energy efficiency
- Required test platform (HW/OS/Tools)
- Certification
- Device endurance, availability, data integrity

- Performance Test Specification v1.0 – Section 1.4

Basic Test Flow

1. Purge

- Security Erase, Sanitize, Format Unit, other proprietary methods where indicated

2. Set Conditions

- Set user selectable test parameters, such as Active Range, Data Pattern, Demand Intensity

3. Pre-Condition

- Workload independent (WIPC)
- Workload dependent (WDPC)

4. Run Until SS

- Reiterate loops until Steady State is reached, or run to a prescribed maximum number of loops

5. Collect Data

- Collect data from Steady State Measurement Window

6. Generate Reports

- Use standard report formats and include required and optional elements

Key Concepts Used in the Spec.

- A. Purge
- B. Pre-Condition
  - Workload independent
  - Workload dependent
- C. Active Range
  - Pre-conditioning
  - Test
- D. Steady State
  - Measurement window
  - Data excursion condition
  - Slope excursion condition

A: Purge

- As per the PTS V1.0 Specification, purge is defined as:

> *“ The process of returning an SSS device to a state in which subsequent writes execute, as closely as possible, as if the device had never been used and does not contain any valid data”*

- Example implementation includes: ATA Security Erase, Sanitize, SCSI Format Unit

B: Pre-Conditioning

- Pre-Conditioning is a key requirement in getting repeatable, representative results
- Goal is to put drive into "Steady State", using:
  - Workload independent *– PTS v1.0 Section 3.3*
    - Use a prescribed workload unrelated to the test loop
    - Write 2X user capacity using SEQ/128KiB blocks
  - Workload dependent *– PTS v1.0 Section 3.3*
    - Run test workload itself as pre-conditioning (self pre-conditioning)

C: Active Range

- As per the PTS V1.0 Specification, Active Range is defined as:

"… ActiveRange is the range of LBA's that may be accessed by the preconditioning and/or test code..."

- They are normally defined as % of the maximum LBA available to the user
- Note Pre-conditioning and Test can have different Active Ranges

D: Steady State Definition

- Premise is that reported data should be take only AFTER the test loop results shows the drive has reached and maintained "Steady State"
- The Measurement Window is the interval, measured in Rounds, when the test results have entered and maintained Steady State for 5 Rounds

D: Steady State Definition

- Steady State is reached only if BOTH of the following conditions are satisfied (assuming "*y*" is the variable being tracked):

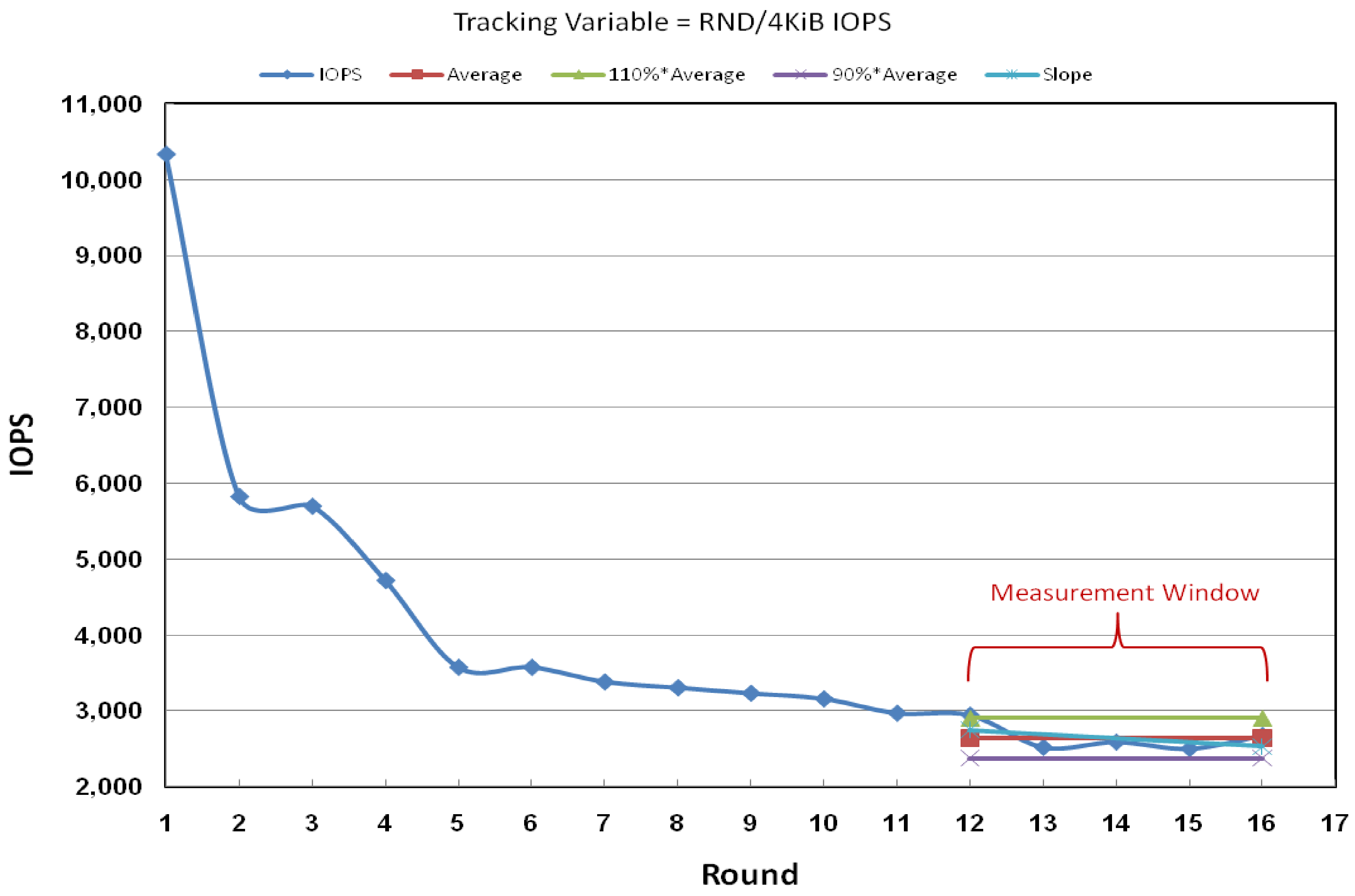
  1. Variation of *y* within the Measurement Windows is within 20% of the Average

     " Max(y)-Min(y) within the Measurement Window is no more than 20% of the Ave(y) within the Measurement Window; and "

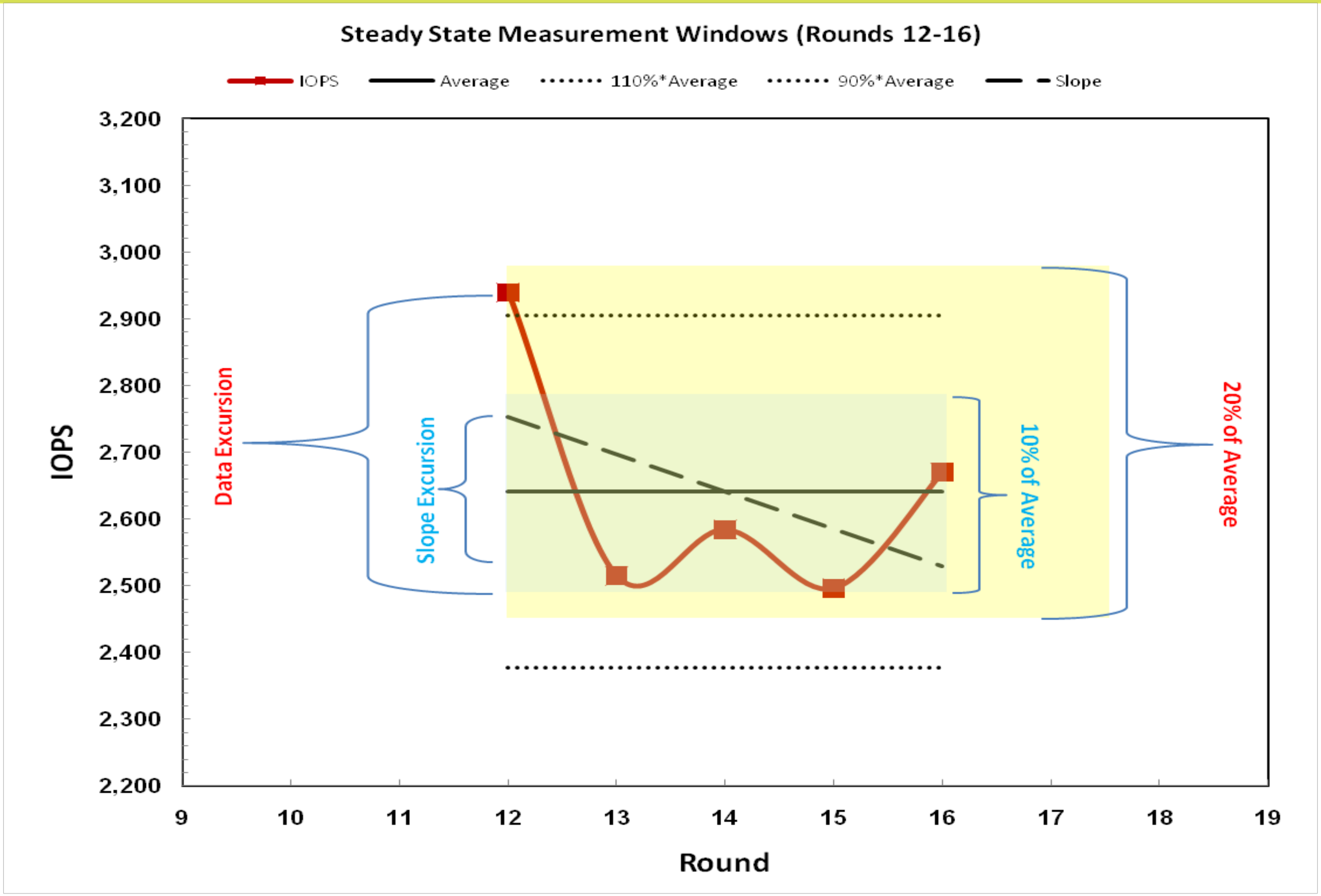
  2. Trending of *y* within the Measurement Windows is within 10% of the Average

     " [Max(y) as defined by the linear curve fit of the data within the Measurement Window] – [Min(y) as defined by the best linear curve fit of the data within the Measurement Window] is within 10% of Ave(y) within the Measurement Window. "

D: SS Measurement Window



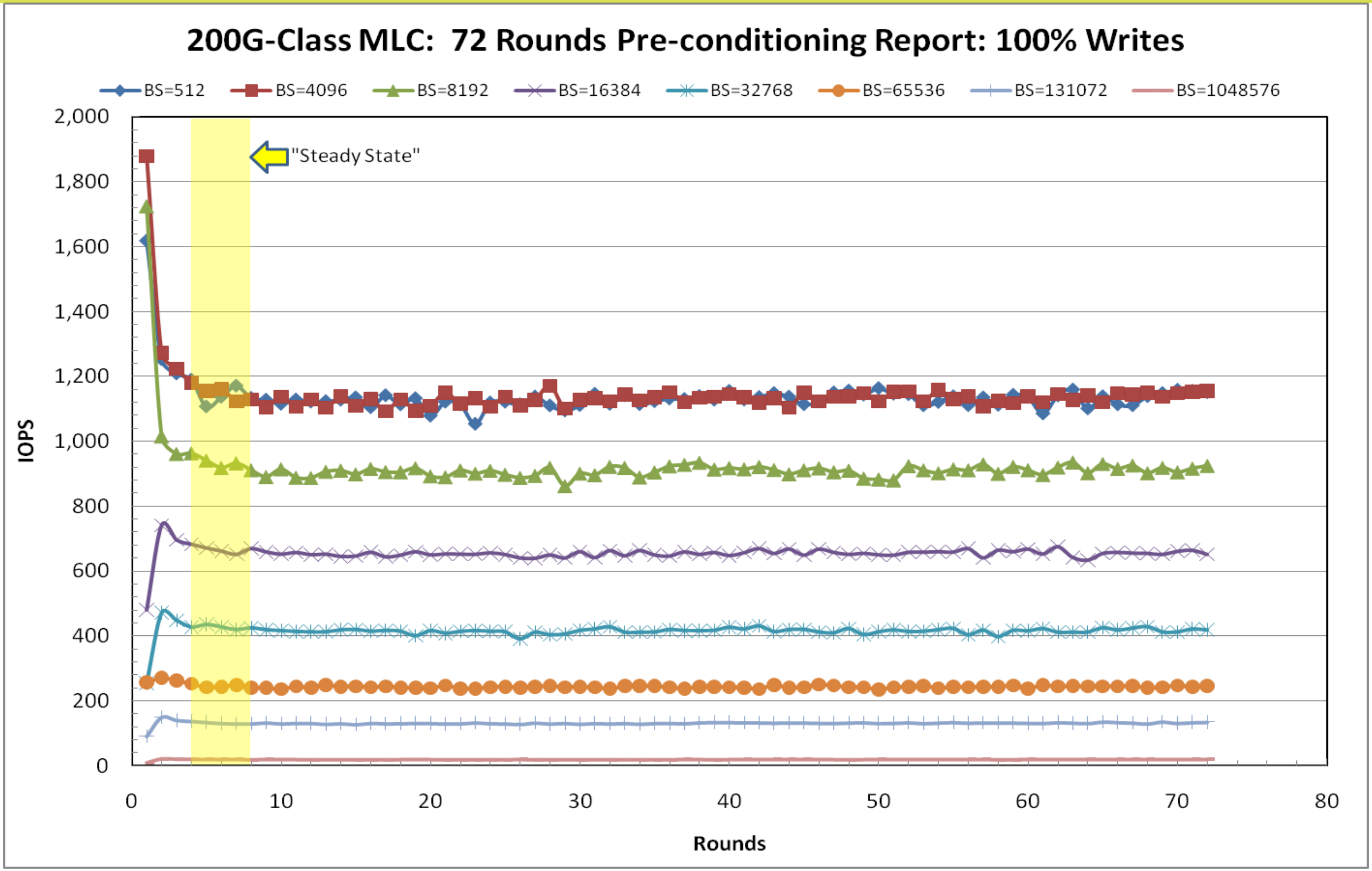
D: SS Measurement Window



D: Steady State Definition

- Compare
  - [Data Excursion] with [20% of Average]
  - [Slope Excursion] with [10% of Average]
- Note
  - This method is slightly more tolerant than +10% and – 10% data excursion method and +5% and – 5% slope excursion method

D: How Good is the Steady State



Workload Schematics

Education SNIA

Write Saturation

- Random Access
- R/W: 100% Writes
- BS: 4K

Enterprise IOPS

- **Random Access**
- **R/W:**
  - 100/0, 95/5, 65/35, 50/50, 35/65, 5/95, 0/100
- **BS:**
  - 1024K, 128K, 64K, 32K, 16K, 8K, 4K, 0.5K

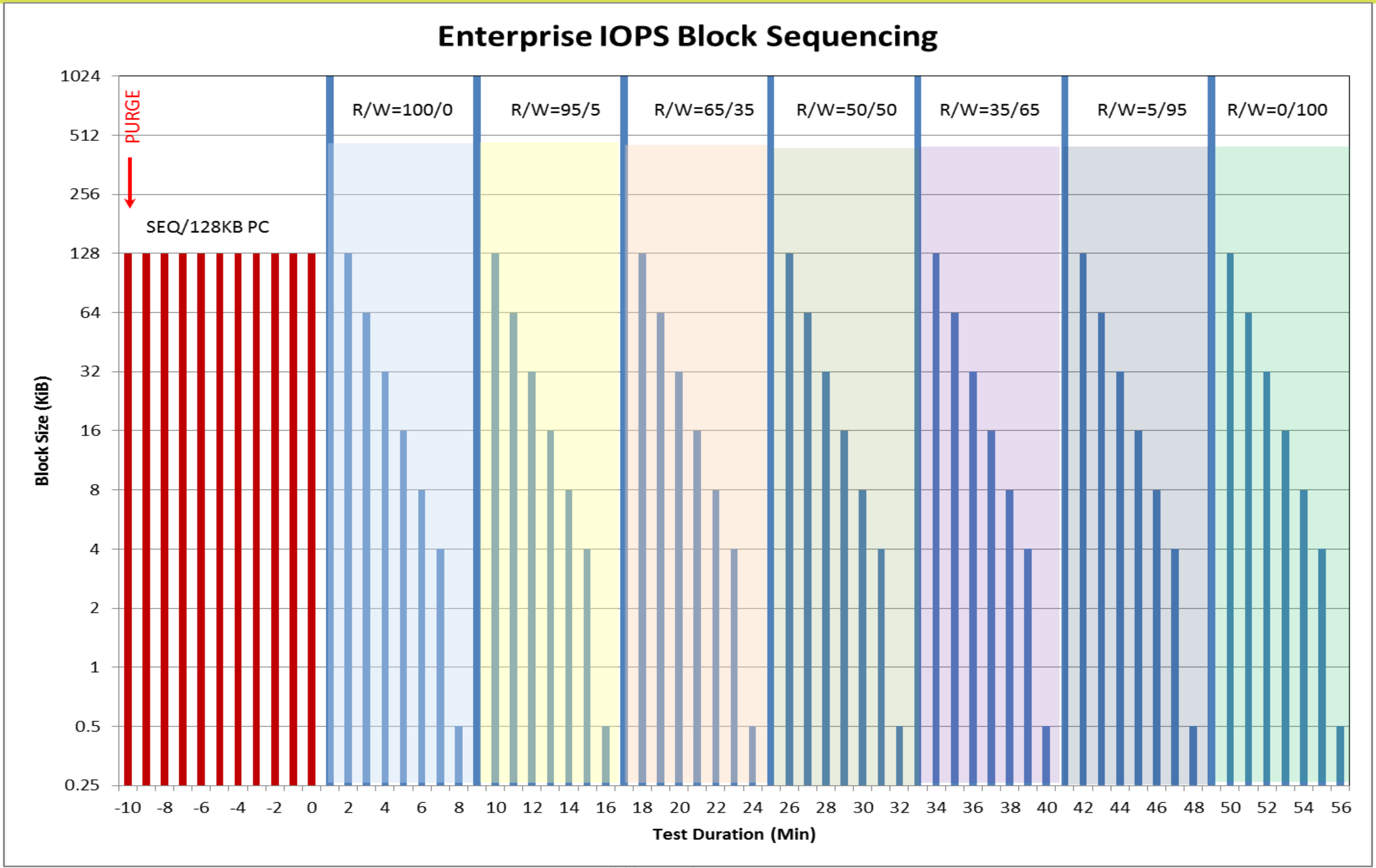
Enterprise TP

- **Sequential Access**
- **R/W:**
  - 100/0, 0/100
- **BS:**
  - 1024K, 64K, 8K, 4K, 0.5K

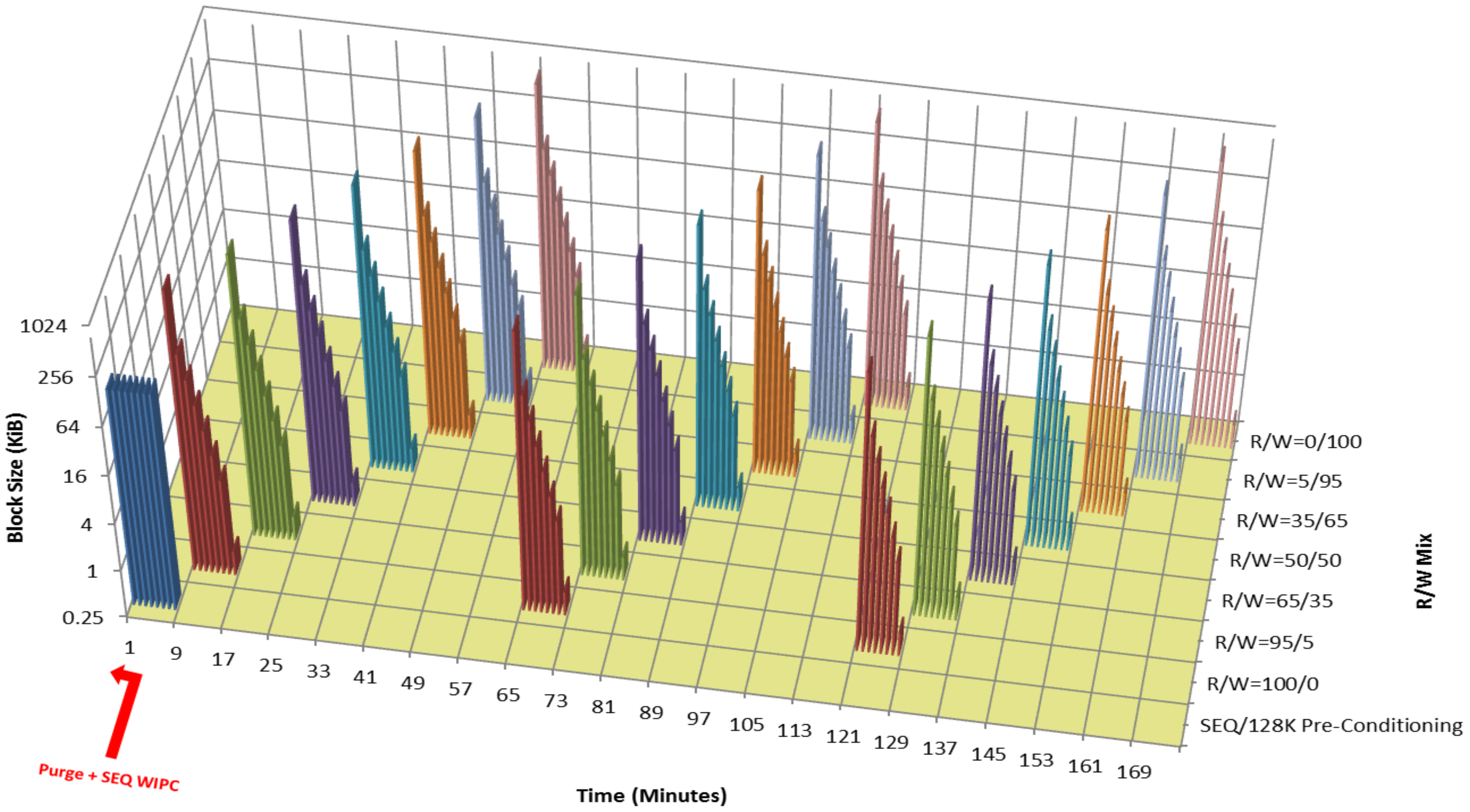
Enterprise Latency

- **Random Access**
- **R/W:**
  - 100/0, 65/35, 0/100
- **BS:**
  - 8K, 4K, 0.5K

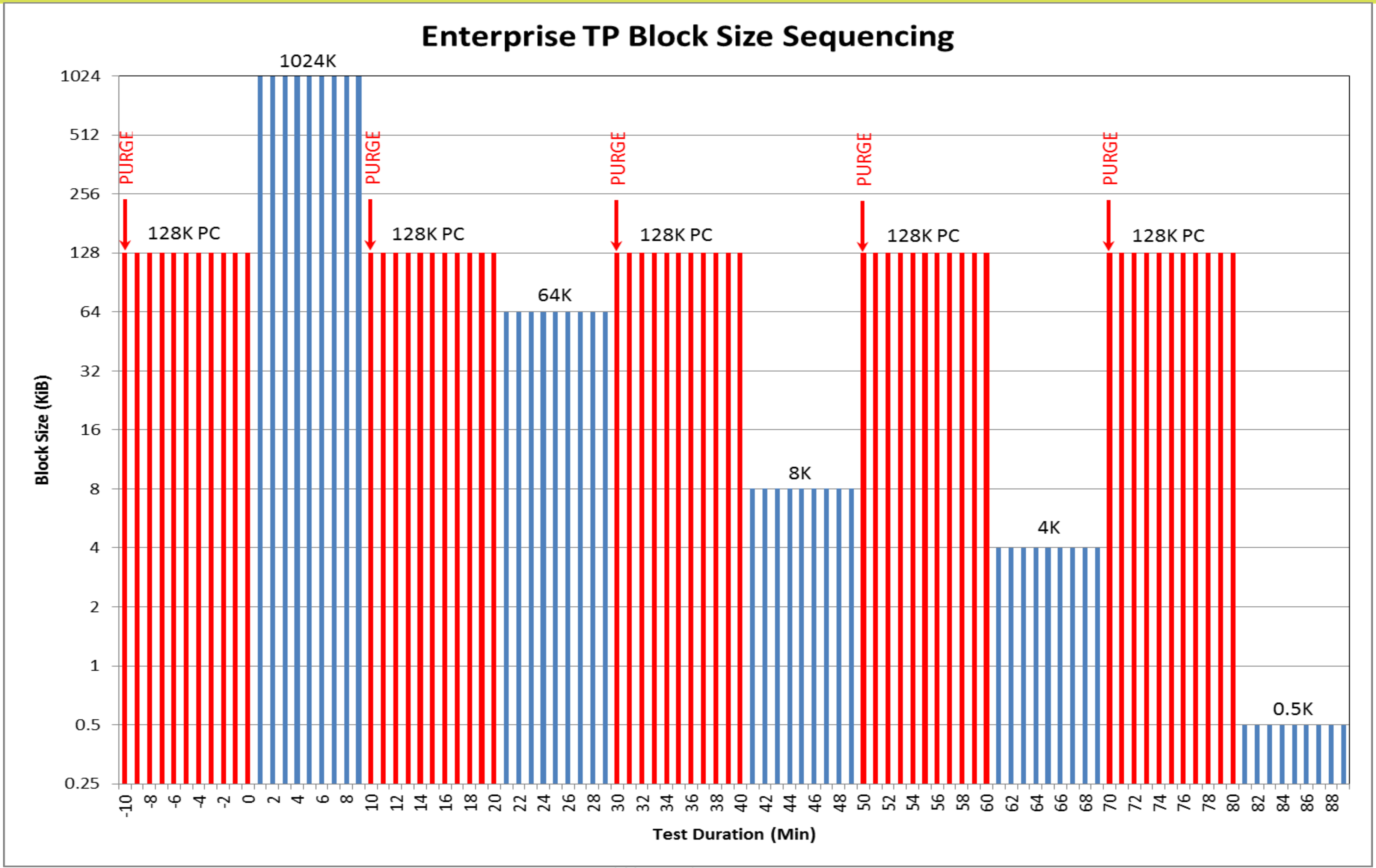
IOPS RW/BS Sequence



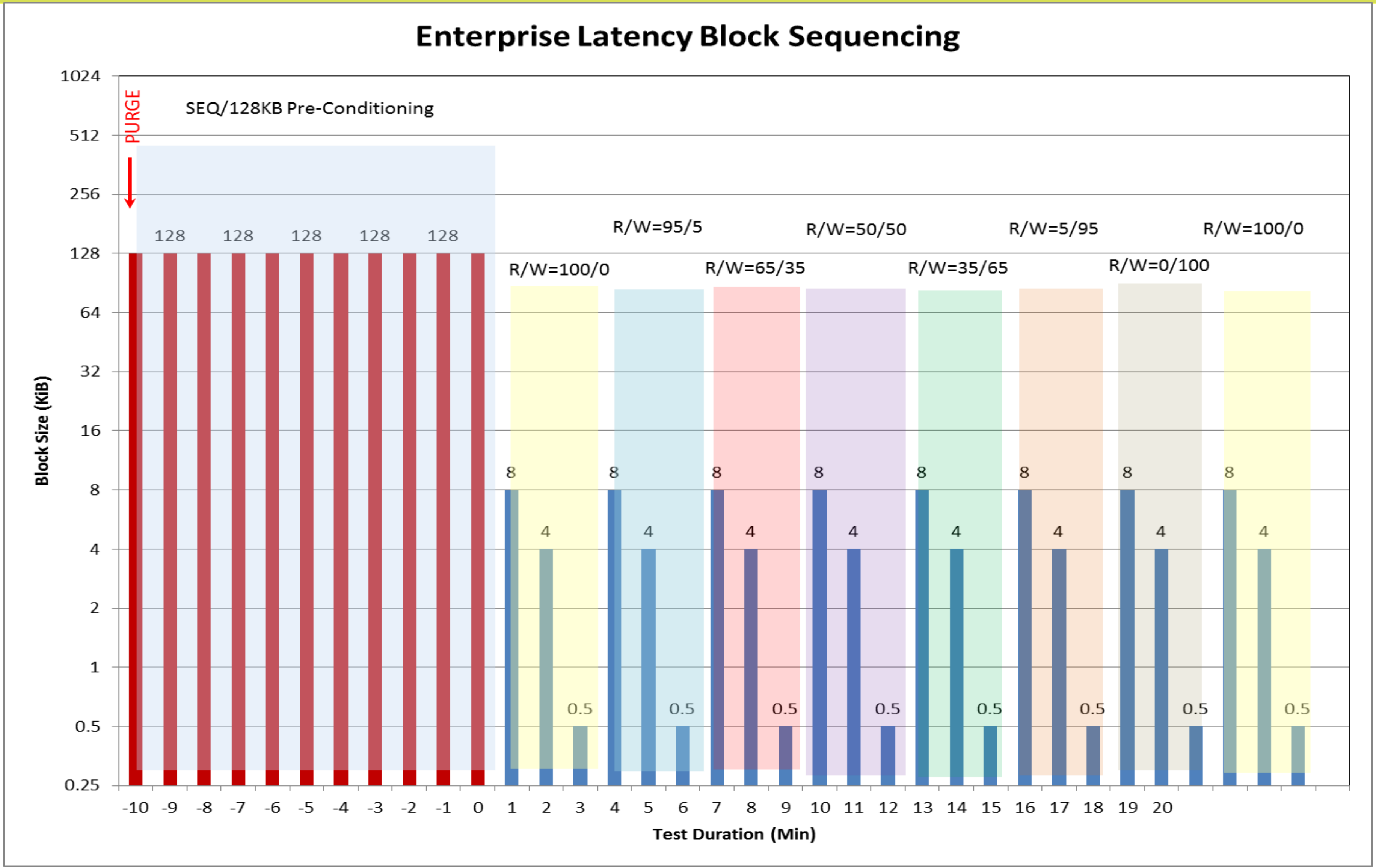
Enterprise IOPS RW/BS Sequence



TP RW/BS Sequence



Enterprise Latency RW/BS Sequence



Example: Enterprise IOPS

- DUT:
  - 100GB-Class Enterprise SLC drive
- Test Parameters:
  - Active Range = [0,100%]
  - Thread Count=2
  - Queue Depth (Outstanding IO/Thread)=16
  - DP=RND

Enterprise IOPS Draft Formatted Report, 1/6

Enterprise IOPS Draft Formatted Report, 2/6

Enterprise IOPS Draft Formatted Report, 3/6

Enterprise IOPS Draft Formatted Report, 4/6

Enterprise IOPS Draft Formatted Report, 5/6

Enterprise IOPS Draft Formatted Report, 6/6

PTS Follow-On Work (PTS-E 1.1)

Idle Recovery	• See how the drive responds to host idle time amidst continuous access
Cross Stimulus Recovery	• See how drive handles switching between sustained access patterns
Demand Intensity	• See how drive responds to increasing host demands
Response Time Histogram	• Get detailed response time statistics during specific stimulus
IOPS/W	• Measures power efficiency of the device
Trace-Based Workloads	• Captures or uses captured workloads traces and provide a consistent way to playback such traces
Enterprise Composite Synthetic Workload	• Synthetic composite workload for Enterprise environments similar to JEDEC workload for endurance testing

Q&A / Feedback

- Please send any questions or comments on this presentation to SNIA: tracksolidstate@snia.org

Many thanks to the following individuals for their contributions to this tutorial.

- SNIA Education Committee

Eden Kim
Easen Ho
Esther Spanjer