Flash Memory SUMMIT

SNIA Solid State Storage Performance Test Specification (PTS)

Session 202-C

Wednesday August 14, 2013

9:50 – 10:50 am

Agenda

- Part 1: SNIA PTS Introduction
  – Eden Kim (9:50 – 10:00 am)
- Part 2: PTS 1.1 – Basic Tests
  - Eden Kim (10:00 – 10:20 am)
- Part 3: PTS-E 1.1(e) *Draft* – Advanced Tests
  - Easen Ho (10:20 – 10:40 am)
- Questions
  •(10:40 – 10:50 am)

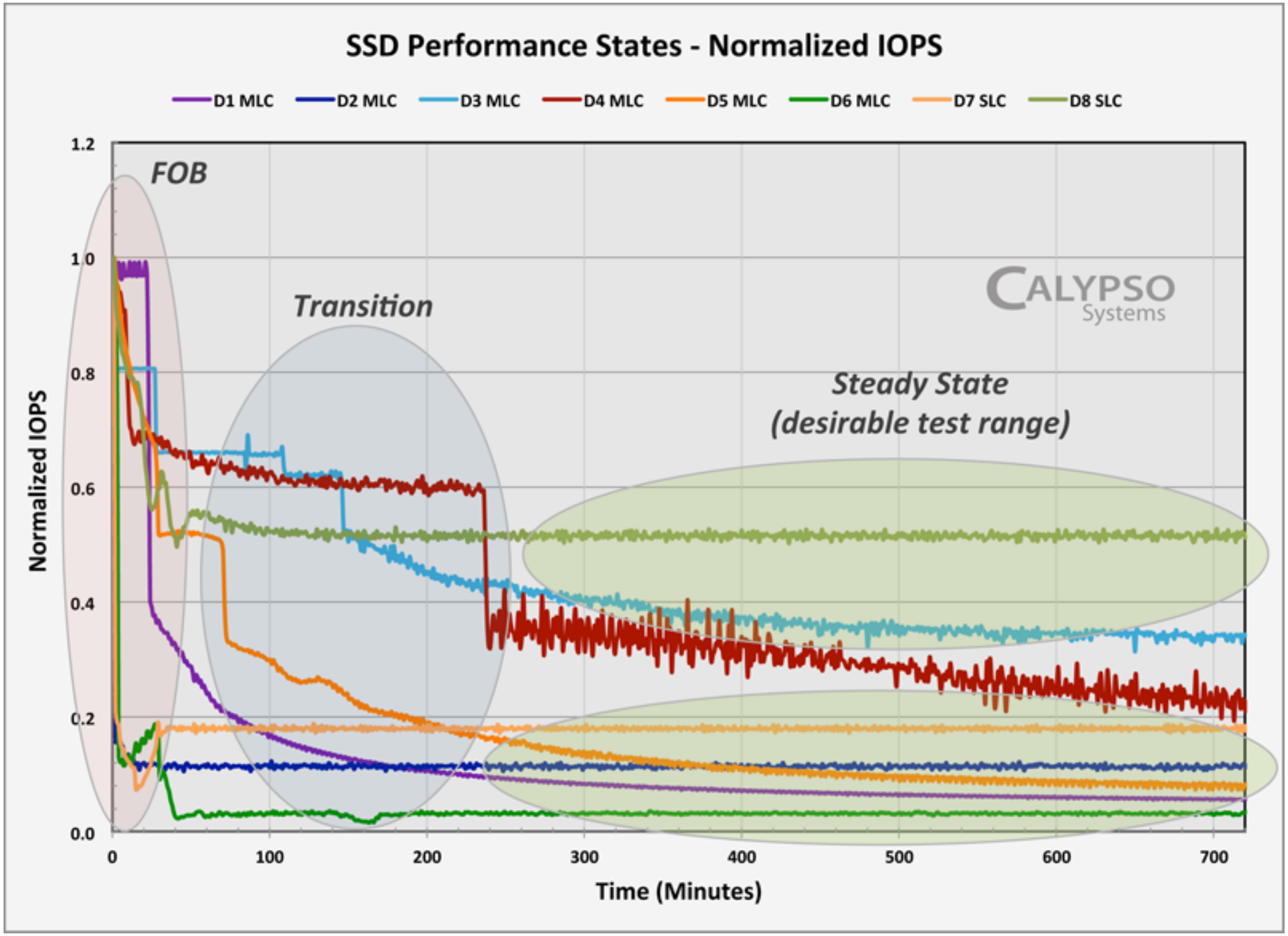
Part 1: SNIA PTS - Introduction

- NAND Flash Characteristics – Performance changes over time

- Factors Affecting Performance Test

- PTS Standardized Performance Test Methodologies

Performance Changes over Time



Factors Affecting Performance

- Write History
- Parameter Settings
- Workloads

Why do we care?

- Benchmarking – accuracy, repeatability
- Validation – properly characterize performance
- Marketing – “up to / sustained Performance”
- Isolate SSD for test - not the system or file system cache
- Workloads – accurately emulate application specific workloads

Factors Affecting Testing

- Hardware Platform – anything in data path
- Operating System – file system, cache, drivers
- Test Software – stimulus generation & measurement
- Test Stimulus - Access Patterns applied to DUT
- Set-up Conditions – parameter settings

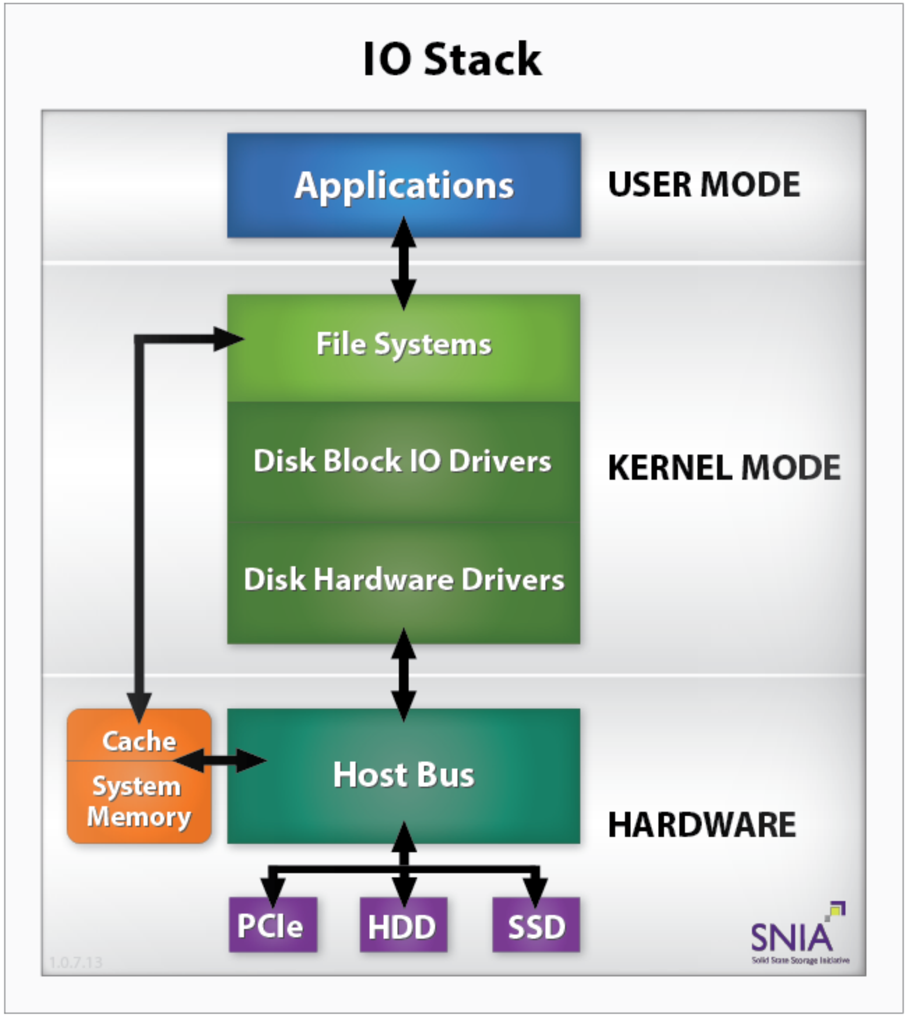
SNIA Reference Test Platform

- Standardized Hardware Platform
- Specified Test Software Capabilities
- Common PTS test Methodologies
- Device Level Test

- SSSI Defines PTS Reference Test Platforms
  - SSSI TechDev Committee: RTP 3.0
    - Selection of Gen 3 motherboard, cpu & RAM
    - Qualification of 12Gb/s SAS HBA cards, SFF 8639 HBA cards
  - SSS PTS:
    - Listing of RTP 3.0 in Annex A

IO Stack Affects Access Patterns

Where the measurement is taken makes a difference



How the IO Stack Affects IOs

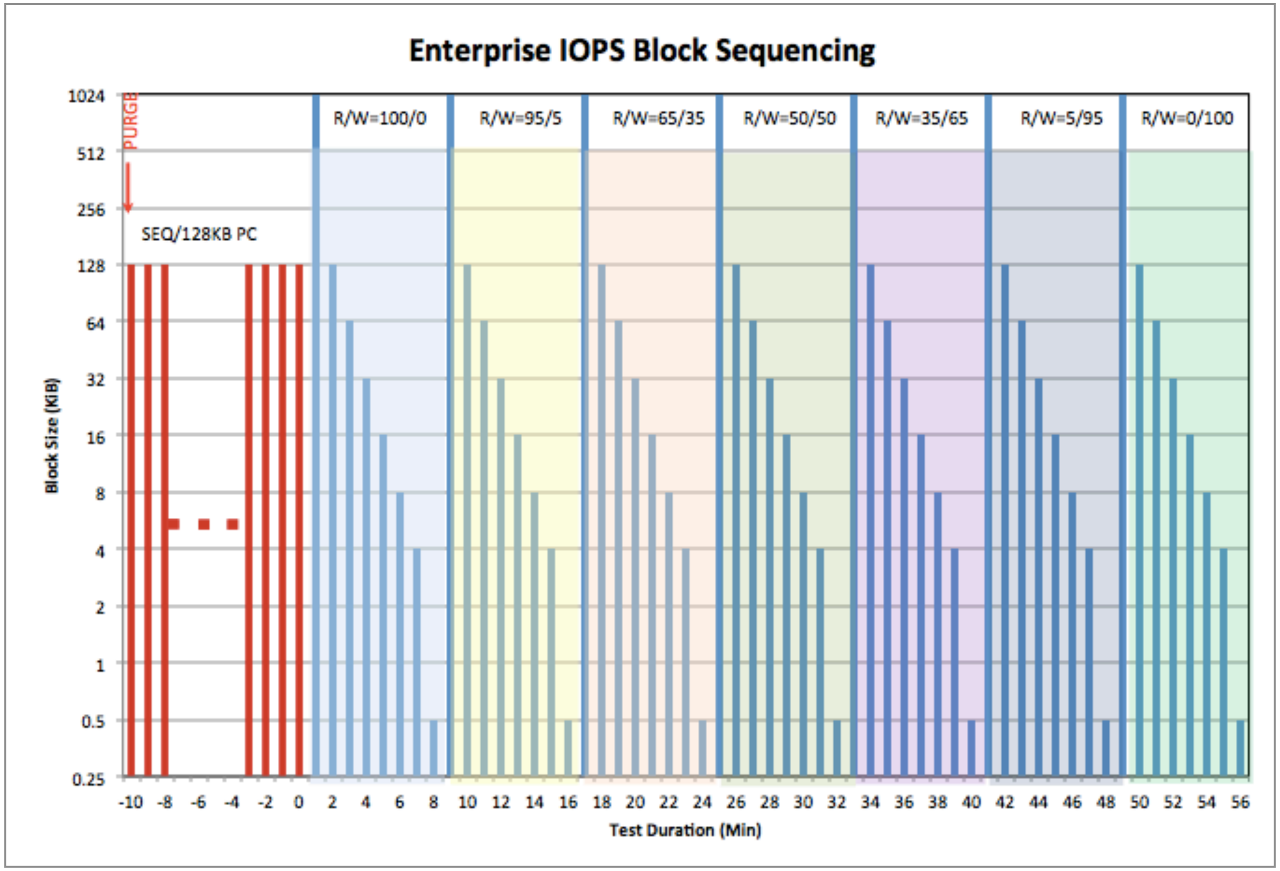
- **Coalescing** – combining small IO data transfers into larger IO data transfers
- **Splitting** – breaking large SEQ IO data transfers into multiple concurrent RND IOs
- **System Cache** – using faster file system cache to defer commits to NAND flash

Standardized Pre-Conditioning

- PURGE – a Known & Repeatable Test Starting Point
- Put the Device Under Test (DUT) in a state "as if no writes have occurred."
- Commands:
  - SECURITY ERASE – ATA Command
  - FORMAT UNIT – SCSI Command
  - Proprietary Command that resets NAND cells to "no write" state

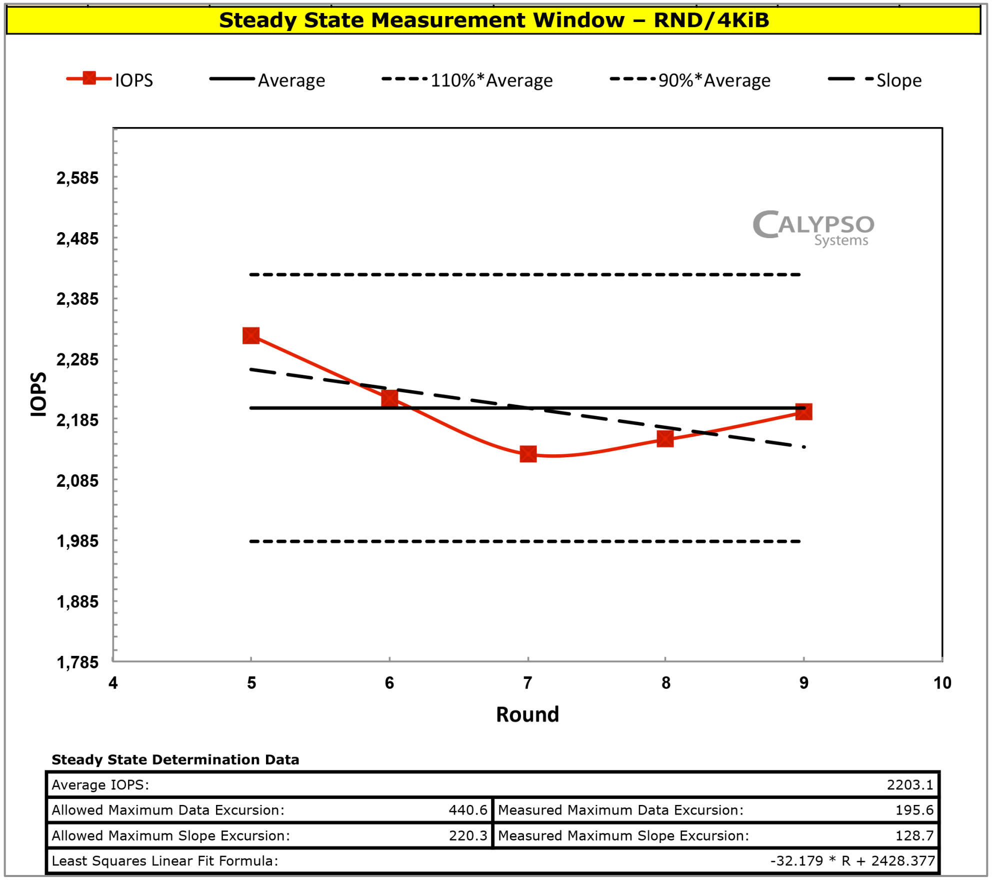
Standardized Pre-Conditioning

Basic Tests – IOPS, Throughput & Latency



- PURGE the DUT
- Write Twice the User Capacity in SEQ 128KiB Ws
- Apply Workload Dependent PC (test access pattern to be measured)

Standardized Steady State



Steady State Determination Data

Average IOPS:			2203.1
Allowed Maximum Data Excursion:	440.6	Measured Maximum Data Excursion:	195.6
Allowed Maximum Slope Excursion:	220.3	Measured Maximum Slope Excursion:	128.7
Least Squares Linear Fit Formula:			-32.179 * R + 2428.377

- Determine Steady State (5 Point Formula – 20% excursion/10% slope)
- Take data from Steady State Window

Pre-Conditioning & Steady State

PTS Variants

WSAT – Three ways to determine Steady State

1. Time
2. Total GB Written
3. 5 Point Formula (one-minute average separated by 30 min of Writes)

Host Idle Recovery – after WSAT Steady State

Cross Stimulus Recovery – time based segments

DIRTH Tests – Thread Count x Queue Depth Loops

- Pre-writes and Inter-loop writes
- Tracking variable 5 point Steady State formula

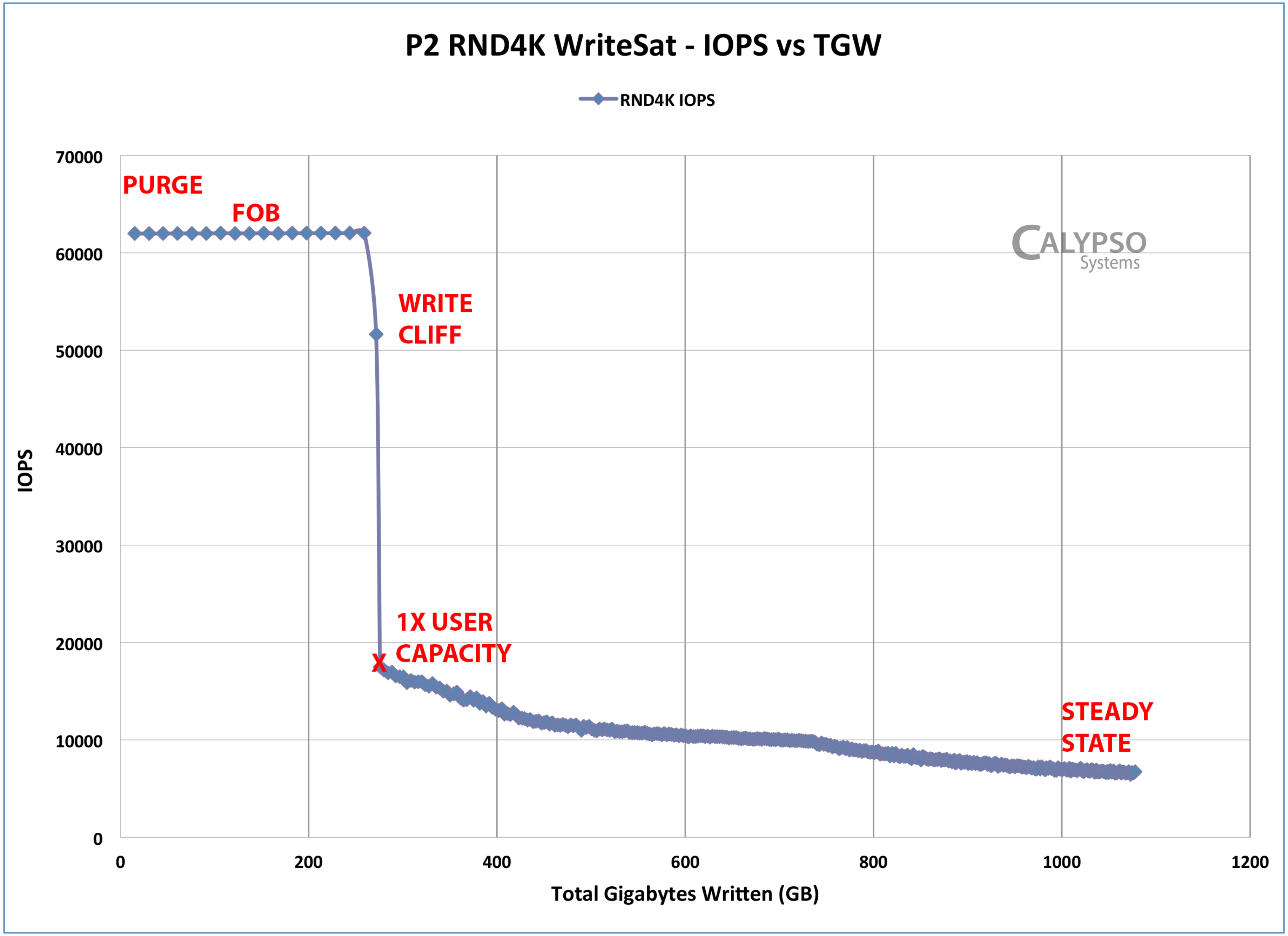
Part 2: PTS - Basic Tests

- Basic PTS Tests – 4 Dimensions of Performance
- Industry Baseline for Comparative Test
- Effects of Changing Test Parameters

4 Dimensions of Performance

- Evolution over Time – Write Saturation Test
- Transaction Rate – IOPS Test
- Bandwidth – Throughput Test
- Response Time – Latency Test

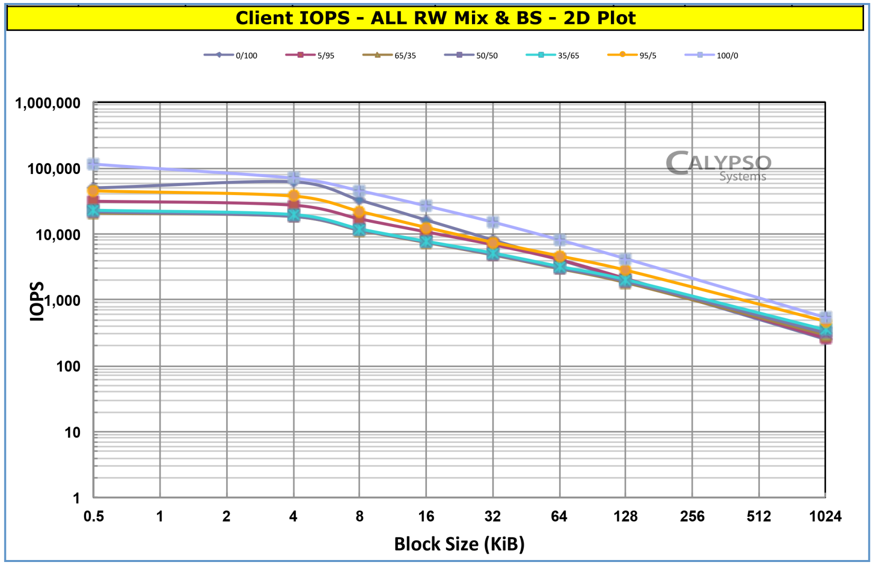
WSAT – Evolution over Time

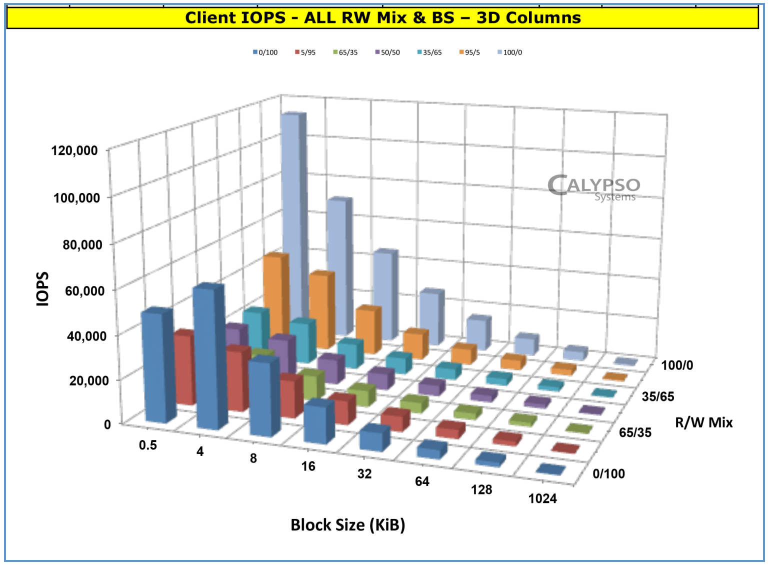


Transaction Rate – IOPS

Client IOPS - ALL RW Mix & BS – Tabular Data

Block Size (KiB)	Read / Write Mix %						
	0/100	5/95	65/35	50/50	35/65	95/5	100/0
0.5	49,053.4	32,137.8	21,205.1	21,452.2	22,868.5	45,001.1	112,880.9
4	62,079.1	27,433.1	18,450.9	18,515.8	19,760.5	37,734.1	70,630.4
8	32,683.6	16,954.9	11,394.2	11,547.9	11,968.9	22,078.8	45,403.6
16	16,306.5	10,776.8	7,430.6	7,536.9	7,756.2	12,587.9	26,747.9
32	8,137.5	6,903.3	4,821.4	4,894.0	5,156.5	7,500.5	15,215.7
64	4,070.8	4,097.6	2,980.1	3,044.3	3,218.5	4,650.9	8,169.2
128	2,034.1	2,113.2	1,830.5	1,912.9	2,034.1	2,827.4	4,224.2
1024	253.4	263.6	293.6	317.5	352.9	474.9	540.6


Client IOPS - ALL RW Mix & BS - 2D Plot

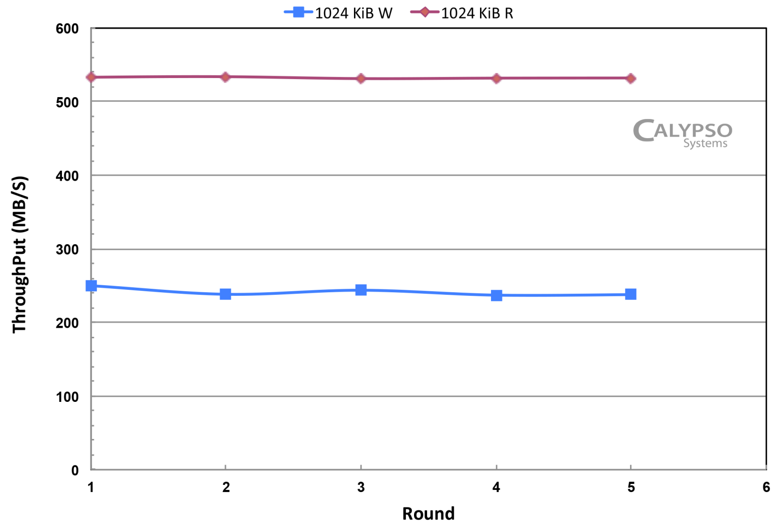

Client IOPS - ALL RW Mix & BS – 3D Columns

Bandwidth – Throughput

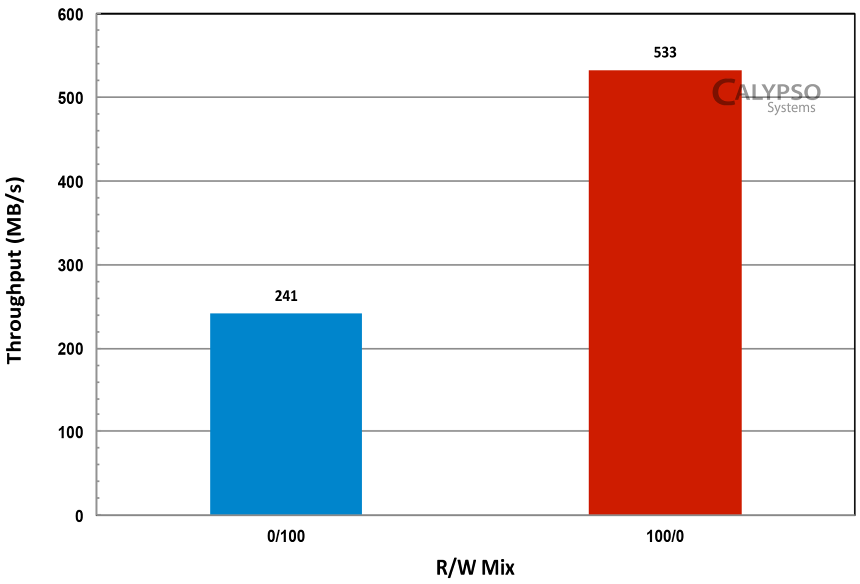
Client Throughput - ALL RW Mix & BS – Tabular Data

Block Size (KiB)	Read / Write Mix %	
	0/100	100/0
1024	241.5	532.6

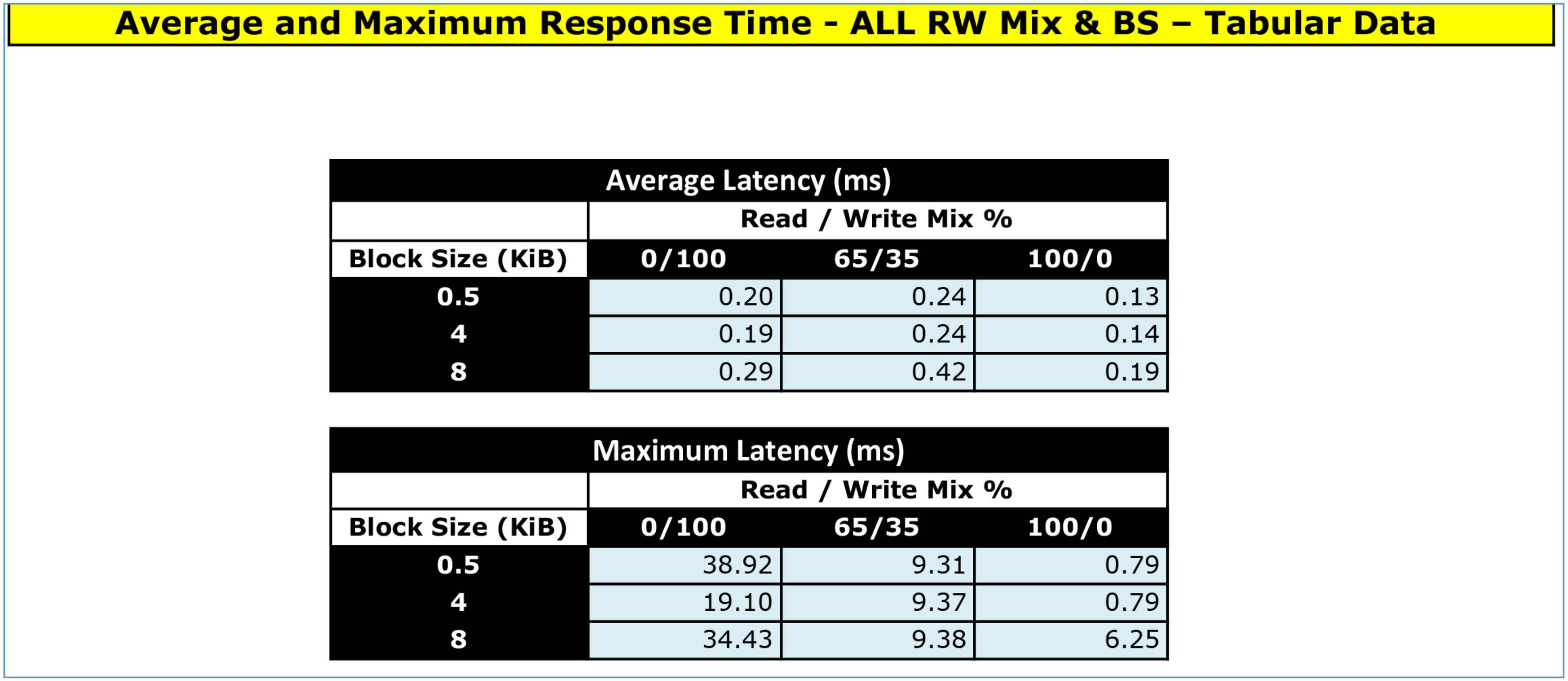
Throughput Test - SS Convergence - 100% Read & 100% Write



Throughput - ALL RW Mix & BS - 2D Plot 1024KiB



Response Time – Latency

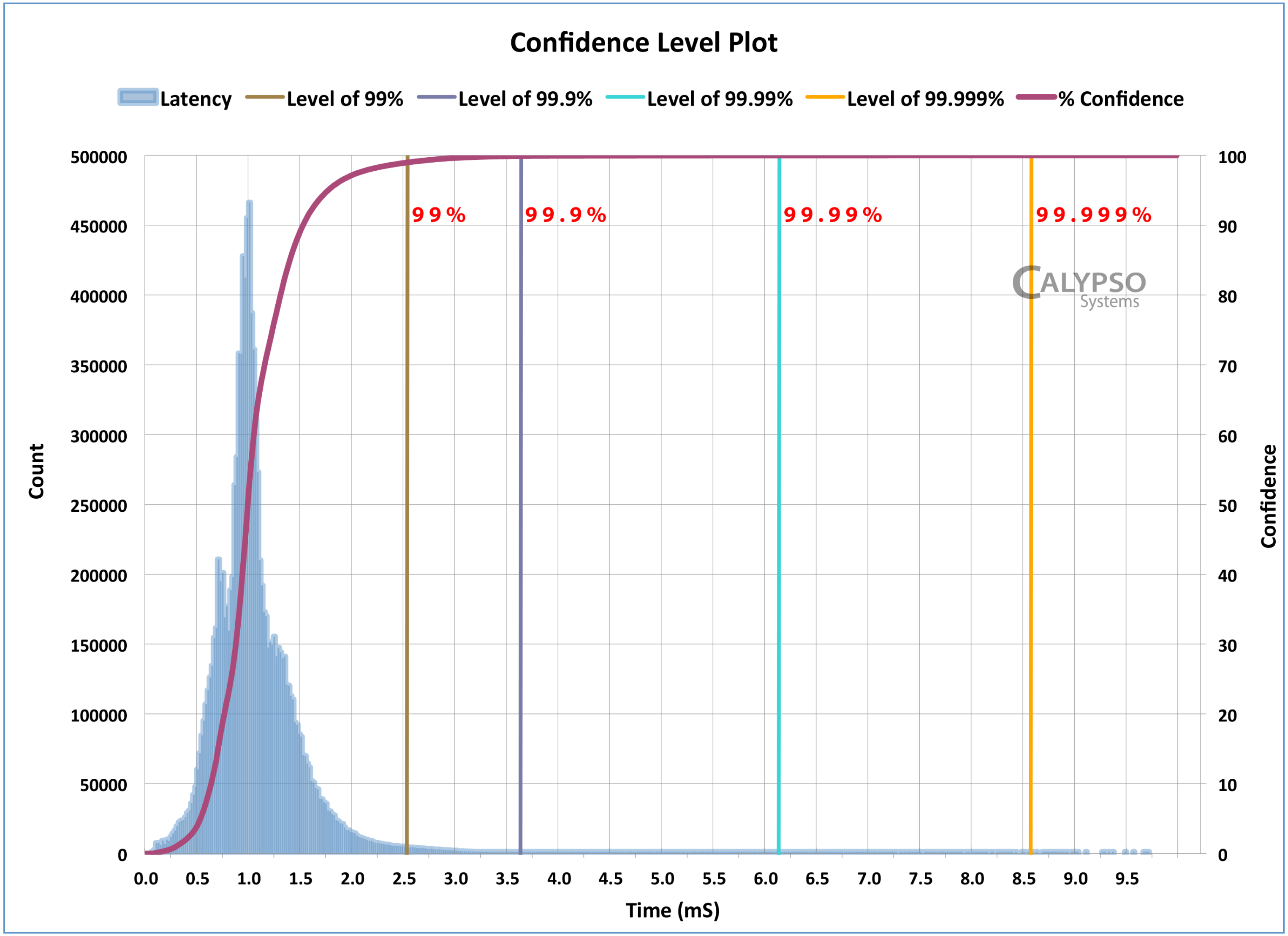


Average and Maximum Response Time - ALL RW Mix & BS – Tabular Data

Average Latency (ms)			
	Read / Write Mix %		
Block Size (KiB)	0/100	65/35	100/0
0.5	0.20	0.24	0.13
4	0.19	0.24	0.14
8	0.29	0.42	0.19

Maximum Latency (ms)			
	Read / Write Mix %		
Block Size (KiB)	0/100	65/35	100/0
0.5	38.92	9.31	0.79
4	19.10	9.37	0.79
8	34.43	9.38	6.25

Response Time – Histogram

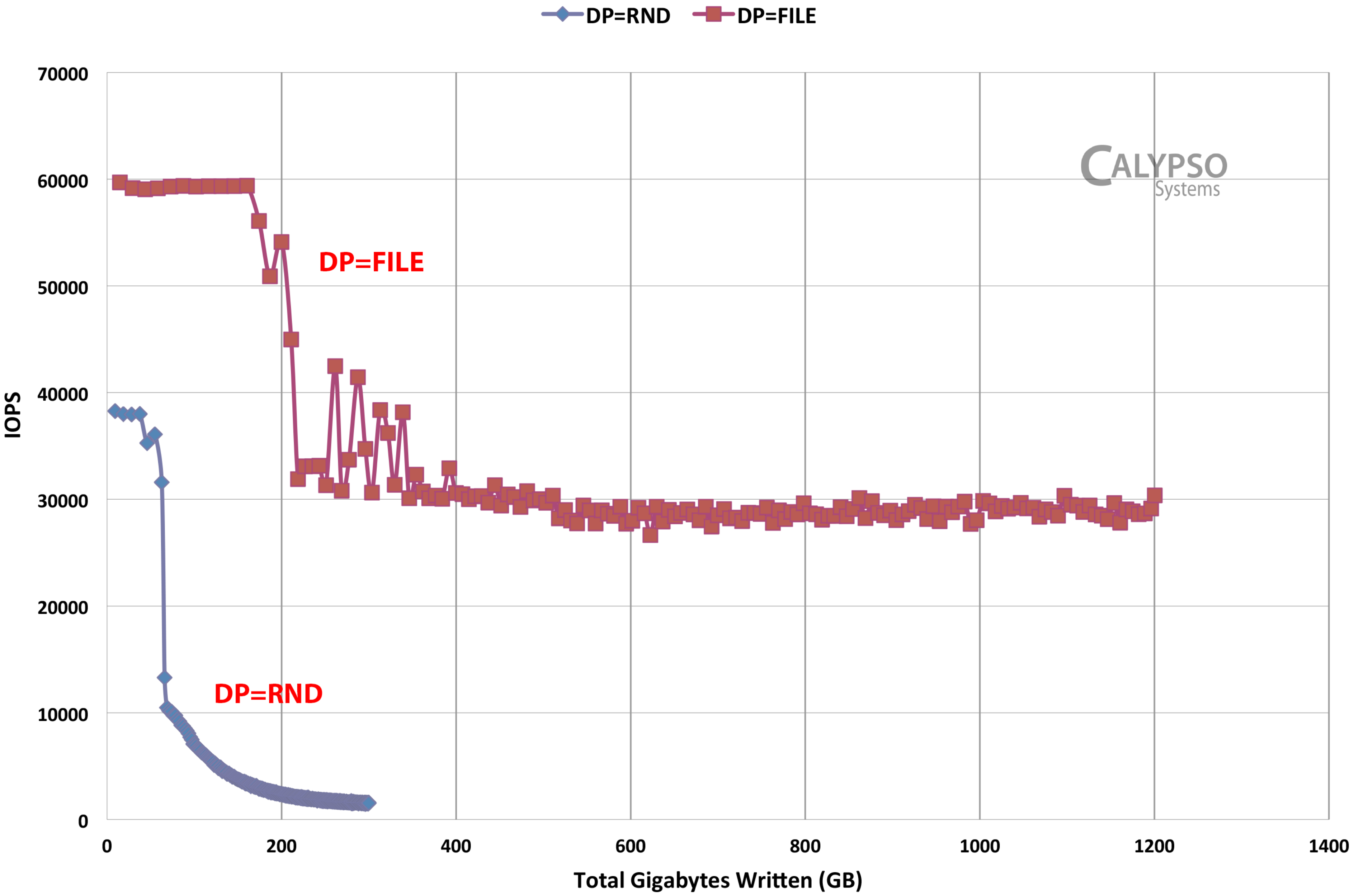


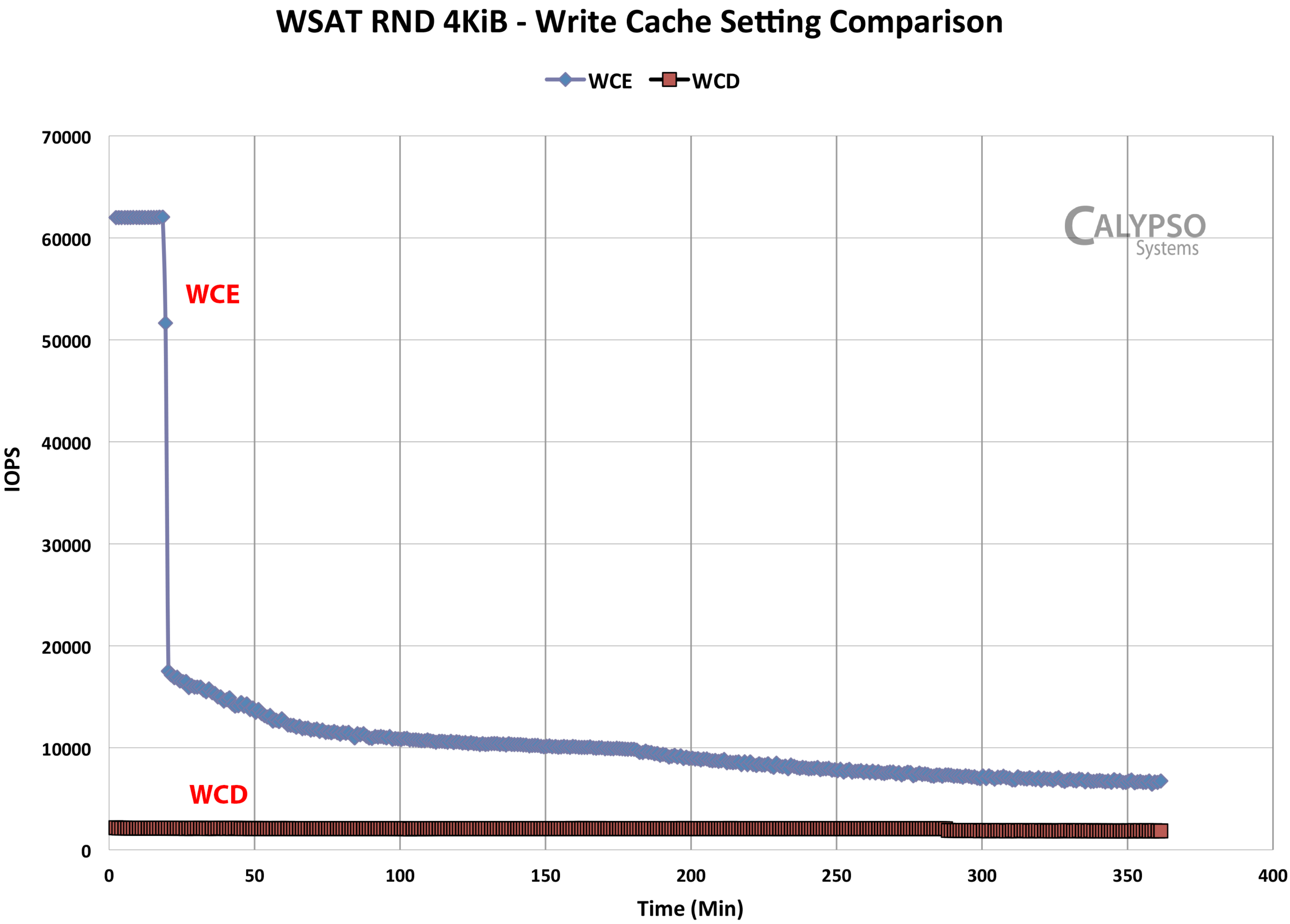
Effects of Parameter Settings

- Data Pattern – RND v Non-RND
- Write Cache Setting – WCE v WCD
- OIO Throttling – Limiting Threads & Queues
- Over Provisioning – Limiting PC and Test Active Ranges
- Active Range – Enterprise v Client PTS Setting

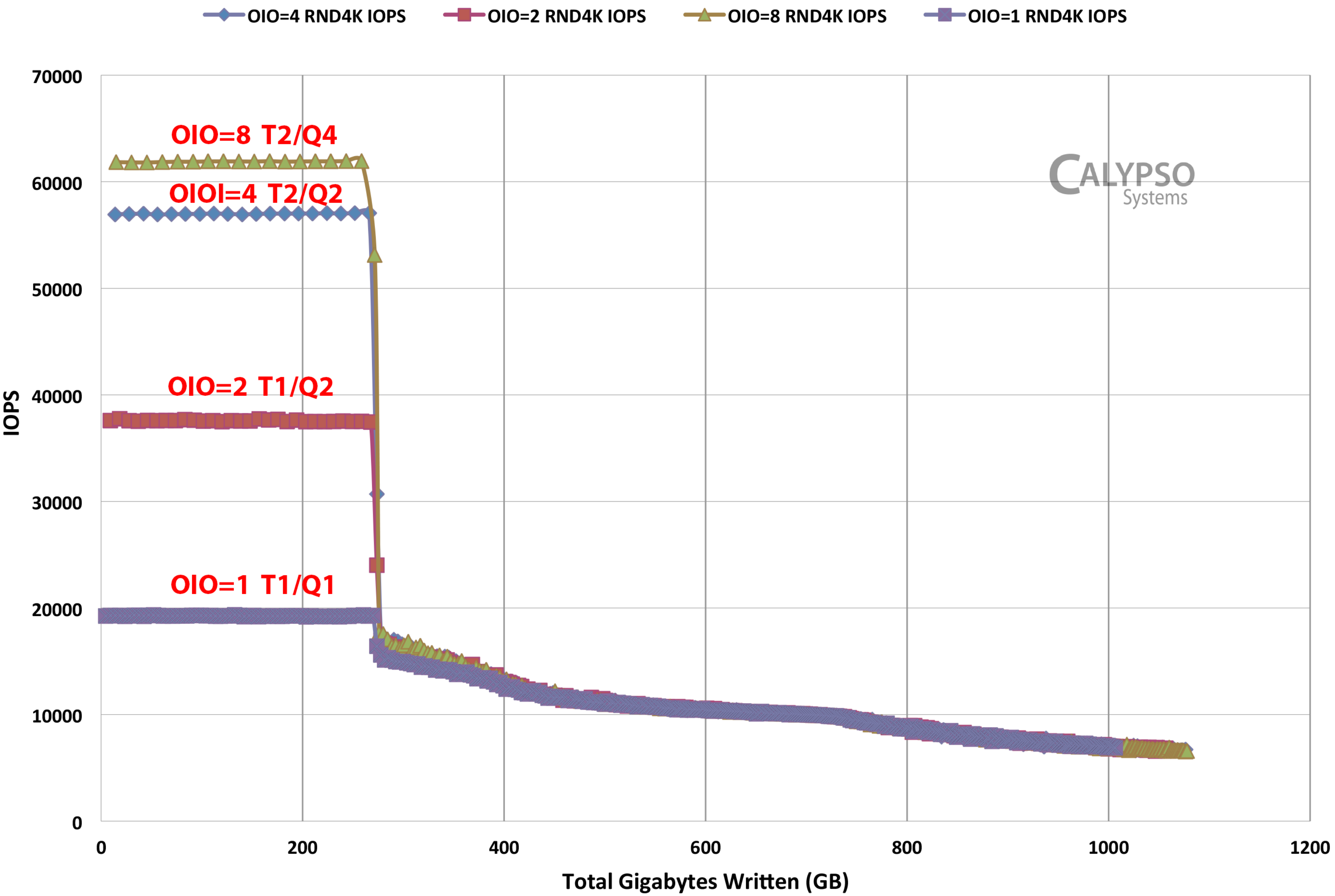
NOTE: PTS sets forth required and optional parameter settings to ensures that test conditions match intended workloads and that tests are repeatable & comparable.

DATA PATTERN COMPARISON - WSAT RND4KiB - IOPS v TGBW

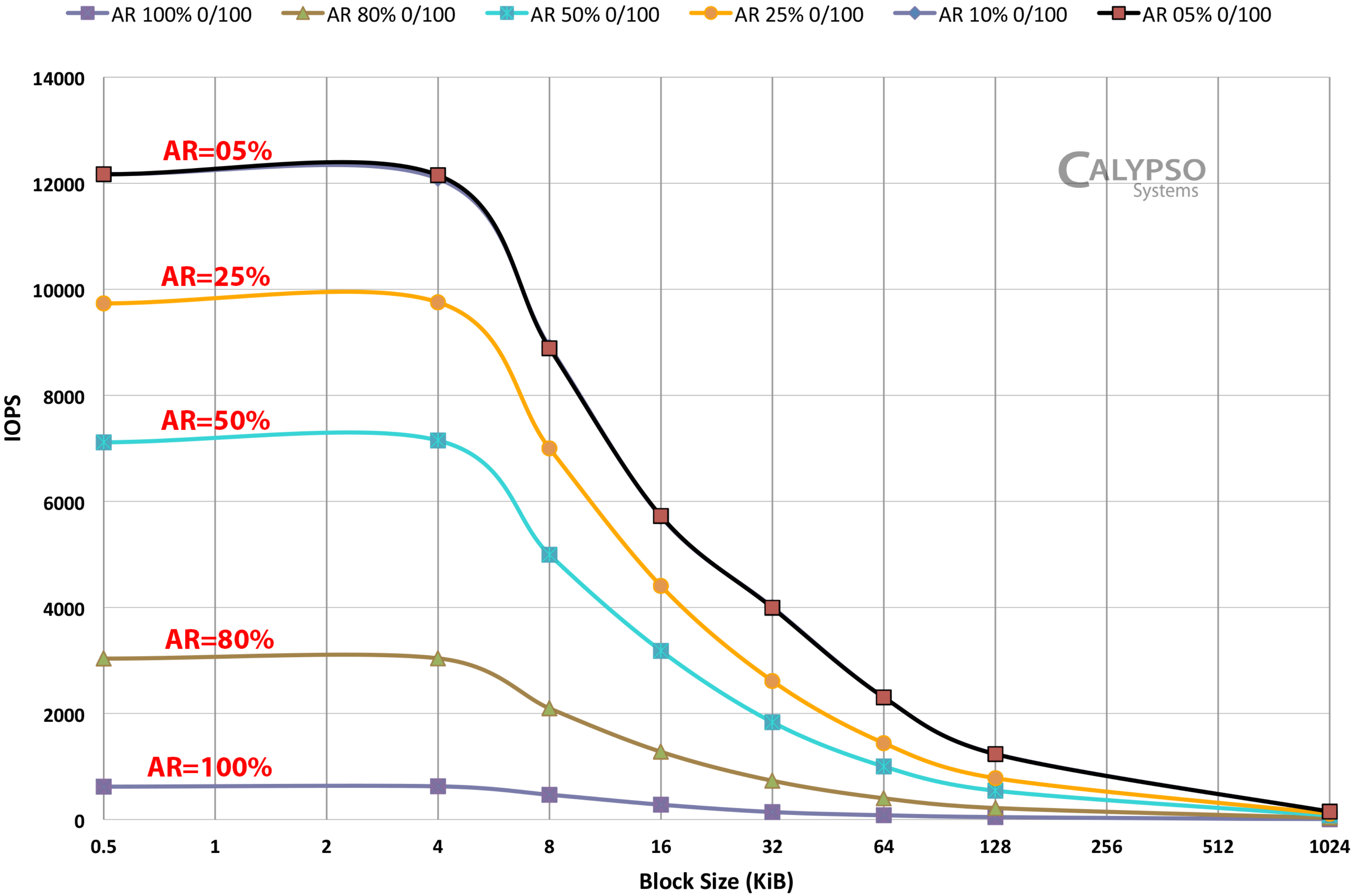



WSAT RND 4KiB - Write Cache Setting Comparison
WCE
WCD
70000
60000
50000
40000
30000
20000
10000
0
IOPS
0
50
100
150
200
250
300
350
400
Time (Min)
WCE
WCD
CALYPSO Systems

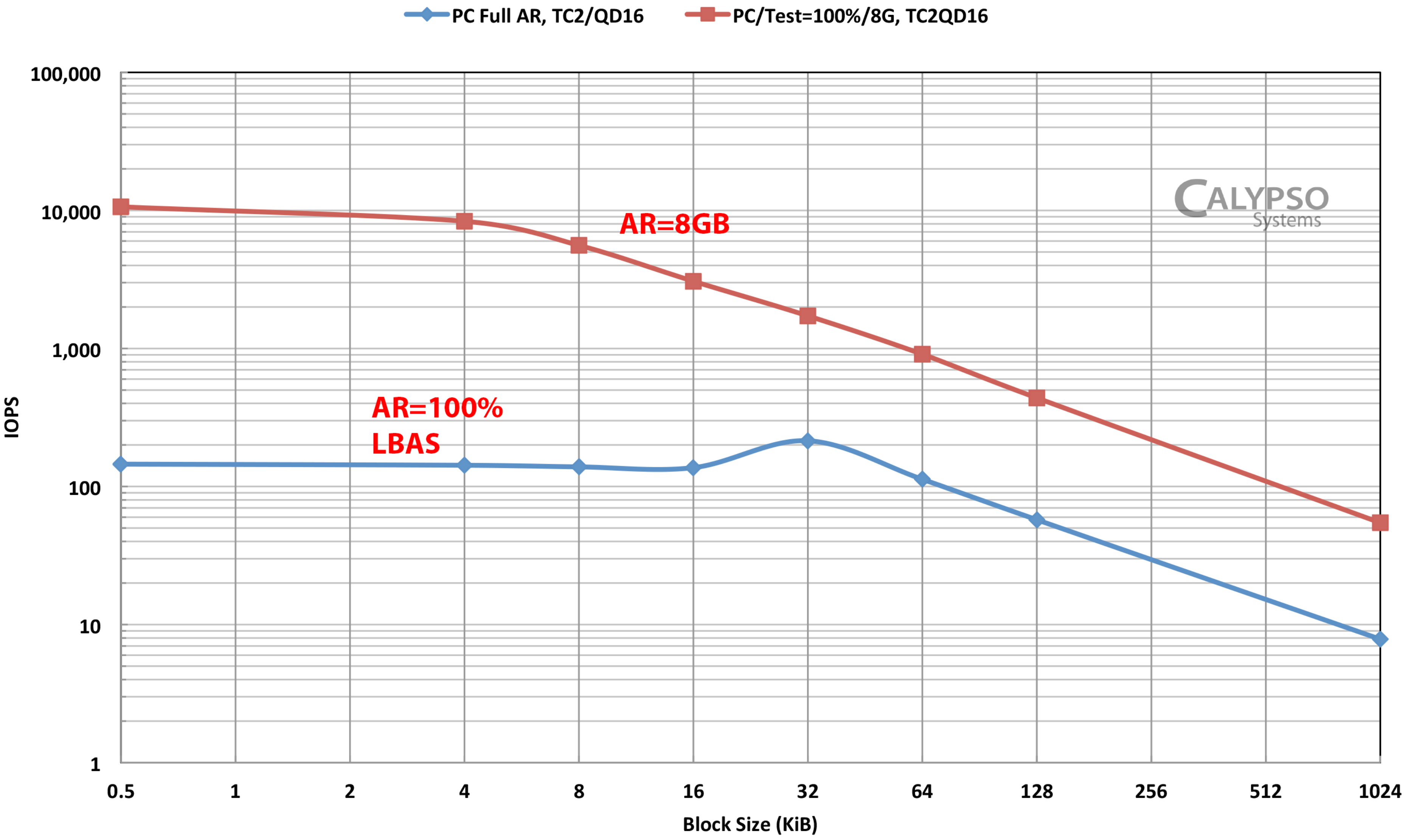
Throttling OIO - WSAT RND 4KiB - IOPS v TGBW



Data Compare SSD B P4 100% W - Over Provisioning / AR Amount



IOPS Full AR versus 8GB AR, WRITE IOPS



Part 3: PTS - Advanced Tests

- Specialized Pre-conditioning Methodology
- Host Idle Recovery – HIR
- Cross Stimulus Recovery – XSR
- Demand Intensity Response Time Histograms
- Examples

Tests Contained In PTS-E 1.0 SPEC

- Enterprise Performance Test Specification (PTS-E) V1.0 encompasses:
  - A suite of basic SSS performance tests
  - *Preconditioning* and *Steady State* requirements
  - Standard test procedures and *reporting requirements*

Write Saturation

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

Enterprise IOPS

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

Enterprise TP

- **Sequential Access**
- **R/W:**
  - 100/0, 0/100
- **BS:**
  - 1024KiB, 128KiB

Enterprise Latency

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

Flash Memory SUMMIT

Tests Contained In PTS-E 1.1

- PTS-E 1.1 adds:

Host Idle Recovery

- Examines effect of idle (no IO) on small block RND writes
- RND/4KiB Writes

Cross Stimulus Response

- Examines switching between large block SEQ and small block RND writes
- SEQ/1024KiB & RND/8KiB Writes

Demand Intensity – Response Time Histograms

- Performance and detailed response time statistics under various workload types
- R/W=65/35 %, RND/8K
- R/W=90/10 %, RND/128K
- Response Time Histograms at various operating points

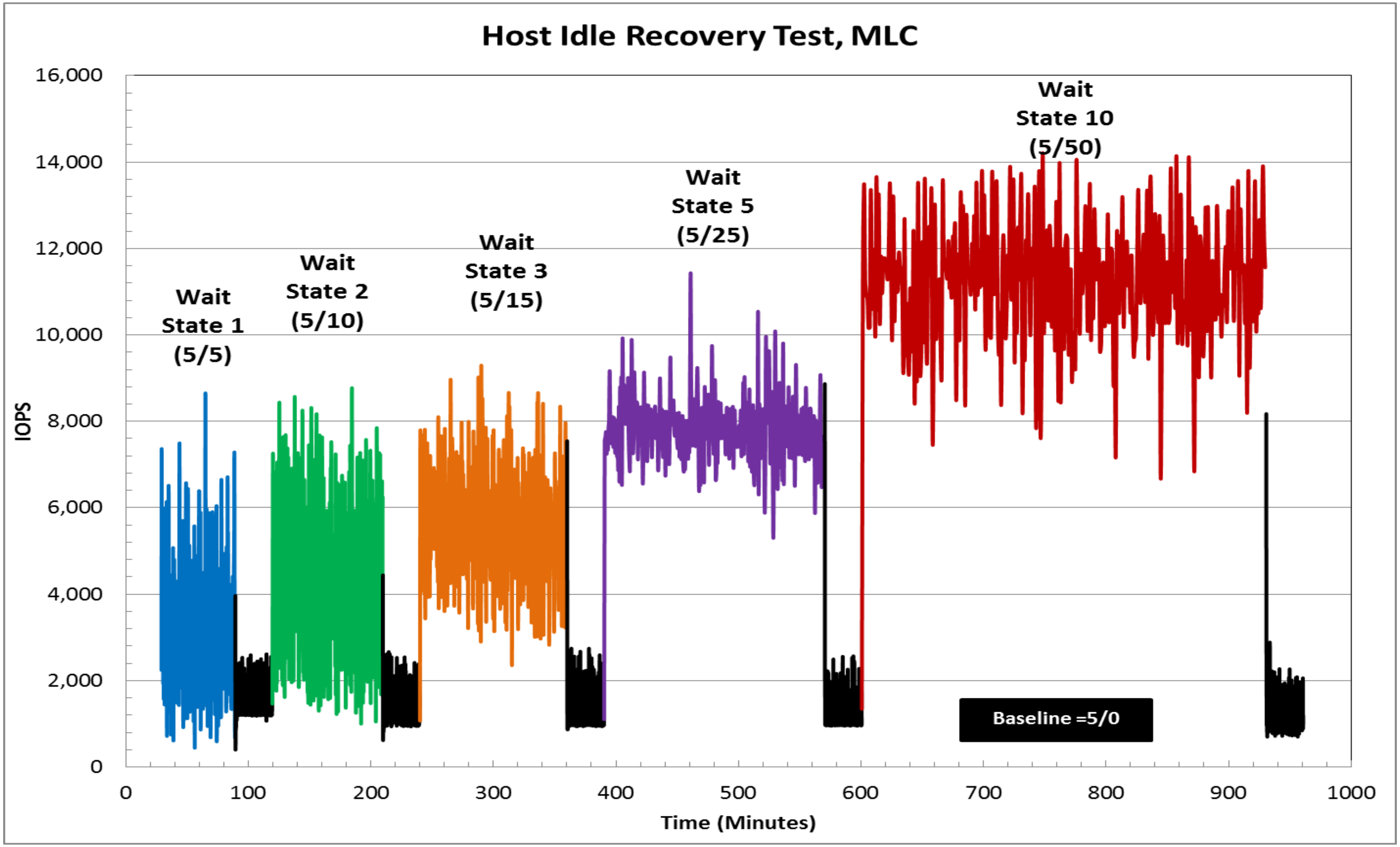
Enterprise Composite Workload

- Performance and detailed response time in a mixed IO Enterprise environment
- R/W=60/40 %
- BS from 0.5-64KiB
- Three LBA probability groups

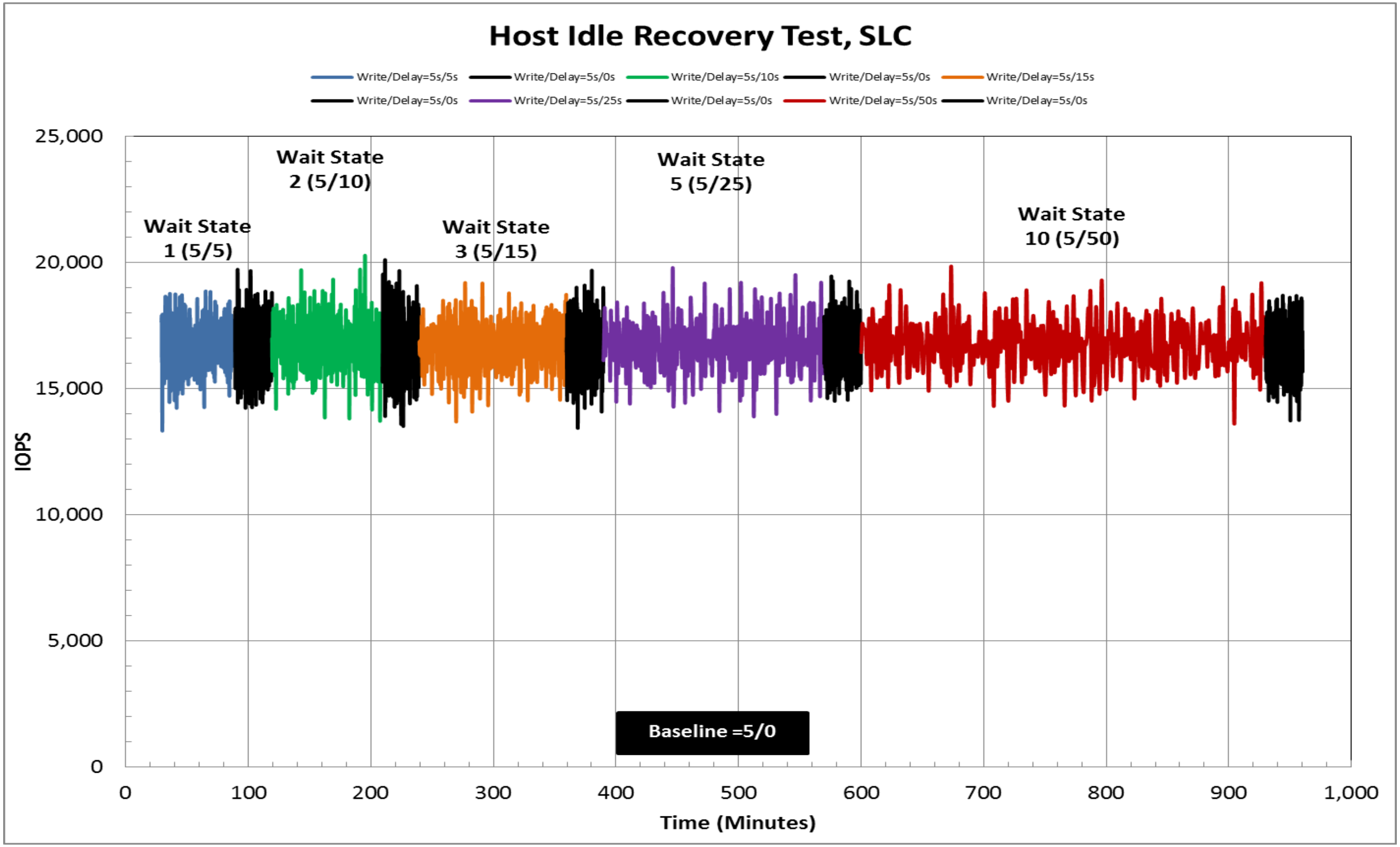
Host Idle Recovery Test (HIR)

Purpose	Examines Effect of Host Idle Period (No Host IO) On The Performance of RND Small Block Writes
Test Setup	
Preconditioning	RND/4KiB Writes to Steady State
Test	Insert various amount of idle time (no IO from host) between periods of 5 second RND/4KiB writes: Segment 1 (Wait State 1): 360 x (5S Write + 5S Idle) + 360 x (5S Write) Segment 2 (Wait State 2): 360 x (5S Write + 10S Idle) + 360 x (5S Write) Segment 3 (Wait State 3): 360 x (5S Write + 15S Idle) + 360 x (5S Write) Segment 4 (Wait State 5): 360 x (5S Write + 25S Idle) + 360 x (5S Write) Segment 5 (Wait State 10): 360 x (5S Write + 50S Idle) + 360 x (5S Write)

HIR Example: MLC/SATA



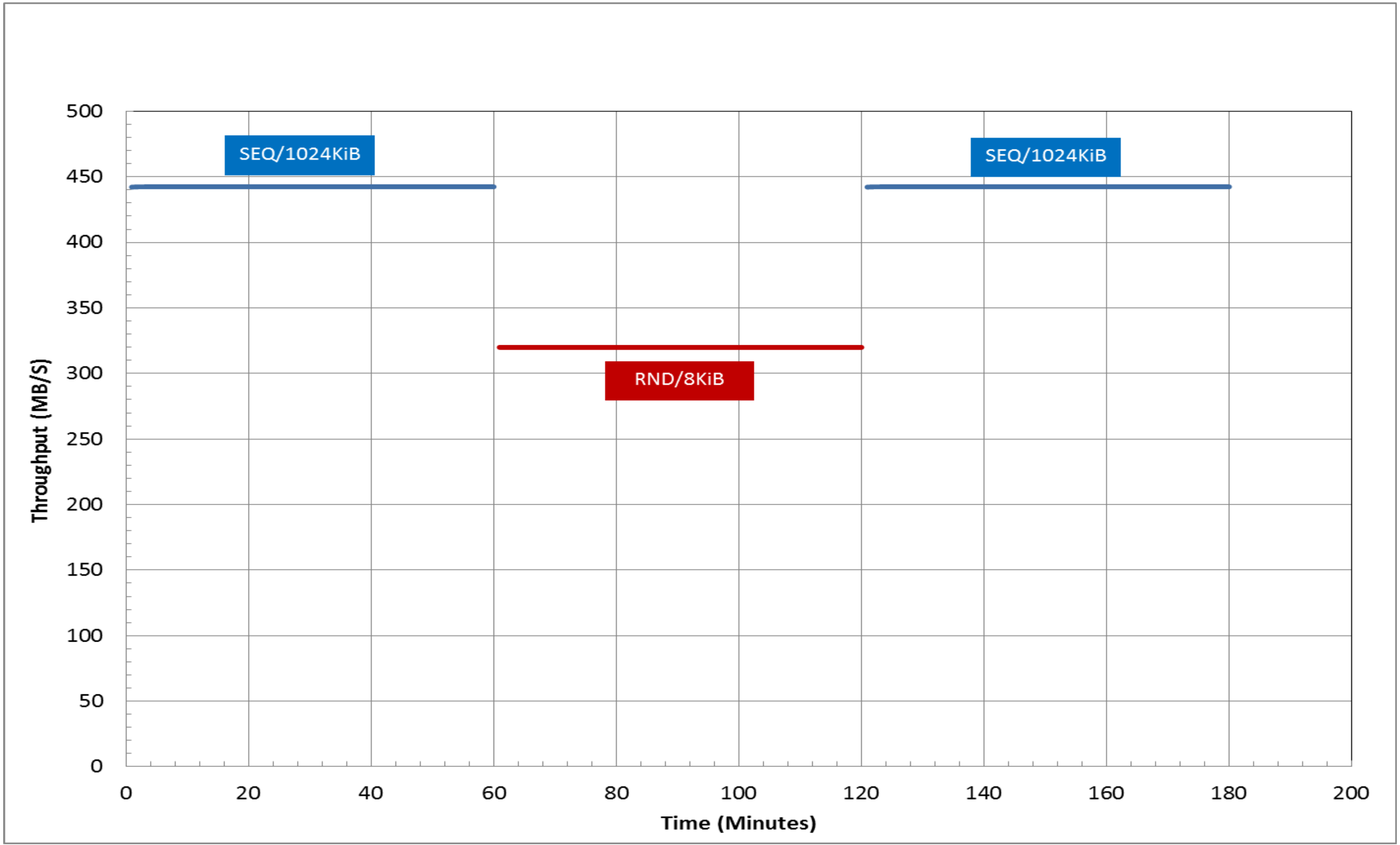
HIR Example: SLC/SAS



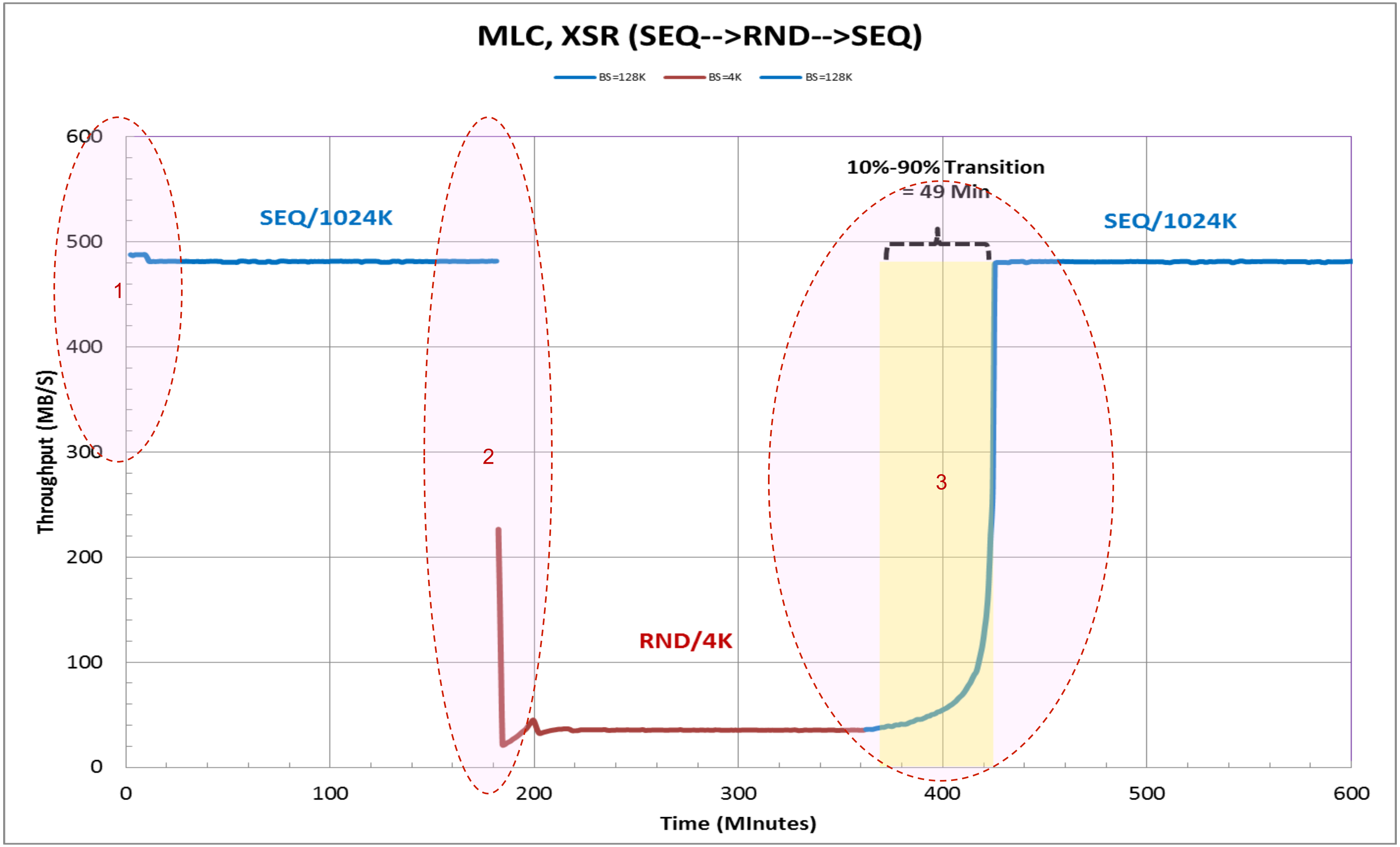
Cross Stimulus Response Test (XSR)

Purpose	Examines Switching Between Sustained Large Block SEQ and Sustained Small Block RND Writes
Test Setup	
Preconditioning	None
Test	Apply three Access Groups: Access Group 1 (Large Block SEQ): 100% SEQ Write, Block Size=1024 KiB Access Group 2 (Small Block RND): 100% RND Write, Block Size=4 KiB Access Group 3 (Large Block SEQ): 100% SEQ Write, Block Size=1024 KiB

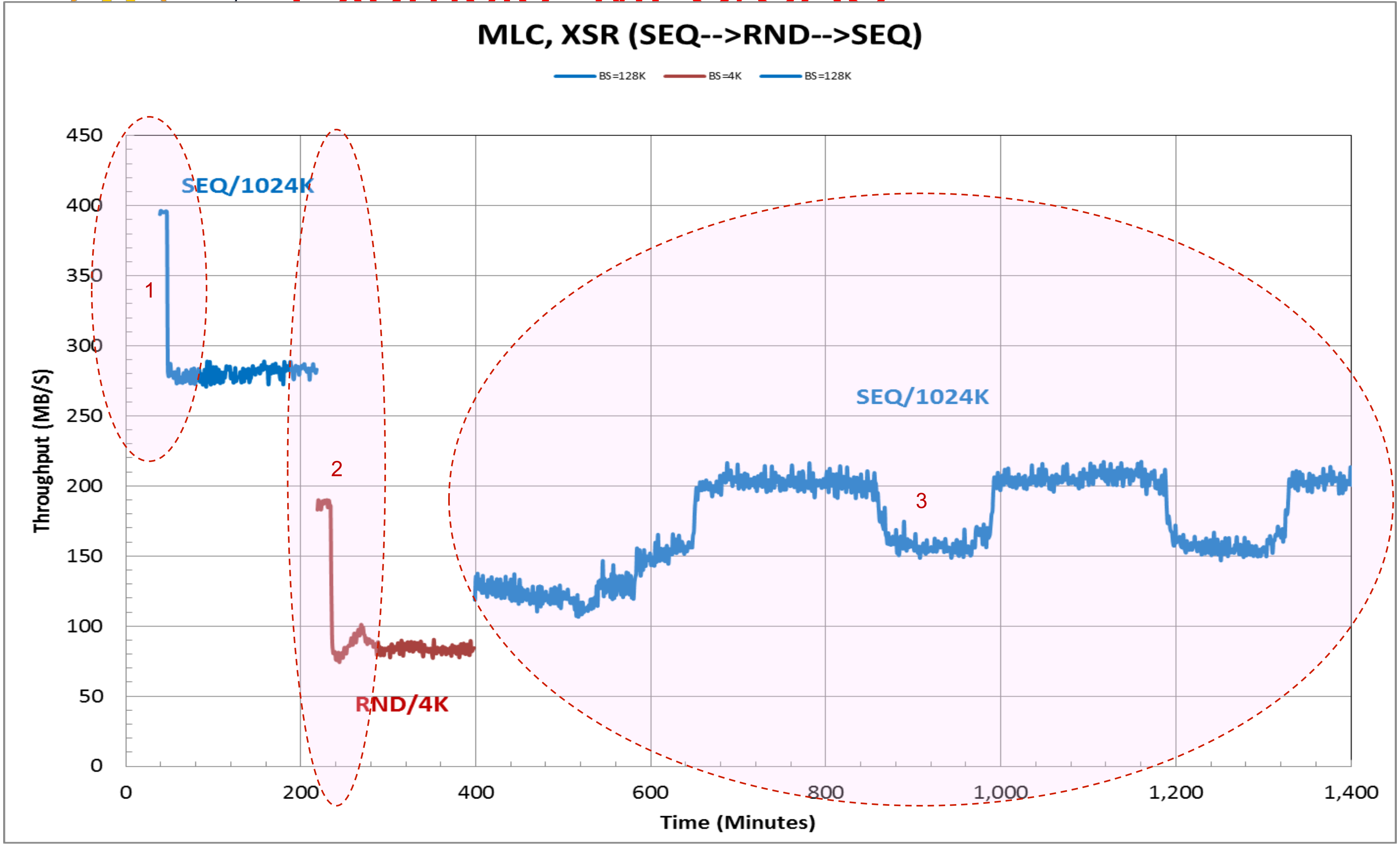
Baseline: Non-NAND-Flash Drive



Example: MLC/SATA



Example: MLC/SAS



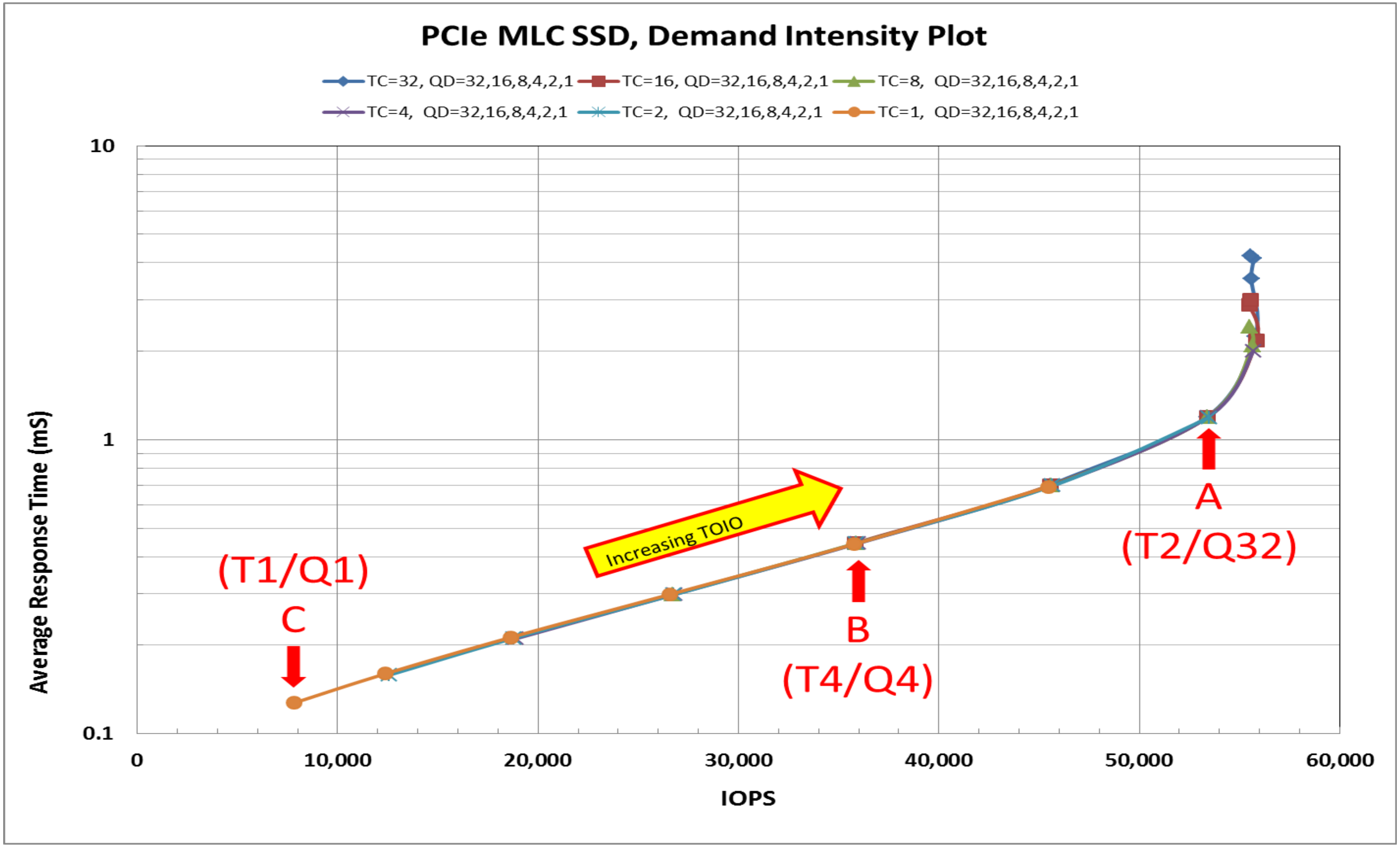
Demand Intensity, Response Time Histogram Test (DIRTH)

Purpose	Examines IOPS and Response Time Characteristics of Various Enterprise Workloads
Test Setup	
Preconditioning	Access Pattern, 100% Writes, until Steady State
Test	1. Using TC=[1,2,4,6,8,16,32] and OIO/Thread=[1,2,4,6,8,16,32], apply ECW using order of decreasing total OIO, until Steady State is reached for (32,32) 2. Manually determine the following operating points: MaxIOPS: operating point with maximum IOPS while maintaining an ART < 5 mS MinIOPS: operating point with minimum measured IOPS MidIOPS: a minimum of one or more operating point(s) that has IOPS values between and equally divides the IOPS value spanned by MaxIOPS and MinIOPS 3. Perform Response Time Histograms, capturing all IO completion times for 10 Min at each operating points.

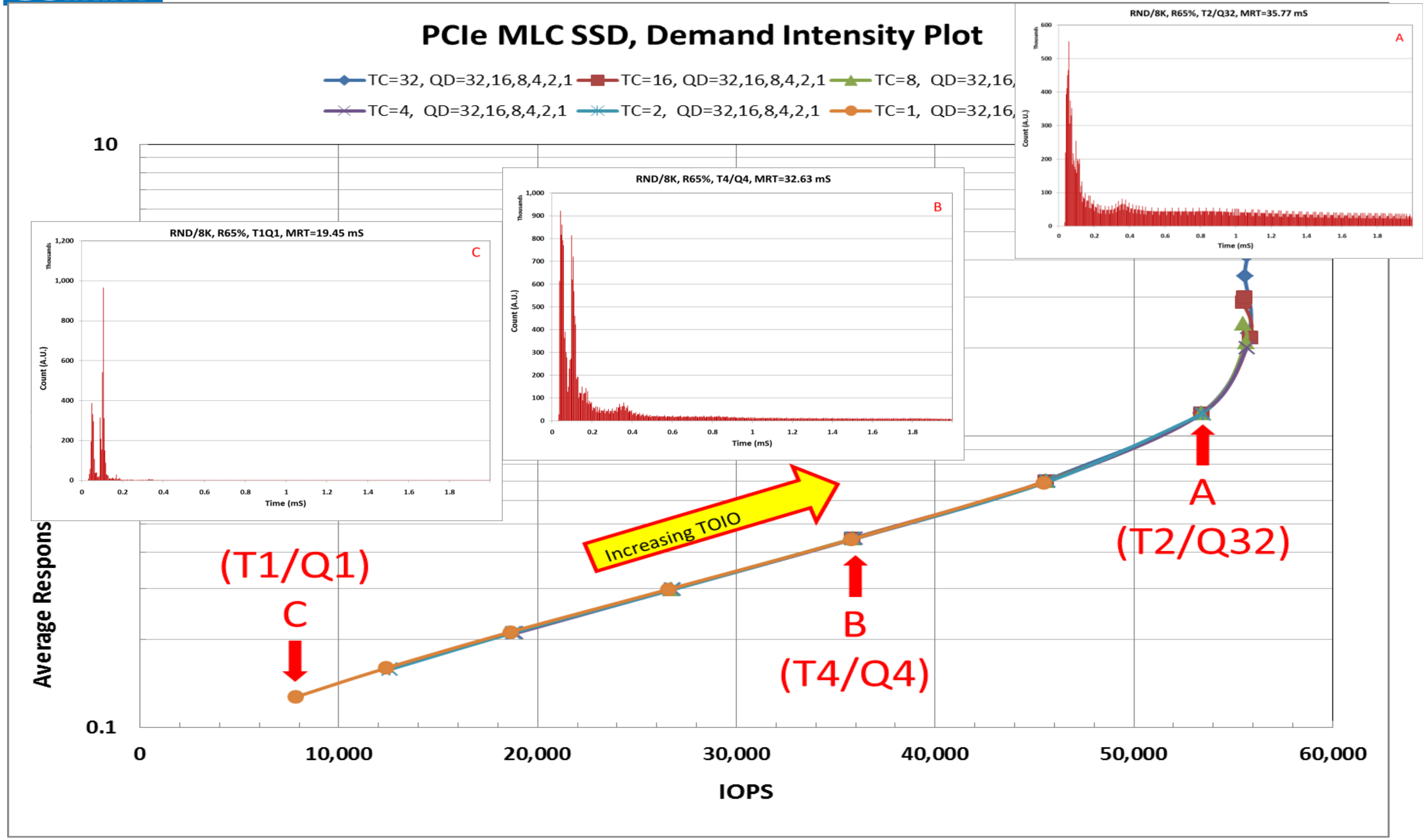
DIRTH Test

- Currently there are two Access Patterns specified for the DIRTH test:
  - OLTP-Like:
    - BS= 8 KiB
    - R/W= 65/35 %
    - Random Access, Random Data
    - Full Drive Access
  - Video-Server-Like
    - BS= 128 KiB
    - R/W= 90/10 %
    - Random Access, Random Data
    - Full Drive Access

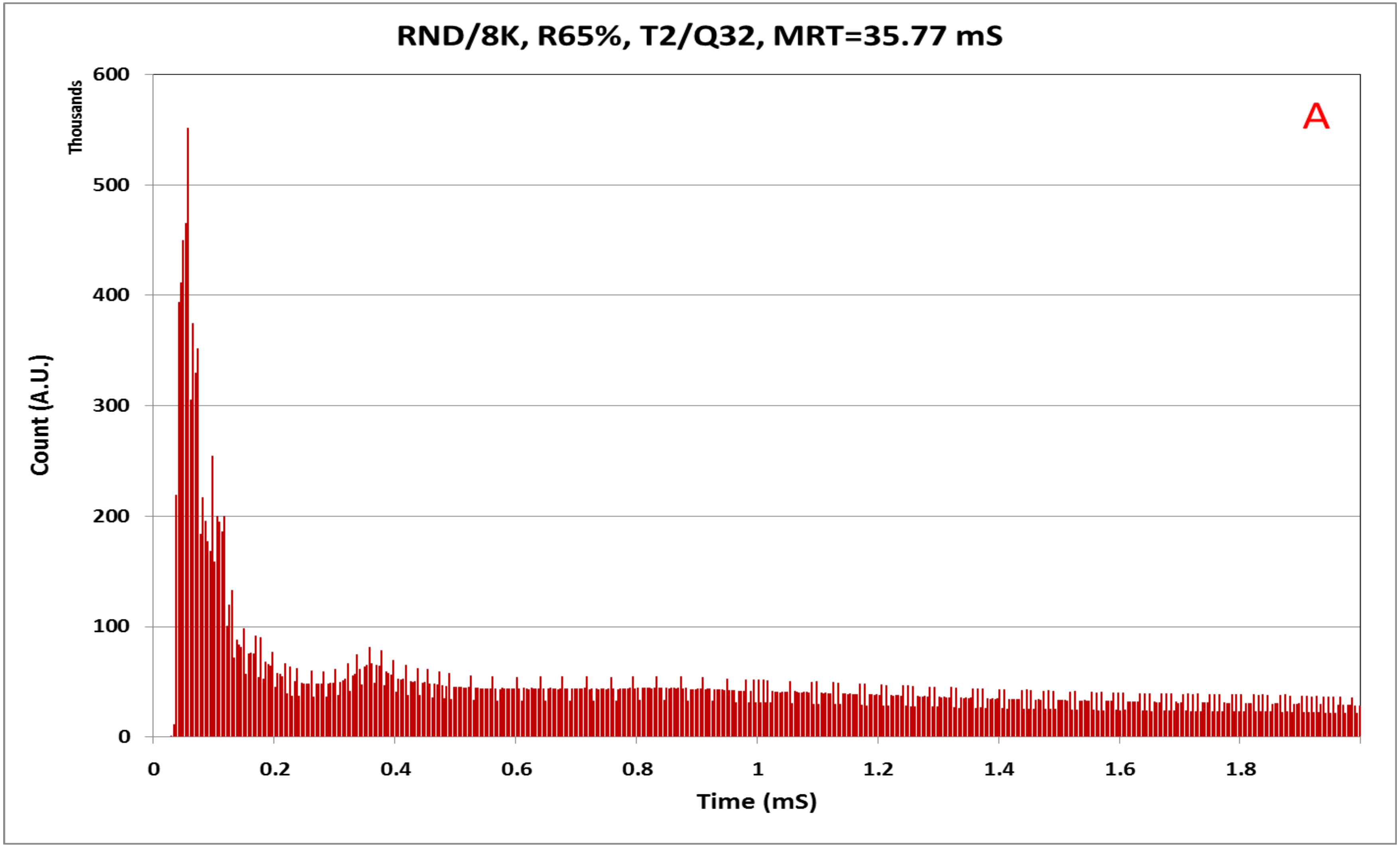
DIRTH (OLTP): Demand Intensity



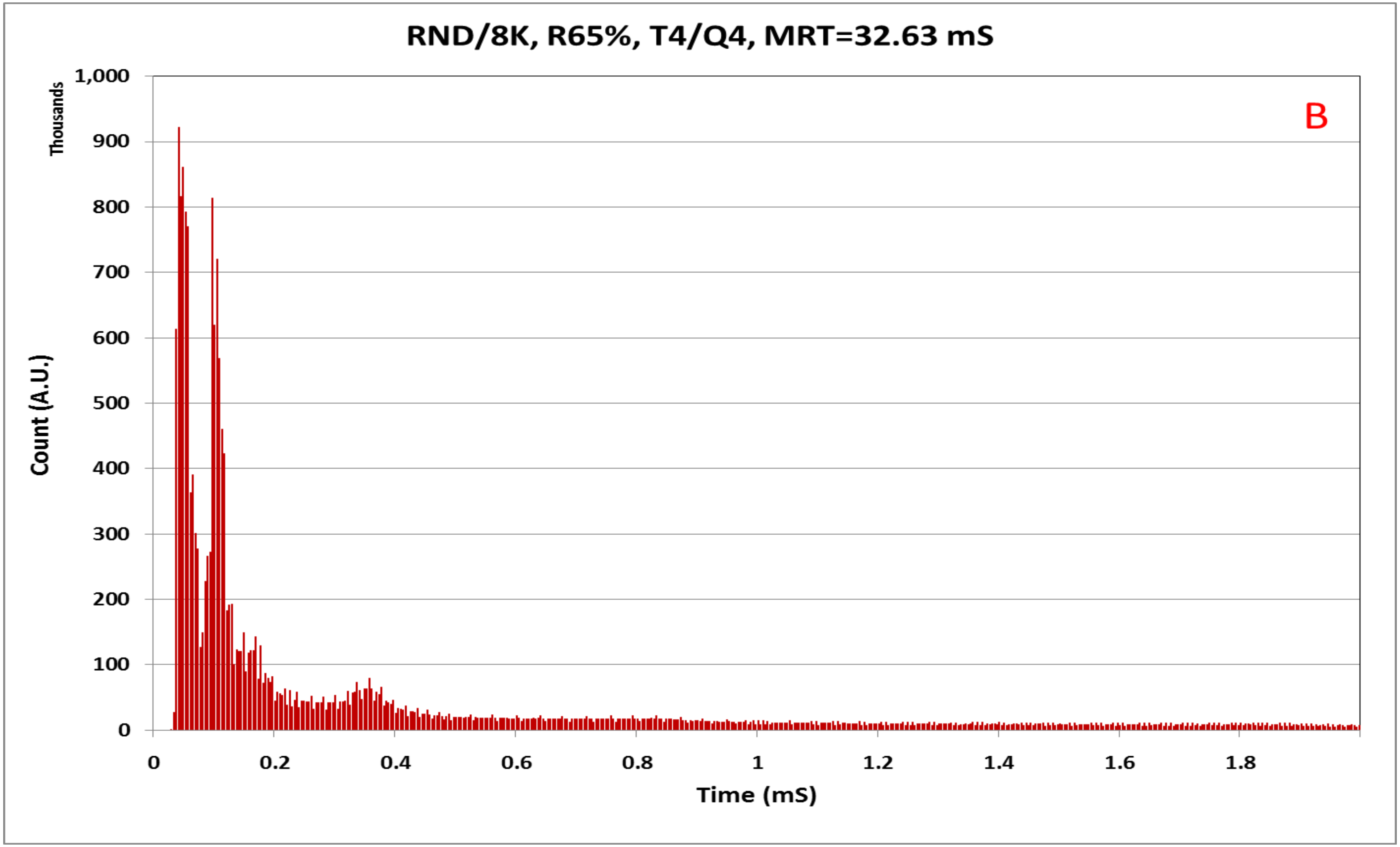
DIRTH (OLTP): Demand Intensity



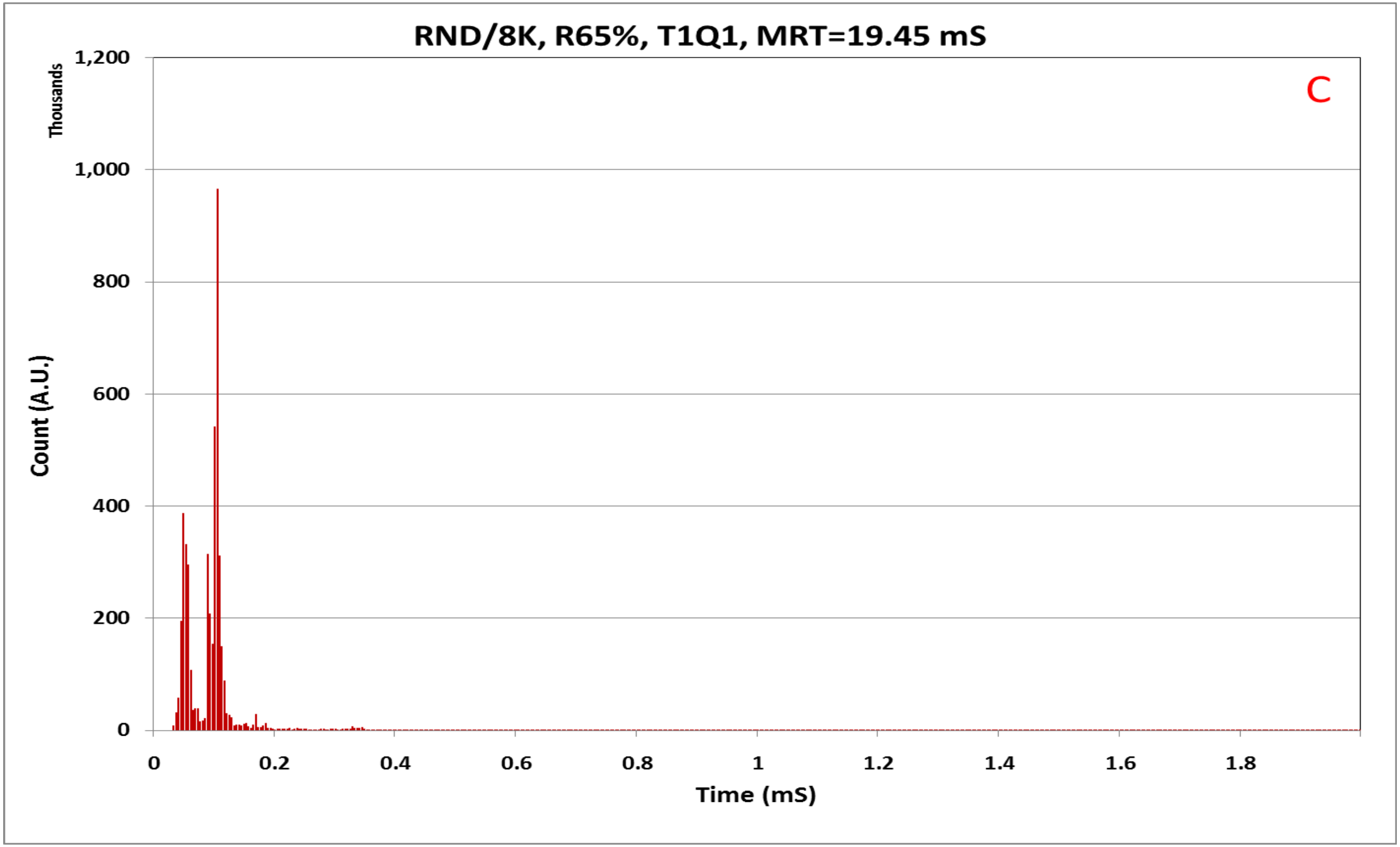
DIRTH (OLTP): MaxIOPS Point



DIRTH (OLTP): MidIOPS Point



DIRTH (OLTP): MinIOPS Point



Enterprise Composite Workload (ECW) Test

Purpose	Examines IOPS and Response Time Characteristics Using a Mixed IO Workload
Test Setup	
Preconditioning	ECW, 100% Write, to Steady State
Test	1. Using TC=[1,2,4,6,8,16,32] and OIO/Thread= [1,2,4,6,8,16,32], apply ECW using order of decreasing total OIO, until Steady State is reached for (32,32) 2. Manually determine the following operating points: MaxIOPS: operating point with maximum IOPS while maintaining an ART < 5 mS MinIOPS: operating point with minimum measured IOPS MidIOPS: a minimum of one or more operating point(s) that has IOPS values between and equally divides the IOPS value spanned by MaxIOPS and MinIOPS 3. Perform Response Time Histograms, capturing all IO completion times for 10 Min at each operating points.

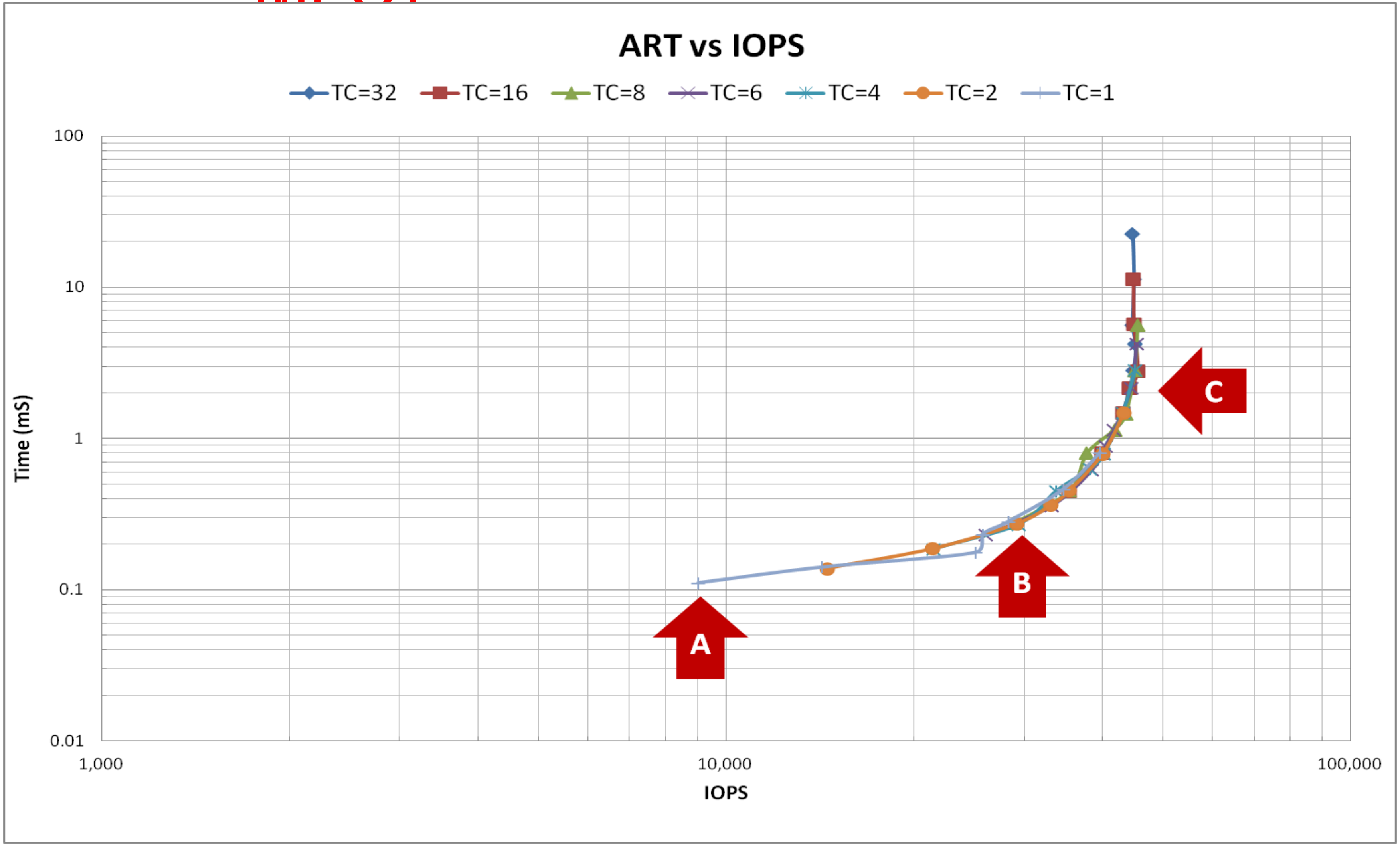
The Enterprise Composite Workload

- The ECW is a R/W=40/60%, random access pattern with a distribution of Block Sizes, each with a pre-defined Access Probability, plus restrictions on Access Range Probability Distribution

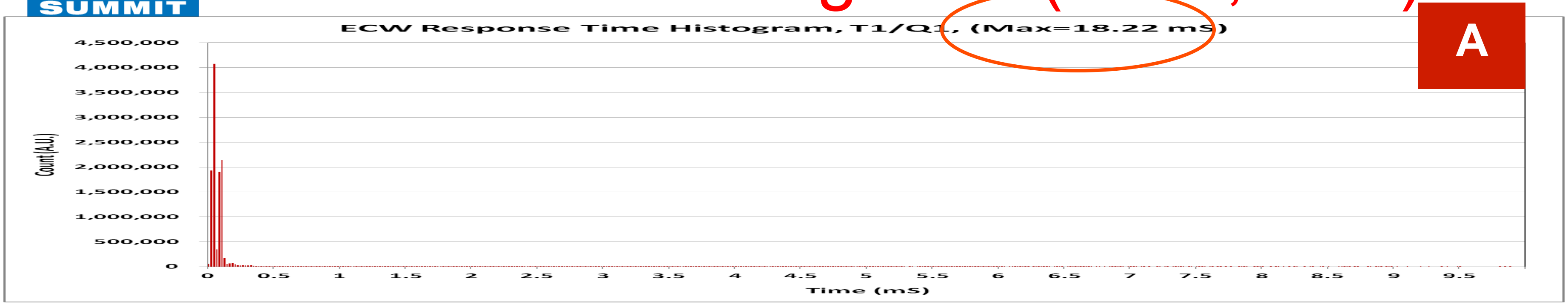
Block Size in Bytes (KiB)	Access Probability Within Each Measurement Period
512 bytes (0.5 KiB)	4%
1024 bytes (1 KiB)	1%
1536 bytes (1.5 KiB)	1%
2048 bytes (2 KiB)	1%
2560 bytes (2.5 KiB)	1%
3072 bytes (3 KiB)	1%
3584 bytes (3.5 KiB)	1%
4096 bytes (4 KiB)	67%
8192 bytes (8 KiB)	10%
16,384 bytes (16 KiB)	7%
32,768 bytes (32 KiB)	3%
65,536 bytes (64 KiB)	3%
Total	100%

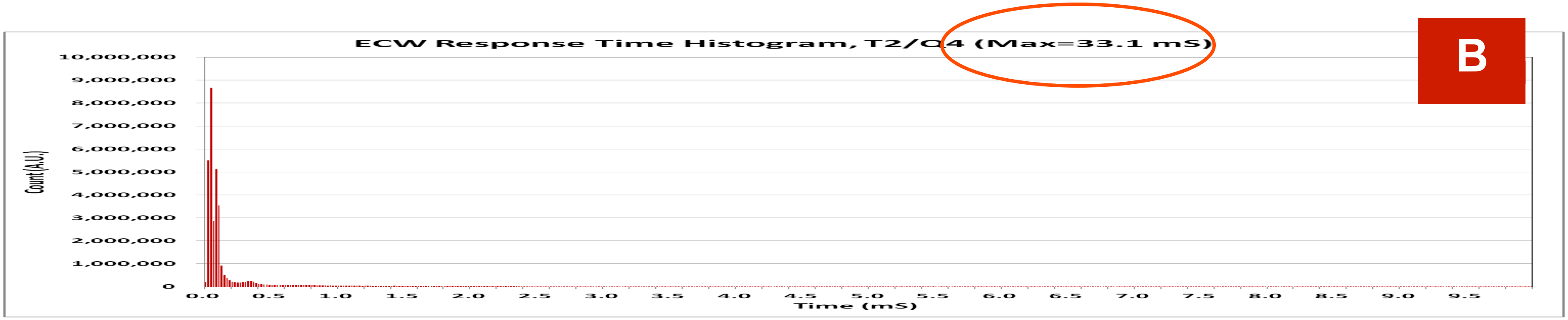
% of Access within 1 Measurement Period	Active Range Restriction	Label
50%	First 5%	LBA Group A
30%	Next 15%	LBA Group B
20%	Remaining 80%	LBA Group C

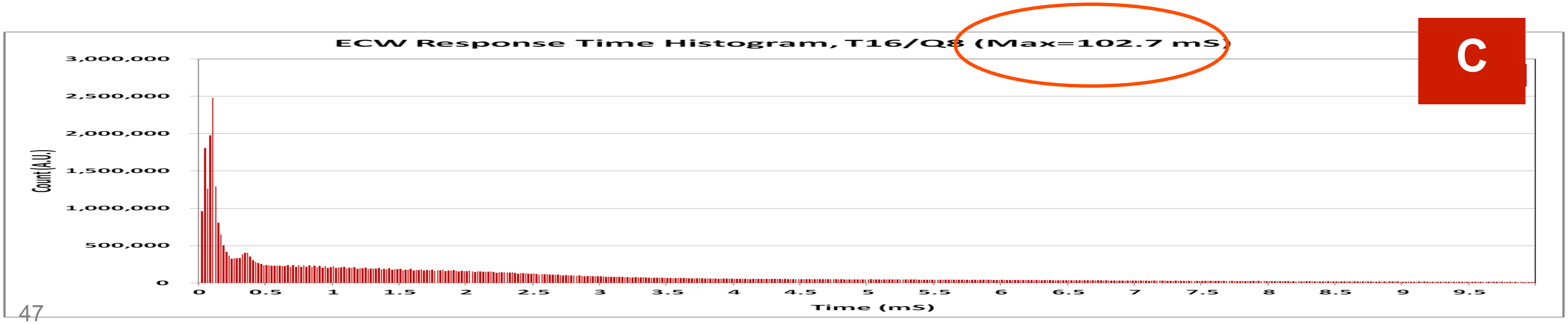
ECW: Demand Intensity (PCIe, MLC)



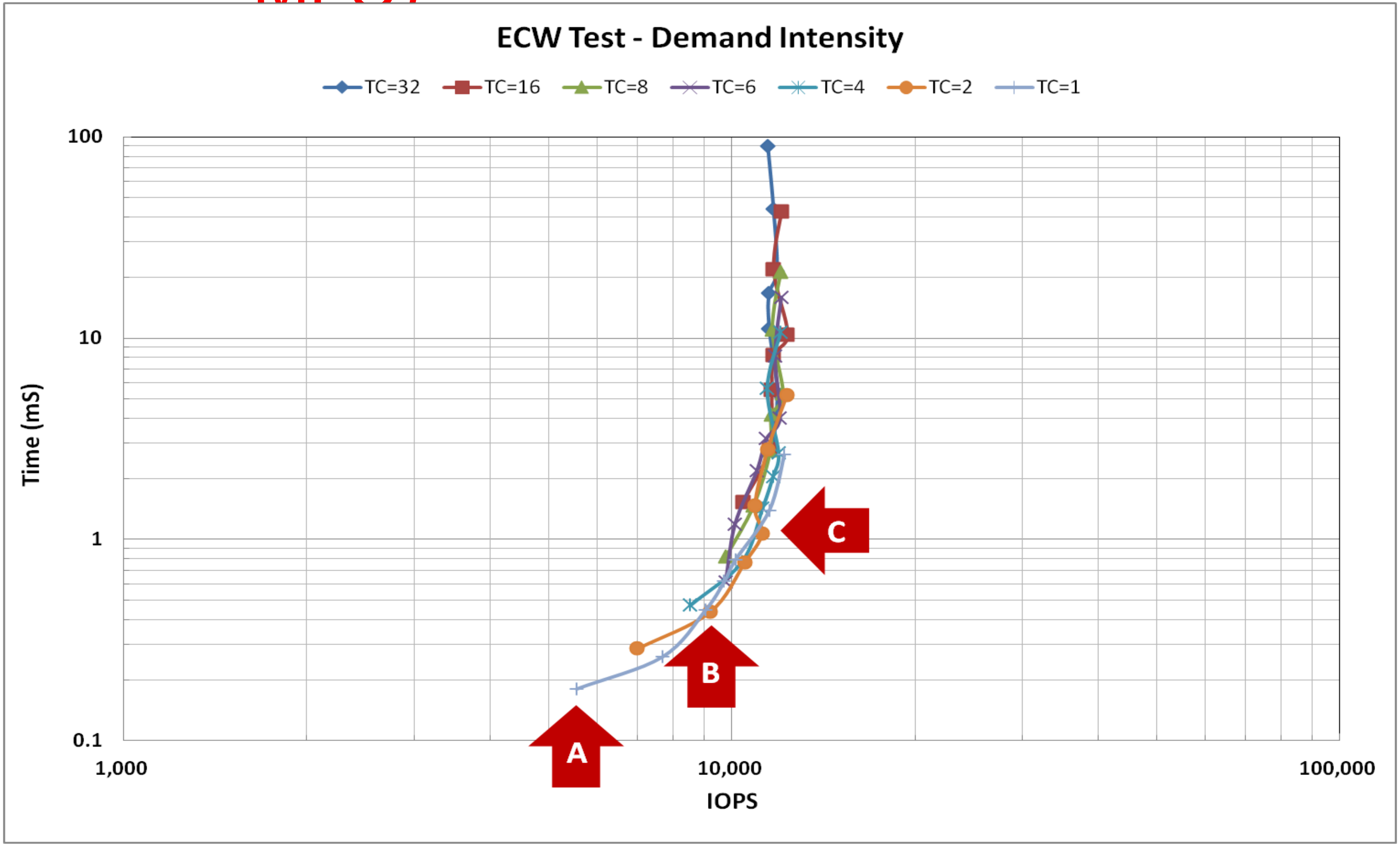
ECW: TRSP Histograms (PCIe,MLC)



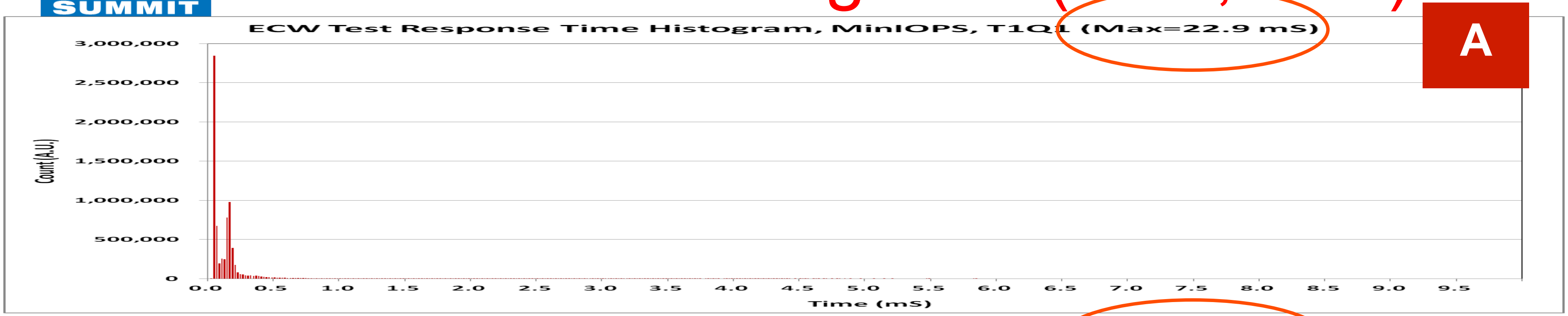


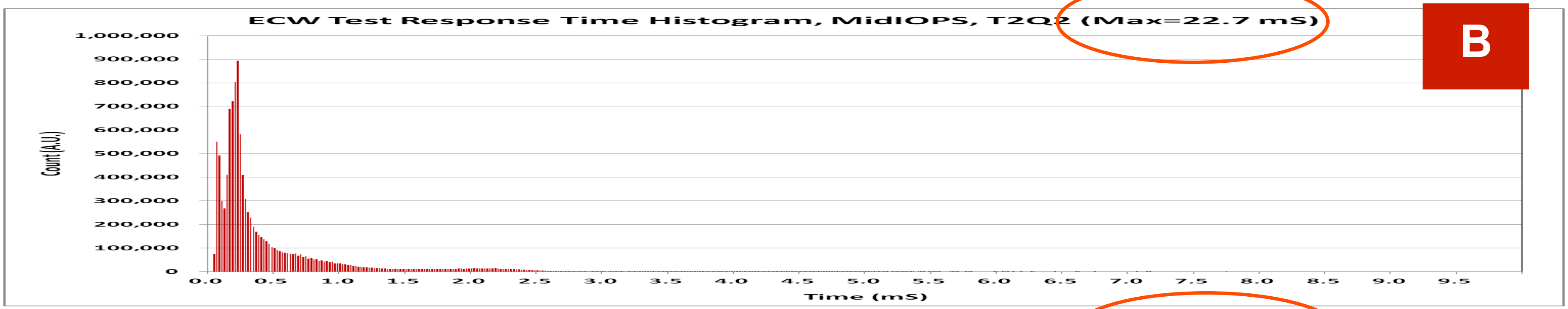


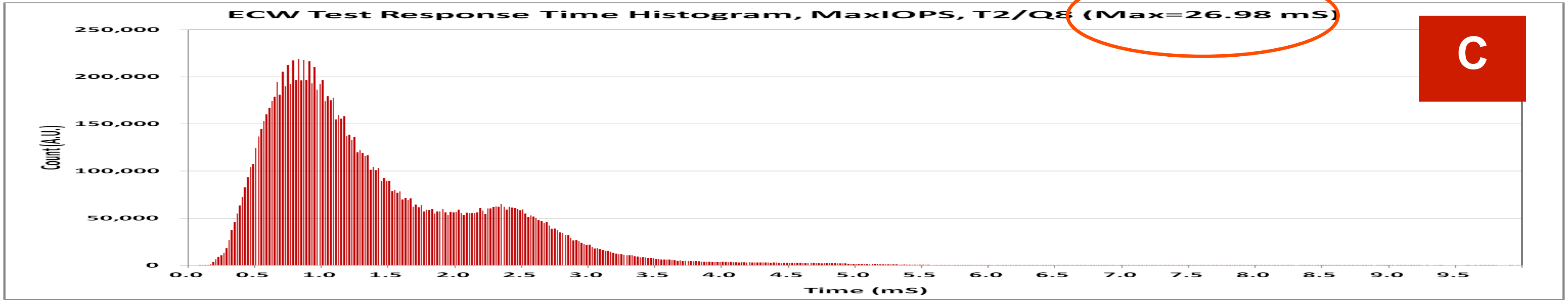
ECW: Demand Intensity (SAS, MLC)



ECW: T_{RSP} Histograms (SAS,MLC)







Summary

- Lack of standard test methodologies for SSS has driven the formation of SNIA's SSS TWG
- Reasonably wide participation and support for the SSS PTS TWG by member companies
- PTS-E 1.0 and PTS-C 1.0 have been released in 2011
- PTS-E 1.1 pending

Agenda

- Questions (10:40 – 10:50 am)