SNIA | NETWORKING
NSF | STORAGE

Optimizing NVMe over Fabrics Performance with Different Ethernet Transports: Host Factors

Live Webcast

September 16, 2020

10:00 am PT

Today’s Presenters

Tom Friend
Moderator
Illuminosi

Fred Zhang
Presenter
Intel

Eden Kim
Presenter
Calypso Systems

David Woolf
Presenter
University of New Hampshire

SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted
- 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
- Neither the authors nor the presenters are attorneys 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 authors' personal opinion and current understanding of the relevant issues involved
    - The authors, the presenters, 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

SNIA-At-A-Glance

SNIA. NSF | NETWORKING STORAGE

Technologies We Cover

- Ethernet
- iSCSI
- NVMe-oF
- InfiniBand
- Fibre Channel, FCoE
- Hyperconverged (HCI)
- Storage protocols (block, file, object)
- Virtualized storage
- Software-defined storage

SNIA NSF | NETWORKING STORAGE

SNIA NSF | NETWORKING STORAGE

Introduction to NVMe over Fabrics

David Woolf, University of New Hampshire

Agenda

- Recap of NVMe over Fabrics
- Factors Impacting Different Ethernet Transport Performance for NVMe over Fabrics
- Data Comparison of NVMe over Fabrics test with iWARP, RoCEv2 and TCP
- Conclusion

NVMe-oF What is it?

- NVMe is a storage protocol optimized for flash memory.
  - The transport used for SSDs is PCIe
- NVMe-oF maps that protocol onto fabric transports.
  - iWARP
  - RoCEv2
  - TCP
  - Fibre Channel
  - Infiniband

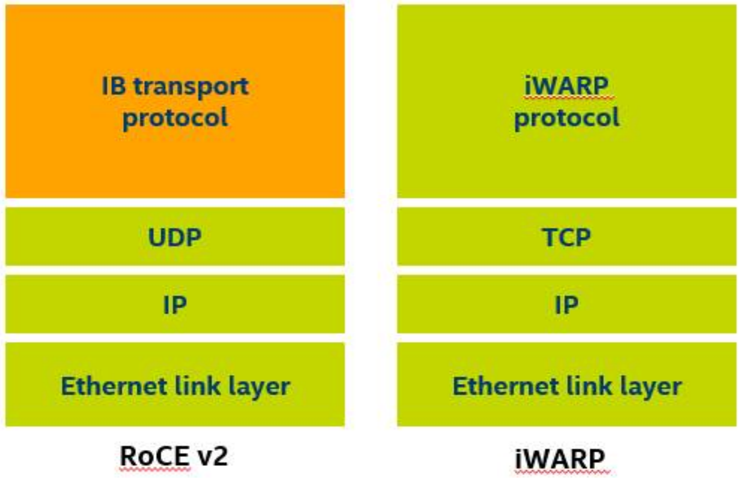
Ethernet based

We’ll focus our presentation here

SNIA. NSF | NETWORKING STORAGE

Know your NVMe-oF Transports: What's the difference?

- iWARP: RDMA over TCP
  - Low latency, Scales well on large datacenter networks
  - Requires iWARP capable adapters
  - Increase performance with TCP Offload
- RoCEv2: RDMA over Converged Ethernet
  - Low latency, best suited for Rack scale
  - Requires RoCE capable NICs/Switches
- NVMe/TCP: NVMe mapped directly on TCP
  - Low latency, Scales well on large datacenter networks
  - Easily supported on simple NICs/Switches
  - Increase performance with TCP Offload



SNIA. NSF | NETWORKING STORAGE

NVMe-oF: How Mature is It?

- Specification Maturity
  - NVMe-oF v1.1 specification released in October 2019
  - NVMe-oF v1.0 specification released in June 2016
- Driver support
  - Linux support for Initiator and Target drivers
  - Starwind produced NVMe-oF Windows Initiator
- Testing
  - Dozens of products on UNH-IOL Integrators List for NVMe-oF and continuing plugfests
- Many Ethernet based NVMe-oF product launches in 2020

Factors Impacting Different Ethernet Transport Performance

Fred Zhang, Intel

Scope of the Discussion of Factors Impacting NVMe-oF Performance: Host Factors

- Many factors impact NVMe over Fabric performance
  - On Host: CPU, NVMe drive
  - Switch: settings for congestion management
  - Network: over-subscribed, different fan-in ratio
- This presentation will focus on Host factors
  - Offload vs. non-offload
  - NVMe drive attributes
  - MTU (Maximum Transmission Unit or Frame)
  - Others, e.g. Workload, Thread Count, Queue Depth, pre-conditioning, test flow
- Focusing on 100GbE throughput
- Additional factors will be discussed in future topic:
  - Switch setup
  - Number of nodes in the storage network
  - Network topology
  - Congestion

Offload vs. Non-Offload

- RDMA is a host bypass and offload technology that need less CPU utilization
- Traditional TCP relies on protocol stack and consumes CPU cycles, especially for high speed Ethernet, e.g. 100GbE
  - New technologies are coming up to optimize on top of standard TCP stack to archive high performance with less CPU utilization, e.g. Application Device Queues
- Other software like SPDK work in user space, using polling, to get the high performance, with dedicated CPU cores
- Offloaded TCP engine will save CPU cycles and get to the desired performance with less CPU utilization

NVMe Drives with different Read/Write IOPs

- NVMe over Fabric performance is very much reliant on NVMe drives performance for directly attached storage target, especially when network speed moved up to 100GbE

- For storage target with head nodes connected to storage clients, or storage array on network, there are additional layers of complexity: RAID, QoS, etc.

- NVMe drive characteristics:

Manufacturer Specs	Random Reads	Random Writes
SSD 1	550K IOPS	550K IOPS
SSD 2	550K IOPS	250K IOPS
SSD 3	1M IOPS	130K IOPS
SSD 4	285K IOPS	41K IOPS

- Real world workload IO patterns are very different:
  - Heavy read/light write
  - Balanced read/write
  - Light read/heavy write
  - Small IO size vs. large IO size

SNIA. NSF | NETWORKING STORAGE

MTU – ~1500B vs. ~9000B

- MTU(Maximum Transmission Unit), in the context of Internet Protocol, is the maximum size of IP packet allowed without fragmentation
  - Default MTU is 1500B, Ethernet frame adds 18B, if not tagged, to make it 1518B maximum Ethernet frame
  - If Jumbo Frame is supported, maximum Ethernet frame can be 9000B
- Many modern OSes force the MTU transmission, regardless of upper level data size, for efficiency
- Caveat: Jumbo Frame needs to be enabled on both Initiator, Target and all network devices along the path

Test Comparison iWARP, ROCEv2 & TCP

Eden Kim, Calypso

PRELIMINARY REVIEW – Phase I Data v 1.0 – 09-14-2020

Discussion: This Preliminary Review is intended to report Phase I data on comparison testing of iWARP, ROCEv2 and TCP transports with MTU Regular 1500B v Jumbo 9000B frames and comparison of high performance six drive LUNs identified as SSD-1 and lower performance six drive LUN identified as SSD-2. Data summary slides are set forth in the presentation.

Test Runs: The first pass tests are comprised of six test runs, each of which has 4 tests. SSD-1 runs take 6 hours while SSD-2 take 9 hours. Variables are MTU setting [1500B v 9000B], Transport [iWARP v ROCEv2 v TCP] and SSD [SSD-1 v SSD-2]

Assumptions: The intent of testing is to measure differences in performance for iWARP, ROCEv2 & TCP
Each transport is tested at MTU 1500B & 9000B to determine if there is a difference
Each transport is also tested against Optane 6 drive LUN and NVME 6 drive LUN

Test Runs:

1. Synthetic RND 4K RW & SEQ 128K RW (corner case stress tests)
2. TC/QD Sweep using GPS Nav 9 IO Stream composite workload (to determine OIO saturation)
3. Replay test – Replay of GPS Nav real world workload (sequence of IO Streams & QDs)
4. Individual Streams test – Running each of the 9 IO Streams to Steady State

Note: 1st Pass Testing

Test results presented herein represent first pass testing to compare Transport, MTU settings and Comparison of SSDs. Data has not yet been repeated and further testing, adjustment of test settings and optimization of platforms is intended in Phase II. Accordingly, some data points may represent anomalous conditions and/or system or software that needs optimization.

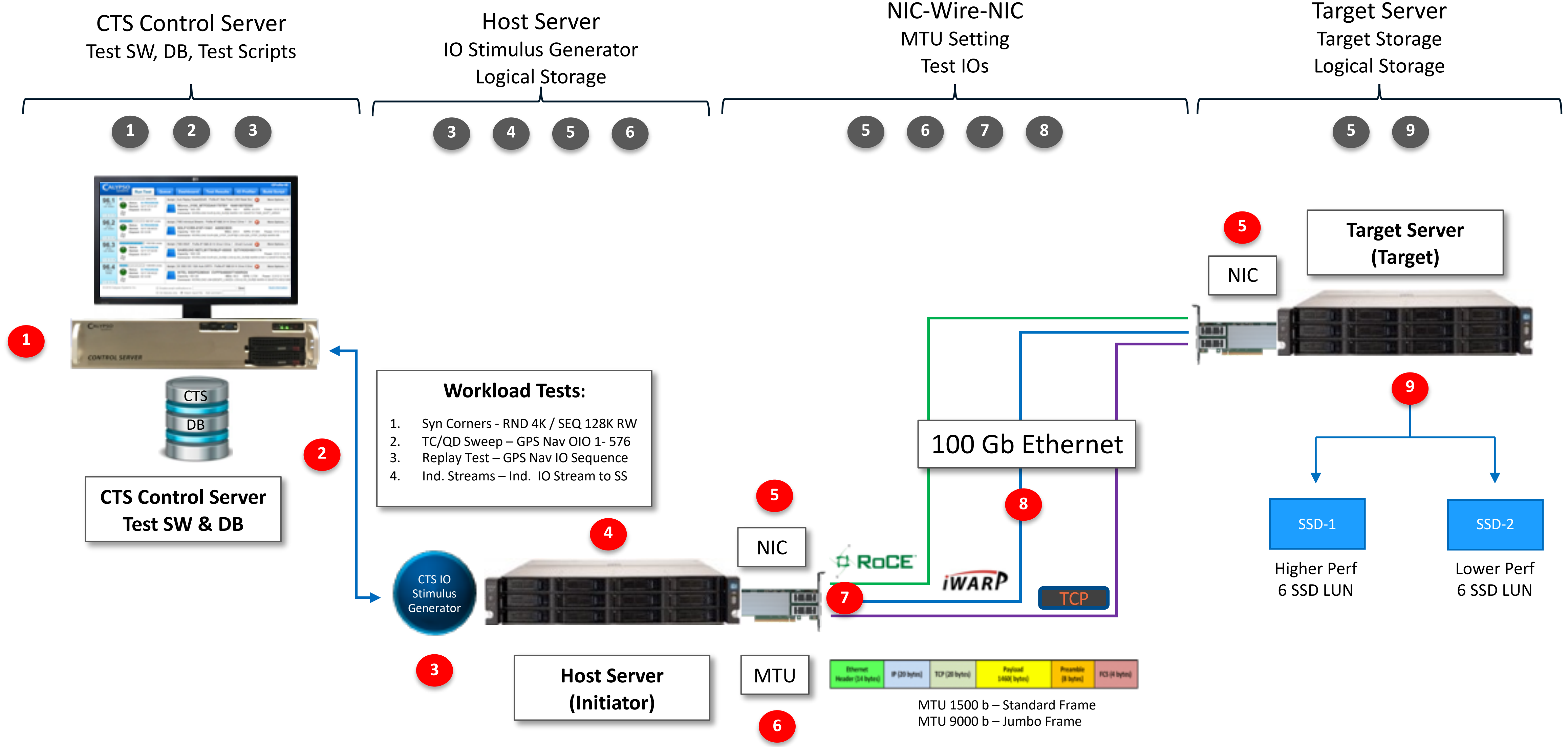
Test Set Up: Objectives, Test Platform & Test Settings

- Objectives: Saturate various Transports across 100Gb wire with different workloads & storage to compare performance
    1. Compare MTU IOPS & QoS: 1500 byte v 9000 byte with Synthetic 4K/128K
    2. Compare Transport: iWARP, RoCEv2, TCP with GPS Nav Demand Intensity & CPU Usage - TC/QD Sweep
    3. Compare Real World GPS Nav Portal Workload: Replay Test & Individual Streams
    4. Compare Drives: SSD-1 (high R/W IOPs) x6, SSD-2 (low R/W IOPs) x6

- Test Platform – High Performance (CPU, RAM):
    - CPU: Intel® Xeon® Platinum 8280L CPU @ 2.70GHz
    - Memory on Target: Two different types of RAM: 12 x 32GB @ 2933MHz; 12 x 256GB @ 2666MHz.
    - Memory on Initiator: 12 x 32GB @ 2933MHz
    - NIC – 100Gb – NIC, No Switch, Link Flow Control ON

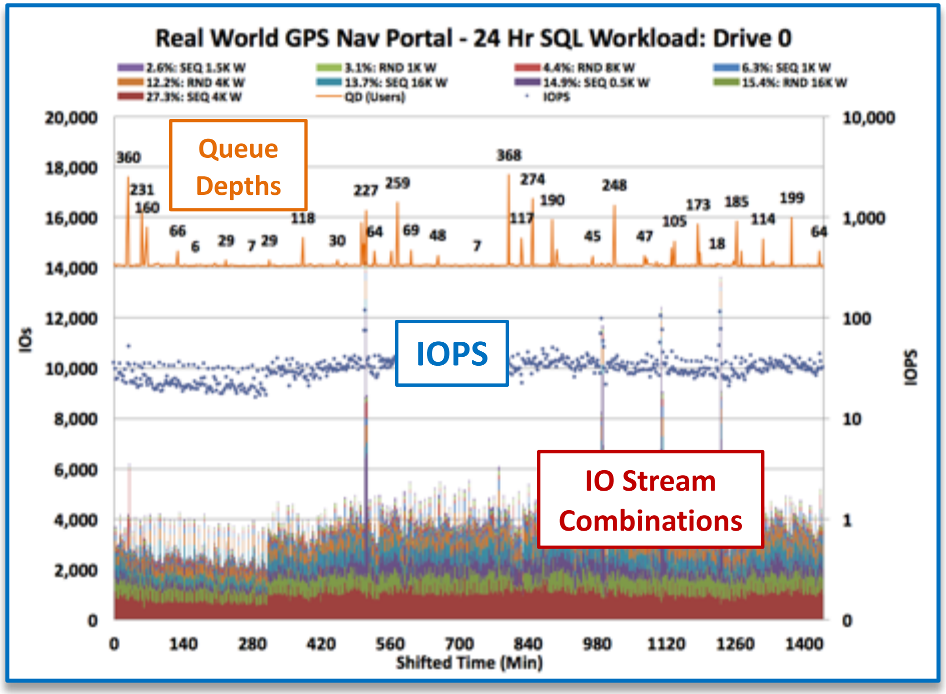
- Tests: Synthetic & Real World Workloads
    - Asynchronous IO libaio stimulus generator & IO Traffic
        - Calypso CTS Stimulus Generator is on the host initiator server and applies test IOs across 100Gb ethernet wire to target storage server
        - IO traffic on the 100Gb wire are test IOs. There is no other application or driver IO traffic during the test
    - Synthetic Single IO Stream Workloads
        - Synthetic RND 4K, SEQ 128K RW – T4Q32, Total OIO=128
        - GPS Workload 9 IO Streams - T4Q32, Total OIO=128
    - Real World GPS Nav Workload
        - Replay test - Replay observed combinations of IO Streams and QDs
        - TC/QD Sweep (DIRTH) – Fixed 9 IO Streams at TC/QD sweep 1 – 576

Phase I – Set-up and Test Plan



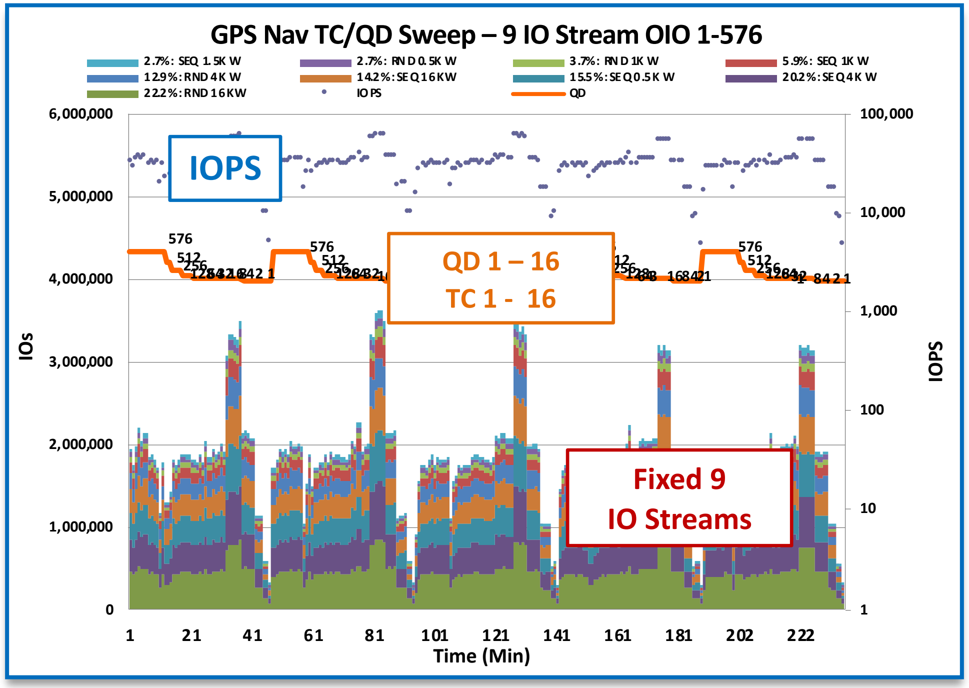
Workloads & Tests: Real World GPS Nav Portal & Synthetic Corner Case

Real World GPS Nav Replay Test, TC/QD Sweep Test, Synthetic Corner Case Test, GPS Nav Individual Streams Test



9 IO Streams = 78% of Total IO Streams
100% Write IOs
Replay Sequence & Combination IO Stream & QDs

Ave QD = 15
Median QD = 8
Max QD = 368



9 IO Streams = 100%
100% Write IOs
Fixed 9 IO Stream Composite for each step

QD Range = 1-36
TC Range = 1-36
Total Max OIO = 576

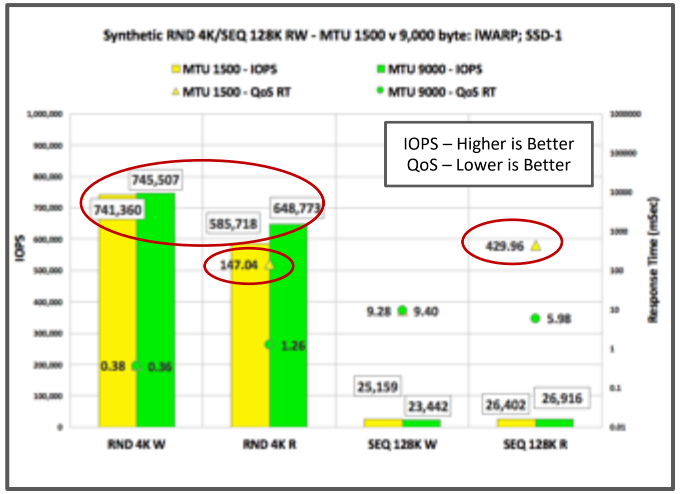
A0	GPS 9 IO IO Streams	
SEQ 4K W	24.8%	28.0
RND 16K W	17.7%	19.9
SEQ 16K W	14.4%	16.2
SEQ 0.5K W	13.9%	15.7
SEQ 1K W	5.8%	6.5
RND 4K W	4.5%	5.1
RND 1K W	2.65%	2.98
RND 28K W	2.54%	2.86
SEQ 1.5K W	2.44%	2.75

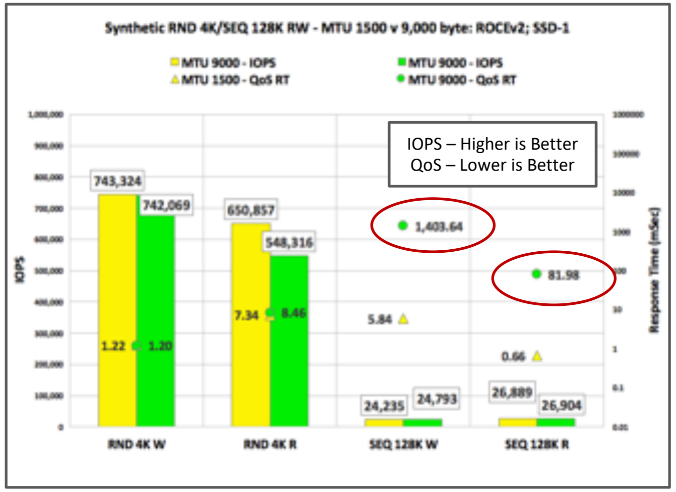
- **Workload Capture:** Each step of the workload has a different combination of the 9 IO Streams and QDs
- **Replay test**: Sequence and Combination of IO Streams and QDs are replayed
- **TC/QD Sweep test** - applies a fixed composite of all 9 IO Streams for each step of the test while running a range of TC/QDs from 1 - 16
- **Synthetic tests** - run a single, or few, IO Stream at a fixed QD and duration after pre-conditioning and steady state

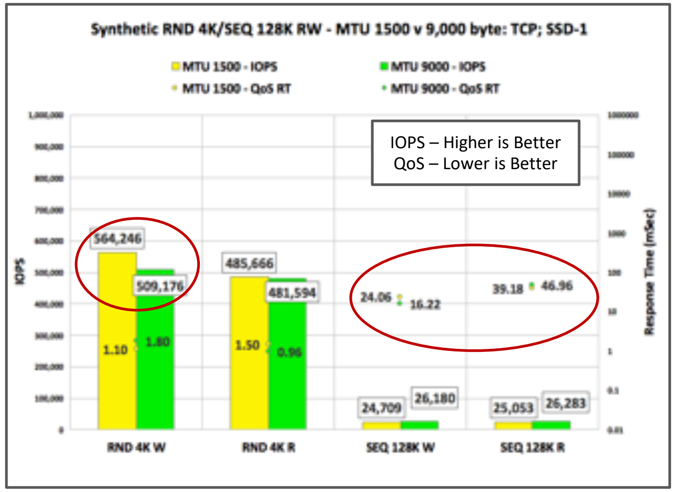
SNIA NSF | NETWORKING STORAGE

1. Compare MTU IOPS & QoS: 1500B v 9000B – SSD-1 - iWARP v ROCEv2 v TCP

- iWARP v ROCEv2 v TCP – 1500B v 9000B
  - Comparison Plot - Synthetic RND 4K RW; SEQ 128K RW







Note: Demand Intensity OIO=128 (T4/Q32)
Pre-conditioned to SNIA PTS Steady State
Each Workload Segment = 5 Minutes

Observations

IOPS – Substantially equivalent for all Transports
QoS – High QoS spikes are observed for Read workloads
CPU – CPU System Usage % are very low – typically less than 2%

SNIA. NSF | NETWORKING STORAGE

1a. MTU: 1500B v 9000B IOPS & QoS – SSD-1: All Workloads, iWARP

Drive Family | Auto-generated Multi WSAT IOPS TP ART 5 9s MRT Power | Auto DIRTH IOPS TP ART 5 9s | Auto Individual Streams IOPS TP ART 5 9s MRT Power | Replay Native Max

NVMeoFx6-1500

SSD-1
MTU 1500
iWARP
2250 GB

Compare IOPS

RND 4K/ SEQ 128K RW

R30.16-10016	T4Q32	WCE	PTS-E	DP=RND
Workload	RND 4KiB W	SEQ 128KiB W	SEQ 128KiB R	RND 4KiB R
IOPS	741,360	25,159	26,402	585,718
ART mSec	0.173	5.087	4.847	0.220
99.999% mSec	0.380	9.280	429.960	147.040

TC/QD Sweep

R30.16-10000	PTS-E
WCE	100% W
Max OIO	T9/Q16
IOPS	435,740
MB/s	2,905
ART	0.33
99.999%	1.000

9 Individual IO IO Streams

R30.16-10011	T4Q32	WCE	PTS-E	DP=RND	Align=4K				
Workload	SEQ 4KiB W	RND 16KiB W	SEQ 0.5KiB W	SEQ 16KiB W	RND 4KiB W	SEQ 1KiB W	RND 8KiB W	RND 1KiB W	SEQ 1.5KiB W
IOPS	769,132	199,003	105,445	190,379	742,917	154,495	366,534	1,232,129	113,124
ART mSec	0.166	0.643	1.214	0.672	0.172	0.828	0.349	0.104	1.152
99.999% mSec	0.300	2.000	2.000	1.400	1.100	1.000	1.900	0.900	1.500

Replay Test

R30.16-10001	T1Q12
Workload	Drive0 Cumulative Workload
IOPS	122,980
ART mSec	0.090
99.999% mSec	0.400

Compare QoS

NVMeoFx6-9000

SSD-1
MTU 9000
iWARP
2250 GB

Auto-generated Multi WSAT

R30.26-10024	T4Q32	WCE	PTS-E	DP=RND
Workload	RND 4KiB W	SEQ 128KiB W	SEQ 128KiB R	RND 4KiB R
IOPS	745,507	23,442	26,916	648,773
ART mSec	0.172	5.461	4.755	0.197
99.999% mSec	0.360	9.400	5.980	1.260

Auto DIRTH Drive0 Cumulative Workload

R30.26-10021	PTS-E
WCE	100% W
Max OIO	T9/Q16
IOPS	435,818
MB/s	2,909
ART	0.33
99.999%	1.450

Auto Individual Streams

R30.26-10023	T4Q32	WCE	PTS-E	DP=RND	Align=4K				
Workload	SEQ 4KiB W	RND 16KiB W	SEQ 0.5KiB W	SEQ 16KiB W	RND 4KiB W	SEQ 1KiB W	RND 8KiB W	RND 1KiB W	SEQ 1.5KiB W
IOPS	776,684	201,024	105,516	178,962	745,234	121,024	365,822	1,243,764	142,453
ART mSec	0.165	0.637	1.213	0.715	0.172	1.058	0.350	0.103	0.899
99.999% mSec	0.300	1.800	1.840	1.400	0.380	1.600	0.920	1.040	1.420

Replay Native Max

R30.26-10022	T1Q12
Workload	Drive0 Cumulative Workload
IOPS	122,154
ART mSec	0.090
99.999% mSec	0.400

Key Pts.

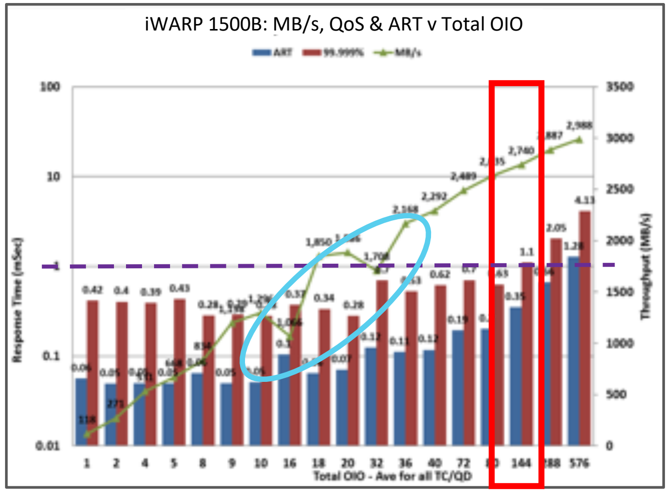
IOPS – Generally no significant difference
QoS – Higher RTs for RND 4K/SEQ 128K Reads

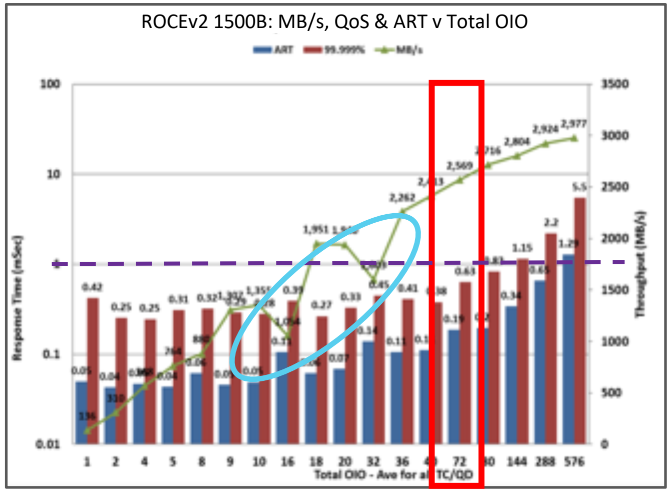
Note: Synthetic Demand Intensity = 128 (T4/Q32)
Replay Workload QD Varies between 7 – 368
TC/QD Sweep DIRTH QD range 1 – 576
Pre-conditioned to SNIA PTS Steady State

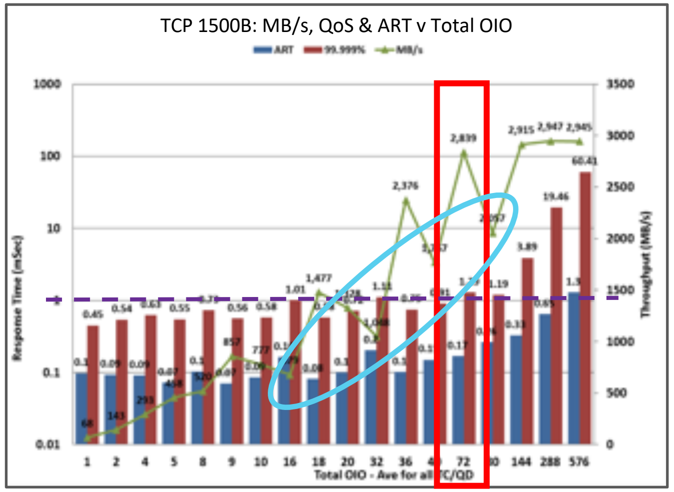
SNIA. NSF | NETWORKING STORAGE

2. Compare Transport: Demand Intensity TC/QD Sweep, SSD-1, iWARP v ROCEv2 v TCP

- TC/QD Sweep DIRTH – Demand Intensity comparison
  - Optimal Demand Intensity, MB/s, QoS comparison (red box) where QoS is the Figure of Merit
  - QoS Ceiling (purple dotted line), IOPS & MB/s dips (teal circle)







Observations

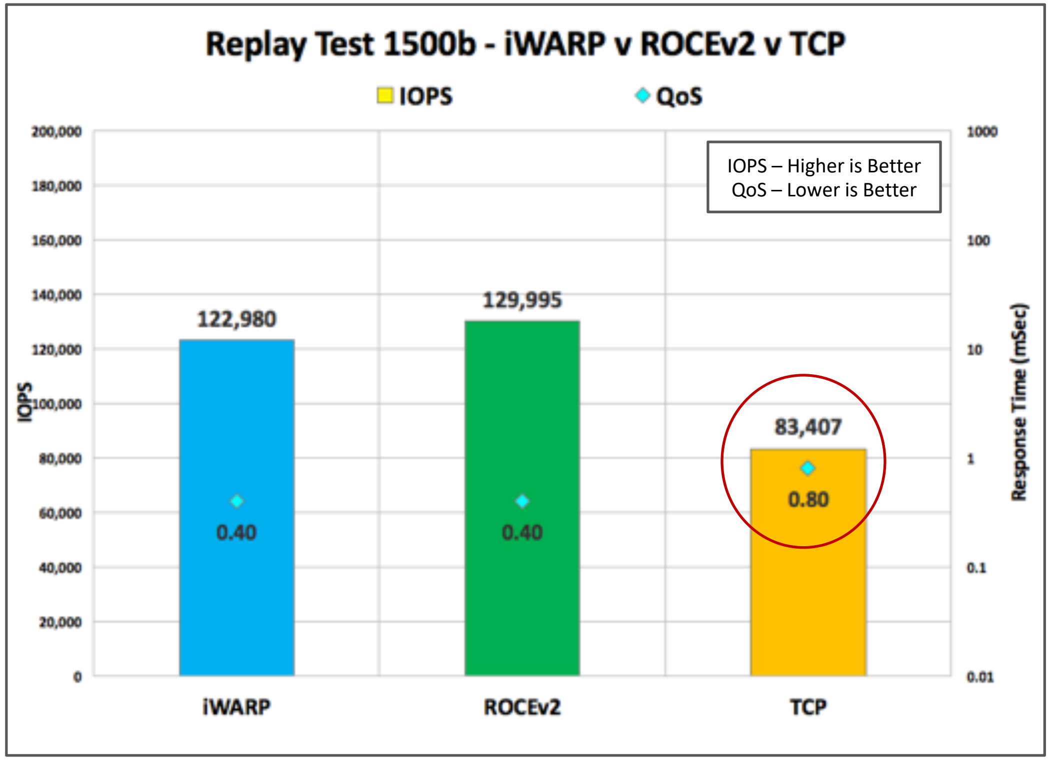
TC/QD test: MB/s & QoS are plotted as Demand Intensity increases over a range of OIO 1 – 576
Figure of Merit: QoS is just before QoS dramatically increases – here at OIO 72-144
Saturation: QoS saturates where RTs exceed 1 mS Ceiling – here at RDMA OIO 288 and TCP OIO 144
Note: RDMA continues to increase is MB/s with small increase in QoS whereas TCP has high QoS across all OIO

3. Compare Real World Workload GPS Nav: Replay Test – iWARP v ROCEv2 v TCP

Replay Test: Replays the Sequence of IO Stream combinations and QDs observed in the workload capture

iWARP v ROCEv2 v TCP

- iWARP & ROCEv2 substantially similar IOPS & QoS
- TCP has lower IOPS and higher QoS than iWARP & ROCEv2
- Note: Replay Test Values are average across total Replay Test. i.e. Variations due to different IO Stream combinations and QDs are averaged



GPS - 9 IO Streams

A0	GPS Nav Auto DIRTH	
SEQ 4K W	24.8%	28.0
RND 16K W	17.7%	19.9
SEQ 16K W	14.4%	16.2
SEQ 0.5K W	13.9%	15.7
SEQ 1K W	5.8%	6.5
RND 4K W	4.5%	5.1
RND 1K W	2.65%	2.98
RND 28K W	2.54%	2.86
SEQ 1.5K W	2.44%	2.75

Replay Test: 720 IO Capture Steps are applied after Steady State for a 1 min interval or 9 min duration

Key Pts.

iWARP & ROCEv2: Substantially similar IOPS & QoS
TCP: Lower IOPS & Higher QoS than iWARP & ROCEv2

3a. Compare GPS Nav: Individual Streams Test – iWARP v ROCEv2 v TCP

Individual Streams Test: Each IO Stream measured at SNIA PTS Steady State for 5 min

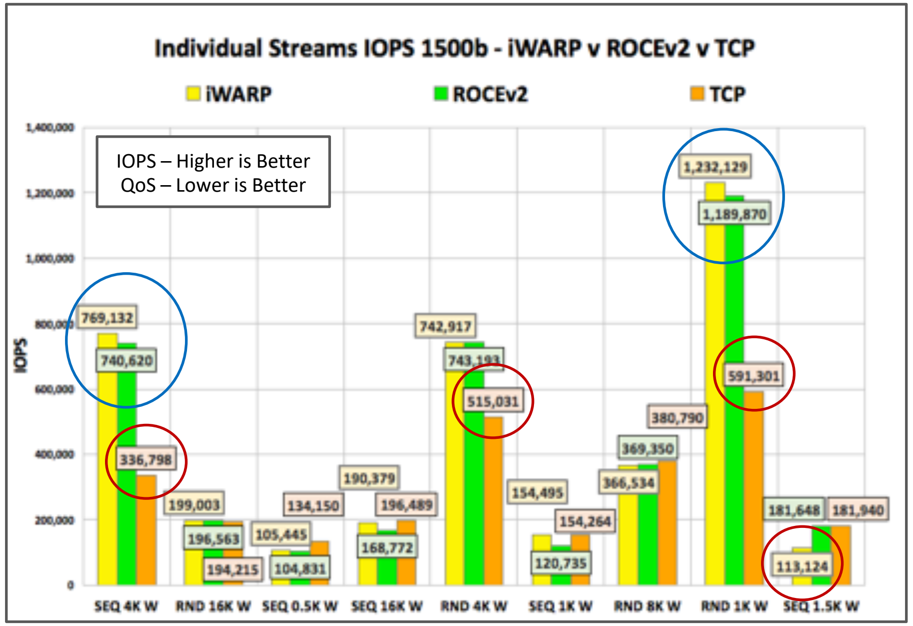
iWARP v ROCEv2 v TCP

IOPS:

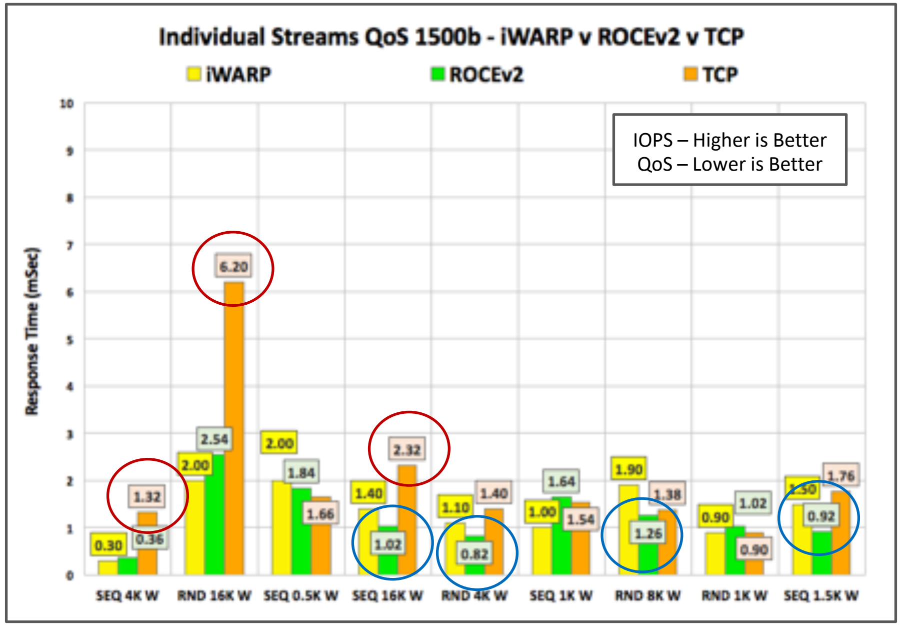
- TCP Lower for all block sizes than iWARP & ROCEv2
- For 0.5K, 1K, 1.5K – iWARP slightly higher than ROCEv2

QoS:

- ROCEv2 Lower for all block sizes than iWARP & TCP
- TCP higher for 4K, 16K



IOPS – Higher Value is Better



QoS – Lower Value is Better

Key Pts.

IOPS – iWARP generally higher ROCEv2, TCP Lower especially for smaller block sizes
QoS – ROCEv2 generally lower than iWARP. TCP higher for 4K & 16K block sizes
Note: Smaller block IO Streams negatively affect TCP IOPS & QoS, less impact on RDMA

4. Compare Drives: Synthetic RND 4K, SEQ128K RW - iWARP v ROCEv2 v TCP

- SSD-1 v SSD-2: RND 4K RW & SEQ 128K RW

iWARP

RND 4K/ SEQ 128K RW

NVMeoFx6-1500
SSD-1
MTU 1500
iWARP
2250 GB

R30.16-10016	T4Q32	WCE	PTS-E	DP=RND
Workload	RND 4KiB W	SEQ 128KiB W	SEQ 128KiB R	RND 4KiB R
IOPS	741,360	25,159	26,402	585,718
ART mSec	0.173	5.087	4.847	0.220
99.999% mSec	0.380	9.280	429.960	147.040

IOPS
QoS

NVMeoFx6-1500
SSD-2
MTU 1500
iWARP
24004 GB

Auto-generated Multi WSAT				
R30.16-10059	T4Q32	WCE	PTS-E	DP=RND
Workload	RND 4KiB W	SEQ 128KiB W	SEQ 128KiB R	RND 4KiB R
IOPS	564,849	17,654	26,659	571,311
ART mSec	0.227	7.364	4.801	0.224
99.999% mSec	7.560	35.680	7.220	0.820

IOPS
QoS

4K W | 128K W | 128K R | 4K R

ROCEv2

RND 4K/ SEQ 128K RW

NVMeoFx6-ROCE-1500
SSD-1
MTU 1500
ROCEv2
2250 GB

R30.36-10030	T4Q32	WCE	PTS-E	DP=RND
Workload	RND 4KiB W	SEQ 128KiB W	SEQ 128KiB R	RND 4KiB R
IOPS	743,324	24,235	26,889	650,857
ART mSec	0.172	5.281	4.760	0.197
99.999% mSec	1.220	7.340	5.840	0.660

NVMeoFx6-ROCE-1500
SSD-2
MTU 1500
ROCEv2
24004 GB

Auto-generated Multi WSAT				
R30.36-10067	T4Q32	WCE	PTS-E	DP=RND
Workload	RND 4KiB W	SEQ 128KiB W	SEQ 128KiB R	RND 4KiB R
IOPS	573,024	17,935	26,895	523,214
ART mSec	0.224	7.196	4.759	0.245
99.999% mSec	6.860	30.960	7.560	1.260

4K W | 128K W | 128K R | 4K R

TCP

RND 4K/ SEQ 128K RW

NVMeoFx6-TCP-1500
SSD-1
MTU 1500
TCP
2250 GB

R30.56-10051	T4Q32	WCE	PTS-E	DP=RND
Workload	RND 4KiB W	SEQ 128KiB W	SEQ 128KiB R	RND 4KiB R
IOPS	564,246	24,709	25,053	485,666
ART mSec	0.227	5.180	5.109	0.264
99.999% mSec	1.100	24.060	39.180	1.500

NVMeoFx6-TCP-1500
SSD-2
MTU 1500
TCP
24004 GB

Auto-generated Multi WSAT				
R30.56-10075	T4Q32	WCE	PTS-E	DP=RND
Workload	RND 4KiB W	SEQ 128KiB W	SEQ 128KiB R	RND 4KiB R
IOPS	487,366	11,395	24,949	413,774
ART mSec	0.263	11.256	5.130	0.309
99.999% mSec	2.820	42.620	56.660	1.480

4K W | 128K W | 128K R | 4K R

Key Pts.

IOPS – SSD-1 has higher or substantially equivalent IOPS compared to SSD-2
QoS – SSD-1 has lower or similar QoS compared to SSD-2 except for iWARP Reads

4a. TC/QD Sweep: SSD-1 v SSD-2, MTU 1500B, iWARP

Optimal OIO (red box), QoS Ceiling (purple dotted line), IOPS & MB/s dips (teal circle)

Optimal OIO = 144

- IOPS, ART, QoS increase to Max OIO
- SSD-1 - 2,740 MB/s, 1.1 mS QoS
- SSD-2 - 449 MB/s, 27.68 mS QoS

MB/s Dips

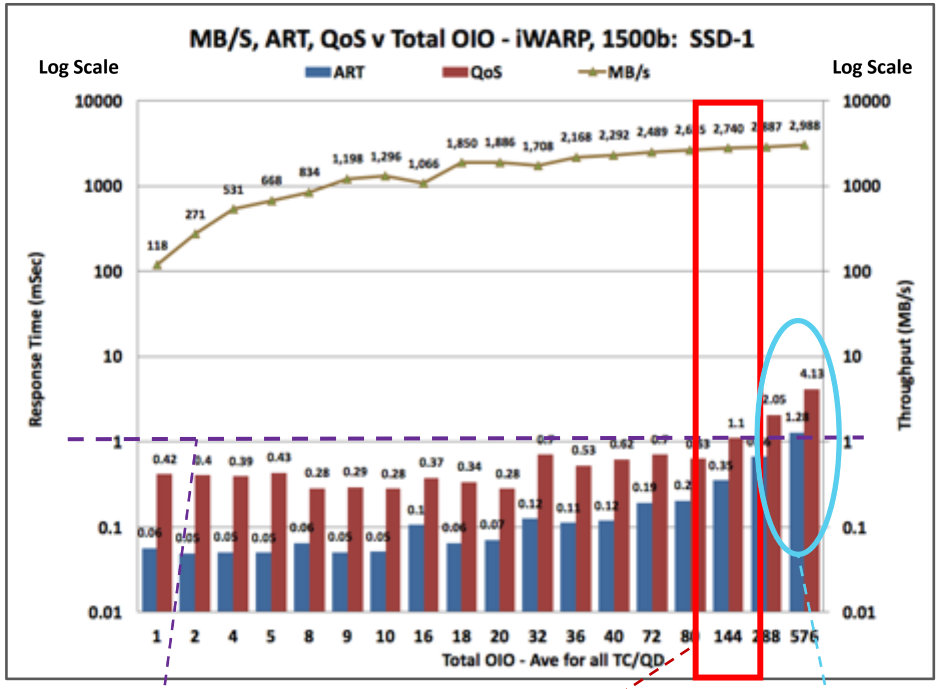
- SSD-1 - MB/s dip at OIO 16
- SSD-2 – MB/s dips at OIO 8,16,32,40

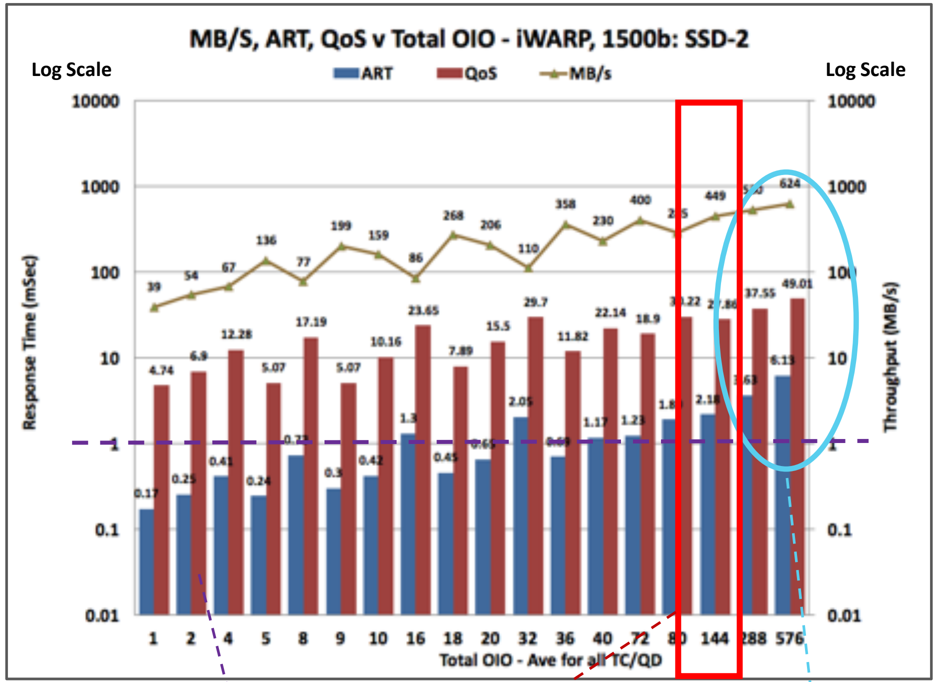
QoS Saturation

- SSD-1 - OIO=288 at 2.05 mS
- SSD-2 - OIO=288 at 37.55 mS

Max OIO=576

- SSD-1 - 2,988 MB/s, 4.13 mS QoS
- SSD-2 – 624 MB/s, 49.01 mS QoS





Response Time Ceiling – 1 mSec

Optimal OIO T9/Q16 = 144

Max OIO QoS Saturation

Response Time Ceiling – 1 mSec

Optimal OIO T9/Q16 = 144

Max OIO QoS Saturation

Key Pts.

IOPS & MB/s continue to increase at Max OIO for both SSDs (absence of plateau or leveling off)
SSD-1 shows higher MB/s and lower QoS. Max OIO=576 2,988 MB/s, 4.13 mS
SSD-2 shows lower MB/s and higher QoS. Max OIO=576 624 MB/s, 49.01 mS
NOTE: SSD-2 shows lower MB/s and higher QoS across all OIO

TC/QD Sweep test runs a fixed composite 9 IO Stream combination from the GPS Nav workload across a range from OIO=1 to OIO=576

SNIA NSF | NETWORKING STORAGE

4b. TC/QD Sweep: SSD-1 v SSD-2, MTU 1500B, ROCEv2

Optimal OIO (red box), QoS Ceiling (purple dotted line), IOPS & MB/s dips (teal circle)

Optimal OIO = 144

- IOPS, ART, QoS increase to Max OIO
- SSD-1 - 2,977 MB/s, 1.15 mS QoS
- SSD-2 - 427 MB/s, 29.52 mS QoS

MB/s Dips

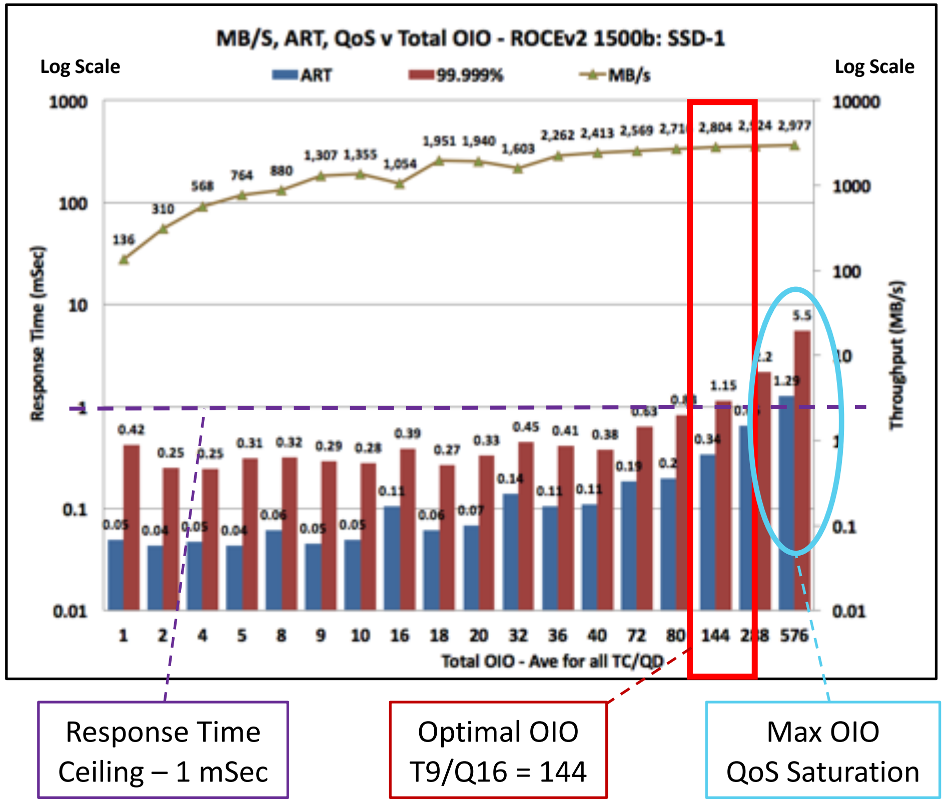
- SSD-1 - MB/s dip at OIO 16
- SSD-2 – MB/s dips at OIO 8,16,32,40

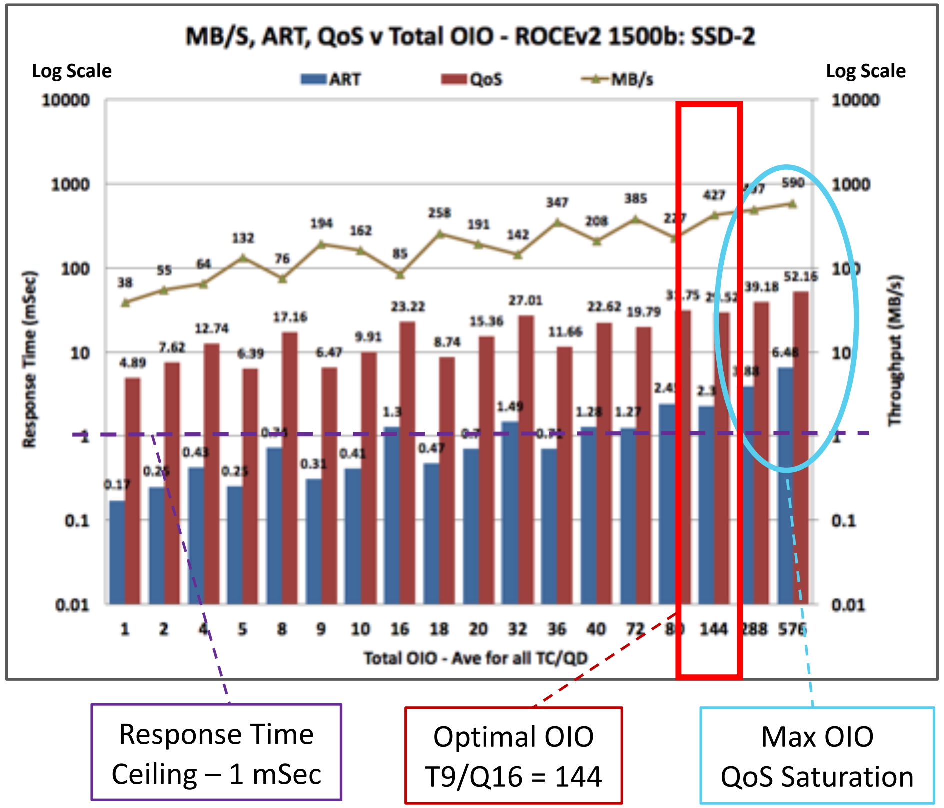
QoS Saturation

- SSD-1 - OIO=288 at 2.2 mS
- SSD-2 - OIO=288 at 39.18 mS

Max OIO=576

- SSD-1 - 2,977 MB/s, 5.5 mS QoS
- SSD-2 – 590 MB/s, 52.16 mS QoS





Key Pts.

SSD-1 ROCEv2 shows higher MB/s and lower QoS. Performance substantially similar to iWARP SSD-1
SSD-2 ROCEv2 shows lower MB/s and higher QoS. Performance substantially similar to iWARP SSD-1
NOTE: ROCEv2 SSD-2 shows lower MB/s and higher QoS across all OIO

TC/QD Sweep test runs a fixed composite 9 IO Stream combination from the GPS Nav workload across a range from OIO=1 to OIO=576

SNIA. NSF | NETWORKING STORAGE

4c. TC/QD Sweep: SSD-1 v SSD-2, MTU 1500B, TCP

Optimal OIO (red box), QoS Ceiling (purple dotted line), IOPS & MB/s dips (teal circle)

Optimal OIO = 72

- IOPS, ART, QoS increase to Max OIO
- SSD-1 - 2,839 MB/s, 1.29 mS QoS
- SSD-2 - 390 MB/s, 18.31 mS QoS

MB/s Dips

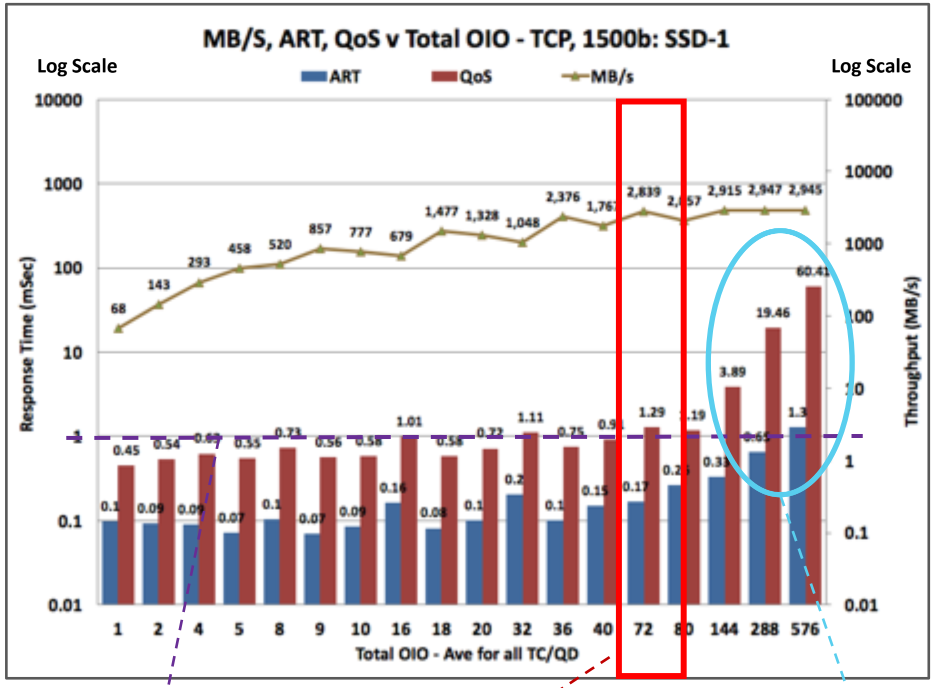
- SSD-1 - MB/s dip at OIO 16,32
- SSD-2 – MB/s dips at OIO 8,16,32,40

QoS Saturation

- SSD-1 - OIO=144 at 3.89 mS
- SSD-2 - OIO=80 at 28.9 mS

Max OIO=576

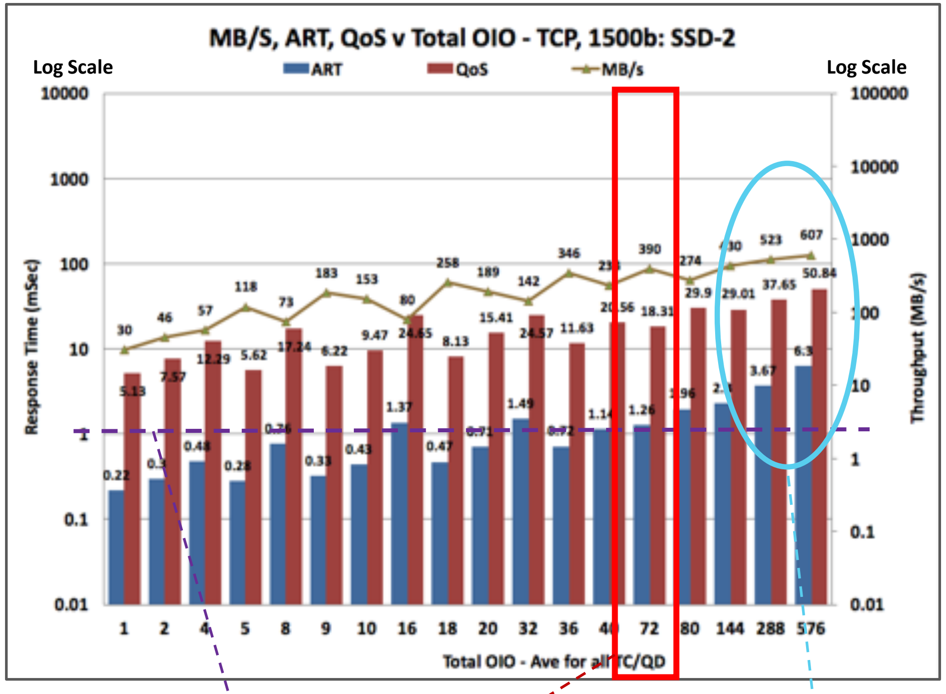
- SSD-1 - 2,945 MB/s, 60.41 mS QoS
- SSD-2 – 607 MB/s, 50.84 mS QoS



Response Time Ceiling – 1 mSec

Optimal OIO T9/Q8 = 72

Max OIO QoS Saturation



Response Time Ceiling – 1 mSec

Optimal OIO T9/Q8 = 72

Max OIO QoS Saturation

Key Pts.

SSD-1 TCP shows higher MB/s and lower QoS. Performance substantially Lower to RDMA SSD-1
SSD-2 TCP shows lower MB/s and higher QoS. Performance substantially Similar to RDMA SSD-1
NOTE: TCP SSD-2 shows lower MB/s and higher QoS across all OIO

TC/QD Sweep test runs a fixed composite 9 IO Stream combination from the GPS Nav workload across a range from OIO=1 to OIO=576

5. Test Data: Preliminary Observations & Conclusions

CPU: Overall low CPU utilization on storage target: 1%-4%, regardless of offload (iWARP, RoCEv2) or non-offload(TCP)

MTU: There appears to be nominal difference between 1500B & 9000B MTU
MTU 1500B has better response times for small block IO Streams (0.5K W, 1K W, 1.5K W)

SSD-1 v SSD-2: SSD-1 has significantly Higher IOPS, MB/s & Lower QoS than SSD-2
However, SSD-2 shows higher performance for small block 0.5K, 1K & 1.5K and SEQ 128K Reads
Note: additional testing is planned to validate findings, iterate test settings and conditions and to optimize test platform set up

TC/QD Sweep: GPS Nav Workload shows increasing IOPS & MB/s at highest OIO
However, ART and QoS increase at high OIO, probably higher than an acceptable application RT Ceiling
Synthetic workload OIO are up to 576 for TC/QD, 128 for RND 4K/SEQ 128K & Ind Streams, and 576 for Replay test.
Note: GPS Nav is a 100% W workload. Future tests should run a real world workload with 35% Reads (e.g. Retail Web Portal 68:32 RW)

Transport:

1. iWARP & ROCEv2 are substantially similar and faster than TCP
2. iWARP optimal OIO=144, ROCEv2 and TCP optimal OIO=72
3. OIO saturation occurs at 288 – 576 for GPS Nav 9 IO Stream workload
4. iWARP & ROCEv2 have lower QoS response times than TCP for GPS Nav workload
5. ROCEv2 has lower QoS response times than iWARP

SNIA. NSF | NETWORKING STORAGE

Questions?

- Please submit to Q & A panel during the session
- Otherwise, please submit to Brighttalk Portal
- Contact fred.zhang@Intel.com or info@calypsotesters.com

Thank You!

Additional SNIA Resources on NVMe-oF

- SNIAVideo “Intro to NVMe-oF” YouTube Playlist:
  - https://www.youtube.com/playlist?list=PLH_ag5Km-YUapfuug7nnwCpaeGJVO2DZE
- SNIA Educational Library:
  - https://www.snia.org/educational-library?search=NVMe+over+Fabrics

After this Webcast

- Please rate this webcast and provide us with your feedback
- This webcast and a copy of the slides will be available at the SNIA Educational Library https://www.snia.org/educational-library
- A Q&A from this webcast, including answers to questions we couldn't get to today, will be posted on our blog at https://sniansfblog.org/
- Follow us on Twitter @SNIANSF

Thank You