Table of Contents

White Paper

Supermicro Ready to Deploy Solutions for NVMe-oF TCP with Solarflare High Performance NICs

Supermicro Hardware Reference Architectures and Performance Evaluations

Executive Summary

- NVMe over Fabric (NVMe-oF) is gaining momentum for server shared networked storage. *NVMe-oF RDMA* (NVMe-oF RDMA over Converged Ethernet or NVMe-oF RoCE) is the current ratified Ethernet NVMe over Fabric standard. However this requires investment in RDMA compatible DCB switches and NICs.
- The new *NVMe-oF TCP* is the new technology which doesn't require special investment in a separate RDMA compatible storage hardware. NVMe-oF TCP (NVMe-oF TCP) is expected to be ratified in 2H 2018.
- We present several recommended configurations based on the Supermicro Ultra SuperServers and Solarflare NVMe-oF TCP optimized NICs that can deliver RDMA equivalent performance in both latency and IOPs.

Super Micro Computer, Inc.
980 Rock Avenue
San Jose, CA 95131 USA
www.supermicro.com

Introduction

Scope of this white paper

Supermicro and Solarflare collaborated to provide real-world performance evaluations of NVMe-oF TCP, and hardware configuration recommendations based on Supermicro's high performance and versatile 1U and 2U Ultra SuperServer platforms.

We present 3 different configurations optimized for varies uses cases in data centers. Each configuration is accompanied with storage throughput benchmarks.

Supermicro's First-to-the-Market Leadership in NVMe Technology

Supermicro was one of the earliest to recognize the significant impact that NVMe would make to modern data center workloads, and has been dedicated to be the first-to-the-market leader in NVMe and NVMe over Fabric technologies with the largest NVMe-ready product portfolio in the industry.

Solarflare selected Supermicro as their premier partner from the early product development stage for the availability of a wide range of NVMe enabled systems in various form factors, and flagship storage systems that can enable an extremely high drive density of up to 48 NVMe SSD drives in a compact 2U form factor.

Build Application Optimized Solutions with Supermicro Ultra and SuperStorage Systems with Solarflare NICs for NVMe over Fabrics

The Ultra SuperServer family offers unparalleled 2-socket performance in the industry supporting the highest performance Intel® Xeon® Scalable processors (up to 205W TDP) with maximum 24 DIMM slots in either 1U or 2U rackmount chassis.

Ultra systems are configurable with support of 20, 10, 4, or 2 hot-swappable NVMe drives that can be optimized for different types of deployment cases at the best cost effectiveness for the customer.

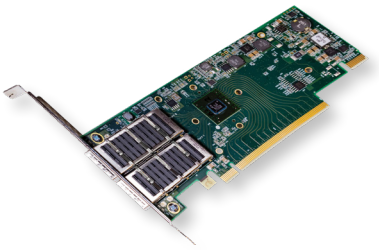
Ultra platforms can support not only a large number of NVMe drives, but also ample PCI-E lanes to accommodate network interface cards (NICs) to provide sufficient bandwidth for external access through fabrics.

NVMe-oF TCP: the Most Cost-Effective and Easiest to Deploy NVMe-oF Option

We will demonstrate in this white paper that with the same level of throughput performance as NVMe over RDMA, NVMe-oF TCP can be a highly cost effective and time-saving solution as one of the NVMe-oF options to consider.

For instance, NVMe-oF TCP can be easily deployed rapidly on existing TCP Ethernet switches and routers with minimal data center downtime.

Supermicro and Solarflare Introduce World's First NVMe-oF TCP Solutions Portfolio



Solarflare SFN8542 40Gb/s dual-port high performance NIC

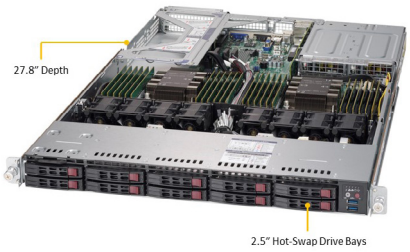
Solarflare Best NVMe-oF TCP Performance NIC

The XtremeScale NIC platforms are built on the XtremePacket Engine enabling thousands of flows/virtual NICs per adapter. The Solarflare architecture delivers sub microsecond latencies ensuring outstanding NVMe-oF TCP storage networking performance for highly demanding applications such as Online Transaction Processing and E-commerce.

Solarflare provides an end-to-end NVMe-oF TCP solution with Linux kernel drivers for both the storage initiator and the storage target.

For more information on Solarflare's innovative technology please visit:
http://solarflare.com/scale-out-flash-storage

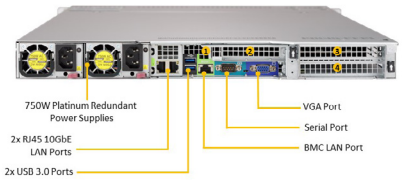
Use Cases	Supermicro System	Solarflare NIC	Performance	Benefits
Cost effective configuration for caching and application acceleration	SYS-1029U-T Series with 2 hot-swappable NVMe drive bays in 1U	1x dual port 40GbE NIC (SFN8542) or 3x dual port 10GbE NIC (SFN8522)	Network bandwidth can support full storage I/O throughput for up to 2 NVMe drives.	Users can utilize 2x NVMe drives as a high-speed caching pool and up to 8x SATA/SAS drives as the capacity pool for better application responsiveness.
Capacity and I/O performance balanced configuration for mainstream workloads	SYS-6019U-TN4 Series with 4 hot-swappable NVMe drive bays in 1U	2x dual port 40GbE NIC (SFN8542)	Network bandwidth can support full storage I/O throughput for up to 4 NVMe drives.	Up to 4x NVMe drives can be used for high-performance cache storage when combined with other caching or capacity oriented servers to create a flexible storage infrastructure over the network.
	SYS-2029U-T Series with 4 hot-swappable NVMe drive bays in 2U	2x dual port 40GbE NIC (SFN8542)	Network bandwidth can support full storage I/O throughput for up to 4 NVMe drives.	Up to 4x NVMe drives and 20x SATA/SAS drives in a 2U space.
All-flash NVMe systems for best storage density and highest performance	SYS-1029U-TN10 Series With 10 NVMe in 1U	2x dual port 40GbE NIC (SFN8542)	Up to 15GB/s thoughput and 3M IOPS NVMe-oF TCP performance with 10x NVMe in 1U	Up to 10x NVMe drives in 1U optimized for applications that are I/O bound.
	SYS-2029U-TN24 Series With 24 NVMe in 2U	2x dual port 40GbE NIC (SFN8542)	Up to 15GB/s thoughput and 3M IOPS NVMe-oF TCP performance with 24x NVMe in 2U	Deploy up to 24x NVMe drives for applications that require high-performance storage at a larger scale.



Slot	Description
0 ~ 7	2.5" Hot-Swap SATA3/SAS3* Drive Bays
8 ~ 9	2.5" Hot-Swap SATA3/SAS3*/NVMe(CPU1)** Drive Bays

*SAS3 support requires additional parts in optional parts list

**NVMe support requires additional parts in optional parts list



Slot	Description
1	PCI-E 3.0 x8 Internal LP Slot HL, 6.6" (CPU1)
2	PCI-E 3.0 x8 LP Slot LP Slot HL, 6.6" (CPU2)
3	PCI-E 3.0 X16 Slot FH, 10.5"L (CPU2)
4	PCI E 3.0 x16 Slot FH, 10.5"L (CPU2)

1U Ultra 2 NVMe

Detailed specifications of each server model shipped with different onboard networking modules can be found at: www.supermicro.com/ultra

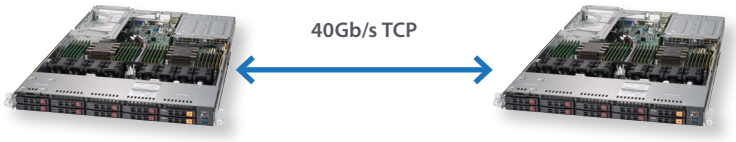
1U 2x NVMe Cost Effective Configuration for Caching and Application Acceleration

This reference configuration is based on the Supermicro Ultra SYS-1029U-T series that supports 2 front-loading hot-swappable NVMe drives and Intel® Xeon® Scalable processors.

Test results have shown that maximum transfer rate of NVMe over PCI-E x4 on both drives can be realized on the fabric through a single Solareflare 40Gb/s SFN8542 NIC utilizing both 40Gb/s ports.

Test Configurations

We built two systems with idential processors and memory based on the SYS-1029U-T series as storage target and initiator as the figure shown below:



Target

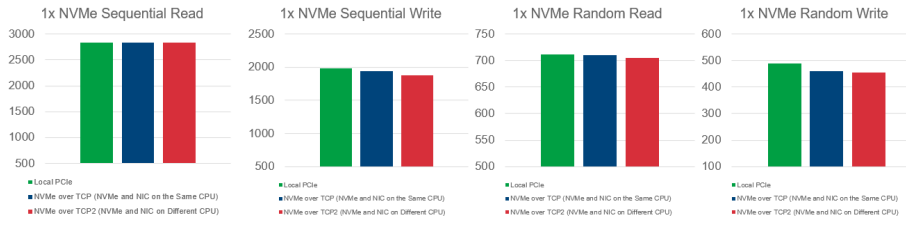
- SYS-1029U-T Series
- 1x Solarflare SFN8542 NIC
- 2x Intel® DC P4500 4TB NVMe SSDs
- 2x Intel Xeon Platinum 8160T
- 24x 32G DDR4-2666MHz RDIMMs

Initiator

- SYS-1029U-T Series
- 1x Solarflare SFN8542 NIC
- 2x Intel® Xeon® Platinum 8160T
- 24x 32G DDR4-2666MHz RDIMMs

1x NVMe Drive Over TCP Test Results

In this simple test with only one NVMe drive, we aim to investigate throughput characteristics by comparing locally accessed drives with accessing through NVMe-oF TCP, and also with the NIC and NVMe drive connected to the same or different CPUs.



		NVMe and SFC Both on Same CPU	NVMe on CPU1 and SFC on CPU2
1x NVMe	Local PCIe	NVMe over TCP	NVMe over TCP
SeqRead/128k/2nj/128iod	2840 MB/s	2840 MB/s	2829 MB/s
SeqWrite/128k/2nj/128iod	1983 MB/s	1938 MB/s	1876 MB/s
RandRead/4k/16nj/32iod	711KIOPs	710 KIOPs	705 KIOPs
RandWrite/4k/16nj/32iod	490KIOPs	461 KIOPs	455 KIOPs

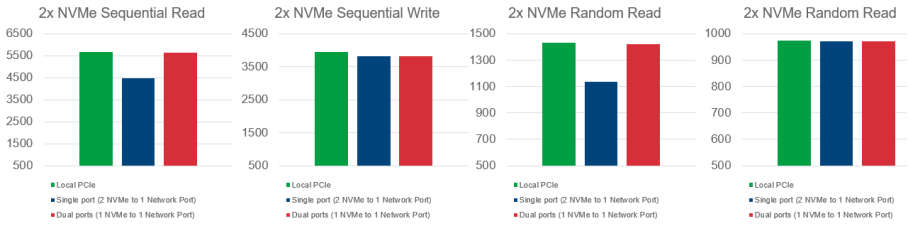
Test Parameter: Sequential Blocksize=128k, iodepth=128, threads=2, MTU=1500
Random R/W blocksize=4k, iodepth=32, threads=16, MTU=1500

Here are the key findings:

- NVMe-oF TCP can achieve the same level of throughput performance when compared with direct local access.
- With the NIC and the NVMe drive both connected to the same CPU and on separate CPUs interconnected by UPI, the performance difference is found to be negligible.
- This is due to the fact that 2 UPI links can support up to 41.6 GB/s, far more than 4 saturated PCI-E 3.0 lanes.

2x NVMe Drive Over TCP Test Results

Next, we examine how well Solarflare's SFN8542 NIC scales over multiple NIC ports. As shown in the previous test, one PCI-E x4 NVMe drive did not saturate the NIC, so in this test we performed test with 2 NVMe drives mapped to a single NIC port, compared with a 1:1 drive-to-port mapping setup.

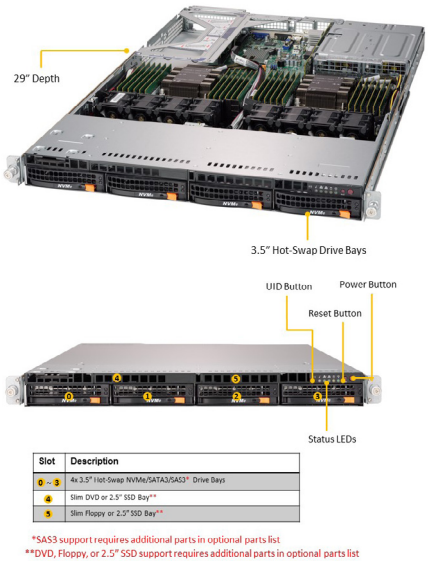


		NVMe and SFC Both on Same CPU	
		Single port (2 NVMe to 1 Port)	Daul Port (1 NVMe to 1 Port)
2x NVMe	Local PCIe	NVMe over TCP	NVMe over TCP
SeqRead/128k/2nj/128iod	5687 MB/s	4467 MB/s	5631 MB/s
SeqWrite/128k/2nj/128iod	3950 MB/s	3820 MB/s	3830 MB/s
RandRead/4k/16nj/32iod	1430KIOPs	1135 KIOPs	1422 KIOPs
RandWrite/4k/16nj/32iod	973KIOPs	971 KIOPs	971 KIOPs

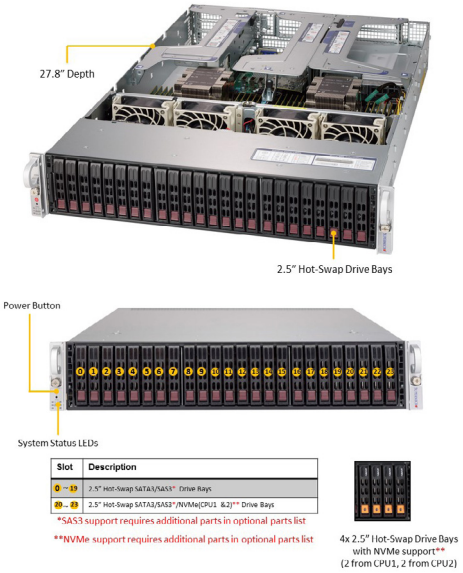
Test Parameter: Sequential Blocksize=128k, iodepth=128, threads=2, MTU=1500
Random R/W blocksize=4k, iodepth=32, threads=16, MTU=1500

Key Results Summary

- By assigning 2 NVMe drives to a single NIC port, the read throughput reached the port's maximum data rate of 4.5GB/s and 1.2M IOPS.
- When each NVMe drive is assigned to a dedicated NIC port, SFN8542 delivered maximum PCI-E x4 throughput of 2.8GB/s on each drive or 5.6GB/s total bandwidth over fabric.
- Since the UPI provide 41.6GB/s throughput, it is irrelevant which CPU that the 2x NVMe drives are placed.



1U Ultra 4 NVMe



2U Ultra 4 NVMe

Detailed specifications of each server model shipped with different onboard networking modules can be found at: www.supermicro.com/ultra

1U/2U 4x NVMe Capacity and I/O Performance Balanced Configurations for Mainstream Workloads

Supermicro Ultra SYS-6019U-TN4 series features 4 hot-swappable 3.5" SATA3/NVMe hybrid drive bays for extra flexibility designed for workloads that might require mixed warm and hot storage in a space-constrained environment.

Supermicro Ultra 2029U-T series is a 2U platform features 24 hot-swappable 2.5" SAS3 drive bays, with optional support of NVMe drives on 4 of the bays. The platform also offers more I/O expansion capabilities for higher degree of customizations to fit different applications.

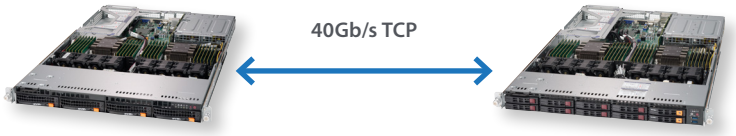
Supermicro Ultra 2029U-E1CR series is a 2U platform similar to the 2029U-T series that ships with SAS Expander and add-on-card.

All of the systems above can support up to 4 high performance NVMe devices, the 2U platforms can also support additional SATA3/SAS3 solid-state or spindle storage devices for data that are accessed less frequently.

We again recommend Solarflare's 40Gb/s dual-port SFN8542 NIC for this type of deployments since it scales well with the number of NVMe drives.

Test Configurations for 1x NIC and 4x NVMe Drives over TCP

We keep the initiator system unchanged and utilized SYS-6019U-TN4 series as the storage target with 4 NVMe drives installed. 2 NVMe drives are connected to CPU1 with the other two drives on CPU2. The NIC card is connected to CPU2.



Target

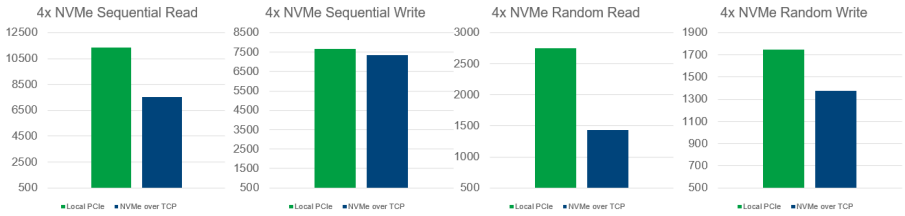
- SYS-6019U-TN4 Series
- 1x Solarflare SFN8542 NIC
- 4x Intel® DC P4500 4TB NVMe SSDs
- 2x Intel Xeon Platinum 8160T
- 24x 32G DDR4-2666MHz RDIMMs

Initiator

- SYS-1029U-T Series
- 1x Solarflare SFN8542 NIC
- 2x Intel® Xeon® Platinum 8160T
- 24x 32G DDR4-2666MHz RDIMMs

1x NIC and 4x NVMe Drive Over TCP Test Results

With a single 40Gb/s dual-port Solarflare NIC and a 1 ports to 2 drives mapping scheme, maximum storage throughput is capped at 7.5GB/s over TCP compared to 10GB/s when accessed locally.



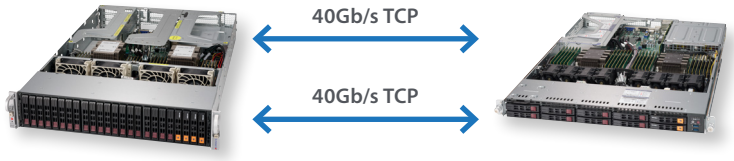
		2 NVMe on CPU1, 2 NVMe on CPU2, and SFC on CPU2
		Daul Port (2 NVMe to Port 1, 2 NVMe to port 2)
4x NVMe	Local PCIe	NVMe over TCP
SeqRead/128k/2nj/128iod	11372 MB/s	7501 MB/s
SeqWrite/128k/2nj/128iod	7658 MB/s	7340 MB/s
RandRead/4k/16nj/32iod	2747 KIOPs	1437 KIOPs
RandWrite/4k/16nj/32iod	1745 KIOPs	1375 KIOPs

Test Parameter: Sequential Blocksize=128k, iodepth=128, threads=2, MTU=1500
Random R/W blocksize=4k, iodepth=32, threads=16, MTU=1500

Next we will examine how the performance can be further scaled with 2x SFN8542 NICs.

Test Configurations for 2x NIC and 4x NVMe Drives over TCP

In this test, we doubled the number of NICs on the intiator and target systems and used the 2U SYS-2029U-T series system as the storage target shown below.



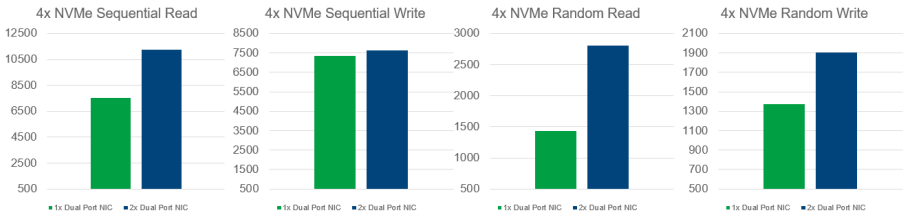
Target

- SYS-2029U-T Series
- 2x Solarflare SFN8542 NIC
- 4x Intel® DC P4500 4TB NVMe SSDs
- 2x Intel Xeon Platinum 8160T
- 24x 32G DDR4-2666MHz RDIMMs

Initiator

- SYS-1029U-T Series
- 2x Solarflare SFN8542 NIC
- 2x Intel® Xeon® Platinum 8160T
- 24x 32G DDR4-2666MHz RDIMMs

1x NIC and 4x NVMe Drive Over TCP Test Results



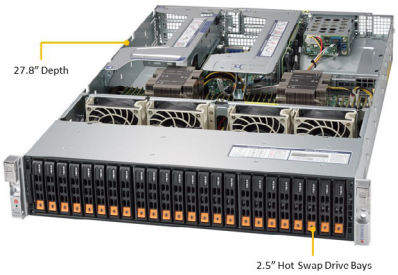
	2 NVMe on CPU1, 2 NVMe on CPU2, and SFC on CPU2	2 NVMe on CPU1, 2 NVMe on CPU2, and 2x SFC
	Daul Port (2 NVMe to Port 1, 2 NVMe to port 2)	2x Daul Port cards (1 NVMe to 1 Port)
4x NVMe	NVMe over TCP	NVMe over TCP
SeqRead/128k/2nj/128iod	7501 MB/s	11260 MB/s
SeqWrite/128k/2nj/128iod	7340 MB/s	7646 MB/s
RandRead/4k/16nj/32iod	1437 KIOPs	2806 KIOPs
RandWrite/4k/16nj/32iod	1375 KIOPs	1906 KIOPs

Test Parameter: Sequential Blocksize=128k, iodepth=128, threads=2, MTU=1500
Random R/W blocksize=4k, iodepth=32, threads=16, MTU=1500

Key Results Summary

- With a single dual port Solarflare NIC, we can max out both network ports with 4 NVMe drives. In practice, deploying a single NIC for 4 drives may be an option for some appliactions since it can still support very high thoughput on read performance of 7.5GB/s and around 1.5M IOPS.
- The locally accessed sequential write throughput is found to be 7.5GB/s, we found almost identical performance using either one or two NICs.
- When 2 NICs are deployed in the system, NVMe-oF TCP provides the same level of performance as locally accessed drives.
- The 4x NVMe drives with 2x NICs configuration can provide up to 11.2GB/s thoughput and 2.8M IOPS over the fabric.

All-flash NVMe Systems for Best Storage Density and Highest Performance



Supermicro Ultra SYS-2029U-TN24R4T system is design for applications that require higher NVMe storage density with exceptional processing performance at optimal balance.

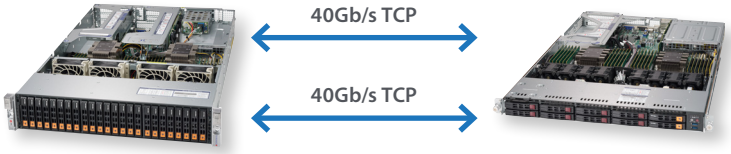
In order to support 24 NVMe drives or 4x 24 PCI-E 3.0 lanes across two CPUs, the SYS-2029U-TN24R4T system incorporates dual PLX switches that are placed between each CPU and the storage blackplane through a total of 16 PCI-E 3.0 lanes. Each CPU is also connected to a Solarflare 40Gb/s SFN8542 NIC through full x16 lanes, hence ensuring that maximum performance can be obtained in this type of configuration.

Test Configurations for 2x NIC and 4x NVMe Drives over TCP

In this test, we populated 24 NVMe drives on the target system, and installed two SFN8542 NICs on the system for a total of 4x 40Gb/s ports.

Slot	Description
0 ~ 11	12x 2.5" Hot-Swap NVMe Drive Bays (CPU1)
12 ~ 23	12x 2.5" Hot-Swap NVMe Drive Bays (CPU2)
19 ~ 23	4x 2.5" Hot-Swap Drive Bays with SAS3/SATA3 support*

*SAS3/SATA3 support requires additional parts in optional parts list

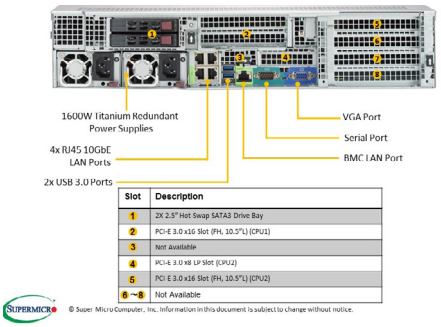


Target

- SYS-2029U-TN24R4T
- 2x Solarflare SFN8542 NIC (one NIC on each CPU)
- 24x Intel® DC P4500 4TB NVMe SSDs
- 2x Intel Xeon Platinum 8160T
- 24x 32G DDR4-2666MHz RDIMMs

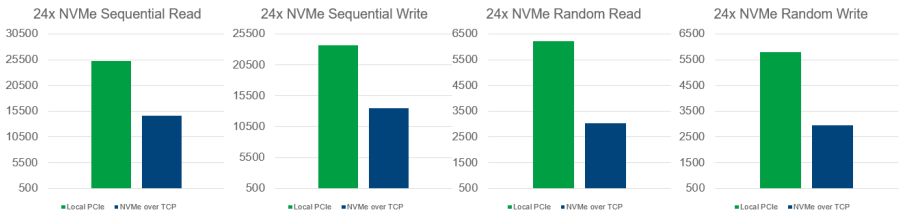
Initiator

- SYS-1029U-T Series
- 2x Solarflare SFN8542 NIC
- 2x Intel® Xeon® Platinum 8160T
- 24x 32G DDR4-2666MHz RDIMMs



Slot	Description
1	2X 2.5" Hot Swap SATA3 Drive Bay
2	PCI-E 3.0 x16 Slot (FH, 10.5"L) (CPU1)
3	Not Available
4	PCI-E 3.0 X8 LP Slot (CPU2)
5	PCI E 3.0 x16 Slot (FH, 10.5"L) (CPU2)
6 ~ 8	Not Available

2x NIC and 24x NVMe Drive Over TCP Test Results



		12 NVMe on CPU1, 12 NVMe on CPU2, and 1x SFC on each CPU
		2x Daul Port (12 NVMe to NIC1, 12 NVMe to NIC2)
4x NVMe	Local PCIe	NVMe over TCP
SeqRead/128k/2nj/128iod	25120 MB/s	14656 MB/s
SeqWrite/128k/2nj/128iod	23677 MB/s	13450 MB/s
RandRead/4k/16nj/32iod	6206 KIOPs	3036 KIOPs
RandWrite/4k/16nj/32iod	5792 KIOPs	2960 KIOPs

Test Parameter: Sequential Blocksize=128k, iodepth=128, threads=2, MTU=1500
Random R/W blocksize=4k, iodepth=32, threads=16, MTU=1500

Detailed specifications of each server model shipped with different onboard networking modules can be found at: www.supermicro.com/ultra

Key Results Summary

- This configuration is optimized for workloads with larger dataset in environments where a balanced storage density and performance is considered the top priority.
- The system can provide up to 15GB/s thoughput and 3.1M IOPS over fabric.

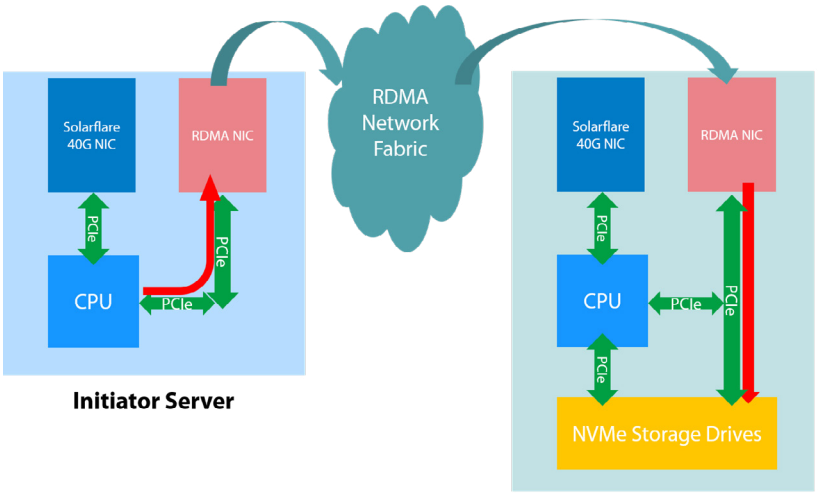
Latency Comparison of TCP and RDMA of NVMe-oF

Latency is considered a key performance indicator for OLTP applications such as SAP HANA, Memcached, SQL Server and etc. The lowest possible latency is when NVMe drives are attached directly to local CPU's PCI-E lanes. We use this as the baseline of our benchmaks.

It is common to assume that NVMe over RDMA (NVMe-oF RDMA over Converged Ethernet or RoCE) solution should deliver much lower latency from by-passing the OS kernel and CPU as shown in the figure below:

Dataflow Diagram: NVMe over RDMA

- Requires purchasing of dedicated RDMA Ethernet switches
- Data flows from/into NVMe drives directly through a RDMA NIC without going through local processor and OS kernel process.
- Commonly believed for having a superior latency performance than a NVMe-oF TCP architecture



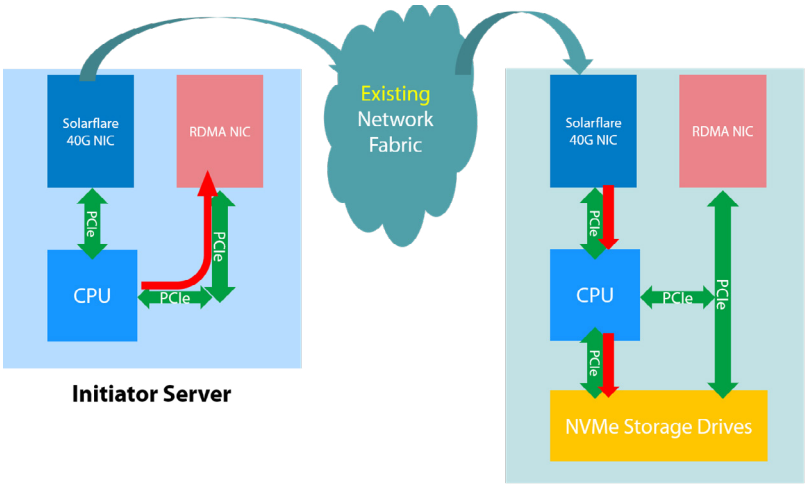
Initiator Server

Storage Target Server

However, the test results have shown that a NVMe-oF TCP solution can deliver comparable latency performance. The figure below illustrates the data flow on a NVMe-oF TCP solution.

Dataflow Diagram: NVMe-oF TCP

- Simply install a NIC and the system is ready for NVMe-oF TCP in the existing networking infrastructure
- Data flows from/into NVMe drives through CPU and NVMe-oF TCP supported NIC
- Significant cost savings compared to NVMe over RDMA



Initiator Server

Storage Target Server

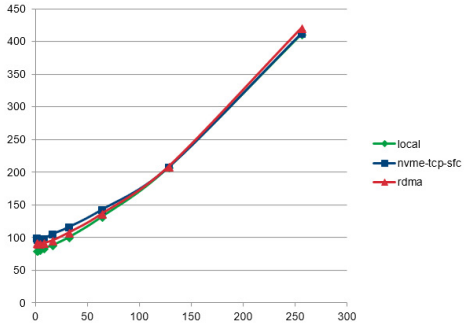
Latency Benchmark

We ran FIO utility for latency testing with different concurrent thread count from 1 to 256 for both random read and random write.

Random Read Latency Results

With a small number of threads, the overall latency with NVMe-oF TCP is only around 7% higher then RDMA; With a larger number of threads typically seen in production environments, NVMe-oF TCP is almost the same as RDMA on read latency performance.

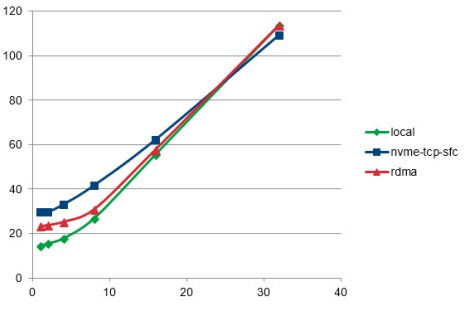
# of Threads	Local PCI-E	NVMe over TCP	RDMA
1	79.72	99.31	92.13
2	80.01	96.6	90.95
4	80.78	96.85	90.95
8	83.38	98.38	92.54
16	88.1	105.66	95.74
32	100.6	115.98	108.1
64	132.25	142.65	135.95
128	207.19	207.89	208.81
256	411.39	412.75	420.53



Random Write Latency Results

With a small number of threads, the overall latency with NVMe-oF TCP is around 20% higher then RDMA; With a larger number of threads typically seen in production environments, again NVMe-oF TCP is almost the same as RDMA on write latency performance.

# of Threads	Local PCI-E	NVMe over TCP	RDMA
1	14.54	29.9	23.21
2	15.43	29.89	23.66
4	17.75	33.02	25.34
8	26.9	41.5	30.96
16	55.5	62.31	57.75
32	113.6	109.38	113.76



Latency Performance Summary

The results have proven that NVMe-oF TCP matches RDMA's latency performance when handling a large number of concurrent processes, which resembles real-world applications.

Also, NVMe-oF TCP is more cost economic to deploy compared to NVMe over RDMA. You can simply install a regular Solarflare TCP NIC, and use existing switch infrastructure for NVMe-oF TCP.

Conclusions

In this white paper we presented serveral performance validated configurations for different types of workloads based on Supermicro's Ultra SuperServers and Solarflare's SFN8542 dual-port 40Gb/s NIC solution.

The benchmark results have shown that NVMe-oF TCP can bring significant cost savings from the utilization of an exsiting networking infrastructure, while delivering NVMe-oF RoCE equivalent latency performance and IOPS performance equivalent or similar to local attached storage.

NVMe-oF TCP (NVMe-oF TCP) standard is expected to be ratified in the second half of 2018. Solarflare is on standards committee along with Intel, VWware, Facebook and other industry vendors.

About Super Micro Computer, Inc.

Supermicro® (NASDAQ: SMCI), the leading innovator in high-performance, high-efficiency server technology is a premier provider of advanced server Building Block Solutions® for Data Center, Cloud Computing, Enterprise IT, Hadoop/Big Data, HPC and Embedded Systems worldwide. Supermicro is committed to protecting the environment through its "We Keep IT Green®" initiative and provides customers with the most energy-efficient, environmentally-friendly solutions available on the market.

Learn more on www.supermicro.com

Printed in USA ♻ Please Recycle 14_Ultra-SolarFlare_180801_FA