SOLUTION BRIEF

Solid State Drives
NVMe* over Fabrics
Intel® Data Center Builders

Accelerated SSD Infrastructure for the Cloud

Learn how to share NVMe* using NVMe over Fabrics with a collaborative solution from Attala Systems, Supermicro, and Intel delivering local NVMe performance with cloud elasticity that scales to support multiple big data customers running data-intensive workloads such as Hadoop* or Spark*.

ATTALA SYSTEMS

Solution Summary

Solid state drives (SSDs) are commonly featured in cloud storage offerings. Further differentiation is now possible as non-volatile memory express (NVMe) standards have arrived to streamline the interface between the user and storage device. The result is higher throughput and lower latency. The performance available to direct-attached PCIe flash is welcome, but cloud total cost of ownership (TCO) is better when storage is shared to many hosts while offering the elasticity to meet dynamic tenant demands.

The maturing of the NVMe over Fabrics (NVMe-oF) specification enables local NVMe performance over a network. With remote direct memory access (RDMA), NVMe actions and data can be transported over a fabric such as Ethernet.

Flash storage arrays integrating the intelligence to handle NVMe-oF to deliver high-performance IO can address data center needs of today and into the future. Intel is working with Attala Systems and Supermicro to demonstrate the value of one such solution.

This brief illustrates how a complex big data analytics workload is well served by Attala Composable Storage on a Supermicro chassis with Intel® 3D NAND SSDs. This workload—which once struggled to scale to support concurrent customers when serviced by an iSCSI appliance or enterprise storage area network (SAN), is shown to match local NVMe performance and scale independently for two customers with no degradation in application runtime when using the NVMe-oF target.

The Test Environment

The system under test was a collection of four two-socket servers with Intel® Xeon® Platinum 8180 processors with 28-cores and populated with 768GB of RAM. The onboard 10GbE port serviced regular Ethernet traffic. The RDMA functionality to support NVMe-oF was enabled with either a RoCEv2 RDMA NIC (RNIC) or an Attala FPGA-powered Host NVMe-oF Adaptor (HNA) connecting to the 100GbE switch as 40GbE ports. Each server had two 1.6TB Intel® SSD DC P4600 NVMe SSDs used in establishing the local NVMe baseline.

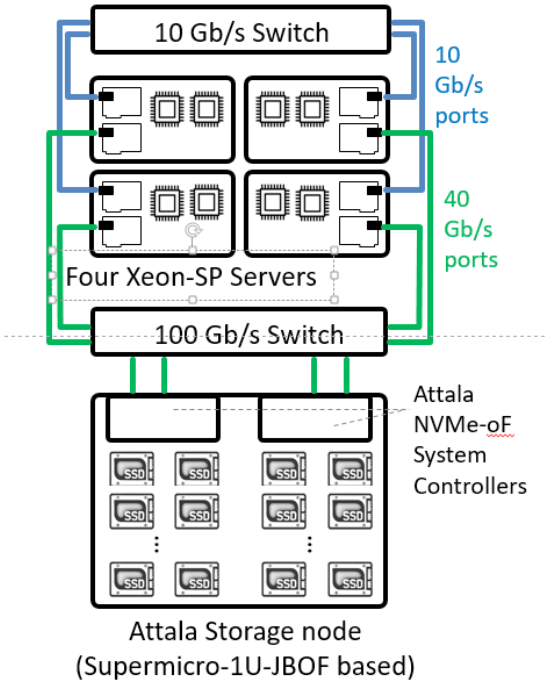

Figure 1. Logical Diagram of Test Environment

The remote Attala storage node features a new Supermicro all-flash NVMe 1U chassis configured for 32 u.2 drive capacity. Two 1.6TB P4600 NVMe SSDs were assigned to each physical server to mimic the local remote configuration.

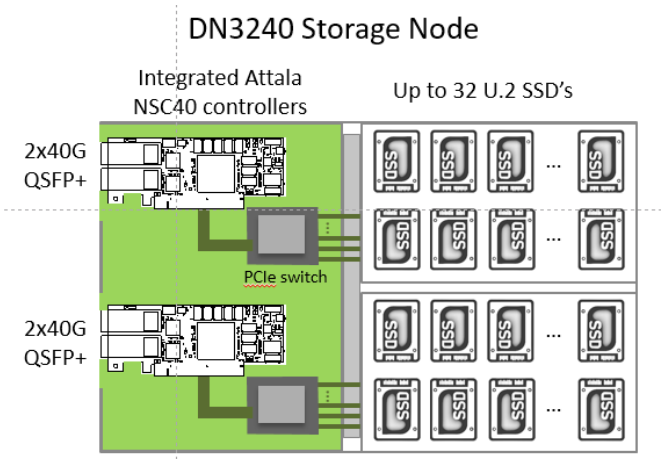

Figure 2. Supermicro Storage Chassis with Attala Dn3240

NVMe-oF Pool Advantages

Provision networked NVMe capacity to cloud infrastructure while matching performance levels of local PCIe-attached SSDs

Lower TCO and increase scalability and elasticity via pooled NVMe – versus local PCIe-attached SSDs. Increase performance versus iSCSI appliances or SAN storage

Options to connect using RDMA-capable NIC or an FPGA-powered host network adaptor from Attala

Industry standard NVMe-over-Fabrics delivers efficient NVMe performance on existing Ethernet infrastructure

Users and applications simply see SSDs and need no specific treatment or host software to realize benefits

From a shared NVMe storage target, the cloud can support concurrent customers at the same quality of service seen with dedicated NVMe

Leapfrog iSCSI/SATA solutions by leveraging NVMe over Fabrics and remove the IO bottleneck

Two Attala NVMe-oF System Controllers (NSC) give NVMe-oF functionality with two 40GbE ports per NSC (Figure 2) for 160Gbps total bandwidth. Attala's HNA and NSC leverage the Intel® Arria® 10 FPGA to achieve optimal performance.

For each host server, KVM was the hypervisor on which CentOS 7.2 guest virtual machines were employed in deploying the Hadoop data and name nodes. Each server hosted 6 data node VMs per Figure 3. Two servers also hosted a name node VM.

The Hadoop cluster was built using Hortonworks Data Platform with more complete software configuration details in Table 1.

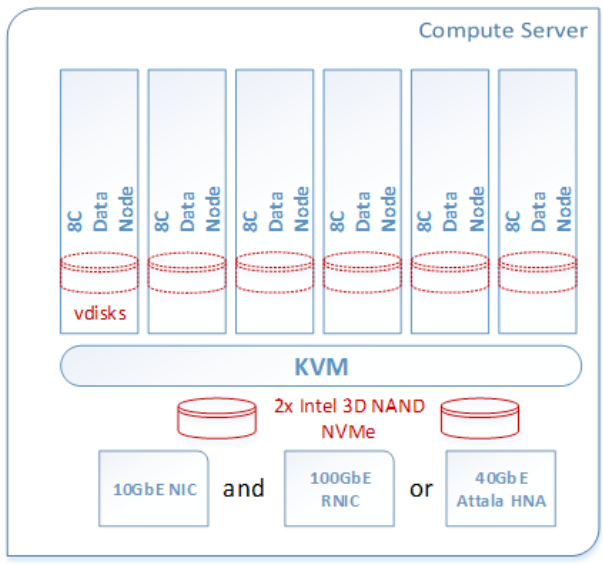

Figure 3. Compute Host Example

Big Data Workload Terasort

HiBench Terasort, for a 1TB data set size, reads 10 billion 100-byte rows which are written back, once sorted, to HDFS.

Software	
Hypervisor	KVM QEMU 1.5.3
Guest OS	CentOS 7.2
Hadoop	HDP 2.4
Spark	1.6
Java	1.8.0
HDFS Replication Factor	2
Name Node VMs	2 x 4 VCPUs, 24GB RAM
Data Node VMs	24 x 16 VCPUs, 112GB RAM
Data Node Logical Volumes	375GB data + 100GB temp

Table 1. Single Cluster Software Configuration

Terasort is a read, compute, and write intensive workload over three stages (see Figure 4). Especially when remote storage is employed, Terasort stresses the network too. This work is carried out by the Hadoop cluster using Spark executors.

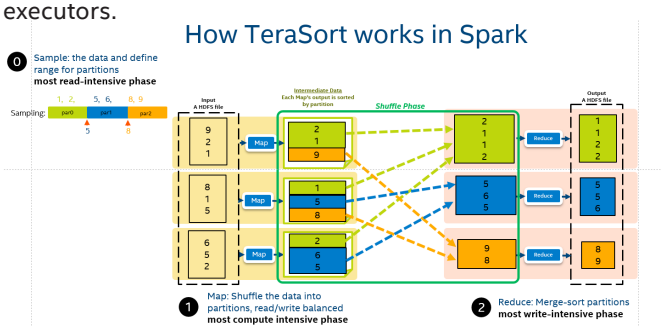

Figure 4. Terasort on Spark

The Local NVME Challenge

Putting NVMe devices directly on the compute host is a natural choice and also provides a reference against any remote NVMe solution. The 4-server, 24-data node HDP cluster took 9.3 minutes to complete the 1TB Terasort run.

The same servers and cluster again ran Terasort identically but over NVMe-oF with either an RNIC or Attala's FPGA-based HNA to handle RDMA against the remote JBOF where two NVMe SSDs were shared to each compute host. The results (Table 2) are exciting.

Software	Local NVMe	NVMe-oF-RNIC	NVMe-oF-Attala
Fabric	PCIe	Ethernet	Ethernet
Adapter	None	RNIC	Attala HNA
Target	Local	Supermicro JBOF	Supermicro JBOF
			Terasort Runtimes (minutes)
Stage 0	0.5	0.4	0.4
Stage 1	3.5	3.6	3.5
Stage 2	5.3	5.4	5.3
Total	9.3	9.5	9.3

Table 2. NVMe-oF vs Local Terasort Result Summary

Remote NVMe performance even close to local NVMe would have been compelling alone having the TCO and operational flexibility to pool storage resources, but these two NVMe-oF solutions essentially matched that seen when local NVMe SSDs were employed.

A look at CPU utilization (see Table 3) confirms that the NVMe-oF solutions are as efficient as local NVMe. Important performance counters, such as CPU utilization and iowait, were almost identical.

To the application, remote NVMe-oF storage performs just as if it were local NVMe in every way including the way it resources.

	Average CPU Utilization		
KVM Host	Local	RNIC	HNA
Total	62	63	63
Sys/kernel	3	4	3
guest	58	59	59
Data Node VMs	Local	RNIC	HNA
total	85	86	85
Sys/kernel	10	11	10
usr	70	70	70
iowait	2	2	1

Table 3. Single Cluster CPU Utilization During Terasort

The Scaling Challenge

Other networked storage solutions exist and are deployed in clouds today. Consider a cloud relying on iSCSI for block storage on an enterprise-class SAN.

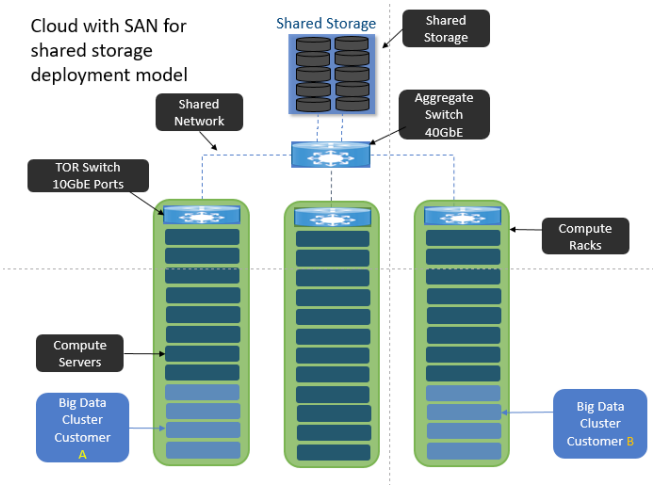

Figure 5. Cloud Environment Having iSCSI SAN

A cloud is expected to support multi-tenancy. To test this environment's capacity to handle multiple big data customers we deployed two Hadoop clusters onto distinct grouping of four servers in separate racks to control for any top of rack switch sharing effects as shown in Figure 5.

Before proceeding to concurrent Terasort runs, a single cluster 1TB Terasort run on an otherwise idle cloud served as a baseline. This enterprise SAN delivered sufficient IOPS, even if leveraging HDD drives, to result in a runtime of 24.3 minutes.

But, when two identical clusters concurrently shared the SAN to service the Terasort workload, the very same cluster experienced a 67% increase in runtime.

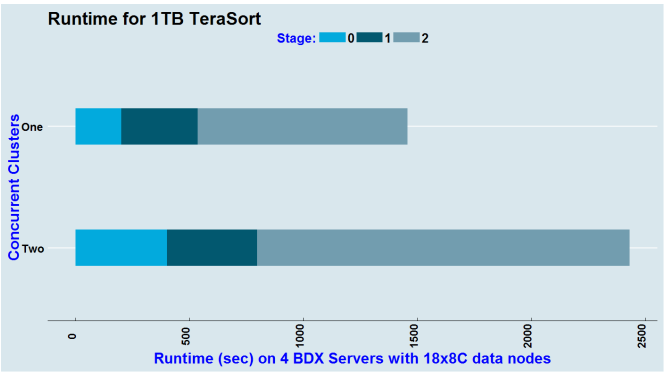

Figure 6. Terasort Runtime Scaling with iSCSI SAN

The runtime increase resulted as IO demand doubled but the SAN was unable to increase IOPs accordingly.

Number of Competing Customers	Runtime (min)	ISCSI IO MB/s
1	24.3	639
2	40.5	740
increase 2:1	1.67	1.16

Table 4. Terasort Scaling Results with iSCSI SAN

The Scaling Response

While placing local NVMe on compute servers would improve performance and mitigate multi-tenancy concerns, the cloud operator would lose elasticity that networked storage allows and potentially strand expensive flash, or risk under-provisioning storage thereby forcing customers to add more compute nodes to meet storage capacity requirements.

Earlier, a customer using 4 nodes performed well using NVMe-oF. A second customer was added to share the Attala storage node. Both customers concurrently ran Terasort at 1TB.

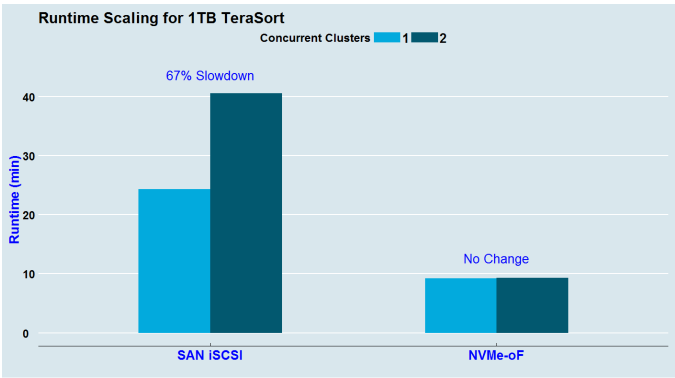

Figure 7. Terasort Runtime Scaling (Lower is Better)

The NVMe-oF solution delivered performance independently to both customers as it was able to scale IOPS linearly. Figure 7 illustrates that NVMe-oF solution—already 2.5x faster than the iSCSI SAN when serving a single customer—especially shines relative to iSCSI (or 4.4x faster) when asked to support multiple tenants.

NVMe-oF Solutions Deliver Local Performance with Cloud Flexibility

NVMe over Fabrics offers elastic storage capacity to cloud infrastructure while matching performance levels of local NVMe. Cloud users and applications simply see drives and operate as before, only faster. Even in the presence of other IO-intensive customers, Attala Composable Storage on Supermicro's All-Flash NVMe Storage Chassis is shown to allow big data cloud tenants to scale independently.

Compute hosts leverage high performance NVMe-oF Attala Composable Storage targets using an industry standard RNIC or Attala's Intel® Arria®10 FPGA-powered HNA card. Either interface encapsulates the NVMe commands across RDMA as efficiently as processing IO operations on NVMe locally.

NVMe-over-Fabric delivers NVMe efficiency on existing Ethernet infrastructure. Combined with proven, integrated storage JBOFs such as that offered by Attala and Supermicro, operators can offer guests local NVMe performance with cloud flexibility, scalability and efficiency.

About Supermicro

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.

About Attala Systems

Our mission is to reinvent cloud infrastructure; using an FPGA-based fabric, we excavate through legacy layers to liberate the performance and cost efficiencies of the underlying hardware and make it available to analytics workloads and multi-tenant cloud applications to create a breakthrough in cloud infrastructure performance, cost, agility and scalability. Learn more on www.attalasystems.com.

Learn More

The Solutions Library on the Intel® Builders home page can help you find reference architectures, white papers, and solution briefs that can help you build and enhance your data infrastructure - https://builders.intel.com/solutionslibrary.

For more details about Supermicro, visit www.supermicro.com.

For more details about Attala, visit www.attalasystems.com.

You can follow Intel® Builders on Twitter by using #IntelBuilders.

Printed in USA 0418/MK/DCG/PDF ♻ Please Recycle