CASE STUDY

# Pooling NVMe within GPFS NSDs Achieves Unprecedented Burst Buffer Bandwidth Level

## BACKGROUND

SciNet is Canada's largest supercomputer center, providing Canadian researchers with computational resources and expertise necessary to perform their research at massive scale. The center helps power work from the biomedical sciences and aerospace engineering to astrophysics and climate science.

The new InfiniBand based SciNet supercomputer located at the University of Toronto, needs to meet high levels of availability to ensure high ROI for its users. Levering the EDR InfiniBand technology, the supercomputer offers multiple technology innovations and networking accelerations that delivers world-class applications performance. Furthermore, the system leveraged the Dragonfly+ topology, which enables seamless extensions as the compute and storage needs grow. Another interesting technology innovation is the usage of burst buffer for smart and fast checkpointing.

## CHALLENGE & SOLUTION

High-performance computing applications consist of complex sets of processes that may run for weeks. Interrupting any of these processes can destroy the results of an entire compute job. The problem becomes more pronounced as supercomputers become more powerful – imagine the challenge for Canada's largest supercomputer. Hence, parallel computing applications use a *checkpoint restart* in case of an interruption- a technique that allows compute jobs to be restarted from the most recently saved checkpoint.

Checkpoints are typically saved in a shared, parallel file system; SciNet has chosen GPFS. But as clusters become larger and the amount of memory per node increases, each individual checkpoint becomes larger and either takes more time to complete or requires a higher-performance file system. When a system is checkpointing it's not computing, which reduces the availability score of the system. To shorten those moments of unavailability, SciNet decided to implement an InfiniBand-based burst buffer leveraging Mellanox interconnect accelerations and Excelero's NVMesh.

A joint solution combining Excelero's NVMesh with Mellanox's world-leading, end-to-end InfiniBand networking solution has enabled SciNet to build a petabyte-scale unified pool of distributed high-performance NVMe as a burst buffer for checkpointing. The NVMe pool delivers 230GB/s of throughput and well over 20M random 4k IOPS and enables SciNet to meet its availability SLAs.

## HIGHLIGHTS

### USE CASE

Large-scale modeling, simulation, analysis and visualization

### CHALLENGE

Complete checkpoints within 15 minutes to meet availability SLAs

### SOLUTION

NVMesh enables a petabyte-scale unified pool of distributed high-performance NVMe flash as burst buffer for checkpointing

### RESULTS

- 80 pooled NVMe devices
- 148 GB/s of write burst (device-limited)
- 230GB/s read throughput (network-limited)
- Well over 20M random 4k IOPS

### BENEFITS

- Meets 15-minute checkpoint window
- Extremely cost effective
- Unprecedented burst buffer bandwidth

*"For SciNet, NVMesh is an extremely cost-effective method of achieving unheard-of burst buffer bandwidth."*

*Dr. Daniel Gruner*
*Chief Technical Officer*
*SciNet High Performance Computing Consortium*

## ABOUT BURST BUFFERS

A burst buffer is a fast and intermediate storage layer between the non-persistent memory of the compute nodes and persistent storage – the parallel file system. This layer is configured to take a burst of write IO at a very high rate. Once the burst (checkpoint) is complete, the written data is "drained" to the parallel file system, using the GPFS policy engine. This allows checkpoints to finish rapidly so that systems meet availability SLAs. When flash storage is used as the burst buffer pool it has the added advantage of facilitating a faster restart (when needed) as checkpoint restarts often impose a very large random read load on the underlying storage. With those benefits in mind, the SciNet burst buffer was sized to hold two checkpoints so the most recent completed checkpoint would be available for restart. To maximize performance and reduce the checkpoint window, SciNet decided to leverage higher performance NVMe SSDs.

## ACHIEVING UNPRECEDENTED NVMESH® BURST BUFFER BANDWIDTH

*"For SciNet, NVMesh is an extremely cost-effective method of achieving unheard-of burst buffer bandwidth," said Dr. Daniel Gruner, chief technical officer, SciNet High Performance Computing Consortium. "By adding commodity flash drives and NVMesh software to compute nodes, and to a low-latency network fabric that was already provided for the supercomputer itself, NVMesh provides redundancy without impacting target CPUs. This enables standard servers to go beyond their usual role in acting as block targets – the servers now can also act as file servers."*

Excelero enables customers to build high-performance burst buffers without needing additional proprietary arrays. NVMesh customers can use standard NVMe drives in their application servers to build a local burst buffer or build converged file system/block server appliances.

This method includes the advantage of adding redundancy with centralized management, while at the same time preserving all compute resources for the applications themselves.

NVMesh and its patented Remote Direct Drive Access (RDDA) technology, riding on top of Mellanox Remote Direct Memory Access (RDMA), allows customers to logically disaggregate NVMe drives in the compute nodes away from CPU resources. Thus, remote compute nodes can use local NVMe drives without consuming the local CPU. As a result, every compute node's local NVMe drives are pooled for use by the cluster. In the simplistic form, half of each drive can be used as a local burst buffer while the other half remains reserved for the redundant copy of a peer. Thus, when a node fails, its scratch is preserved and accessible by an alternate node – any node on the fabric.

*"In supercomputing any unavailability wastes time, reduces the availability score of the system and impedes the progress of scientific exploration. We're delighted to provide SciNet and its researchers with important storage functionality that achieves the highest performance available in the industry at a significantly reduced price – while assuring vital scientific research can progress swiftly," said Lior Gal, CEO and co-founder at Excelero.*

## MELLANOX'S END-TO-END EDR INFINIBAND

Mellanox's end-to-end EDR InfiniBand enables the world's fastest storage networking technology, supporting the highest bandwidth (EDR 100Gb/s) and the lowest latency (<90ns port-to-port) of any major networking fabric.

Mellanox EDR solution simplifies the deployment and management of high-speed networking for high performance data centers, optimizing the overall performance, power and density of the most demanding application workloads. Delivering the highest data speeds and performance-enhancing CPU offloads, Mellanox InfiniBand (IB) solutions maximize data center return on investment and reduce cost of ownership.

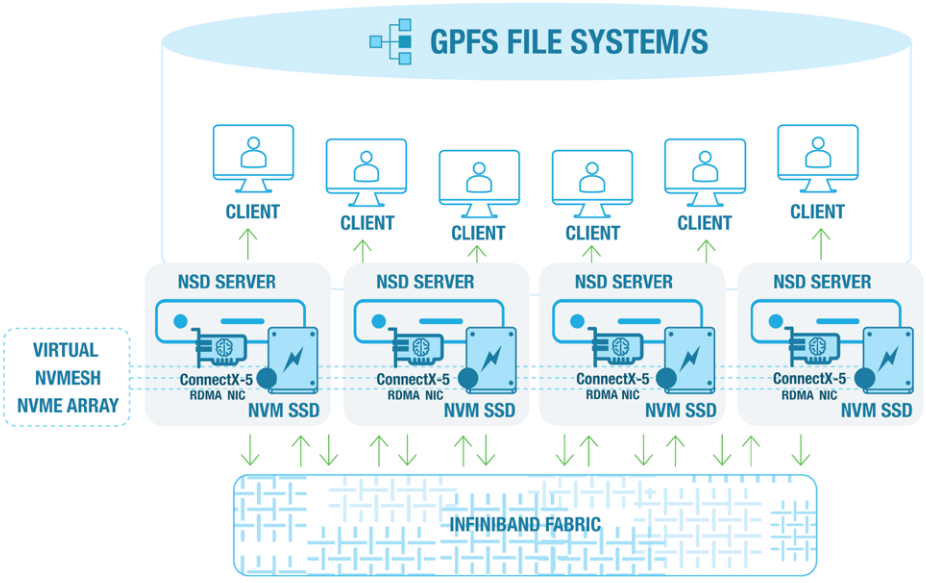


***Figure 1. Joint Excelero-Mellanox solution, providing high-speed, low latency, shared storage with in-server flash performance***

## WORLD'S FIRST SMART NETWORK SWITCH

Mellanox's industry-leading Smart Switch-IB™ 2 is the world's first smart network switch and optimized for ultra-low latency lossless fabrics. Built to support SciNet and other performance-demanding data centers, Smart Switch-IB™ 2 is is the ideal top-of-rack switch for hyper-converged infrastructure deployments, enabling the highest fabric performance available in the market.

Mellanox's RDMA-enabled InfiniBand ConnectX®-5 network adapter cards, wquipped with intelligent ASICs, provide advanced NVMe over Fabric (NVMe-oF) target offload capabilities, increasing NVMe storage access efficiency levels, without any CPU intervention, and reaching sub-600ns latency levels. By bypassing the CPU, RDMA frees up the CPU's resources, especially in a CPU-bound environment, for storage and supercomputing tasks, which allows for higher scalability and data center efficiency levels.

*"Mellanox's smart and scalable NVMe accelerations enable users to maximize their storage performance and efficiency," said Gilad Shainer, Vice President of Marketing at Mellanox Technologies. "Leveraging the advantages of InfiniBand, Excelero delivers world leading NVMe platforms that are accelerating the next generations of supercomputers."*

## INNOVATIVE CO-DESIGN SHARP™ TECHNOLOGY

Mellanox's industry-leading 100Gb/s InfiniBand smart switches empower SciNet with in-network computing through Co-Design SHARP™ technology, enabling up to 7.2Tb/s of non-blocking bandwidth with 90ns port-to-port latency. Powered by Mellanox's own ASICs, the 36-port non-blocking EDR 100Gb/s IB Smart Switch cuts through switching at line speeds of 40/56/100 Gb/s with no packet loss. The dynamically shared switch buffer provides the best microburst adoption, thus it enables the storage solution to deliver guaranteed throughput and latency. Combined with faster congestion notification, Mellanox switches form a network fabric that unleashes the maximal power of NVMe storage pools across the network.

## JOINT SOLUTION

The joint Mellanox-Excelero solution provides remote, high-speed, low latency shared storage with in-server flash performance. Deploying the solution on existing supercomputing application clusters, transforms them into a converged infrastructure with extremely high levels of compute, storage and application performance.

While Excelero's NVMesh RDDA technology enables performance-demanding supercomputing applications to enjoy the full performance, capacity and processing power of underlying servers and storage, Mellanox's InfiniBand RDMA technology boosts the joint-solution with smart accelerations, enabling in-network-computing that ensures faster data processing, higher performance, and efficiency for the various applications workloads. Maximizing overall system performance, the joint solution enables customers to achieve the highest performance levels at the best price/performance ratio.

## SCINET'S NVMESH BURST BUFFER IMPLEMENTATION

Excelero's NVMesh enables SciNet to create a petabyte-scale unified pool of distributed high-performance NVMe flash, retaining the speeds and latencies of directly-attached media. The NVMe pool, consisting of 80 NVMe devices in just 10 NSD servers gives about 148GB/s of write burst (device limited) and 230GB/s (network limited) of read throughput and well over 20M random 4k IOPS. This configuration is more than sufficient to meet the 15 minutes checkpoint window that is required to meet availability SLAs that were defined for the new supercomputer.

For SciNet, NVMesh has been able to achieve unprecedented burst buffer bandwidth by adding commodity flash drives and NVMesh software to compute nodes and low latency network fabric. NVMesh provides redundancy without impacting target CPUs, enabling standard servers to act not only as block targets but also as file servers. Moreover, as NVMesh appears like a simple block device, integration with SciNet's parallel file system was very straightforward.

## NVMESH BENEFITS FOR BURST BUFFER

- Petabyte-scale unified pool of high-performance flash retaining the speeds and latencies of directly-attached media.
- Supports large-scale modeling, simulation, analysis and visualization.
- Visualizes supercomputer simulation data on 100s of compute nodes.
- Enables fast checkpointing and computer job restarts
- Achieves highest performance at the lowest price.
- Leverages the full performance of your NVMe SSDs at scale, over the network.
- Scales your performance and capacity linearly.
- Easy to manage & monitor, reduces the maintenance TCO.
- Utilizes hardware from any server, storage and network vendor. No vendor lock-in.

## About SciNet

SciNet is Canada's largest supercomputer centre, providing Canadian researchers with computational resources and expertise necessary to perform their research on scales not previously possible in Canada. SciNet powers work from the biomedical sciences and aerospace engineering to astrophysics and climate science. SciNet is part of Compute Canada, a national infrastructure for supercomputing-powered innovation, and is funded by CFI, NSERC, the Ontario Government, Fed Dev Ontario, and the University of Toronto. More information is available at: www.scinethpc.ca/about-scinet

## About Excelero

Excelero enables enterprises and service providers to design scale-out storage infrastructures leveraging standard servers and high-performance flash storage. Founded in 2014 by a team of storage veterans and inspired by the tech giants' shared-nothing architectures for web-scale applications, the company has designed a software-defined block storage solution that meets performance and scalability requirements of the largest web-scale and enterprise applications. More information is available at: www.excelero.com

## About Mellanox

Mellanox Technologies is a leading supplier of end-to-end InfiniBand and Ethernet interconnect solutions and services for servers and storage. Mellanox interconnect solutions increase data center efficiency by providing the highest throughput and lowest latency, delivering data faster to applications and unlocking system performance capability. Mellanox offers a choice of fast interconnect products: adapters, switches, software, cables and silicon that accelerate application runtime and maximize business results for a wide range of markets including high-performance computing, enterprise data centers, Web 2.0, cloud, storage and financial services. More information is available at: www.mellanox.com

## About NVMesh

With Excelero's NVMesh, customers can build distributed, high-performance server SAN for mixed application workloads. Customers benefit from the performance of local flash, with the convenience of centralized storage while avoiding proprietary hardware lock-in and reducing the overall storage TCO.

The solution has been deployed for hyper-scale Industrial IoT services, machine learning applications and massive-scale simulation visualization. More information is available at: www.excelero.com/product/nvmesh

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com



060077CS
Rev1.1