Micron NVMe SSDs and Excelero NVMesh® Shared Storage Speed AI Innovation for InstaDeep

High-performance flash storage using NVMe over Fabrics keeps up with the voracious appetite of InstaDeep's GPUs

Artificial intelligence (AI) as a service. Possible? It's a pioneering innovation now being brought to life. And it's on a highly efficient data center infrastructure built from a scalable pool of high-performance Micron® SSDs with NVMe™ using the NVM Express™ over Fabrics™ (NVMe-oF™) protocol.

The project was announced by Excelero, a company offering software-defined block storage solutions for web-scale applications.

InstaDeep Ltd., a global innovator of AI products and services for the enterprise, chose Excelero's NVMesh® software on Micron flash storage to run its AI and machine learning (ML) applications on GPU-based systems efficiently and with low latency. InstaDeep's AI as a Service (AIaaS) solution is designed to bring AI/ML workloads within reach of more organizations, helping them gain the benefits of AI without the investment of time, costs, and expertise to run their own AI stacks.

> *"The GPU systems powering the AI and ML explosion have an amazing appetite for data, but many organizations are finding they quickly create a storage bottleneck."*
>
> *−Yaniv Romem, Excelero CTO*

NVMe-oF Unlocks Flash Storage Performance

Micron's high-performance NVMe SSDs utilized in Excelero's NVMesh Reference Architecture help unlock the full performance and value of NVMe and share it across applications. NVMesh is a 100% software-only solution for distributed NVMe, so it is hardware agnostic.

Storage experts estimate, on average, companies are using less than 50% of an NVMe SSD's IOPS and capacity[1] in application servers. With Excelero's NVMesh, customers can more efficiently build distributed, high-performance Server SAN for AI/ML workloads.

AI/ML Workloads Need Fast, Flash Storage

InstaDeep offers a host of AI solutions, ranging from optimized pattern-recognition platforms, to GPU-accelerated insights, to self-learning decision-making systems.

InstaDeep wanted speed and low latency for its service platform. “Finding a storage infrastructure that would scale modularly and was highly efficient for AI and ML workflows is no small challenge,” explained Amine Kerkeni, Head of AI Product at InstaDeep. “Our clients simply will not achieve the performance they need if an infrastructure starves the GPUs with slow storage or wastes time copying data to and from systems.”

As InstaDeep built out the AI/ML solution architecture for its AIaaS offering, it searched for the right technology to serve its customers, maximize utilization, and reduce overall costs. The choice: GPU-optimized servers to access remote Micron high-performance NVMe SSDs as if they were local flash—with full IOPS and bandwidth capabilities. InstaDeep’s team can now provide streamlined workflow management and faster time to insights from its new AI services.

Feeding the GPU Means No Bottlenecks

“The GPU systems powering the AI and ML explosion have an amazing appetite for data, but many organizations are finding they quickly create a storage bottleneck,” explained Yaniv Romem, Excelero’s CTO. “The only storage that is fast enough to keep up with these GPUs is local NVMe flash, due to the high competition for valuable PCIe connectivity amongst GPUs, networking, and storage.”

The InstaDeep platform also benefited from the massive network connectivity of NVIDIA’s DGX® nodes. A DGX node can ingest as much as 48 GB/s of bandwidth via 4-8 x 100Gb ports[2]. Excelero’s NVMesh software enables its customers to maximize GPU utilization with this connectivity and low-latency (5µs), delivering the high IOPS of local NVMe in a distributed and linearly scalable architecture.

NVMe-oF Evolves to AI/ML Use Cases

Historically, the large storage needs of high-performance computing (HPC) focused primarily on streaming large files between parallel file systems. Now, NVMe-oF helps fulfill critical needs for high capacity, low latency storage in AI/ML, big data, and other workloads that demand intensive I/O. Organizations can now accelerate HPC applications with readily available storage.

Excelero NVMesh enables IT to pursue both performance and storage functionality. Organizations can easily share data and protect it, as well as deliver new levels of storage performance. Excelero NVMesh and Micron NVMe storage enable InstaDeep to deliver huge bandwidth and IOPS in a very small and cost-effective storage form factor.

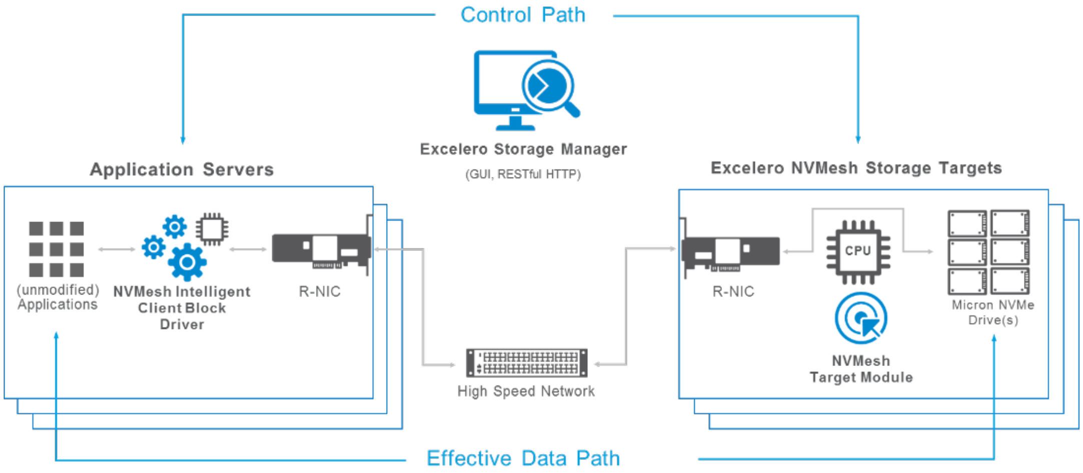

Figure 1: Excelero NVMesh with Micron SSDs Architecture Overview

What’s Under the Hood

InstaDeep’s AIaaS system includes a 2U Boston Flash-IO Talyn server and Excelero NVMesh software to provide access to up to 100TB of Micron NVMe SSDs, which are the external high-performance storage. The Mellanox 100Gb Infiniband network cards in the NVIDIA DGX-1® node enable local performance from the remote NVMe storage with equal or better performance than the local cache in the DGX.

Fast Facts: Excelero

- **Industry**: Software-defined block storage solutions
- **Challenges**: Meeting low-latency performance and scalability requirements of large web-scale and enterprise applications
- **Solution**: NVMesh for NVM Express over Fabrics (NVMe-oF)
- **What made the difference**: Enabling customers to run NVMesh on any file system, with lowered latency and larger bandwidth, as if the NVMe storage were local
- **Result**: NVMesh enables shared NVMe across any network and supports any local or distributed file system, providing performance of local flash with the convenience of centralized storage while avoiding proprietary hardware lock-in and reducing overall storage TCO

Fast Facts: InstaDeep

- **Industry**: AI software and services
- **Challenges**: Delivering AI/ML benefits for all sizes of customers who might not have the expertise and resources to set up AI stacks
- **Solution**: An AIaaS platform with the speed, performance and efficiency to handle multiple projects in dynamic and complex environments such as mobility, logistics, manufacturing, and energy
- **What made the difference**: High-performance Micron NVMe SSDs linked with Excelero NVMesh to enable avoidance of storage starvation issues in GPUs
- **Result**: One of the industry’s first and most capable AIaaS solutions

Learn More

Accelerating infrastructure is foundational for AI, ML and deep learning. Talk with your Micron representative or visit micron.com/AI or micron.com/flashfabric to learn more.

[1] Chris Evans, Architecting IT blog series, for example. https://blog.architecting.it/avoiding-all-flash-missing-iops/

[2] https://www.excelero.com/wp-content/uploads/2019/05/InstaDeep%E2%84%A2-powers-AI-as-a-Service-with-shared-NVMe.pdf

Reference herein to any specific third-party commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by Micron or the referenced customer. This case study was prepared for informational purposes only as a general account of certain assistance provided by Micron to the referenced customer. Many factors may have contributed to the results and benefits described in this case study, and Micron does not guarantee comparable results elsewhere. The information in this case study is provided "as is" and does not constitute any representation or warranty, either express or implied, by Micron or the referenced customer regarding any information, apparatus, product, or process discussed herein, or regarding the accuracy, completeness, or usefulness of any information, apparatus, product, or process discussed herein, and all such representations and warranties are hereby expressly disclaimed, including without limitation those respecting merchant ability or fitness for a particular purpose. Micron products are warranted only to meet Micron's production data sheet specifications. Micron products and specifications are subject to change without notice. Information in this case study is subject to change without notice. Any dates or timelines referenced in this case study are estimates only.©2019 Micron Technology, Inc. All rights reserved. All information is provided on an “AS IS” basis without warranties of any kind. Micron, the Micron logo, and all other Micron trademarks are the property of Micron Technology, Inc. All other trademarks are the property of their respective owners. Rev. A 7/19 CCM004-676576390-11339