WHITE PAPER–MAY 2020

AI Composability and Virtualization: Mellanox Network Attached GPUs

vSphere Bitfusion Solution Brief

vmware®

Table of Contents

SOLUTION HIGHLIGHTS

- Make all GPUs in your network visible to all workloads, clients, and containers
- Run AI/ML frameworks by attaching remote GPUs from anywhere in the network at run-time
- Any GPU server can be attached to the network, and instantaneously be used by any AI/ML remote client
- Industry's first composable AI and elastic GPUs with Mellanox 10, 25, 40, 50, 100, or 200Gb/s end-to-end intelligent interconnect solutions

Pool, Share, and Virtualize Your GPU Cluster Solution Highlights With Mellanox Low-Latency and High-Throughput Network

Accelerated compute (GPUs, FPGAs, AI ASICs) is needed to augment CPUs to efficiently run Artificial Intelligence (AI) and Machine Learning (ML) workloads. However, GPUs are a scarce resource, 10-20x more expensive, and are deployed in very small quantities in the network.

Now with vSphere Bitfusion software and Mellanox end-to-end high-performance Ethernet solutions, any GPU cluster can be remotely attached to clients, containers, or workloads–essentially any compute across the network. Much like storage area networks, or NVME over Fabric, GPUs can be disaggregated and consumed on-demand by remote clients. The solution works with any software environment (bare-metal, ESXi, containers, etc.) and with any type of GPU server (e.g., any GPU type, GPU density, NVLink, PCIe, RoCE networks, InfiniBand networks, etc.).

Challenge

Today, GPUs are being deployed as an isolated hardware resource, dedicated to very narrow and specialized workloads in an organization. Since few developers have privileged access to the GPU servers, not only do organizations see low utilization, but large bodies of AI researchers, ML developers, and data scientists cannot get access to GPU resources.

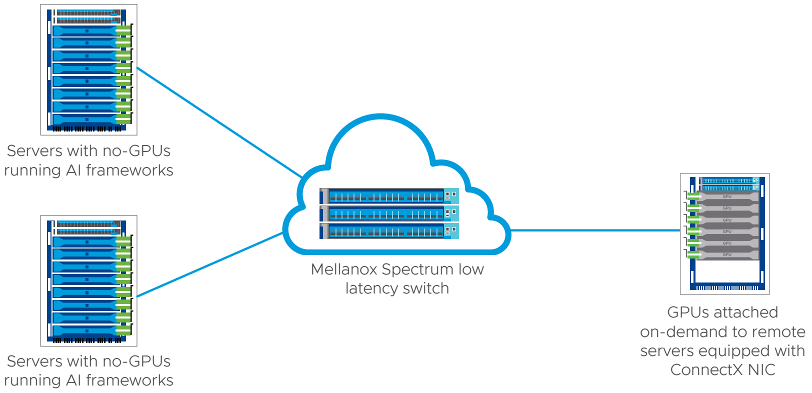
The few fortunate AI developers that are allocated GPU run-time must often exercise a painful migration of their application's environment and data to the GPU server–not a productive process—which further hogs the GPU server, and lowers its utilization.

To further aggravate the problem, there is no viable way to sub-segment physical GPUs into small "entities" and allow a user or workload to consume less than one GPU remotely (e.g., for development and test, inference workloads, etc.). Workloads are forced to use GPUs as designed in hardware, with no ability to virtualize and share a GPU's resources.

Solution

With a vSphere Bitfusion elastic software platform and Mellanox networking solutions, GPUs can be sliced and diced, and connected remotely to any client across the network. Very much like NVMe over Fabric, GPUs become a composable resource, reachable and accessible by any remote node. And much like storage, which needs low latency and high throughput access, Mellanox technology can provide a networking fabric that meets the requirements. The vSphere Bitfusion elastic software platform works with all Mellanox technologies: Ethernet, RoCE, and InfiniBand.

Once the GPU servers and the compute servers are connected with Mellanox NICs and switches, an Elastic AI architecture is formed, allowing any user and any AI/ML application to connect to one or more GPU servers for the application's run-time and to disconnect when done. Metrics show that utilization goes up, as well as flexibility, agility, productivity, and sharing. The IT organization gets an AI uplift with the ability to share, pool, and automate resources.


Servers with no-GPUs running AI frameworks
Servers with no-GPUs running AI frameworks
Mellanox Spectrum low latency switch
GPUs attached on-demand to remote servers equipped with ConnectX NIC

How Does It Work

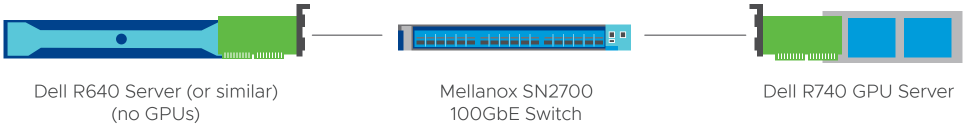
Implementing Elastic AI with vSphere Bitfusion and Mellanox is straightforward. vSphere Bitfusion provides a VMware appliance (pre-packaged VM) for the GPU servers and OS-independent software for each compute server; all the vSphere Bitfusion software runs in user space. Mellanox provides the network fabric (network adapters and switches).

Any GPU server and any compute server will work (there is no need for any particular hardware or memory design/configuration). With minimal steps implemented, users can run AI workloads from any one of the servers. With vSphere Bitfusion software under the hood, one or more GPUs from the cluster can be attached on-demand for the duration of the CUDA execution.

With Mellanox infrastructure and vSphere Bitfusion in place, users can share and pool common GPU resources. Neither are users bounded by their physical location or software environment. All they need to do is launch their ML or AI workload (unmodified), and vSphere Bitfusion will attach the remote GPUs.

Organizations can now plan ahead and provide GPU resources—efficiently, with speed and agility—to all developers and production teams. The Mellanox Spectrum™ switches, ConnectX® NICs, BlueField SmartNICs, and LinkX® cable and transceivers interconnect products offer a seamless experience to the users, as if each user has a local GPU (or GPUs) attached to their servers. vSphere Bitfusion software works with any environment: containers, virtual machines, and bare-metal. The broad Mellanox portfolio of InfiniBand, RoCE, and Ethernet technologies are a perfect match for Elastic AI, and complement the vSphere Bitfusion software platform. Very much like storage, where extreme performance is needed for demanding workloads, vSphere Bitfusion performance is enhanced in High Performance Computing (HPC) environments through InfiniBand infrastructure and delivers top-notch performance.

When superior performance is needed in scale-out environments (with many users), Mellanox and vSphere Bitfusion offer RoCE for remote direct memory access to boost network and host performance through lower latency, lower CPU requirements, and higher bandwidth. Finally, as Ethernet is the most ubiquitous network protocol in Enterprise, vSphere Bitfusion delivers a full gamut of Elastic AI infrastructure configuration options that enables higher GPU efficiency. In addition, the solution can also operate in heterogeneous networking environments. For example, GPUs can be connected with InfiniBand or RoCE for high-profile customers, while others could use common Ethernet for access.

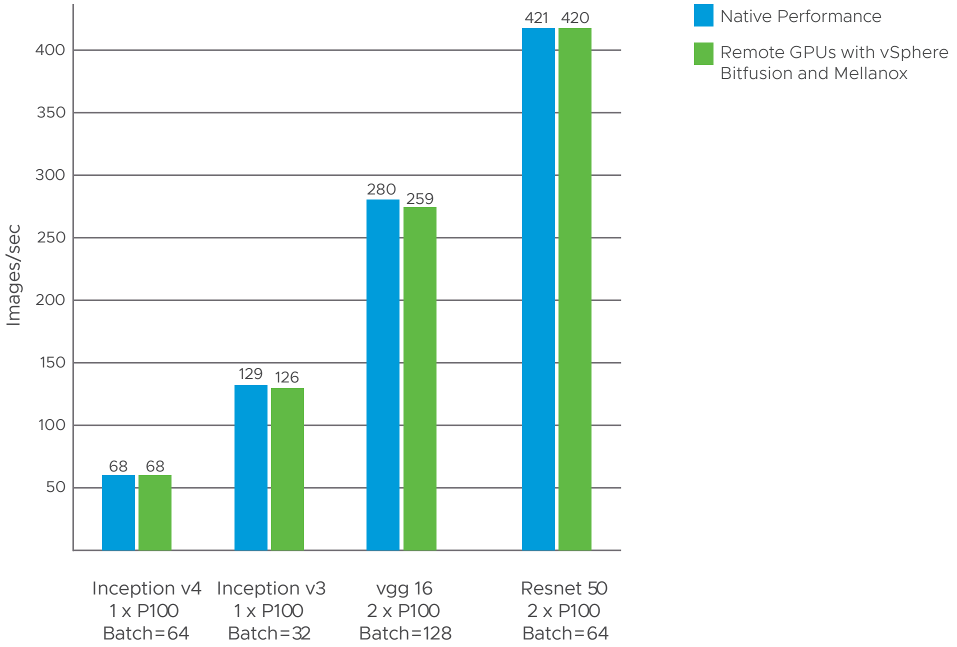

Dell R640 Server (or similar) (no GPUs)
Mellanox SN2700 100GbE Switch
Dell R740 GPU Server

Benchmark and Tests

ML and AI workloads stress the throughput and latency of the network. Mellanox and vSphere Bitfusion set up the following configuration to emulate a real-life Elastic AI infrastructure. The test bed relied on generally available servers (both compute and GPU servers), and on a generic OS and software packages (Ubuntu 16.04, publicly available NIC drivers, CUDA 9.1, etc.).

It is worthwhile mentioning that the implementation of vSphere Bitfusion does not necessitate any changes to the OS, drivers, kernel modules, or AI frameworks.

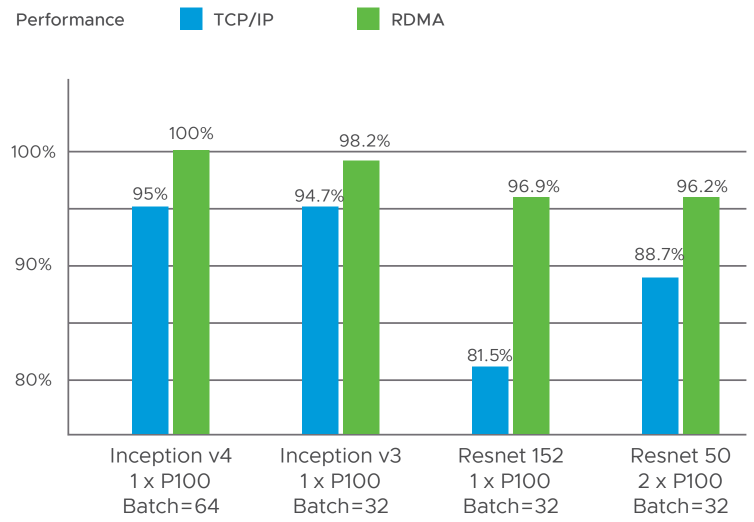
The intent of the tests was to prove the AI developer experience—executing a workload—was the same whether the GPUs were attached locally or were accessed from remote servers using vSphere Bitfusion. Industry benchmarks were used with a variety of models, batch sizes and configurations. The results are shown in the chart below. Across models, batch sizes, and tests, Mellanox and vSphere Bitfusion demonstrated that remote GPUs deliver similar execution and user experience. There were many other configurations included in the testing. For a full list of database benchmarks, please contact Mellanox or vSphere Bitfusion.



Intelligent Interconnects

The Mellanox portfolio of Intelligent Interconnects is the best-designed network fabric for AI and ML infrastructure. All technologies, InfiniBand, RoCE, and Ethernet, serve remote storage and remote GPU resources to any client on the network. The conceptual parallel to NVMe over Fabric for AI/ML is the CUDA over Fabric Architecture implementation that vSphere Bitfusion has developed and has recently honed to production availability.

Additional unique capabilities from Mellanox can further enhance Elastic AI deployments, such as Socket Direct, which can remove latency experienced from dual-socket CPUs passing traffic through the coherent bus. To provide additional color on the effective Mellanox intelligent interconnect capabilities, a comparison test bed was set up to show relative performance of RoCE and TCP/IP. The chart above shows how effective RoCE can be while deploying machine learning workloads. With RoCE, the user can get a similar experience (depicted as 100% bar) running the workload as in a native configuration. While TCP/IP (no RoCE) may provide sufficient performance in some instances, RoCE can be used to address the most demanding GPU training workloads.



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.

To find out more, visit our website: *www.mellanox.com*

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 vmware.com

Item No: VMW-0520-2210_VMW_CPBU Technical White Papers_BitfusionDocs_16 AI Composability and Virtualization_1.4_MMT