SOLUTION BRIEF

bitfusion

Elastic AI
Network Attached GPUs

AI Composability and Virtualization: Mellanox Network Attached GPUs

POOL, SHARE AND VIRTUALIZE YOUR GPU CLUSTER WITH MELLANOX LOW LATENCY AND HIGH THROUGHPUT NETWORK

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

Now with Bitfusion software platform 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 a very narrow and specialize workloads in an organization. Due to few developers having privileged access to the GPU servers it drives low utilization and large bodies of AI researchers, ML developers and data scientists cannot get access to the precious GPUs resource.

The few fortunate AI developers that were allocated GPUs run-time, now need to exercise a painful migration of their applications environment and data to the GPU server – not a productive process, which will further hog the GPU server, and lower their utilization.

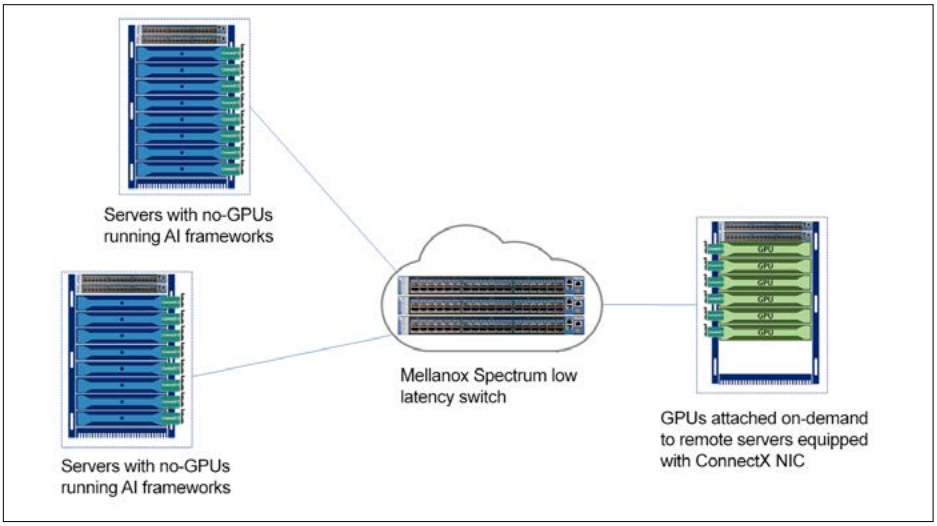
To further aggravate the problem, there is no viable way to sub segment physical GPUs to small 'quantities' and allow different users/workloads to consume less than one GPU remotely (e.g. for development and test, inference workloads, etc.). Workloads are forced to use GPUs as designed in hardware, without the ability to virtualize a GPUs resources.

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 in 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, 200Gb/s End-to-End Intelligent Interconnect Solutions

SOLUTION

With Bitfusion Elastic software platform and Mellanox networking solutions, GPUs can be sliced and diced and be connected remotely to any client across the network. Very much like NVME over Fabric, GPUs become composable resource reachable and accessible by any remote node. And much like storage, where the network has to be low latency and high throughput, Mellanox technology provides the networking Fabric. 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 applications run time, and thereafter disconnect. Utilization metrics go 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.

HOW DOES IT WORK

Implementing Elastic AI with Bitfusion and Mellanox is straight forward. Bitfusion provide OS and Hypervisor independent software which is installed in user space in each GPU server and each compute server. Mellanox provides the network fabric (NIC and switch).

Any GPU server and any compute server will work (there is no need for particular hardware or memory design/configuration). With minimal steps implemented, users can run AI workloads from any one of the servers. With 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 Bitfusion in place, users can share and pool common GPU resources. Users are also not bounded by their physical location and their software environment. All they need to do is launch any ML or AI workload (unmodified), and Bitfusion will attach the remote GPUs.

Organizations now can plan ahead and provide GPU resources, efficiently, with speed and agility to all developers and production teams. Mellanox Spectrum™ switches, ConnectX® NICs, BlueField SmartNICs and LinkX® cable and transceivers interconnect products offer seamless experience to the users, as if each user has a local GPU (or GPUs) attached to their servers. Bitfusion software works with any environment: containers, virtual machines and bare-metal. It also can connect two dissimilar eco systems (e.g. VMware at the client server side, and bare-metal on the GPU server side). The broad Mellanox portfolio of InfiniBand, RoCE and Ethernet technologies are a perfect match for Elastic AI and complement the Bitfusion software platform. Very much like storage, where extreme performance is needed for very demanding workloads, Bitfusion performance delivers 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 Bitfusion offer RoCE for remote direct memory access to boost network and host performance through lower latency, lower CPU requirements and higher bandwidth.

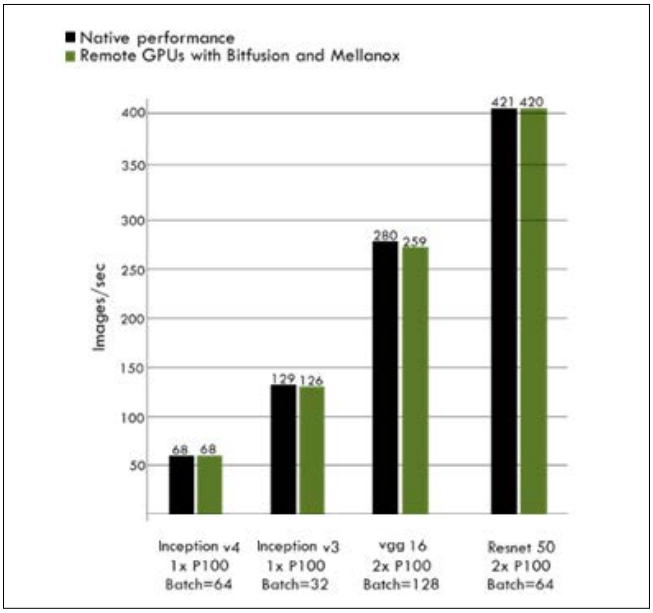
Finally, as Ethernet is the most ubiquitous network protocol in Enterprises, Bitfusion delivers a full gamut of Elastic AI infrastructure configuration options that enables higher GPU efficiency. In addition the solution can also operate in heterogeneous networking environments. For example, GPUs can be connected with InfiniBand, while high profile customers can utilize RoCE for access, and others can use common Ethernet for access.

BENCHMARK AND TESTS

ML and AI workloads stress the throughput and latency of the network. Mellanox and Bitfusion set the following configuration to emulate a real life Elastic AI infrastructure. The test bed relied on general available servers (both compute and GPU servers) and 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 Bitfusion doesn't necessitate any changes to the OS, drivers, kernel modules or AI frameworks.

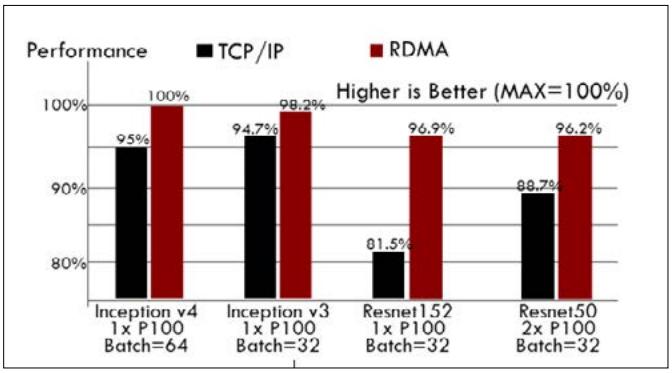


The intent of the tests were to prove the AI developer experience is the same, as if the GPUs are attached locally to the servers where the workloads are being executed, compared with executing the CUDA calls in a remote GPU (or GPUs). Industry benchmarks were used with a variety of models, batch sizes and configurations. The results are shown in the chart above. Across models, batch sizes and tests, Mellanox and Bitfusion demonstrated that remote GPUs will accomplish similar user experience and execution. There were many other configurations included in the testing. For a full list of database benchmarks, please contact Mellanox or 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 Bitfusion has been working on for the last few years, and has recently honed to production availability.

Additional unique capabilities from Mellanox can further enhance Elastic AI deployments such as GPUDirect which is used to tie multiple GPU servers together and offers remote clients GPU resources from separate GPU servers. Another technology, Socket Direct, 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 to show 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



similar experience (depicted as 100% bar) running the workload 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

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

060244
Rev1.1