Reference Architecture: Acceleration over PCIe for Dell EMC PowerEdge MX7000

Tech Note by

Ramesh Radhakrishnan
Seamus Jones

Summary

Many of today's most demanding applications can make use of PCIe acceleration. Liqid designed on Dell Technologies, enables the rapid and dynamic provisioning of PCIe resources such as GPU, FPGA, or NVMe to Dell EMC PowerEdge MX7000 compute sleds. Ensuring workload performance needs are met for the most accelerator hungry applications.

Request a demo or quote from a Dell Technologies Design Solutions Expert
Design Solution Portal

Background

The Dell EMC PowerEdge MX7000 Modular Chassis simplifies the deployment and management of today's most challenging workloads by allowing IT administrators to dynamically assign, move and scale shared pools of compute, storage and networking resources. It provides IT administrators the ability to deliver fast results, eliminating managing and reconfiguring infrastructure to meet ever-changing needs of their end users. The addition of PCIe infrastructure to this managed pool of resources using Liqid technology designed on Dell EMC MX7000 expands the promise of software-defined composability for today's AI-driven compute environments and high-value applications.

GPU Acceleration for PowerEdge MX7000

For workloads like AI that require parallel accelerated computing, the addition of GPU acceleration within the PowerEdge MX7000 is paramount. With Liqid technology and management software, GPUs of any form factor can be quickly added to any new or existing MX compute sled via the management interface, quickly delivering the resources needed to manage each step of the machine learning workflow including data ingest, cleansing, training, and inferencing. Spin-up new bare-metal servers with the exact number of accelerators required and then dynamically add or remove them as workload needs change.

Essential PowerEdge Expansion Components

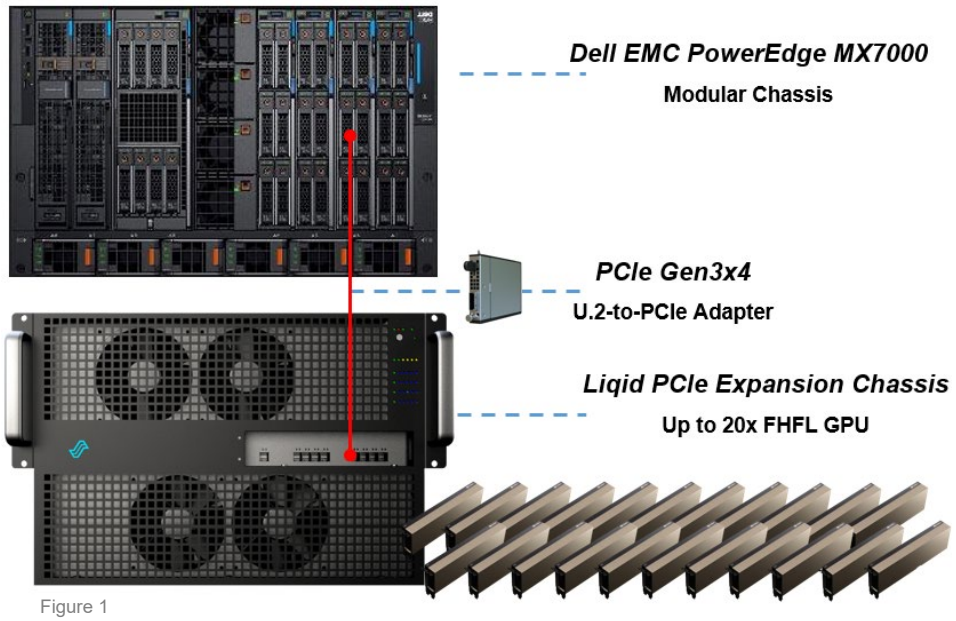

Figure 1

GPU Expansion Over PCIe	
Compute Sleds	Up to 8 x Compute Sleds per Chassis
GPU Chassis	PCIe Expansion Chassis (Contain GPU or other devices Direct Connect to Compute Sled)
Interconnect	PCIe Gen3 x4 Per Compute Sled (Multiple Gen 3 x4 Links Possible)
GPU Expansion	20x GPU (FHFL)
GPU Supported	V100, A100, RTX, T4, Others
OS Supported	Linux, Windows, VMWare and Others
Devices Supported	GPU, FPGA, and NVMe Storage
Form Factor	14U Total = MX7000 (7U) + PCIe Expansion Chassis (7U)

Table 1

Implementing GPU Expansion for MX

GPUs are installed into the PCIe expansion chassis. Next, U.2 to PCIe Gen3 adapters are added to each compute sled that requires GPU acceleration, and then they are connected to the expansion chassis (Figure 1). Liqid Command Center software enables discovery of all GPUs, making them ready to be added to the server over native PCIe. FPGA and NVMe storage can also be added to compute nodes in tandem. This PCIe expansion chassis & software are available from the Dell Design Solutions team.

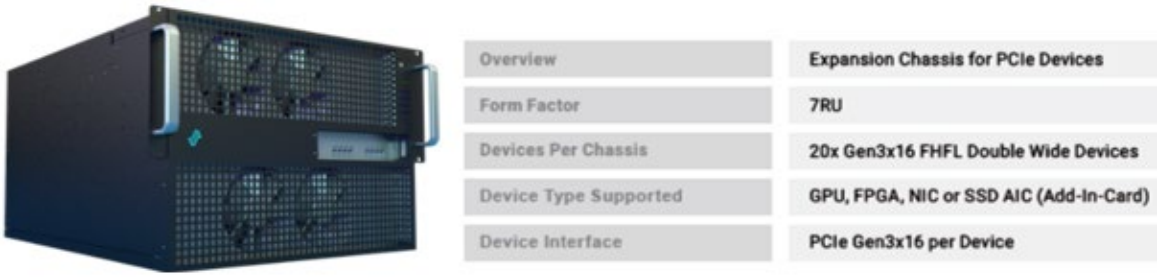

Figure 2

Software Defined Composability

Once PCIe devices are connected to the MX7000, Liqid Command Center software enables the dynamic allocation of GPUs to MX compute sleds at the bare metal (GPU hot-plug supported). Any amount of resources can be added to the compute sleds, via Liqid Command Center (GUI) or RESTful API, in any ratio to meet the end user workload requirement To the operating system, the GPUs are

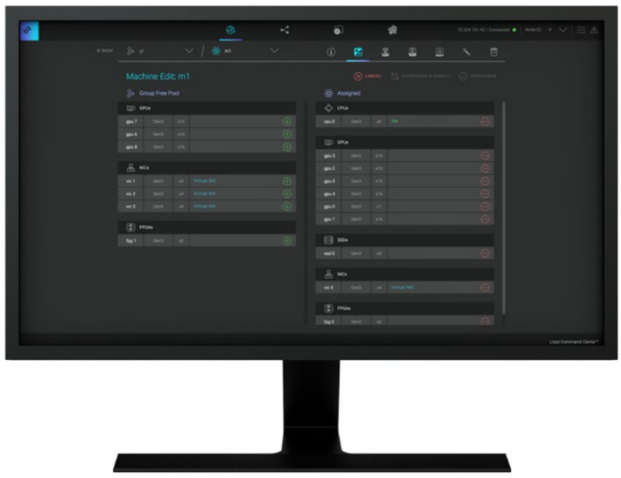

Figure 3 Liqid Command Center

presented as local resources direct connected to the MX compute sled over PCIe (Figure 3). All operating systems are supported including Linux, Windows, and VMware. As workload needs change, add or remove resources on the fly, via software including NVMe SSD and FPGA (Table 1).

Enabling GPU Peer-2-Peer Capability

A key feature included with the PCIe expansion solution for PowerEdge MX7000 is the ability for RDMA Peer-2-Peer between GPU devices. Direct RDMA transfers have a massive impact on both throughput and latency for the highest performing GPU-centric applications. Up to 10x improvement in performance has been achieved with RDMA Peer-2-Peer enabled. Below is the overview of how PCIe Peer-2-Peer functions (Figure 4).

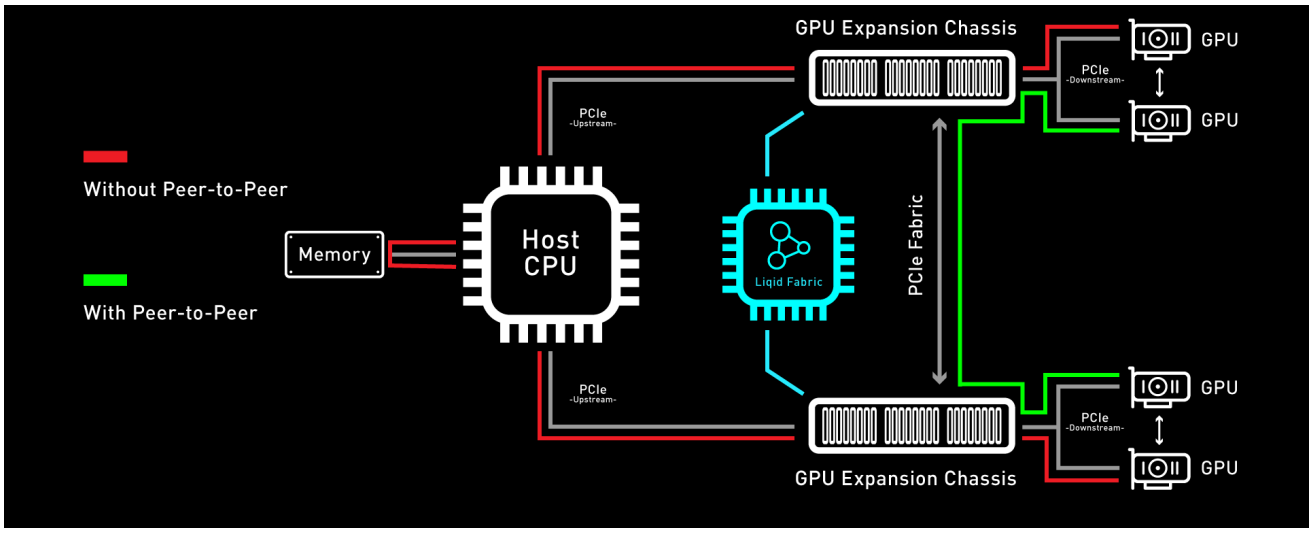

Figure 4 PCIe Peer-2-Peer

Bypassing the x86 processor and enabling direct RDMA communication between GPUs, realizes a dramatic improvement in bandwidth and in addition a reduction in latency is also realized. This chart outlines the performance expected for GPUs that are composed to a single node with GPU RDMA Peer-2-Peer enabled (Table 2).

	Bandwidth (GB/s)	Latency (µs)
Peer-To-Peer Disabled	8.59	33.65
Peer-To-Peer Enabled	25.01	3.1
Delta	291%	1085%

Table 2

Application Level Performance

RDMA Peer-2-Peer is a key feature in GPU scaling for Artificial Intelligence, specifically machine learning based applications. Figure 5 outlines performance data measured on mainstream AI/ML applications on the MX7000 with GPU expansion over PCIe. It further demonstrates the performance scaling from 1-GPU to 8-GPU for a single MX740c compute sled. High scaling efficiency is observed for ResNet152, VGG16, Inception V3, and ResNet50 on MX7000 with composable PCIe GPUs measured with Peer-2-Peer enabled. These results indicate a near-linear growth pattern, and with the current capabilities of the Liqid PCIe 7U expansion sled one can allocate up to 20 GPUs to an application running on a single node.

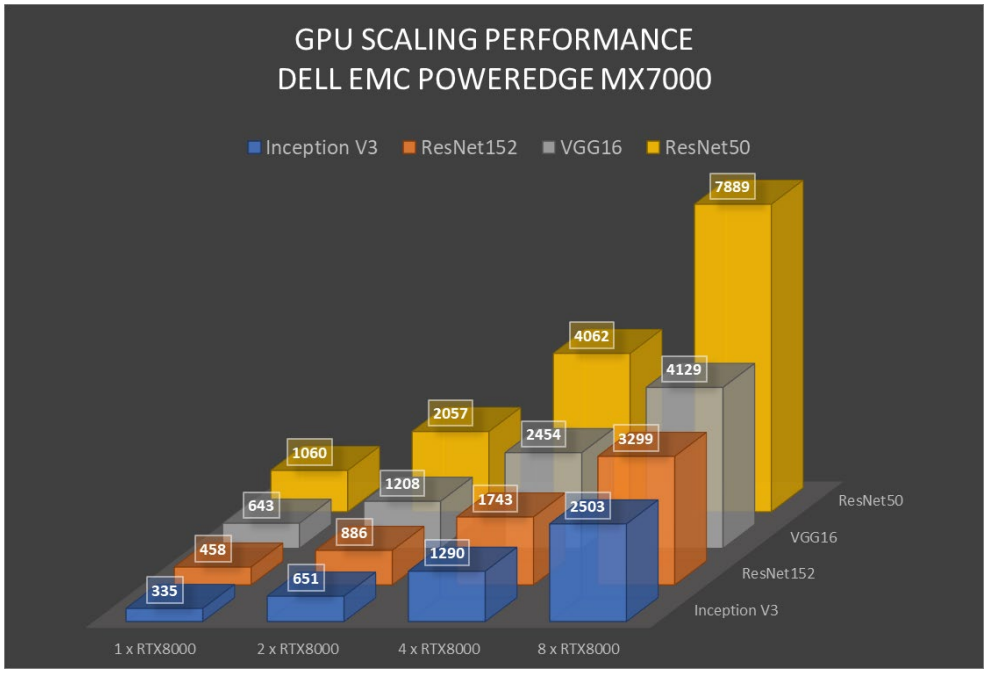

Figure 5 GPU Performance Scaling Comparison

- MX7000 Leverages RTX8000 in PCIe expansion chassis measured with P2P Enabled

Conclusion

Liqid PCIe expansion for the Dell EMC PowerEdge MX7000 unlocks the ability to manage the most demanding workloads in which accelerators are required for both new and existing deployments. Liqid collaborated with Dell Technologies Design Solutions to accelerate applications by through the addition of GPUs to the Dell EMC MX compute sleds over PCIe.

Learn More | See a Demo | Get a Quote
This reference architecture is available as part of the Dell Technologies Design Solutions.
Contact a Design Expert today https://www.delltechnologies.com/en-us/oem/index2.htm#open-contact-form