Introduction

Research and development (R&D) for machine learning (ML) has blossomed with large investments from public cloud giants such as Amazon Web Services (AWS), Microsoft Azure, Baidu Cloud, Google Cloud Platform (GCP), and others. However, due to the size of ML training data sets, plus regional and vertical market compliance regulations, many customers are opting to deploy private cloud-based ML solutions.

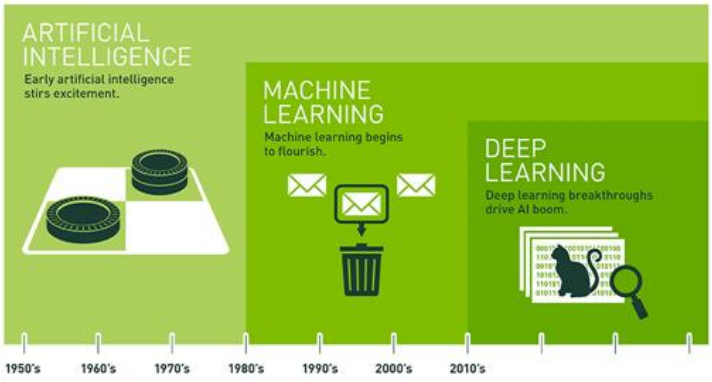
In November 2014, Dell was the first datacenter solution provider to bring to market a dense, four graphics processing unit (GPU) solution in a 1U form factor, the PowerEdge C4130. Three years later, in November 2017, Dell EMC launched its PowerEdge C4140 successor to the C4130, along with its Ready Solutions for Machine and Deep Learning. These ML solutions use Dell EMC PowerEdge C4140, R740, and T640 servers to host a variety of accelerators, including GPUs, with an array of configuration options.

On the silicon side, NVIDIA invested early in both software tools development and new GPU architectures to enable and accelerate ML. In 2017, Dell EMC and NVIDIA announced a strategic agreement to jointly develop new datacenter products based on NVIDIA's Volta generation GPUs, specifically for high-performance computing (HPC), data analytics, and artificial intelligence (AI) workloads. TIRIAS Research believes the agreement, coupled with continuous product line updates, will generate momentum for Dell EMC in ML applications.

Hardware Plus Software Drives Machine Learning Progress

Machine learning is a key component of artificial intelligence (AI). Deep learning (DL) is a set of learning techniques within ML, but ML contains a more diverse set of learning techniques.

Figure 1: AI, ML, & DL Relationships



Source: NVIDIA

New DL techniques underlie recent ML advances, and therefore AI breakthroughs, such as natural language processing and autonomous vehicles. The heart of DL and more general ML algorithms is matrix multiply operations. Traditional CPU cores are much slower and less energy efficient at matrix multiplies than GPUs. CPUs are optimized to run branchy, lightly-threaded code, while GPUs are optimized to run highly parallel workloads.

ML has two phases. Each phase has different workload characteristics:

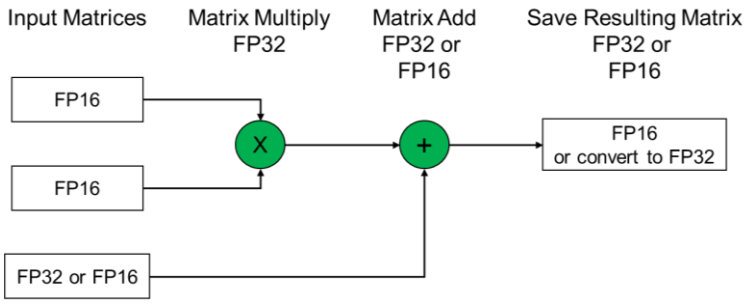
- **Training** feeds massive amounts of representative data through a neural network to train the network to recognize patterns in the data and to optimize the network framework. The training phase often requires higher-precision floating point math (FP32 32-bit single-precision) to maintain enough accuracy across each network layer and through the many layers of a deep neural network.
- **Inference** is the production end of a neural network, where a service presents data to a trained neural network, and the trained network then identifies patterns in the data. Less accuracy is required for inference, and many models can use lower-precision floating-point math (FP16 16-bit half-precision and INT8 single-byte integer precision). A trained model can incrementally learn during inference, but that learning enables only minor optimizations to the network model. While trained network models can be loaded into endpoints to enable local inference, many cloud-based services perform inference at cloud scale.

NVIDIA GPUs & Software Platform

Over the past two decades, NVIDIA has tuned its GPU designs to accelerate 3D graphics by processing matrix multiplies faster and more efficiently with each subsequent GPU core generation. NVIDIA's Pascal generation GPUs introduced FP16 and (on Tesla P4 and P40 products) INT8 operations. NVIDIA's current Volta generation introduced a "Tensor Core" that is designed to very efficiently process 4x4 FP16 matrix warps with FP32 accumulates.

Tensor Core and INT8 operations are DL specific innovations. INT8 helps accelerate inference tasks by 2-3x performance, with little loss in accuracy, while Tensor Core enables Volta generation GPUs to process DL training tasks at about 3x the rate of Pascal generation GPUs.

Figure 2: NVIDIA Volta Generation Tensor Core Flow Diagram



Source: TIRIAS Research

Software developers access NVIDIA's GPU compute capabilities through NVIDIA's CUDA parallel computing platform and programming model. ML software developers have relied on CUDA 8 as their baseline DL platform. In September 2017, NVIDIA launched CUDA 9 support for programming Volta's Tensor Cores. NVIDIA CUDA Deep Neural Network library (cuDNN) provides a higher-level deep learning application programming interface (API) that is used by

leading deep learning frameworks, such as Caffe 2, TensorFlow, Theano, Torch, MXNet, Microsoft Cognitive Toolkit (CNTK), and more.

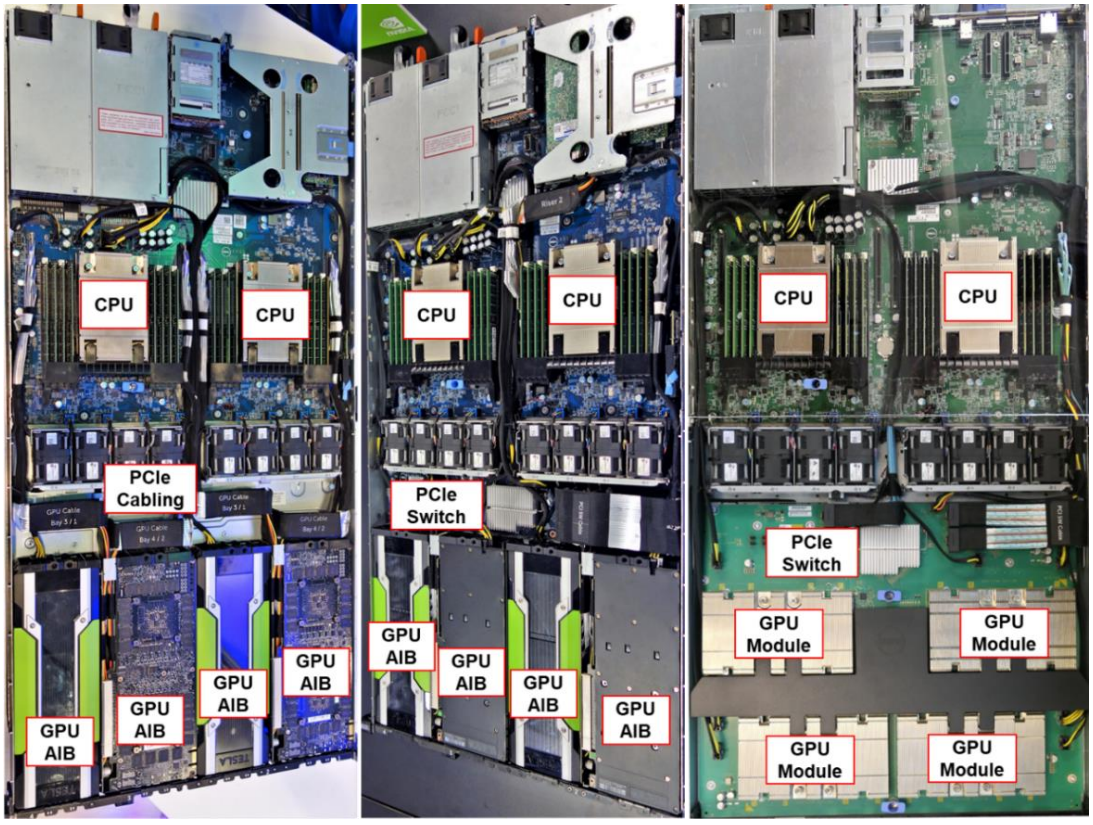
NVIDIA also offers TensorRT 3, a programmable inference accelerator. TensorRT 3 is designed to optimize, validate, and deploy trained neural networks for inferencing at scale. Target markets for TensorRT include cloud-scale inference-as-a-service, as well as embedded and automotive inferencing products.

NVIDIA's GPU Cloud (NGC) is a set of GPU-optimized software development and deployment tools for DL and HPC. NGC's container registry offers NVIDIA tuned, tested, certified, and maintained containers for deploying the most widely used DL frameworks and TensorRT 3.

Dell EMC PowerEdge C4140 Server

The workhorse of Dell EMC's GPU-accelerated line-up is its new 1U PowerEdge C4140 refresh. The PowerEdge C4140 has a modular design, available in three configuration options based on the CPU-to-accelerator board interconnect and the accelerator form factor used.

Figure 3: PowerEdge C4130 / C4140 CPU to GPU Connection Options



Source: TIRIAS Research

Left and middle photos are prototype PowerEdge C4130 systems photographed in 2016; right photo is a production system photographed in July 2017. Visually, the C4130 and C4140 are difficult to tell apart. The differences between the above systems include:

- **Left**: Direct PCIe cabling between CPU sockets and GPU AIBs
- **Middle**: PCIe switch between CPUs and GPU AIBs
- **Right**: PCIe switch between CPUs and either P100 or V100 SXM2 modules, SXM2 modules are connected to each other via NVLink; black shroud between P100 modules directs air flow

The PowerEdge C4140 configuration using a PCIe switch implements all four GPU add-in boards (AIBs) or modules. The added cost and power consumption of the switch is minor compared to the four GPUs, so it makes economic sense to include the switch in the highest performance options, such as the V100 SXM2 configuration where each module is directly connected to the switch via PCIe. The four V100 SXM2 modules are also directly connected with each other by NVLink in a fully-connected fashion, which enables each GPU to communicate with both CPUs and the other three GPUs with the lowest latency and highest bandwidth possible.

Table 1: Dell EMC Machine & Deep Learning Reference Configurations

Configuration	Inference	Training		Medium "K"		Large "K" *	
PowerEdge Server	**R740**	**T640**	**R740**	**C4140**		**C4140**	
Node Count	1	1	1	1	1	4	4
Accelerator Count	3	4	3	4	4	4	4
Accelerator Type	PCIe	PCIe	PCIe	SXM2		SXM2	
NVIDIA Tesla Model	**P40**	**V100**	**V100**	**P100**	**V100**	**P100**	**V100**
Total FP32 TFLOPS	36	56	42	42	63	170	251
Total FP16 TFLOPS	N/A	448	336	85	500	339	2,000
Total INT8 TFLOPS	141	228	171	N/A	252	N/A	1,008
Total GPU Power	**750**	**1,000**	**750**	**1,200**	**1,200**	**4,800**	**4,800**

* Includes Dell Storage MD1280, PowerEdge R740xd head node, and Mellanox InfiniBand cluster networking

Source: Dell EMC & NVIDIA

PowerEdge C4140 machine learning reference designs include:

- NVIDIA Tesla P100 or V100 SXM2 modules, each with 16GB of memory
- Dual Intel Xeon Gold Scalable 6148 processors (Skylake / Purley generation)
- 384GB of DDR4 memory at 2667MHz
- Two 120GB M.2 SSDs
- Bright Computing's Bright Cluster Manager

Large "K" four-node reference designs include a Mellanox InfiniBand EDR 100Gbps PCIe NIC plus a Mellanox ConnectX-4 Virtual Protocol Interconnect (VPI) PCIe card in each node, plus a Mellanox SwitchIB2-based EDR InfiniBand 1U switch.

Customers using Ethernet will most likely insert their preferred NICs into any of these configurations.

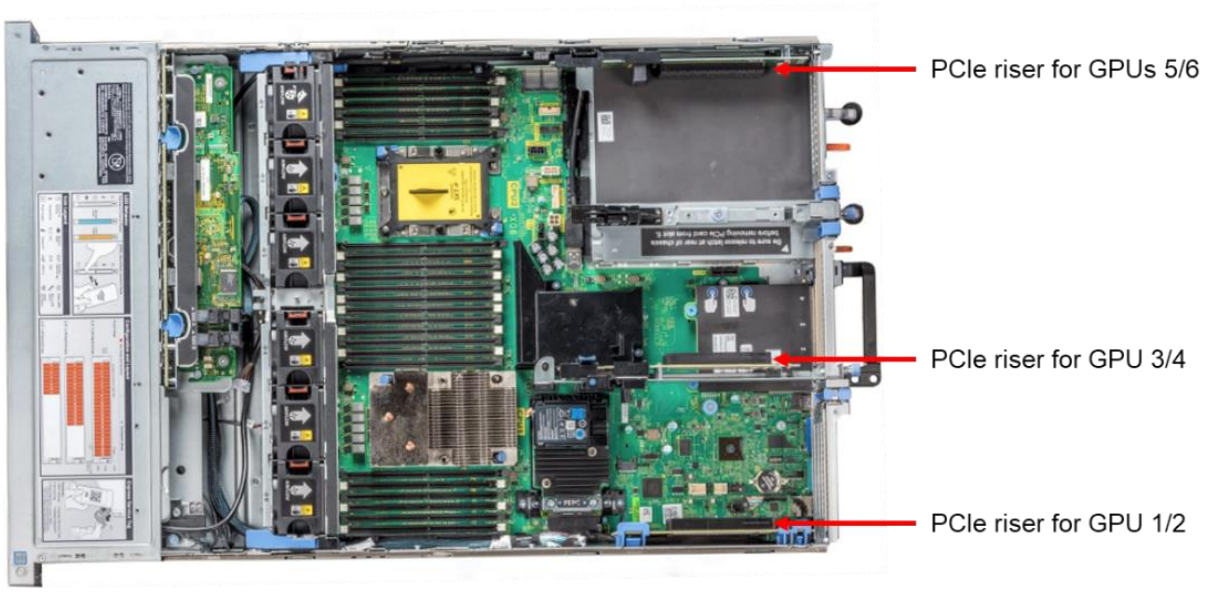
Dell EMC has a longstanding relationship with Bright Computing, whose mission is to make smaller HPC installations (under 500 compute nodes) manageable by non-HPC specialist IT staff with minimal exposure to HPC systems. Dell EMC already pre-loads Bright Cluster Manager (BCM) for HPC customers.

Dell EMC PowerEdge R740 Server

Dell EMC's 2U PowerEdge R740 dual-socket Intel Xeon Scalable server sports three full-sized, double wide AIB bays, each with PCIe x16 connectors (Figure 4). For smaller, experienced data analytics and machine learning customers, the PowerEdge R740 may be a good entry-level choice compared to the PowerEdge C4140 directly-cabled PCIe AIB baseline configuration.

The PowerEdge R740 can host three full-length, double-wide PCIe x16 GPU AIBs, such as NVIDIA's Tesla P40, P100, and V100 AIBs.

Figure 4: PowerEdge R740 Chassis Showing PCIe Risers for GPUs



Source: Dell EMC

Dell EMC PowerEdge R7425 Server

Dell EMC's 2U PowerEdge R7425 dual-socket AMD EPYC server also sports three full-sized, double wide AIB bays, each with PCIe x16 connectors, and can host three full-length, double-wide PCIe x16 GPU AIBs. Like Dell EMC's PowerEdge R740, its PowerEdge R7425 may be a good entry-level choice compared to the PowerEdge C4140 PCIe baseline configuration.

Dell EMC PowerEdge T640 Server

Like the PowerEdge C4140, Dell EMC's 5U PowerEdge T640 Tower Server can host up to four 300W PCIe accelerator AIBs (or up to nine smaller AIBs). The T640 is a good low-density choice for data scientists and machine learning researchers and modelers who prefer a desk-side appliance that can natively host up to 18 3.5-inch SATA and / or SAS storage drives (plus more with an optional "flex bay"). Typically, these customers are do-it-yourself upgraders who will buy and install new GPU AIBs as they can afford to or as newer, faster AIBs enter the market.

Other Dell EMC & Dell Machine Learning Solutions

Dell's Precision 7000 series dual-socket workstations are also an entry-level choice for students and researchers. Dell's Precision workstations host a wide variety of NVIDIA GPU AIBs, including NVIDIA Pascal generation AIBs and future Volta generation AIBs.

Summary

NVIDIA's Pascal and Volta generation GPUs are the *de facto* standard for accelerating DL today. Most case studies for accelerating DL use NVIDIA GPUs. NVIDIA's Volta GPU, with its CUDA 9 enabled Tensor Core, should help NVIDIA maintain its market leading position.

Dell EMC's joint development and Volta launch agreement with NVIDIA has helped both Dell EMC and NVIDIA. Dell EMC has early access to NVIDIA architectural improvements, while NVIDIA has access to Dell EMC's enterprise datacenter marketing acumen and reach. In addition, Dell EMC's partnership with Bright Computing will enable mainstream enterprise customers to start evaluating how to make ML work for them.

TIRIAS Research recommends that machine learning customers evaluate Dell EMC's PowerEdge C4140 ML reference configurations using NVIDIA Tesla P40, P100 or V100 SXM2 modules. Based on intended training and/or inference workloads, customers new to ML can start off with P100 modules and then decide whether to and when to move to P40 or V100 products as they characterize their ML workloads. For the most flexibility across DL training and inference, as well as HPC workloads, Tesla V100 is appropriate. For better scale-out performance for inference, Tesla P40 may be the right choice.

At this early stage of AI, ML, and DL market evolution, customers will need to experiment with different system configurations for different ML models—there are no good guides to matching ML models to specific hardware configurations for optimal performance. It may take a decade or more of modeling experience to determine the right balance of processors and GPUs for different applications.