DSS 8440: More raw power for machine learning

The machine learning challenge

Data center workloads continue to evolve in challenging ways as the computing landscape responds to the rapid advancement of new technologies. The availability of massive amounts of data – both structured and unstructured – and the emergence of cloud native applications – with their demands for higher throughput and parallel computing – are driving data centers to look for more advanced processing solutions to incorporate into their existing infrastructures. In particular, they are looking for accelerator solutions that deliver more computing horsepower than the general-purpose CPUs that are becoming a bottleneck for overall processing.

Introducing the DSS 8440: a dynamic machine learning server

The DSS 8440 is a 4U 2-socket accelerator-optimized server designed to deliver exceptionally high compute performance. Its open architecture, based on a high performance switched PCIe fabric, maximizes customer choice for machine learning infrastructure while also delivering best of breed technology (#1 server provider & the #1 GPU provider). It lets you tailor your machine learning infrastructure to your specific needs – with open, PCIe based components.

With a choice of 4, 8 or 10 industry-leading NVIDIA® Tesla® V100 Tensor Core GPUs, combined with 2 Intel CPUs for system functions, the PCIe fabric for rapid IO and up to 10 local NVMe and SAS drives for optimized access to data, this server has both the performance and flexibility to be an ideal solution for machine learning training - as well as other compute-intensive workloads like simulation, modeling and predictive analysis in engineering and scientific environments.

The DSS 8440 and machine learning

Machine learning encompasses two distinctly different workloads: training and inference. While each benefits from accelerator use, they do so in different ways, and rely on different accelerator characteristics that may vary from accelerator to accelerator. The initial release of the DSS 8440 is specifically targeted at complex, training workloads. It provides more of the raw
compute horsepower needed to quickly process the increasingly complicated models that are being developed for complex workloads like image recognition, facial recognition and natural language translation.

At the simplest level, machine learning **training** involves “training” a model by iteratively running massive amounts of data through a weighted, multi-layered algorithm (thousands of times!), comparing it to a specifically targeted outcome and iteratively adjusting the model/weights to ultimately result in a “trained” model that allows for a fast and accurate way to make future predictions. **Inference** is the production or real-time use of that trained model to make relevant predictions based on new data.

**Training workloads** demand extremely high-performance compute capability. To train a model for a typical image recognition workload requires accelerators that can rapidly process multiple layers of matrices in a highly iterative way - accelerators that can scale to match the need. NVIDIA® Tesla® V100 Tensor Core GPUs are such an accelerator. The DSS 8440 with NVIDIA GPUs and a PCIe fabric interconnect has demonstrated scaling capability to near-equivalent performance to the industry-leading DGX-1 server (within 5%) when using the most common machine learning frameworks (i.e., TensorFlow) and popular convolutional neural network (CNN) models (i.e., image recognition).

Note that Dell EMC is also partnering with the start-up accelerator company Graphcore to achieve new levels of training performance. Graphcore is developing machine learning specific, graph-based technology to enable even higher performance for training workloads. Graphcore accelerators will be available with DSS 8440 in a future release. See the Graphcore sidebar for more details.

**Inference workloads**, while still requiring acceleration, do not demand as high a level of performance, because they only need one pass through the trained model to determine the result.

However, inference workloads demand the fastest possible **response time**, so they require accelerators that provide lower overall latency. While this release of the DSS 8440 is not targeted for inference usage, note that the accelerator card that Graphcore is developing can support both training
and inference. (It lowers overall latency by loading the full machine learning model into accelerator memory.)

Exceptional throughput performance

With the ability to scale up to 10 accelerators, the DSS 8440 can deliver higher performance for today’s increasingly complex computing challenges. Its low latency, switched PCIe fabric for GPU-to-GPU communication enables it to deliver near equivalent performance to competitive systems that are based on the more expensive SMX2 interconnect. In fact, for the most common type of training workloads, not only is the DSS 8440 throughput performance exceptional, it also provides better power efficiency (performance/watt).

Most of the competitive accelerator optimized systems in the marketplace today are 8-way systems. An obvious advantage of the DSS 8440 10 GPU scaling capability is that it can provide more raw horsepower for compute-hungry workloads. More horsepower that can be used to concentrate on increasingly complex machine learning tasks, or conversely, may be distributed across a wider range of workloads – whether machine learning or other compute-intensive tasks. This type of distributed, departmental sharing of accelerated resources is a common practice in scientific and academic environments where those resources are at a premium and typically need to be re-assigned as needed among dynamic projects.

Better performance per watt

One of the challenges faced as accelerator capacity is increased is the additional energy required to drive an increased number of accelerators. Large scale data centers understand the importance of energy savings at scale. The DSS 8440 configured with 8 GPUs has proven to be more efficient on a performance per watt basis than a similarly configured competitive SMX2-based server – up to 13.5% more efficient.

That is, when performing convolutional neural network (CNN) training for image recognition it processes more images than the competitive system, while using the same amount of energy. This testing was done using the most common machine learning frameworks – TensorFlow, PyTorch and MXNet – and in all three cases the DSS 8440 bested the competition. Over time, and at data center scale, this advantage can result in significant operational savings.
Accelerated development with NVIDIA GPU Cloud (NGC)

When the DSS 8440 is configured with NVIDIA V100 Tensor Core GPUs you get the best of both worlds - working with the world's #1 server provider (Dell EMC) and the industry's #1 provider of GPU accelerators (NVIDIA). In addition, you can take advantage of the work NVIDIA has done with NVIDIA GPU Cloud (NGC), a program that offers a registry for pre-validated, pre-optimized containers for a wide range of machine learning frameworks, including TensorFlow, PyTorch, and MXNet. Along with the performance-tuned NVIDIA AI stack these pre-integrated containers include NVIDIA® CUDA® Toolkit, NVIDIA deep learning libraries, and the top AI software. They help data scientists and researchers rapidly build, train, and deploy AI models to meet continually evolving demands.

Balanced design for better performance

Fast, efficient compute

The NVIDIA V100 Tensor Core is the most advanced data center GPU ever built to accelerate machine learning, high performance computing (HPC), and graphics. It supports a PCIe interconnect for GPU-to-GPU communication – enabling scalable performance on extremely large machine learning models, comes in 16 and 32GB configurations, and offers the equivalent performance of up to 100 CPUs in a single GPU. The PCIe-based GPU runs at 250W – 50W lower than the SMX2-based GPU – allowing for better power efficiency at maximum capacity than an 8-way SMX2-based system.
Wide range of IO options
Access to data is crucial to machine learning training. To that end, the DSS 8440 has 8 full height and 1 low profile x16 PCIe slots available for use in the rear of the server. (A tenth slot is reserved for a RAID storage controller.)

Extensive, high speed local storage
The DSS 8440 provides flexible local storage options for faster access to training data, with up to 10 drives, 2 fixed as SATA, 2 fixed as NVMe and 6 that can be either SATA or NVMe. For maximum performance it can be configured with up to 8 NVMe drives. (NVMe drives are 7 times faster than SATA SSDs.)

GRAPHCORE Innovative acceleration
Graphcore’s Intelligence Processing Unit (IPU) is completely different from today’s CPU and GPU processors. It is a highly flexible parallel processor that has been designed from the ground up to support the full range of machine learning workloads.

With 2432 independent IPU-Cores™ on each processor, it can deliver up to 2 PetaFLOPs of compute, enabling new levels of performance, and allows for features like:

Massive on-chip memory Avoids traditional latency penalties by utilizing up to 4.8GB of combined on-chip memory

Low Latency Make decisions faster with 720TB/s memory bandwidth

Poplar C++ programming framework provides an interface to standard machine learning frameworks, so applications written for existing frameworks may be easily optimized.

Graphcore uses high speed IPU-LINKS™ (2.5 Tb/s bandwidth) to connect the 8 IPU cards in the DSS 8440 to enable a shared pool of compute.

Future Release the Graphcore C2 accelerator card will be available in DSS 8440 servers later this year.
More power, more efficiency – the DSS 8440

Solve tougher challenges faster. Reduce the time it takes to train machine learning models with the scalable acceleration provided by the DSS 8440. Whether detecting patterns in online retail, diagnosing symptoms in the medical arena, or analyzing deep space data, more computing horsepower allows you to get better results sooner - improving service to customers, creating healthier patients, advancing the progress of research.

Now you can meet those challenges while simultaneously gaining greater energy efficiency for your datacenter. The DSS 8440 is the ideal machine learning solution for data centers that are scaling to meet the demands of today’s applications and want to contain the cost and inefficiencies that typically come with scale.

Contact the Dell EMC Extreme Scale Infrastructure team for more information about the DSS 8440 accelerator-optimized server. (ESI@dell.com)