DSS 8440: Flexible machine learning for data centers

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.

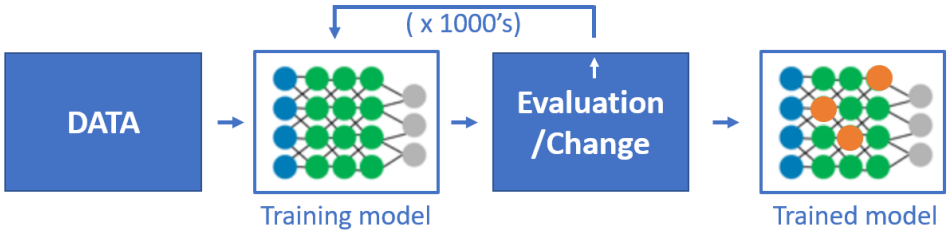
The DSS 8440: a highly flexible machine learning server

The DSS 8440 is a 4U 2-socket accelerator-optimized server designed to deliver exceptionally high compute performance for both training and inference. 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.

Choose between 4, 8 or 10 NVIDIA® **V100S GPUs** for the highest performance training of machine learning models or 4, 8, or 10 NVIDIA® Quadro™ **RTX GPUs** for a lower cost alternative with almost the same high performance, or select 8, 12 or 16 NVIDIA **T4 GPUs** to optimize in the inferencing phase. (Note, the lower cost, lower energy consuming T4 card is also an excellent option for *training* environments that do not require the highest levels of performance.) Combined with 2 second generation Intel Xeon CPUs for system functions, a 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 the widest range of machine learning solutions - 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 was specifically targeted at complex, **training** workloads. By implementing up to 10 V100 GPUs it provided more of the raw compute horsepower needed to quickly process the increasingly complicated models being developed for complex workloads like image recognition, facial recognition and natural language translation.



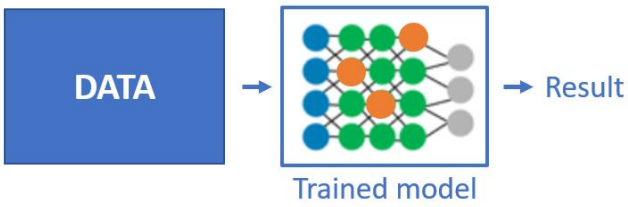
Machine learning training flow

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® V100S 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, that is developing machine learning specific, graph-based technology to enable even higher performance for ML workloads. The DSS 8440 with Graphcore accelerators is available to early adopter customers with extensive experience in machine learning. 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.

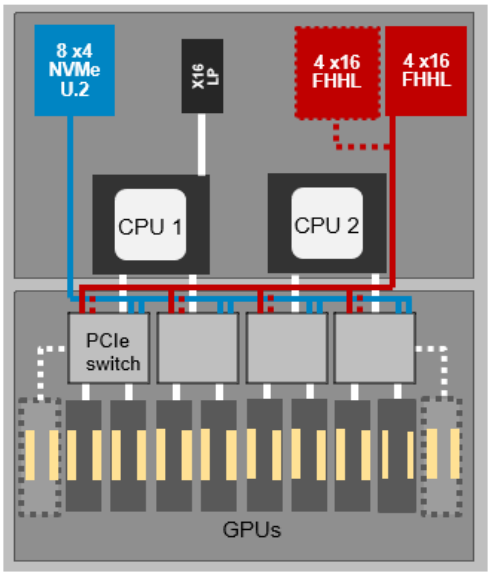


However, inference workloads demand the fastest possible *response time*, so they require accelerators that provide lower overall latency. Dell EMC is now supporting the use of up to 16 NVIDIA T4 GPUs for use with the DSS 8440.

While the T4 GPU provides less overall performance than the V100 (640 cores vs 320 cores), it supplies more than enough to deliver superb inference performance – and it does so while using less than 30% of the energy, only 70 watts per GPU.

V100S TRAINING: 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).



DSS 8440 topology – up to 10 V100S GPUs

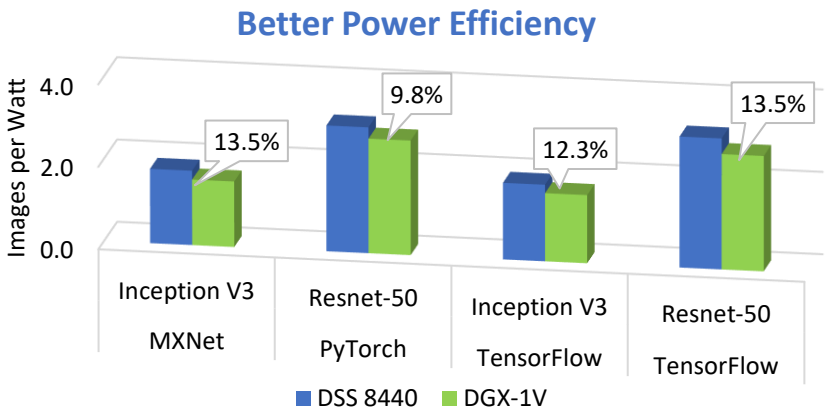
Most of the competitive accelerator optimized systems in the marketplace today are 8-way systems. An obvious advantage of the DSS 8440 10 GPU scaling capability is that it can provide more raw horsepower for compute-hungry workloads. More horsepower that can be used to concentrate on increasingly complex 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 V100 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.



The NVIDIA Quadro RTX GPUs

The DSS 8440 also supports NVIDIA Quadro RTX GPUS. The NVIDIA® RTX 8000 passive GPU is dual wide with **48 GB** GDDR6 memory and a 250W maximum power limit, passively cooled. The NVIDIA® RTX 6000 passive GPU is dual wide with **24 GB** GDDR6 memory and a 250W maximum power limit, also passively cooled. The DSS 8440 will support 4, 8 and 10 GPU configurations. They offer a lower cost Machine Learning training alternative for customers that want high performance ML, but don't require the absolute top end performance of V100S GPUs. These two mid-range accelerators are roughly 30% less cost than the V100 GPU while still being approximately 70% as performant.

Rendering Workloads

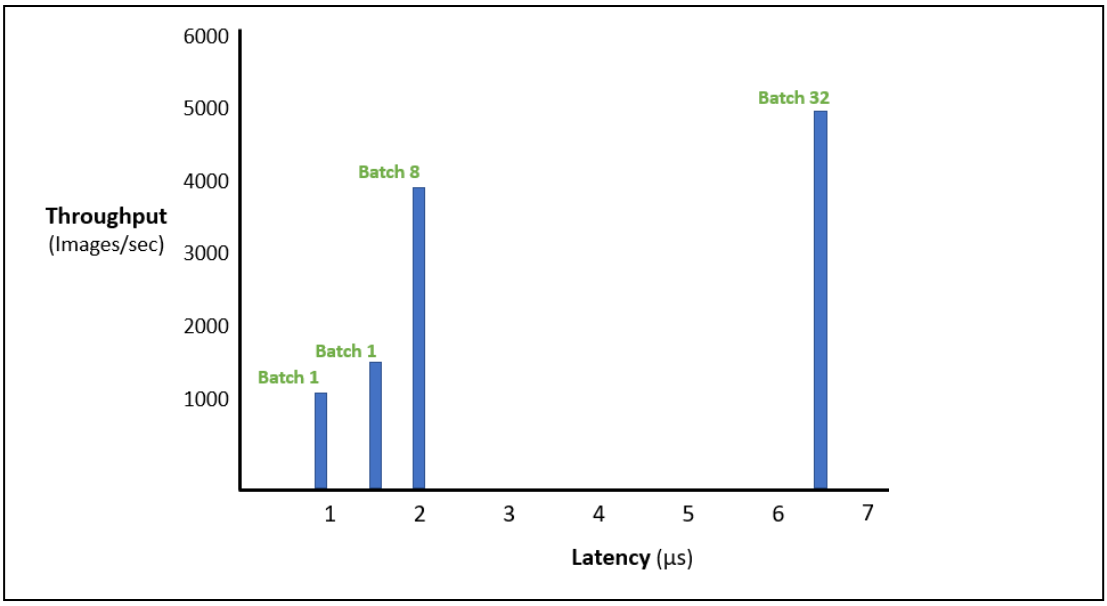
Based on the NVIDIA Turing™ architecture and the NVIDIA RTX™ platform, RTX 6000 and RTX 8000 GPUs feature RT Cores and multi-precision Tensor Cores for real-time ray tracing, AI, and advanced graphics capabilities. These specialized graphics functions also enable large scale graphics rendering capability on the dense, up to 10-way, DSS8440. The exceptionally large video memory on the RTX GPUS enable them to make better use of the high number of CUDA cores they have for rendering.

Note: The DSS 8440 does NOT support the NVIDIA NVLink™ bridge (typically used to connect 2 RTX GPUs to scale performance and memory capacity).

T4 INFERENCE

The DSS 8440 with NVIDIA® T4 GPUs offers high capacity, high performance machine learning inference with exceptional energy and cost savings. Customers can choose to implement 8, 12 or 16 T4 GPUs for compute resource, and because inference is typically a single accelerator operation (no need to scale across GPU cards) the DSS 8440's high capacity for accelerators enables an extremely flexible multi-tenancy environment. It allows data centers to share those inference resources among multiple users and departments - easily and with no loss of performance.

T4 GPUs in the DSS 8440 have demonstrated average throughput of nearly 3900 images per second at a 2.05 millisecond latency (at batch size 8) using the ResNet50 model. As with Training, performance for inference can be significantly impacted by batch size. Latency and throughput fluctuate depending on the amount of data being processed simultaneously. This can be seen in the chart below, where a batch size of 32 exhibits 3 times higher latency than a batch size of 8, while delivering relatively similar throughput. So, for results that deliver *both* high throughput *and* low latency a batch size of 8 is the optimum choice.



Optimized Inferencing performance with NVIDA TensorRT™

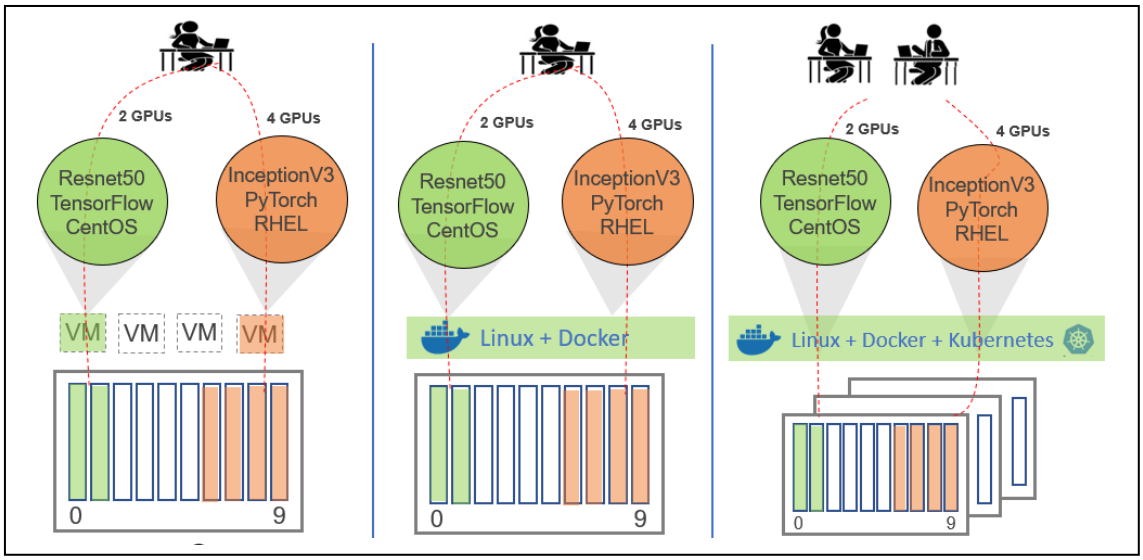
NVIDIA TensorRT is a platform that optimizes inference performance by maximizing utilization of GPUs and seamlessly integrating with deep learning frameworks. It leverages libraries, development tools and technologies in CUDA-X AI for artificial intelligence, autonomous machines, high-performance computing, and graphics. It also provides INT8 and FP16 precision optimizations for inference applications such as video streaming, speech recognition, recommendation and natural language processing. Reduced precision inference significantly reduces application latency, which is a requirement for many real-time services, auto and embedded applications.

Accelerated development with NVIDIA GPU Cloud (NGC)

When the DSS 8440 is configured with NVIDIA 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. The DSS 8440 is certified to work with NGC.

Multi-tenancy for higher productivity and greater flexibility

As mentioned above, the high accelerator capacity of the DSS 8440 makes it an ideal multi-tenancy solution. It can provide Training or Inference resource across multiple workloads, multiple users and departments, or multiple systems. It gives users the flexibility to run different stacks of machine learning software (i.e., models, frameworks, OS) simultaneously on the same server using different numbers of accelerators, as needed. And multi-tenancy also lets data centers simplify the management of machine learning services.



In a multi-tenant environment, you can use NVIDIA NGC and NVIDIA-Docker for ease of use and performance. As mentioned above, NGC includes a container runtime library and utilities to automatically configure containers to leverage NVIDIA GPUs and Python can be used at runtime to indicate the number of GPUs needed. Additionally, a distributed cluster can be managed by an administrator using Kubernetes, and multiple users can make resource requests through that interface.

Balanced design for better performance

Industry leading Training GPU

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.

NVIDIA V100 Tensor Core GPUs in the DSS 8440

Powerful, energy efficient Inference GPU

The NVIDIA T4 Tensor Core GPU delivers responsive, state-of-the-art performance in real-time and allows customers to reduce inference costs by providing high performance in a lower power accelerator. In small batch size jobs multiple T4 GPUs can outperform a single V100, at nearly equivalent power. For example, four T4's can provide more than 3 times the performance of a single V100 – at similar cost – and two T4's can deliver almost twice the performance of a single V100 using roughly half the energy and at half the cost.

First time accelerator customers who choose T4 GPUs will be able to take advantage of it a 40X improvement in speed over CPU-only systems for inference workloads when used in with NVIDIA's TensorRT runtime platform.

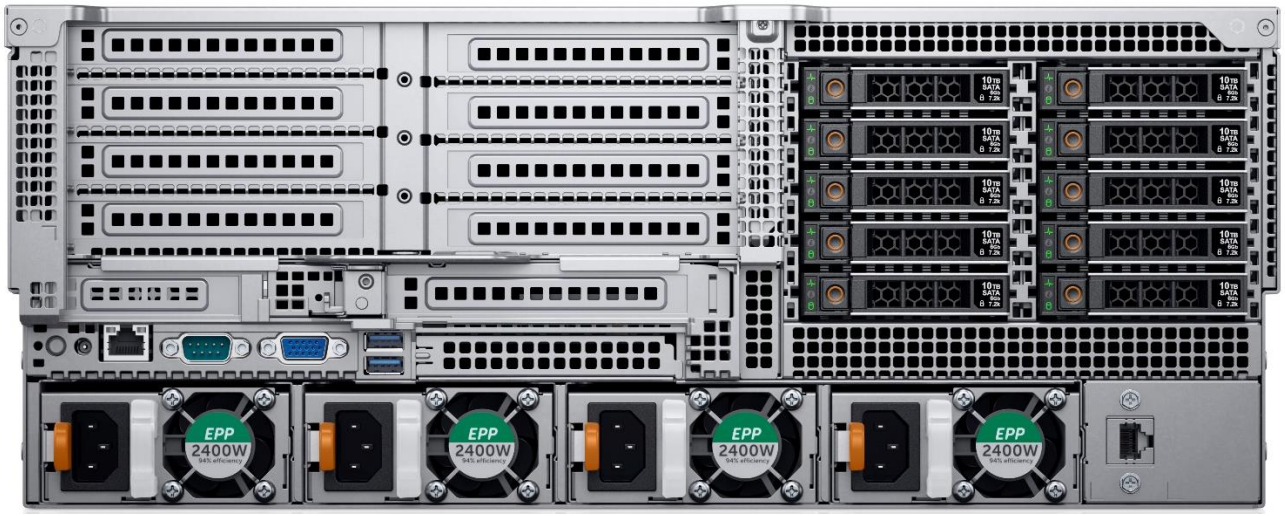
The T4 GPU is also an excellent option for Training environments that don't require top-end performance and want to save on GPU and energy costs. You can save 20% for cost-optimized training workloads and get better performance per dollar. A single T4 GPU gives you 80% of V100 performance at 25% cost)

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

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.

Initial Release The DSS 8440 with 8 Graphcore C2 accelerator cards is available for purchase by a limited number of early adopter customers. Ask your Dell Sales representative for more information.

More power, more efficiency, more flexibility – 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 with V100S and RTX GPUs, get inference results faster with the low latencies available using the DSS 8440 with T4 GPUs, or take advantage of the RTX GPU graphics capabilities for rendering. 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 your Dell Sales representative for more information about the DSS 8440 accelerator-optimized server.