Accelerating AI through the Power of DGX Software

In the last few years, much focus has been placed on the performance boost provided to deep learning through the affinity that exists between modern GPU’s and the massively parallelized algorithms powering today’s deep neural networks. While the hardware-accelerated benefits are unmistakable, increased focus is now being placed on the substantial gains afforded by the custom-engineered software powering NVIDIA® DGX™ Systems.

TUNED FOR DEEP LEARNING PERFORMANCE AND SCALE

NVIDIA DGX Systems deliver an integrated hardware and software solution that’s been optimized for delivering the fastest time-to-solution in a GPU-powered appliance. To make this possible, NVIDIA has made an unmatched investment in engineering the deep learning software stack found on every DGX System, as well as the NVIDIA GPU Cloud.

HOW WE GOT HERE

Years of technology interlock between NVIDIA Engineering and the developers of the leading deep learning frameworks has resulted in steady, incremental progress, advancing the art and science of deep learning performance. Early on, we started with the popular Caffe and Torch, co-developing our stack with each development team. For example, in the first year since the introduction of NVIDIA DGX-1™, this effort has netted a two-fold performance improvement in Caffe training of AlexNet, going from 1,265 images per second, to 2,568 images per second. We’ve since expanded our focus to test framework performance using more modern workloads with deeper neural networks using more layers and delivering lower error rates. At the same time, we’ve included newer frameworks as they’re introduced to the marketplace. The result is a continued set of partnerships that NVIDIA uniquely holds, influencing framework development, and workload benchmarking that pushes the performance bar ever higher.
Targeted Improvements

With each new framework release comes enhancements to DGX software, that have been planned early on in the development cycle, ahead of the actual framework release to the general public. These improvements typically fall into several pillars of targeted development that NVIDIA engineers have perfected over time. Let’s review each of these.

**cuDNN – Convolution Algorithm Choices**

> One of the most important variables in training performance is the selection and optimization of the algorithms used to perform a neural net convolution. cuDNN offers multiple algorithm variants (seven in total)–including matrix-multiply, fast fourier transform (FFT), Winograd and more, each with their own strengths, as well as trade-offs in memory footprint. NVIDIA engineering tunes the heuristic search (native to cuDNN) to select the optimal algorithm for the model running on DGX Systems, and available memory.

**Managing Data Flow and CPU Usage**

> A significant portion of NVIDIA’s engineering effort is focused on the disparity between GPU compute capacity, and the CPU’s ability to keep the GPU utilized. With GPU performance continually improving, CPU performance remains a bottleneck and can lead to “starving” the GPU of data.

In reality, there is currently no I/O subsystem that is able to match the image processing speed that modern GPU’s like NVIDIA Tesla® V100 are actually capable of. NVIDIA Engineering therefore expends significant effort optimizing code to offload CPU tasks to GPU to help reduce the overhead burden on CPU. This includes significant engineering effort to optimize the process of retrieving image data from the data store, and performing JPEG decoding (and transfer to GPU) across multiple CPU threads, using every available core. This asynchronous processing of tasks that might typically wait on CPU cycles is referred to as de-serialization. Fusing multiple kernels together during the weight update step performed after backpropagation, helps reduce the number of data transfers to and from global memory.

All of this is done in order to minimize valuable cycle time lost to I/O and CPU to GPU data transfer, resulting in better overall utilization.

**Memory Management**

> GPU’s have an affinity for larger training batch sizes, and consequently GPU performance and efficiency is significantly impacted by data and parameter storage, as well as I/O buffers and workspace for convolutional layers. Memory management techniques implemented in DGX software enable training of increasingly larger batches, and include optimizations for memory allocation, workspace and memory footprint optimization, and use of batch normalization.
DGX software is also optimized for data parallel execution, which is essential for training at scale and improving the linearity of performance as more GPU’s are tasked. This is achieved on several fronts, including re-writing framework code to improve its ability to support data parallelism automatically. The NVIDIA Collective Communications Library (NCCL) is core to this, and takes into account the physical topology connecting GPU’s and determines the optimal, highest-bandwidth communication patterns possible over PCIe and NVIDIA NVLink™. This increases the speed of the all-reduce operation, central to data-parallel deep learning.

Testing with Today’s Modern Workloads

As mentioned earlier, NVIDIA’s engineering focus has expanded along two axes, namely, support for the latest, most popular frameworks used by deep learning practitioners, as well as testing with modern workloads that are indicative of the environments being deployed in the field. For example, while our initial efforts used AlexNet as our baseline from which to develop specific optimizations to accelerate training speed, neural networks have since grown dramatically deeper in size. ResNets (Residual Neural Networks) are a technique developed by Microsoft in 2015, winning the ImageNet competition with a 3.57% error rate. In addition to using AlexNet as a comparative to show the extent of improvements yielded over time, NVIDIA engineering now includes much deeper neural networks like ResNet50, which provide real-world workload representation against which our framework optimizations can be measured.

Enabling User Flexibility and Choice of Deep Learning Frameworks

DGX software benefits from the investment of engineering expertise and the accumulated knowledge from many years of deep learning research and development. DGX customers benefit from continuous improvement in the performance of their frameworks, updated monthly, via software updates delivered to their DGX System. The cumulative effect of the optimizations discussed here is an approximately 30% performance advantage when compared with similar hardware that does not run on DGX software.

However, beyond the performance uplift of DGX software, is the additional advantage associated with the continual improvement across the portfolio of supported frameworks. Through the technology interlock between NVIDIA and its framework partnerships, customers are now seeing greater uniformity between the expected performance of one framework versus another. This enables teams to use the framework they’re most comfortable with, without significant disparity from the performance yielded on another, resulting in greater flexibility and choice for the deep learning practitioner.