



NVIDIA and Microsoft Accelerate AI Together

GPU-Accelerated Microsoft Cognitive Toolkit Now Available in the Cloud on Microsoft Azure and On-Premises with NVIDIA DGX-1

SC16 -- To help companies join the AI revolution, NVIDIA today announced a collaboration with Microsoft to accelerate AI in the enterprise.

Using the first purpose-built enterprise AI framework optimized to run on [NVIDIA® Tesla® GPUs](#) in Microsoft Azure or on-premises, enterprises now have an AI platform that spans from their data center to Microsoft's cloud.

"Every industry has awoken to the potential of AI," said Jen-Hsun Huang, founder and chief executive officer, NVIDIA. "We've worked with Microsoft to create a lightning-fast AI platform that is available from on-premises with our DGX-1™ supercomputer to the Microsoft Azure cloud. With Microsoft's global reach, every company around the world can now tap the power of AI to transform their business."

"We're working hard to empower every organization with AI, so that they can make smarter products and solve some of the world's most pressing problems," said Harry Shum, executive vice president of the Artificial Intelligence and Research Group at Microsoft. "By working closely with NVIDIA and harnessing the power of GPU-accelerated systems, we've made Cognitive Toolkit and Microsoft Azure the fastest, most versatile AI platform. AI is now within reach of any business."

This jointly optimized platform runs the new Microsoft Cognitive Toolkit (formerly CNTK) on NVIDIA GPUs, including the [NVIDIA DGX-1™ supercomputer](#), which uses [Pascal™ architecture](#) GPUs with [NVLink™ interconnect technology](#), and on Azure N-Series virtual machines, currently in preview. This combination delivers unprecedented performance and ease of use when using data for deep learning.

As a result, companies can harness AI to make better decisions, offer new products and services faster and provide better customer experiences. This is causing every industry to implement AI. In just two years, the number of companies NVIDIA collaborates with on deep learning has jumped 194x to over 19,000. Industries such as healthcare, life sciences, energy, financial services, automotive and manufacturing are benefiting from deeper insight on extreme amounts of data.

The Microsoft Cognitive Toolkit trains and evaluates deep learning algorithms faster than other available toolkits, scaling efficiently in a range of environments -- from a CPU, to GPUs, to multiple machines -- while maintaining accuracy. NVIDIA and Microsoft worked closely to accelerate the Cognitive Toolkit on GPU-based systems and in the Microsoft Azure cloud, offering startups and major enterprises:

- **Greater versatility:** The Cognitive Toolkit lets customers use one framework to train models on premises with the [NVIDIA DGX-1](#) or with NVIDIA GPU-based systems, and then run those models in the cloud on Azure. This scalable, hybrid approach lets enterprises rapidly prototype and deploy intelligent features.
- **Faster performance:** When compared to running on CPUs, the GPU-accelerated Cognitive Toolkit performs deep learning training and inference much faster on NVIDIA GPUs available in Azure N-Series servers and on premises.⁽¹⁾ For example, NVIDIA DGX-1 with Pascal and NVLink interconnect technology is 170x faster than CPU servers for the Cognitive Toolkit.
- **Wider availability:** Azure N-Series virtual machines powered by NVIDIA GPUs are currently in preview to Azure customers, and will be generally available soon. Azure GPUs can be used to accelerate both training and model evaluation. With thousands of customers already part of the preview, businesses of all sizes are already running workloads on Tesla GPUs in Azure N-Series VMs.

NVIDIA and Microsoft plan to continue their collaboration to help optimize the Cognitive Toolkit for NVIDIA GPUs in Azure and as part of a hybrid cloud AI platform, when connected to NVIDIA DGX-1 on premises.

More Resources

- [Deep Learning on Azure with GPUs](#)
- [Microsoft Cognitive Toolkit](#)
- [Azure N-Series](#)
- [NVIDIA Deep Learning](#)
- [NVIDIA DGX-1](#)
- [The Intelligent Industrial Revolution](#) by NVIDIA CEO Jen-Hsun Huang

(1) AlexNet training batch size 128, dual-socket E5-2699v4, 44 cores CNTK 2.0b2 for CPU compared to NVIDIA DGX-1 system. Latest CNTK 2.0b which includes cuDNN 5.1.8, NCCL 1.6.1.

Keep Current on NVIDIA

Subscribe to the [NVIDIA blog](#), follow us on [Facebook](#), [Google+](#), [Twitter](#), [LinkedIn](#) and [Instagram](#), and view NVIDIA videos on [YouTube](#) and images on [Flickr](#).

About NVIDIA

[NVIDIA](#) (NASDAQ: NVDA) is the AI computing company. Its invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI -- the next era of computing -- with the GPU acting as the brain of computers, robots and self-driving cars that can perceive and understand the world. More information at <http://nvidianews.nvidia.com/>.

Certain statements in this press release including, but not limited to the impact and benefits of NVIDIA's and Microsoft's AI acceleration collaboration, Tesla GPUs, DGX-1, the Pascal architecture, NVLink interconnect technology and the Microsoft Cognitive Toolkit; the availability of Azure N-Series virtual machines; and the continuation of NVIDIA's and Microsoft's collaboration are forward-looking statements that are subject to risks and uncertainties that could cause results to be materially different than expectations. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners' products; design, manufacturing or software defects; changes in consumer preferences or demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems; as well as other factors detailed from time to time in the reports NVIDIA files with the Securities and Exchange Commission, or SEC, including its Form 10-Q for the fiscal period ended July 31, 2016. Copies of reports filed with the SEC are posted on the company's website and are available from NVIDIA without charge. These forward-looking statements are not guarantees of future performance and speak only as of the date hereof, and, except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances.

© 2016 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, Tesla, Pascal, NVLink and DGX-1 are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. Features, pricing, availability and specifications are subject to change without notice.

Ken Brown
Corporate Communications
+1-408-486-2626
kebrown@nvidia.com