NVIDIA DGX A100

FOR BUILDING CONVERSATIONAL AI APPLICATIONS

Conversational AI is on the Rise

- Worldwide spending on conversational AI is expected to top $13.8 billion by 2023.[1] – IDC
- Early benefits of conversational AI adoption yield an average of 30% annual growth in business value.[2] – Deloitte Digital

The Challenge for Building State-of-the-Art Conversational AI

The successful adoption of conversational AI hinges on the delivery of natural interactions that feel human. This requires contextual awareness, the capacity to understand sentiment, and the ability to hold simultaneous conversations—all delivered in milliseconds. Also, in order to realize ROI, developers need AI expertise, access to large amounts of industry or product-specific data, and the infrastructure and tools to speed model iteration and accuracy. NVIDIA® DGX™ A100 offers the high-performance AI infrastructure needed to deliver state-of-the-art conversational AI, making it easier for enterprises to implement AI-powered assistants, messaging apps, and chatbots with superhuman levels of language understanding.

Leverage Ready-to-use Tools and Optimized Models

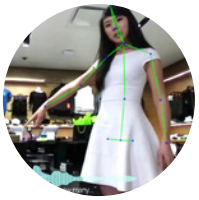
NVIDIA Jarvis, an application framework developed to build end-to-end GPU-optimized conversational AI services, includes hundreds of pre-trained models. Jarvis boasts optimized, end-to-end pipelines for speech, vision, and natural language understanding (NLU) tasks. These models were trained for more than 100,000 hours, across open and proprietary datasets, using **NVIDIA DGX systems**. Developers can fine-tune them with their domain-specific data using a simple API with the **NVIDIA NeMo** toolkit on their DGX system. They can also build and train state-of-the-art models, such as Quartznet, Jasper, BERT, Tacotron2, and WaveGlow from scratch using NeMo. More natural interactions can be achieved on conversational AI applications by building multimodal skills that fuse speech and vision.

Achieve Highest Levels of Accuracy and Human-Like Conversation

NVIDIA DGX A100 features eight NVIDIA A100 Tensor Core GPUs—the most advanced data center accelerator ever made. Third generation Tensor Cores

VIDEOCONFERENCE TRANSLATION, TRANSCRIPTION

200M Meetings per day



RETAIL ASSISTANTS

12M Retail stores

CALL CENTER

500M Calls per day

IN-CAR ASSISTANTS

75M New cars per year

SMART SPEAKERS

150M Sold per year

are optimized to accelerate matrix-multiplication calculations, which are at the heart of AI training and inference. Tensor Float (TF32) precision enables support for both larger magnitudes and precision, delivering 10X the AI performance improvement than previous generations—without any code change. And because natural language processing (NLP) is typically faced with large and sparse input matrices, as the number of useful words is a tiny fraction of the size of the dictionary[3], the new structural sparsity feature within the A100 GPU can accelerate computations by an additional 2X across a variety of common AI networks. Third generation NVIDIA NVLink®, NVIDIA NVSwitch™, and NVIDIA Mellanox InfiniBand delivers ultra-high bandwidth and low latency connections between all the GPUs and enables scaling of multiple DGX A100 systems to train the largest NLP models.

Deliver Interactive Response

The latency threshold for real-time performance and human-like conversations is under 300 milliseconds (ms). Using the NVIDIA Jarvis conversational AI framework, developers can optimize state-of-the-art models for inference and offer real-time services that run in 150 ms (vs. 25 seconds on CPU-only platforms), delivering a 160X performance improvement. Serve several models on NVIDIA Triton Inference Server, running efficiently with NVIDIA TensorRT™ optimizations, and deploy services using a single command through Helm charts on Kubernetes clusters. Use **Multi-Instance GPU (MIG)** innovation to maximize throughput, and enable up to 56 simultaneous inference servers on a single DGX A100—each fully isolated at the hardware level with high bandwidth memory, cache, and compute cores.

Universal Conversational AI Platform: From Prototype to Production

NVIDIA DGX A100 is the universal system for AI infrastructure, ranging from analytics to training to inference. DGX A100 sets a new bar for compute density, packing 5 petaFLOPS of AI performance into a 6U form factor, replacing inflexible legacy compute infrastructure with a single, unified system that can do it all. Train and fine-tune large models, such as Megatron-BERT on your DGX A100, using up to eight NVIDIA A100 Tensor Core GPUs or divide each GPU into 7 separate instances to run inference. This MIG innovation allows users to mix and match multiple training and inference jobs in parallel, on the same system, with dedicated resources for optimal utilization.

Get to Success Faster with Advice From DGXperts

NVIDIA DGX A100 is a complete hardware and software platform, backed by thousands of NVIDIA AI experts, and is built upon the knowledge gained from the world's largest DGX proving ground, NVIDIA DGX SATURNV. Owning a DGX A100 gives you direct access to **NVIDIA DGXperts**, a global team of AI-fluent practitioners who offer prescriptive guidance and design expertise to help fast-track AI transformation. This ensures mission-critical applications get up and running quickly and stay running smoothly, dramatically improving time to insights.

Achieving the Highest Accuracy at Scale

Kensho, the innovation hub for S&P Global, developed a speech recognition solution for finance and business using NVIDIA's conversational AI framework. With models trained on NVIDIA DGX SuperPOD™, an architectural design cluster of DGX A100 systems, they improved accuracy for transcribing earnings calls and financial audio, compared to commercial solutions, by up to 20%.

Get Going In Three Easy Steps

STEP 01
Leverage pre-trained models from **NGC** (Trained for 100,000 hours on DGX systems)

STEP 02
Re-train and fine-tune models on your data on your DGX A100

STEP 03
Deploy services with one line of code through Helm charts on Kubernetes clusters

Learn more about NVIDIA DGX A100: **www.nvidia.com/dgxa100**

Explore conversational AI with NVIDIA Jarvis: **developer.nvidia.com/nvidia-jarvis**

1 David Schubmehl. IDC Worldwide Artificial Intelligence Software Platforms Forecast, 2020-2024. June 2020. Market Forecast - Doc # US45724520
2 Deloitte. Conversational AI: The Next Wave of Customer and Employee Experiences. Q4 2019.
3 Luis Filipe Kopp, José Barbosa da Silva Filho, Claudio Miceli de Farias, and Priscila Machado Vieira Lima. Modeling Sparse Data as Input for Weightless Neural Network. April 2019.

NVIDIA