Meta Works with NVIDIA to Build Massive AI Research Supercomputer

Meta's AI supercomputer — the largest NVIDIA DGX A100 customer system to date — will deliver Meta AI researchers 5 exaflops of AI performance and features cutting-edge NVIDIA systems, InfiniBand fabric and software enabling optimization across thousands of GPUs.

Author: Charlie Boyle

Meta Platforms gave a big thumbs up to NVIDIA, choosing our technologies for what it believes will be its most powerful research system to date.

The AI Research SuperCluster (RSC), announced today, is already training new models to advance AI.

Once fully deployed, Meta's RSC is expected to be the largest customer installation of NVIDIA DGX A100 systems.

"We hope RSC will help us build entirely new AI systems that can, for example, power real-time voice translations to large groups of people, each speaking a different language, so they could seamlessly collaborate on a research project or play an AR game together," the company said in a blog .

When RSC is fully built out, later this year, Meta aims to use it to train AI models with more than a trillion parameters. That could advance fields such as natural-language processing for jobs like identifying harmful content in real time.

In addition to performance at scale, Meta cited extreme reliability, security, privacy and the flexibility to handle "a wide range of AI models" as its key criteria for RSC.

The new AI supercomputer currently uses 760 NVIDIA DGX A100 systems as its compute nodes. They pack a total of 6,080 NVIDIA A100 GPUs linked on an NVIDIA Quantum 200Gb/s InfiniBand network to deliver 1,895 petaflops of TF32 performance.

Despite challenges from COVID-19, RSC took just 18 months to go from an idea on paper to a working AI supercomputer (shown in the video below) thanks in part to the NVIDIA DGX A100 technology at the foundation of Meta RSC.

Penguin Computing is our NVIDIA Partner Network delivery partner for RSC. In addition to the 760 DGX A100 systems and InfiniBand networking, Penguin provided managed services and AI-optimized infrastructure for Meta comprised of 46 petabytes of cache storage with its Altus systems. Pure Storage FlashBlade and FlashArray//C provide the highly performant and scalable all-flash storage capabilities needed to power RSC.

■ ■

It's the second time Meta has picked NVIDIA technologies as the base for its research infrastructure. In 2017, Meta built the first generation of this infrastructure for AI research with 22,000 NVIDIA V100 Tensor Core GPUs that handles 35,000 AI training jobs a day.

Meta's early benchmarks showed RSC can train large NLP models 3x faster and run computer vision jobs 20x faster than the prior system.

In a second phase later this year, RSC will expand to 16,000 GPUs that Meta believes will deliver a whopping 5 exaflops of mixed precision AI performance. And Meta aims to expand RSC's storage system to deliver up to an exabyte of data at 16 terabytes per second.

NVIDIA AI technologies are available to enterprises of any size.

NVIDIA DGX, which includes a full stack of NVIDIA AI software , scales easily from a single system to a DGX SuperPOD running on-premises or at a colocation provider . Customers can also rent DGX systems through NVIDIA DGX Foundry .

Original URL: https://blogs.nvidia.com/blog/2022/01/24/meta-ai-supercomputer-dgx/