Why the New NVIDIA Grace Hopper Superchip Is Ideal for Next-Gen Recommender Systems

Performance of the massive AI models that help users personalize the internet will hit new levels of accuracy with the Grace Hopper Superchip.

Author: Paresh Kharya

Recommender systems, the economic engines of the internet, are getting a new turbocharger: the NVIDIA Grace Hopper Superchip .

Every day, recommenders serve up trillions of search results, ads, products, music and news stories to billions of people. They're among the most important AI models of our time because they're incredibly effective at finding in the internet's pandemonium the pearls users want.

These machine learning pipelines run on data, terabytes of it. The more data recommenders consume, the more accurate their results and the more return on investment they deliver.

To process this data tsunami, companies are already adopting accelerated computing to personalize services for their customers. Grace Hopper will take their advances to the next level.

Pinterest, the image-sharing social media company, was able to move to 100x larger recommender models by adopting NVIDIA GPUs. That increased engagement by 16% for its more than 400 million users.

"Normally, we would be happy with a 2% increase, and 16% is just a beginning," a software engineer at the company said in a recent blog . "We see additional gains — it opens a lot of doors for opportunities."

The next generation of the NVIDIA AI platform promises even greater gains for companies processing massive datasets with super-sized recommender models.

Because data is the fuel of AI, Grace Hopper is designed to pump more data through recommender systems than any other processor on the planet.

Grace Hopper achieves this because it's a superchip — two chips in one unit, sharing a superfast chip-to-chip interconnect. It's an Arm-based NVIDIA Grace CPU and a Hopper GPU that communicate over NVIDIA NVLink-C2C .

What's more, NVLink also connects many superchips into a super system, a computing cluster built to run terabyte-class recommender systems.

NVLink carries data at a whopping 900 gigabytes per second — 7x the bandwidth of PCIe Gen 5, the interconnect most leading edge upcoming systems will use.

That means Grace Hopper feeds recommenders 7x more of the embeddings — data tables packed with context — that they need to personalize results for users.

The Grace CPU uses LPDDR5X, a type of memory that strikes the optimal balance of bandwidth, energy efficiency, capacity and cost for recommender systems and other demanding workloads. It provides 50% more bandwidth while using an eighth of the power per gigabyte of traditional DDR5 memory subsystems.

Any Hopper GPU in a cluster can access Grace's memory over NVLink. It's a feature of Grace Hopper that provides the largest pools of GPU memory ever.

In addition, NVLink-C2C requires just 1.3 picojoules per bit transferred, giving it more than 5x the energy efficiency of PCIe Gen 5.

The overall result is recommenders get a further up to 4x more performance and greater efficiency using Grace Hopper than using Hopper with traditional CPUs (see chart below).

The Grace Hopper Superchip runs the full stack of NVIDIA AI software used in some of the world's largest recommender systems today.

NVIDIA Merlin is the rocket fuel of recommenders, a collection of models, methods and libraries for building AI systems that can provide better predictions and increase clicks.

NVIDIA Merlin HugeCTR , a recommender framework, helps users process massive datasets fast across distributed GPU clusters with help from the NVIDIA Collective Communications Library .

Learn more about Grace Hopper and NVLink in this technical blog . Watch this GTC session to learn more about building recommender systems.

You can also hear NVIDIA CEO and co-founder Jensen Huang provide perspective on recommenders here or watch the full GTC keynote below.

Original URL: https://blogs.nvidia.com/blog/2022/09/20/grace-hopper-recommender-systems/