Nvidia reveals Blackwell B200 GPU, the ‘world’s most powerful chip’ for AI

Nvidia’s must-have H100 AI chip made it a multitrillion-dollar company, one that may be worth more than Alphabet and Amazon, and competitors have been fighting to catch up. But perhaps Nvidia is about to extend its lead — with the new Blackwell B200 GPU and GB200 “superchip.”

a:hover]:text-gray-63 [&>a:hover]:shadow-underline-black dark:[&>a:hover]:text-gray-bd dark:[&>a:hover]:shadow-underline-gray [&>a]:shadow-underline-gray-63 dark:[&>a]:text-gray-bd dark:[&>a]:shadow-underline-gray”>Image: Nvidia
Nvidia says the new B200 GPU offers up to 20 *petaflops* of FP4 horsepower from its 208 billion transistors. Also, it says, a GB200 that combines two of those GPUs with a single Grace CPU can offer 30 times the performance for LLM inference workloads while also potentially being substantially more efficient. It “reduces cost and energy consumption by up to 25x” over an H100, says Nvidia.

Training a 1.8 trillion parameter model would have previously taken 8,000 Hopper GPUs and 15 megawatts of power, Nvidia claims. Today, Nvidia’s CEO says 2,000 Blackwell GPUs can do it while consuming just four megawatts.

On a GPT-3 LLM benchmark with 175 billion parameters, Nvidia says the GB200 has a somewhat more modest seven times the performance of an H100, and Nvidia says it offers four times the training speed.

a:hover]:text-gray-63 [&>a:hover]:shadow-underline-black

dark:[&>a:hover]:text-gray-bd dark:[&>a:hover]:shadow-underline-gray [&>a]:shadow-underline-gray-63 dark:[&>a]:text-gray-bd dark:[&>a]:shadow-underline-gray”>Image: Nvidia
Nvidia told journalists one of the key improvements is a second-gen transformer engine that doubles the compute, bandwidth, and model size by using four bits for each neuron instead of eight (thus, the 20 petaflops of FP4 I mentioned earlier). A second key difference only comes when you link up huge numbers of these GPUs: a next-gen NVLink switch that lets 576 GPUs talk to each other, with 1.8 terabytes per second of bidirectional bandwidth.

That required Nvidia to build an entire new network switch chip, one with 50 billion transistors and some of its own onboard compute: 3.6 teraflops of FP8, says Nvidia.

a:hover]:text-gray-63 [&>a:hover]:shadow-underline-black dark:[&>a:hover]:text-gray-bd dark:[&>a:hover]:shadow-underline-gray [&>a]:shadow-underline-gray-63 dark:[&>a]:text-gray-bd dark:[&>a]:shadow-underline-gray”>Image: Nvidia
Previously, Nvidia says, a cluster of just 16 GPUs would spend 60 percent of their time communicating with one another and only 40 percent actually computing.

Nvidia is counting on companies to buy large quantities of these GPUs, of course, and is packaging them in larger designs, like the GB200 NVL72, which plugs 36 CPUs and 72 GPUs into a single liquid-cooled rack for a total of 720 petaflops of AI training performance or 1,440 petaflops (aka 1.4 *exaflops*) of inference. It has nearly two miles of cables inside, with 5,000 individual cables.

a:hover]:text-gray-63 [&>a:hover]:shadow-underline-black dark:[&>a:hover]:text-gray-bd dark:[&>a:hover]:shadow-underline-gray [&>a]:shadow-underline-gray-63 dark:[&>a]:text-gray-bd dark:[&>a]:shadow-underline-gray”>Image: Nvidia
Each tray in the rack contains either two GB200 chips or two NVLink switches, with 18 of the former and nine of the latter

per rack. In total, Nvidia says one of these racks can support a 27-trillion parameter model. GPT-4 is rumored to be around a 1.7-trillion parameter model.

The company says Amazon, Google, Microsoft, and Oracle are all already planning to offer the NVL72 racks in their cloud service offerings, though it’s not clear how many they’re buying.

And of course, Nvidia is happy to offer companies the rest of the solution, too. Here’s the DGX Superpod for DGX GB200, which combines eight systems in one for a total of 288 CPUs, 576 GPUs, 240TB of memory, and 11.5 exaflops of FP4 computing.