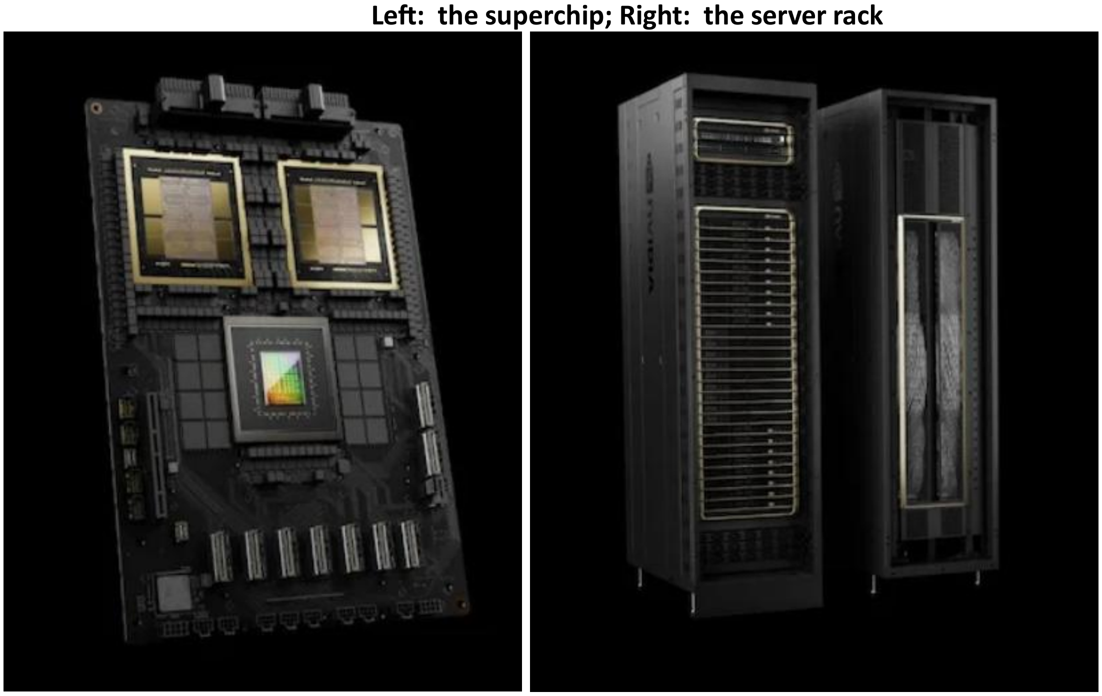
Inside NVIDIA: Last Quarter's Progress

Jun. 12, 2025

It's impossible to review Nvidia's latest earnings report without standing in awe at the strength, inventiveness and momentum of this company. Nvidia is the engine behind the "AI Boom," which might very well rank with the invention of electricity as an engine of global economic growth.

NVIDIA's total revenue growth of 69% and data center revenue growth of 73% since last year supports the idea that demand for AI is not slowing down. This past quarter, NVIDIA ramped up its Blackwell GPU, the newest release of its graphics processing unit (GPU) built specifically for data center scale generative AI. NVIDIA combines its Grace CPU with 2 Blackwell GPUs to create one superchip, then further combines 36 of these superchips in a stacked design in a liquid-cooled enclosure, creating its newest server rack, the GB200 NVL72. This server rack acts as a single GPU, even though it contains 72 Blackwell GPUs, and delivers a 30x performance increase and 25x decrease in total cost of ownership over NVIDIA's previous server, the H100. Companies like Microsoft have been buying and installing 1,000 racks per week and are expected to ramp that number up over time.

Left: the superchip; Right: the server rack



This kind of usage goes hand in hand with what we heard from Microsoft last quarter when talking about the increase in inference demand. Microsoft's Azure OpenAI processed more than 100 trillion tokens last quarter, which was 5x the amount from a year ago. ("Tokens" are words or parts of words which AI models "read" in the process of coming up with their responses to user requests.)

As AI workloads turn more towards inference, more AI factories are needed, which will drive growth for NVIDIA. There are roughly 100 NVIDIA-powered AI factories around the world today. This is double the number from a year ago, and each new factory has double the number of GPUs as older factories.

The one pitfall from last quarter came from the government's April 9th issued export controls on H20, the GPU chip that NVIDIA made exclusively for China. NVIDIA made H20 to comply with earlier restrictions, and H20 is essentially a weaker version of the last generation Hopper GPU. The new export controls went into effect immediately, meaning NVIDIA couldn't fulfill existing orders and couldn't get rid of existing inventory, leading to weaker profitability for the quarter and lower expected revenue for the next quarter. CEO Jensen Huang and his team don't believe these new export controls are in the best interest of the US, but if they remain in effect, NVIDIA is looking at limited ways of working around the export controls. Even in the week since the earnings call, where management had no news of a solution yet, there are rumors that a new chip is being made to comply with the new restrictions.

This past quarter's earnings showed that the AI boom is only in early innings, and strong growth should be expected to continue. The pop in the stock price was nice to see after Outlook's recent purchase of NVIDIA. We have no doubt there will be many more pops, both up and down in the future, but the general direction should still be emphatically "up".