Nvidia sells tiny new computer that puts big AI on your desktop

For the OS, the Spark is an ARM-based system that runs Nvidia's DGX OS, an Ubuntu Linux-based operating system built specifically for GPU processing. It comes with Nvidia's AI software stack preinstalled, including CUDA libraries and the company's NIM microservices.

Prices for the DGX Spark start at US $3,999. That may seem like a lot, but given the cost of high-end GPUs with ample video RAM like the RTX Pro 6000 (about $9,000) or AI server GPUs (like $25,000 for a base-level H100), the DGX Spark may represent a far less expensive option overall, though it's not nearly as powerful.

In fact, according to The Register, the GPU computing performance of the GB10 chip is roughly equivalent to an RTX 5070. However, the 5070 is limited to 12GB of video memory, which limits the size of AI models that can be run on such a system. With 128GB of unified memory, the DGX Spark can run far larger models, albeit at a slower speed than, say, an RTX 5090 (which typically ships with 24 GB of RAM). For example, to run the 120 billion-parameter larger version of OpenAI's recent gpt-oss language model, you'd need about 80GB of memory, which is far more than you can get in a consumer GPU.

A callback to 2016

Nvidia founder and CEO Jensen Huang marked the occasion of the DGX Spark launch by personally delivering one of the first

units to Elon Musk at SpaceX’s Starbase facility in Texas, echoing a similar delivery Huang made to Musk at OpenAI in 2016.

“In 2016, we built DGX-1 to give AI researchers their own supercomputer. I hand-delivered the first system to Elon at a small startup called OpenAI, and from it came ChatGPT,” Huang said in a statement. “DGX-1 launched the era of AI supercomputers and unlocked the scaling laws that drive modern AI. With DGX Spark, we return to that mission.”

https://arstechnica.com/ai/2025/10/nvidia-sells-tiny-new-computer-that-puts-big-ai-on-your-desktop/