



NVIDIA DGX Spark

Also known as: DGX Spark Founders Edition · GB10 desktop · Project DIGITS

DGX Spark is a desk-side box built around NVIDIA's GB10 Grace Blackwell superchip, 128 GB of unified memory, and the full CUDA stack, aimed at a single developer prototyping and fine-tuning models locally before moving to real GPU cluster capacity. It is not a cluster node: its 273 GB/s memory bandwidth is well behind a datacenter GPU, and while two units can be linked over ConnectX-7 to handle models up to 405B parameters, that is a two-box desk setup, not a scalable rack architecture. Buy it to develop against CUDA and iterate on model code without renting cloud GPU time; don't buy it expecting DGX B200-class throughput.

1 PFLOP FP4 (sparse) AI performance · GB10 Grace Blackwell superchip	20-core Arm CPU · 10x Cortex-X925 + 10x Cortex-A725	128 GB LPDDR5x unified Memory · 273 GB/s bandwidth, 256-bit interface
1 TB or 4 TB NVMe Storage · M.2 2242, self-encrypting	ConnectX-7, up to 200 Gbps Networking · dual QSFP; links two units together	240 W external PSU Power · GB10 SoC rated 140 W TDP

COMPUTE01/05	
Superchip	GB10 Grace Blackwell · Grace CPU + Blackwell GPU on one SoC
CPU cores	20-core Arm · 10x Cortex-X925 (4 GHz) + 10x Cortex-A725 (2.8 GHz)
GPU	Blackwell, 5th-gen Tensor Cores, 4th-gen RT Cores · 6,144 CUDA cores
FP4 performance	1 PFLOP (sparse)
FP32 performance	31 TFLOPS
MEMORY02/05	
Capacity	128 GB LPDDR5x · coherent unified CPU+GPU memory
Bandwidth	273 GB/s · 256-bit interface, 16 channels
Inference ceiling	up to 200B parameters · single unit
Fine-tuning ceiling	up to 70B parameters · single unit, full-parameter
Clustered inference ceiling	up to 405B parameters · two units linked over ConnectX-7

CONNECTIVITY		03/05
Scale-out NIC	ConnectX-7, 200 Gbps · 2x QSFP; PCIe lane sharing caps aggregate at 200 Gbps despite two ports	
Ethernet	1x RJ-45, 10 GbE · standard LAN	
USB	4x USB-C · one used for power delivery	
Display	1x HDMI 2.1a · plus up to 3x DisplayPort via USB-C alt mode	
Wireless	Wi-Fi 7, Bluetooth 5.4	
POWER & PHYSICAL		04/05
Power supply	240 W external · included	
GB10 TDP	140 W · of the 240 W budget	
Dimensions	150 x 150 x 50.5 mm · L x W x H	
Weight	1.2 kg	
Operating temperature	5C to 30C · non-condensing, 10-90% humidity	
SOFTWARE		05/05
OS	NVIDIA DGX OS · Ubuntu-based	
Stack	NVIDIA AI Enterprise, NIM · preloaded	
Ecosystem	Full CUDA access · distinguishes it from Apple Silicon or other ARM desktop AI boxes	
PROCUREMENT		
Typical lead time Ships from stock through Micro Center and Amazon. NVIDIA's own online store has run intermittently sold out since the October 15, 2025 launch, reported; retail channel stock has been more reliable than NVIDIA's direct store. Pricing \$3,999 to \$4,699 for the 4 TB Founders Edition, reported. NVIDIA raised its own-store MSRP from \$3,999 to \$4,699 in early 2026, citing memory supply constraints; Micro Center has continued selling at the original \$3,999.99 price point. OEM-branded GB10 desktop boxes (Dell, HP, Lenovo, ASUS) are a separate SKU and price point. - Price has moved once already since launch; confirm current pricing at order time rather than quoting from memory. - The 1 TB storage configuration is a lower-cost variant of the same Founders Edition SKU where available. - Two units plus a ConnectX-7 link cable roughly doubles the cost for the 405B-parameter clustered ceiling; evaluate against a single cloud GPU rental hour before committing. - No enterprise support contract is bundled at retail; NVIDIA AI Enterprise licensing for production use is a separate purchase.		

FIELD NOTES

- 273 GB/s memory bandwidth is the real ceiling, not the 128 GB capacity. Prefill-heavy and fine-tuning workloads are compute-bound and look great; decode-heavy inference is memory-bound and noticeably slower than the raw PFLOP figure suggests.
- It is not a cluster building block. The dual QSFP ConnectX-7 ports look like 400 Gbps of fabric, but PCIe lane sharing caps actual throughput at 200 Gbps aggregate, and NVIDIA's own clustering story tops out at two linked units.
- The 2242 M.2 storage form factor has been reported to lock up under sustained heavy I/O in third-party testing; treat local SSD as working storage, not the system of record for large datasets.
- For raw token throughput on models that fit in 32 GB, a workstation RTX 5090 is faster and cheaper. DGX Spark's reason to exist is the 128 GB unified memory footprint for models a 5090 cannot hold, not raw speed.
- Buy this for CUDA-native local development and fine-tuning iteration, not as a production inference box. Budget cloud GPU time for anything latency-sensitive or high-concurrency.

QUESTIONS WE GET ON THIS PART

How much does DGX Spark cost?

The 4 TB Founders Edition has listed between \$3,999 and \$4,699, reported. NVIDIA raised its own-store price to \$4,699 in early 2026 citing memory supply constraints, while retailers like Micro Center have continued to sell at \$3,999.99.

What is DGX Spark used for?

Local AI development: prototyping, full-CUDA-stack coding, and fine-tuning models up to roughly 70B parameters or running inference on models up to roughly 200B parameters, thanks to its 128 GB unified memory. It targets a single developer's desk, not production inference serving.

Can I cluster two DGX Spark units together?

Yes. Two units connected over their ConnectX-7 networking can handle inference on models up to roughly 405B parameters. This is a two-box desk setup for larger model capacity, not a scalable multi-node cluster architecture.

Is DGX Spark good for inference or training?

It handles both, but memory bandwidth (273 GB/s) is the bottleneck for decode-heavy inference. It is strongest for full-parameter fine-tuning and prefill-heavy workloads where compute, not memory bandwidth, is the limiting factor.

REFERENCES

- [1] <https://www.nvidia.com/en-us/products/workstations/dgx-spark/>
- [2] <https://docs.nvidia.com/dgx/dgx-spark/hardware.html>
- [3] <https://www.tomshardware.com/desktops/mini-pcs/nvidia-dgx-spark-gets-18-percent-price-increase-as-memory-shortages-bite-founder-s-edition-now-usd4-699-up-from-usd3-999>
- [4] <https://www.storagereview.com/review/nvidia-dgx-spark-review-the-ai-appliance-bringing-datacenter-capabilities-to-desktops>
- [5] <https://robert-mcdermott.medium.com/the-nvidia-dgx-spark-0e2ca7833c2c>
- [6] <https://www.networkworld.com/article/4072947/nvidias-dgx-spark-desktop-supercomputer-is-on-sale-now-but-hard-to-find-2.html>