# GB10 Boxes Compared: DGX Spark vs Dell vs ASUS vs MSI

February 6, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** All four GB10 machines use the same NVIDIA Grace Blackwell chip with 128GB unified memory. Performance is identical — Qwen 34B runs at 61 tok/s on every single one. The differences are chassis quality, storage speed, and thermals. The DGX Spark ($3,999, Gen 5 NVMe) loads models 25% faster. The ASUS GX10 ($3,099, 1TB) is cheapest but heaviest and the only one that triggers thermal slowdown events. The MSI EdgeXpert ($2,999, 1TB) is the cheapest entry point but feels plastic. The Dell Pro Max ($3,699, 2TB) is basically the Spark with better labeling. For most local AI builders, a used RTX 3090 at $900 with 24GB VRAM is still a better deal — it runs quantized 70B models faster than any GB10. The GB10's value is loading unquantized 70B-200B models that nothing else can fit. But memory bandwidth is 273 GB/s — expect 2-3 tok/s on 70B models. A Mac Studio M3 Ultra delivers 3x more bandwidth for similar money.

📚 **More on this topic:** GPU Buying Guide · VRAM Requirements · Multi-GPU Local AI · What Can You Run on 24GB VRAM · Planning Tool

Four companies are selling boxes with the same chip inside. NVIDIA's DGX Spark, Dell's Pro Max GB10, ASUS's Ascent GX10, and MSI's EdgeXpert all use the NVIDIA Grace Blackwell GB10 superchip — 20 ARM cores, a Blackwell GPU with 6,144 CUDA cores, and 128GB of unified LPDDR5X memory. Same silicon, same 1 PFLOP sparse FP4 compute, same DGXOS.

The pitch is compelling: a petaflop AI computer on your desk for $3,000-4,000. Load 70B models unquantized. Run 200B models in FP4. No multi-GPU complexity, no 850W power supply, no rack mount.

But "same chip" doesn't mean same machine. The chassis matters for thermals. The NVMe generation matters for model loading. The build quality matters if you're spending $4,000. And most importantly — should you spend $3,000-4,000 on any of these when a used RTX 3090 costs $900?

Here's what 45-minute heat soak tests, real inference benchmarks, and side-by-side comparisons actually show.

## The Comparison Table

| | DGX Spark | Dell Pro Max | ASUS GX10 | MSI EdgeXpert |
|---|---|---|---|---|
| **Price (1TB)** | N/A | N/A | $3,099 | $2,999 |
| **Price (2TB)** | N/A | $3,699 | ~$3,200 | N/A |
| **Price (4TB)** | $3,999 | $3,999 | $4,150 | $3,999 |
| **NVMe gen** | Gen 5 | Gen 4 | Gen 4 | Gen 4 |
| **SSD size** | 2242 | 2242 | 2242 | 2242 |
| **SSD upgradeable** | Yes | Yes | Yes (harder) | Yes |
| **Weight** | 1,255g | 1,256g | 1,474g | 1,257g |
| **Back panel** | Magnetic | Magnetic (6 magnets) | Screws | Screws |
| **Port labeling** | None | Minimal | Good | Best |
| **Front power button** | No | No | Yes | No |
| **Chassis** | Metal | Metal | Metal + plastic top | Mostly plastic |
| **Noise** | Quietest | Very quiet | Quiet | Loudest |
| **Thermal events** | None | None | 2 slowdowns | None |

All four machines run DGXOS (Ubuntu-based with pre-installed AI tools), have ConnectX-7 200GbE SmartNICs, and support linking two units for up to 256GB combined memory.

## The GB10 Chip — What You're Actually Getting

Every GB10 box contains the same SoC, co-designed by NVIDIA and MediaTek on TSMC's 3nm process:

| Spec | Value |
|---|---|
| **CPU** | 20 ARM v9.2 cores (10 performance + 10 efficiency) |
| **GPU** | Blackwell, 48 SMs, 6,144 CUDA cores |
| **Tensor Cores** | 5th-gen, 192 total |
| **Memory** | 128GB LPDDR5X unified (shared CPU+GPU) |

| Spec | Value |
|---|---|
| **Memory bandwidth** | 273 GB/s actual |
| **Compute** | 1 PFLOP sparse FP4, ~100 TFLOPS FP16 |
| **TDP** | 140W rated, ~100W observed under load |
| **Networking** | ConnectX-7 (200GbE), Wi-Fi |

The 273 GB/s memory bandwidth is the number that matters most for LLM inference. Token generation is memory-bandwidth-bound — the GPU needs to read every model weight from memory for each token. At 273 GB/s, you're limited to roughly 2-3 tok/s on a 70B model at FP8 and about 5 tok/s at FP4 with TensorRT-LLM optimization.

For comparison: an RTX 5090 has 1,792 GB/s bandwidth (6.5x more), and a Mac Studio M3 Ultra has 819 GB/s (3x more). Both generate tokens significantly faster on models that fit in their memory. The GB10's advantage is pure capacity — 128GB lets you load models that 32GB or even 48GB GPUs can't touch.

## Performance — They're All The Same

This is the headline: when running the same model on all four machines, token generation is identical. The chip is the chip.

### Qwen 34B (Quantized)

| Machine | PP 4096 (tok/s) | TG 8192 (tok/s) |
|---|---|---|
| DGX Spark | ~1,976 | 61 |
| Dell Pro Max | ~1,976 | 61 |
| ASUS GX10 | ~1,976 | 61 |
| MSI EdgeXpert | ~1,976 | 61 |

Prompt processing and token generation are functionally identical across all four machines. This ran via llama.cpp at ~96% GPU utilization for 45 minutes.

### Neotron 30B (Unquantized)

| Machine | PP (tok/s) | Avg GPU Power (W) |
|---|---|---|
| DGX Spark | 1,070 | 66.0 |
| Dell Pro Max | 1,068 | 62.7 |
| ASUS GX10 | 1,068 | 60.0 |
| MSI EdgeXpert | 1,068 | 60.0 |

Same story. The Spark draws slightly more GPU power (~66W average vs ~60W on the others), but performance is indistinguishable. Clock speeds stayed consistent across all machines with no throttling during normal inference workloads.

### What About Bigger Models?

The GB10's marketing claim is "up to 200B parameters." Here's the reality:

| Model | Format | Fits in 128GB? | Decode Speed |
|---|---|---|---|
| 7B | FP16 | Yes (14GB) | Fast |
| 34B | FP16 | Yes (68GB) | ~61 tok/s |
| 70B | FP16 | No (needs 140GB) | N/A |
| 70B | FP8 | Yes (70GB) | ~2.7 tok/s |
| 70B | FP4 (TRT-LLM) | Yes (35GB) | ~5.2 tok/s |
| 120B | FP4 | Yes (~60GB) | Slow |
| 200B | FP4 | Yes (~100GB) | Very slow |

A 70B model does not fit unquantized at FP16 — it needs ~140GB. It fits at FP8, but generates tokens at 2.7 tok/s. That's usable for batch processing but painful for interactive chat. Time-to-first-token on a 90B model is around 2 minutes.

## Thermals and Power — Where the Chassis Matters

All four machines were heat-soaked for 45+ minutes running continuous inference. Under normal LLM workloads, they all behave similarly:

- GPU temperature: ~80°C after heat soak
- Surface temperature: ~50°C
- Total system power: 140-160W
- GPU power alone: ~60-66W

The interesting differences appear under stress testing (GPU burn, maximizing compute beyond typical inference):

### The Software Power Cap

Every GB10 box hits a software-imposed power cap at ~100W GPU draw. This is what John Carmack flagged when he noted the DGX Spark “appears to be maxing out at only 100 watts power draw, less than half of the rated 240 watts.”

Carmack was right about the 100W cap. But calling it “thermal throttling” (as many headlines did) is wrong. The CPU clock speeds stay consistent across all machines when the power cap engages. It’s a software limit, not a thermal event. NVIDIA could raise this cap via a firmware update — and they’ve already shipped performance-improving software updates at CES 2026 that delivered up to 2.6x improvements for some workloads.

The 240W number is the power adapter rating, not the chip’s expected draw. The GB10 chip is rated at 140W TDP.

### The ASUS Exception

Under sustained stress testing, the ASUS GX10 is the only machine that triggers actual thermal slowdown events. Two instances were recorded where the GPU power dropped from 96W to 76W and the OS flagged a software thermal slowdown signal. This happened at GPU temperatures around 95°C — while the Dell reached 99°C without triggering the same event.

The practical impact? Negligible. Looking at the performance charts before and after the thermal events, the output is visually identical. But it’s worth knowing: if you’re running sustained heavy workloads 24/7 (training, continuous agent orchestration), the ASUS is the one machine that shows signs of thermal limits.

The ASUS weighs 1,474g — 220g more than the others. Extra weight usually means extra cooling hardware, but in this case it doesn't translate to better thermals.

### The Acer Note

Storage Review tested a fifth GB10 machine — the Acer Veriton GN100 ($3,999) — and found it ran coolest of all, peaking at just 76°C during demanding prefill workloads where other systems climbed into the mid-to-upper 80s. We haven't tested the Acer ourselves, but if thermals are your top priority, it's worth investigating.

## Storage — The One Real Difference

This is where the DGX Spark actually justifies its premium. It's the only machine with a Gen 5 NVMe drive.

| Machine | NVMe Gen | Sequential Read | Neotron 30B Cold Load |
|---|---|---|---|
| DGX Spark | Gen 5 | ~13,000 MB/s | 8.49s |
| Dell Pro Max | Gen 4 | ~7,000 MB/s | ~11.5s |
| ASUS GX10 | Gen 4 | ~7,000 MB/s | ~11.5s |
| MSI EdgeXpert | Gen 4 | ~7,000 MB/s | ~11.5s |

The Spark loads models 25% faster from cold start. For a 30B model, that's a 3-second difference. For larger models, it scales up — a 70B FP8 model (~70GB) would save 7-8 seconds on the Spark.

This matters if you're swapping models frequently — running agents that pull different models for different tasks, or serving multiple users who need different models. If you load one model and run it all day, you'll never notice.

All four machines use 2242 M.2 SSDs, and all are upgradeable. You can buy the cheapest ASUS GX10 (1TB, $3,099) and drop in a 4TB Gen 4 drive yourself to save $500-1,000 versus buying a 4TB configuration. Getting into the ASUS is slightly harder (requires more disassembly) compared to the Spark and Dell, which use magnetic back panels.

## Physical Design

### Build Quality

The Spark and Dell are nearly twins — both are metal chassis, both use magnetic back panels for SSD access, and they weigh within 1 gram of each other (1,255g vs 1,256g). Dell kept NVIDIA's reference design mostly intact and added a cleaner front grille.

The ASUS is all metal except for a plastic top panel. It's the heaviest at 1,474g and has a front power button — the only one to offer this. If you're rack-mounting these, reaching around the back to hit a power button gets old fast.

The MSI feels like plastic all around. It's the lightest build quality of the four but has the best port labeling — every port is marked with its speed and function. The Spark has no labels at all.

### Rack Mounting

None of these fit in a single rack unit. They're all too tall by a few millimeters. If you're planning a rack deployment, you'll need custom shelving or 2U spacing.

### Noise

All four are dramatically quieter than any desktop GPU. The Spark is the quietest, the MSI is the loudest, but the difference is marginal — none of them are louder than a laptop under moderate load.

## The Budget Reality Check

Here's the honest take from a budget-hardware perspective.

### GB10 vs Used RTX 3090

| | Used RTX 3090 | GB10 (any) |
|---|---|---|
| **Price** | ~$900 | $2,999-$3,999 |
| **VRAM/Memory** | 24GB GDDR6X | 128GB LPDDR5X |
| **Memory bandwidth** | 936 GB/s | 273 GB/s |
| **Llama 70B Q4** | ~18 tok/s | ~5 tok/s (FP4) |

| | Used RTX 3090 | GB10 (any) |
|---|---|---|
| **Llama 7B** | ~80+ tok/s | ~61 tok/s |
| **Power draw** | ~350W | ~100W |
| **Needs host PC** | Yes | No (standalone) |

The RTX 3090 is faster on every model that fits in 24GB, including quantized 70B. It has 3.4x more memory bandwidth. It costs a quarter of the price.

The GB10's only advantage: it loads models that 24GB can't hold. Unquantized 70B at FP8. Unquantized 34B at FP16. Models in the 100-200B range at FP4. If you need that, nothing in this price range competes. But if you're running 7B-34B quantized models — which is what most local AI hobbyists do — the 3090 wins on speed and cost.

### GB10 vs Mac Studio

| | Mac Studio M3 Ultra (128GB) | GB10 (any) |
|---|---|---|
| **Price** | ~$5,000-6,000 | $2,999-$3,999 |
| **Memory** | 128GB unified | 128GB unified |
| **Memory bandwidth** | 819 GB/s | 273 GB/s |
| **Llama 70B (FP8)** | ~8 tok/s | ~2.7 tok/s |
| **Ecosystem** | macOS, MLX, Ollama | DGXOS (Ubuntu), CUDA |

The Mac Studio M3 Ultra has 3x the memory bandwidth and delivers roughly 3x faster token generation on the same models. It costs more, but you get a fully functional desktop computer — not just an inference box.

The GB10 wins on CUDA compatibility and price. If your workflow depends on CUDA-specific tools (TensorRT, vLLM, PyTorch CUDA), the GB10 runs them natively. The Mac requires Metal or MLX.

## The GB10 Paradox

The GB10 has a fundamental tension: it loads models that nothing else in its price range can fit, but runs them slowly. A 70B model at FP8 generates 2.7 tokens per second. That's technically functional but hardly interactive. Time-to-first-token on a 90B+ model can hit 2 minutes.

The machines that run models fast (RTX 5090 at 1,792 GB/s, RTX 3090 at 936 GB/s) can't load the models the GB10 can. The machines that match the GB10's capacity (Mac Studio M3 Ultra) run them 3x faster.

The GB10's sweet spot is narrow: researchers and developers who need unquantized 70B-200B models in a CUDA environment, running batch inference or agent orchestration where 2-5 tok/s is acceptable. If that's you, any of these four boxes will do the job identically.

## Which One to Buy

If you've decided a GB10 is right for your workload, here's how to choose between the four:

| Your Priority | Best Choice |
|---|---|
| Fastest model loading | **DGX Spark** ($3,999) — Gen 5 NVMe |
| Cheapest entry | **MSI EdgeXpert** ($2,999, 1TB) or **ASUS GX10** ($3,099, 1TB) |
| Best build quality | **DGX Spark** or **Dell Pro Max** — metal, magnetic panels |
| Best thermals | **DGX Spark** or **Dell Pro Max** — no thermal events |
| Front power button | **ASUS GX10** — only one that has it |
| Best port labeling | **MSI EdgeXpert** — every port labeled |
| Upgrade the SSD yourself | **DGX Spark** or **Dell Pro Max** — magnetic back, easy access |
| Coolest running | **Acer Veriton GN100** ($3,999) — peaked at 76°C in third-party testing |

If you want the reference design with the fastest storage, get the Spark. If you want to save $900-1,000 and performance is identical anyway, get the MSI or ASUS 1TB and upgrade the SSD later. If build quality matters and you want something between the Spark's price and the budget options, the Dell at $3,699 for 2TB is a reasonable middle ground.

Don't pay a premium for performance differences — there aren't any. You're paying for storage speed, chassis materials, and convenience features.

## Related Guides

- GPU Buying Guide for Local AI

- VRAM Requirements for Local LLMs
- Running LLMs on Mac M-Series
- Multi-GPU Local AI
- What Can You Run on 24GB VRAM
- Razer AIKit Guide
- Local AI Planning Tool — VRAM Calculator

Get notified when we publish new guides.

Subscribe — free, no spam

---

Source: https://insiderllm.com/guides/gb10-boxes-compared/

Free guides for running AI locally