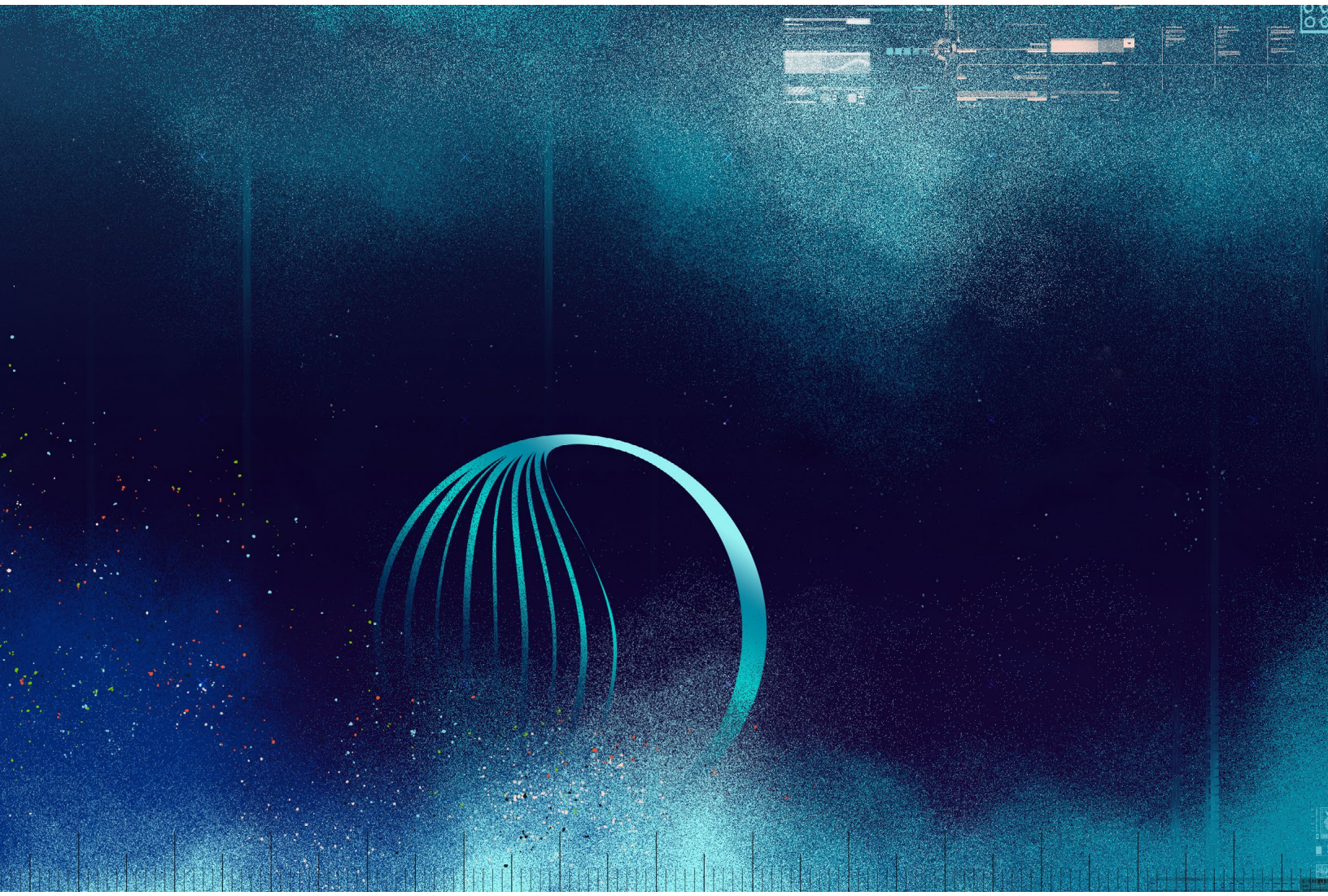
signal65



The NVIDIA DGX Spark Platform: Arm and NVIDIA Reinvent the Workstation

AUTHOR

Ryan Shrout
President and GM | Signal65

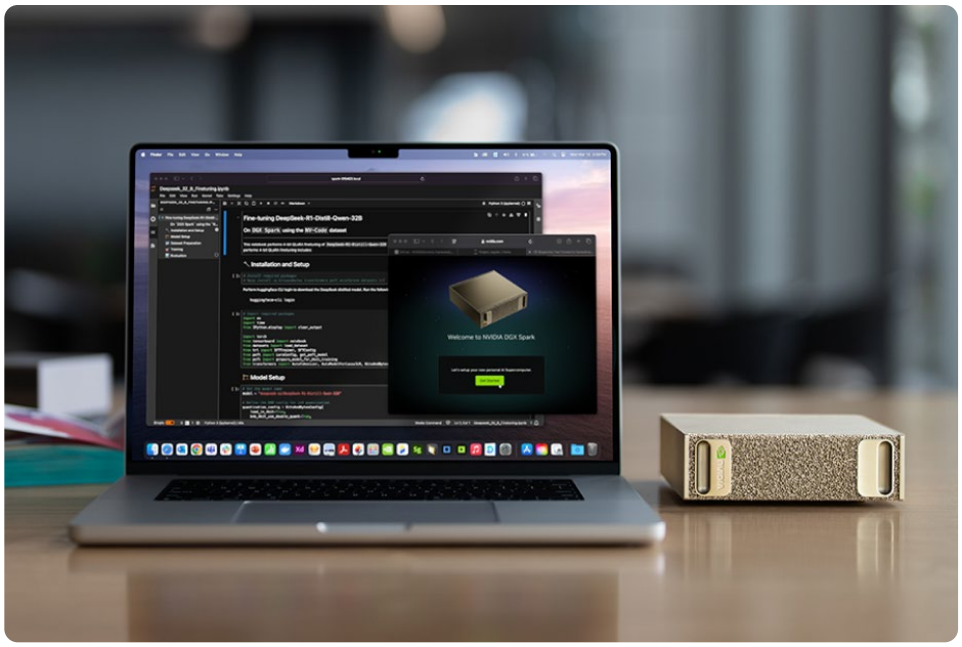
IN PARTNERSHIP WITH

arm

MARCH 2026

Introduction

NVIDIA DGX Spark represents a meaningful inflection point in the personal workstation market. Originally introduced as part of the NVIDIA Project DIGITS initiative, DGX Spark is a compact, desktop-class AI supercomputer built around the GB10 Grace Blackwell chip, a co-designed processor from NVIDIA and MediaTek that pairs an Arm-based CPU complex with a full Blackwell GPU. Its stated purpose is to bring data-center-class AI capability to the desk of every developer, researcher, and engineer who needs serious local compute without a cloud invoice.



The GB10 CPU complex marks a significant differentiator: it is essentially the first Arm-based processor to enter the professional workstation class in a commercially shipping product targeting developer and engineering workflows. Built around ten Arm Cortex-X925 cores and ten Cortex-A725 cores, the GB10 CPU does not arrive as an experiment.

Instead, it shows up as a product positioned against established x86 workstations from HP, Dell, and others. The question the market is now asking is whether the Arm instruction set architecture, long dominant in mobile and now ascendant in hyperscale data centers, is ready to compete at the workstation level, where software optimization depth and raw throughput have historically favored x86.

The Signal65 evaluation situates the DGX Spark against two directly competing small form factor (SFF) workstations: the HP Z2 Mini G1a, powered by the Ryzen AI MAX+ Pro 395 from AMD (Strix Halo), and the HP Z2 Mini G1i, which pairs an Intel Core Ultra 7 265 with a discrete NVIDIA RTX 4000 SFF Ada GPU. These platforms represent the strongest x86 SFF competition currently available: one leveraging the high-bandwidth integrated GPU architecture from AMD, the other relying on a purpose-built discrete GPU.

Together they define a credible performance ceiling for the x86 SFF category, making them the appropriate reference points for assessing the DGX Spark competitive position.

Three narrative threads run through this paper. First, we examine whether the Arm + unified memory architecture of the GB10 delivers on its architectural promise across a broad range of traditional

CPU workloads (rendering, compilation, scientific computing, and more) where x86 has decades of software optimization momentum.

Second, we quantify the practical impact of the DGX Spark and its 128GB unified memory pool, which enables local execution of models that are simply out of reach for competing platforms constrained by GPU VRAM.

Third, we assess the DGX Spark as a developer platform: not just a benchmark runner, but a viable daily-use machine for AI developers who want the full NVIDIA software ecosystem (vLLM, NeMo, CUDA) running locally on hardware that is architecturally continuous with NVIDIA cloud infrastructure.

System Details and Competitive Comparisons

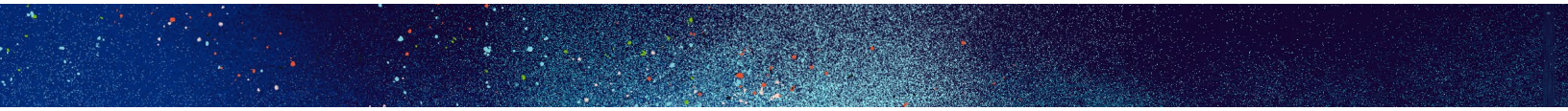
This evaluation covers three small form factor workstations selected to represent the leading available configurations in this market segment. Each system reflects a distinct architectural strategy: the DGX Spark pairs an Arm CPU with a Blackwell-generation integrated GPU and a large unified memory pool; the HP Z2 Mini G1a relies on the Strix Halo APU from AMD, which integrates a capable RDNA 3.5 GPU alongside a Zen 5 CPU core complex within a shared 128GB memory architecture.

The HP Z2 Mini G1i pairs a conventional Intel Core Ultra 7 CPU with a discrete NVIDIA RTX 4000 SFF Ada Lovelace GPU offering 20GB of dedicated GDDR6 VRAM.

Each design involves meaningful tradeoffs in memory capacity, bandwidth, and AI software ecosystem maturity.

Feature	NVIDIA DGX Spark	HP Z2 Mini G1a	HP Z2 Mini G1i
CPU	ARMv8 Cortex-X925 (10C) Cortex-A725 (10C)	AMD Ryzen AI MAX+ Pro 395	Intel Core Ultra 7 265
GPU	NVIDIA GB10 (Blackwell)	AMD Radeon (integrated)	NVIDIA RTX 4000 SFF Ada 20GB
Memory	128GB LPDDR5X-8533 (unified)	128GB (unified)	System + 20GB GDDR6 VRAM
Architecture	Arm + unified memory	x86 + unified memory	x86 + discrete GPU
Est. Price	~$4,500	~$4,000	~$4,300

From a pricing perspective, the three systems occupy a similar bracket. The NVIDIA DGX Spark is priced at approximately $4,500 USD today. The HP Z2 Mini G1a and G1i configurations in the 128GB memory tier are comparably priced in the $3,700 to $4,400 range depending on configuration. This pricing proximity is meaningful: buyers in this segment are making architectural and ecosystem decisions, not just raw cost comparisons. At similar price points, the memory architecture, GPU compute capability, and software ecosystem compatibility become the primary differentiating factors.



CPU Compute Performance

The goal of this section is to assess how the GB10 Arm-based CPU performs across a broad set of CPU-bound workloads relative to the competing x86 platforms. Modern workstations are evaluated on more than GPU tasks: rendering, compilation, scientific simulation, and productivity workflows remain integral to daily professional use. Establishing the GB10 CPU's competitiveness here is important because the DGX Spark is positioned as a complete workstation replacement, not a dedicated inference appliance.

CPU Compute Rendering

CPU rendering workloads stress floating-point throughput, cache hierarchy efficiency, and the ability to saturate all available cores with parallelizable ray-casting or related tasks. These tests are relevant to 3D artists, visualization engineers, and any workflow that leans on CPU-side rendering as a fallback or primary path. In our testing, C-Ray and AOBench were used as representative benchmarks across this workload class.

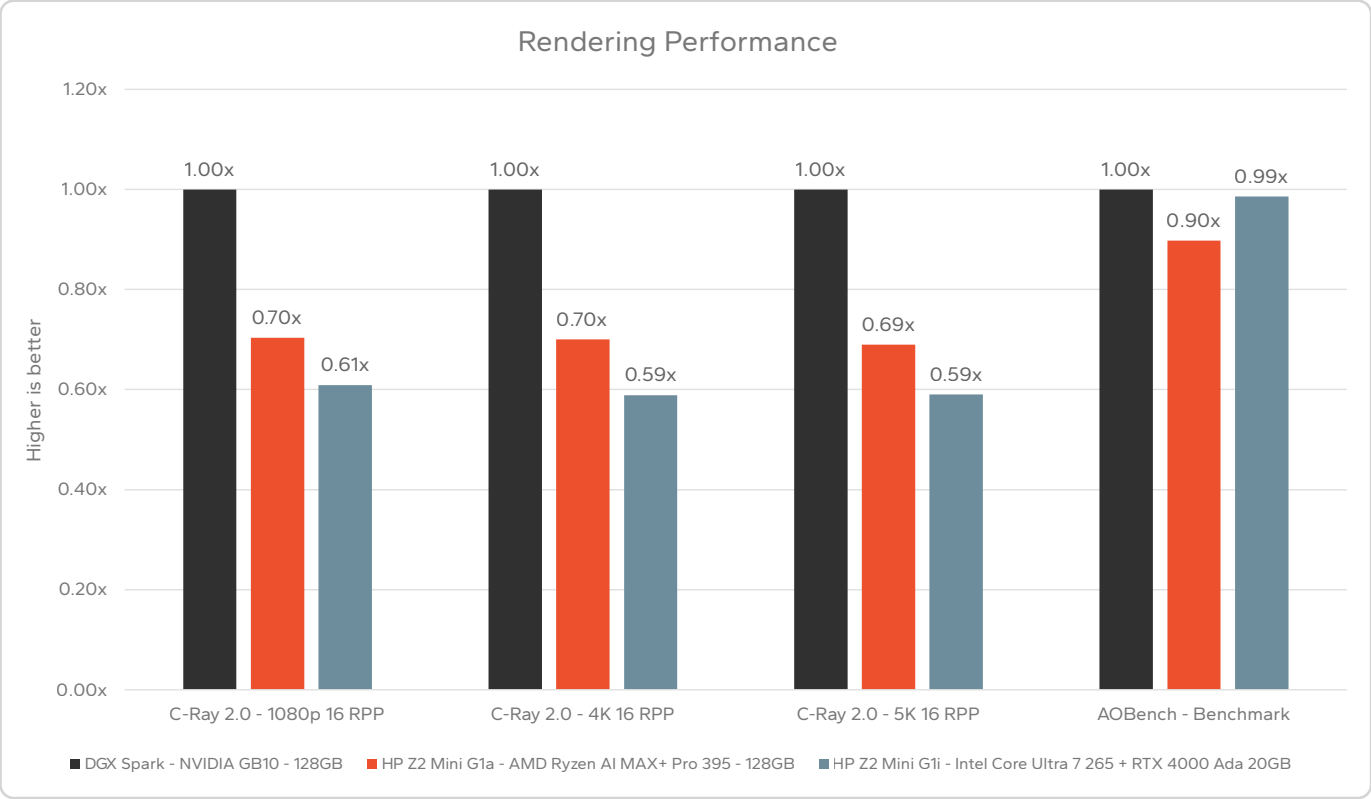


Figure 1: *CPU Rendering Benchmark Result*

In our testing, the DGX Spark posted C-Ray 2.0.0 scores 30% to 41% higher than the HP Z2 Mini G1a (AMD) and 39% to 41% higher than the HP Z2 Mini G1i (Intel) across all three tested resolutions (1080p, 4K, and 5K). The consistency of this lead across resolutions indicates a clear advantage in this workload class. This result indicates that for CPU-bound ray tracing, the GB10 is in a different performance tier than either competing x86 platform.

In the AOBench ambient occlusion benchmark, the DGX Spark led the AMD system by approximately 10%, while the Intel result normalizes to 0.99x (effectively at parity). This convergence on AOBench versus the larger C-Ray gap is notable: AOBench is a less memory-bandwidth-sensitive workload, and the tighter spread here suggests that x86 platforms can close the rendering gap in workloads that are more compute-bound than bandwidth-bound. The DGX Spark still leads outright, but the Intel system's near-parity performance illustrates the nuance that not all CPU rendering tasks yield the same architectural advantage.

Code Compilation

Software compilation is a critical workload for many developers: it measures how quickly a machine can parse, transform, and link large source trees under realistic conditions. Compilation is generally latency-sensitive (developers wait for builds) and exercises a combination of single-threaded parse depth and multi-threaded parallelism. In our testing, we used FFMPEG 7 compilation time as the primary benchmark, as FFMPEG is a large, well-known open-source project with a complex dependency graph that stresses the full build toolchain.

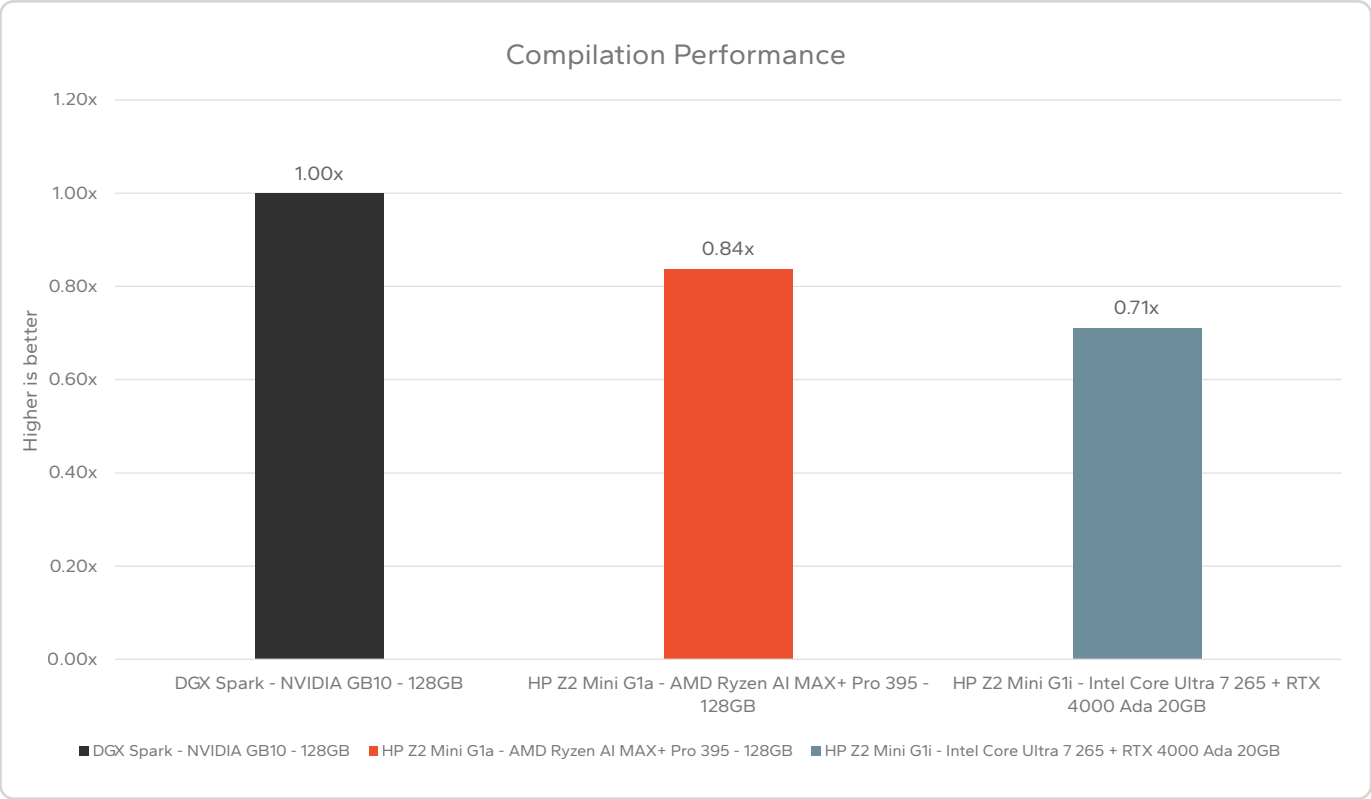


Figure 2: *Code Compilation Benchmark Results*

In our testing, the DGX Spark compiled FFMPEG 7 approximately 16% faster than the HP Z2 Mini G1a (AMD) and 29% faster than the HP Z2 Mini G1i (Intel). For a developer who runs several incremental builds per day, a reduction in compile time represents a meaningfully faster iteration cycle. This result indicates that the GB10 Cortex-X925 cores, combined with the high-bandwidth LPDDR5-8533 memory subsystem, are well-suited to compilation workflows that mix sequential parsing with parallelized linking stages.

Video Encoding

Video encoding measures a system's ability to transcode video content from one format or resolution to another, a workload relevant to content creators, media production pipelines, and any development environment where video assets need processing. We used Handbrake to transcode a 4K H.264 source to 1080p H.265 using the GPU-accelerated pathway, a common production workflow that exercises both decode/encode pathways and memory bandwidth.

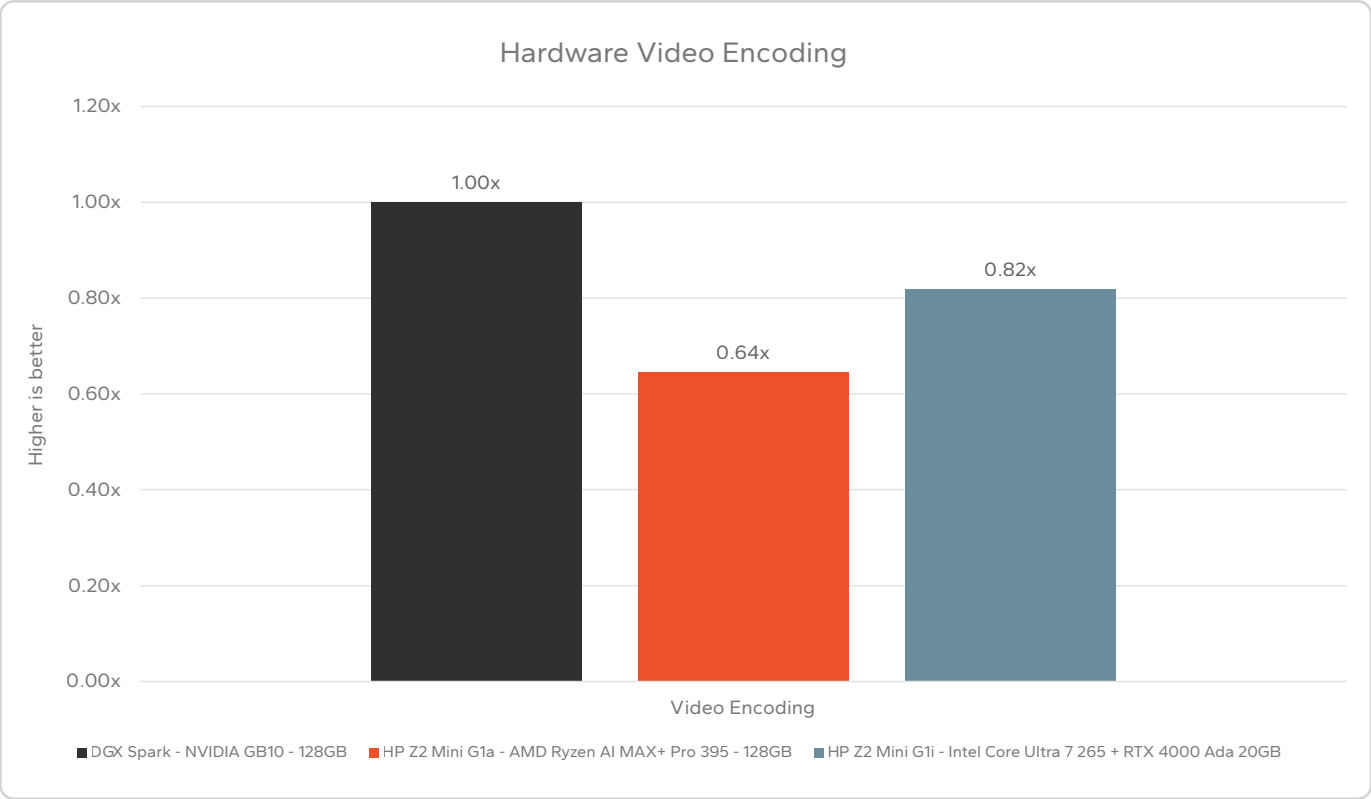


Figure 3: *Video Encoding Benchmark Results*

In our testing, the DGX Spark delivered the fastest Handbrake transcode times across the test, with the AMD system scoring 0.64x and the Intel system scoring 0.82x of DGX Spark performance. The AMD result in particular is notable: despite the Ryzen AI MAX+ Pro 395's strong multi-core profile, it trails the DGX Spark significantly on this H.264-to-H.265 pipeline. These results suggest the GB10 CPU handles the combination of frame preparation, pipeline orchestration, and memory movement stages in this specific Handbrake configuration more efficiently than either x86 competitor.

Scientific Computing

Scientific computing benchmarks simulate real-world numerical workloads used in research, engineering simulation, and life sciences. These tests exercise double-precision floating-point arithmetic, memory access patterns, and vectorized computation paths, the same characteristics that determine performance in finite element analysis, computational fluid dynamics, and bioinformatics pipelines. In our testing, we used Dolfyn (CFD), HMMer Search (bioinformatics), and LULESH (hydrodynamics simulation).

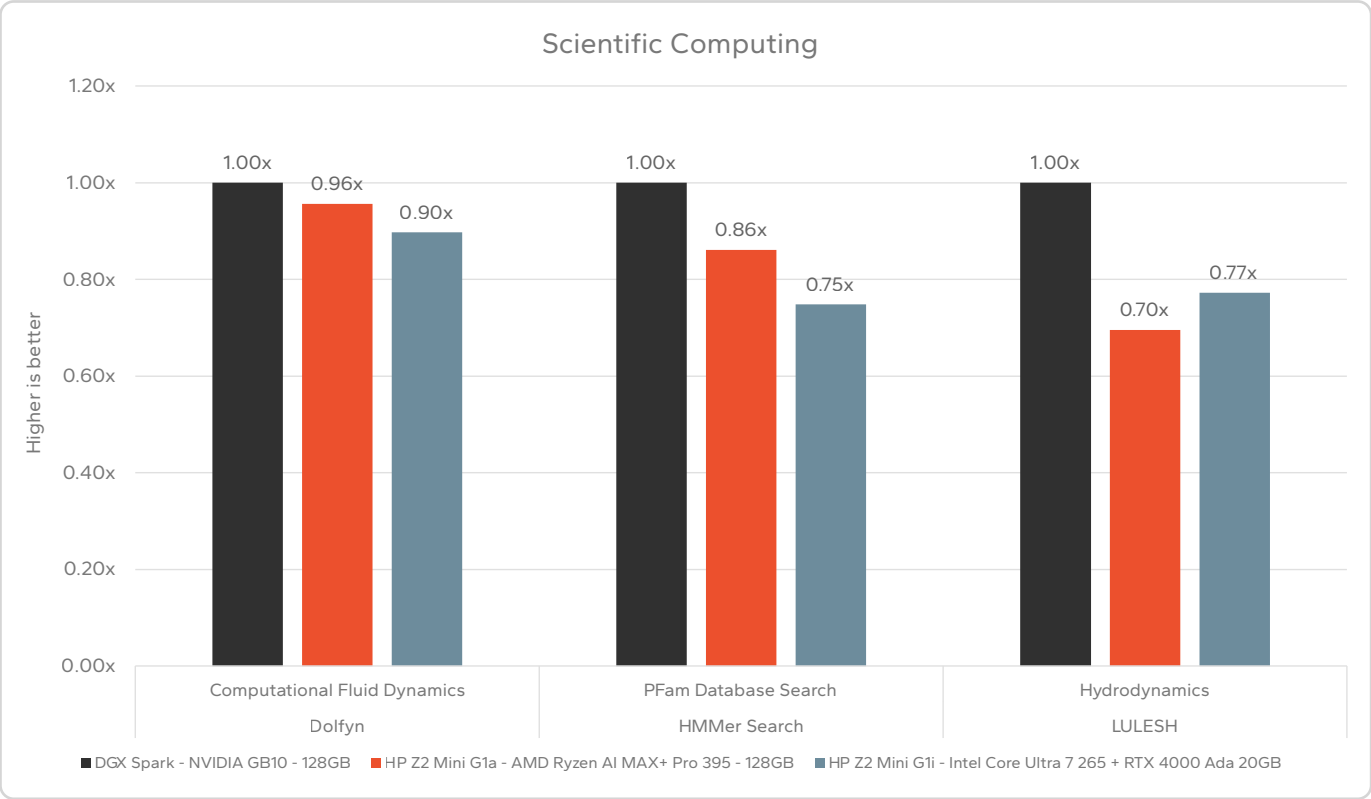


Figure 4: *Scientific Computing Benchmark Results*

In our testing, the DGX Spark led all three of these scientific computing workloads. On Dolfyn 0.527 (computational fluid dynamics), AMD scored 0.96x and Intel scored 0.90x relative to the DGX Spark, a modest but consistent lead that reflects the GB10's strong memory bandwidth in a workload dominated by sparse matrix operations. The near-parity on Dolfyn specifically is worth noting: CFD workloads are well-optimized on x86 platforms, and the DGX Spark maintaining a lead here speaks to the growing maturity of the Arm scientific computing software stack.

On HMMer Search 3.4, a sequence database search tool widely used in bioinformatics, the DGX Spark lead widens: AMD scored 0.86x and Intel scored 0.75x. HMMer is sensitive to both compute throughput and memory access latency, and the LPDDR5-8533 memory subsystem's bandwidth advantage over the competing configurations appears to contribute here.

The LULESH hydrodynamics simulation shows the largest DGX Spark advantage in this section: AMD scored 0.70x and Intel scored 0.77x, a 30%+ lead for DGX Spark on a workload that represents the type of structured-mesh Lagrangian simulation used in materials science and weapons physics research. Across all three scientific workloads, the DGX Spark establishes a consistent and defensible performance advantage over both x86 competitors.

Memory Performance

Memory bandwidth is increasingly the binding constraint in some modern computational workloads: AI inference, scientific simulation, large-data analytics, and even compilation are all limited not always by raw compute throughput but by how quickly data can move between memory and processing units. The GB10's LPDDR5X-8533 unified memory subsystem, operating at significantly higher bandwidth than the DDR5 configurations in the competing HP systems, creates a structural advantage that should be visible in memory-bound workloads. We measured this using RAMspeed SMP in both Integer Add and Integer Copy configurations.

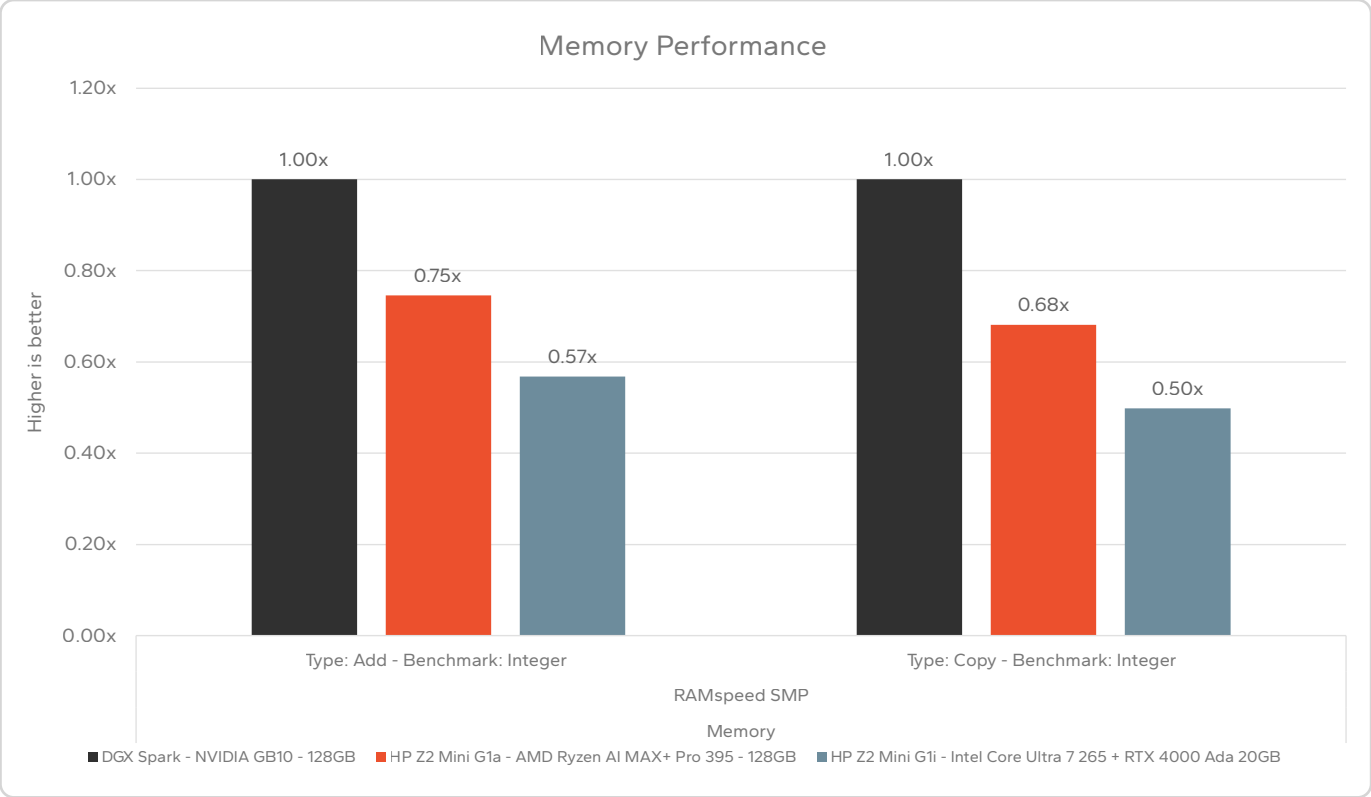


Figure 5: *Memory Performance Benchmark Results*

In our testing, the DGX Spark demonstrated a 25% to 32% memory bandwidth advantage over the AMD system and a 43% to 50% advantage over the Intel system on RAMspeed SMP Integer Add and Copy operations. These margins are not minor: they reflect an architectural difference in memory subsystem design. The LPDDR5-8533X interface at 128GB capacity delivers aggregate bandwidth that neither the HP Z2 Mini G1a nor the G1i can match.

This result is practically significant because memory bandwidth constraints manifest directly in AI inference throughput, scientific simulation time-to-solution, and any workload that operates on large in-memory datasets, meaning this advantage compounds across the DGX Spark's full workload portfolio.

Productivity and Engineering

Productivity and desktop engineering workloads represent the daily-use dimension of workstation performance: document processing, office automation, and circuit simulation. While these workloads may not headline a benchmark report, they determine whether a machine feels fast in practical day-to-day use. We tested document conversion throughput with LibreOffice and circuit simulation performance with NGSpice across two distinct circuit configurations.

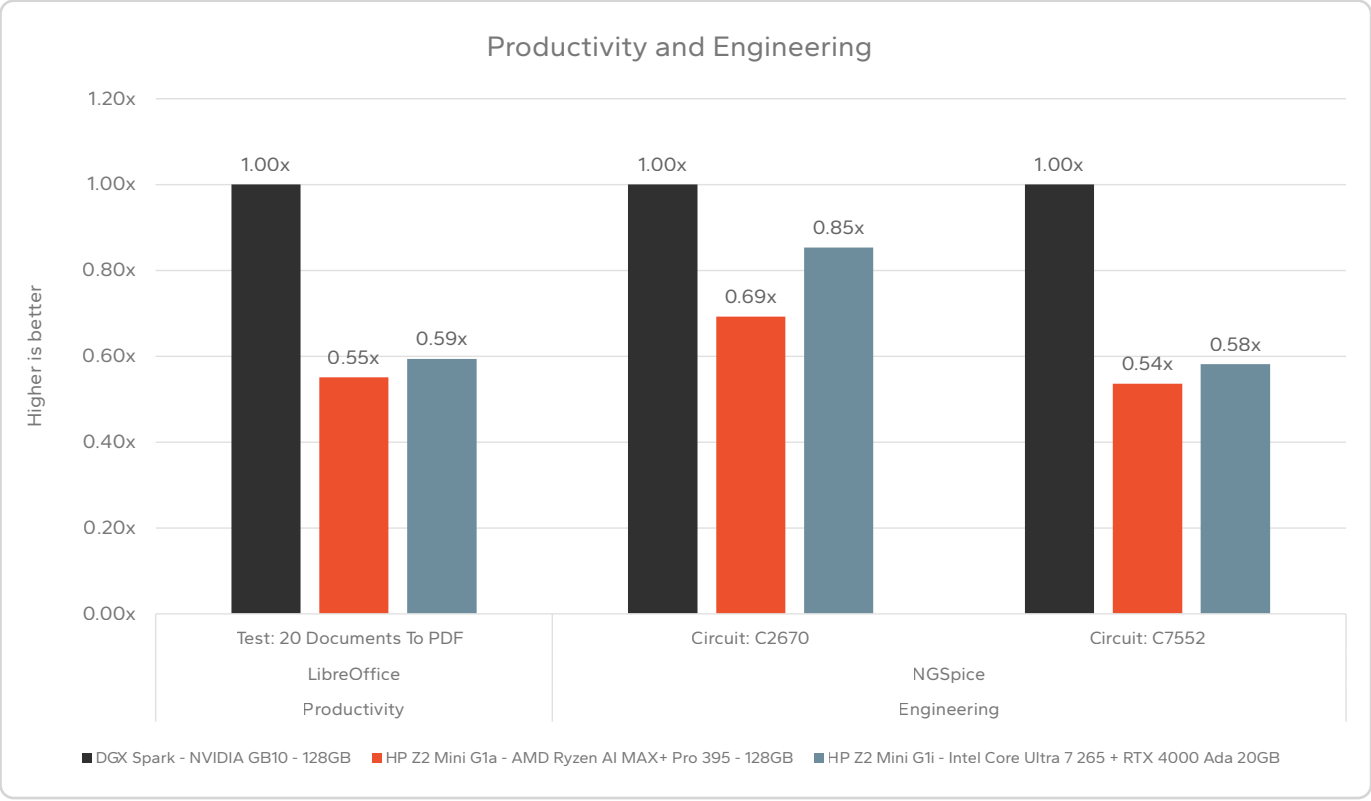


Figure 6: *Productivity and Engineering Benchmark Results*

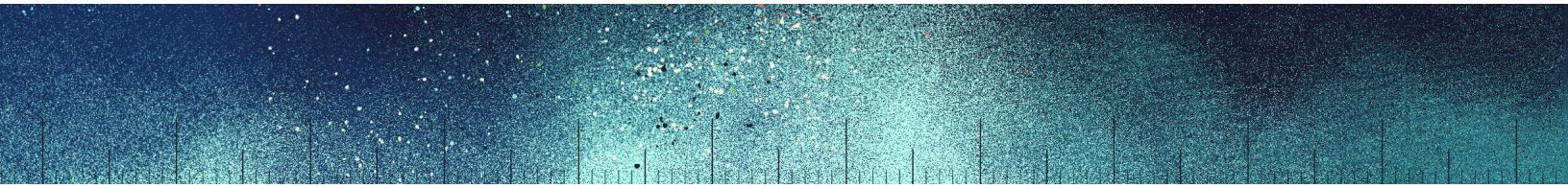
In our testing, the DGX Spark led the LibreOffice document-to-PDF batch conversion, with the AMD system scoring 0.55x and Intel scoring 0.59x, meaning the competing platforms took roughly 70% to 80% longer to complete the same 20-document batch. For productivity users running large document workflows or scripted office automation pipelines, this difference is immediately tangible. On NGSpice circuit simulation, the DGX Spark again led both configurations tested: AMD scored 0.54x to 0.69x and Intel scored 0.58x to 0.85x across the two circuits. Circuit simulation is sensitive to both floating-point throughput and memory access patterns, and the DGX Spark architecture handles both dimensions effectively here.

Section Key Insights

The DGX Spark and its Arm-based GB10 CPU is broadly competitive, and frequently superior, across the full range of traditional CPU workloads tested. The strongest DGX Spark advantages appear in workloads that are memory-bandwidth-sensitive (RAMspeed, HMMer, LULESH), floating-point-intensive (C-Ray), and cache-efficient (LibreOffice, NGSpice). In these categories, the Arm CPU core design combined with the LPDDR5X-8533 memory subsystem produces a consistent and measurable lead over both the AMD Ryzen AI MAX+ Pro 395 and the Intel Core Ultra 7 265 platforms.

x86 platforms retain competitive parity or narrow advantages in a subset of workloads, particularly those with deeply optimized SIMD code paths (such as video codec processing) or workloads that scale primarily with raw thread count rather than per-thread throughput or memory bandwidth. These gaps are a reflection of the software optimization history of the x86 ecosystem rather than a fundamental architectural limitation of Arm. As the Arm workstation software ecosystem matures and ISV optimization efforts accumulate, these remaining x86 advantages are likely to narrow.

These results establish that the GB10 CPU provides more than sufficient daily-use performance to serve as a workstation-class processor. Developers, engineers, and researchers who adopt the DGX Spark as their primary workstation will not be making a CPU performance compromise; in most benchmarked workload categories, they will be gaining a measurable performance advantage over these x86 SFF alternatives.



AI Workload Performance

The goal of this section is to establish if the DGX Spark is really leading local AI developer platform in the SFF workstation class as NVIDIA has marketed. This means measuring if the DGX Spark executes the same AI workloads faster than competing platforms, and also that its 128GB unified memory architecture unlocks model capabilities that are out of reach for the Intel configuration and, for some workloads, deliver a qualitatively different experience from what the AMD system can provide. Workloads tested include LLM inference at multiple model scales, multi-user concurrency, image generation, video generation, and model fine-tuning.

LLM Inference: Model Capability

Large language model inference is the most consequential local AI workload for developers and knowledge workers in 2026. Performance in this category determines how quickly a developer can iterate on prompts, how much context a model can process in a single session, and whether a given model can run locally at all. We tested four models at varying parameter counts and quantization levels to profile each platform's inference capability across the full range of practically relevant model sizes.

GPT-OSS 20B (FP4)

At 20B parameters in FP4 quantization, GPT-OSS 20B represents a capable, production-quality reasoning model that all three tested platforms can execute locally. This provides a direct apples-to-apples comparison across all three systems. The critical distinction at this model scale is not whether a platform can run the model, but how efficiently it processes context; prompt processing speed directly determines how quickly a developer can work with large inputs, code repositories, or extended conversation histories.

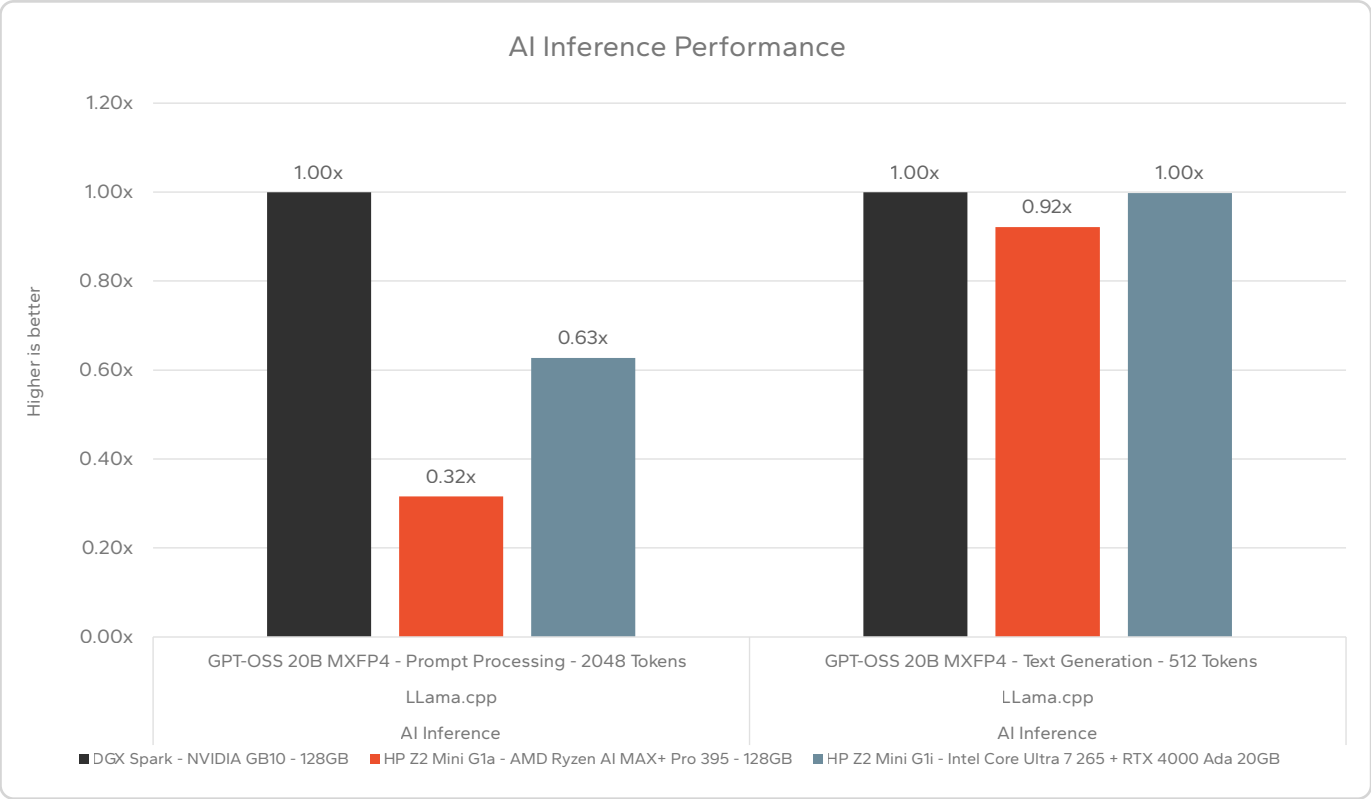


Figure 7: *GPT-OSS 20B Inference Results*

In our testing, the DGX Spark processed prompts at 3.2x the rate of the AMD system and 1.6x the rate of the Intel system on GPT-OSS 20B FP4. Text generation speeds, by contrast, were roughly equivalent across all three platforms (AMD at 0.92x, Intel at 1.00x relative to DGX Spark). This asymmetry is architecturally meaningful: prompt processing is highly memory-bandwidth-bound, while token generation is more compute-bound and involves smaller active data footprints. The DGX Spark's prompt processing performance translates directly into a better end-user experience (faster time-to-first-token on complex, long-context inputs) even when final token generation rates converge.

GPT-OSS 120B (FP4)

At 120B parameters in FP4 quantization, GPT-OSS 120B represents a frontier-class reasoning model, the type of large, capable model that organizations typically access exclusively through cloud APIs due to the memory requirements for local inference. In our testing configuration, only the DGX Spark and the HP Z2 Mini G1a (AMD) were able to load and execute this model. The HP Z2 Mini G1i (Intel) with its 20GB RTX 4000 Ada GPU has insufficient GPU VRAM to host a 120B FP4 model, and the Intel system's CPU-only path cannot operate the model at practical inference speeds.

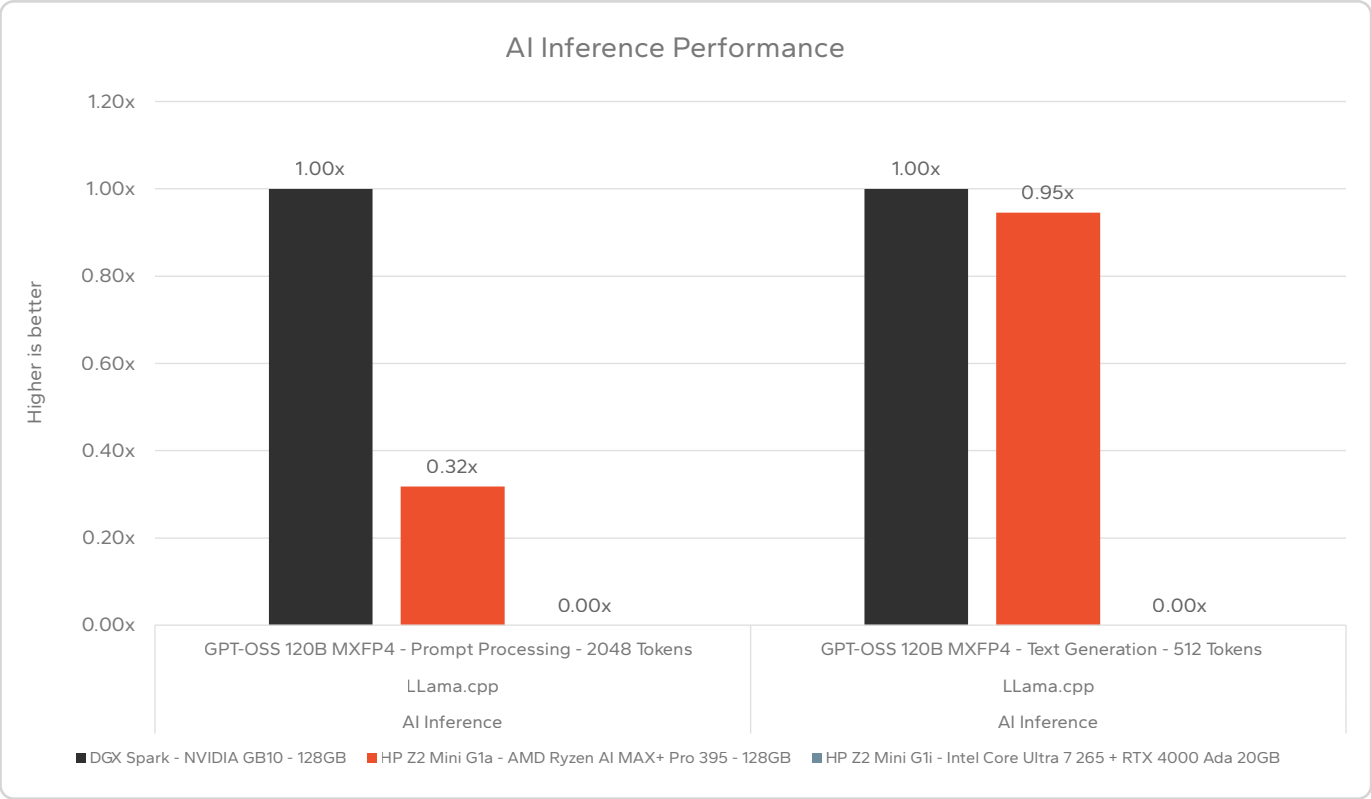


Figure 8: *GPT-OSS 120B Inference Results*

In our testing, the DGX Spark's prompt processing advantage on the 120B model is 3.14x over the AMD system, virtually identical to the 3.2x lead at 20B, indicating that the architectural bandwidth advantage scales consistently with model size. Text generation rates remain close between the two platforms (AMD at 0.95x), confirming the same bandwidth-vs.-compute dynamic observed at smaller model scales. These results suggest that for organizations considering local 120B-class model deployment, the DGX Spark offers not only the ability to run the model (which the Intel platform cannot match at all) but a significantly better interactive experience due to its superior context processing throughput.

LLaMA 3.3 70B Instruct (Q4_K_M)

LLaMA 3.3 70B Instruct in Q4_K_M quantization is one of the most widely deployed open-weight models among developers. It represents the mainstream upper bound of what most local inference setups attempt to run: capable enough for serious production use cases, large enough that memory becomes the binding constraint. As with GPT-OSS 120B, the Intel system's 20GB VRAM prevents it from running this model, making this a two-platform comparison.

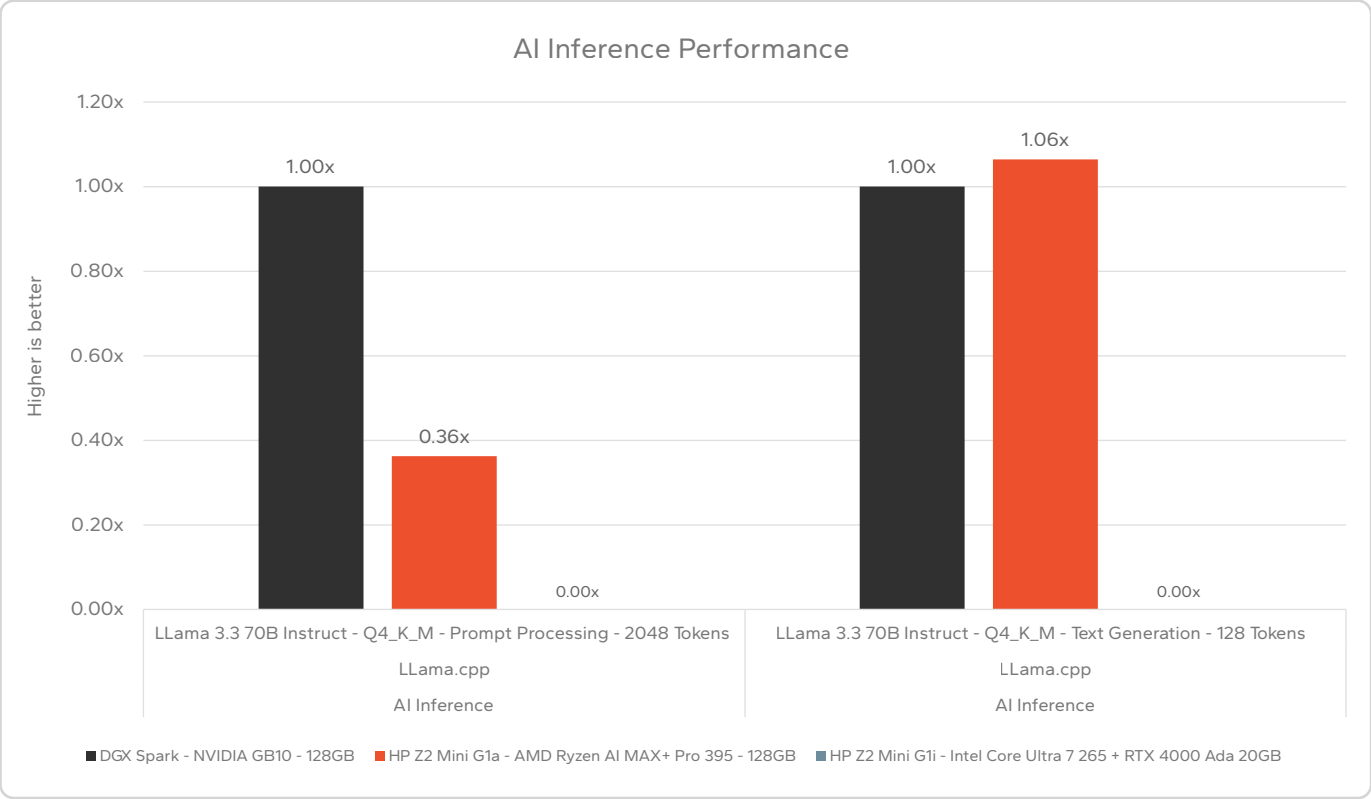


Figure 9: *LLaMA 3.3 70B Instruct Inference Results*

In our testing, the DGX Spark led LLaMA 3.3 70B prompt processing by 2.76x over the AMD system, a commanding advantage for workloads involving large context windows, long documents, or multi-turn conversations with extensive history. The AMD system posted a slight 1.06x advantage on text generation speed, meaning it generates output tokens marginally faster once prefill is complete. In practice, this small generation-speed advantage is largely offset by the DGX Spark's massive prefill lead: for any interaction involving a non-trivial prompt (code files, documents, system prompts), the DGX Spark delivers first-token significantly faster, which is much of the latency that users actually perceive.

Total throughput across a full prompt-plus-generation interaction favors the DGX Spark.

Qwen3 VL 32B Thinking (Q4_K_M)

Qwen3 VL 32B Thinking is a multimodal vision-language model with reasoning capabilities, representing an emerging class of workloads that combine visual input processing with extended chain-of-thought inference. At 32B parameters in Q4_K_M, all three tested platforms can execute this model, providing another full three-way comparison. This model class is increasingly relevant for developers building vision-capable AI applications or multimodal reasoning pipelines.

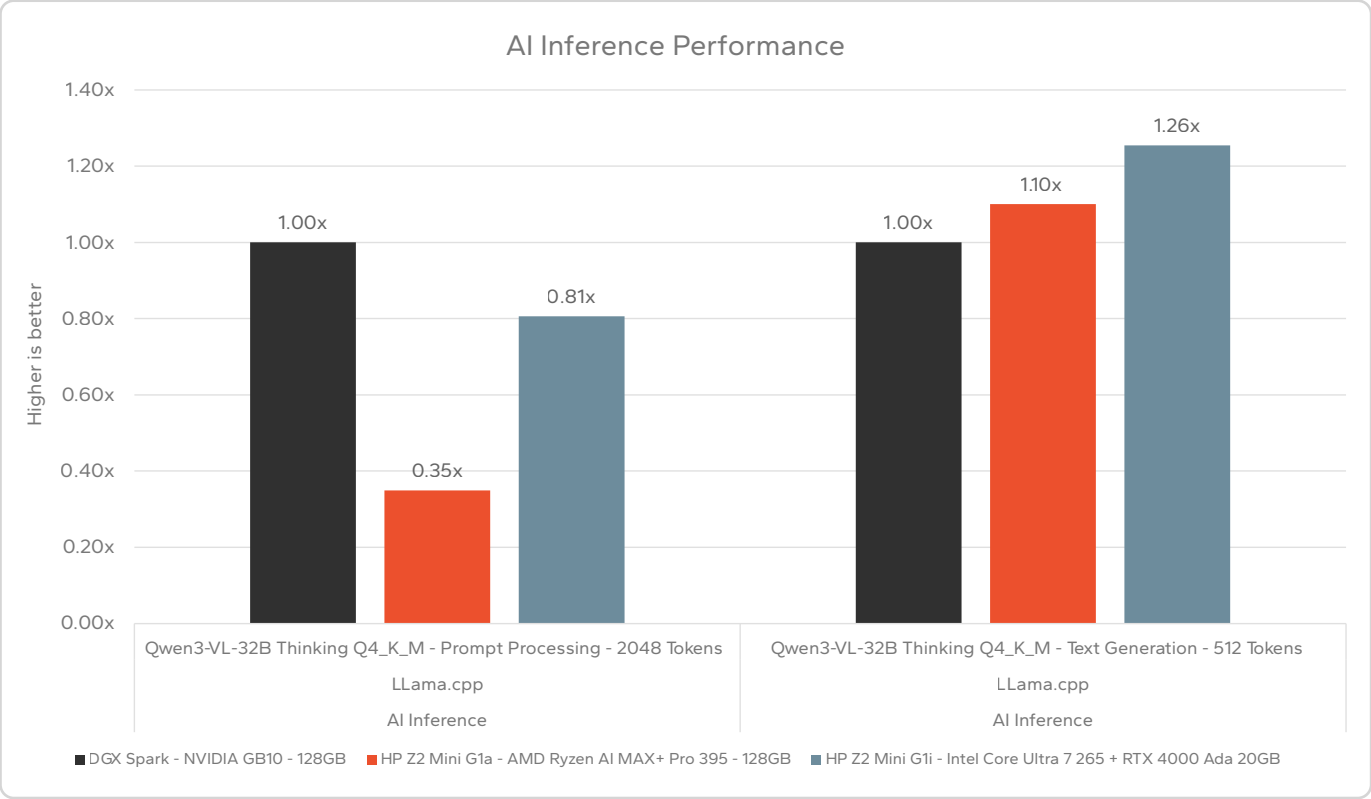


Figure 10: *Qwen3 VL 32B Thinking Inference Results*

In our testing, the DGX Spark again led prompt processing by 3.2x over AMD and 1.6x over Intel, a consistent ratio that appears to be a reliable architectural signature of the GB10's memory subsystem advantage on this class of workload. The text generation picture is more nuanced here: the AMD system is approximately 10% faster on token generation, and the Intel RTX 4000 Ada discrete GPU holds a roughly 25% advantage. This result is notable and worth contextualizing: the RTX 4000 Ada's higher memory bandwidth relative to its active compute footprint during generation gives it a discrete-GPU advantage on this specific path.

Even so, the DGX Spark's throughput on this model is fully production-capable, and its prompt processing performance means interactions involving large multimodal inputs (images, long system prompts, extended context) will still feel faster end-to-end on the DGX Spark for most real-world usage patterns.

Key Insight: Memory-Gated Model Access

The 128GB unified memory architecture on the DGX Spark is not merely a performance differentiator; it is a capability gate. On GPT-OSS 120B and LLaMA 3.3 70B, the Intel HP Z2 Mini G1i is simply unable to run the model at all. Its 20GB RTX 4000 Ada GPU, while capable in FP16 and FP8 workloads, cannot host the memory footprint of these scale models. This means that an Intel-based SFF workstation user who wants to run a 70B+ model locally has no path forward without cloud dependency, regardless of budget.

The DGX Spark, like the AMD system, clears this threshold; but only the DGX Spark pairs that model access with a 2.7x to 3.1x prompt processing advantage that makes the experience genuinely responsive at these scales.

LLM Inference: Multi-User Concurrency

For team environments (developer pods, research groups, or small engineering teams sharing a local inference server), single-user inference performance is necessary but not sufficient. The more relevant question is how a platform behaves under concurrent load: can it sustain low latency and reasonable throughput when multiple users are querying the same model simultaneously? We designed a concurrency test using Qwen3 Coder 30B A3B in 4-bit quantization, a code-focused model representative of AI-assisted development tooling, and compared the DGX Spark against the AMD Strix Halo system (HP Z2 Mini G1a) across simulated workloads scaling from 1 to 8 concurrent developers. In this testing we used the vLLM inference engine, the CUDA code path for DGX Spark and ROCm for AMD Strix Halo.

Two scenarios were tested: a Light Workload scenario modeled at 2 requests per second per developer, representing typical interactive coding assistant use, and a Heavy Workload scenario at 10 requests per second per developer, simulating an intensive or bursty workflow. Time-to-first-token (TTFT) and request throughput (req/s) were the primary metrics.

Scenario	DGX Spark TTFT	Strix Halo TTFT	DGX Spark Throughput	Strix Halo Throughput
1 Dev (20 prompts)	77 ms	354 ms	0.21 req/s	0.06 req/s
2 Devs (40 Prompts)	123 ms	464 ms	0.30 req/s	0.08 req/s
4 Devs (80 Prompts)	181 ms	677 ms	0.43 req/s	0.12 req/s
8 Devs (160 Prompts)	246 ms	768 ms	0.68 req/s	0.16 req/s

Figure 11: *Multi-User Concurrency*

In the Light Workload scenario, the DGX Spark posted time-to-first-token of 77ms with a single developer, scaling gracefully to 246ms at 8 concurrent developers. The Strix Halo system started at 354ms TTFT with a single user and degraded to 768ms at 8 developers, a baseline that already exceeds the DGX Spark's 8-developer degraded result. From a user experience standpoint, 246ms TTFT is within the range of imperceptible latency for interactive tools; 768ms is a noticeable pause that compounds per request in a busy coding session.

Throughput in the light scenario ranged from 0.21 req/s (1 dev) to 0.68 req/s (8 devs) for DGX Spark versus 0.06 to 0.16 req/s for Strix Halo, a roughly 4x throughput advantage that holds consistently across all concurrency levels.

Scenario	DGX Spark TTFT	Strix Halo TTFT	DGX Spark Throughput	Strix Halo Throughput
1 Dev (20 prompts)	71 ms	488 ms	0.23 req/s	0.07 req/s
2 Devs (40 Prompts)	114 ms	471 ms	0.33 req/s	0.09 req/s
4 Devs (80 Prompts)	159 ms	535 ms	0.47 req/s	0.13 req/s
8 Devs (160 Prompts)	220 ms	712 ms	0.74 req/s	0.18 req/s

Figure 12: *Multi-User Concurrency*

The Heavy Workload scenario is where the DGX Spark's architecture demonstrates its most decisive concurrency advantage. Under heavy load at a single user, the Strix Halo system's TTFT jumps to 488ms (already over 6x the DGX Spark's 71ms result) and climbs to 712ms at 8 developers. The DGX Spark's TTFT under the same heavy load scales from 71ms (1 dev) to 220ms (8 devs), remaining below 250ms even at peak concurrency.

Throughput followed the same pattern: DGX Spark delivered 0.23 to 0.74 req/s versus Strix Halo's 0.07 to 0.18 req/s across the 1-to-8-developer range. These results indicate that the DGX Spark can meaningfully serve as a shared team inference server in small-office or developer-pod configurations, a use case where the AMD platform's latency would result in a noticeably slower interactive experience.

Key Insight: Concurrency and Team Productivity

At scale, the DGX Spark's multi-user concurrency advantage is 3x to 7x faster time-to-first-token and approximately 4x higher request throughput compared to the AMD Strix Halo platform across both light and heavy workload scenarios. Critically, the DGX Spark's TTFT degradation curve is significantly shallower: scaling from 1 to 8 concurrent developers adds roughly 170ms of TTFT in the light scenario versus over 400ms on the Strix Halo system.

For a team of developers sharing a local model server, the DGX Spark provides a qualitatively solid experience; interactive latency remains within productive bounds even under load.

Image Generation

Generative image workloads exercise GPU compute throughput, memory bandwidth, and floating-point format support in ways that directly reflect the GPU's generation and architecture. We tested four image generation pipelines across varying precision levels to assess both absolute performance and model compatibility: Stable Diffusion FP16, Flux1.Dev FP8, Z-image Turbo BF16, and Z-image BF16. The AMD Strix Halo system could not execute the FP16 or FP8 configurations of Stable Diffusion and Flux1.Dev, limiting it to the BF16 workflows.

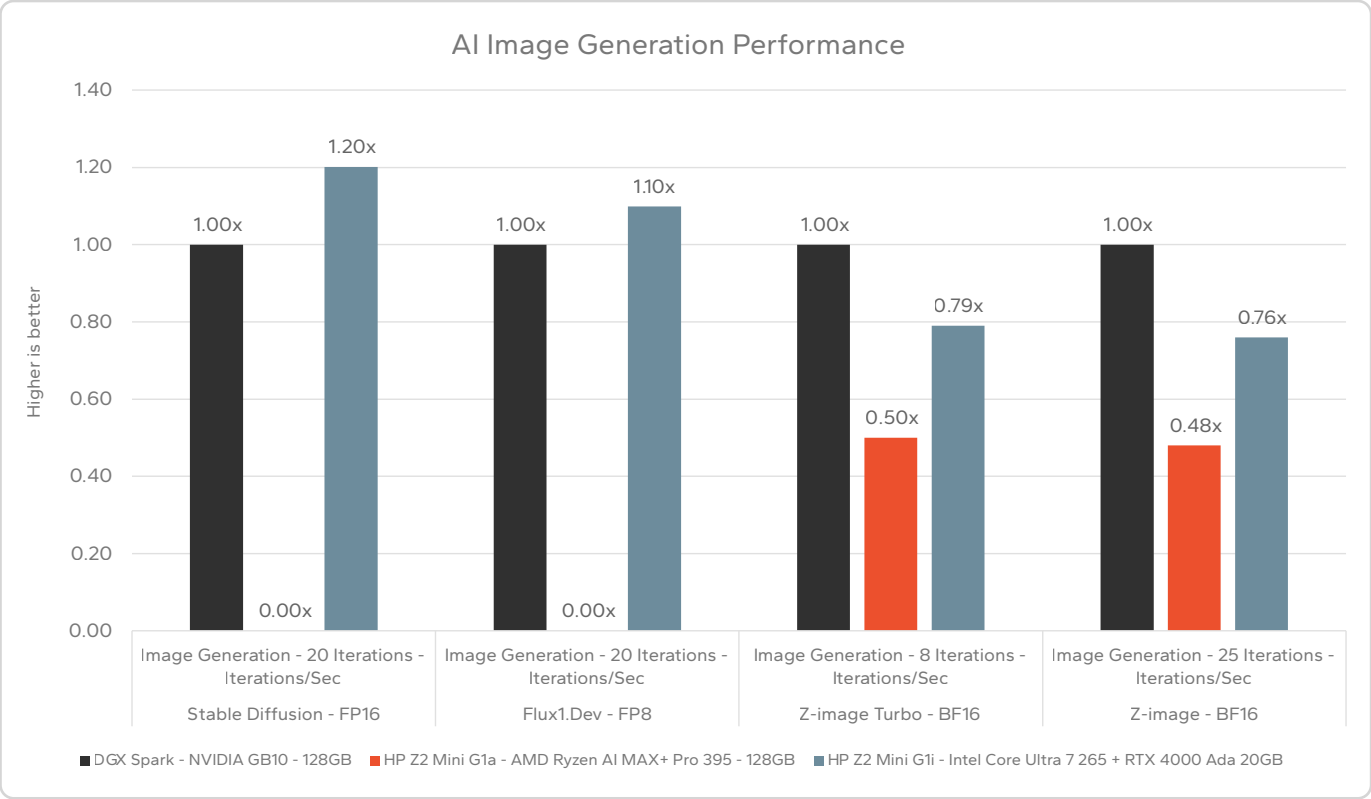


Figure 13: *Image Generation Benchmark Results*

In our testing, Stable Diffusion FP16 (20 iterations) showed the Intel RTX 4000 Ada discrete GPU with a 1.20x advantage over the DGX Spark (14.21 it/s vs. 11.83 it/s). Flux1.Dev FP8 followed a similar pattern with discrete NVIDIA GPU system posting 1.67 it/s versus the DGX Spark's 1.52 it/s, a 10% lead. These results reflect the discrete Ada GPU's optimized FP16/FP8 execution path, where dedicated Tensor Core throughput and GDDR6 bandwidth on the RTX 4000 Ada produce a focused advantage.

Importantly, the AMD system could not run either workload. The DGX Spark runs them both at competitive speeds while also supporting the BF16 pipelines that represent the forward direction for generative workloads.

On the BF16 workloads (Z-image Turbo BF16 and Z-image BF16), the performance picture reverses. For Z-image Turbo BF16 (8 iterations), AMD scored 0.50x and Intel scored 0.79x relative to DGX Spark, meaning the DGX Spark completes the same number of iterations roughly twice as fast as AMD and approximately 27% faster than Intel. Z-image BF16 (25 iterations) showed DGX Spark at

0.50 it/s, with AMD at 0.48x and Intel at 0.76x; again, DGX Spark leading by approximately 2x over AMD and 1.3x over Intel.

These results indicate that as the generative AI ecosystem continues its migration toward BF16 and mixed-precision formats, the DGX Spark's GPU compute architecture is well-positioned to maintain or extend its advantage over both competing platforms.

Video Generation

Video generation is among the most compute- and memory-intensive generative AI workloads available today. It requires sustained GPU throughput across many diffusion steps on large activation tensors, placing simultaneous demands on compute, memory bandwidth, and memory capacity. We tested Wan 2.2 14B text-to-video generation at 4 iterations as the representative workload.

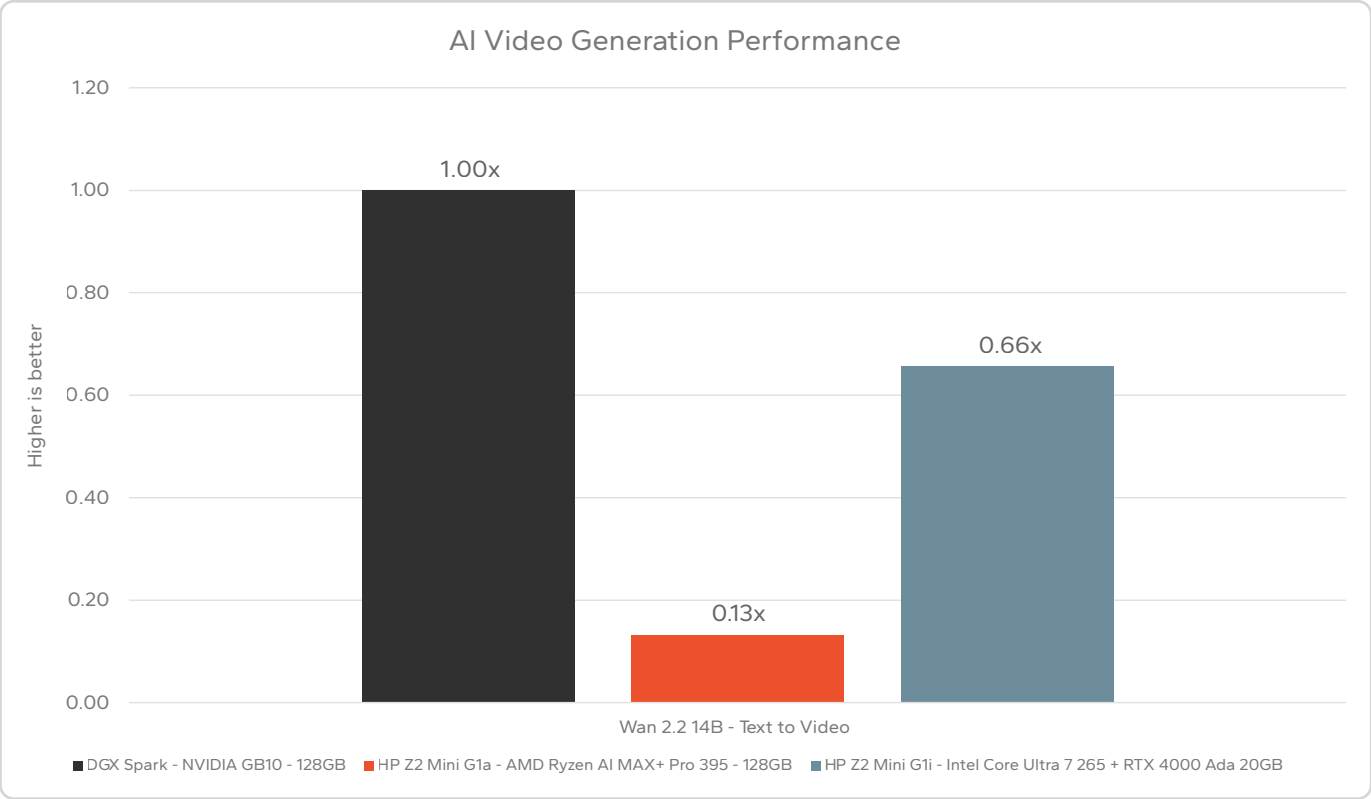


Figure 14: *Video Generation Benchmark Results*

In our testing, the DGX Spark completed Wan 2.2 14B video generation at 20.56 seconds per iteration. The AMD Strix Halo system required 156 seconds per iteration (7.6x slower), a gap large enough to render the AMD platform impractical for iterative video generation workflows. The Intel system, running on the discrete RTX 4000 Ada GPU, posted 31.32 seconds per iteration, 1.52x slower than the DGX Spark.

The magnitude of the AMD gap here is particularly significant: it reflects the combination of Wan 2.2 14B's large memory footprint stressing the AMD GPU's integrated memory bandwidth and the AMD GPU compute architecture's disadvantage on this class of diffusion workload. These results indicate that for video generation use cases, the DGX Spark is operating in a different performance tier, particularly relative to integrated GPU platforms.

Model Fine-Tuning

Model fine-tuning, adapting a pre-trained foundation model to a specific task using supervised training data, is a qualitatively different workload from inference. Using the Unsloth framework, we ran LoRA fine-tuning on OpenAI's GPT-OSS models loaded in MXFP4 precision, training rank-8 adapters on the HuggingFaceH4/Multilingual-Thinking multilingual reasoning dataset over 30 steps. Fine-tuning requires gradient computation, optimizer state storage, and activation checkpointing that multiply memory requirements well beyond inference, making it the most memory-intensive workload in our test suite.

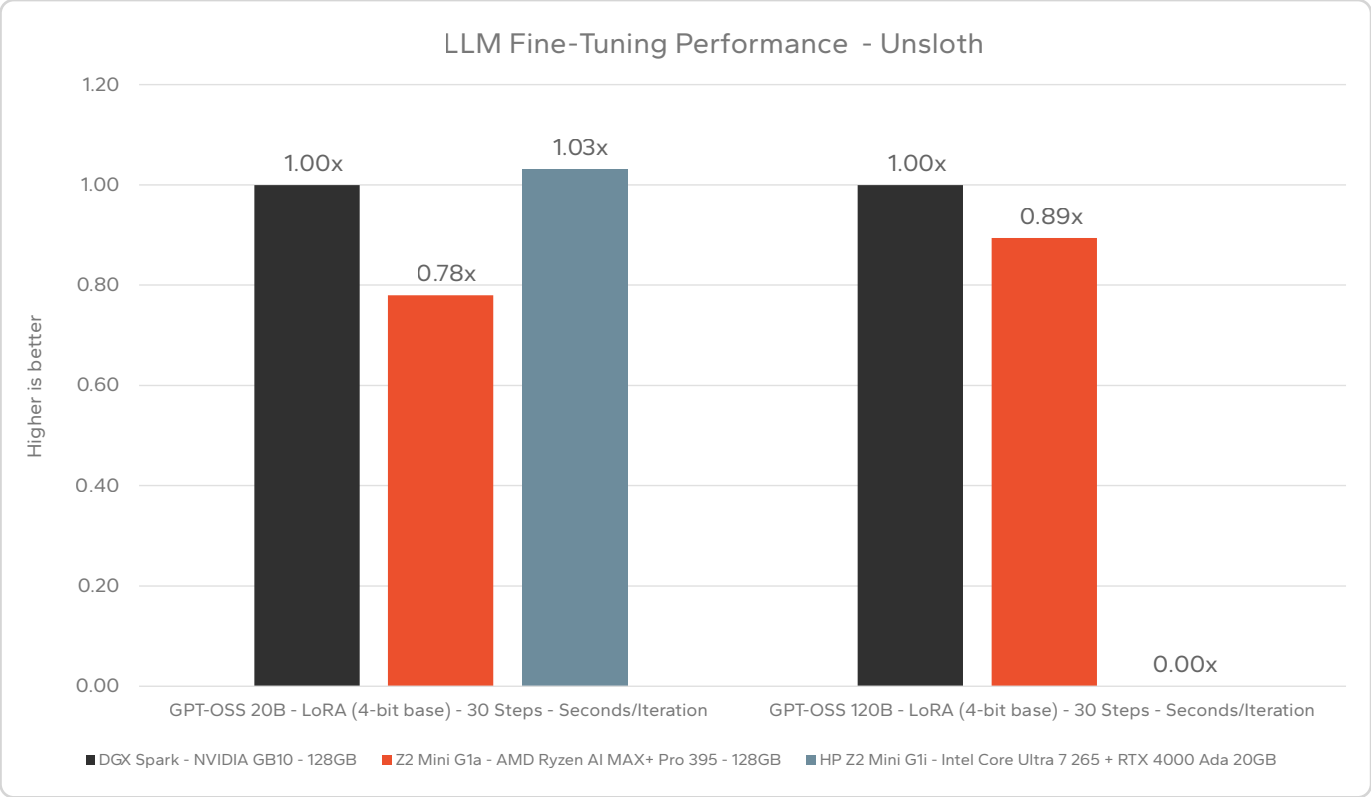


Figure 14: *Video Generation Benchmark Results*

LoRA fine-tuning with a 4-bit base quantization represents a more memory-efficient approach to model adaptation than full-precision training, enabling larger models to fit within constrained GPU memory budgets. At the 20B parameter scale, all three tested platforms successfully completed the workload. In our testing, the Intel HP Z2 Mini G1i posted a slight 3% advantage over the DGX Spark, while the AMD HP Z2 Mini G1a scored 0.78x relative to DGX Spark, making it approximately 22% slower per iteration.

The competitive picture shifts substantially at 120B parameters. The Intel HP Z2 Mini G1i cannot execute this workload at all since the RTX 4000 Ada GPU with 20GB of dedicated VRAM lacks sufficient memory to host a 120B LoRA fine-tuning job even with 4-bit base quantization. The AMD system completes the task at 0.89x relative to the DGX Spark, a narrower gap than on inference workloads at this scale.

The 120B result reinforces the memory-gated capability theme observed throughout the AI workload section. At smaller model scales where all platforms can operate, the performance spread is modest and the Intel discrete GPU is competitive. At frontier model scales, the Intel platform drops out entirely and only the DGX Spark and AMD Strix Halo systems with 128GB unified memory pools can execute the workload.

Section Key Insights

Taken together, the AI Workload Performance results establish the DGX Spark as a unique platform from the competing x86 SFF workstations in this evaluation. Its most decisive advantage is in prompt processing and context throughput: across every tested LLM configuration where both platforms could run the model, the DGX Spark led AMD prompt processing by 2.76x to 3.2x. This is not a marginal improvement; it changes how developers interact with large models, enabling faster iteration on long-context inputs and more responsive tool-integrated workflows.

The 128GB unified memory architecture unlocks model access that the Intel platform cannot match. GPT-OSS 120B and LLaMA 3.3 70B are unavailable to the Intel HP Z2 Mini G1i entirely. This is a capability gap, not a performance one: Intel users running SFF workstations in this class must choose between running smaller models locally or paying for cloud inference on the models that actually meet production quality thresholds. The DGX Spark closes that gap.

On specialized generative AI workloads, the DGX Spark's advantages vary by format. In BF16 image generation, it leads by 1.3x to 2x. In video generation (Wan 2.2 14B), its 7.6x lead over AMD and 1.5x lead over Intel reflect the GPU's advantage on large, memory-intensive diffusion workloads. The Intel RTX 4000 Ada discrete GPU holds a narrow edge in FP16/FP8 image generation (Stable Diffusion, Flux1.Dev), the one area where discrete GPU VRAM bandwidth produces a measurable advantage. However, the DGX Spark is the only platform that can run all four tested image generation workloads with competitive or leading performance across the full set.

AI Developer Focus

This section examines the DGX Spark not as a benchmark subject but as a developer platform, one that changes what kinds of AI development workflows are possible, economically viable, and practical on local hardware. Findings are drawn from Signal65 hands-on use of the platform and from a quantitative analysis of developer TCO.

Running Large Models at the Edge

The ability to run a 120B parameter model locally (on a device that fits on a desk, under a monitor and costs approximately $4,500) is not a capability that existed in many options in the SFF workstation market before the DGX Spark. The Intel-based HP Z2 Mini G1i cannot run GPT-OSS 120B at all. For any developer whose workflow depends on frontier-scale reasoning quality, the only historical alternative to the DGX Spark has been a cloud API. Running GPT-OSS 120B locally on the DGX Spark eliminates that dependency entirely.

Local model execution means that model weights, prompts, and outputs never leave the device. For developers working with proprietary codebases, customer data under processing agreements, or sensitive internal documents, this is not a convenience feature; it is a compliance and security requirement. The DGX Spark satisfies data residency constraints that cloud-hosted model APIs fundamentally cannot, regardless of the cloud provider's contractual terms.

Offline development capability is a practical productivity benefit that becomes apparent quickly in daily use. The DGX Spark's model execution requires no network connectivity once the model weights are cached locally. Developers in facilities with restricted internet access or simply working through connectivity interruptions maintain full access to their local model server. Cloud-dependent workflows have no equivalent fallback.

The economic dimension of local model execution is direct: every query to GPT-OSS 120B or LLaMA 3.3 70B running on the DGX Spark carries zero marginal API cost. Cloud-hosted frontier models are priced on token throughput; high-frequency development usage (iterating on prompts, running eval suites, processing large codebases) can generate cloud costs that compound quickly. On the DGX Spark, those queries are free at the margin after the hardware purchase. For developers with high query volumes, the cost structure shifts from variable to fixed.

Network round-trip latency to cloud inference endpoints (typically 100 to 500ms or more depending on the provider and request complexity) is eliminated on local hardware. For interactive applications where time-to-first-token is user-facing, this architectural difference can be perceivable. The DGX Spark's TTFT on GPT-OSS 120B in our testing (measured locally) reflects purely computational latency; there is no network overhead to subtract.

Cloud-to-Local Migration: TCO Analysis

The DGX Spark's approximately $4,500 purchase price reaches capital return in roughly one calendar month when compared to the cost of equivalent cloud-hosted GPU inference. An H100 instance capable of running 120B-parameter models at comparable inference throughput costs approximately $6.00 per hour at current cloud market rates. At 24/7 utilization (the mode in which a developer workstation is effectively available even when idle), that rate translates to $1,008 per week and approximately $4,380 per calendar month (730-hour standard billing cycle). **At these numbers, the DGX Spark reaches breakeven on its purchase price in approximately 31 days of displaced cloud usage.**

Beyond raw payback period, the economic model of cloud GPU usage imposes a documented behavioral constraint on developers: per-token or per-hour pricing can create hesitation around complex, exploratory, or high-iteration queries. Developers running on cloud credits self-censor their use of AI tools in ways that reduce the quality and frequency of their model interactions. The DGX Spark eliminates this "metered anxiety" by changing the cost structure to fixed. A developer with an idea at 1:00 AM can run a GPT-OSS 120B reasoning task immediately, at zero incremental cost, without waiting for the cloud console to provision an instance or accumulating charges on a corporate card.

The shift from cloud compute to owned hardware is also a shift from utility to asset. A cloud GPU hour is consumed and gone; a DGX Spark purchase represents a capital asset that depreciates over its useful life (typically three to five years) while delivering compute capacity in every one of those hours. For finance purposes, this distinction affects how AI development costs are classified, budgeted, and amortized. Organizations moving AI development spend from cloud opex to hardware capex may find the DGX Spark's economics compelling beyond the simple monthly breakeven comparison.

This "Sovereignty of Compute" framing extends beyond economics. Having 128GB of locally accessible GPU memory, available at any hour, with no queue, no provision delay, and no network dependency, changes how developers scope their work. Queries that would be reserved for "when it's worth the API cost" become routine. Models that would require a cloud reservation become immediately available. The DGX Spark does not just reduce the cost of AI development; it removes the friction that causes developers to use AI tools less than they otherwise would.

NVIDIA-Validated Platform Advantage

One of the DGX Spark's most significant attributes is its software ecosystem continuity with NVIDIA cloud infrastructure. The platform runs the same CUDA toolkit, NeMo framework, and serving stack toolchain that operate on H200 and B200 data center GPUs. Containers built and tested on the DGX Spark are guaranteed to run (without modification) on NVIDIA cloud infrastructure. This is not a compatibility claim; it is an architectural advantage derived from the shared GB/Blackwell architecture and NVIDIA unified software stack.

Signal65 demonstrated this continuity directly: we ran a vLLM-based GPT-OSS 120B chatbot inference workload on the DGX Spark, measured baseline performance, and then migrated the

identical workload to an H200 GPU in our Signal65 AI Lab environment. The migration required less than 15 minutes and no software modifications. The same container, the same model format, the same serving configuration, transferred directly between local and cloud hardware.

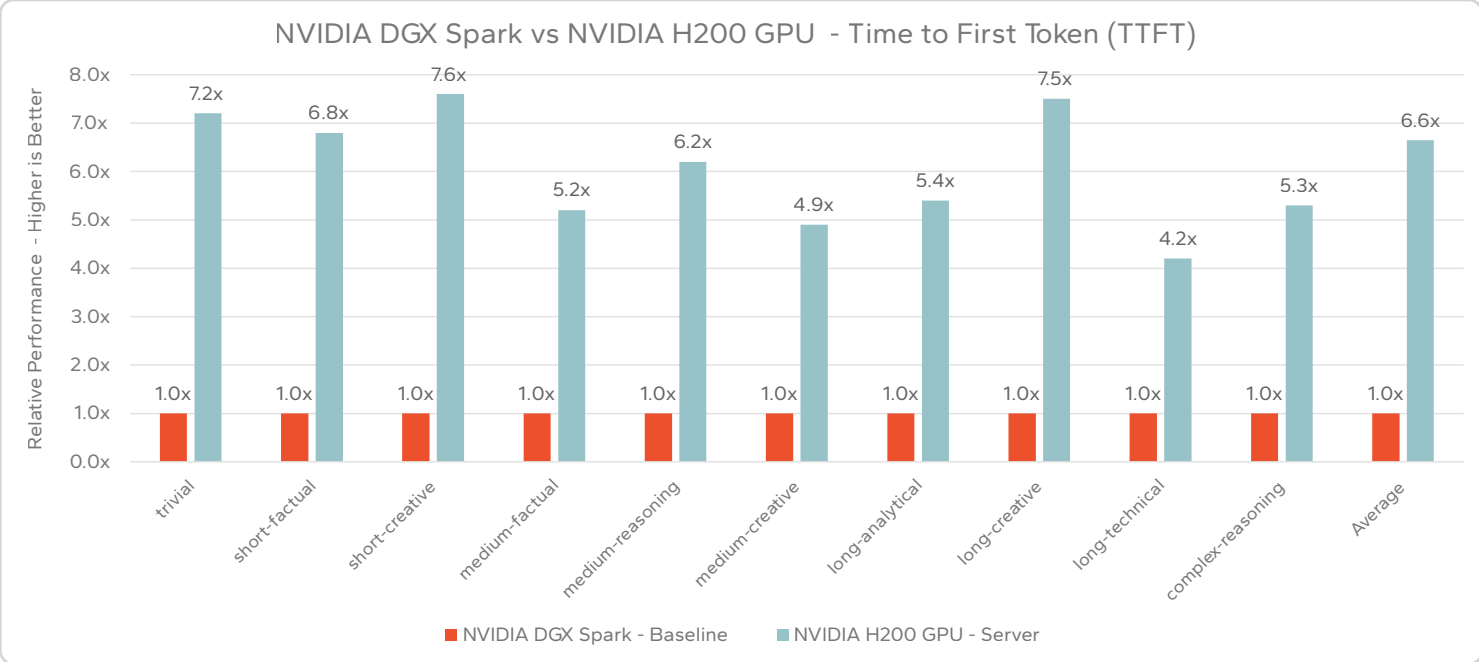


Figure 15: *TTFT Comparison*

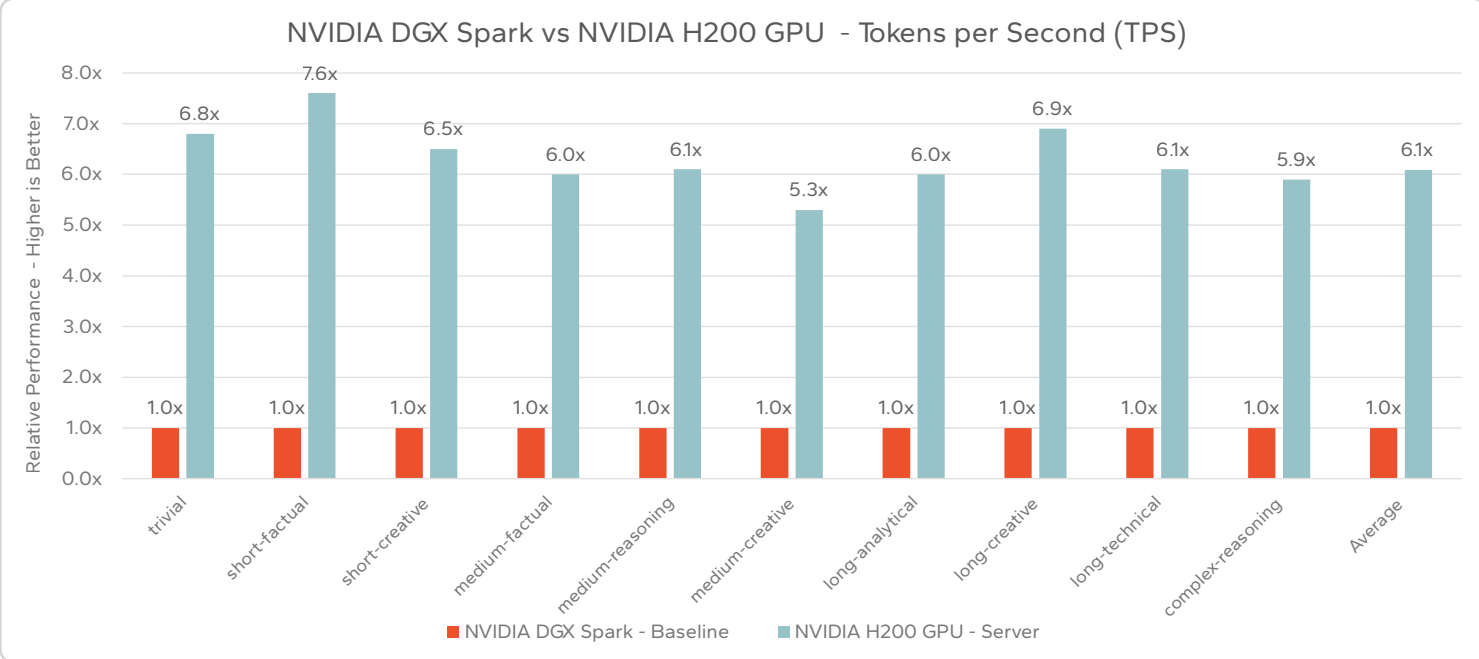


Figure 16: *Tokens Per Second Comparison*

In our testing, the H200 posted approximately 6x faster TTFT and approximately 6x higher token generation throughput than the DGX Spark on identical GPT-OSS 120B inference configurations. This gap is expected and architectural: the H200 is a $30,000+ data center accelerator with vastly higher HBM3 bandwidth and compute throughput. The relevant insight is not the magnitude of the gap but

the frictionlessness of the transition. **The DGX Spark functions as a local prototype and validation environment for workloads destined for H200 or B200 clusters, a local on-ramp to the Blackwell cloud ecosystem.**

Developer velocity is the compound benefit of this ecosystem. Cold-start delays on cloud instances, environment inconsistencies between local and cloud containers, and network round-trips during the iteration phase of development are all eliminated when the local development environment is architecturally aligned to the deployment target. Developers prototype and debug on the DGX Spark's 128GB unified memory (where they have full, instant, zero-cost access), then push validated workloads to cloud infrastructure when they are ready to scale. This train-local-deploy-cloud workflow reduces the risk of environment-driven bugs at deployment and allows developers to defer cloud spend until the workload is genuinely ready for production scale.

Section Key Insights

The DGX Spark's developer value proposition operates on multiple simultaneous dimensions: capability (run models that most competing SFF platform can execute), economics (one-month cloud breakeven, zero marginal query cost), data governance (model weights and data remain on-device), and ecosystem (zero-friction migration to NVIDIA cloud infrastructure). No competing SFF workstation in this evaluation combines all four of these dimensions in a single platform.

The NVIDIA software ecosystem continuity is the most underappreciated dimension of the DGX Spark's value. The ability to prototype locally on hardware that is architecturally continuous with H200 and B200 cloud infrastructure is a workflow capability that has no equivalent on AMD or Intel SFF platforms. An AMD Strix Halo-based workstation and an H200 cloud GPU occupy different software ecosystems; an AMD-based developer environment cannot guarantee that local workloads will migrate without modification to NVIDIA cloud infrastructure. The DGX Spark closes this gap by design.

As the NVIDIA Blackwell cloud ecosystem grows (more models optimized for NVFP4, more NeMo workflows, more serving configurations), the DGX Spark's value as a local Blackwell-compatible development platform grows proportionally. The platform is an investment in a software ecosystem with a clear, heavily capitalized roadmap, not just in a hardware configuration.

Power Efficiency Analysis

Power efficiency is an increasingly important dimension of workstation evaluation, both for total cost of ownership (electricity costs over the system lifecycle) and for thermal and deployment constraints in office and home environments. In this section, we measure GPU subsystem power draw under AI inference workloads and total wall power for the full platform, comparing the DGX Spark against the AMD and Intel competing systems.

GPU Power Consumption

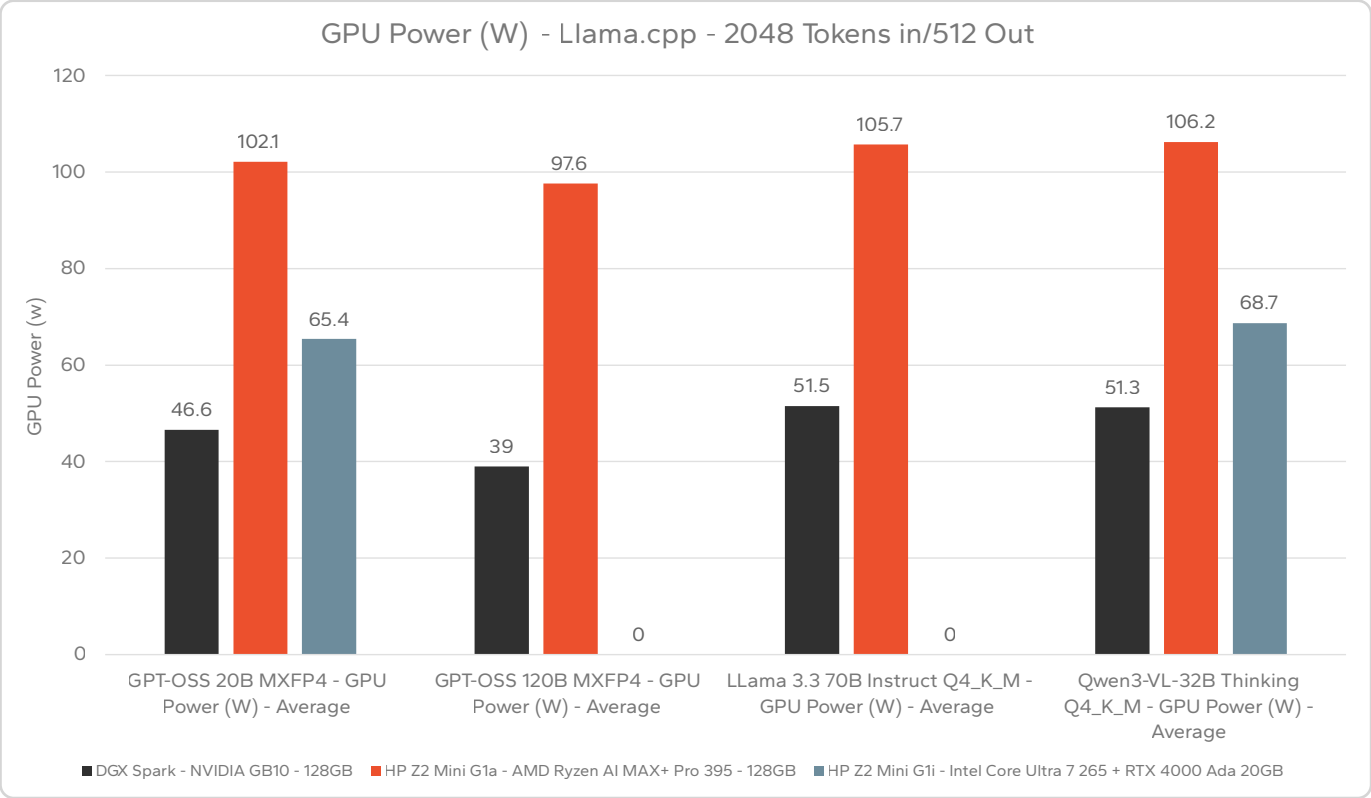


Figure 17: *GPU Power Consumption Under AI Inference*

In our testing, the DGX Spark's GPU subsystem drew 46.6W during GPT-OSS 20B inference, compared to 102.1W for the AMD system and 65.4W for the Intel RTX 4000 Ada. This means the DGX Spark delivered the fastest prompt processing of all three platforms while consuming approximately 55% less GPU power than the AMD system and approximately 29% less than Intel. On GPT-OSS 120B (the two-platform comparison), the DGX Spark drew 39.0W versus AMD 97.6W, the GB10 GPU drawing less power on the larger model than on the 20B model, indicating efficient memory access patterns at this scale.

LLaMA 3.3 70B and Qwen3-VL-32B showed DGX Spark GPU power in the 51 to 52W range versus AMD 105 to 106W, a consistent 2x GPU power efficiency advantage across the full tested AI inference portfolio.

Wall Power Consumption

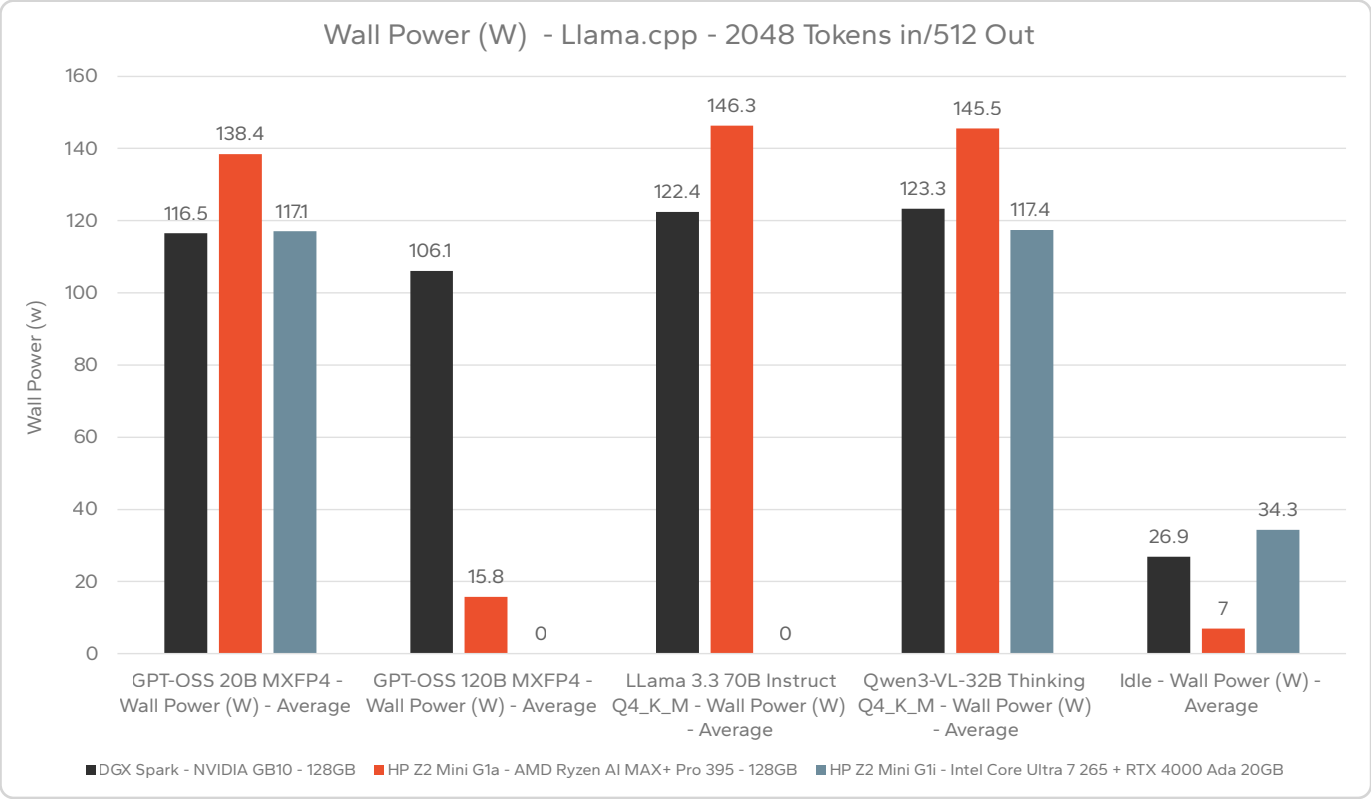


Figure 18: *Total System (Wall) Power Consumption*

In our testing, total wall power for the DGX Spark under AI inference ranged from 106.1W to 123.3W across the tested workloads. The AMD system drew 138.4 to 146.3W on workloads where both platforms were active. This represents a 17 to 20% total platform power advantage for the DGX Spark, meaningful for deployments where power budget or thermal management constraints apply. The Intel system's wall power under the two workloads it could run (GPT-OSS 20B and Qwen3-VL-32B) was 117.1W and 117.4W respectively, comparable to the DGX Spark's range despite the Intel system delivering substantially lower AI performance.

At idle, the DGX Spark draws 26.9W, the AMD system draws 7.0W, and the Intel system draws 34.3W. For deployments running the DGX Spark as a shared inference server (a use case directly supported by the concurrency results in an earlier section), the active power figures are more relevant than idle power, and the DGX Spark's efficiency advantage under load is clear and consistent.

Conclusions

The NVIDIA DGX Spark, built on the Arm-based NVIDIA GB10 Grace Blackwell Superchip, establishes a credible new category in the SFF workstation market: a local AI supercomputer that matches or exceeds x86 competition across traditional CPU workloads while simultaneously delivering AI capabilities that no competing x86 SFF platform can match.

Across the full CPU benchmark suite, the DGX Spark's Arm-based CPU complex wins the majority of tested workloads, with strongest advantages in C-Ray ray tracing (30 to 41% over AMD and Intel), RAMspeed memory bandwidth (25 to 50% over both platforms), LULESH hydrodynamics simulation (30%+ over AMD), and productivity workflows (LibreOffice, NGSpice). x86 platforms retain advantages in workloads with heavily optimized code paths and workloads that scale primarily with raw thread count. These are expected gaps that reflect software optimization history rather than fundamental architectural limitations.

For a developer or engineer using the DGX Spark as a daily-use workstation, the CPU performance is not a compromise; it is competitive with and frequently superior to the best available x86 SFF alternatives. The Arm-based GB10 achieves this while drawing up to 20% less total system

power than the x86 platforms under AI inference workloads. Its LPDDR5-8533 unified memory subsystem, delivering up to 50% higher bandwidth than either x86 configuration, is the foundation for the DGX Spark's advantages across memory-sensitive workloads.

The DGX Spark's unified 128GB memory architecture and GB10 GPU deliver AI capabilities that establish a capability boundary over competing SFF platforms. The Intel HP Z2 Mini G1i cannot run GPT-OSS 120B or LLaMA 3.3 70B at all. Against the AMD system (the only comparable configuration that also has 128GB), the DGX Spark delivers 2.7x to 3.2x faster prompt processing across all tested LLM workloads, leads by 1.3x to 2x on BF16 image generation, processes video generation 7.6x faster, and is capable of faster local model fine-tuning on 7B to 8B parameter models.

For team and shared server use cases, the DGX Spark's multi-user concurrency results are decisive: 3x to 7x faster TTFT and approximately 4x higher throughput versus the AMD Strix Halo platform across both light and heavy workload scenarios, with graceful latency scaling from 1 to 8 concurrent developers. This positions the DGX Spark not just as a single-user developer machine but as a viable shared local inference server for small teams.

The NVIDIA software ecosystem continuity between the DGX Spark and NVIDIA cloud infrastructure (H200, B200, and beyond) is a strategic platform advantage in the x86 SFF competitive set. Through Signal65 direct validation: migrating a GPT-OSS 120B workload from DGX Spark to H200 in under 15 minutes with no software changes, demonstrates that the DGX Spark functions as a genuine local prototype environment for cloud-scale AI workloads. This workflow continuity, combined with the 31-day cloud-equivalent breakeven and zero marginal query cost economics, makes the DGX Spark a financially and technically compelling alternative to cloud-only AI development infrastructure.

Across all tested dimensions (CPU compute, AI inference, generative AI workloads, multi-user concurrency, developer platform value, and power efficiency), the NVIDIA DGX Spark outperforms or differentiates from the leading x86 SFF alternatives in its price class. For developers and engineering teams evaluating a local AI compute investment, the DGX Spark's combination of Arm CPU competitiveness, GPU AI performance, 128GB memory capacity, NVIDIA software ecosystem continuity, and power efficiency represents a uniquely capable platform that the current x86 SFF market cannot replicate.

System Configurations

	NVIDIA DGX Spark	HP Z2 Mini G1a	HP Z2 Mini G1i
CPU	NVIDIA GB10	AMD Ryzen AI Max+ PRO 395	Intel Core Ultyra 7 265
Graphics	NVIDIA GB10	AMD Radeon 8060S	NVIDIA RTX 4000 Ada SFF 20GB
RAM	128GB LPDDR5X-8533	128GB LPDDR5X-8000	64GB DDR5-6400
Storage	4TB Samsung MZALC4T0HBL1-00B07	2TB Samsung MZVL22T0HBLB-00BH1	1TB Samsung MZVL21T0HCLR-00BH1
System BIOS	5.36_0ACUM018	01.04.00	1.07.01
Operating System	DGX OS (Ubuntu 24.04.4)	Fedora 43 (Workstation Edition)	Fedora 43 (Workstation Edition)
Kernel Version	6.17.0	6.18.12	6.18.12

Software Tested

llama.cpp (b7723)

vLLM 0.16 with ROCm 7.3.53390 (Z2 Mini G1a)

vLLM 0.15.2 with CUDA 13.0.r13 (DGX Spark and HP Z2 Mini G1i)

ComfyUI 0.13.0

NVIDIA NeMo-Automodel 0.2.0

HandBrake 20260129081203-5305d441f-master

C-Ray 2.0.0

AOBench 20180207

Dolfyn 0.527

HMMer Search 3.4

LULESH 2.0.3

RAMSpeed SMP 3.5.0

LibreOffice 25.8.4.2

NGSpice 45.2

Phoronix Test Suite v10.8.6

Models Used

GPT-OSS 20B (MXFP4)

GPT-OSS 120B (MXFP4)

LLama 3.3 70B Instruct - Q4_K_M

Qwen3VL 32B Thinking - Q4_K_M

Qwen3 Coder 30B A3B Instruct (AQW 4-bit)

Stable Diffusion 1.5 (FP16)

Flux1.Dev (FP8)

Z-image Turbo (BF16)

Z-image (BF16)

Wan 2.2 14B - Text to Video (FP8)

Important Information About this Report

CONTRIBUTORS

Ryan Shrout
President and GM | Signal65

Ken Addison
Client Performance Director | Signal65

INQUIRIES
Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS
This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

DISCLOSURES
Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT SIGNAL65
Signal65 is an independent research, analysis, and advisory firm, focused on digital innovation and market-disrupting technologies and trends. Every day our analysts, researchers, and advisors help business leaders from around the world anticipate tectonic shifts in their industries and leverage disruptive innovation to either gain or maintain a competitive advantage in their markets.

CONTACT INFORMATION
Signal65 I signal65.com