NVIDIA — Datacenter GPU Performance Study (2006–2025)

A MIT FutureTech analysis of how AI accelerators outran Moore's Law, and what's narrowing the lead

Explainer · June 3, 2026

An academic study measuring two decades of NVIDIA datacenter GPU performance, tracing how AI-specialized hardware drove compute gains far beyond transistor scaling.

COMPANY	NVIDIA (subject of the study)	HEADQUARTERS	Santa Clara, California, USA
OWNERSHIP	Public	TICKER	NASDAQ: NVDA
CEO	Jensen Huang		

ABSTRACT

This paper analyzes 102 NVIDIA datacenter GPUs spanning Tesla (2006) to Blackwell Ultra (2025), authored by MIT FutureTech and CSAIL researchers (Del Sozzo, Fleming, Flamm, Thompson). Its central finding is that AI-relevant compute has grown dramatically faster than Moore's Law: FP16 (half-precision) dense performance now doubles roughly every 1.43 years versus the ~2-year cadence of classic transistor scaling. The mechanism is architectural, not lithographic — specialized Tensor Cores and high-bandwidth memory (HBM) did the heavy lifting, while double-precision (FP64) math was deliberately deprioritized as workloads shifted to lower-precision AI arithmetic. The study also quantifies how US export controls engineered a 23.6× peak-performance gap between top chips and export-compliant ones, and pushes back on louder industry claims about the true pace of progress. The implications: memory, power, and price are increasingly the binding constraints; NVIDIA's dominance persists even as AMD overtakes it on two of three precision metrics; and the "rate of AI hardware progress" is itself contested.

Keywords: NVIDIA GPUs; Moore's Law; Tensor Cores; HBM; FP16; export controls; AI compute; performance scaling

1. Snapshot

This is a secondary academic synthesis rather than a vendor benchmark: MIT FutureTech and CSAIL researchers (Del Sozzo, Fleming, Flamm, and Thompson) assembled and regressed a dataset of NVIDIA datacenter GPUs to characterize how performance, memory, price, and power have evolved over nearly two decades. The corpus covers 102 GPUs from the Tesla generation (2006) through Blackwell Ultra (2025). The headline figures rest on exponential-fit regressions across that population, separating "all datacenter GPUs" from "top-performing GPUs" at each point in time. NVIDIA itself is the subject, not the author; the work reads as an independent measurement and recompute of the company's trajectory and of competing claims about it. What the paper does not contain is real-world benchmark data — every headline rate uses theoretical peak performance.

2. Thesis: What the Paper Argues

The core argument is that NVIDIA datacenter GPU performance grew far faster than Moore's Law from the mid-2000s to 2025, and that the engine of that growth was AI-oriented design rather than raw transistor scaling. Tensor Cores and HBM, not smaller transistors, carried the curve. The paper then layers in two complications. First, the raw-performance lead over competitors is narrowing without yet compelling a market shift. Second, US export controls have reshaped *who* can access top performance far more than they have changed the underlying pace of progress.

What makes this timely is the gap between marketing and measurement. As AI compute demand explodes, the rate at which accelerators improve is a load-bearing input to everything from datacenter buildout to national policy, yet that rate is asserted very differently depending on who is counting. This study is an attempt to pin the number down empirically and to flag where it remains genuinely uncertain. Treat the result as a measurement baseline for 2006–2025, not an extrapolation engine; the authors themselves are doubtful that the trends should be projected forward.

3. The Core Idea in Plain English

For decades, chips got faster mostly because transistors got smaller and more numerous, a roughly two-year doubling known as Moore's Law. This paper shows that NVIDIA's AI accelerators broke away from that cadence by changing *what kind of math the chip is built to do*, not just how many transistors it packs.

Think of a GPU as a factory floor. NVIDIA rebuilt the floor around a new machine optimized specifically for the arithmetic neural networks need — that machine is the Tensor Core, purpose-built for low-precision matrix math — and then widened the loading docks with HBM so the machine never starves for data. The old FP64 factory is now a legacy wing that gets proportionally less investment each generation. The new low-precision factory is where the headline numbers come from, and those numbers are real, but they measure a different thing than "GPU performance" meant a decade ago.

4. The Technical Space

The yardstick for accelerator progress is not a single number but a small set of dimensions that have to move together. The ones that matter here are compute throughput at a given numerical precision, memory size and bandwidth to feed that compute, power draw, and price. A chip that doubles its math throughput but cannot feed data fast enough, or cannot be powered and cooled, delivers little realized gain.

Precision is central. FP64 (double-precision) is the gold standard for scientific simulation; FP32 (single-precision) is the traditional graphics and general-compute workhorse; FP16 (half-precision) and lower are sufficient for most neural-network training and inference. The strategic insight of the modern era is that AI tolerates low precision, so hardware can trade exactness for raw throughput. Sparsity, the technique of skipping zero-valued computations, can lift effective throughput further without adding silicon.

"Good" therefore means improving the precision tiers that matter for AI faster than the rest, while keeping memory, power, and cost from becoming bottlenecks. The paper's contribution is to measure exactly how unevenly those dimensions have advanced, and to attribute the compute gains to specific architectural inflection points rather than to generic process improvement.

5. How the Performance Was Achieved (and What Drove It)

The study decomposes the gains into compute, memory, and the architectural events behind both.

- **Compute outran Moore's Law.** Across all datacenter GPUs, FP16 dense performance doubles every 1.43 years (1.27 with sparsity exploited), and FP32 every 1.67 years. For top-performing GPUs, FP16 compute grows at a 68.7% CAGR, a 1.32-year doubling. By contrast, FP64 doubling times stretch to

between 2.05 and 3.79 years.

- **Specialized hardware, not transistors, did it.** The architectural inflection points are concrete: FP64 hardware support arrived in 2008 with the Tesla GT200, FP16 CUDA-core support with Pascal in 2016, and Tensor Cores debuted with Volta in 2017. These purpose-built units, not lithographic shrink, explain the steep AI-precision curves.
- **HBM was a structural accelerator.** Starting with Pascal in 2016, HBM alone yielded a 31.8% CAGR for memory size and 26.8% for bandwidth, roughly 13.7 and 14.1 percentage points higher than GDDR alone. Without that memory leap, the compute gains would have starved.
- **FP64 was deliberately deprioritized.** As workloads shifted to lower-precision AI math, double-precision units became, in the authors' framing, a lower-priority resource. This is a design choice with downstream competitive consequences.

The proprietary edge here is genuine but bounded: the architectural sequencing and the HBM integration are hard engineering, yet the same techniques are available to competitors. The durable advantage lives less in the silicon than in the surrounding ecosystem, addressed below.

6. The Divergence of Dimensions

The most consequential finding for anyone planning datacenter capacity is that compute growth has vastly outstripped every other dimension. Memory size and bandwidth double only every 3.29 to 3.41 years, release prices roughly every 5.03 years, and power consumption roughly every 15 years. The divergence compounds: by the time FP16 compute has doubled four times (about 5.7 years), memory bandwidth has not yet doubled twice. That gap is the structural setup for a bottleneck.

Power is becoming the load-bearing constraint. Top-GPU TDP rose from 170.9W on the Tesla C870 to 1100W on Blackwell Ultra B300, and GPU power demand now represents approximately 40% of total power usage in AI datacenters. Performance-per-watt did improve, by 1.3× to 2.9× more than performance-per-dollar, but flagship chips now draw roughly 1.1 kW each. The trajectory points toward energy, not transistors, as the limiting input. One caution: the ~40% figure reads as an externally cited statistic rather than an output of the 102-GPU regression, so it should be treated as context, not as a measured result of this study.

7. Competition and the Export-Control Gap

NVIDIA's dominance is intact while its raw-performance lead erodes. The study reports NVIDIA holding over 80% market share in desktop GPUs and over 90% in datacenter GPUs.

The lead is narrowing, not collapsing. As of 2025, AMD has actually surpassed NVIDIA in FP16 by 1.12× and offers 62.91× higher FP64, the latter a direct consequence of NVIDIA's choice to deprioritize double precision. NVIDIA retains a 7.15× FP32 advantage. The striking implication is that a competitor now leads on two of three precision metrics, yet market share has not moved. CUDA, NVIDIA's software ecosystem, is often credited with the stickiness, but that is an attribution rather than a measured result in this dataset; the paper documents the share and the gaps more firmly than it isolates the locking mechanism.

Export controls are the other axis of competition. Using TPP (total processing performance, defined as 2 × MAC_TOPS × bitwidth), the study finds US export controls created a potential peak-performance gap of 23.6× under the 2025 regulations. The regulatory anchors are concrete: the October 2022 rule (ECCN 3A090) gated chips at aggregate bidirectional I/O bandwidth at or above 600 GB/s and TPP at or above 4800. A recent proposed change allowing H200 export could shrink the gap to 3.54×, though that outcome is conditional, not realized. The paper also notes there is little current or historical evidence that such engineered gaps endure, citing domestic substitution and chip-aggregation workarounds, so it presents both the gap and the doubt without reconciling them.

8. Caveats and Open Questions

The most important interpretive caveats, framed as the questions a careful reader should press:

- **Are the headline rates real-world or theoretical?** Every figure uses theoretical peak performance; the paper notes boost clocks can only be sustained briefly, and no benchmarks are analyzed, so realized gains may be lower.
- **How reliable is the newest data?** Blackwell and Blackwell Ultra values were partly inferred because NVIDIA's public documentation is no longer exhaustive, and shipped boards differ from fabricated-chip specs.
- **Is the power story precise?** Findings rest on an exponential-fit regression, and the TDP-based power figures carry approximation error that the clean numbers can obscure.
- **Will the export gap persist?** The 23.6× gap (or 3.54× if H200 export proceeds) is a regulatory artifact, and the paper itself doubts such gaps endure given domestic substitution and chip aggregation.
- **Whose "rate of progress" is correct?** The authors recompute Jensen Huang's COMPUTEX 2024 claim of a 1000× gain over 8 years as a 138% CAGR (0.8-year doubling), 65 percentage points above the paper's best identified rate. Meanwhile Epoch AI finds FP16 progress 26 points lower (36% versus 62%) and FP32 23 points lower (28% versus 51%). The true rate depends on definitional choices the field has not standardized.

9. Bottom Line

Three takeaways. First, AI accelerator performance genuinely broke from Moore's Law, and the cause is architectural specialization (Tensor Cores, HBM) rather than transistor scaling. Second, the single biggest tension is divergence: compute has outrun memory, price, and especially power so sharply that energy is becoming the binding constraint, even as AMD overtakes NVIDIA on FP16 and FP64 without dislodging it. Third, the thing to watch is policy and definitions, the export-control gap (23.6×, or 3.54× if H200 export proceeds) and the unresolved dispute over what the true rate of progress even is.

10. For the Nerds

The deepest tension in the paper is definitional, and it determines whether any of the headline rates survive scrutiny. The TPP metric (2 × MAC_TOPS × bitwidth) is itself a regulatory construct, which means the 23.6× export gap is partly an artifact of how the rule chose to weight bitwidth, not purely a physical performance delta. That matters because sparsity, precision tier, and the theoretical-versus-sustained distinction each shift the CAGR by tens of percentage points; the spread between the authors' 62% FP16 figure, Epoch AI's 36%, and Huang's implied 138% is mostly a fight over those choices, not over the underlying hardware.

The sparsity figure deserves a specific caution. The 1.27-year FP16 doubling time (with sparsity) versus 1.43 years (without) assumes workloads can actually exploit structured sparsity, which depends on model-level zero patterns and compiler support that vary significantly by task. The with-sparsity rate is therefore a ceiling many production workloads cannot reach, making the 1.43-year figure the more conservative and probably more representative benchmark for real deployments.

A second open question is data integrity at the frontier. The divergence between fabricated-chip specs and shipped boards (GH100's 18432 FP32 cores versus the 16896 or 14592 on actual products) means peak figures for the newest parts are estimates, and the HBM die-area portions of the 2024 and 2025 regulations were largely set aside for lack of public data. The frontier is exactly where the measurement is weakest.

WHERE SOURCES DISAGREE

The "true" rate of AI-hardware progress

One reading (this paper) places FP16 growth near 62% CAGR. Another account (Epoch AI) finds it roughly 26 ppts lower (~36%), while NVIDIA's own "Huang's Law" framing implies ~138%

Unresolved: the figures rest on different definitional choices — theoretical peak versus benchmarks, inclusion of sparsity, and what counts as "AI compute" — so no single canonical rate is defensible

Whether the export-control performance gap will persist

The regulatory calculation shows a real engineered gap (23.6×, or 3.54× if the H200 allowance proceeds). Another reading within the same source argues there is "little current or historical evidence" such gaps endure, citing domestic substitution and chip-aggregation workarounds

Unresolved: this is a forward policy judgment, not a measured quantity, and the source presents both the gap and the doubt without reconciling them

Why the market hasn't shifted despite AMD leading on two of three precision metrics

One reading attributes NVIDIA's stickiness chiefly to the CUDA software ecosystem. Another notes that with a competitor already ahead in FP16 and far ahead in FP64, the hardware case for switching exists yet share has not moved — implying the cause is asserted more than demonstrated

Unresolved: the available evidence documents the market share and the metric gaps but does not isolate the mechanism keeping customers locked in

Theoretical peak versus real-world performance

All headline figures use theoretical peak values; the source notes boost clocks "can only be used for a limited time" and real benchmarks are not analyzed

Unresolved: the corpus contains no benchmark data to quantify the gap between peak and sustained throughput, so headline rates may overstate realized gains

Reliability of the newest-chip data

Values for Blackwell and Blackwell Ultra were partly inferred because NVIDIA's public documentation is "no longer as exhaustive," and shipped boards differ from fabricated-chip specs (e.g., GH100's 18432 FP32 cores vs. 16896/14592 on actual boards)

Unresolved: vendors do not publish the underlying design details, and the HBM die-area portions of the 2024/2025 regulations were "mainly ignored" for the same lack of public data