GPU SMITH
COMPUTE ENGINEERING

GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

B200 Confidential Computing Inference: Security and SLO Gate

> Plan and build private AI compute with GPU Smith. We connect workload requirements with GPU hardware, networking, deployment and operating decisions so your infrastructure fits the work it must perform.

Published by GPU Smith · 2026-11-16 · b200 confidential computing inference · nvidia confidential computing · blackwell confidential computing · gpu attestation · intel tdx · cc-on cc-off benchmark · confidential ai inference · tensorrt-llm · private llm inference · gpu procurement checklist

Original article: https://gpusmith.com/articles/en/b200-confidential-inference-security-performance-gate



In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes When a B200 Serves Inference in Confidential Mode
4. Implementation Considerations and Process Changes
5. Data Analysis and Evidence
6. Implications and Future Directions
7. Frequently Asked Questions (FAQs)
8. Conclusion

Executive Summary

A confidential **NVIDIA B200** inference deployment can meet a production service level objective (**SLO**), but the public evidence supports a gated, measured acceptance test rather than a single penalty figure. NVIDIA's engineering post, dated September 22, 2026, reports that on one **eight-GPU DGX B200** running **DeepSeek-R1-0528-NVFP4** under **TensorRT-LLM**, confidential computing (CC) on retained **96.1% to 98.2%** of CC-off output-token throughput across concurrency 1 to 16 (developer.nvidia.com). Mean time per output token (TPOT) stayed within **1.2% to 4.3%** of the baseline (developer.nvidia.com). NVIDIA tested one model, one framework build and one input shape (**32K** input tokens, 1K output tokens), and it chose that shape because long inputs, long outputs and low concurrency make CC cost easier to see (developer.nvidia.com). The result is a documented vendor measurement, not a purchasing guarantee.

A non-NVIDIA cross-check exists but needs scrutiny. A preprint posted to arXiv in August 2026 (identifier 2608.26575) measured **paired** CC-on and CC-off runs on a single host with eight B200 GPUs under Intel Trust Domain Extensions (TDX) (arxiv.org). It reports low single-digit overhead for correctly configured inference, and **30% to 40%** penalties on stock inference stacks (arxiv.org). Its arXiv listing shows no journal reference, and the authors themselves put run-to-run noise at roughly plus or minus 2 percentage points (arxiv.org). The two sources agree on mechanism more than on a number: bounce-buffer copies, kernel-launch cost and collective-communication changes decide the outcome (arxiv.org).

The security boundary is wider than the GPU. NVIDIA's documentation states that a successful GPU **attestation** does not cover the CPU trusted execution environment, and that CPU evidence alone does not establish GPU state (docs.nvidia.com). Attestation also does not prove that approved code is free of vulnerabilities (docs.nvidia.com), and it supplies evidence for a trust decision without replacing the policy that makes the decision (docs.nvidia.com). Encryption alone is therefore not a compliance finding; the **GDPR** text asks for measures that are regularly tested and evaluated for effectiveness (eur-lex.europa.eu).

The recommended gate has four parts. First, pin the stack: driver, firmware, CUDA, framework image, kernel and **TDX module**. As of November 2026, NVIDIA's **R595** release pairs **CUDA 13.2** with **driver 595.58.03** (docs.nvidia.com). Second, run paired CC-on and CC-off tests at the traffic shape that will actually be served, computing **throughput retained** as 100 times CC-on tokens per second divided by CC-off tokens per second. Third, convert measured retention and SLO behavior into GPU headroom: at NVIDIA's lowest retention of **96.1%** the **capacity multiplier** is about **1.04** (the lowest of five tested concurrency points on one model, not a lower bound), but a tight TPOT SLO can bind earlier than throughput does. Fourth, preserve attestation evidence and a rollback path. Vendor percentages should inform the test plan, not replace it.

Introduction and Background

Regulated-industry platform teams increasingly ask a narrow question: after confidential computing is switched on for a B200 inference service, does the private LLM inference service still meet its latency and throughput targets, and can the security claim be evidenced? This report examines that question for the NVIDIA **Blackwell B200** generation as of November 2026. It does not survey private-LLM architectures in general, and it is not a hardware compatibility matrix; NVIDIA publishes its own matrix and release notes, and those documents change with each driver release (www.nvidia.com).

Several terms recur in the sections that follow:

- **Trusted execution environment (TEE):** a hardware-isolated region whose contents the host cannot read. The Confidential Computing Consortium defines confidential computing as protection of data in use by performing computation in a hardware-based, attested TEE (confidentialcomputing.io).
- **Intel Trust Domain Extensions (TDX):** hardware-isolated virtual machines called trust domains (www.intel.com). A confidential virtual machine (CVM) is such a guest.
- **TPOT and TTFT:** TPOT is the average time to generate each output token after the first, and time to first token (TTFT) covers queuing, prefill and network latency (nvidia.github.io) (docs.nvidia.com).
- **NVLS:** NVLink SHARP, a feature that offloads collective operations to the NVSwitch domain.
- **Retention and overhead:** throughput retained is CC-on output tokens per second as a percentage of CC-off, and **TPOT overhead** is the percentage by which CC-on TPOT exceeds CC-off.

The commercial context is procurement. The decision at stake is whether a regulated workload can run on shared or third-party infrastructure with attestable isolation. NVIDIA also documents production use of CC as a licensed feature (docs.nvidia.com), which is a line item buyers should confirm in writing.

This report is published by GPU Smith, an independent private AI infrastructure engineering consultancy that advises on GPU procurement, workload sizing, networking, inference optimization and air-gapped deployments. GPU Smith does not manufacture GPUs or operate a cloud, and it is not an option in any comparison here. Every benchmark figure below is attributed to its originator, with its configuration, and no result from a GPU Smith deployment is claimed.

The analytical distinction that organizes the report is between three kinds of evidence. Vendor measurements, such as NVIDIA's DGX B200 post, document a specific stack. Non-NVIDIA preprints, such as the arXiv B200 study (identifier 2608.26575), test other stacks; they are unreviewed, and that study reports its own paired measurements on a different stack. Deployment-specific measurements, which only the buyer can produce, are the only ones that carry an SLO decision. The sections that follow explain what changes in a confidential deployment, how to run a paired acceptance test, and how to convert the result into capacity and evidence.

Key Changes When a B200 Serves Inference in Confidential Mode

The security boundary: CPU CVM, GPU CC and encrypted NVLink

NVIDIA describes its CC stack as memory-encrypted CVMs, confidential GPUs and encrypted NVLink, and states that on **HGX B200** and **HGX B300** CC spans up to eight GPUs with NVLink encryption (developer.nvidia.com). The CPU side matters as much as the GPU side. Intel's attestation documentation states that a GPU on its own is not a complete TEE and relies on a CPU CVM (docs.trustauthority.intel.com). In a TDX guest, memory starts private and the hypervisor cannot read it (docs.kernel.org), while device traffic uses shared bounce buffers: the kernel documentation notes that TDX uses SWIOTLB (software I/O translation buffer) for most DMA allocations (docs.kernel.org).

The GPU memory protection model differs from the CPU's. NVIDIA's threat model says GPU memory encryption is not the protection mechanism, and that each GPU is assigned whole to one CVM. Several features are outside the documented configurations: multi-instance GPU, multi-process service and GPUDirect

RDMA are not supported there. The non-NVIDIA preprint reports the same RDMA limit for the stack it tested, which matters for any plan that depends on cross-node GPU networking.

Availability is a different question from confidentiality. The Consortium's technical analysis treats availability attacks as not addressed by current hardware TEEs (confidentialcomputing.io), so an infrastructure owner who can stop the CVM, interrupt networking, detach devices or remove power remains outside what attestation addresses. A security architect should therefore write the threat model as confidentiality and integrity of data in use, not uptime.

Blackwell and **Hopper** differ in the multi-GPU case. NVIDIA's VM documentation lists multi-GPU CC mode on Blackwell and a separate protected PCIe mode for multi-GPU Hopper (docs.nvidia.com). That is the main structural answer to the **H100**-versus-B200 question: the Hopper multi-GPU mode is a different design, and the Hopper overhead studies cited in the data section cover a single GPU.

Supported SKU, firmware and driver stack

Table 1 below lists the layers that must be pinned and recorded. The values are the ones NVIDIA documents or used in its test; a buyer must confirm the exact combination against NVIDIA's current compatibility matrix, because NVIDIA states that it supports only the combinations in its tables.

Layer	Documented requirement or value used in vendor test	Why it belongs in the acceptance record
CPU TEE	**Intel TDX**; NVIDIA lists Emerald Rapids and Granite Rapids (docs.nvidia.com). TDX module 1.x for Emerald Rapids, 2.x for Granite Rapids (docs.nvidia.com).	The CPU quote is half of the attestation evidence.
Host OS and kernel	R595 notes list an Intel host of **Ubuntu 25.10** with kernel 6.17.0 or later (docs.nvidia.com). Canonical documents TDX host support from Ubuntu 25.10 (ubuntu.com).	Kernel and TDX module versions change copy and scheduling behavior; record them.
Guest OS	**Ubuntu 24.04** LTS guest support (ubuntu.com). Google recommends version 580 or later of the NVIDIA open kernel module for its confidential GPU VMs (docs.cloud.google.com).	The guest kernel carries the driver and Linux Kernel Crypto API (LKCA) dependency.
GPU and mode	HGX B200 8-GPU SXM6 180GB listed for multi-GPU passthrough (docs.nvidia.com); up to eight GPUs per CVM. Canonical's TDX repository documents only full-GPU passthrough into a trust domain (github.com).	Mode must match the tested topology; GPU mode is set only from the host (docs.nvidia.com).
VBIOS, driver, CUDA	R595: CUDA 13.2 with driver 595.58.03; B200 firmware family **FW 1.4.X** (docs.nvidia.com). NVIDIA's DeepSeek test used driver **595.71.05** and CUDA 13.2, a different driver build from the 595.58.03 named in the release notes, so confirm which build your supplier deploys.	Firmware and driver versions are policy inputs; a valid signature alone does not make a version acceptable.
Interconnect	**Fabric Manager** and a symmetric rail policy are required for multi-GPU CC on Blackwell. **NVLS multicast** is unavailable in B200 CC configurations (developer.nvidia.com).	Collective algorithm selection changes when NVLS disappears.
Framework	TensorRT-LLM container release **1.3.0rc22** with NCCL v2.30 in the vendor test (developer.nvidia.com).	Performance fixes are framework-version specific.

Layer	Documented requirement or value used in vendor test	Why it belongs in the acceptance record
Attestation tooling	Attestation SDK Local GPU Verifier 2.6.3 or later for multi-GPU passthrough (2.6.0 or later for single-GPU passthrough) in R595; CLI form documented as `nvattest attest --device gpu --verifier local` (docs.nvidia.com).	The verifier version determines which claims can be evaluated.

The matrix shows how many independent version axes a single confidential deployment has. It also exposes documentation drift: NVIDIA's R595 note names Ubuntu 25.10 for Intel hosts, while the v1.0.0 supported-platforms page lists a different host release for its B200 rows. Readers should cite the exact document and version in their own acceptance record rather than assuming one NVIDIA page supersedes another. The release history also shows a short maturity curve for B200: R575 offered single-GPU passthrough as early access, R580 reached general availability for single-GPU passthrough with a bounce buffer (docs.nvidia.com), and multi-GPU passthrough followed in R590 (docs.nvidia.com).

Attestation: what it evidences and what it does not

NVIDIA's attestation suite consists of the Remote Attestation Service (**NRAS**), the Reference Integrity Manifest (**RIM**) service and an online certificate status protocol (OCSP) service (docs.nvidia.com). NRAS issues signed Entity Attestation Tokens, and RIM bundles are signed by NVIDIA. In the vendor's description, a verifier compares signed GPU measurements with known-good reference measurements, and a confidential workload should undergo remote attestation before it receives any secrets (developer.nvidia.com). The general model is the one standardized in the IETF's RATS architecture, where a verifier appraises evidence and produces attestation results for a relying party (www.rfc-editor.org), and a composite device such as a CPU plus GPUs is appraised through its sub-entities (www.rfc-editor.org).

Three practical points follow:

- **Composite evidence needs both sides.** In TDX the CPU side has two steps, report generation and quote generation (docs.kernel.org). Intel documents a path in which **Intel Trust Authority** and NRAS act as two verifiers, and composite CPU plus GPU attestation as stronger than GPU-only attestation (docs.trustauthority.intel.com); one request can cover up to eight GPUs (docs.trustauthority.intel.com). Cloud platforms follow the same pattern: Azure documents secure key release bound to successful attestation (learn.microsoft.com) and configurable attestation policies (learn.microsoft.com).
- **Verifier availability is a dependency.** Intel's documentation warns that if NRAS is slow or unavailable an error occurs and no token is returned (docs.trustauthority.intel.com), and NVIDIA's own services collect the client IP address and GPU information (docs.nvidia.com), which matters for air-gapped designs and privacy review. NVIDIA also documents a local verifier option, but marks the Python SDK and stand-alone local verifier as deprecated in favor of the C++ SDK (github.com).
- **The claims must be consumed.** The GPU claims report debug-facility status and a nonce match for freshness (docs.nvidia.com), and NVIDIA states that **CC-DevTools** mode permits profiling and must be excluded from production policy (docs.nvidia.com).

Two statements bound the security claim. NVIDIA's overview says protection depends on the actual topology, trusted software and enforced policy. The Confidential Computing Consortium's compliance paper positions the technology as one part of a broader data-protection strategy (confidentialcomputing.io) that complements existing measures (confidentialcomputing.io). NIST's draft IR 8320E notes that its attestation prototype is only one possible way to implement the flow (nvlpubs.nist.gov). In GDPR terms, Article 32 lists encryption as one example measure (eur-lex.europa.eu) and separately requires a process for regularly testing and evaluating effectiveness (eur-lex.europa.eu). This report does not offer legal advice; a compliance determination belongs to the buyer's counsel and auditors.

NVIDIA also states that attestation adds no latency to individual inference requests once the workload is running, and describes the handshake as typically a one-time startup event (developer.nvidia.com). That is a vendor statement. The non-NVIDIA preprint explicitly did not measure attestation or confidential-VM startup overhead (arxiv.org), so no public source reviewed quantifies startup, re-attestation or relaunch cost for B200. The buyer must measure it.

Runtime path changes: copies, autotuning and collectives

NVIDIA's B200 post says secure execution changes assumptions about memory movement, timing, scheduling and multi-GPU communication. Four changes explain most observed behavior.

- **Copies.** Host-to-device transfers pass through a software-encrypted bounce buffer because the GPU cannot directly access protected CVM memory (developer.nvidia.com). Pinned memory loses its usual asynchronous-transfer advantage, and TensorRT-LLM now chooses pageable memory on affected paths (developer.nvidia.com). CUDA's runtime documentation explains why this matters: for transfers from pageable host memory, a stream sync is performed before the copy is initiated (docs.nvidia.com).
- **Token readback.** NVIDIA reports that CC can make many host-to-device and device-to-host copies effectively synchronous during cudaMemcpyAsync, so frameworks move token readback off the critical path (developer.nvidia.com). In the preprint, the unpatched equivalent was forced synchronous through the bounce buffers (arxiv.org).
- **Autotuner timing.** In the tested configuration, CUDA-event timestamps produced an unstable timing signal, so TensorRT-LLM measures tactics with the GPU global timer under CC (developer.nvidia.com). The preprint reports that event timing can return negative values on the bounce-buffer path (arxiv.org). A kernel tactic chosen under bad timing can silently lower throughput.
- **Collectives.** NVLS multicast is unavailable in B200 CC configurations, so frameworks should detect it and choose algorithms by message size and topology. The preprint reports multicast over NVSwitch is blocked under CC (arxiv.org) and measures NVLink encryption costing about 11% of Copy Engine bandwidth and about 18% of SM-based bandwidth.

A fifth mechanism is submission cost. NVIDIA states that secure work submission latency is often the larger factor for inference, and that smaller units of work are affected more (developer.nvidia.com). The preprint's reading is the same: its high serving penalties were launch overhead, not bandwidth (arxiv.org), and capturing the whole step in one CUDA graph made the extra submissions vanish in a synthetic probe. For a buyer, the consequence is that small-batch, short-output and prefill-heavy traffic can behave very differently from the long-context sample NVIDIA chose.

Implementation Considerations and Process Changes

Choose the traffic samples before choosing the benchmark

Two samples are needed. The first should resemble the vendor's stress case: long context, long generation, low concurrency. NVIDIA reasons that long context stresses data movement during prefill, long generation amplifies small per-token CC overhead during decode, and high request volume can amortize fixed encryption costs. The second sample should be normal production traffic taken from real request logs: the actual input and output length distributions, the actual arrival pattern, the actual concurrency at the 50th and 95th percentile of load. The preprint shows why this second sample matters. In a single-GPU sweep of one model (Qwen2.5-72B-AWQ on one B200, not the eight-GPU setup above), its patched stack at concurrency 32 showed a median penalty of **1.2%** across ten shapes but a worst case of **6.5%** at 4096 input and 1024 output tokens (arxiv.org). The preprint's separate production-default MoE results (MiniMax-M2.7-NVFP4 at TP8 on eight B200s) are a different case: they show about 14% at 256 output tokens with 1024 input tokens and concurrency 32, and the authors say short, bursty, prefill-heavy traffic is the one corner that reaches the low teens (arxiv.org). Output length changed the penalty steeply, which the authors attribute to fixed per-step cost amortizing.

The paired benchmark protocol

NVIDIA's method is to run the same workload with CC off and on, changing only CC state (developer.nvidia.com). A buyer-grade version adds controls:

- **Hold fixed:** model weights and quantization, container image digest, framework version and flags, tensor, expert and pipeline parallelism, concurrency, input and output token lengths, driver, VBIOS, CUDA, NCCL, kernel and TDX module. NVIDIA's table caption states that one configuration was used for both CC-on and CC-off runs.
- **Change only CC state.** The preprint's definition is instructive: CC-on and CC-off differ only by the GPU CC bit and the TDX guest object.
- **Reboot between runs.** The preprint warns that the system under test should always be rebooted for clean measurements (arxiv.org); stale state can inflate the apparent tax.
- **Warm up and sweep.** NVIDIA's benchmarking guide recommends a warm-up test before a full concurrency sweep, and TensorRT-LLM's documentation recommends applying its documented GPU settings when conducting performance comparisons (nvidia.github.io); the same documentation labels its benchmarking suite a work in progress (nvidia.github.io).
- **Repeat.** Single passes in the preprint carry roughly plus or minus 2 percentage points of noise; in one pinned pair of five-repeat sessions the TP8 penalty was 3.6% in session one and 2.8% in session two. Report paired differences with a confidence interval; NIST's engineering handbook defines paired data and gives the interval for the mean difference (www.itl.nist.gov).
- **Fix metric definitions.** Tool implementations differ, and results are comparable only when definitions align (docs.nvidia.com). Tools also differ on whether TTFT enters the per-token average, and with speculative decoding one streamed output can carry several tokens (docs.vllm.ai); vLLM's documentation advises comparing measurement points and formulas rather than metric names (docs.vllm.ai).

Two formulas are used throughout. Throughput retained is 100 times CC-on tokens per second divided by CC-off tokens per second (developer.nvidia.com). TPOT overhead is 100 times (CC-on TPOT divided by CC-off TPOT, minus 1). Both should be reported per concurrency level, with the number of repeats, and per traffic sample.

Diagnose before accepting or rejecting

When a paired result misses the SLO, the failure usually has an identifiable path. Work through four checks in order:

- **Copy behavior:** confirm that readback is asynchronous and that pageable memory is used on affected paths, as in NVIDIA's description.
- **Autotuner:** confirm that tactic selection is not using unstable event timing.
- **Collective path:** NCCL documents that setting `NCCL_NVLS_ENABLE` to 0 disables NVLink SHARP and allocates no NVLS resources (docs.nvidia.com), so logs should show which algorithm is actually in use rather than assuming.
- **Memory offload:** the preprint advises avoiding CC-mode KV-cache offload, CPU expert offload and weight streaming on the hot path (arxiv.org).

Tooling is constrained: kernel profiling is disabled by design under CC, and performance counters are disabled in full CC-On mode. Diagnosis therefore needs a separate DevTools environment with synthetic data, and its results must never be promoted into the production policy.

Add capacity headroom and decide

Headroom has three components, and each can be computed from the paired data:

- **Throughput retention:** required CC capacity is CC-off capacity divided by measured retention. At NVIDIA's lowest published retention of 96.1% (the lowest of five tested points on one model, not a lower bound) the multiplier is 1.04; at a hypothetical 85% it would be 1.18.
- **SLO binding:** test whether the TPOT SLO binds before throughput does. In NVIDIA's concurrency-16 results, CC-on mean TPOT was **18.039 ms** at a ratio of 103.1% to baseline, which implies a CC-off value near **17.50 ms** (derived arithmetic, subject to rounding). An SLO of 17.8 ms would pass CC-off and fail CC-on at that concurrency, forcing lower per-node concurrency and a capacity cost larger than the 3.2% throughput loss (**Hypothetical Example**).
- **Resilience:** include spare capacity for CVM relaunch, because a reboot command in a CVM terminates the VM (docs.nvidia.com), and a relaunch needs fresh attestation before secrets are released (developer.nvidia.com).

A sizing worksheet for one service therefore has these rows:

- **Baseline nodes:** CC-off nodes needed at the SLO.
- **Retention:** measured retention at the production concurrency.
- **Adjusted nodes:** nodes after dividing by retention and rounding up.
- **SLO adjustment:** additional nodes if the TPOT SLO binds.

- **Spares:** nodes for rolling updates and attestation outages.

In a hypothetical fleet needing five eight-GPU nodes at CC-off (Hypothetical Example), retention of 96.1% gives 5.20 and rounds up to six nodes, while a measured retention of 80% gives 6.25 and rounds up to seven, before spares.

Preserve evidence and keep a rollback path

An acceptance decision needs an evidence package, not only a chart. The following list is a minimum security-evidence checklist:

- **Attestation tokens:** signed CPU and GPU results for each launched CVM, logged for audit as the Consortium recommends for digitally signed attestation messages (confidentialcomputing.io).
- **Policy file:** the allow-list of driver, VBIOS, TDX module, kernel and image digests, with the CC-DevTools mode excluded.
- **Freshness proof:** a nonce match recorded in the GPU claims.
- **Verifier dependency record:** which verifiers were used, with failure behavior and retry policy.
- **Benchmark artifacts:** raw per-run outputs for both CC states, configuration dumps, and the derived paired statistics.
- **Mode control record:** who can change GPU mode, since it is set from the host and persists across reboots and power cycles.
- **Persistence setting: persistence mode** kept enabled for the driver.

Rollback should mean returning to the last known-good CC-on configuration, not running production data with CC off. Firmware changes can be one-way: NVIDIA's compatibility matrix notes that for the HGX Hopper firmware family, rollback to 1.7.1 or earlier is not supported (www.nvidia.com). A change window should therefore keep the previous driver, image and policy together, and test them as a unit before an upgrade is promoted. TensorRT-LLM release-candidate containers are published weekly from the main branch (nvidia.github.io), which is another reason to pin the digest used for acceptance.

Data Analysis and Evidence

What NVIDIA measured on B200

Table 2 below reproduces NVIDIA's published DGX B200 results from the figure descriptions on the page and adds derived columns. Throughput retained comes from NVIDIA's chart text (developer.nvidia.com), TPOT ratios from the second chart (developer.nvidia.com), and absolute CC-on TPOT values from the same chart (developer.nvidia.com). The implied CC-off TPOT is this report's arithmetic (CC-on TPOT divided by the ratio) and the page does not state CC-off absolutes.

Concurrency	Throughput retained, %	Throughput loss, % (derived)	CC-on TPOT vs CC-off, %	TPOT overhead, %	CC-on mean TPOT, ms	Implied CC-off TPOT, ms (derived)
1	98.2	1.8	101.2	1.2	3.225	3.187

Concurrency	Throughput retained, %	Throughput loss, % (derived)	CC-on TPOT vs CC-off, %	TPOT overhead, %	CC-on mean TPOT, ms	Implied CC-off TPOT, ms (derived)
2	98.0	2.0	103.2	3.2	4.571	4.429
4	96.1	3.9	104.3	4.3	6.658	6.384
8	96.8	3.2	103.5	3.5	10.759	10.395
16	96.8	3.2	103.1	3.1	18.039	17.497

The setup was one DGX B200 with eight B200 GPUs, TensorRT-LLM 1.3.0rc22, driver 595.71.05 and CUDA 13.2 (developer.nvidia.com). The workload table is captioned as the configuration for the CC-on and CC-off comparison. For hardware context, NVIDIA lists 1,440 GB of total GPU memory and 64 TB/s of HBM3e bandwidth for the DGX B200 (www.nvidia.com), with eight Blackwell GPUs on fifth-generation NVLink (www.nvidia.com), and the DeepSeek-R1 NVFP4 model card states that quantization reduces disk and GPU memory requirements by approximately 1.6x (huggingface.co). The checkpoint's model card states that the sample deployment needs eight B200 GPUs (huggingface.co).

Table 2 shows two patterns. Retention is lowest at concurrency 4 and recovers slightly at 8 and 16, which is consistent with, but does not prove, NVIDIA's statement that higher request volume can amortize fixed costs. And TPOT overhead and throughput loss are close but not equal at each level, because output-token throughput in a benchmark includes more than steady-state decode. The decision-relevant reading is conditional: for this model, framework build and 32K/1K shape, the overhead was single-digit percent. NVIDIA itself concludes that confidential computing does not remove the need for performance engineering.

What other sources add

NVIDIA's separate July 2, 2026 post tested HGX B300, not B200, with Qwen 3.5 397B-A17B at FP8 (developer.nvidia.com). It reports a relative performance impact of approximately -1% to -8% across concurrency levels and sequence lengths (developer.nvidia.com), with all other variables held constant. The headline "**up to 98%**" is a best-case figure from that table, and the NVIDIA product page repeats it as "up to 98%" with the caveat that results depend on workload, platform and configuration (www.nvidia.com). The B300 post used **SGLang**, so it is evidence on a second framework and GPU generation rather than a replication of the B200 result.

The non-NVIDIA preprint adds a wider surface. Its authors say it covers seven production models, and reports that GPU compute, energy draw and usable memory capacity are unaffected by CC. Its configured-stack results include a TP8 penalty plateau near **5.4%** by concurrency 64 for one mixture-of-experts model at 1024 input and 2048 output tokens, and long context at 32k input within measurement noise in one multi-GPU case. Its stock-stack results are far worse: 30% to 40% penalties attributed to avoidable configurations (arxiv.org). The authors state they did not audit any security property (arxiv.org) and that cross-node transfer, including RDMA and prefill/decode disaggregation, is out of scope.

Older H100 studies are a different regime. A Hopper study that used vLLM v0.5.4 found the average overhead less than 9% (arxiv.org) and attributed most of the penalty to data transfer. A single H100 VM study reported throughput penalties of 4% to 8% that diminish as batch and input sizes grow (arxiv.org), on a rented VM without bare-metal access. Workloads dominated by model swapping or offload showed much larger gaps, such as 45% to 70% higher throughput without CC in a model-swapping scenario (arxiv.org) and up to 52.8% and 88.2% throughput drops for memory-oversubscribed OPT models on H100 (arxiv.org). These do not transfer to a B200 model that fits in GPU memory, and they are a reason to keep offload off the hot path.

Are the vendor and preprint numbers comparable?

Table 3 below sets the sources side by side on the variables that decide comparability.

Source	Hardware and TEE	Stack	Workload and concurrency	Reported result	Use for a B200 buyer
NVIDIA B200 post	One DGX B200, eight GPUs, Intel TDX	TensorRT-LLM 1.3.0rc22, driver 595.71.05	DeepSeek-R1-0528-NVFP4, 32K in, 1K out, concurrency 1 to 16	96.1% to 98.2% throughput retained (developer.nvidia.com)	Best template for the paired method; one model and shape only
NVIDIA B300 post	HGX B300 with Intel TDX	SGLang with CC-specific changes	Qwen 3.5 397B FP8, concurrency 4 to 256	-1% to -8%	Different GPU and framework; shows range at higher concurrency
arXiv 2608.26575 (v2)	Eight B200 on Xeon 6767P with TDX	SGLang and a cc-fixes branch, driver 595.71.05	Seven models; concurrency and shapes vary	Low single-digit for configured stacks; 30% to 40% for stock	Corroborates mechanism; not the same stack
Hopper studies	H100 or H200, single GPU	vLLM v0.5.4 in the cited study	Single-GPU sweeps	Under 9% average	Different GPU generation and CC mode; context only

Three conclusions follow. First, the method is compatible between NVIDIA's B200 post and the preprint: both use paired CC-on and CC-off runs on eight-GPU B200 systems under TDX with the same driver version. Second, the stack is not: TensorRT-LLM with DeepSeek-R1 versus SGLang and vLLM with other models, and the preprint itself notes that its rows were collected over several weeks so cross-row comparisons are directional rather than controlled (arxiv.org). Third, the preprint reports its measurements alongside NVIDIA's published low single-digit figure. Because the preprint shows no journal reference on its arXiv listing and the two sources share a driver version, they should be read as two consistent data points rather than a statistical confirmation. Neither source licenses a universal penalty, in either direction.

Implications and Future Directions

The central implication is procedural. The B200 CC overhead can be small when copy paths, timing and collective selection are CC-aware, and large when they are not. A procurement document should therefore require the supplier or internal platform team to deliver a paired acceptance report at the buyer's traffic shape, with pinned versions and raw data. Vendor percentages belong in the plan as hypotheses.

TEE-IO is the main thing to track, and the sources reviewed describe it as a path rather than a deployed default for B200. NVIDIA calls Blackwell the first TEE-I/O capable GPU in the industry (www.nvidia.com), and its Secure AI whitepaper says that enabling TDISP and IDE end-to-end requires both GPU and CPU support (docs.nvidia.com) and describes inline encryption as higher performance than the bounce buffer. Intel states that TDX 2.0 extends TDX to TEE-IO (www.intel.com), but lists the newest TDX Connect interface documents as not final (www.intel.com), and its enabling guide says additional features may be added later (cc-enabling.trustedservices.intel.com). NVIDIA's measured B200 results described above use the software bounce-buffer path. A move to a hardware-protected I/O path would change the copy economics and require a repeat of the paired test; no source reviewed provides a B200 TEE-IO benchmark.

Availability is also still developing. Among the cloud documentation reviewed, Microsoft's confidential GPU VMs pair AMD SEV-SNP with H100 (learn.microsoft.com), Google lists H100 with Intel TDX on A3 (docs.cloud.google.com) and announced a Blackwell RTX PRO 6000 confidential G4 preview (cloud.google.com), and none of the pages reviewed lists a B200 confidential VM. Google also documents that its GPU-accelerated confidential VMs do not support creating clusters (docs.cloud.google.com). NVIDIA also publishes a reference architecture for confidential AI factories that covers Hopper and Blackwell GPUs and uses a Key Broker Service to release the model decryption key into protected memory only after attestation verification (developer.nvidia.com). Buyers evaluating a hosted service should ask for the paired report and the attestation policy, not only the announcement.

A procurement checklist follows from this analysis:

- **Stack pinning:** the supplier states the driver, firmware, kernel, TDX module and framework image, matching Table 1.
- **Paired report:** CC-on and CC-off results at the buyer's traffic shape, with repeats and raw data.
- **Attestation policy:** CPU and GPU evidence both verified, debug modes excluded, evidence retained.
- **Licensing and support:** the production CC license and the supported version window confirmed in writing.
- **Scope statement:** a written note that availability and application vulnerabilities fall outside what attestation addresses.
- **Change control:** the paired test repeated after each version change, with a tested rollback path.

Finally, operational drift is a cost. Driver, firmware, kernel and framework versions each change the CC overhead, as the autotuner and readback fixes illustrate. A standing regression job that repeats the paired test after each upgrade turns a one-time acceptance into a maintained control. Disclosure: GPU Smith's role in such an engagement is advisory, defining the test protocol, assumptions and acceptance thresholds, while the measurements themselves must come from the buyer's own hardware.

Frequently Asked Questions (FAQs)

What is the performance overhead of confidential computing on a B200? In NVIDIA's DGX B200 test, CC-on retained 96.1% to 98.2% of CC-off throughput for DeepSeek-R1 across concurrency 1 to 16. A non-NVIDIA, unreviewed preprint reports low single-digit overhead for configured stacks and 30% to 40% for stock stacks. The figure for a given deployment must be measured.

Is there a public GPU confidential computing inference benchmark a team can reuse? NVIDIA's B200 post documents a paired method and stack that can be copied, and NVIDIA's NIM benchmarking guide documents metric definitions. Reuse the method, not the percentages.

Does attestation slow each request? NVIDIA says attestation does not add latency to individual inference requests once the workload is running. The non-NVIDIA preprint did not measure attestation or startup overhead, so startup and relaunch time should be measured locally.

How does the Blackwell attestation flow work? The GPU collects and signs measurements and a verifier compares them with known-good reference values; NRAS validates GPU evidence and issues signed tokens, and the CPU quote is verified separately for TDX (docs.kernel.org).

Is TEE-IO available for B200 inference today? NVIDIA describes Blackwell as TEE-I/O capable, but its B200 benchmark describes a software-encrypted bounce-buffer path. Ask the supplier which path is deployed and test that path.

H100 versus B200: what changes? Blackwell supports multi-GPU CC with encrypted NVLink, while Hopper's multi-GPU option is a protected PCIe mode. Published H100 overhead studies are mostly single-GPU and use older frameworks.

Which verifier should a regulated team use? NVIDIA offers remote and local options, and Intel documents a composite flow with two verifiers. Choose based on network policy, token format, retention requirements and failure behavior, and test that failure mode.

What belongs on a procurement checklist? Use the checklist in the implications section: pinned stack versions from Table 1, a paired acceptance report, an attestation policy with evidence retention, the production CC license, and a rollback plan.

Conclusion

On the limited public record reviewed here, two B200 sources and one of them NVIDIA's own, confidential B200 inference looks neither free nor prohibitively expensive. NVIDIA's documented DGX B200 test shows 96.1% to 98.2% throughput retained and 1.2% to 4.3% higher mean TPOT for one model, one framework build and one long-context shape. A non-NVIDIA, unreviewed preprint on similar hardware reports low single-digit overhead for configured stacks and far larger penalties for stock configurations. The shared lesson is that the copy path, autotuner timing, kernel-launch cost and collective algorithm decide the result.

For a regulated buyer, the acceptance gate has a performance half and a security half. The performance half is a paired CC-on and CC-off run at the buyer's traffic shape, with pinned versions, repeated measurements and an SLO check that includes TPOT, not only throughput. The security half is attestation of both CPU and GPU, a policy that rejects debug modes and unapproved firmware, and retained evidence; attestation provides evidence for a trust decision and does not replace the policy. Neither half is satisfied by a vendor percentage or by encryption alone.

Current information is dated November 2026 and the stack is moving: release notes, firmware families and TEE-IO support change over time, and the non-NVIDIA preprint was still at its second version when reviewed. The practical recommendation is to treat vendor results as the template for a test, run that test on the intended

hardware, add measured headroom, and repeat it after every material version change.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://developer.nvidia.com/blog/enabling-private-high-performance-production-ai-inference-with-nvidia-confidential-computing
2. https://arxiv.org/html/2608.26575
3. https://arxiv.org/abs/2608.26575
4. https://docs.nvidia.com/confidential-computing/virtual-machines/v1.0.0/threat-model
5. https://docs.nvidia.com/confidential-computing/overview
6. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML?uri=CELEX:32016R0679
7. https://docs.nvidia.com/confidential-computing/release-notes/r595-trd1
8. https://www.nvidia.com/en-us/data-center/solutions/confidential-computing/secure-ai-compatibility-matrix
9. https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC-A-Technical-Analysis-of-Confidential-Computing-v1.3_unlocked.pdf
10. https://www.intel.com/content/www/us/en/developer/tools/trust-domain-extensions/overview.html
11. https://nvidia.github.io/TensorRT-LLM/latest/deployment-guide/deployment-guide-for-deepseek-r1-on-trtllm.html
12. https://docs.nvidia.com/nim/benchmarking/llm/latest/metrics.html
13. https://docs.nvidia.com/confidential-computing/resources/licensing
14. https://developer.nvidia.com/blog/hardware-rooted-ai-security-that-wont-slow-you-down
15. https://docs.trustauthority.intel.com/main/articles/articles/ita/concept-gpu-attestation.html
16. https://docs.kernel.org/arch/x86/tdx.html
17. https://docs.nvidia.com/confidential-computing/virtual-machines/v1.0.0
18. https://docs.nvidia.com/confidential-computing/virtual-machines/v1.0.0/supported-platforms
19. https://docs.nvidia.com/confidential-computing/virtual-machines/v1.0.0/deployment-guide/host
20. https://ubuntu.com/server/docs/how-to/virtualisation/intel-tdx
21. https://docs.cloud.google.com/confidential-computing/confidential-vm/docs/create-a-confidential-vm-instance-with-gpu
22. https://github.com/canonical/tdx
23. https://docs.nvidia.com/confidential-computing/virtual-machines/v1.0.0/deployment-guide/gpu-mode
24. https://docs.nvidia.com/attestation/nv-attestation-sdk-cpp/latest/overview.html
25. https://docs.nvidia.com/confidential-computing/release-notes/r580-trd1
26. https://docs.nvidia.com/confidential-computing/release-notes/r590-trd1
27. https://docs.nvidia.com/attestation
28. https://www.rfc-editor.org/rfc/rfc9334.html
29. https://learn.microsoft.com/en-us/azure/confidential-computing/confidential-vm-overview
30. https://docs.nvidia.com/attestation/cloud-services/latest/nras/nras_introduction.html
31. https://github.com/NVIDIA/nvtrust

32. https://docs.nvidia.com/attestation/advanced-documentation/latest/claims-guide/gpu_claims.html
33. https://confidentialcomputing.io/wp-content/uploads/sites/10/2025/09/ccc_gdpr_comp_091025.pdf
34. https://nvlpubs.nist.gov/nistpubs/ir/2026/NIST.IR.8320E.ipd.pdf
35. https://docs.nvidia.com/cuda/cuda-runtime-api/api-sync-behavior.html
36. https://nvidia.github.io/TensorRT-LLM/performance/perf-benchmarking.html
37. https://www.itl.nist.gov/div898/handbook/prc/section3/prc311.htm
38. https://docs.vllm.ai/en/latest/benchmarking/cli
39. https://docs.nvidia.com/deeplearning/nccl/user-guide/docs/env.html
40. https://docs.nvidia.com/confidential-computing/virtual-machines/v1.0.0/deployment-guide/create-vm
41. https://www.nvidia.com/en-us/data-center/dgx-b200
42. https://huggingface.co/nvidia/DeepSeek-R1-0528-FP4
43. https://www.nvidia.com/en-us/data-center/solutions/confidential-computing
44. https://arxiv.org/html/2409.03992v4
45. https://arxiv.org/abs/2509.18886
46. https://arxiv.org/abs/2505.16501
47. https://arxiv.org/abs/2411.03357
48. https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture
49. https://docs.nvidia.com/nvidia-secure-ai-with-blackwell-and-hopper-gpus-whitepaper.pdf
50. https://www.intel.com/content/www/us/en/developer/tools/trust-domain-extensions/documentation.html
51. https://cc-enabling.trustedservices.intel.com/intel-tdx-enabling-guide
52. https://learn.microsoft.com/en-us/azure/confidential-computing/gpu-options
53. https://cloud.google.com/blog/products/identity-security/verifiable-trust-in-the-ai-era-whats-new-in-confidential-computing
54. https://docs.cloud.google.com/confidential-computing/confidential-vm/docs/supported-configurations
55. https://developer.nvidia.com/blog/building-a-zero-trust-architecture-for-confidential-ai-factories

THE PUBLISHER

About GPU Smith

GPU Smith provides independent engineering and research for private AI infrastructure. We help technical buyers and operators reason about GPU workload sizing, hardware selection, procurement, deployment and inference operations, from the compute system to the networking, power and cooling requirements around it.

Start with the workload

Useful infrastructure decisions begin with the models, data, throughput, latency and operating constraints a team actually has. GPU Smith helps connect those requirements with system design and deployment choices. The goal is a privately controlled AI environment whose capacity and operational demands are understood before a purchasing decision.

Hardware and supplier research

Our public reference library covers GPUs, complete systems and networking components. The server-vendor directory and manufacturing research help teams investigate suppliers, compare options and follow sources behind technical claims. We distinguish manufacturer specifications from measured benchmarks and advertised capabilities from tested configurations. Reference pages are research resources, not a live inventory listing or a binding equipment quote.

Deployment and operations

GPU Smith's engineering scope includes infrastructure integration, networking, power, cooling and the practical operation of inference workloads. Our published articles explain the tradeoffs behind deployment, procurement and ongoing operations so teams can ask better questions and document decisions.

Work with GPU Smith

Explore the hardware library, GPU references, networking references and vendor research. Contact GPU Smith with your workloads and deployment constraints to discuss a project and confirm current scope, pricing and availability.

Inclusion of a manufacturer or product in our research does not imply a partnership, certification or endorsement.

Website: https://gpusmith.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.