GPU SMITH
COMPUTE ENGINEERING

GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

HP ZGX Fury RHEL Certification and Procurement

> 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 · 2026-09-21 · hp zgx fury rhel certification · hp zgx fury · nvidia gb300 · red hat enterprise linux · red hat ai factory · enterprise ai infrastructure · private ai · gpu procurement · ai inference

Original article: https://gpusmith.com/articles/en/hp-zgx-fury-rhel-procurement

In this article

1. Executive Summary
2. Introduction and Background
3. Key Changes
4. Implementation Considerations and Process Changes
5. Data Analysis and Evidence
6. Procurement Decision and Acceptance Plan
7. Implications and Future Directions
8. Frequently Asked Questions (FAQs)
9. Conclusion

Executive Summary

As of **September 21, 2026**, the procurement answer is narrower than the marketing story. **HP ZGX Fury is orderable**, and HP says it is certified to run Red Hat Enterprise Linux (RHEL) (www.hp.com) (catalog.redhat.com). The Red Hat record covers **RHEL 10.2 through 10.x on aarch64**, but that hardware certificate is not evidence that the announced HP implementation of Red Hat AI Factory with NVIDIA is already a generally available, jointly validated production stack. HP's current product page states that ZGX Fury is powered by Red Hat AI Enterprise (www.hp.com). Separately, HP's September 9 launch release calls the Red Hat AI Factory with NVIDIA integration planned and says timing, locations, eligibility, supported configurations, and sandbox access will be disclosed later (www.hp.com) (www.hp.com).

The hardware case is substantial but workload-specific. The Grace Blackwell Ultra design combines a **72-core Grace CPU**, **496 GB of LPDDR5X**, and **252 GB of HBM3e**, for **748 GB of coherent addressable memory** (docs.nvidia.com) (docs.nvidia.com). Coherent does not mean uniform: the HBM3e is rated at **7.1 TB/s**, the LPDDR5X at **396 GB/s**, and NVIDIA states that the CPU and GPU retain dedicated physical memory (docs.nvidia.com) (docs.nvidia.com). Therefore, neither the **20-petaFLOPS sparse FP4** headline nor the **one-trillion-parameter** claim predicts production latency, concurrency, or fine-tuning time.

The recommended decision is conditional. **Buy now** only when a single node fits the model and concurrency target, local data control matters, and the quote closes the missing operational details. **Pilot now** when memory fit appears plausible but latency, KV-cache pressure, power, or software ownership remains unmeasured. **Wait** when the Red Hat AI Factory experience, its supported configuration, or a signed joint escalation route is mandatory. **Use rack or cloud infrastructure** when the workload needs redundant nodes, scale beyond one coherent-memory system, large external storage, or multi-site availability. Production Red Hat AI Factory guidance itself recommends at least **three control-plane machines and two compute workers**, which is materially different from one station (docs.nvidia.com) (docs.nvidia.com).

Before acceptance, capture the exact bill of materials, firmware, RHEL subscription, NVIDIA entitlement, warranty, delivery window, measured power, storage endurance, management interface, and service terms. Then run model-residency, representative concurrency, network and storage, recovery, patching, and air-gap tests. Public pricing, maximum continuous power, thermals, acoustics, and independent ZGX Fury application benchmarks were not available in the sources reviewed, so a defensible three-year total cost must use the buyer’s quote, measured kilowatt-hours, actual tariff, facility overhead, labor, support, and a comparable cloud bill.

Introduction and Background

The **HP ZGX Fury RHEL certification** matters because the system sits between two familiar purchasing categories. It has workstation-like placement and acquisition characteristics, yet its **GB300 Grace Blackwell Ultra Desktop Superchip** gives it a memory footprint and networking configuration intended for shared AI development and inference. That makes the correct buying question operational: can this exact configuration meet the organization’s workload, availability, governance, and support requirements?

The status changed on **September 9, 2026**. HP announced that the station was available to order and RHEL certified, while describing its Red Hat AI Factory integration as future work. The distinction is crucial because Red Hat hardware certification is bounded. Red Hat states that each listing applies to specific architectures and major releases, and that support remains limited by the customer's subscription service-level agreement (access.redhat.com) (access.redhat.com). It does not, by itself, license Red Hat AI Enterprise, license NVIDIA AI Enterprise, establish a high-availability architecture, or prove a workload service-level objective (SLO).

This report uses a procurement-gate approach. It separates what is **shipping**, **planned**, and **unknown**, explains why 748 GB is not a single-speed memory pool, and turns the public evidence into acceptance tests. GPU Smith's relevant posture is that of an adjacent independent engineering advisor, not a station vendor. Its published method covers specification, procurement, integration, and validation of private AI infrastructure against written criteria (gpusmith.com) (gpusmith.com). That perspective appears here as a testable decision framework, not as a competing product row.

Key Changes

September 9 created a procurement gate

Table 1 separates the three evidence states that a purchase committee should record on its decision date.

Status	What the evidence supports	Procurement treatment
Shipping or current	ZGX Fury is orderable; HP states RHEL certification; the catalog records **RHEL 10.2 to 10.x on aarch64** (www.hp.com) (catalog.redhat.com).	Obtain the regional SKU, quote, delivery commitment, warranty, exact RHEL image and subscription, firmware baseline, and entitlement certificates.
Planned	HP plans Red Hat AI Factory with NVIDIA integration and a sandboxed evaluation (www.hp.com).	Do not place planned components on the accepted bill of materials. If required, make delivery and support evidence a purchase condition.
Unknown publicly	Sandbox timing and eligibility, public system price, delivery window, maximum continuous power, acoustics, thermal output, U.S. SKU, storage endurance, detailed BMC capabilities, and ZGX-specific joint support route.	Require written supplier answers. Use measured values and contractual terms, not analogous products or speculative estimates.

The table changes the default interpretation. A buyer can procure **certified RHEL hardware now**, but cannot infer that the planned HP and Red Hat production experience is included. Red Hat AI Factory with NVIDIA exists as a broader product, and Red Hat said it was available on **February 24, 2026** (www.redhat.com). The gap is product-specific validation and packaging for ZGX Fury, not the existence of the general Red Hat offering.

The hardware boundary is large capacity, not uniform performance

The station combines **496 GB LPDDR5X CPU memory** and **252 GB HBM3e GPU memory**. HP lists four 128 GB SOCAMM modules for the CPU side and 252 GB HBM3e for the GPU side (www.hp.com) (www.hp.com). NVIDIA describes a shared address space accessible by both processors, but it also says GB300 access to

CPU memory is not cached and cross-domain bandwidth is limited by NVLink (docs.nvidia.com) (docs.nvidia.com).

Consequently, **748 GB coherent memory** is valuable for fitting larger working sets and enabling controlled spill beyond HBM, but it is not equivalent to 748 GB of HBM3e. Default system allocations do not automatically migrate into GB300 memory, and NVIDIA recommends Unified Virtual Memory primarily when the working set exceeds HBM capacity (docs.nvidia.com) (docs.nvidia.com). Procurement should therefore demand a memory-placement test, not simply a successful model load.

Connectivity is similarly capable but conditional. The reference platform describes **two 400 GbE ports through ConnectX-8** (docs.nvidia.com). Link rate does not establish switch compatibility, optics, cabling, storage throughput, redundancy, or application scaling. Those belong in the quoted design and acceptance plan.

Certification is a support boundary, not a benchmark

Red Hat's certificate is meaningful. It records platform and architecture compatibility, and Red Hat describes certified hardware in terms of tested interoperability, lifecycle management, and support (catalog.redhat.com). It also has limits. Red Hat says the hardware vendor performs certification and that certification applies to a major RHEL release and processor architecture (access.redhat.com) (docs.redhat.com). The policy further says that a certificate remains valid from its posted minor version until the next major version (docs.redhat.com).

Most importantly, performance remains the vendor's responsibility under the policy (docs.redhat.com). A certificate therefore should answer "is this identified model supported on this RHEL line?" It does not answer "will a 70-billion-parameter model meet a 95th-percentile first-token target for 24 users?" That second question requires a controlled workload test.

Implementation Considerations and Process Changes

Workload-fit and model-residency gate

A useful sizing worksheet begins with model weights, then adds memory that marketing capacity figures omit. NVIDIA's heuristic is **parameters multiplied by bytes per parameter, divided by tensor parallelism** (docs.nvidia.com). FP8 halves weight memory relative to 16-bit storage, but weights are only the first line item (docs.nvidia.com). Add key-value (KV) cache, activations, runtime reservations, CUDA graph compilation, fragmentation, adapters, and safety margin. Research on PagedAttention characterizes per-request KV cache as large and dynamically sized (arxiv.org). The cache grows with live sequences and context, and TensorRT-LLM identifies batch, beam width, KV heads, sequence length, and head dimension in its cache shape (nvidia.github.io).

Use these worksheet rows for each candidate model:

- **Weights:** parameter count times storage bytes, using the actual checkpoint and quantization format.
- **Runtime overhead:** engine workspace, CUDA graphs, kernels, allocator reserve, adapters, and framework processes.

- **KV cache:** layers, KV heads, head dimension, cache precision, active tokens, and concurrent sequences.
- **HBM target:** the latency-critical share intended to remain inside the **252 GB HBM3e** domain.
- **CPU-memory spill:** the explicitly tested share in LPDDR5X, including transfer behavior and latency impact.
- **Growth reserve:** longer contexts, higher concurrency, model revisions, and parallel workloads.

A simple worked estimate illustrates the gate without pretending to benchmark the ZGX. A 70-billion-parameter checkpoint stored at two bytes per parameter needs about **140 GB** for weights before runtime and KV cache. At one byte per parameter, the weight file is about **70 GB**. These are arithmetic estimates, not measured residency, and quantization metadata can add overhead. The acceptance criterion should be observed steady-state placement under the target context and concurrency, plus a margin agreed before purchase.

Single-node serving is attractive when the entire latency-sensitive working set fits and locality is valuable. It becomes less attractive when the required KV-cache concurrency exceeds the node. vLLM's own scaling guidance says to add GPUs or nodes when KV-cache-derived concurrency is insufficient (docs.vllm.ai). Multi-node work also introduces identical-environment requirements and shared-storage or orchestration concerns. KServe's documented multi-node vLLM setup requires at least two Kubernetes nodes and ReadWriteMany storage (kserve.github.io) (kserve.github.io).

Software, licensing, and support ownership

HP's native software layer is distinct from HP's stated Red Hat AI Enterprise configuration and the separately planned Red Hat AI Factory with NVIDIA integration. **HP Z Runtime** is preinstalled, and HP also offers it as a Snap package (www.hp.com) (www.hp.com). **HP Z Toolkit** is listed without a separate charge, but its client must be x86-based and run Windows 11 or Ubuntu 24.04 or later (www.hp.com) (www.hp.com).

Red Hat AI Enterprise combines OpenShift Container Platform, OpenShift AI, and Red Hat AI Accelerators into a unified stack (docs.redhat.com). Its OpenShift entitlement is restricted to AI workloads, so mixed clusters need placement controls (docs.redhat.com). **NVIDIA AI Enterprise** is generally licensed per GPU, and NVIDIA says Blackwell DGX licenses are purchased separately (docs.nvidia.com) (docs.nvidia.com). That does not prove the same commercial treatment for ZGX Fury, but it makes an entitlement certificate an essential quote item.

The support matrix should name the first contact and handoff rule for every layer:

- **HP:** chassis, storage devices, cooling, optional display GPU, firmware, rails, and onsite service.
- **Red Hat:** RHEL, OpenShift, and OpenShift AI under the purchased subscription.
- **NVIDIA:** driver, NIM, GPU and Network Operators, and NVIDIA AI Enterprise components under entitlement.
- **Customer or integrator:** model artifacts, serving configuration, identity, application code, data, observability, backup, and network integration.

NVIDIA documents a collaborative Red Hat and NVIDIA support model, routing NVIDIA software issues to NVIDIA Enterprise Support and RHEL or OpenShift issues to Red Hat (docs.nvidia.com) (docs.nvidia.com). A ZGX purchase should turn that general model into a signed escalation path including HP.

Production operations gate

A single ZGX Fury is still a single system. Red Hat defines high availability around eliminating single points and moving services between cluster nodes; its basic example needs two RHEL nodes and fencing for each (docs.redhat.com) (docs.redhat.com). If the service cannot tolerate station maintenance or component outage, the solution needs another serving path.

Production acceptance should cover:

- **Remote operations:** verify power control, console access, telemetry, authentication, audit logs, firmware inventory, and recovery without local hands. Redfish defines a multivendor out-of-band interface, but public ZGX evidence reviewed here does not establish its BMC implementation (www.dmtf.org). Its schema can represent BIOS, management-controller firmware, device firmware, drivers, and provider software (redfish.dmtf.org).
- **Configuration control:** preserve SKU, serials, NICs, SSDs, optional GPU, BIOS, BMC, device firmware, RHEL image, drivers, containers, models, and checksums. NIST calls for a current baseline under configuration control (nvlpubs.nist.gov).
- **Storage:** confirm capacity, RAID behavior, endurance, encryption, backup destination, checkpoint time, and restore time. SNIA notes that SSD endurance is finite and workload-dependent (www.snia.org). NIST recommends that backups be created regularly and tested (csrc.nist.gov).
- **Observability:** retain resource, application, queue, model, and SLO metrics. RHEL Performance Co-Pilot supports monitoring and historical retrieval (docs.redhat.com). Prometheus recommends alerting on high latency and error rates high in the stack (prometheus.io).
- **Patching:** stage firmware, kernel, driver, container, and model changes with rollback criteria. RHEL 10 live patching begins at 10.2, but Red Hat says it does not cover every critical or important CVE (docs.redhat.com) (docs.redhat.com).
- **Disconnected operation:** mirror all images, charts, packages, models, and licenses through an approved transfer path. Red Hat's workflow mirrors to disk, then manually transfers the archive to the disconnected registry network (docs.redhat.com). NVIDIA's air-gap tooling records SHA-256 checksums in a manifest (docs.nvidia.com).

Data Analysis and Evidence

Specification and support matrix

Table 2 converts the specification sheet into operational consequences. Vendor peak values are retained as specifications, not application results.

Dimension	Published evidence	Decision implication
Compute	72-core Grace CPU; up to **20 petaFLOPS sparse FP4** (www.hp.com)	Use FP4 only as a hardware mode descriptor. Measure the actual model, precision, engine, batch, and SLO.

Dimension	Published evidence	Decision implication
GPU memory	**252 GB HBM3e**, vendor-rated **7.1 TB/s** (docs.nvidia.com)	Place latency-critical weights and cache here when possible. Record observed residency.
CPU memory	**496 GB LPDDR5X**, vendor-rated **396 GB/s** (www.hp.com)	Treat spill as a separate performance tier and benchmark transfer-heavy paths.
Storage	HP offers **2 TB or 4 TB NVMe M.2** self-encrypting choices; selection occurs at purchase (www.hp.com)	Size model registry, checkpoints, logs, containers, and scratch separately. Obtain endurance and replacement terms.
Network	Two **400 Gbps** QSFP112 ports	Validate optics, switch mode, link redundancy, RDMA configuration, and end-to-end storage throughput.
Placement	Tower or standard **5U rack** using included rails (www.hp.com)	Confirm regional SKU, rail compatibility, service clearance, weight, branch circuit, cooling, and acoustic limits.
RHEL	Certified for **RHEL 10.2 to 10.x, aarch64**	Freeze the accepted OS, driver, firmware, and repo combination. Verify every optional component in the ordered SKU.
AI platform	HP states that ZGX Fury is powered by Red Hat AI Enterprise (www.hp.com); the separately announced Red Hat AI Factory with NVIDIA integration and sandbox remain planned (www.hp.com).	Price subscriptions and support separately. Do not accept a roadmap statement as a delivered configuration.

The matrix exposes the main tradeoff. ZGX Fury offers unusually large single-node coherent capacity in a compact placement, but the fast domain is one-third of the marketed 748 GB. Its two high-rate ports create expansion options, not automatic scale-out. The RHEL record lowers operating-system compatibility uncertainty, not workload or topology uncertainty.

Concurrency and SLO test sheet

The throughput test should use a fixed model digest, precision, serving engine, prompt-length distribution, output-length distribution, and arrival process. Report at least:

- **Time to first token, P50 and P95:** user-perceived queue plus prefill delay.
- **Inter-token latency, P50 and P95:** generation smoothness after the first token.
- **Per-user tokens per second:** experience at each concurrency level.
- **System tokens per second:** aggregate throughput across all users.
- **Sustained concurrency:** simultaneous requests while all latency objectives remain satisfied.
- **Memory placement:** HBM, LPDDR5X, KV cache, allocator reserve, and observed peak.
- **Power and thermals:** idle, representative, and peak test-window measurements.
- **Stability:** error rate, queue depth, throttling, restarts, and recovery over an extended run.

MLCommons defines concurrency as simultaneous queries sustained in flight and system throughput as total tokens produced across all users (mlcommons.org) (mlcommons.org). Its framework highlights the expected tradeoff: total throughput often rises as concurrency increases while per-user speed falls (mlcommons.org).

MLPerf's Llama 2 method uses tokens per second for throughput because input and output lengths vary (mlcommons.org). This is why a single maximum tokens-per-second number is inadequate.

Three-year cost template

No credible universal total exists without a customer quote and measurements. Build the model as:

Three-year owned cost = acquisition + subscriptions + support + deployment labor + 36-month energy + facility overhead + network and storage + spares and backup + operating labor + expected downtime cost minus residual value.

For energy, calculate measured average IT kilowatts times operating hours times the site electricity rate, then apply measured facility overhead. ENERGY STAR defines Power Usage Effectiveness (PUE) as total annual source energy divided by annual IT source energy (portfoliomanager.energystar.gov). The 2025 U.S. average retail electricity price was **13.63 cents per kWh**, but state averages ranged from **8.20 to 35.72 cents**, so the actual tariff is the correct input (www.eia.gov) (www.eia.gov).

Measure power with an analyzer between the AC source and system under test. SPEC calls for true-RMS watts, uncertainty of **1% or better**, and at least one measurement set per second (ftp.spec.org) (ftp.spec.org). Compare that owned model with an effective cloud bill at the same workload unit, including compute, storage, outbound transfer, support, engineering time, idle capacity, and commitments. The FinOps definition of total cost includes acquisition, support, communications, labor, training, and downtime opportunity cost (framework.finops.org). Its FOCUS specification says effective cost reflects applicable pricing adjustments (focus.finops.org). Google's method separately includes recurring patching, monitoring, and scaling costs (docs.cloud.google.com).

Procurement Decision and Acceptance Plan

Buy now, pilot, wait, or select another platform

Table 3 maps evidence to action. A committee can apply it after the model-residency worksheet and SLO test have explicit pass thresholds.

Path	Choose it when	Evidence required before commitment
Buy ZGX Fury now	One node fits weights, KV cache, overhead, and growth; planned Red Hat AI Factory packaging is not a dependency; a maintenance window is acceptable; local control has measurable value.	Regional SKU and delivery commitment, accepted RHEL image, entitlements, measured pilot results or contractual acceptance test, power and environment schedule, service route.
Pilot ZGX Fury now	Capacity appears sufficient, but LPDDR5X spill, concurrency, data path, air-gap flow, or support handoffs remain uncertain.	Time-boxed unit, representative data and model, instrumentation, pass/fail SLO, rollback and data-erasure process.
Wait for HP and Red Hat stack	A supported Red Hat AI Factory configuration, sandbox evidence, OpenShift operating model, or joint escalation path is mandatory.	Generally available SKU/configuration, subscription BOM, compatibility matrix, lifecycle statement, and signed multivendor escalation map.

Path	Choose it when	Evidence required before commitment
Use a rack server	Redundant power, extensive NVMe, out-of-band management, PCIe expansion, or several accelerators are hard requirements.	Rack design, facility capacity, fabric, storage, cluster software, support, and workload benchmark. A current eight-GPU rack example offers redundant hot-swap power, hardware RAID boot, and Redfish management (www.dell.com) (www.dell.com) (www.dell.com).
Use a cloud service	Demand is bursty, capacity must scale beyond one node, geographic resilience matters, or the organization prefers infrastructure consumption over ownership.	Reserved capacity, full bill at target use, data-transfer model, guest-OS duties, quota, regions, SLO, exit path, and data-control review. AWS identifies compute, storage, and outbound transfer as core cost drivers (docs.aws.amazon.com).

The right comparison is not one ZGX Fury against the theoretical maximum of a rack or cloud. It is the smallest complete architecture that meets the same requirement. AWS offers GB300 configurations with multi-terabyte GPU memory and multi-terabit networking, while Azure documents NVLink domains up to 72 GPUs (aws.amazon.com) (aws.amazon.com) (learn.microsoft.com). Google documents **3,200 Gbps** of GPU networking for A4X Max (docs.cloud.google.com). Those are different scale classes and operating models. They become relevant only when the workload needs them.

Acceptance sequence

The procurement record should contain objective evidence at each gate:

1. **Commercial baseline:** capture quote, SKU, serializable BOM, delivery date, warranty geography, onsite response, subscriptions, renewal terms, and return conditions.
2. **Firmware and software baseline:** record BIOS, management controller, NIC, storage, GPU driver, RHEL build, kernel, CUDA, containers, model digests, and licenses. NIST recommends inventories be updated during installs, removals, and system updates (nvlpubs.nist.gov).
3. **Facility test:** verify rack fit, rail and service clearance, circuit, grounding, measured idle and load power, inlet temperatures, airflow, noise, and shutdown behavior. ENERGY STAR lists temperature, input power, utilization, inlet temperature, and airflow as useful measurements (www.energystar.gov). ASHRAE's published recommended temperature range is **18°C to 27°C** (www.ashrae.org).
4. **Memory-residency test:** load the production checkpoint, reach target context and concurrency, record HBM and CPU-memory use, then compare all latency percentiles with the SLO.
5. **Network and storage test:** validate each port, failover design, east-west bandwidth, model load, checkpoint write, sustained log volume, backup, and restore.
6. **Concurrency test:** replay representative arrivals and prompt sizes. Report TTFT, inter-token latency, per-user and total throughput, error rate, queue depth, and power at each level. MLPerf models random server arrivals with a Poisson distribution (mlcommons.org).
7. **Recovery test:** restart serving, restore a checkpoint, replace or logically isolate a storage path, expire a credential, and confirm alarms and runbooks. Measure scheduled and unscheduled interruption because both contribute to availability (docs.aws.amazon.com).
8. **Patch and air-gap test:** import signed artifacts, verify checksums and software bill of materials, stage an update, roll it back, and prove operation without unintended external dependencies. CISA defines an SBOM as a formal component and supply-chain relationship record (www.cisa.gov).

9. **Support drill:** open a non-urgent test case and verify HP, Red Hat, NVIDIA, and integrator ownership, required logs, entitlement recognition, handoff rules, and closure evidence.

Acceptance should end with an as-built package and signed exceptions. That is especially important for an early product whose public pages do not yet answer power, noise, endurance, and detailed service questions.

Implications and Future Directions

ZGX Fury indicates a meaningful shift in department-level infrastructure: memory capacity once associated with larger multi-accelerator systems can now sit in a tower or **5U** footprint. This favors private inference, local development, controlled fine-tuning, and remote-site deployments where data movement or intermittent connectivity is a constraint. It also moves data-center disciplines closer to the team buying the node. Monitoring, patch management, artifact provenance, backup, power, cooling, and escalation cannot be deferred merely because the device resembles a workstation.

The planned Red Hat integration could reduce that operational discontinuity if HP publishes a validated configuration, lifecycle, subscription BOM, and multivendor support path. Red Hat AI 3.5 was generally available on **September 9, 2026**, and is part of the broader Red Hat AI Factory with NVIDIA offering (www.redhat.com) (www.redhat.com). The remaining uncertainty is ZGX-specific delivery, not Red Hat's general platform roadmap.

Independent workload evidence will matter more than peak arithmetic. MLCommons describes reproducible, architecture-neutral inference testing and separates available systems from preview systems (mlcommons.org) (docs.mlcommons.org). NIST recommends reporting benchmark versions, model versions, protocol details, and uncertainty estimates (nvlpubs.nist.gov). Until comparable ZGX Fury results appear, procurement teams should publish their own model version, engine, precision, context distribution, concurrency, latency percentiles, throughput, power, and software bill. A result without those conditions is not portable.

For GPU Smith's audience, the opportunity is therefore procedural. An adjacent advisor can make the decision legible by reconciling the quote with the workload, facility, network, software, and acceptance record. It should not imply an HP partnership or substitute advisory judgment for vendor support. The result is a build, pilot, wait, or alternate-platform recommendation with explicit assumptions and measurable release criteria.

Frequently Asked Questions (FAQs)

Is HP ZGX Fury certified for Red Hat Enterprise Linux?

Yes, within a defined boundary. The Red Hat catalog records HP ZGX Fury G1n for **RHEL 10.2 through 10.x on aarch64**. Buyers should match the ordered SKU, optional devices, firmware, and OS image to the certificate and support statement. Hardware certification does not establish workload performance or include every AI platform subscription.

Is Red Hat AI Factory with NVIDIA shipping on ZGX Fury?

HP's current product page states that ZGX Fury is powered by Red Hat AI Enterprise (www.hp.com). HP's September 9 launch release separately calls the Red Hat AI Factory with NVIDIA integration and sandbox planned; timing, locations, eligibility, supported configurations, and access will be shared when available (www.hp.com).

Does 748 GB mean every byte performs like HBM3e?

No. The capacity is coherent and addressable across CPU and GPU, but the system has **252 GB HBM3e** and **496 GB LPDDR5X** with different bandwidth and access behavior. Buyers should test placement, transfer behavior, and SLO impact under the real workload.

Can ZGX Fury run a one-trillion-parameter model?

HP makes an **up to one trillion parameters** vendor claim (www.hp.com). The cited passage does not disclose precision, offload, runtime overhead, context, batch, concurrency, or achieved speed. Treat the number as a capacity-positioning claim until the exact model passes the residency and SLO tests.

What must be in an HP ZGX Fury procurement package?

At minimum: regional SKU, price, delivery date, warranty and onsite response, complete BOM, rails, power and environmental schedule, storage endurance, firmware, accepted RHEL build, HP software terms, Red Hat and NVIDIA subscriptions, network components, backup design, acceptance criteria, and a signed support escalation map.

When is a rack server or cloud service more appropriate?

Choose another platform when the requirement exceeds one node, needs redundant service during maintenance, requires much larger local storage or PCIe expansion, or demands multi-zone geographic operation. Google, for example, requires reserved capacity for its A4X Max instances, illustrating that cloud removes hardware ownership but not capacity planning (docs.cloud.google.com). Compare complete architectures at the same SLO and utilization, not nominal accelerator names.

Conclusion

HP ZGX Fury is a buyable, RHEL-certified system whose current product page states that it is powered by Red Hat AI Enterprise; the separately announced Red Hat AI Factory with NVIDIA integration and sandbox remain planned procurement conditions. The current certificate provides a useful RHEL 10.2-to-10.x aarch64 support anchor. It does not bundle AI subscriptions, prove application throughput, provide redundancy, or convert a planned sandbox into a production platform.

The system’s defining advantage is **748 GB of coherent capacity** around a GB300, paired with two 400 Gbps ports and tower or 5U placement. Its defining sizing constraint is equally clear: only **252 GB** is HBM3e, while **496 GB** is a separate LPDDR5X domain. Model weights can fit while KV cache, runtime overhead, concurrency, and memory traffic still miss the service objective.

The practical recommendation is to buy only against evidence. Close the commercial and support gaps in writing, measure power rather than estimating it, test the exact model under representative arrivals, verify backup and recovery, and preserve a complete as-built baseline. Pilot when the outcome depends on memory spill or concurrency. Wait when the ZGX-specific Red Hat AI Factory configuration is mandatory. Select rack or cloud capacity when high availability or scale cannot be satisfied by one node. This turns a new-product announcement into a defensible infrastructure decision.

Source links

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

1. https://www.hp.com/us-en/newsroom/press-releases/2026/hp-extends-data-center-ai-architecture-to-the-edge.html
2. https://catalog.redhat.com/en/hardware/system/detail/318527
3. https://www.hp.com/us-en/workstations/ai-stations/zgx-fury-ai-station.html
4. https://docs.nvidia.com/dgx/dgx-station-development-guide/overview.html
5. https://docs.nvidia.com/dgx/dgx-station-development-guide/Intro.html
6. https://docs.nvidia.com/dgx/dgx-station-development-guide/coherency.html
7. https://docs.nvidia.com/ai-enterprise/deployment/red-hat-ai-factory/latest/prerequisites.html
8. https://access.redhat.com/solutions/7494
9. https://gpusmith.com/
10. https://www.redhat.com/en/about/press-releases/red-hat-ai-factory-nvidia-accelerates-path-scalable-production-ai
11. https://gpusmith.com/articles/en/hbm3e-vs-hbm4-ai-gpus
12. https://gpusmith.com/articles/en/nvlink-vs-infinity-fabric-gpu-interconnects
13. https://docs.nvidia.com/dgx/dgx-station-development-guide/porting/software-requirements.html
14. https://docs.redhat.com/en/documentation/red_hat_hardware_certification/2024/html-single/red_hat_hardware_certification_program_policy_guide/index
15. https://gpusmith.com/articles/en/llm-inference-hardware-sizing-guide
16. https://docs.nvidia.com/nim/large-language-models/2.0.5/troubleshooting/memory.html
17. https://arxiv.org/abs/2309.06180
18. https://nvidia.github.io/TensorRT-LLM/python-api/tensorrt_llm.functional.html
19. https://docs.vllm.ai/en/latest/serving/parallelism_scaling/
20. https://kserve.github.io/website/docs/model-serving/generative-inference/multi-node
21. https://www.hp.com/us-en/workstations/ai-stations/zgx-ai-stations-software.html
22. https://docs.redhat.com/en/documentation/red_hat_ai_enterprise/3/html-single/introduction_to_red_hat_ai_enterprise/introduction_to_red_hat_ai_enterprise
23. https://docs.nvidia.com/ai-enterprise/planning-resource/licensing-guide/latest/licensing.html
24. https://docs.nvidia.com/ai-enterprise/deployment/red-hat-ai-factory/latest/support.html
25. https://docs.redhat.com/en/documentation/red_hat_enterprise_linux/10/html-single/configuring_and_managing_high_availability_clusters/configuring_and_managing_high_availability_clusters
26. https://gpusmith.com/articles/en/gpu-cluster-acceptance-testing-runbook

27. https://www.dmtf.org/sites/default/files/standards/documents/DSP0266_1.23.0.html
28. https://redfish.dmtf.org/redfish/schema_index
29. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/800-171r3/NIST.SP.800-171r3.html
30. https://www.snia.org/node/13837
31. https://csrc.nist.gov/pubs/sp/1339/final
32. https://docs.redhat.com/en/documentation/red_hat_enterprise_linux/10/html/monitoring_and_managing_system_status_and_performance/overview-of-performance-monitoring-options
33. https://prometheus.io/docs/practices/alerting/
34. https://docs.redhat.com/en/documentation/red_hat_enterprise_linux/10/html/managing_monitoring_and_updating_the_kernel/applying-patches-with-kernel-live-patching
35. https://docs.redhat.com/en/documentation/openshift_container_platform/4.17/pdf/disconnected_environments/index
36. https://docs.nvidia.com/mission-control/docs/nmc-software-installation-guide/2.3.0/airgapped-installation/downloading-artifacts.html
37. https://www.hp.com/us-en/workstations/ai-stations.html
38. https://mlcommons.org/benchmarks/endpoints/
39. https://mlcommons.org/2024/03/mlperf-llama2-70b/
40. https://portfoliomanager.energystar.gov/pm/glossary
41. https://www.eia.gov/energyexplained/electricity/prices-and-factors-affecting-prices.php
42. https://ftp.spec.org/power/docs/specpower_ssj2008-run_reporting_rules/
43. https://framework.finops.org/assets/terminology/
44. https://focus.finops.org/docs/specification/v1-4/columns/cost-and-usage/effective-cost/
45. https://docs.cloud.google.com/architecture/framework/cost-optimization/align-cloud-spending-business-value
46. https://www.dell.com/en-us/shop/ipovw/poweredge-xe9685l
47. https://docs.aws.amazon.com/whitepapers/latest/how-aws-pricing-works/key-principles.html
48. https://aws.amazon.com/ec2/instance-types/p6/
49. https://learn.microsoft.com/en-us/azure/virtual-machines/sizes/gpu-accelerated/nd-gb300-v6-series
50. https://docs.cloud.google.com/compute/docs/gpus/gpu-network-bandwidth?hl=en
51. https://www.energystar.gov/products/data_center_equipment/16-more-ways-cut-energy-waste-data-center/use-sensors-and-controls
52. https://www.ashrae.org/File%20Library/Technical%20Resources/Publication%20Errata%20and%20Updates/2011-Gaseous-and-Particulate-Guidelines.pdf
53. https://docs.aws.amazon.com/wellarchitected/latest/reliability-pillar/availability.html
54. https://www.cisa.gov/topics/cyber-threats-and-advisories/sbom/sbomresourceslibrary
55. https://www.redhat.com/en/about/press-releases/red-hat-puts-safety-and-observability-core-enterprise-ai-red-hat-ai-35
56. https://gpusmith.com/articles/en/mlperf-inference-6-1-gpu-procurement
57. https://mlcommons.org/2026/04/mlperf-inference-v6-0-results/
58. https://docs.mlcommons.org/inference/submission/
59. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-2.ipd.pdf

60. https://www.hp.com/us-en/newsroom/blogs/2026/unlock-ai-with-zgx-fury.html

61. https://docs.cloud.google.com/compute/docs/gpus

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