GPU SMITH
COMPUTE ENGINEERING

GPU SMITH / RESEARCH & PRACTICAL GUIDANCE

NVIDIA AI Enterprise 8.2 vs 7.8 LTSB: Upgrade Decision

> 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 · 2026-09-25 · nvidia ai enterprise 8.2 vs 7.8 ltsb · nvidia ai enterprise · infrastructure 8.2 · infrastructure 7.8 · ltsb · gpu drivers · production upgrade · support matrix

Original article: https://gpusmith.com/articles/en/nvidia-enterprise-8-2-vs-7-8-ltsb

In this article

Executive Summary

As of September 25, 2026, the production choice between **NVIDIA AI Enterprise Infrastructure 8.2** and **7.8 Long-Term Support Branch (LTSB)** is primarily a choice of supported configuration and driver lifecycle. NVIDIA lists both as August 2026 releases. Infrastructure 8.2 is the **Production Branch** on **R595**, with a planned **April 2027** end of life. Infrastructure 7.8 is the **LTSB** on **R580**, with a planned **July 2028** end of life. Counting calendar months from publication, those plans imply roughly **7** and **22 months** of runway, respectively, a **15-month** difference; NVIDIA gives EOL months rather than exact days and can revise its schedule. (docs.nvidia.com) (docs.nvidia.com)

The higher release number does not mean a uniformly newer stack. Both version-pinned matrices list **DOCA Driver 3.4.0**, **Container Toolkit 1.19.1**, **Run:ai 2.26**, **DPU Operator 26.4.0**, **GPU Operator 26.3.3**, **Network Operator 26.4.1**, **NIM Operator 3.1.2**, and **Base Command Manager 11.33.1**. The consequential delta is the host, virtual GPU (vGPU), guest-driver and Fabric Manager path. NVIDIA lists data-center driver **595.91.07** and vGPU Manager **595.91.04** for 8.2, versus **580.178.04** and **580.178.05** for 7.8. (docs.nvidia.com) (docs.nvidia.com)

Select **8.2** when a needed, explicitly supported R595 hardware or virtualization configuration justifies a nearer migration date. Select **7.8 LTSB** when its exact matrix covers the deployed stack and a longer planned support window has greater value. **Hold** when neither target has a documented row and footnote set for the complete deployment, or when a reproducible canary and rollback cannot fit the change window. The matrices show that 8.2 does **not** automatically lift the single-node KVM wording, Grace platform bare-metal restriction, or HGX GPU Operator restriction with KubeVirt and OpenShift Virtualization. Both branches exclude Red Hat Enterprise Linux (RHEL) 9.7. (docs.nvidia.com) (docs.nvidia.com)

The practical upgrade gate is an evidence pack: current and proposed versions, the exact matrix row and footnotes, hypervisor and operating-system approvals, test outputs, pass criteria, rollback artifacts, a named owner, and confirmed support entitlement. NVIDIA's published 8.2 checklist addresses **8.1 to 8.2**, so a **7.8-to-8.2** project needs an explicit cross-branch plan rather than assuming that checklist is a direct recipe. The

independent engineering perspective is to validate the configured workload and fabric against written acceptance criteria, then choose the branch that passes those criteria within its lifecycle. (docs.nvidia.com) (gpusmith.com)

Introduction and Background

Private GPU estates rarely upgrade as a single software package. A production node may combine firmware, a host operating system, an NVIDIA data-center driver, a hypervisor, vGPU host and guest components, Kubernetes, operators, network adapters, container runtimes, model-serving containers, and observability. A release number can label a coherent support bill without guaranteeing that every adjacent product's version is interchangeable. This report therefore treats **8.2 versus 7.8 LTSB** as a configuration decision, not a feature contest. NVIDIA's own release notes direct readers to a support matrix for platforms, hypervisors, operating systems and orchestration software, and to an explorer for full-stack checks. (docs.nvidia.com)

The audience includes cluster owners, virtualization teams, security and change-control reviewers, and subscription owners. Each needs a different answer: whether the desired hardware works, whether a host and guest pair is supported, whether the orchestration layer can be maintained safely, and whether support continues through the organization's next maintenance window. A purchase or entitlement does not certify a configuration. NVIDIA licenses AI Enterprise on a **per-GPU** basis and describes production support as part of a subscription; the matrix separately defines technical compatibility. (docs.nvidia.com)

Terminology is a first source of confusion. NVIDIA divides AI Enterprise into **application software** and **infrastructure software**, each with independent branches and lifecycles. Application Feature Branch (FB), Production Branch (PB), and LTSB language must not be used to infer the driver and operator support date. NVIDIA's infrastructure policy describes ordinary driver branch support and separately extends designated infrastructure LTSBs. The release index, rather than a generic statement about application branches, supplies the planned EOL months for **Infra 8** and **Infra 7 LTSB**. (docs.nvidia.com)

GPU Smith describes its role as independent engineering for private AI infrastructure, including specification, integration and validation against written acceptance criteria. In this comparison that is an advisor's method, not a third branch or a claim to sell NVIDIA software: require a version-pinned acceptance record before changing the production stack. (gpusmith.com)

NVIDIA AI Enterprise Infrastructure 8.2 Production Branch

Capabilities

Infrastructure **8.2** is NVIDIA's **R595 Production Branch** release in the August 2026 index. Its notes call the branch a standard-cadence PB with feature additions and “quarterly support.” The lifecycle documentation separately describes an infrastructure minor-release cadence of roughly **three months** and says actual timing is guidance. The phrase in the release note should therefore be read in context; it is not evidence that application PB security updates, infrastructure releases, and a particular security-fix schedule are identical. The application-branch page describes **monthly** PB and **quarterly** LTSB security updates for that different layer. (docs.nvidia.com)

The version-pinned 8.2 support matrix lists **GPU driver 595.91.07**, **Fabric Manager 595.91.07**, **vGPU Manager 595.91.04**, **Linux vGPU guest driver 595.91.07**, and **Windows guest driver 596.86**. NVIDIA's 8.2 release notes say Fabric Manager binaries are included in the AI Enterprise drivers and no longer require separate installation. Packaging and operations thus change even when an operator version stays constant; an inventory should record the installed host packages and their service behavior, not only a cluster-level release label. (docs.nvidia.com)

The release highlights describe vGPU for Compute on **HGX B200** and **HGX B300** with VMware vSphere, but limit that mode to **1:1** and multi-vGPU virtual machines; fractional vGPU virtual machines are excluded. The 8.2 matrix also qualifies certain **RHEL 10.0, 10.1 and 10.2** vGPU guests on VMware ESXi to **ESXi 9 Update 1**. These are examples of why a GPU name or broad hypervisor family is insufficient evidence of a supported deployment. Verify profile, host release, guest release and all footnotes together. (docs.nvidia.com)

Adoption and fit

There is no public, representative adoption count in the official sources reviewed for this report. The practical adoption question is narrower: does **R595** unlock a configuration the estate actually needs, and can that configuration be certified before the planned **April 2027** EOL? A team already on **8.1** has an NVIDIA checklist explicitly written for **8.1-to-8.2** upgrades. A team on **7.8** should not treat that document as a direct cross-branch migration guide. (docs.nvidia.com)

Where R595 enables a needed virtualization path, the change case should show a matrix row for the exact GPU, hypervisor, guest and orchestration layer. The case should also name the workload that benefits, the canary that proves it, and the date by which the estate will leave the branch. Kubernetes maintenance guidance recommends draining nodes to evict workloads while respecting disruption budgets; that is an operational precondition, not a substitute for NVIDIA compatibility proof. (docs.nvidia.com) (kubernetes.io) (kubernetes.io) (docs.redhat.com) (knowledge.broadcom.com) (docs.redhat.com)

Strengths and limitations

- **Driver path:** R595 is the branch-specific reason to consider 8.2 when a required device or vGPU mode appears only in its supported configuration.
- **Operational change:** Bundled Fabric Manager binaries simplify the package list, but teams must test service start, GPU fabric enumeration and rollback of the actual package set.
- **Time horizon:** The index's planned April 2027 EOL gives less calendar runway than 7.8 LTSB as of publication.
- **Footnotes:** Its KVM single-node, Grace bare-metal and HGX virtualization limits remain material.
- **Host selection:** RHEL 9.7 is absent from the supported host hypervisor and guest lists.

NVIDIA AI Enterprise Infrastructure 7.8 LTSB

Capabilities

Infrastructure **7.8** is the **R580 LTSB** release shown in NVIDIA's August 2026 index. Its planned **July 2028** EOL makes it the longer-runway option at publication. That date is a maintained vendor plan, not a guarantee that every component or every adjacent operating system will remain supported until the same day. NVIDIA's policy says an infrastructure driver branch is normally supported for **one year** and a designated infrastructure LTSB for **three years**, while the index provides the specific planned date to use in a change record. (docs.nvidia.com)

The 7.8 matrix lists **GPU driver and Fabric Manager 580.178.04**, **vGPU Manager 580.178.05**, **Linux guest driver 580.178.04**, and **Windows guest driver 582.78**. NVIDIA's release highlights describe the driver and Fabric Manager moving in lockstep. This does not imply that every running virtual machine can be moved across host driver families without a planned guest-driver and licensing sequence. Record the current host and guest pair and test the actual VM operations used in production. (docs.nvidia.com)

The same version-pinned matrix lists many software components at the same release numbers as 8.2, including the GPU, Network, DPU and NIM operators, Container Toolkit, DOCA, Run:ai and Base Command Manager. That overlap narrows the question: if the workload does not need an R595-only capability, 7.8 can preserve a longer planned driver support window without forgoing those listed component versions. The shared number, however, does not override platform support rows or footnotes. (docs.nvidia.com)

Adoption and fit

No reliable public adoption split between these two infrastructure releases was verified. A 7.8 recommendation should rest on the estate's inventory and its support proof. The attractive cases are stable, supported combinations whose maintenance policy places high value on a longer branch horizon. The branch label is not permission to freeze all dependencies: an organization still needs to track operating-system, hypervisor, orchestration and application lifecycles independently. NVIDIA expressly distinguishes its infrastructure and application layers. (docs.nvidia.com)

For virtualized deployments, **7.8** has explicit constraints that often dominate the decision. Its release notes say vGPU for Compute on Linux KVM is supported only on a **single node**. Grace Hopper and Grace Blackwell are **bare metal only**, and GPU Operator is not supported with **KubeVirt** or **OpenShift Virtualization** on HGX. The 8.2 static matrix contains corresponding boundaries, so these are not automatic reasons to leave 7.8. (docs.nvidia.com)

Strengths and limitations

- **Support horizon:** NVIDIA lists a planned July 2028 EOL for the R580 LTSB line.
- **Component parity:** Several named operator and toolkit releases match 8.2.
- **Virtualization boundary:** The KVM single-node wording and HGX operator restriction require configuration-level review.
- **Hardware boundary:** Grace Hopper and Grace Blackwell are bare-metal only in the matrix.
- **Host selection:** RHEL 9.7 remains unsupported after its earlier removal.

Feature Comparison

Table 1 compares the supported infrastructure bill as published in the version-pinned matrices. A matching component number describes the bundled version; it does not certify every platform or application combination. The table separates host and guest drivers because an upgrade of the hypervisor host alone can leave a production VM on a different compatibility path. (docs.nvidia.com) (docs.nvidia.com)

Component	Infra 8.2 PB	Infra 7.8 LTSB	Decision consequence
GPU driver and Fabric Manager	R595, 595.91.07; Fabric Manager binaries bundled with driver.	R580, 580.178.04 for both driver and Fabric Manager.	Inventory host packaging, fabric service and rollback artifacts.
vGPU Manager and guest drivers	Manager 595.91.04; Linux guest 595.91.07; Windows guest 596.86.	Manager 580.178.05; Linux guest 580.178.04; Windows guest 582.78.	Test the installed host-guest pair and every production VM operation.
DOCA and Container Toolkit	DOCA Driver 3.4.0; Toolkit 1.19.1.	DOCA Driver 3.4.0; Toolkit 1.19.1.	Equal listed versions do not erase NIC, OS or runtime footnotes.
Scheduling and operators	Run:ai 2.26; DPU 26.4.0; GPU 26.3.3; Network 26.4.1; NIM 3.1.2.	Run:ai 2.26; DPU 26.4.0; GPU 26.3.3; Network 26.4.1; NIM 3.1.2.	Validate operator health and a representative scheduled job.
Cluster management	Base Command Manager 11.33.1.	Base Command Manager 11.33.1.	Check provisioning and node-recovery workflows on the candidate branch.

The table points to a useful test-design economy. The shared component versions permit reuse of some application tests, but the changed driver family still requires fresh GPU discovery, scheduler allocation, CUDA workload, network and virtualization checks. NVIDIA's own release notes describe compatibility verification as a full-stack exercise, and the matrices are version pinned precisely because individual rows carry platform qualifications. (docs.nvidia.com)

Table 2 is a support-boundary checklist, not a replacement for a configuration search. “Listed” means the branch matrix has relevant rows; the buyer must still capture the exact GPU, host, guest, operating-system, operator, runtime and footnote combination. Where the two tables show similar boundaries, switching branches alone is not a remediation. (docs.nvidia.com) (docs.nvidia.com)

Deployment question	8.2 evidence	7.8 evidence	Acceptance condition
Bare-metal Kubernetes and OpenShift	Matrix lists upstream Kubernetes 1.32 to 1.36 and OpenShift 4.18 to 4.22, subject to OS/operator columns.	Same displayed orchestration ranges, subject to OS/operator columns.	Save the exact row, minor release, runtime and all footnotes.
KVM and vGPU	RHEL with KVM virtualized row says single-node only.	Matrix and release notes state single-node support for KVM workloads.	Do not infer supported multi-node KVM from GPU visibility alone.

Deployment question	8.2 evidence	7.8 evidence	Acceptance condition
Grace Hopper and Grace Blackwell	GH200, GB200 and GB300 platform rows are bare-metal only.	GH200, GB200 and GB300 are limited to bare-metal deployments.	Match the exact platform variant and deployment model.
HGX with KubeVirt or OpenShift Virtualization	GPU Operator unsupported for this combination.	Same matrix restriction.	Design a documented alternative before approving a rollout.
VMware vGPU	Matrix lists ESXi 8.0 and later and 9.0 and later, with additional guest and profile qualifications.	Matrix lists those ESXi families with branch-specific qualifications.	Record hypervisor build, vGPU profile, guest driver and migration method.
Network and Government Ready	ConnectX and BlueField families are listed; Government Ready varies by component and architecture.	ConnectX and BlueField families are listed; Government Ready varies by component and architecture.	Preserve networking row, RoCE requirement and component-level Government Ready cells.

These rows are deliberately conditional. The matrices state that multi-node setups need an Ethernet network interface controller (NIC) supporting **RDMA over Converged Ethernet (RoCE)**; a present GPU cannot establish network support. The OpenShift patch-release footnote also depends on continued Red Hat support. The final approval record should contain both NVIDIA's row and the platform vendor's active support statement. (docs.nvidia.com) (docs.redhat.com) (docs.redhat.com) (kubernetes.io) (kubernetes.io) (knowledge.broadcom.com)

Performance and Benchmarks

No controlled public benchmark in the sources reviewed isolates **R595 in 8.2** against **R580 in 7.8** on otherwise identical hardware, firmware, network, model and workload. It would be unsound to infer a throughput gain from the version number or from the shared operator versions. A production benchmark should instead ask whether the intended configuration meets a predeclared service objective and whether the candidate changes latency, throughput, error rate or recovery time relative to the currently accepted baseline. Prometheus recommends high-level latency and error-rate alerts, while OpenTelemetry describes a vendor-neutral collector path for telemetry; neither source is an NVIDIA branch performance claim. (prometheus.io) (prometheus.io) (opentelemetry.io) (kubernetes.io) (opentelemetry.io) (opentelemetry.io) (prometheus.io) (prometheus.io)

- **GPU discovery:** Record device count, model, memory, topology, driver and Fabric Manager status before and after the change.
- **Scheduling:** Place one representative multi-GPU job and one inference service; record pending time, allocation and cleanup.
- **Network:** For distributed jobs, measure RDMA availability and the expected fabric path on the actual supported NIC.
- **Virtualization:** Where used, boot each guest-driver pair, exercise the configured vGPU profile and run the site's migration or failover procedure. (www.libvirt.org) (www.libvirt.org) (knowledge.broadcom.com) (learn.microsoft.com)

- **Serving:** Replay a fixed prompt and model mix with stated concurrency; compare throughput, tail latency and errors under the same acceptance thresholds. (prometheus.io) (prometheus.io) (opentelemetry.io) (kubernetes.io)
- **Observability:** Verify that metrics and alerts survive the node and operator sequence, not merely that a container reports Ready. (opentelemetry.io) (opentelemetry.io) (prometheus.io) (prometheus.io) (prometheus.io) (prometheus.io) (opentelemetry.io) (kubernetes.io)
- **Rollback:** Restore the captured host and guest state in a lab and repeat the same acceptance tests.

Benchmark results should name workload, input distribution, GPU SKU, memory, CPU, network, software build, warm-up and measurement window. This is an evidence specification, not a claim that either branch is faster. If only a marketing demonstration or an unmatched before-and-after run is available, mark performance as **unmeasured** in the scorecard. A neutral result can still favor 7.8 when lifecycle matters more; a measured workload benefit can justify 8.2 only when the support configuration and maintenance plan also pass.

Data Analysis and Evidence

The most defensible quantitative comparison is the vendor's dated branch table. On **September 25, 2026**, an **April 2027** EOL is seven calendar months away, while **July 2028** is twenty-two calendar months away. The relative runway is **fifteen calendar months**. Because NVIDIA publishes EOL months, these are month differences, not precise remaining service days. The index also says the release date for each is **August 2026**; the release-note “last updated” stamps of **September 2** for 8.2 and **August 12** for 7.8 are document dates and should not be relabeled as product launch days. (docs.nvidia.com)

That longer 7.8 horizon is one input, not a universal score. A weighted decision can be made reproducible without pretending that a subjective weighting is measured product performance. Define five buyer-chosen weights that sum to **100**: **configuration support**, **required capability**, **lifecycle runway**, **change-window feasibility**, and **rollback confidence**. Score each candidate **0** if the requirement fails, **1** if evidence is incomplete, and **2** if the requirement is documented and tested. The weighted result is the sum of weight multiplied by score divided by **2**. A candidate with a failed mandatory support row is rejected regardless of arithmetic. These weights and scores are a decision method proposed here, not NVIDIA-published data.

- **Configuration support:** Mandatory. Capture the exact matrix row and every numbered footnote.
- **Required capability:** Name the hardware, vGPU profile, guest or operator behavior that motivates change.
- **Lifecycle runway:** Use the vendor-maintained EOL month, then compare it with the enterprise's next approved migration window.
- **Change-window feasibility:** Plan node draining before maintenance. (kubernetes.io)
- **Rollback confidence:** Retain previous versions of baseline configurations to support rollback. (NIST SP 800-53)

(Hypothetical Example). An estate might assign weights of **35**, **20**, **20**, **15** and **10** in that order. If both branches pass configuration and workload tests, but the organization cannot schedule another migration before April 2027, its own lifecycle and change-window scores will tend toward 7.8. If a required, supported

R595 vGPU mode is absent from 7.8, the mandatory capability gate can favor 8.2 despite the shorter planned runway. The example is a scoring illustration, not a benchmark or a recommendation for an unspecified estate. (docs.nvidia.com)

The evidence must be dated. NVIDIA's explorer was updated with **8.2** and **7.8** data in **August 2026**, and its documentation says to check the complete stack from the GPU driver. Store the query inputs and results with the change record. A source-linked matrix copied into a ticket can otherwise outlive the page's support status. NIST's configuration-management guidance supports monitoring system configuration as an ongoing process; SLSA's provenance model provides a useful vocabulary for preserving the origin of software artifacts. (docs.nvidia.com) (csrc.nist.gov) (csrc.nist.gov) (slsa.dev) (www.suse.com) (documentation.ubuntu.com) (slsa.dev) (csrc.nist.gov) (www.libvirt.org) (containerd.io) (csrc.nist.gov)

Implications and Future Directions

The branch choice should be revisited when a new supported GPU, host release, hypervisor, operating system or orchestration minor release becomes necessary, and when the planned EOL changes. NVIDIA calls its cadence general guidance and reserves discretion to change it. A calendar reminder based only on the current **April 2027** or **July 2028** row is weaker than a periodically refreshed compatibility record. (docs.nvidia.com)

An upgrade owner should run a staged gate with the current configuration as the control. The 8.2 NVIDIA checklist covers **8.1 to 8.2**; it calls for a compatibility review and a per-component rollback path. For **7.8 to 8.2**, the team should map each host, guest, operator and application transition and confirm the sequence with its support channel for the specific estate. Kubernetes documents node draining and disruption budgets, while Red Hat's OpenShift guidance emphasizes spare worker capacity and cluster upgradeability. Those are concrete reasons to reserve a canary pool and an abort point before touching the full cluster. (docs.nvidia.com) (kubernetes.io) (docs.redhat.com)

Table 3 is a compact evidence ledger to attach to the change record. The cells marked “site value” must be filled from the buyer's own inventory and tests. A blank support proof or rollback field is a failed gate, even when the target branch is newer. (docs.nvidia.com) (csrc.nist.gov) (csrc.nist.gov) (slsa.dev) (www.suse.com) (documentation.ubuntu.com)

Item and owner	Current and candidate	Support proof	Test and pass criterion	Rollback artifact
Host GPU and fabric, platform owner	Site driver, firmware, Fabric Manager to target R595 or R580.	Exact GPU, OS and fabric rows with footnotes.	GPU count, topology, fabric and distributed-job result match site baseline.	Signed packages, config export, boot image. (slsa.dev) (csrc.nist.gov)
Virtualization, VM owner	Site hypervisor, vGPU Manager and guest pair to target pair.	Host, profile, guest and migration qualification .	Boot, license check and every used VM move succeed.	Host and guest image plus VM configuration. (www.libvirt.org) (www.libvirt.org)

Item and owner	Current and candidate	Support proof	Test and pass criterion	Rollback artifact
Orchestration, cluster owner	Site Kubernetes, OpenShift and operators to supported versions.	Matrix row, runtime and patch-support condition.	Drain, reschedule, GPU allocation and operator recovery pass.	Manifests and images; take an etcd backup before updating the cluster. (docs.redhat.com)
Serving and telemetry, workload owner	Site NIM or other serving image and metrics pipeline to candidate.	Image provenance record. (slsa.dev)	Replay a representative serving request mix at recorded concurrency; pass only when measured p95/p99 latency and error rate meet the site-defined thresholds. Alert on high latency and error rates. (prometheus.io)	Prior image digest, model and alert configuration. (opentelemetry.io) (opentelemetry.io) (prometheus.io) (prometheus.io)
Entitlement and sign-off, support owner	Site subscription and candidate support case.	Valid entitlement plus written configuration answer.	Support contact and rollback trigger named in ticket.	Archived approvals and case reference. (csrc.nist.gov)

The ledger also prevents a common category error: a valid software entitlement and a passing GPU smoke test are useful evidence, but neither proves that a specific host and guest pairing is on the support matrix. Conversely, a fully supported matrix row does not prove the organization's serving objective. The change decision requires both, along with a tested rollback path. GPU Smith's own described method uses written acceptance and as-built records; the same discipline is appropriate here without treating the consultancy as a competing software option. (docs.nvidia.com) (gpusmith.com)

Upgrade sequence and hold criteria

- **Freeze inventory:** Export running versions, GPU and NIC identifiers, VM profiles, cluster distribution, operator charts, container digests and firmware. (slsa.dev) (csrc.nist.gov) (www.libvirt.org) (containerd.io) (csrc.nist.gov)
- **Verify support:** Save the version-pinned matrix row, footnotes, lifecycle explorer result and hypervisor or operating-system vendor approval.
- **Protect availability:** Configure a PodDisruptionBudget to help keep workloads available during maintenance. (kubernetes.io)
- **Rehearse:** Restore one representative node and VM from the saved artifacts before the maintenance window. (www.libvirt.org) (www.libvirt.org) (knowledge.broadcom.com) (learn.microsoft.com)
- **Canary:** Change the smallest workload-bearing slice that exercises the production topology; capture GPU, RDMA, scheduling, serving and alert results. (prometheus.io) (prometheus.io) (opentelemetry.io) (kubernetes.io)
- **Expand or stop:** Roll forward only when the written pass criteria hold; otherwise execute the recorded rollback. (csrc.nist.gov) (csrc.nist.gov) (slsa.dev) (www.suse.com) (documentation.ubuntu.com)
- **Close the record:** Retain records of configuration-controlled changes for an organization-defined period. (NIST security controls)

- **Schedule review:** Review and update the system component inventory on an organization-defined schedule. (NIST security controls)

Frequently Asked Questions (FAQs)

Is 8.2 simply the newer and therefore preferred branch?

No. Its infrastructure driver is **R595**, while 7.8 uses **R580**; several operators and toolkits have the same listed versions. The supportable decision depends on exact matrix rows and the planned EOL. A site with no R595-only requirement may rationally prefer the longer 7.8 LTSB runway. (docs.nvidia.com)

Can a production estate upgrade directly from 7.8 LTSB to 8.2?

The official 8.2 checklist verified here is written for **8.1 to 8.2**. It is useful as a list of checks, but does not itself certify a **7.8-to-8.2** migration. Build a cross-branch component map, validate host and guest sequencing, rehearse rollback and obtain a support answer for the configured environment before a production change. (docs.nvidia.com)

Does 8.2 remove 7.8's KVM, Grace and HGX limits?

The 8.2 matrix still labels RHEL/KVM virtualized rows **single-node only**, GH200/GB200/GB300 platforms **bare metal only**, and GPU Operator with KubeVirt or OpenShift Virtualization on HGX **unsupported**. Read the exact row and footnote because the scope of a limit matters. (docs.nvidia.com)

What happens to RHEL 9.7 and support entitlement?

Both release-note sets exclude **RHEL 9.7** from relevant supported host or guest lists. A subscription is separately licensed on a per-GPU basis and includes production support during its term; it does not make an excluded operating-system combination supported. Check the technical matrix and the actual contract together. (docs.nvidia.com) (docs.nvidia.com)

Conclusion

The **8.2 versus 7.8 LTSB** decision is a versioned support decision. Both arrived in **August 2026**, but 8.2 pairs **R595** with a planned **April 2027** EOL, while 7.8 pairs **R580** with a planned **July 2028** EOL. Many named operator and toolkit versions match. Choose 8.2 for a required, proven R595 configuration that can be operated and migrated within the shorter horizon. Choose 7.8 when its exact matrix covers the workload and the longer lifecycle has greater operational value. Hold if support proof, canary evidence or rollback cannot be produced. (docs.nvidia.com)

Before approval, preserve the exact support row and footnotes, current and candidate versions, entitlement, representative workload measurements and an exercised reversal path. Recheck the living NVIDIA index and matrices at the change date. A decision that can be reproduced from those artifacts will remain explainable when the next branch, host release or workload arrives. (docs.nvidia.com) (csrc.nist.gov) (csrc.nist.gov) (slsa.dev) (www.suse.com) (documentation.ubuntu.com)

Source links

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

1. https://docs.nvidia.com/ai-enterprise/index.html
2. https://docs.nvidia.com/ai-enterprise/release-8/latest/support/support-matrix-8/8.2.html
3. https://docs.nvidia.com/ai-enterprise/release-7/latest/support/support-matrix-7/7.8.html
4. https://docs.nvidia.com/ai-enterprise/release-8/latest/releases/upgrading.html
5. https://gpusmith.com/
6. https://docs.nvidia.com/ai-enterprise/release-8/latest/overview/release-notes-8/8.2.html
7. https://docs.nvidia.com/ai-enterprise/planning-resource/licensing-guide/latest/licensing.html
8. https://docs.nvidia.com/ai-enterprise/release-8/latest/index.html
9. https://kubernetes.io/docs/tasks/administer-cluster/safely-drain-node/
10. https://kubernetes.io/docs/concepts/workloads/controllers/deployment/
11. https://docs.redhat.com/en/documentation/openshift_container_platform/4.21/html/backup_and_restore/control-plane-backup-and-restore
12. https://knowledge.broadcom.com/external/article/427499/cns-volume-attachment-issues-on-kubernet.html
13. https://docs.redhat.com/en/documentation/openshift_container_platform/4.20/html/updating_clusters/performing-a-cluster-update
14. https://docs.nvidia.com/ai-enterprise/release-7/latest/overview/release-notes-7/7.8.html
15. https://prometheus.io/docs/practices/alerting/
16. https://prometheus.io/docs/prometheus/latest/configuration/configuration/
17. https://opentelemetry.io/docs/collector/
18. https://opentelemetry.io/docs/platforms/kubernetes/collector/
19. https://www.libvirt.org/migration.html
20. https://www.libvirt.org/kbase/snapshots.html
21. https://learn.microsoft.com/en-us/windows-server/virtualization/hyper-v/gpu-partitioning
22. https://csrc.nist.gov/CSRC/media/Projects/risk-management/800-53%20Downloads/800-53r5/SP_800-53_v5_1-derived-OSCAL.pdf
23. https://docs.nvidia.com/ai-enterprise/lifecycle/latest/index.html
24. https://csrc.nist.gov/pubs/sp/800/128/upd1/final
25. https://csrc.nist.gov/glossary/term/configuration_control
26. https://slsa.dev/spec/v1.2/provenance
27. https://www.suse.com/lifecycle/
28. https://documentation.ubuntu.com/security/security-features/platform-protections/secure-boot/
29. https://containerd.io/docs/2.1/cri/config/
30. https://docs.nvidia.com/ai-enterprise/lifecycle/latest/choosing-a-branch.html
31. https://www.nccoe.nist.gov/publication/1800-16/VolB/vol-b-appendix.html

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