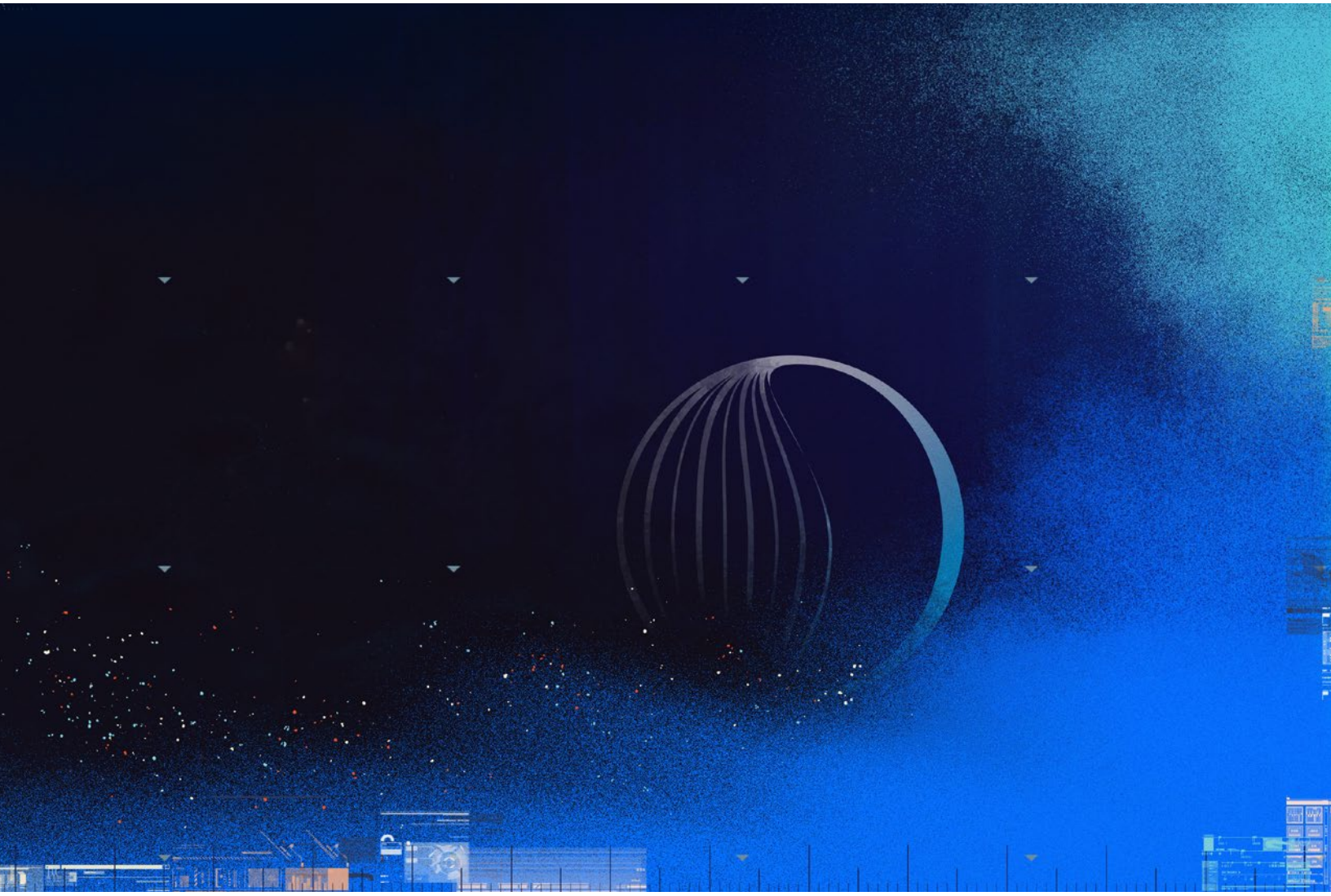
signal65



Enterprise Inference Considerations on Dell PowerEdge XE7745

Balancing Response Time and Memory with DeepSeek and Gemma 4

AUTHOR

Brian Martin
AI Data Center Performance | Signal65

IN PARTNERSHIP WITH

DELL Technologies

OCTOBER 2026

Executive Summary

Enterprise inference services do more than load models onto a server. They must answer the right questions within an acceptable time while handling expected peak demand. How GPUs are assigned to model instances can affect that service, even when the hardware and model stay the same.

Signal65 evaluated two deployment strategies for DeepSeek-V4.1-Flash on Dell PowerEdge XE7745 with eight NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, networked through Broadcom-based Dell PowerSwitch Z9864F-ON switches running Enterprise SONiC Distribution by Dell Technologies. Two independent four-GPU instances favored concurrent work. One eight-GPU instance provided more memory capacity for active contexts and shorter wait times for individual long prompts.

Key Highlights

37%

More output throughput with two instances at 64 streams vs a single instance

48%

Less elapsed time for 12-session million-token workloads than a single instance

1.6–1.7x

More context capacity with one instance relative to two at the same context length

The two-instance layout delivered 2,756 output tokens per second at 64 streams, with lower tail latency, compared with 2,017 for the full-node instance. Concurrent long-context testing showed the dual-instance configuration completing the million-token workload in 738 seconds compared with 1,408 seconds for the single instance on the same hardware. These results make the split layout a strong starting point when deploying busy shared services.

One primary tradeoff is memory; the full-node instance retained more space for active requests and used half as much host memory for DeepSeek Engram tables. Buyers supporting many persistent long-context workloads should weigh that headroom against the higher output rate and lower latency in a dual-instance configuration.

For less demanding applications, Gemma 4 adds smaller-model deployment options. Single-GPU variants served successfully, and eight quantized 26B instances coexisted on a single node. At the time of testing, Gemma 4 compiles and serves cleanly on stock vLLM, while DeepSeek-V4.1-Flash required custom configuration (localinferencelab/vllm:jovian-judgement-community-20260914-r38).

It remains essential to evaluate the complete application and service configuration alongside the server and network. Model quality, instance layout, context limits, routing and the serving image determine how much useful work the system can support. A representative acceptance workload can guide the tradeoffs before deployment.



© 2026 Signal65. All rights reserved.

Why Deployment Layout Matters

An eight-GPU system gives an enterprise several ways to deploy inference services. One large instance can spread a model across the entire node. A model that fits on fewer GPUs can instead run in multiple instances, allowing separate schedulers to work on different requests. The better choice depends on how requests arrive and how much memory each conversation retains.

In this evaluation, both layouts used the same DeepSeek-V4.1-Flash model, serving image and eight GPUs. Tensor parallelism divides a model across GPUs: TP4 uses four GPUs per instance, while TP8 uses eight. The split layout ran two TP4 instances behind a router; the full-node layout ran one TP8 instance.

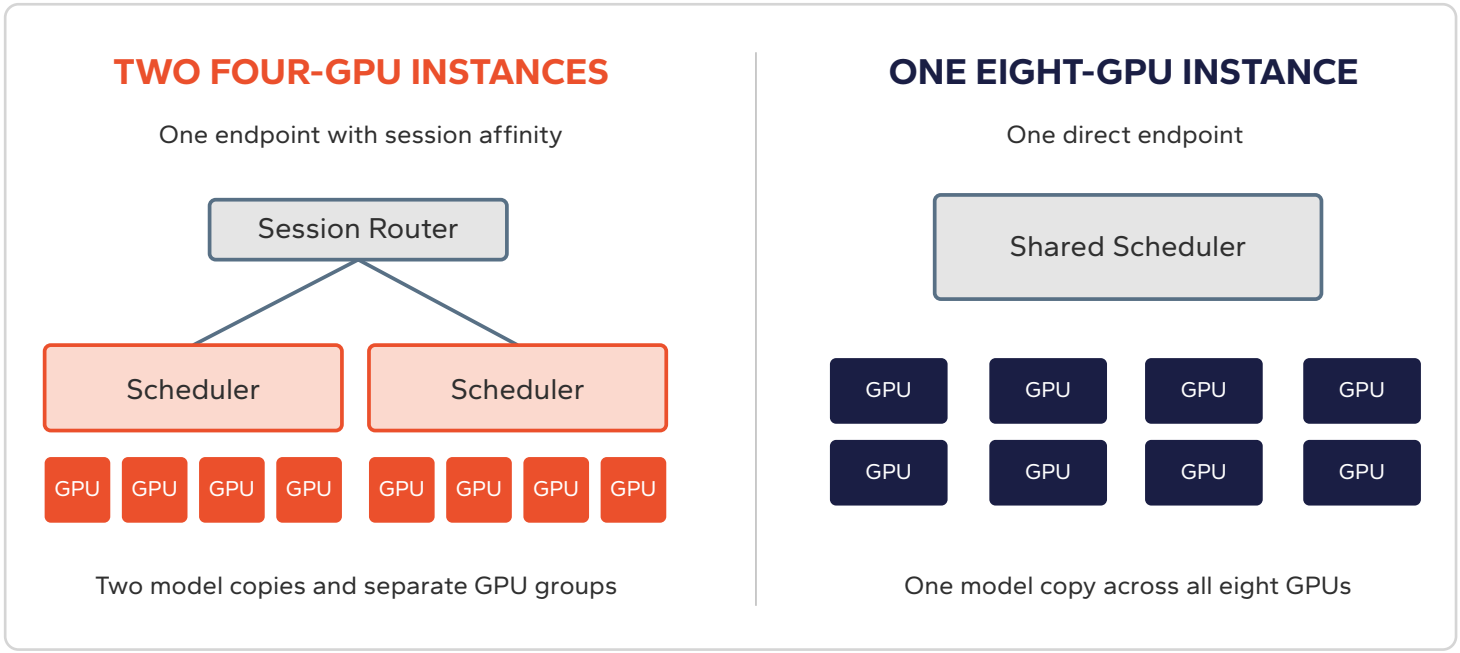


Figure 1: *Two deployment layouts on the same eight-GPU server*

Independent Instances Change Scheduling

Each four-GPU instance can process a different prompt while the other is busy. That independence matters when multiple users submit large documents together. A router keeps a conversation attached to the instance holding its cached context, so follow-up turns can reuse the cached prefix.

The split design creates separate memory pools for each instance. A request assigned to one instance cannot take advantage of unused context memory in the other. Monitoring must include the queues and memory usage of all instances, and admission controls must account for uneven demand.

One Instance Changes the Memory Balance

A single eight-GPU instance distributes one model copy across more accelerators, leaving more memory per GPU for request state. On this host, that instance spanned both CPU sockets; the split layout kept each instance with its local group of GPUs and host resources.

These differences explain why GPU count alone is an incomplete capacity specification. The proposed configuration should include the instance layout and routing behavior, with the workload used to establish its supported operating limits.

© 2026 Signal65. All rights reserved.

Test Configuration and Scope

The measurements were collected in September and October 2026 in Signal65's two-node Dell PowerEdge XE7745 cluster. Each deployment layout used the eight GPUs of one server, with the same DeepSeek model and serving image.

Each XE7745 combines two AMD EPYC 9555 processors, approximately 2.3 TB of DDR5 memory and eight NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs. Each server includes eight Broadcom BCM57608 400GbE network controllers with RoCEv2 support. The lab network uses Dell PowerSwitch Z9864F-ON switches built on Broadcom Tomahawk 5 silicon and running Enterprise SONiC 4.4.0.

Recorded Inference Configuration

Component	Recorded configuration
Accelerators	8 × RTX PRO 6000 Blackwell Server Edition 97,887 MiB reported per GPU; ECC on; no NVLink
CPU and host memory	2 × AMD EPYC 9555; 64 cores per socket 2,267 GB reported host memory
Operating environment	Ubuntu 24.04; Linux 6.8.0-87; NVIDIA driver 615.71.09
DeepSeek model	deepseek-ai/DeepSeek-V4.1-Flash
DeepSeek serving image	localinferencelab/vllm:jovian-judgement-community-20260914-r38
GPU and CPU placement	GPUs 0–3 on NUMA node 0; GPUs 4–7 on node 1 Split instances pinned to their local CPU sets
Baseline serving settings	FP8 KV cache; GPU memory utilization 0.95; batch budget 4,096 One parallel prefill per instance; decode context parallelism 1
Request and context settings	64 sequences per TP4 instance; 128 for TP8 Context ceilings 131,072 and 1,048,576 tokens

Measurement Design

Matched output and long-context results report medians of three deployment restarts. Selected tuning and GSM8K checks used six restarts per layout or setting. Long-context workloads used 12 sessions and 24 requests, with 1,024-token replies. GSM8K used the first 250 test questions, zero-shot, at temperature 0 and seed 0 with a 2,048-token answer allowance. The targeted truncation recheck used 8,192 tokens.

The split router preserved session affinity and aggregated measurements from both backends. A change of load-generation host was verified with bridging runs; the largest difference across four control cells was 0.59%. Two later corrections to the test journal apply: the split layout's 128-stream cell was skipped due to insufficient GPU memory, and the prompt budget used for the near-ceiling recall check was corrected.

© 2026 Signal65. All rights reserved.

Concurrency Changes the Value of the Server

A shared assistant or document service spends much of its time handling overlapping requests. The split DeepSeek deployment delivered more output from the same eight GPUs. It led at every matched stream count from one through 64, with its advantage reaching 37% at 64 simultaneous streams. The split configuration did not support more than 64 streams, while the single instance did Context and Answer Length Share the Budget.

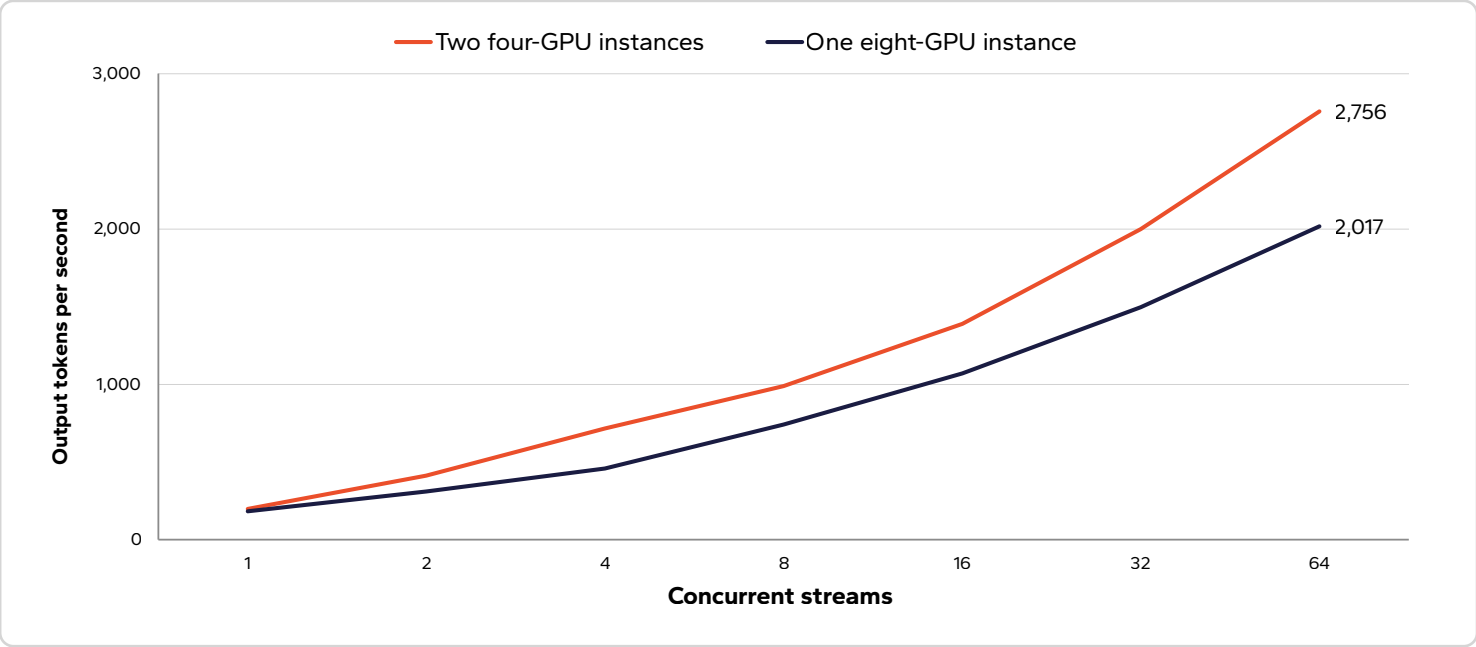


Figure 2: *Aggregate output throughput as simultaneous streams increase*

Short-prompt output generation with a 131,072-token context ceiling. Values are medians of three restarts; the horizontal axis doubles at each step. Source: Signal65 matched configuration tests.

Response at 64 simultaneous streams	Two four-GPU instances	One eight-GPU instance
Output throughput	2,756 tokens/s	2,017 tokens/s
p99 time to first token	0.56 s	1.28 s
p99 gap between output tokens	24.3 ms	33.5 ms

Throughput and User Experience Improved Together

Higher aggregate throughput is useful only if individual requests remain responsive. Here, the split layout also improved the two waits a user experiences: the delay before an answer begins and the gaps as the answer arrives. The p99 measurements describe the slow end of the observed response distribution, where a service can feel inconsistent even when its average looks acceptable.

© 2026 Signal65. All rights reserved.

For organizations supporting many simultaneous interactions, the split layout offers a practical way to get more from the same eight-GPU server. At 64 concurrent requests, it delivered 37% more throughput than the single instance while also improving response times, a compelling reason to evaluate two instances with the organization's own workloads and service targets.



Long Context Creates Different Kinds of Demand

A large context window can support document analysis, but the experience depends on how that context arrives. A service receiving many new documents faces a different scheduling problem from one processing a single document or answering follow-up questions on a cached conversation.

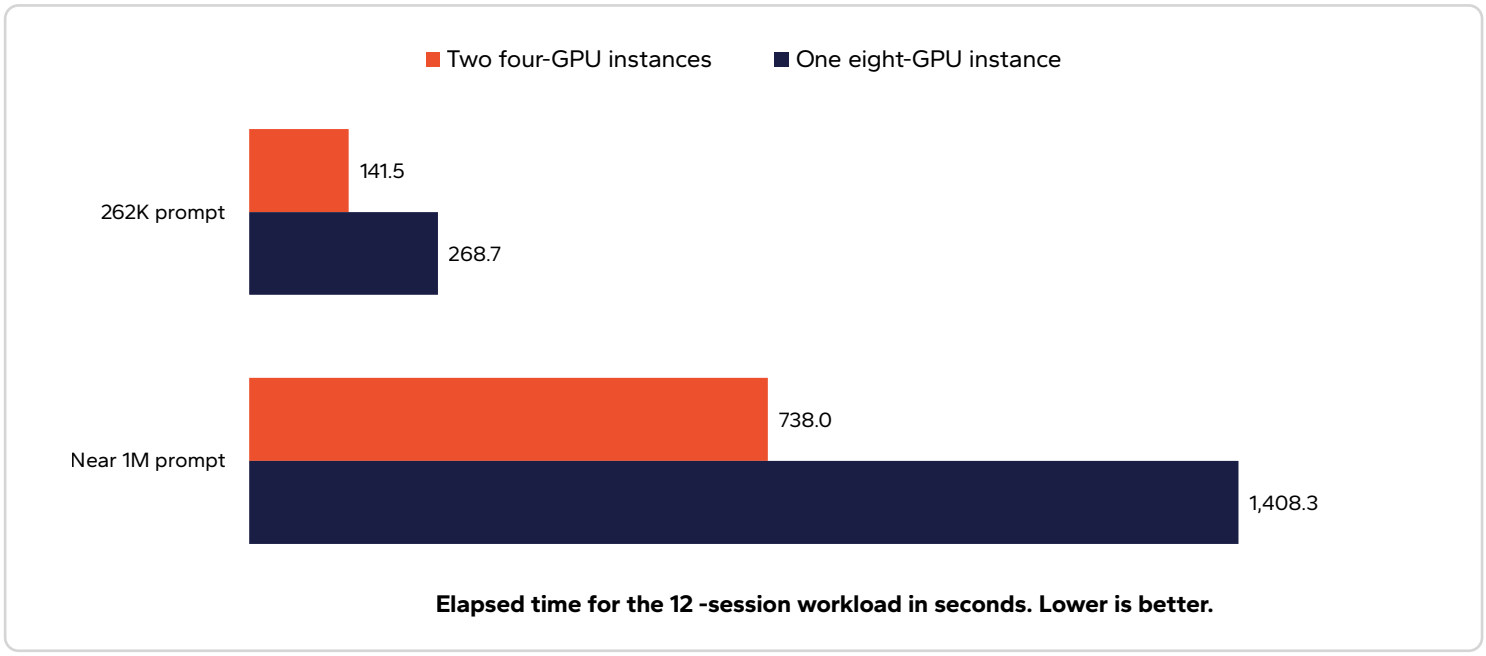


Figure 3: *Elapsed time for 12 concurrent long-context sessions*

Medians of three restarts; each workload contained 12 sessions and 24 requests with 1,024-token replies. 1M means 1,046,528 input tokens. These are whole-workload completion times. Source: Signal65 matched tests.

With 12 sessions queued, two instances completed both tested workloads in roughly half the time. Each instance had its own prompt-processing lane, allowing the pair to advance separate requests. The full-node instance's larger context-memory pool did not remove the scheduling bottleneck in this configuration.

© 2026 Signal65. All rights reserved.

Interaction pattern	Two instances	One instance
One near-1M prompt Time to first token	119.42 s	113.58 s
One conversation grown over eight turns Total time to final context	136.0 s	131.7 s

The full-node instance started answering a single long prompt about 5–9% sooner across the tested prompt lengths. The growing conversation showed a much smaller layout difference. Prefix reuse avoids processing all prior content again, so later turns do not repeat the work of a new bulk prompt.

This makes the request pattern a procurement and deployment input. A document queue with a constrained processing window should test concurrent ingestion. Interactive analysis tools should also test follow-up turns, session affinity and user wait times. Application retrieval and tool calls belong in the final end-to-end measurement; the timings above isolate model serving.



Memory Capacity Determines the Operating Envelope

Faster answers and more room for active conversations are different forms of capacity. The full-node DeepSeek layout offered more GPU memory headroom because it spread model weights across eight GPUs. Its engine-reported request-capacity estimate was about 1.6–1.7 times the combined split estimates at matched context lengths.

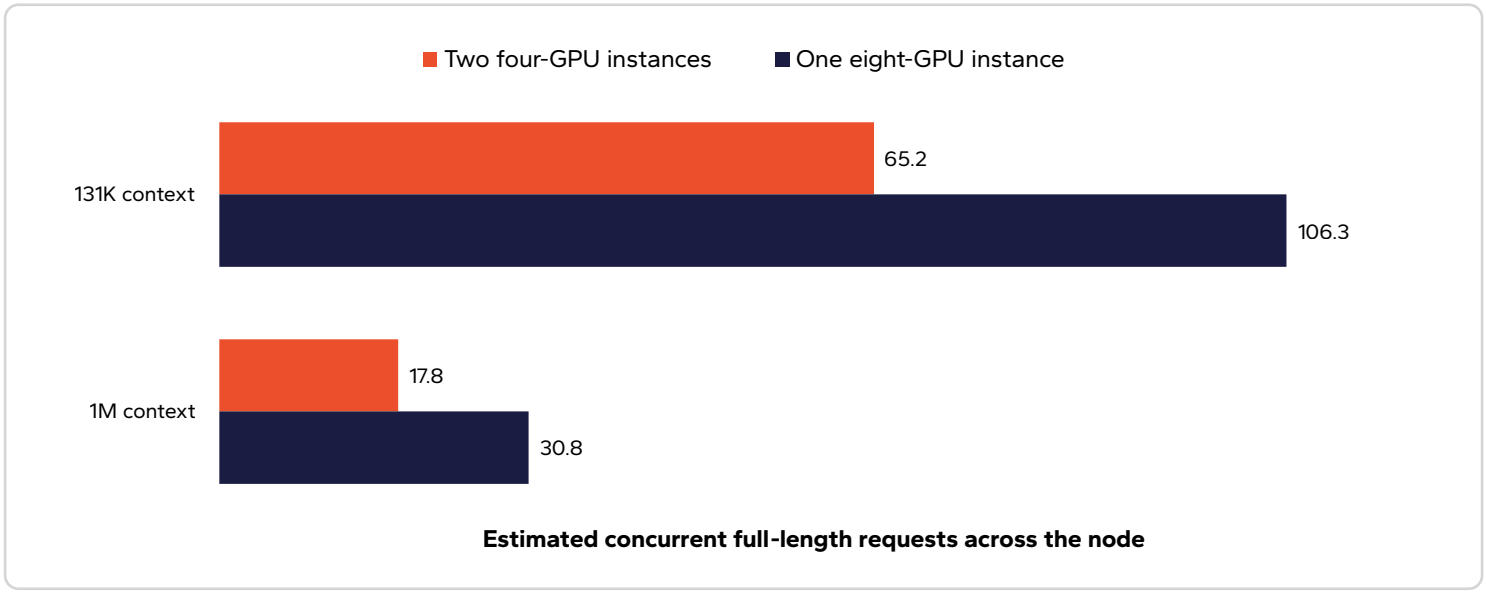


Figure 4: *Estimated capacity for full-length requests at matched context ceilings*

Memory-based allocation estimates. The split total combines two separate pools and assumes requests can be placed appropriately. Sources: Signal65 matched tests and Phase 1 findings.

© 2026 Signal65. All rights reserved.

Context and Answer Length Share the Budget

At the 131,072-token ceiling, the estimates were 65 parallel requests across the pair and 106 for the single instance. Near the one-million-token ceiling, they fell to approximately 18 and 31. Longer retained contexts consume more memory even when users are not continuously generating output.

Answer limits matter as well. The full-node layout completed a 128-stream decode test, while the split-instance layout could not execute that test with an 8,192-token answer allowance due to insufficient GPU memory. A smaller output-token allowance might have fit within the budget, so consider application requirements and usage patterns in any evaluation.

Host Memory Is Part of the Configuration

DeepSeek's Engram tables pinned 190 GiB of host memory for one instance and 380 GiB for two. The extra model copy in the dual-instance configuration ensured NUMA (Non-Uniform Memory Access) locality between each instance's CPUs and GPUs, but it doubled the host-memory requirement and divided the GPU context space.

The useful capacity figure is the peak workload the complete service can sustain within its response-time and quality requirements. Test workloads with realistic input and answer lengths, and measure accepted, queued and rejected requests on each instance.



Tuning Has a Capacity Cost

LLM serving settings can improve a benchmark while reducing headroom available to the application. The clearest example in this evaluation was the batch token budget. Doubling it from 4,096 to 8,192 tokens shortened the repeated long-context workload by about 2.5–2.7%, while reducing estimated request capacity by 20–48% across the tested contexts and layouts.

262K workload completion time	4,096-token budget	8,192-token budget
Two four-GPU instances	140.94 s	137.07 s
One eight-GPU instance	268.55 s	261.72 s

Means of six restarts per setting, using 12 sessions, up to 1M-token context. Reported reductions were 2.74% and 2.54%. Source: Signal65 Phase 4 confirmation.

© 2026 Signal65. All rights reserved.

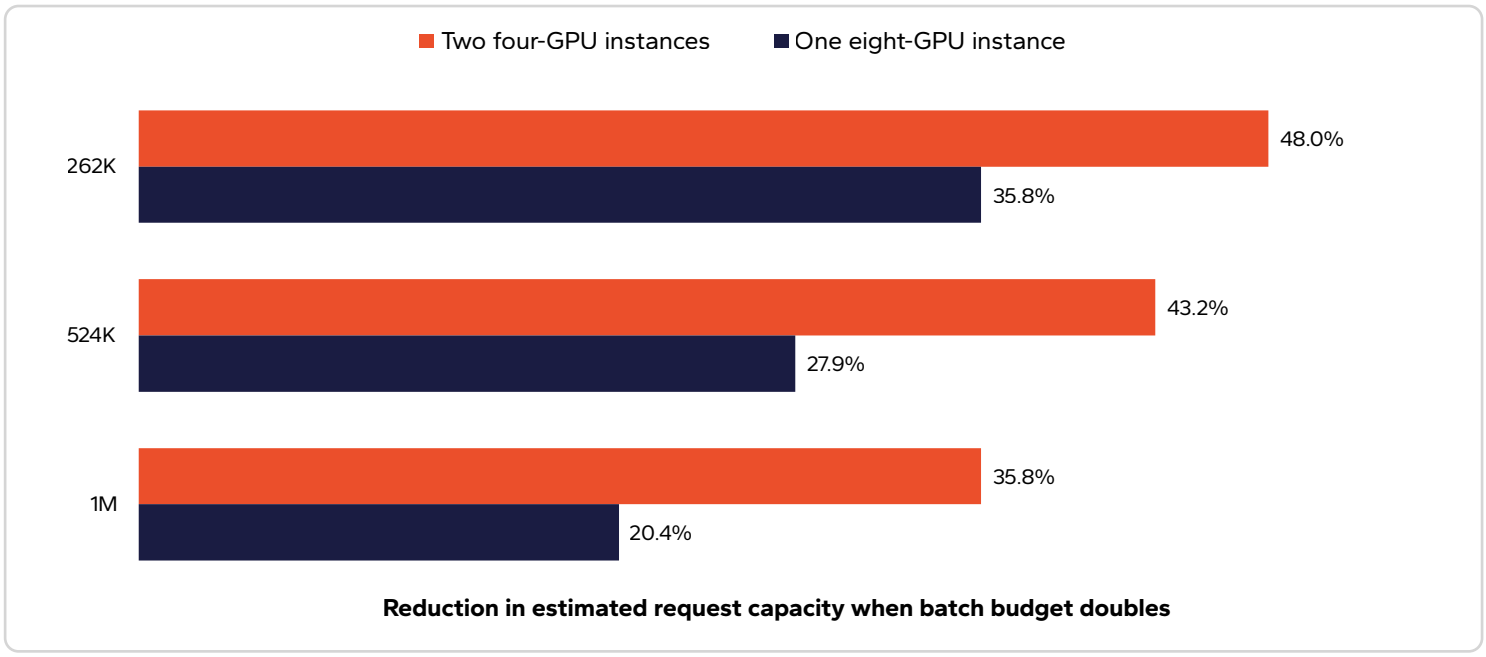


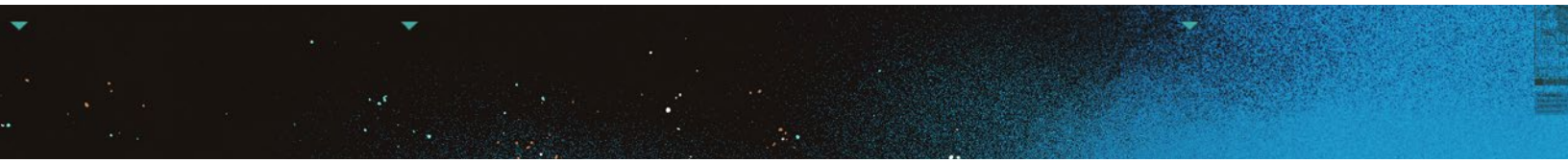
Figure 5: *Reduction in estimated request capacity when the batch budget doubled*

A Small Speed Gain Can Consume Useful Headroom

The larger budget nearly doubled the fixed memory reservation per request in the tested hybrid-cache implementation. That cost remained relevant even when a prompt did not use the full context window. A service already constrained by active-context memory could lose more operating flexibility than it gains from slightly shorter completion time.

The broader exploratory sweep did not reverse the layout ranking. Allowing a second simultaneous prompt-processing lane within an engine worsened tail latency in that run. Adding an independent instance and raising a scheduler limit are different changes and are not interchangeable.

The baseline is the stronger starting point when it already meets the response target and memory is the constraint. A larger batch budget deserves consideration when the workload has spare headroom and benefits from faster completion.



Quality Belongs in the Deployment Decision

A faster service creates value when its answers still meet the application's requirements. The DeepSeek checks did not reveal a material quality difference between layouts at the resolution tested. Across six restarts, average accuracy on the same 250 GSM8K questions was 96.60% for two instances and 97.07% for one.

© 2026 Signal65. All rights reserved.

Quality check	Two instances	One instance
GSM8K average across six restarts 2,048-token answer allowance	96.60%	97.07%
Targeted recheck of truncated items Corrected initial-run score	245 / 250	245 / 250
Planted-fact recall at each tested length	9 / 9	9 / 9

GSM8K reused the same 250 questions on every restart. The recheck covered 11 truncated items at an 8,192-token answer allowance; it was not a full six-restart evaluation at that allowance. Recall checked planted facts, not general document comprehension. Sources: Signal65 matched and Phase 4 tests.

Answer Limits Can Look Like Model Failures

The average layout difference was about 1.2 questions, compared with a three-question score range within each layout. Agreement between layouts was also similar to agreement across repeated runs of the same layout. That supports treating serving behavior as the primary input to the layout decision while retaining application quality as an acceptance requirement.

Eleven items in the initial paired run reached the answer ceiling. Rechecking those items with more room to finish produced a corrected initial-run score of 245 out of 250 for both layouts. The practical issue is that answer length affects both measured quality and available capacity. Reducing the output allowance can make the system appear smaller and faster while leaving important work unfinished.

Test the Work the Application Must Complete

A domain or reasoning benchmark is useful evidence, but it does not qualify an enterprise workflow. A document assistant should use the correct source material; a service that produces structured records should return valid fields; a tool-using application should respect the actions it is allowed to take. These behaviors need representative tests with explicit success criteria.

Long documents and follow-up turns belong in that set. Both DeepSeek layouts passed the limited planted-fact checks, including near the context ceiling, but the result does not establish broad comprehension of every long document. The buyer should evaluate grounded answers and task completion with the intended retrieval, prompts and runtime configuration.

Consider PINNACLE benchmark from Kamiwaza and Signal65 to bring application PINNACLE benchmark brings this application-level perspective to model and platform evaluation. It tests how reliably models complete multistep enterprise tasks, ground answers in supplied information and produce verifiable results. By connecting task quality with performance and cost per correct task, PINNACLE helps buyers assess which model and infrastructure combination can deliver the most useful work for their investment.

© 2026 Signal65. All rights reserved.



Gemma 4 Broadens the Deployment Options

A smaller model that meets the task requirement can change how a GPU-dense server is used. Instead of assigning four or eight GPUs to one model instance, it becomes possible to run one instance per GPU. Gemma 4 feasibility tests demonstrated that this is a configuration worth evaluating on Dell PowerEdge XE7745.

Testing served 31B and 26B A4B variants on single GPUs, including NVFP4 26B A4B instances. Each instance answered a short question when checked in turn, which established coexistence and basic serving rather than performance under simultaneous application load.

The Serving Image Changed the Result

Stock vLLM supported all three Gemma 4 variants. All three compiled, captured CUDA graphs and answered the basic check at a GPU memory utilization setting of 0.90. Earlier testing with eager execution using the same engine as DeepSeek resulted in lower performance.

Gemma 4 variant One GPU per instance	Context Ceiling Tokens	Estimated full-length request capacity
26B A4B BF16	262,144	4.10
26B A4B NVFP4	262,144	15.50
31B NVFP4	262,144	3.05

October 6, 2026 stock vLLM cu134-nightly results, pinned by image digest; memory utilization 0.90. All rows used compiled execution. Source: Signal65 on-premises fleet feasibility record.

At 0.95, the stock configurations ran out of memory during initialization; the 26B records place the failure in fused-MoE autotuning after cache sizing. Reducing the target left room for initialization and allowed serving to complete. Usable configurations depend on the runtime's total memory needs, not simply whether model weights and a calculated cache will fit.

Qualifying a Gemma 4 Service

Quantization increases memory headroom, but its effect on task quality still needs measurement. Testing did not measure the eight-instance coexistence on stock vLLM or compare performance with DeepSeek.

Implementors need to understand and evaluate the exact model, precision, execution mode and image used for proposed services, then measure the complete workload on that combination. Pinning the successful image by digest preserves the tested version. Startup success narrows the shortlist; representative quality and load tests establish whether the smaller-model path delivers the intended service.

© 2026 Signal65. All rights reserved.

Turning Performance into Operational Value

The economic opportunity in these results comes from making a fixed server allocation more useful. If a configuration completes the required work sooner or sustains more accepted requests within the response target, it may absorb growth without an immediate increase in hardware. That opportunity depends on the workload mix and operating requirements, not simply the highest token rate.

More Productive Shared Capacity

Two DeepSeek instances improved concurrent output and completed the tested document queue sooner. For a shared service, that can translate into shorter queues or more room for overlapping work. The value needs to be measured at the application boundary, where retrieval, tool calls and answer quality determine whether a request has actually been completed.

The full-node instance offered a different benefit: more memory available for long contexts and fewer host-resident model tables. A service with many retained conversations may prefer that capacity even when its output benchmark is lower. Buying against the wrong constraint can leave additional compute unable to solve the problem users experience.

Predictable Deployment and Operation

Smaller Gemma 4 models offer further parallelization and more granular deployment. Testing also highlighted the importance of the full software stack in infrastructure decisions. Changing the image removed a compilation limitation, while changing the memory target resolved an initialization failure.

Routing and monitoring are part of the same operating requirement. Split layouts need conversation affinity and visibility into both instances. Node-level totals alone can conceal an overloaded partition, while telemetry that samples only one backend can understate the available activity or capacity.

Build the Business Case Around Accepted Work

When designing inference solutions, define accepted tasks before testing. Count work that meets the required answer quality and response time, and include the resources consumed by rejected requests, retries and incomplete answers. Then evaluate the attributable server, software, support and energy costs over the expected demand profile.

This study identified configurations that deserve priority in a proof of concept. The business case becomes credible when the selected service meets its combined quality, capacity and response requirements at a measured operating cost.

© 2026 Signal65. All rights reserved.

Conclusion

The Dell PowerEdge XE7745 results demonstrate that deployment layout can materially change the service delivered by the same eight GPUs. For DeepSeek-V4.1-Flash, two four-GPU instances were the stronger configuration for concurrent output and multiple long prompts. One eight-GPU instance offered more context-memory headroom, lower host-memory use for Engram tables and a modest advantage when processing a single long prompt.

The choice should follow the application's demand. A busy shared assistant and a service retaining many long conversations can put pressure on different resources. Neither GPU count nor maximum context length is enough to describe usable capacity. The evaluation must keep request lengths, answer budgets and quality requirements consistent while measuring the behavior that users will experience.

Primary requirement	Configuration to evaluate first
Concurrent DeepSeek output or queued large documents	Two four-GPU instances
DeepSeek context-memory headroom or one large prompt at a time	One eight-GPU instance
A smaller model that can meet the task-quality requirement	Single-GPU Gemma 4 instances

More Useful Work from the Same Server

The strongest result in this evaluation is the improvement available within a fixed hardware investment. Two DeepSeek instances delivered 37% more throughput at 64 concurrent requests and completed the tested concurrent long-context workloads in roughly half the time. Those gains came from changing how the existing eight GPUs served the workload.

Dell PowerEdge XE7745 compute with Broadcom-based networking supports multiple ways to allocate inference capacity as application needs change. Split instances deliver greater throughput and a full-node instance provides more active-context capacity, while smaller Gemma 4 models open a path to services running on individual GPUs. Implementors can choose how much of the server a model needs and revisit that choice as their mix of applications evolves.

For organizations expanding enterprise AI, this makes deployment design a practical investment decision. A configuration that handles more of the required work within the application's response and quality targets can accommodate growth before another server is needed. The opportunity is to build that capacity into the deployment from the outset and make each node more productive as inference demand grows.

© 2026 Signal65. All rights reserved.

Important Information About this Report

CONTRIBUTORS

Brian Martin
AI Data Center Performance | Signal65

PUBLISHER

Ryan Shrout
President and GM | Signal65

INQUIRIES

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

CITATIONS

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

LICENSING

This publication may be reproduced, redistributed, or shared, but may not be modified in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and lab services to many high-tech companies, including those mentioned in this paper.

ABOUT SIGNAL65

Signal65 exists to be a source of data in a world where technology markets and product landscapes create complex and distorted views of product truth. We strive to provide honest and comprehensive feedback and analysis for our clients in order for them to better understand their own competitive positioning and create optimal opportunities to market and message their devices and services.

IN PARTNERSHIP WITH

signal65

CONTACT INFORMATION
Signal65 | signal65.com

© 2026 Signal65. All rights reserved.