

Enterprise Inference on Dell PowerEdge XE7745

Balancing response time and memory with DeepSeek and Gemma 4 on Dell PowerEdge servers and Broadcom networking.

Inference services must answer the right questions within an acceptable time while absorbing peak demand. How GPUs are assigned to model instances changes the service they deliver, even when the hardware and model stay the same. Deployment layout is a capacity decision: two four-GPU instances favored concurrent work, one eight-GPU instance kept more memory for active contexts, and single-GPU Gemma 4 instances opened smaller-model paths on the same node.

KEY FINDINGS

+37%

OUTPUT THROUGHPUT

Two four-GPU instances delivered 2,756 tokens/s at 64 simultaneous streams, against 2,017 for one eight-GPU instance, and led at every matched stream count from 1 through 64.

-48%

ELAPSED TIME

Twelve queued million-token sessions completed in 738 s on two instances versus 1,408 s on one. Each instance brought its own prompt-processing lane.

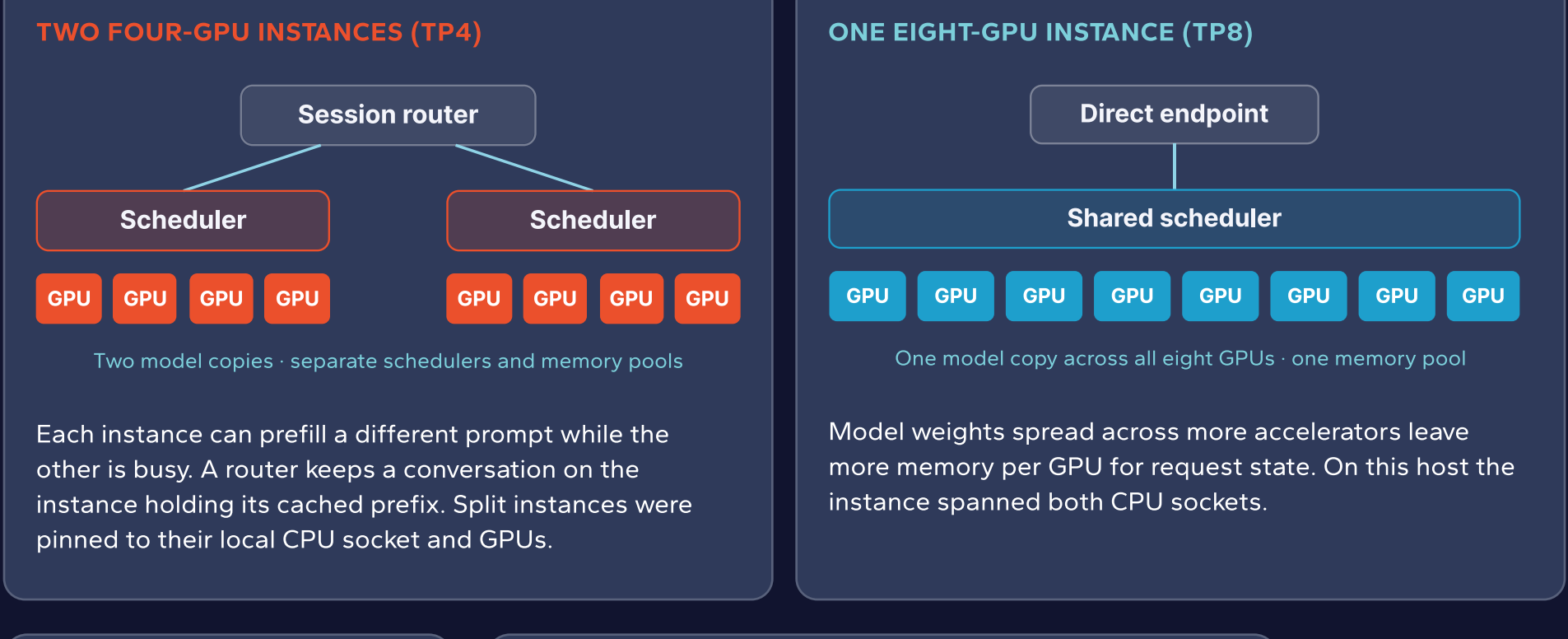
1.6 - 1.7x

CONTEXT CAPACITY

One eight-GPU instance holds 1.6–1.7 times the full-length requests of the split pair at matched context ceilings, and pins half the host memory for Engram tables.

ONE SERVER, TWO DEPLOYMENT LAYOUTS

Tensor parallelism divides a model across GPUs: TP4 uses four GPUs per instance, TP8 uses eight. Both layouts ran the same DeepSeek-V4.1-Flash model and serving image on the same eight GPUs of one Dell PowerEdge XE7745.



- Same model · DeepSeek-V4.1-Flash
- Same serving image · FP8 KV cache, 4,096-token batch budget
- Same GPUs · 8 × NVIDIA RTX PRO 6000 Blackwell Server Edition
- Medians of three deployment restarts

CONCURRENCY: MORE OUTPUT, SHORTER WAITS

At 64 simultaneous short-prompt streams with a 131,072-token context ceiling, the split layout raised aggregate output and improved both waits a user feels: the delay before an answer begins and the gaps as it arrives.



Response at 64 simultaneous streams; medians of three restarts. The split pair (orange bars) did not support more than 64 streams at an 8,192-token answer allowance, where the single instance (blue bars) completed a 128-stream test. Source: Signal65 matched configuration tests.

LONG CONTEXT: QUEUES VS. SINGLE PROMPTS

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.

12 concurrent sessions, whole-workload completion time

seconds · lower is better · 24 requests with 1,024-token replies

Prompt Length	Two four-GPU instances	One eight-GPU instance
262K prompt	141.5 s	268.7 s
Near-1M prompt	738.0 s	1,408.3 s

Near 1M means 1,046,528 input tokens. Medians of three restarts. The full-node instance's larger memory pool did not remove the prompt-processing bottleneck.

ONE NEAR-1M PROMPT · TIME TO FIRST TOKEN

113.6 s vs 119.4 s

One eight-GPU instance started answering a single long prompt about 5–9% sooner across the tested prompt lengths.

EIGHT-TURN CONVERSATION · TOTAL TIME

131.7 s vs 136.0 s

Prefix reuse avoids reprocessing prior content, so follow-up turns showed a much smaller layout difference.

WHAT TO TEST

A document queue with a processing window should test concurrent ingestion. Interactive tools should test follow-up turns, session affinity and user wait times, with retrieval and tool calls in the final end-to-end measurement.

MEMORY CAPACITY SETS THE OPERATING ENVELOPE

Faster answers and more room for active conversations are different forms of capacity. Longer retained contexts consume memory even when users are not generating output, and serving settings that speed a benchmark can spend the headroom an application needs.

Estimated concurrent full-length requests across the node

engine-reported, memory-based estimates · higher is more headroom

Context Length	Two four-GPU instances (combined)	One eight-GPU instance
131K context	65.2	106.3
1M context	17.8	30.8

Split total combines two separate pools and assumes requests can be placed appropriately. Sources: Signal65's matched tests and capacity-sizing runs.

HOST MEMORY FOR DEEPSEEK ENGRAM TABLES

190 GiB one instance · 380 GiB for two

The second model copy keeps NUMA locality for each instance's CPUs and GPUs, but doubles the host-memory requirement and divides GPU context space.

TUNING HAS A CAPACITY COST

-20% to -48% request capacity

Doubling the batch token budget from 4,096 to 8,192 shortened long-context workloads by only 2.5–2.7% while nearly doubling the fixed memory reserved per request.

QUALITY DID NOT SEPARATE THE LAYOUTS

96.60% / 97.07%

GSM8K accuracy across six restarts for two and one instance. Eleven truncated items rechecked with more room scored 245 of 250 on both, planted-fact recall was 9 of 9 at every tested length.

GEMMA 4: ONE INSTANCE PER GPU

A smaller model that meets the task changes how the server is used.

Instead of assigning four or eight GPUs to one model instance, feasibility tested all three, captured CUDA graphs and answered a basic check at a GPU memory utilization of 0.90; at 0.95 the configurations ran out of memory during initialization. Eight quantized 26B instances coexisted on one node.

Quantization widens memory headroom, but its effect on task quality still needs measurement. Pin the successful image by digest, then qualify the exact model, precision and execution mode with the complete workload.

Gemma 4 variant · one GPU each	Context ceiling	Est. full-length requests
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; compiled execution. Coexistence and basic serving were established, not performance under simultaneous load.

THE TESTED PLATFORM

Dell PowerEdge XE7745

GPU-dense inference node; each layout used the eight GPUs of one server in Signal65's two-node cluster.

- 2 × AMD EPYC 9555, 64 cores per socket; 2,267 GB DDR5 reported
- 8 × NVIDIA RTX PRO 6000 Blackwell Server Edition, 97,887 MiB reported per GPU; ECC on; no NVLink
- 8 × Broadcom BCM57608 400GbE network controllers with RoCEv2
- GPUs 0–3 on NUMA node 0, 4–7 on node 1; split instances pinned to their local CPU sets

Dell PowerSwitch Z9864F-ON — the fabric

Built on Broadcom Tomahawk 5 silicon and running Enterprise SONiC Distribution by Dell Technologies 4.4.0, the switch connects the cluster's servers over 400GbE with RoCEv2.

Software stack

Both DeepSeek layouts used the same serving image and baseline settings; Gemma 4 ran on stock vLLM.

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	localinference-lab/vllm:jovian-judgement-community-20260914-r38
Gemma 4 serving image	Stock vLLM cu134-nightly (October 6, 2026), pinned by image digest
Baseline serving settings	FP8 KV cache · batch budget 4,096 tokens · one parallel prefill per instance · decode context parallelism 1
GPU memory utilization	0.95 for DeepSeek · 0.90 for Gemma 4

CHOOSE THE LAYOUT BY THE WORKLOAD

Neither GPU count nor maximum context length describes usable capacity. A busy shared assistant and a service retaining many long conversations put pressure on different resources.

PRIMARY REQUIREMENT

Concurrent DeepSeek output or queued large documents

EVALUATE FIRST

Two four-GPU instances

PRIMARY REQUIREMENT

Context-memory headroom or one large prompt at a time

EVALUATE FIRST

One eight-GPU instance

PRIMARY REQUIREMENT

A smaller model that can meet the task-quality requirement

EVALUATE FIRST

Single-GPU Gemma 4 instances