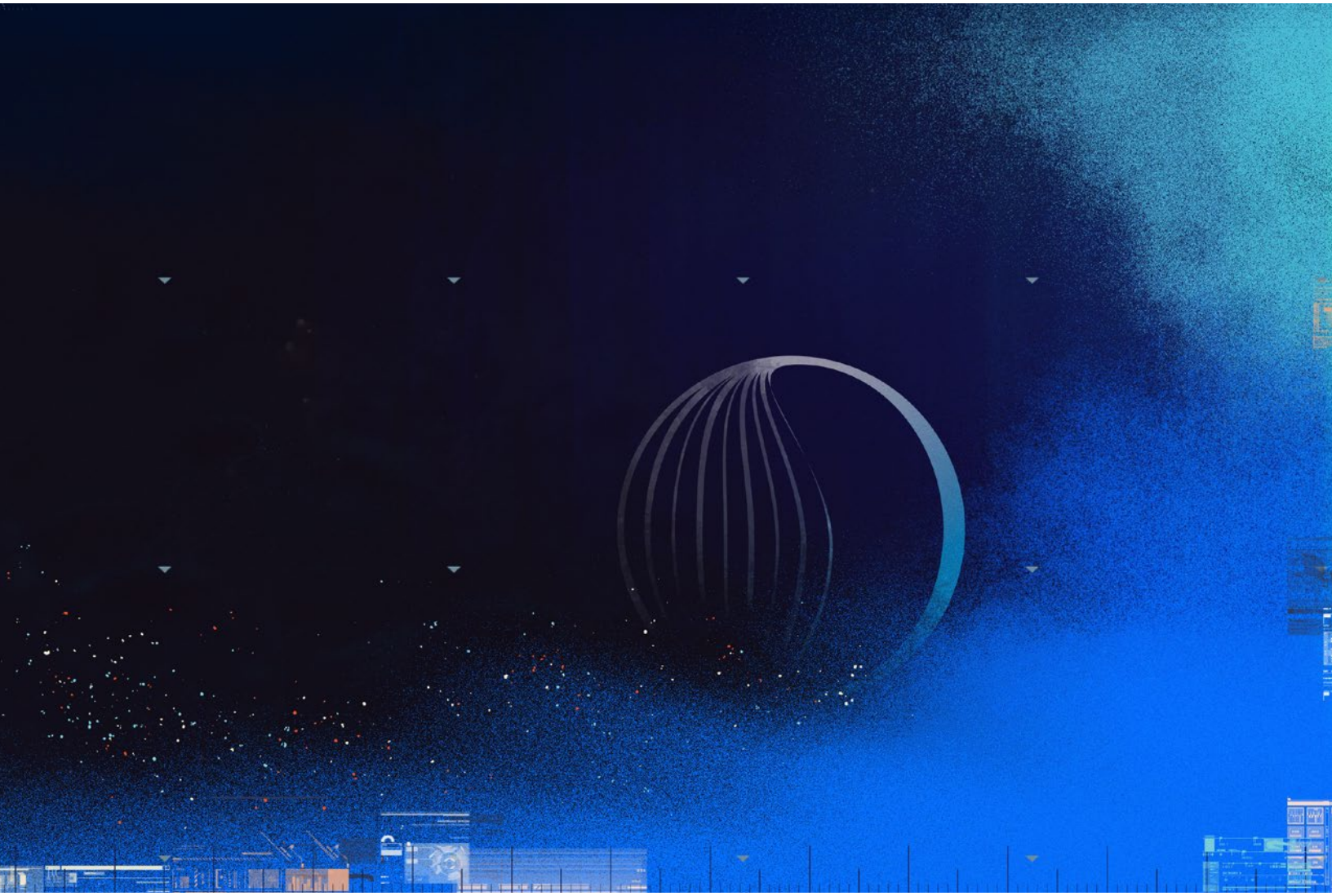
signal65



EXECUTIVE SUMMARY

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.

For less demanding applications, Gemma 4 adds smaller-model options, with single-GPU instances serving on stock vLLM. The right layout depends on how requests arrive and how much context each conversation retains, so a representative acceptance workload should guide the choice before deployment.

Key Highlights

37%

More output throughput with two instances at 64 streams

48%

Less elapsed time for 12-session million-token workloads

1.6–1.7x

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

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.

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 led at every matched stream count from one through 64, reaching a 37% advantage at 64 simultaneous streams, while also shortening the delay before an answer begins and the gaps as it arrives.

© 2026 Signal65. All rights reserved.

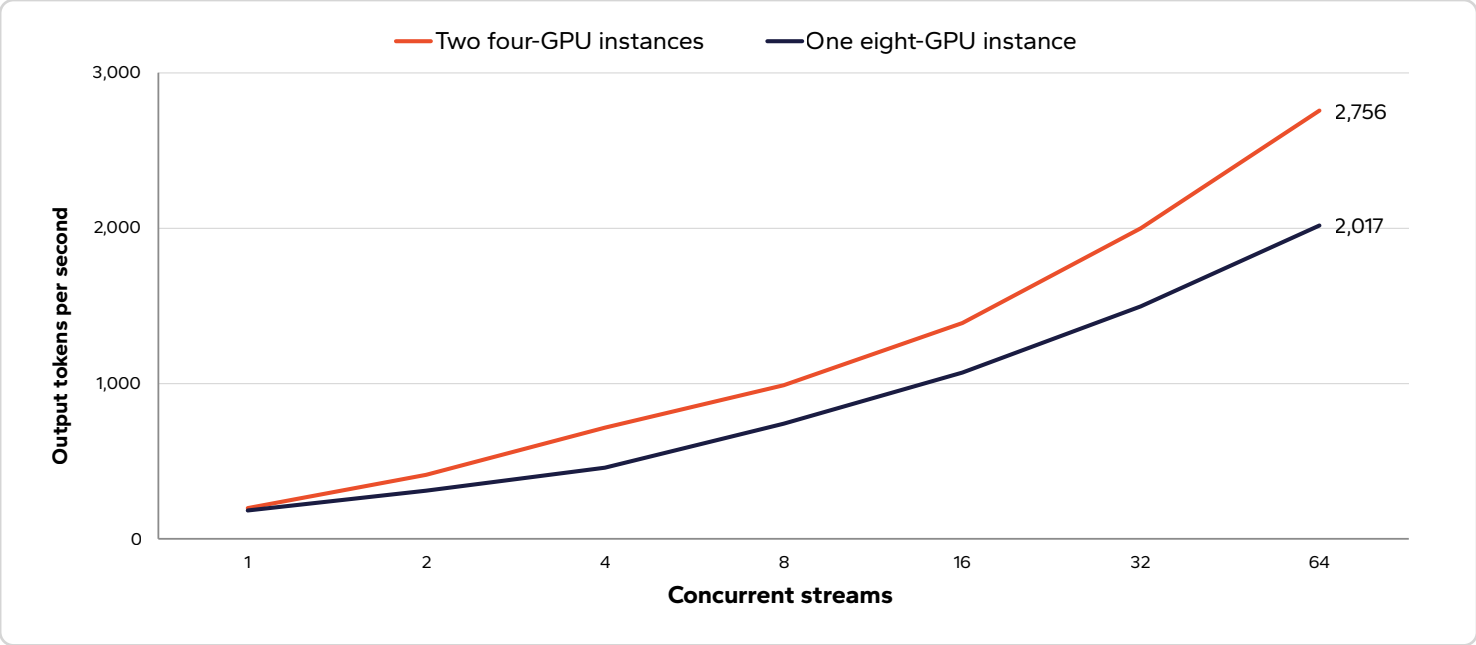


Figure 1: *Aggregate output throughput as simultaneous streams increase*

Short-prompt generation, 131,072-token ceiling; medians of three restarts; x-axis doubles each step. Source: Signal65 matched 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

Long-context queues followed the same pattern: with 12 sessions queued, two instances completed the million-token workload in 738 seconds against 1,408 seconds for the single instance, because each instance had its own prompt-processing lane.

Memory Capacity Sets the Operating Envelope

Faster answers and more room for active conversations are different forms of capacity. With model weights spread across eight GPUs, the full-node instance reported about 1.6–1.7 times the combined request capacity of the split pair at matched context lengths (106 parallel requests against 65 at the 131,072-token ceiling; roughly 31 against 18 near one million tokens). It also pinned half as much host memory for DeepSeek Engram lookup tables (190 GiB rather than 380 GiB) and answered a single near-1M prompt about 5–9% sooner.

Tuning carries its own capacity cost: doubling the batch token budget shortened long-context workloads by about 2.5–2.7% while cutting estimated request capacity by 20–48%. Quality did not separate the layouts; GSM8K accuracy was 96.60% and 97.07% across six restarts.

© 2026 Signal65. All rights reserved.

Conclusion

Deployment layout can materially change the service delivered by the same eight GPUs. For DeepSeek-V4.1-Flash, two four-GPU instances were stronger for concurrent output and queued large documents; one eight-GPU instance offered more context-memory headroom and a modest advantage on a single long prompt; and single-GPU Gemma 4 instances suit smaller models that meet the task-quality requirement. Dell PowerEdge XE7745 compute with Broadcom-based networking supports each allocation, letting implementors size the server to the model and revisit that choice as applications evolve.

Read the full report: Enterprise Inference Considerations on Dell PowerEdge XE7745

© 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.