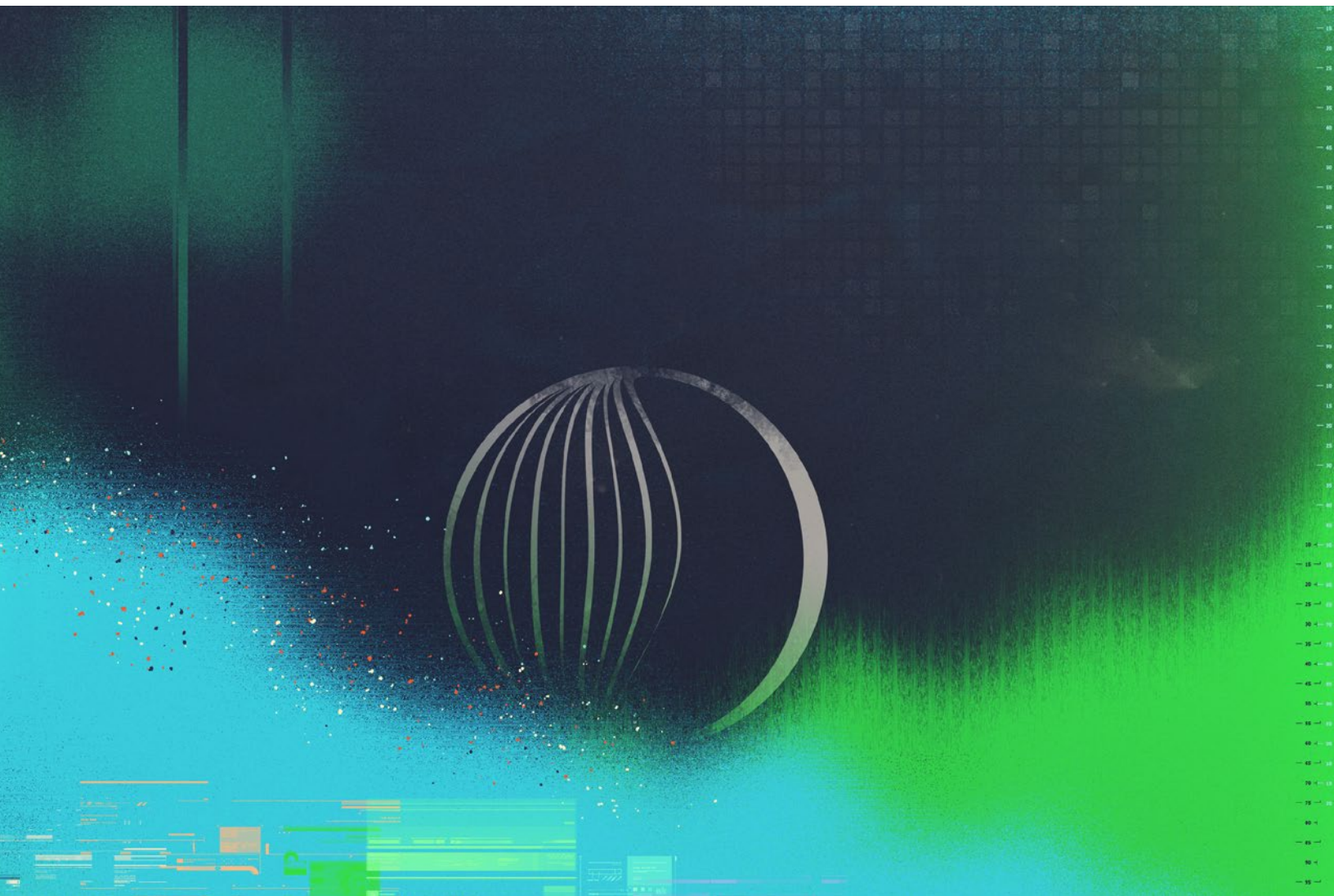




SIGNAL65 EXECUTIVE SUMMARY

Scaling On-Premises Agentic Workloads from Desktop to Datacenter with the Dell AI Factory with NVIDIA

AUTHORS

Mitch Lewis

Performance Analyst | Signal65

Ryan Shrout

President & GM | Signal65

IN PARTNERSHIP WITH

DELLTechnologies

 **NVIDIA**

SEPTEMBER 2026

Overview

AI workloads are quickly shifting from short interactive chats to autonomous agents that run continuously, call tools, and complete real enterprise tasks. That shift changes more than how AI gets used — it also changes where it makes economic sense to run. Agents can tolerate slower responses than a live user waiting on a screen, but they generate far more token volume over the course of a day and can run autonomously beyond typical work hours. That combination of relaxed latency and higher throughput tips the economics toward on-premises infrastructure instead of per-token cloud costs.

This report puts that logic to the test. Signal65 benchmarked eight Dell AI Factory with NVIDIA platforms — from a single-GPU desktop system to a 32-GPU PowerEdge cluster — utilizing the performance data to model three realistic enterprise agent workloads. The economics of each workload were modeled over a two-year cost horizon to compare Dell AI Factory with NVIDIA on-premises solutions against open-weight models on AWS Bedrock and leading frontier model APIs.

VS AWS BEDROCK

Up to 89.8% savings — desktop systems

Up to 67.7% savings — PowerEdge servers

VS LEADING FRONTIER APIs

Up to 98.4% savings — desktop systems

Up to 93.7% savings — PowerEdge servers

8

Dell AI Factory platforms benchmarked, desktop to datacenter

30B - 120B

parameter model range tested across workloads

16,000+

concurrent agents supported on the largest cluster

Dell Delivers a Cost Advantage Across Every Workload

To ground the analysis in real usage, Signal65 modeled three enterprise agent personas of increasing complexity: a knowledge worker handling everyday tasks, a sales agent blending two model sizes, and an always-on software development agent. Each workload showcased the benefits of on-premises agentic AI deployments, and highlighted the versatility of the Dell AI Factory with NVIDIA to scale from desktop to datacenter and from demo to deployment.

AI-agent Assisted Knowledge Worker: Big Savings from Lightweight AI

The lightest workload in the study — internal Q&A, summarization, everyday drafting — runs on the smallest model tested.

- **Model:** NVIDIA Nemotron Nano (3.6B active / 31.6B total parameters)
- **Scale:** 26 to 16,156 agents, from a compact desktop system to a full server cluster — every configuration beat both cloud comparisons
- **Savings:** 26%–90% vs AWS Bedrock · 88%–98% vs leading frontier APIs
- **Breakeven:** under 18 months vs Bedrock · under 3 months vs leading frontier pricing
- **At full cluster scale:** \$4.2M saved vs Bedrock, \$51.8M vs leading frontier APIs over two years

AI-enabled Sales Agent: Balanced Complexity, Consistent Savings

Blending a small model for routine tasks with a larger one for complex requests.

- **Model:** 50/50 blend of Nemotron Nano and Nemotron Super (12B active / 120B total parameters)
- **Scale:** 97 to 8,676 agents
- **Savings:** 41%–85% vs AWS Bedrock · 89%–97% vs leading frontier APIs
- **At full scale:** \$5.4M saved vs Bedrock, \$50.6M vs leading frontier APIs over two years

AI-agent Software Development: Maximum Complexity, Maximum Savings

Running an always-on coding agent on the largest model tested — the most demanding scenario in the study.

- **Model:** NVIDIA Nemotron Super (12B active / 120B total parameters)
- **Scale:** 2 to 4,850 agents — Dell won every configuration tested
- **Savings:** 8%–88% vs AWS Bedrock · 82%–98% vs leading frontier APIs
- **At full scale:** \$9.7M saved vs Bedrock, \$68.8M vs leading frontier APIs over two years — the largest dollar savings of any workload



Built to Scale from Desktop to Datacenter

The Dell AI Factory with NVIDIA portfolio spans from a single-GPU desktop system to multi-server PowerEdge clusters, giving organizations an entry point at every scale. Compact desktop and workstation systems are well suited to pilots, prototyping, and small-team deployments — proving out agentic use cases before committing to larger infrastructure. Mid-range configurations support growing deployments in the hundreds of agents, while PowerEdge servers and multi-server clusters scale to full production fleets of thousands of agents, all without sacrificing the cost advantage found at every tier tested.

The same portfolio also supports the full range of workload complexity. Even the smallest systems in this study ran every workload tested, including complex software development. By leveraging the full Dell AI Factory with NVIDIA portfolio, organizations can start with whatever platform matches their current needs, whether that's a single desktop for a pilot or a full cluster for a production fleet, and grow into a larger deployment later without changing their overall approach to agentic AI.

Bottom Line

As agentic AI reshapes how enterprises deploy AI, direct financial impact — not just productivity — is becoming the metric organizations use to justify AI investment, according to Futurum Research. Across every workload and platform tested, the Dell AI Factory with NVIDIA delivered a cost advantage in the large majority of configurations, with payback as fast as roughly 2 months against open-weight cloud pricing and under a month against leading frontier APIs. With support ranging from a handful of agents on an edge desktop to more than 16,000 on a multi-server cluster, on-premises infrastructure is a credible option at every point on that spectrum — not just at the extremes.

See the full Signal65 Lab Insight report, *"The Economics of Agentic AI: Cost Advantages of Dell AI Factory with NVIDIA vs Public Cloud from Pilot to Scale,"* for detailed methodology, assumptions, and per-configuration results.

Important Information About this Report

CONTRIBUTORS

Mitch Lewis

Performance Analyst | Signal65

Ryan Shrout

President and GM | 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 document, including any supporting materials, is owned by Signal65. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of Signal65.

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT SIGNAL65

Signal65 is a leading research organization specializing in enterprise AI infrastructure optimization and deployment strategies. Our lab focuses on evaluating and optimizing AI hardware and software solutions for real-world enterprise applications, with particular expertise in large language models, retrieval-augmented generation systems, and distributed AI architectures.

For more information, visit signal65.com or contact research@signal65.com



IN PARTNERSHIP WITH



CONTACT INFORMATION

Signal65 | signal65.com