signal65

SIGNAL65 INSIGHTS

The Economics of Agentic AI:

On-premises Deployments with Dell AI Factory with NVIDIA vs. Cloud

AUTHORS

Mitch Lewis
Performance Analyst | Signal65

Ryan Shrout
President & GM | Signal65

IN PARTNERSHIP WITH

DELL Technologies

MAY 2026

Overview

Enterprise AI adoption is rapidly shifting from experimentation to execution, and increasingly, that execution is agentic. According to Futurum Research, 93% of enterprises are already researching, piloting, or deploying AI agents.

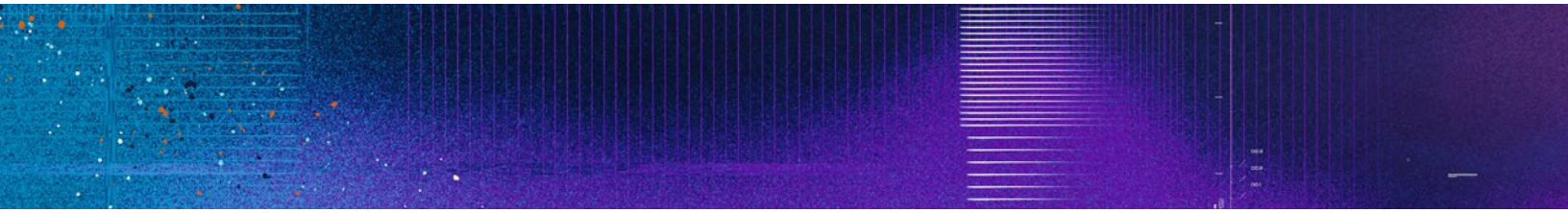
*Signal65 validated analysis shows that The Dell AI Factory with NVIDIA **can breakeven in as few as two months**, compared to public cloud APIs.*

Early enterprise AI deployments have largely relied on cloud-hosted models due to their accessibility, scalability, and consumption-based pricing. While well suited for bursty workloads and specialized services, ongoing agentic AI workloads fundamentally change this economic model. Autonomous AI agents can operate continuously, generating sustained inference demand and significantly increasing long-term token consumption. Compared to standard chat interactions, agentic workloads consume orders of magnitude more tokens with agents easily utilizing 4x to 15x more tokens. As agentic workloads continue to evolve, token growth is expected to grow even further, with autonomous agents driving up to 1000x more inference demand than reasoning AI.

Dell AI Factory with NVIDIA infrastructure provides an alternative to cloud APIs, enabling organizations to run agentic workloads of all sizes on-premises, without per-token pricing. The Dell AI Factory with NVIDIA portfolio ranges from PCs, to workstations, to enterprise grade PowerEdge servers, all capable of producing and consuming tokens to support concurrent agents.

This analysis examines the economics of deploying agentic AI workloads on-premises with Dell AI Factory with NVIDIA infrastructure versus cloud-based APIs. Using Dell Technologies and NVIDIA performance and pricing data, the analysis models three enterprise workload profiles: an AI-agent assisted knowledge worker, an AI-enabled sales agent, and AI-agent assisted software development. Each workload was modeled as a persistent 24-hour deployment for 260 days a year—representing a global organization's workweek—over a two year period. On-premises environments were configured with 60-80% utilization, depending on the workload, with cloud comparisons sized to match the number of agents supported by each hardware platform.

Across all tested profiles, on-premises AI infrastructure demonstrated a commanding financial lead, offering a substantial reduction in TCO (Total Cost of Ownership) compared to cloud deployments for both small-scale assistants and complex agentic fleets.



[1] Futurum Research, 1H 2026 AI Platforms Decision Maker Survey, March 2026.

Key Findings

Dell AI Factory with NVIDIA Delivers Significant Cost Advantages Across All Workloads

Across all workload configurations analyzed, Dell AI Factory with NVIDIA infrastructure reduced the cost of persistent AI agent deployments by **28% to 90%+** compared to cloud-based AI APIs.

AI-agent Assisted Knowledge Worker (Low Complexity)

- Dell Pro Max with NVIDIA GB10 GPU **reduced costs by up to 28%** for deployments supporting 8 agents.
- Dell T2 Workstations with NVIDIA RTX PRO 6000 BW GPUs **reduced costs by up to 71%** at scales of approximately 75 agents.
- Dell Pro Max with NVIDIA GB300 Ultra GPU delivered up to **56% lower costs** for deployments approaching 300 agents.
- Dell PowerEdge XE7740 systems with NVIDIA RTX PRO 6000 BW GPUs achieved up to **94% cost savings** for deployments nearing 18,000 agents.

AI-enabled Sales Agent (Medium Complexity)

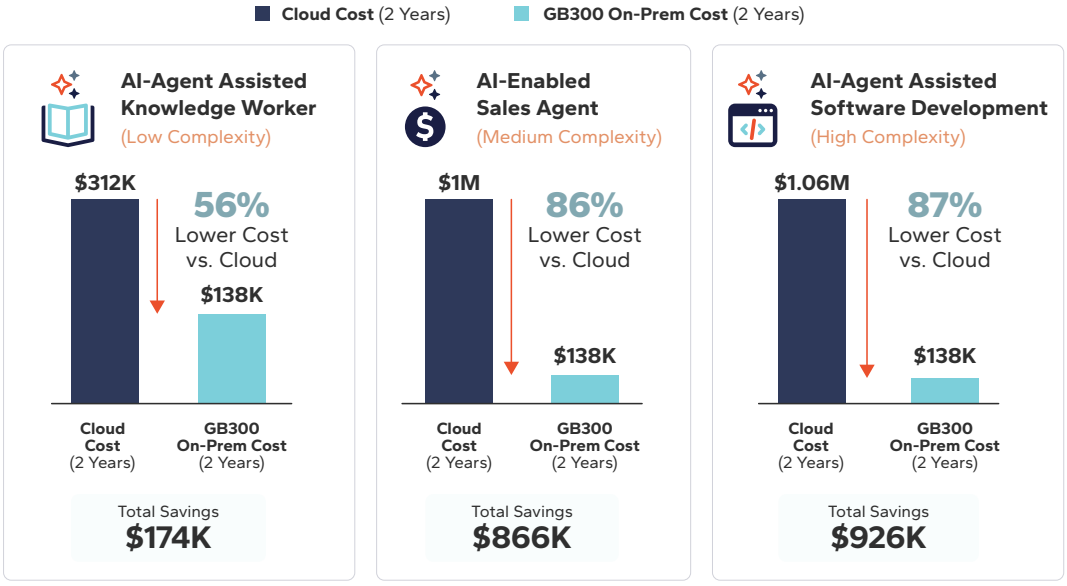
- Dell Pro Max with NVIDIA GB10 GPU delivered up to **76% lower costs** for medium-complexity sales agent deployments supporting 4 agents.
- Dell T2 Workstations with NVIDIA RTX PRO 6000 BW GPUs achieved up to **91% cost savings** for deployments supporting up to 40 agents.
- Dell Pro Max with NVIDIA GB300 Ultra GPU delivered up to **86% lower costs** for AI-enabled sales workloads supporting nearly 150 agents.

AI-agent Assisted Software Development (High Complexity)

- Dell T2 Workstations with NVIDIA RTX PRO 6000 BW GPUs achieved up to **93% cost savings** for deployments supporting approximately 20 agents.
- Dell Pro Max with NVIDIA GB300 Ultra GPU delivered up to **87% cost savings** for software development use cases supporting approximately 60 agents.
- Dell PowerEdge XE7745 systems with NVIDIA H200 NVL GPUs delivered up to **98% lower costs** for large-scale software development deployments supporting more than 5,000 agents.

Dell Pro Max with NVIDIA GB300 Ultra GPU Delivers Significant Cost Savings vs. Cloud

Same NVIDIA GB300 Ultra GPU System. Dramatically Lower Cost Across All Workloads.

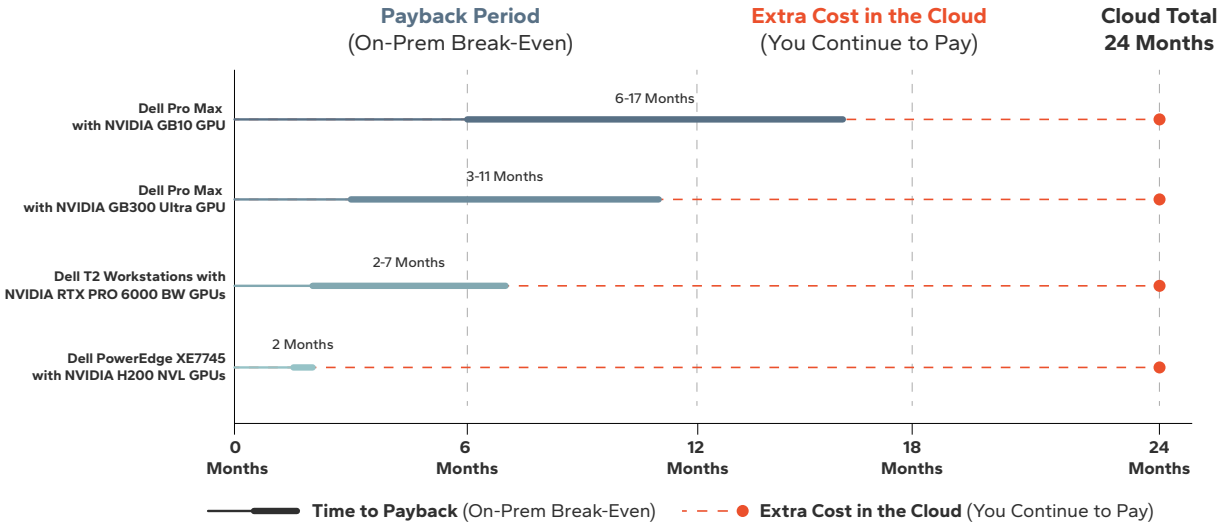


Dell Pro Max with NVIDIA GB300 Ultra GPU provides substantial savings across every workload and complexity level.

On-Premises Infrastructure Achieves Rapid ROI

Most Dell AI Factory with NVIDIA solutions achieved breakeven in **less than one year** under persistent AI agent utilization, depending on agent workload and utilization assumptions.

On-Premises Hardware Pays Off Faster. Every Month After Payback is Extra Cost in the Cloud.



The sooner you pay off your hardware, the more you save.
All months after payback = Extra cost in the cloud.

These findings indicate that infrastructure investments can rapidly offset operational API costs as organizations scale always-on AI agent deployments.

On-premises Infrastructure Provides Significant Long-Term Cost Savings

Over a two-year period, modeled savings ranged from thousands of dollars for individual AI knowledge worker deployments to multimillion-dollar reductions for enterprise-scale autonomous AI environments. The results indicate that organizations can realize meaningful economic advantages from on-premises AI infrastructure across a broad range of agentic AI deployment models.

Hardware	Workload	Modeled Two-Year Savings
Dell Pro Max with NVIDIA GB10 GPU	AI-agent Assisted Knowledge Worker	$2.5K
	AI-enabled Sales Agent	$20K
Dell Pro Max T2 with NVIDIA RTX PRO 6000 BW GPU	AI-agent Assisted Knowledge Worker	$58K
	AI-enabled Sales Agent	$249K
	AI-agent Assisted Software Development	$309K
Dell Pro Max with NVIDIA GB300 Ultra GPU	AI-agent Assisted Knowledge Worker	$174K
	AI-enabled Sales Agent	$866K
	AI-agent Assisted Software Development	$926K
Dell PowerEdge XE7740 with NVIDIA RTX PRO 6000 BW GPUs	AI-agent Assisted Knowledge Worker	$18M
Dell PowerEdge XE7745 with NVIDIA H200 NVL GPUs	AI-agent Assisted Software Development	$90M

The Future Economics of Agentic AI

The transition from interactive AI assistants to persistent autonomous agents is reshaping enterprise AI economics. Traditional chatbots are a sprint; agentic AI is a marathon. Because these systems operate autonomously and continuously, they consume exponentially more tokens than single-turn workloads.

As organizations scale AI adoption, infrastructure strategy is becoming an increasingly important factor in controlling operational AI costs. The findings in this analysis suggest that a portfolio-based approach spanning workstations and enterprise AI infrastructure can help organizations align performance, scalability, and economics across a wide range of agentic workloads.

Cloud-based AI APIs will continue to play an important role in enterprise AI strategies, particularly for burst capacity and rapid deployment. While cloud offers convenience for specialized workflows, it additionally introduces challenges around cost-predictability and data gravity. When considering persistent, high-utilization AI agents, the economics increasingly favor on-premises infrastructure as organizations move from pilot projects toward production-scale deployments.

The results of this analysis demonstrate that Dell AI Factory with NVIDIA can deliver significant long-term economic advantages across multiple deployment models, with rapid breakeven periods and substantial reductions in total AI operating costs compared to cloud-based API consumption.

This research represents an initial economic analysis of agentic AI deployment models using currently available on-premises performance and pricing data from Dell Technologies and NVIDIA. Additional economic improvements are expected as AI hardware, software stacks, and inference optimization techniques continue to evolve. For example, NVIDIA has demonstrated up to a 35x reduction in cost per million tokens between Hopper and Blackwell architectures. Future Signal65 research will expand on these findings through additional benchmarking, real-world workload validation, and deeper analysis of evolving agentic AI deployment architectures.

Signal65 believes that the Dell AI Factory with NVIDIA provides economically strategic options for deploying Agentic AI solutions. To learn more about the Dell AI Factory with NVIDIA, visit Dell.com/NVIDIA-AI.

Appendix

Key Assumptions

- All agents were assumed to operate up to 24 hours a day for 260 days a year
- Hardware utilization was modeled based on workload
  - AI-agent Assisted Knowledge Worker – 60% utilization
  - AI-enabled Sales – 60% utilization
  - AI-agent Assisted Software Development – 80% utilization
- Cost comparisons were completed over a 2 year time horizon

Agent Personas

The workloads modeled in this study are based on three agent personas that represent varying levels of complexity. Personas were modeled based on common AI use cases and the approximate tokens required.

Persona	Input Tokens Consumed /Agent/Day	Output Tokens Consumed /Agent/Day	Total Tokens Consumed /Agent/Day
AI-agent Assisted Knowledge Worker	~13.1M	~205K	~13.3M
AI-enabled Sales Agent	~16M	~250K	~16.3M
AI-agent Assisted Software Development	~21M	~330K	~21.3M

The AI-agent Assisted Knowledge Worker represents a low complexity agent, primarily completing low and moderate complexity tasks, such as email writing or text summarization. This agent has the lowest overall token utilization and can leverage small models.

AI-agent Assisted Knowledge Worker (Low Complexity) **~13.3M Total** Tokens/Agent/Day		
Use Case	**Description**	**Complexity**
Simple Q&A	Basic questions, quick facts	Low
Learning / Tutoring	Explanations, examples, Q&A sessions	Low
Email Writing	Draft emails, responses	Low
Meeting Transcription Summary	Converting hour-long meetings to summaries	Low
Text Summarization	Summarizing documents, articles	Low
Document Editing	Revising large documents	Low
Research Assistant	Detailed explanations, research synthesis	Moderate
Complex Problem Solving	Multi-step reasoning, planning	Moderate
Data Analysis	Processing CSV / datasets, generating insights	Moderate
Content Creation	Blog posts, marketing copy, creative writing	Moderate

The AI-agent Assisted Software Development workload represents a highly complex agent, with tasks such as code generation and API integration. This agent requires the highest overall token utilization and utilizes large models.

AI-agent Assisted Software Development (High Complexity) **~21.3M Total** Tokens/Agent/Day		
Use Case	**Description**	**Complexity**
Documentation Writing	API docs, technical writing	Moderate
Technical Research	Looking up patterns, comparing approaches	Moderate
Code Completion	Inline IDE completions	High
Code Generation	Generating new functions / components	High
API Integration	Connect services, implement endpoints	High
Code Review	Reviewing code, suggestions	High
Bug Troubleshooting	Debugging, error analysis	High
Testing Assistance	Unit tests, test planning	High
SQL Query Generation	Writing / explaining SQL	High
Code Refactoring	Restructuring existing code	High

The AI-enabled Sales agent represents a medium complexity agent with its workload mix derived from the high and low complexity agents. In total, the AI-enabled Sales Agent is comprised of 50% moderate complexity tasks and 50% high complexity tasks and utilizes ~16.3M tokens/agent/day.

Cloud Costs

Cloud API costs utilized represent average cloud pricing for models aligned to the complexity requirements of each persona. All cloud pricing was compiled as of May 1st 2026. Cached input tokens were calculated as 10% of the input token price, with cache hit rates assumed to be 40%, 50%, or 65% depending on the agent complexity. An additional 40% discount was applied to all cloud costs.

Agent	Average Input Cost ($/MTokens)	Cached Input Cost ($/MTokens)	Cache Hit Rate	Average Output Cost ($/MTokens)	Cloud Discount
AI-agent Assisted Knowledge Worker	$0.3666	$0.03666	40%	$1.7753	40%
AI-enabled Sales Agent	$2.1545	$0.21545	50%	$11	40%
AI-agent Assisted Software Development	$5.54	$0.554	65%	$28.97	40%

On-Premises Hardware

On-premises solutions were modeled utilizing pricing and performance data provided by Dell and NVIDIA. For each platform, workloads were sized based on an assumed performance threshold of 10 output tokens/second/agent. Each agent persona was aligned to model sizes to match its complexity.

Agent	On-Premises Model
AI-agent Assisted Knowledge Worker	Nemotron-3-Nano-30B-A4B
AI-enabled Sales Agent	Mixed models (50% Nemotron-3-Nano-30B-A4B, 50% Nemotron-3-Super-120B-A12B)
AI-agent Assisted Software Development	Nemotron-3-Super-120B-A12B

Solution	Price	**Max Output Tokens/s** (Nemotron-3-Nano-30B-A4B, 10 TPS/Agent, 100% Utilization)	**Max Output Tokens/s** (Nemotron-3-Super-120B-A12B, 10 TPS/Agent, 100% Utilization)
Dell Pro Max with NVIDIA GB10 GPU	~$6K	32	5
Dell Pro Max T2 with NVIDIA RTX PRO 6000 BW GPU	~$20K	300	87.5
Dell Pro Max with NVIDIA GB300 Ultra GPU	~$135K	1150	280
Dell PowerEdge XE7740 with NVIDIA RTX PRO 6000 BW GPUs	~$1.1M	71,188	Not yet tested
Dell PowerEdge XE7745 with NVIDIA H200 NVL GPUs	~$1.2M	Not yet tested	24,053

Energy consumption costs were added to desktop pricing, while colocation, software, energy, and infrastructure management costs were added to each server solution's pricing.

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

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

signal65

IN PARTNERSHIP WITH

DELL Technologies

signal65

CONTACT INFORMATION
Signal65 I signal65.com