



Dell Technologies

NVIDIA

Red Hat

Scaling Your AI May Be Easier – and Cost Less – Than You Think

A practical, familiar path to operational AI with
Dell Technologies, NVIDIA, and Red Hat OpenShift

Get started →

Scaling Your AI May Be Easier – and Cost Less – Than You Think

Executive summary

Scaling AI into production is often harder – and more expensive – than expected, not because of the models, but because of operational complexity. For many enterprises, on-premises infrastructure continues to anchor production AI, offering the control, security, and cost predictability required as AI scales. Dell AI Factory with NVIDIA, validated with Red Hat OpenShift, provides a practical, familiar path to operational AI. By extending platforms enterprises already trust, organizations can simplify AI operations, apply consistent governance, and scale AI with greater predictability and lower total cost.

Table of contents

- From AI Momentum to AI Reality 3
- Acceleration Changes the Stakes 4
- Why Orchestration Becomes the Differentiator 5
- The Hybrid Operating Reality of Enterprise AI 6
- Familiar Orchestration, Expanded Capabilities 7
- Built as a System, Not a Stack 8
- From Experiments to Repeatable Operations 9
- Scaling Without Cost Sprawl 10
- Validated Economics in the Real World 11
- Designed for What Comes Next 12
- Operationalize AI with Confidence 13

Key topics in this eBook

What changes when AI scales, and why operations matter more than models

How enterprises can operationalize AI using platforms they already trust

How production AI can deliver up to 38% lower five-year TCO

Where AI gets real →

From AI Momentum to AI Reality

Why so many AI initiatives stall between pilot and production

AI experimentation is accelerating across the enterprise, but successful pilots often mask more difficult challenges ahead: operating AI reliably, securely, and predictably at production scale.

Where things break down

As AI moves closer to core business processes, operational gaps emerge around governance, security, lifecycle management, and cost control – areas rarely tested during early experimentation.



Why this matters now

Without a deliberate operational foundation, AI initiatives fragment quickly, increasing risk and complexity just as organizations attempt to scale what initially looked promising.

What happens when AI scales? →



Acceleration Changes the Stakes

How NVIDIA-powered AI raises both performance and operational demands

AI acceleration dramatically expands what organizations can achieve, enabling more complex models, faster inference, and broader AI adoption across increasingly critical business functions.

Performance comes with pressure

As accelerated AI workloads move into production, expectations rise around reliability, security, governance, and uptime, demands that exceed what most pilot environments were designed to support.

Complexity scales too

Without coordinated operations, acceleration can amplify fragmentation, increasing integration effort, operational overhead, and risk precisely when organizations need AI to be dependable.

Speed alone is not enough →



Why Orchestration Becomes the Differentiator

Managing accelerated AI without reinventing operations

As AI workloads multiply and diversify, the challenge shifts from deploying models to coordinating how those workloads are governed, managed, and operated consistently at scale.

Consistency over customization

Without a common orchestration layer, teams resort to one-off integrations, which increase complexity and make AI environments harder to secure, govern, and maintain over time.

Operations define outcomes

Reliable AI performance depends less on individual components and more on how workloads are orchestrated, monitored, and controlled across infrastructure and environments.

AI rarely runs in one place →



The Hybrid Operating Reality of Enterprise AI

Running AI where data, security, and the business require

Enterprise AI rarely operates in a single environment, as data gravity, security requirements, latency, and regulatory constraints dictate where workloads must run.

For most enterprises, on-premises environments remain the operational anchor, providing the security, control, and predictability required as AI scales across hybrid deployments.

One strategy, many locations

AI workloads increasingly span on-premises, cloud, and edge environments, creating operational complexity when deployment models, tools, and controls differ across locations.

Consistency becomes critical

Without a unified operational approach, hybrid AI environments fragment quickly, increasing risk and slowing the path from experimentation to dependable production.



Hybrid environments demand familiar control →



Familiar Orchestration, Expanded Capabilities

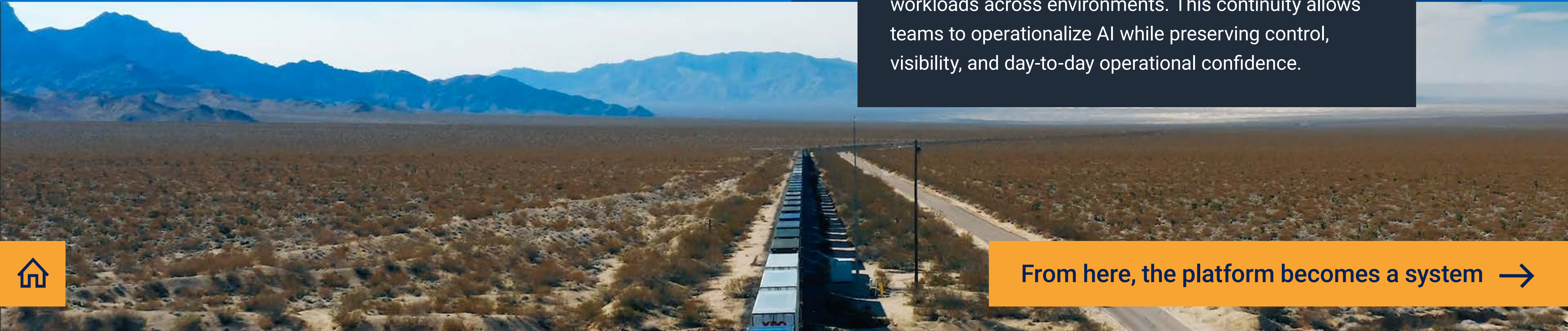
Using Red Hat OpenShift to orchestrate
accelerated AI at scale

As AI workloads accelerate and become more distributed, organizations need a way to scale without introducing entirely new operational models. The most effective path forward builds on platforms teams already trust, particularly in on-premises environments, extending existing tools, processes, and governance to support production AI.



Orchestration without reinvention

Red Hat OpenShift provides a familiar orchestration layer that enterprises already use to run applications at scale, now applied to consistently manage accelerated AI workloads across environments. This continuity allows teams to operationalize AI while preserving control, visibility, and day-to-day operational confidence.



From here, the platform becomes a system →



Built as a System, Not a Stack

Dell infrastructure, NVIDIA acceleration, and OpenShift orchestration, working together

Dell: The trusted foundation

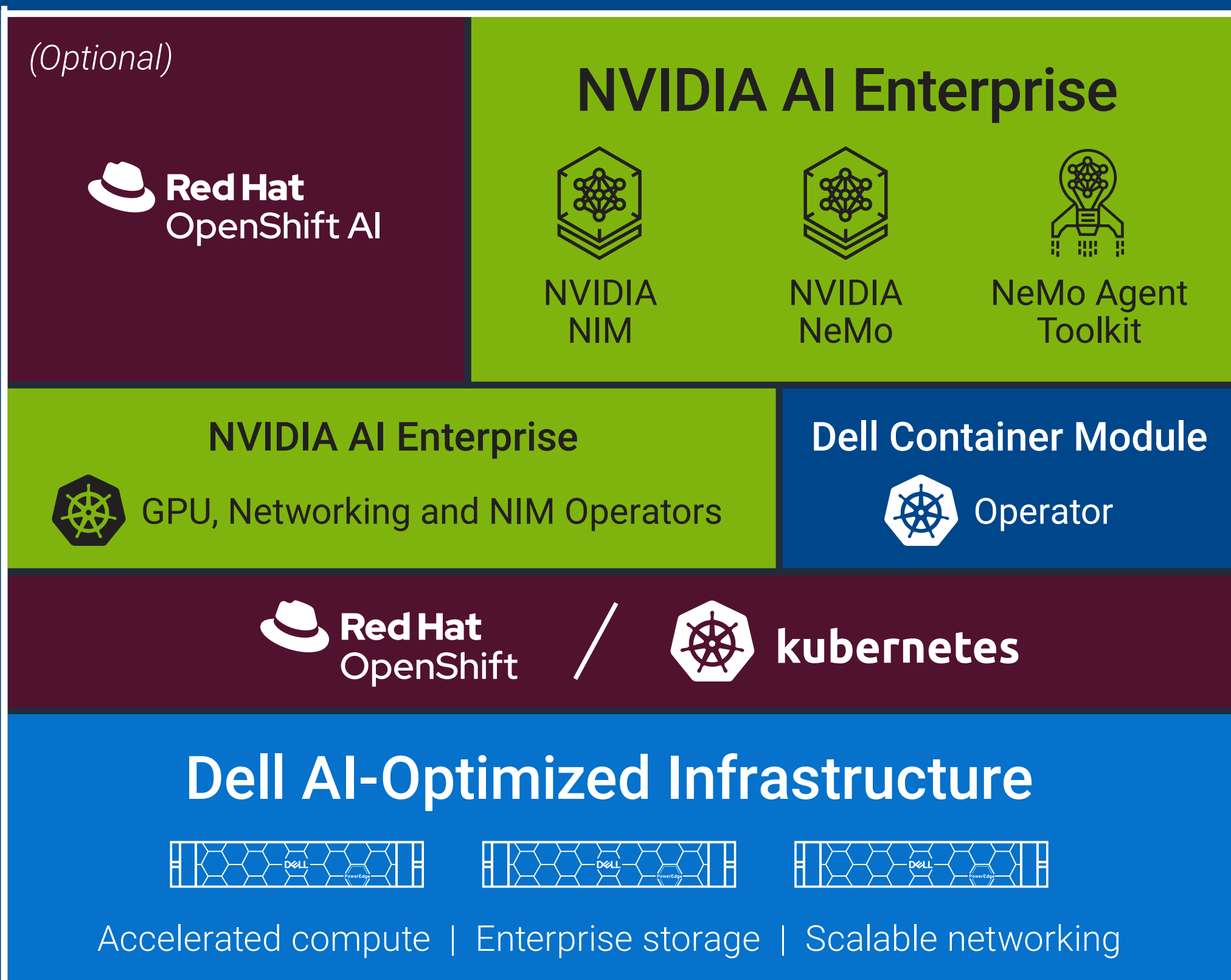
Dell provides the secure, reliable on-premises infrastructure foundation required to run AI as a production system, not a series of experiments. Purpose-built platforms, validated architectures, and enterprise-grade services establish consistency, resilience, and predictability across environments, creating the operational bedrock AI depends on as it scales.

NVIDIA: Acceleration that demands coordination

NVIDIA delivers the acceleration and enterprise AI software that power modern AI workloads, enabling higher performance and more advanced use cases. When integrated into a validated system and orchestrated consistently, acceleration becomes a force multiplier, enhancing capability without introducing operational instability.

With a system in place, operations can repeat →

AI USE CASES



STRATEGIZE

IMPLEMENT

ADOPT

SCALE

Dell AI Factory with NVIDIA

From Experiments to Repeatable Operations

Simplifying AI workflows, governance, and Day 2 management

With the right platform in place, AI can move beyond isolated experiments into repeatable operations that support consistent performance, governance, and reliability at scale.

From one-off to standard practice

Standardized workflows reduce friction across development, deployment, and operations, allowing teams to scale AI without reinventing processes for every new model or use case.

Day 2 becomes the focus

As AI enters production, success depends on monitoring, lifecycle management, and operational discipline—not just initial deployment or model performance.

Consistency changes the economics of scale →



Scaling Without Cost Sprawl

How operational simplicity drives predictable economics

As AI operations scale, complexity often drives unexpected cost growth, making it difficult to predict spending or sustain momentum beyond early successes.



What does AI really cost over time?

Consistent operational models reduce waste, minimize duplication, and improve resource utilization across environments, helping organizations scale AI more efficiently.

Predictability replaces surprises

When AI is operated consistently, cost drivers become visible and manageable, enabling teams to plan capacity, optimize utilization, and avoid runaway spending.



Proof matters when budgets scale →



Validated Economics in the Real World

Independent TCO findings for accelerated AI
with Dell AI Factory and OpenShift

Scaling AI predictably requires more than architectural confidence.
It requires economic validation.

Total Cost of Ownership (TCO) benefits
Up to 38% over 5 years

With an on-premises Dell AI Factory with
NVIDIA solution utilizing Red Hat OpenShift
vs. leading enterprise cloud AI alternative

The cost advantages observed were driven largely by operating
accelerated AI workloads on-premises, where infrastructure,
utilization, and governance can be tightly controlled.

How the analysis was conducted

Principled Technologies modeled a real-world generative
AI workload, including data science development, data
processing, model fine-tuning, and inference. The study
compared on-premises Dell AI Factory infrastructure
against a leading cloud AI service, using consistent
workload assumptions, enterprise-grade components,
and transparent cost modeling across a five-year horizon.

Why this research matters

Independent,
third-party analysis

Realistic enterprise
AI workload scenarios

Five-year cost
modeling, not
short-term estimates

Includes infrastructure,
software, and operational
considerations

Economics for today, adaptability for tomorrow →



Designed for What Comes Next

Evolving AI capabilities without
repeated platform resets

AI platforms must evolve as models, tools, and accelerators change, without forcing organizations to repeatedly redesign architectures or disrupt existing operations.

Change is the constant

New AI capabilities arrive faster than enterprise platforms can be rebuilt, making adaptability essential for sustaining momentum beyond initial deployments.

Continuity protects progress

Platforms that extend familiar operational models allow organizations to adopt what's next without sacrificing governance, reliability, or day-to-day stability.



With adaptability in place, action follows →



Operationalize AI with Confidence

A practical path forward with Dell, NVIDIA, and Red Hat OpenShift

Scaling AI does not require reinventing enterprise operations. By building on trusted on-premises infrastructure, introducing acceleration thoughtfully, and orchestrating workloads with familiar tools, organizations can move from experimentation to production with greater confidence, control, and predictability, both operationally and economically.

Key takeaways



AI scales when operations scale



Familiar orchestration reduces risk



Acceleration changes the equation, not the operating model



Systems beat stacks



Predictability enables progress

Dell

Provides the trusted, secure infrastructure foundation and validated architectures required to operate AI as a production system at enterprise scale.

NVIDIA

Delivers the acceleration and enterprise AI software that powers advanced AI workloads and unlocks higher performance across training and inference.

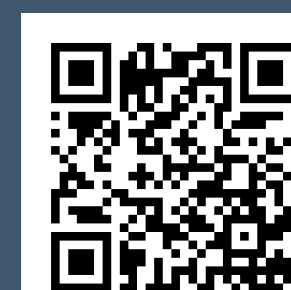
Red Hat OpenShift

Serves as the familiar orchestration layer that brings accelerated AI workloads under consistent management, governance, and control.

Learn more

Scan the QR code or visit:

www.dell.com/nvidia-ai



Explore Dell AI Factory with NVIDIA, validated with Red Hat OpenShift, and learn how to operationalize AI with confidence.

