MODERN ENTERPRISE PRESENTS

The Rack-Scale Readiness Checklist

A CLOUD AND HYBRID CHECKLIST FOR AI LEADERS

For enterprises that rely on cloud infrastructure for their AI workloads, here are three strategic questions to bring to your next provider conversation. These cut through the buzz and get to the real issues of performance, cost, and flexibility.

IF YOU'RE **CLOUD FIRST** WHAT TO ASK YOUR PROVIDER *before* YOU SCALE

01

Are rack-scale nodes offered as a managed service—not just bare metal?

WHAT IT MEANS

A **rack-scale node** is a fully integrated set of hardware: processors, memory, power, cooling, and networking engineered to work as a single, high-performance unit. Unlike traditional server racks that are loosely connected, rack-scale nodes are optimized for massive AI workloads, acting more like one giant machine than a collection of smaller ones.

Bare metal, on the other hand, refers to raw physical servers that cloud providers make available without any pre-installed software or management layers. While powerful, bare metal typically puts the burden of setup, security, orchestration, and scaling on your IT team.

WHY IT MATTERS

Choosing a managed service means the provider handles those complexities for you, so your team can focus on running models, not maintaining infrastructure. For most enterprises, this reduces operational risk, speeds time-to-value, and lowers total cost of ownership.

02

Do we have visibility into energy-per-token or energy-per-inference metrics?

WHAT IT MEANS

A **token** is a unit of text that an AI model reads or generates—usually a word fragment, like “market” or “ing.” When a model processes a prompt or generates a response, it does so **one token at a time**.

Inference is the process of using a trained AI model to make predictions or generate outputs. Per-inference refers to the energy cost of completing a single task, like answering a customer query, generating a forecast, or analyzing a document.

WHY IT MATTERS

These metrics tell you **how much energy is consumed every time your AI model is used**—token by token, task by task. As workloads scale, energy use directly impacts your operating costs, emissions profile, and infrastructure planning.

Without visibility, it’s difficult to control costs, compare providers, or meet ESG goals tied to efficiency and sustainability.

03

Is there native support for fine-tuning and inference, not just model training?

WHAT IT MEANS

Training is the process of teaching an AI model from scratch, using vast amounts of data to help it learn patterns, relationships, and rules. This is how large models like GPT or BERT are initially created—and it typically requires massive compute resources.

Fine-tuning is the next step. It takes a pre-trained model and adapts it to your specific business context—whether it's financial data, legal language, or customer behavior. It's far less resource-intensive than full training, but far more relevant to your use case.

Inference in this context is what happens after the model is trained and fine-tuned: it's the act of using the model to generate answers, make predictions, or take action. In a business setting, this is the stage where AI creates real value—responding to customers, scanning documents, analyzing trends.

WHY IT MATTERS

Most enterprises don't need to train models from scratch, but they absolutely need to fine-tune and run them. Make sure your cloud provider or infrastructure supports these later-stage functions, or you'll be stuck paying for capabilities you don't use and missing the performance you need.

For enterprises exploring a hybrid strategy or maintaining critical infrastructure on-premises, rack-scale readiness isn't just about buying hardware—it's about future-proofing your ecosystem.

IF YOU'RE **HYBRID OR ON-PREM** WHAT TO VET *before* YOU INVEST

01

Does the rack-scale system integrate with our existing network fabric?

WHAT IT MEANS

Your network fabric is the high-speed infrastructure—cables, switches, and software—that connects all the servers and systems within your data center. Think of it like the circulatory system that moves data across your organization.

Integration means the new rack-scale system can plug into and communicate efficiently with what you already have—without requiring major rewiring, hardware replacements, or complex workarounds.

WHY IT MATTERS

Rack-scale systems often rely on specialized, high-bandwidth interconnects to achieve their performance. If these aren't compatible with your existing setup, you could face unexpected delays, cost overruns, or reduced performance.

Confirming integration in advance ensures your investment fits into your broader architecture—without blowing up your timeline or your budget.

02

Is the software stack containerized and vendor-agnostic (e.g., K8s, Slurm)?

WHAT IT MEANS

A **container** is a lightweight, portable software package that includes everything an application needs to run—code, settings, and dependencies. Containers allow AI models and tools to be moved easily between systems without breaking or needing to be reconfigured.

Kubernetes (K8s) is an open-source system for managing and scaling these containers across different machines or environments. Think of it as an automated traffic controller that keeps workloads running smoothly and efficiently, even at scale.

Slurm is another orchestration tool, often used in high-performance computing environments. It schedules and allocates resources—like GPUs or CPUs—for AI and research workloads.

WHY IT MATTERS

A **containerized, vendor-agnostic stack** gives your enterprise flexibility. It means you're not locked into one hardware provider or cloud service, and you can shift workloads between systems as needs evolve. This reduces risk, increases negotiating power, and supports long-term scalability—on your terms.

03

Is there a modular roadmap? Can we scale with add-ons?

WHAT IT MEANS

A **modular roadmap** means your infrastructure can grow piece by piece, rather than requiring a massive upfront investment. You can add components as your needs evolve, rather than overbuilding from the start. Examples can include

Storage refers to the systems that hold your data—everything from training datasets to real-time customer interactions. Scalable storage is essential to feed AI models efficiently.

An **inference-only rack** is built to run models that are already trained. These racks are optimized for fast, cost-effective predictions—like answering queries, analyzing documents, or powering chatbots at scale.

Network accelerators improve how quickly data moves between systems. They reduce bottlenecks in high-speed environments, which is especially important in AI workloads where latency can impact performance and cost.

WHY IT MATTERS

A modular approach lets you invest in what you need now, with the confidence that you can scale later—without starting over. It also allows you to optimize for different parts of the AI lifecycle (e.g., training vs inference) as your use cases mature. This keeps infrastructure aligned with business needs, not vendor roadmaps.

Reminder: Rack-scale isn't the destination. It's a design choice. The best-performing enterprises don't just buy infrastructure, they build strategy. Use this checklist as a **conversation starter** with your vendors, your IT leaders, and your board.

MODERN ENTERPRISE

This is not a marketing agency.

It's a strategic studio and consultancy
for companies at an inflection point.

CLARITY IS CURRENCY
STORY IS STRATEGY
TIMING IS EVERYTHING

Fractional CMO

Go-to-Market Strategy

Narrative Strategy

Executive Voice

modernenterprise.ai
studio@modernenterprise.ai

Meet the Founder

Wallis Mills

You can’t call Olivia Pope, Jessica Pearson, or Annalise Keating, but you can call Wallis Mills. With 25 years of experience shaping narratives and go-to-market strategy for companies like AMD and NVIDIA, Wallis has advised founders, executives, and boards across AI, infrastructure, FinTech and capital markets.

Her work sits at the intersection of strategy, story, and leadership and has shaped trillion and billion-dollar businesses, investor pitches, and executive platforms. Wallis has a reputation for cutting through complexity and delivers quietly and decisively — when it matters most.


ME
MODERN
ENTERPRISE