



Broadcom Announces VMware AI Factory, Enabling Faster Time to Production AI and Greater Control Over AI Tokenomics

August 31, 2026

New AI Infrastructure Automation and Private AI Services Deliver a Single Consistent Environment for Deploying, Governing, and Securing AI Workloads from Bare Metal to Inference

LAS VEGAS, Aug. 31, 2026 (GLOBE NEWSWIRE) -- **VMware Explore 2026** -- Broadcom Inc. (NASDAQ: AVGO), a global technology leader that designs, develops, and supplies semiconductor and infrastructure software solutions, today announced VMware AI Factory, the software-defined foundation of VMware Private AI Cloud. VMware AI Factory provides customers a simplified path to production AI with new automation innovations for deploying AI-ready infrastructure and supporting Day 2 operations. With VMware AI Factory, customers can achieve faster time to first model deployment and better manage AI tokenomics.

"Enterprises want to run AI where their data lives, but the journey from metal to model is slow, complex, and expensive," said Paul Turner, chief product officer, VMware Cloud Foundation Division, Broadcom. "VMware AI Factory changes that. We give customers a software-defined foundation that automates infrastructure deployment, unifies lifecycle management, and lets them choose their preferred hardware and vetted models. The result is faster time to first model, predictable private cloud costs, and better control over AI tokenomics."

VMware AI Factory Accelerates Time to First Model

VMware AI Factory brings AI applications directly to enterprise private data within a secure private cloud environment. VCF's unique infrastructure automation capabilities can reduce the time from bare metal server deployment to serving the first AI model from weeks to a matter of hours. VMware AI Factory streamlines AI infrastructure management by fully automating hardware provisioning, software stack enablement, and end-to-end lifecycle management. By integrating hardware and software operations into a unified, automated solution, organizations can rapidly scale AI workloads while minimizing operational complexity.

As part of the VMware AI Factory, private AI services help make AI operational, governable, and cost-effective. VCF pools and shares GPU resources across the organization so teams can run multiple models on shared hardware instead of dedicating infrastructure to each workload. A unified model gallery gives IT and data science teams a single interface for deploying and managing model inference and RAG workflows across VMs, containers, and GPU resources, with built-in observability into token throughput, latency, and compute and memory utilization. Enterprises can pivot to new models while keeping costs low through shared infrastructure and governed models-as-a-service. New and forthcoming private AI services include:

- **Multi-tenant Model Sharing:** Model Runtime now supports secure sharing of AI models between tenants or lines of business through isolated namespaces, maintaining data privacy while eliminating redundant model deployments that waste GPU allocation and infrastructure resources.
- **AI Gateway:** Unified model governance between on-premises and cloud environments through a single consumption interface, enabling access to locally hosted models with enhancements like intelligent prompt routing, token and usage rate-limiting, and application authorization.
- **Secure AI Sandboxes and Governance:** Secure virtualized container spaces will isolate dynamic agent-generated code execution, with a control layer defining how agents are invoked, what tools they access, and how their outputs are validated before being acted upon.

To further streamline VCF AI Factory deployment, Broadcom is announcing a new partnership with MetalSoft to deliver integrated heterogeneous bare metal automation for VCF that drops bare-metal provisioning time from weeks to minutes. The integration will help IT provision or repave physical servers from multiple vendors directly through the VCF management console, unifying the software and hardware lifecycle into a single operational model and eliminating the need for vendor-specific tools for hardware and firmware management.

VMware AI Factory combines VMware Cloud Foundation (VCF) with certified VCF AI ReadyNodes from Cisco, Dell Technologies, Lenovo, Supermicro and others¹ and customers' preferred AI software and accelerator architectures. Broadcom and AMD are collaborating to deliver a VMware AI Factory that pairs VCF with AMD Instinct GPUs and the open AMD ROCm software ecosystem. Zero-touch provisioning will orchestrate the end-to-end deployment of the entire stack, from vSphere and vSAN through Kubernetes and the AMD GPU operator, and the AMD DVX driver can attach GPUs to large VMs consumed by a VMware vSphere Kubernetes Service cluster.

VMware AI Factory gives enterprises a production-ready path to running leading AI models on-premises. VCF customers can run more than 150 open source and commercial models, including Nemotron 3, Gemma 4, cotomi, Qwen 3.7-Max, and GLM 5.2. Broadcom is working with the world's leading AI model providers to give organizations a clear path to data sovereignty and cost-effective AI at scale, with leading models available securely and delivered as a service to their user community through VCF's built-in services. [Read the full announcement here.](#)

Partner Commentary

"Enterprises adopting AI need freedom to choose their infrastructure without locking into a single vendor stack," said Suresh Andani, corporate vice president, Compute and Enterprise AI Group, AMD. "Pairing AMD Instinct MI350 Series GPUs with the open ROCm software ecosystem on VMware Cloud Foundation gives customers a validated path to private AI with predictable economics and no per-token pricing. This collaboration lets organizations run production AI where their data lives while maintaining choice and control over their full stack."

"Moving AI into production requires more than GPUs. It requires an infrastructure stack that operates as one seamless foundation from the data center to the edge," said Jeremy Foster, Senior Vice President and General Manager, Cisco Compute. "Cisco's validated compute and an open, high-performance network are aligned with NVIDIA Enterprise Reference Architectures and Cisco Validated Designs with AMD. As we work towards

VMware AI Factory's new infrastructure automation, this will give customers a simpler path from bare metal to production AI, while preserving flexibility in the workloads and accelerator architectures they choose."

"As enterprises scale AI, token economics is emerging as a critical driver of adoption and ROI," said Scott Patti, Vice President, Infrastructure Solutions Group, Lenovo. "VMware AI Factory on Lenovo Infrastructure Solutions gives customers the flexibility to run AI where it makes the most economic and operational sense, helping them accelerate production deployments with predictable costs, stronger governance, and greater business value."

"Supermicro delivers its application-optimized Data Center Building Block Solutions® (DCBBS) infrastructure with the industry-leading time to deployment," said Vik Maylaya, chief business officer, Supermicro. "Our SuperCloud Director supports faster implementation of multi-tenancy bare metal provisioning, combining Supermicro's Total IT solutions with VMware AI Factory to give customers factory-validated flexible and cost-efficient building blocks for modern private cloud and AI workloads for the enterprise."

"Managing physical servers has always been a separate operational domain from software, creating silos that slow down AI infrastructure deployment," said Lucas Roh, founder and CEO, MetalSoft. "MetalSoft's bare metal automation now extends VCF's operational workflows down to the physical layer, letting IT treat physical servers like code. Customers can provision or repave heterogeneous hardware directly through the VCF management console, closing the gap between hardware and software operations and turning weeks of manual work into minutes."

[1-Broadcom Compatibility Guide](#)

Additional Resources

- Learn more about [VMware Cloud Foundation](#)
- Follow VMware Cloud Foundation social channels on [LinkedIn](#), [X, formerly known as Twitter](#) and [YouTube](#)

About VMware Explore

VMware Explore is the established cloud event for IT professionals to advance their skills and credentials. VMware Explore 2026 will connect attendees with technology experts and solution architects who are solving real-world challenges. Focused technical sessions, Hands-on Labs, and certification opportunities will enable attendees to apply their learnings and deliver greater value back to their organizations. Attendees will leave with the training, tools, and strategies to build and operate a modern private cloud that is AI-native—running any workload while keeping data local and secure. To learn more about VMware Explore, please visit: <https://www.vmware.com/explore>

About Broadcom

Broadcom Inc. (NASDAQ: AVGO) is a technology leader that designs, develops, and supplies semiconductors and infrastructure software for global organizations' complex, mission-critical needs. Broadcom combines long-term R&D investment with superb execution to deliver the best technology, at scale. Broadcom is a Delaware corporation headquartered in Palo Alto, CA. For more information, visit www.broadcom.com.

Broadcom, the pulse logo, and Connecting everything are among the trademarks of Broadcom. The term "Broadcom" refers to Broadcom Inc., and/or its subsidiaries. Other trademarks are the property of their respective owners.

Media Contact:

Roger T. Fortier
VMware Cloud Foundation Division, Broadcom
+1.408.348.1569
roger.fortier@broadcom.com