Broadcom Announces VMware Cloud Foundation 9.1, Enabling Secure and Cost-Effective Infrastructure for Production AI

May 5, 2026

VCF 9.1 Empowers Enterprises with Accelerated AI Workload Deployment at Lower Costs, Enhanced Security, and an Open Hardware Ecosystem

PALO ALTO, Calif., May 05, 2026 (GLOBE NEWSWIRE) -- Broadcom Inc. (NASDAQ: AVGO), a global technology leader that designs, develops, and supplies semiconductor and infrastructure software solutions, today announced VMware Cloud Foundation (VCF) 9.1, a secure and cost-effective infrastructure platform for production AI workloads. VCF 9.1 delivers an AI and Kubernetes native private cloud platform with integrated security and mixed compute infrastructure support across AMD, Intel, and NVIDIA. This enables enterprises to deploy inference and agentic AI applications with significantly lower costs, enhanced security, and freedom to choose best-of-breed GPU and CPU hardware.

A preview of Broadcom's Private Cloud Outlook 2026 report reveals private cloud continues to be the preferred platform for production AI. More than half of organizations surveyed (56%) are running or planning to run production inferencing in a private cloud. More importantly, public cloud use for production inference was 41%, down 15% year over year. Additionally, 62% of IT leaders reported being very or extremely concerned about generative AI infrastructure costs while 36% report AI is driving new requirements for data protection, privacy, security controls and risk management.

VMware Cloud Foundation provides a better alternative to public cloud for production workloads through intelligent software that maximizes infrastructure efficiency on existing servers while providing architectural control and regulatory compliance capabilities essential for production AI deployments. VMware Cloud Foundation 9.1 will enable enterprises to deploy production workloads including inference and agentic AI with:

- Up to 40% reduction in server costs through intelligent memory tiering for clusters running a mix of AI and non-AI workloads[1];
- Up to 39% lower storage TCO through enhanced compression and deduplication for AI data pipelines[1];
- Up to 46% reduction in Kubernetes operational costs for running AI workloads at scale[1];
- 4x faster cluster upgrades and 2x increased fleet capacity to rapidly scale AI infrastructure[1].

"As more enterprises turn to AI for driving competitive advantage, they face three critical challenges: data and IP privacy concerns, surging infrastructure costs, and their readiness for the world of agentic AI," said Krish Prasad, senior vice president and general manager, VMware Cloud Foundation Division, Broadcom. "VCF 9.1 is a single unified platform that addresses all three and delivers one of the most advanced infrastructure for Private AI. We enable zero-trust security for AI, reduce costs through intelligent infrastructure optimization and hardware choice, and provide the flexibility to run both agentic workflows and accelerated inferencing on the same platform."

Efficient Infrastructure at Scale for AI Workloads

VCF 9.1 maximizes density for both VM and containerized AI workloads on existing infrastructure while dramatically reducing operational complexity. Through intelligent resource management and automated operations, enterprises can deploy more production workloads on current servers, scale efficiently across distributed environments, and eliminate the need for costly infrastructure expansion during a period of hardware shortage and rising costs. Key capabilities include:

- **Intelligent resource optimization** that maximizes infrastructure utilization through advanced memory tiering and next-generation storage compression for AI data pipelines, enabling higher AI workload density without performance compromises or expensive hardware refresh.
- **Automated fleet operations** at scale that deliver doubled management capacity to 5,000 hosts and 4x faster cluster upgrades across distributed and air-gapped environments, eliminating manual patching overhead while supporting rapid AI infrastructure expansion.
- **Multi-tenant infrastructure for AI isolation** that enables enterprises and service providers to run multiple AI projects and customers on shared infrastructure with strict security boundaries, maximizing utilization of expensive GPU and CPU resources while supporting data sovereignty for sensitive models.
- **Open ecosystem integration** that delivers multi-accelerator GPU choice across AMD and NVIDIA, support for leading AMD and Intel CPU platforms, and standards-based EVPN and VXLAN interoperability with Arista Universal Cloud Network, demonstrating VCF's commitment to providing the high-performance connectivity and compute flexibility production AI demands.
- **High speed networking for AI workloads** through VCF support for NVIDIA ConnectX-7 NICs and NVIDIA BlueField-3 with Enhanced DirectPath I/O. With this enhancement high-speed, multi-host AI model training and data transfer, crucial for demanding Gen AI workloads is enabled.
- **Virtualized load balancing and security** with VMware Avi Load Balancer[2] and VMware vDefend[2] eliminate hardware appliance requirements for AI inference endpoints and agentic applications, reducing capital expense while providing enterprise-grade resilience and automated lifecycle management.

High Velocity App Delivery: Modern Workload Platform for AI, Containers, and VMs

VCF 9.1 delivers a unified platform that accelerates AI application deployment by running inference workloads, agentic applications, containerized services, and traditional VMs on a single infrastructure layer. This eliminates operational fragmentation and the cost of managing separate stacks while providing the developer velocity and platform governance that production AI requires. Key capabilities include:

- **Kubernetes scale and performance** for AI that delivers 2.6x increased cluster scale, 70% faster deployments, 75% shorter upgrade windows compared to preview versions[1], and seamless scaling that enables zero downtime for production AI services.
- **Mixed compute management** that efficiently handles both CPU-intensive agentic AI workflows and GPU-accelerated inference on a unified platform, addressing the reality that agentic workloads demand significantly more CPU than GPU capacity for workflow execution and decision orchestration.
- **AI observability and governance** that provides detailed metrics for time to first token, token throughput, and GPU utilization across multiple accelerator types, enabling enterprises to maximize infrastructure ROI through precise hardware utilization monitoring while centralized policy injection and data sovereignty controls enable AI compliance enforcement and secure model access.
- **Live application stack blueprints** that capture multi-VM applications as reusable templates for rapid environment deployment, eliminating manual configuration errors and preventing configuration drift across development, test, and production environments while accelerating infrastructure delivery velocity.

Zero-Trust Architecture for AI Data Sovereignty and Governance

VCF 9.1 integrates security at the infrastructure layer to protect AI workloads, proprietary models, and training data from hypervisor to application. By delivering zero-trust segmentation, sovereign recovery, and continuous patching without bolt-on tools, VCF strengthens the security posture essential for production AI deployments that public cloud environments cannot match. Key capabilities include:

- **On-premises ransomware recovery** that provides isolated recovery environments and integrated validation tools including new CrowdStrike Falcon® Endpoint Security support protect AI models and training data – significant intellectual property – from cross-border movement while avoiding massive bandwidth fees during crisis restoration.
- **Continuous compliance enforcement[2]** that maintains regulatory adherence through centralized monitoring and automated desired state remediation for workloads and VCF stack components, enabling enterprises to demonstrate audit readiness for production AI deployments without manual overhead or separate compliance tools.
- **Zero-downtime live patching** that supports up to 80% of use cases without host evacuation or maintenance windows, eliminating disruption to production AI inference services and agentic applications that require continuous availability for service level agreements[1].
- **Zero-trust lateral security[2]** that extends distributed IDS/IPS protection to Kubernetes AI workloads for the first time, delivering 9 Tbps threat inspection performance for distributed inference and 5x increased application identification for private cloud and internet applications[1].
- **Self-service security with automation[2]** that provides centralized tagging, pre-defined security profiles, delegated firewall configurations and ingress web application security, enabling enterprises and service providers to secure AI deployments without operational complexity or fragmented security toolchains.

Customer and Partner Commentary

"Analyzing years of news archives in the public cloud is cost-prohibitive, with unpredictable pricing that makes AI projects difficult to plan," ***said V V Jacob, Senior General Manager, Systems for Malayala Manorama Co Ltd. "****By deploying VCF Private AI Services on our existing VMware Cloud Foundation infrastructure, we will run AI-powered content summarization, heading generation, and editorial assistance directly on our private cloud. We believe this will give us the privacy and security essential for protecting editorial sources while delivering the cost predictability that on-premises private cloud infrastructure provides."*

"By unifying our VMs and containers on VMware Cloud Foundation, we've achieved greater operational efficiency and raised the overall availability," ***said Alexander Hopfgartner, Head of Technology at Notruf Niederösterreich.*** *"VMware vSphere Kubernetes Service, as the built-in Kubernetes runtime of VCF, empowers our operations team to easily deploy, scale, and manage our most critical applications."*

"As enterprises move AI from experimentation to production, they need infrastructure that delivers performance, efficiency, and flexibility across a broad ecosystem at scale," ***said Kumaran Siva, corporate vice president, Compute and Enterprise AI, AMD.*** *"AMD enterprise AI solutions, along with VMware Cloud Foundation 9.1, enable scalable, cost-efficient AI workloads; helping customers deploy inference and agentic AI with the performance, security, and data sovereignty required for production environments."*

"Arista Networks and Broadcom share a fundamental commitment to open, standards-based networking that gives enterprises true architectural freedom and choice for production AI infrastructure," ***said Jeff Raymond, Vice President and General Manager of EOS Software and Services, Arista Networks.*** *"EVPN and VXLAN interoperability between Arista Universal Cloud Network and VMware Cloud Foundation 9.1 delivers the openness and performance that production AI requires. Through standards-based direct ESX-to-fabric connectivity, enterprises can build scalable network architectures for AI infrastructure while reducing both capital and operational costs."*

"AI workloads are now prime targets, and recovery without validation is a risk enterprises can't afford," ***said Chris Stewart, Vice President, Global Cloud and Technology Alliance Partners, CrowdStrike.*** *"With CrowdStrike integrated with VMware Cloud Foundation, organizations can stop breaches faster, validate that environments are truly clean before restoring, and prevent reinfection – critical to protecting high-value models and data while maintaining full control over sovereignty and compliance."*

"VMware Cloud Foundation 9.1 is further optimized for Intel® Xeon® 6 processors, unlocking the full potential of a high-density, AI-ready platform. Native integration of Intel® QuickAssist Technology accelerates Encrypted vMotion while freeing valuable compute resources," ***said Caitlin Anderson, Corporate Vice President, Americas Sales at Intel Corporation.*** *"Together, we remain committed to delivering continuous innovation with superior total cost of ownership, helping customers accelerate their AI and container modernization journeys."*

"Enterprises need infrastructure that delivers breakthrough AI performance while maintaining data sovereignty and control," ***said John Fanelli, vice president of enterprise software at NVIDIA.*** *"Our collaboration with Broadcom brings NVIDIA Blackwell architecture–including RTX Pro Servers equipped with BlueField-3 and the NVIDIA Blackwell HGX platform–along with high-speed DirectPath I/O to VMware Cloud Foundation. This enables organizations to deploy private AI with the same performance they expect from public cloud, but with complete control over their models and data. This collaboration addresses the reality that production AI requires both extraordinary compute power and enterprise-grade governance."*

Additional Resources

- Read all the VMware Cloud Foundation 9.1 blogs to learn about the new innovations
- Learn more about VMware Cloud Foundation
- Follow VMware Cloud Foundation social channels on LinkedIn, X, formerly known as Twitter and YouTube

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.

1-Based on internal Broadcom estimates or test results, subject to change. April 2026
2-Advanced Service for VCF sold separately

Media Contact:

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

Source: Broadcom Inc.