



Broadcom Introduces VMware Private AI Cloud, Enabling Enterprises to Scale AI Cost-Effectively, Operate More Securely, and Innovate Rapidly

August 31, 2026

Broadcom's Portfolio of Advanced Cloud Infrastructure, Application and Security Software Gives Enterprises a Production-ready Path to Building, Running, and Governing AI Where Their Data Lives

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 introduced VMware Private AI Cloud, a more secure, scalable, and flexible approach to AI that brings the model to the data, not the data to the model. Built on Broadcom's advanced software capabilities, VMware Private AI Cloud gives organizations a production-ready path to securely building, running, and governing inference workloads, agentic applications, and traditional enterprise workloads together on a single private cloud platform. Offering diverse hardware, model, and accelerator choices, VMware Private AI Cloud lets enterprises scale AI cost-effectively, operate more securely, and innovate rapidly.

"VMware Private AI Cloud is the inflection point where enterprise private cloud and private AI infrastructure stop operating as separate disciplines and become one—enabling production inference workloads and agentic AI with the data sovereignty, compliance posture, and cost predictability their business demands," said Ram Velaga, president, Infrastructure Software Group, Broadcom.

Scale AI Cost-Effectively

VMware Private AI Cloud addresses the three core AI cost drivers: hardware CapEx, operational complexity, and token economics (tokenomics). VMware Cloud Foundation (VCF) 9 lowers hardware costs via NVMe memory tiering and cluster-wide storage deduplication. VCF supports GPUs, CPUs, and accelerators from leading vendors, along with server hardware from major OEM and ODM vendors, allowing customers to run heterogeneous clusters cost-effectively. To optimize tokenomics and resource usage, it features token monitoring, multi-tenant model sharing, enhanced GPU/vGPU tracking, and an AI metrics observability dashboard. Infrastructure and operations innovations announced for Private AI Cloud include:

- **Metal to model faster with VMware AI Factory:** [Broadcom announced VMware AI Factory](#), the software-defined foundation of VMware Private AI Cloud, providing customers a simplified path to production with new automation innovations for deploying AI-ready infrastructure and Day 2 operations. With VMware AI Factory, customers can achieve faster time to first model deployment and better manage AI tokenomics.
- **Validated AI Models for VCF enable Model as a Service:** 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.

Operate More Securely

Designed with a defense-in-depth approach aligned to NIST CSF 2.0, VCF protects against AI-accelerated threats by minimizing the attack surface and enabling continuous compliance. Automated, non-disruptive updates keep systems current, while VMware vDefend uses virtual patching and hypervisor-level lateral security with microsegmentation to enforce Zero Trust and block exploits. Furthermore, vDefend's multi-layer threat defense and VMware Avi Load Balancer's web application firewall and API protection prevent sophisticated attacks. Security innovations supporting Private AI Cloud include:

- **TrueSource by Broadcom for verifiably built open source:** Designed to address the acceleration in AI-enabled exploitation, Spring Enterprise delivers secure, curated Spring releases from the team that maintains it, including frontier model scanned, human-verified patches delivered simultaneously across every release line. [TrueSource Trusted Artifacts](#) extends clean-room builds across the Java ecosystem, Python, and Node.js, along with the Bitnami Secure Images catalog while TrueSource Data Services brings the same standard to the data tier: PostgreSQL, RabbitMQ, MySQL, and Valkey.
- **Agentic Zero Trust with vDefend:** New vDefend enhancements will extend Zero Trust lateral security for agentic AI workloads by identifying agentic AI components through continuous monitoring of traffic flows, detecting unauthorized usage of shadow AI, and providing distributed virtual patching through AI-generated Intrusion Detection and Prevention (IDPS) signatures.
- **Agentic Threat Defense with Avi Load Balancer:** Avi web security will broaden its protection of agentic AI workloads by restricting agents from accessing unauthorized tools and preventing misuse; flagging and isolating anomalous behavior to detect zero day attacks; and helping to prevent the unauthorized exfiltration of sensitive data with data protection guardrails.

Innovate Rapidly for the Agentic AI Era

Unlike traditional apps, autonomous AI agents can act unchecked, exceed scope, or misinterpret instructions. Consequently, trust depends on robust controls and data integrity. VMware Tanzu Platform, with VMware vDefend, provides a foundation for trustworthy enterprise agents via a deny-by-default architecture, a prebuilt harness, and a curated marketplace. Agentic AI innovations announced for Private AI Cloud include:

- **AgentMinder by Broadcom helps govern autonomous AI agents at scale:** [Broadcom today unveiled AgentMinder](#), a new solution that provides enterprises with a central control plane for autonomous AI agents. It treats agents as enterprise-grade identities, binding their authority to a specific mission, approved tools, and authorized resources. Additionally, it provides runtime policy enforcement—ensuring least-privileged access for every tool invocation—and delivers compliance-grade auditability, giving enterprises full visibility into what their agents are doing.
- **AI-ready data foundations turn enterprise data into AI agent knowledge:** [New AI-ready data foundations in Tanzu Platform](#) let data owners build and manage dynamic pipelines across structured and unstructured data, producing data products optimized for low-cost agent consumption. Those products publish to the Tanzu Platform marketplace as governed, context-rich services that both agents and developers can discover and use without data ever leaving the enterprise.
- **Tanzu Platform agent foundations to build and run trusted AI agents:** Tanzu Platform agent foundations enforce strict containment through a deny-by-default runtime; agents have zero access to APIs, networks, MCP servers, or the internet unless explicitly granted. New enhancements include an isolated credential store that shields all credentials from agents entirely. Agents can't leak or misuse what they can't see, closing off credential theft and prompt injection attacks.

Additional Resources

- Read all of the announcements in the [VMware Explore 2026 media kit](#)
- Follow VMware 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 Contacts:

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

Eloy Ontiveros
Broadcom Global Communications
+1.650.427.6145
eloy.ontiveros@broadcom.com