

Accelerating Private AI

for a leading hospital system on VMware Cloud Foundation 9



OVERVIEW

Modern hospital systems operating across multiple facilities require secure, high-performance platforms designed to support demanding AI-driven diagnostic imaging workloads. To address escalating public cloud costs and strict HIPAA compliance requirements, ThinkOn can provide a private infrastructure solution capable of delivering GPU-accelerated inference at scale while prioritizing patient privacy.

CHALLENGES

Balancing AI ambition with compliance and cost constraints

Healthcare organizations often face significant hurdles when running diagnostic AI models in public cloud environments.

Challenges include:

- **Unpredictable costs:** Monthly compute expenses can grow substantially as imaging volumes increase, with some hospital systems seeing costs exceed \$180,000.
- **Compliance risks:** Sensitive patient data traversing external networks can raise ongoing regulatory concerns regarding patient privacy.
- **Workflow bottlenecks:** Inconsistent GPU availability during peak periods may lead to processing backlogs that can delay critical radiologist workflows by up to six hours.



SOLUTION

Private AI infrastructure purpose-built for healthcare

ThinkOn can deploy a Private AI environment on VMware Cloud Foundation 9 (VCF 9), offering a fully managed, HIPAA-aligned environment with dedicated NVIDIA GPU resources.

High-performance compute density. The platform can be configured with specific high-density hardware to support real-time medical image analysis across radiology, pathology, and cardiology departments, such as:

- 8 NVIDIA L40S GPUs
- 512 GB of GPU memory
- 2 TB of system RAM
- 120 vCPUs

Accelerated diagnostics with on-premises performance

ThinkOn's Secure Subscriber Portal provides full visibility into GPU utilization, workload scheduling, and compliance posture.

- **Near-bare-metal speed:** VCF 9's native GPU passthrough and vSphere integration allow organizations to run AI inference models with optimized performance.
- **Rapid processing:** This architecture has the potential to reduce diagnostic image processing times from an average of 14 minutes to under 90 seconds per study.
- **Real-time insights:** AI-assisted pre-reads can be made available to radiologists within two minutes of image acquisition.

Data sovereignty and regulatory confidence

By using SOC 2 Type II and HIPAA-certified infrastructure, patient imaging data remains within a secure environment, eliminating third-party data transit risks. VCF 9's micro-segmentation and NSX networking further ensure workload isolation between departments.

POTENTIAL OUTCOMES

Framed as illustrative benefits, a Private AI deployment on VCF 9 offers:

- **Significant cost optimization:** Organizations may achieve up to a 42% reduction in AI compute costs compared to public cloud, potentially saving over \$75,000 per month through dedicated GPU allocation and the elimination of egress fees.
- **Enhanced diagnostic speed:** This solution has the potential to reduce AI-assisted image processing time by up to 89%, accelerating radiologist workflows.
- **Airtight security & Compliance:** Maintain a strong compliance posture through micro-segmented networking and infrastructure designed to keep patient data within the private cloud environment.
- **Reliable high availability:** A 99.99% uptime SLA can help ensure continuous AI processing across all facilities with automated failover and real-time GPU health monitoring.

IMPLEMENTATION STRATEGY

ThinkOn's deployment process is designed to ensure operational continuity and technical alignment:

- **AI workload assessment:** Collaborative profiling of existing AI models and GPU requirements ensures the environment is right-sized from day one.
- **Seamless migration:** Using tools such as VMware HCX, ThinkOn can facilitate a zero-downtime migration of AI workloads and archives; in similar scenarios, transitions have been completed within a 6-week period.
- **Model deployment pipeline:** ThinkOn configures containerized AI model serving on VCF 9 using Kubernetes, enabling data science teams to deploy and scale models independently.
- **Guided onboarding:** Comprehensive onboarding and 24/7 support can allow operations teams to gain full self-service capabilities within two weeks of go-live.

Transformative impact of Private AI on VCF 9

Deploying Private AI on VMware Cloud Foundation 9 through ThinkOn can fundamentally transform how hospital systems approach clinical AI. By bringing GPU-accelerated inference into a managed private cloud, organizations can eliminate public cloud cost unpredictability, achieve sub-two-minute diagnostic turnaround times, and maintain rigorous HIPAA alignment across all facilities.

Reach out. We're here to help! 844.888.4465 | sales@thinkon.com

About ThinkOn

Think On, Inc. is a cloud service provider with a global data center footprint offering creative solutions for complex data problems. ThinkOn's channel-only distribution model empowers value-add resellers and service providers across North America, the UK, Europe, and Australia through solutions built on leading technology platforms.

