

Private AI Inference on Lenovo ThinkAgile VX and FX V4 with Intel Xeon 6 and VMware Cloud Foundation

Solution Brief

Enterprise interest in generative AI has moved past experimentation. Organizations are putting retrieval-augmented generation, intelligent assistants, and automated document processing into production, and for many of them the hard part now has less to do with the model and more to do with where and how it runs.

Regulatory requirements, data residency rules, and internal security policies are pushing organizations toward private AI: inference that stays on infrastructure the organization owns and governs, instead of sending sensitive data to a third-party API. At the same time, infrastructure teams don't want to build a second, parallel stack just for AI.

A private AI platform has to work with the tools, processes, and staff already running virtualization, storage, and networking. Lenovo ThinkAgile VX and FX V4 systems, powered by Intel Xeon 6 processors and built on VMware Cloud Foundation, extend that same operating model to AI inference, without a separate silo of specialized hardware and management tools.

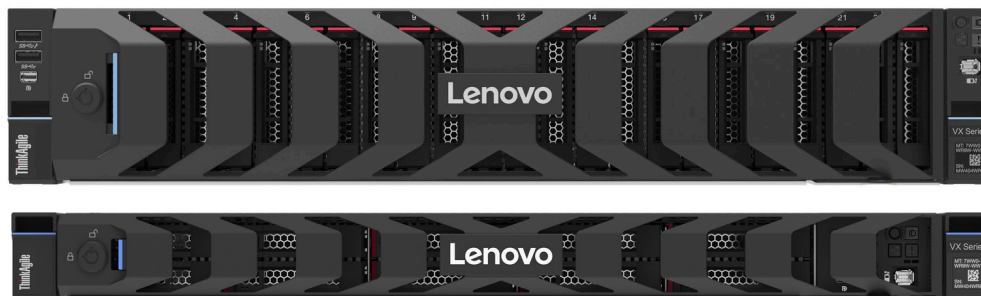


Figure 1. ThinkAgile VX650 V4 (top) and VX630 V4 (bottom) designed for VMware hyperconverged infrastructure

Business Challenge

As organizations move AI initiatives from pilot to production, the requirements change. A proof of concept can run on a single server or a public API with a small, hand-picked dataset; production inference has to serve real users, work with sensitive business data, and meet the same availability, security, and audit requirements as any other business-critical application. Many organizations also find that much of this inference could run on the infrastructure they already operate, but lack a clear way to bring AI onto it without adding complexity.

Organizations making this shift typically face four challenges:

- **Data governance and residency:** sensitive data used for retrieval-augmented generation and document processing often cannot leave the organization's own infrastructure, which rules out public inference APIs for many use cases.

- **Operational silos:** dedicated AI clusters, built on different hardware, hypervisors, and management tools than the rest of the data center, add a second set of skills, processes, and vendors for IT teams to support.
- **Right-sizing:** without a common platform for AI and traditional workloads, organizations either over-provision specialized accelerators for inference that does not need them, or under-provision and hit performance walls as usage grows.
- **Security and multi-tenancy:** different teams and business units need access to different models and datasets, with consistent role-based access control and network isolation between AI services and the applications around them.

None of this is unique to AI. IT organizations already handle governance, operations, and capacity planning for every other workload; AI just adds one more class of application to that list.

Solution

Lenovo, Intel, VMware, and Canonical combine into a stack that runs private AI inference on the same infrastructure, tools, and staff already running enterprise applications.

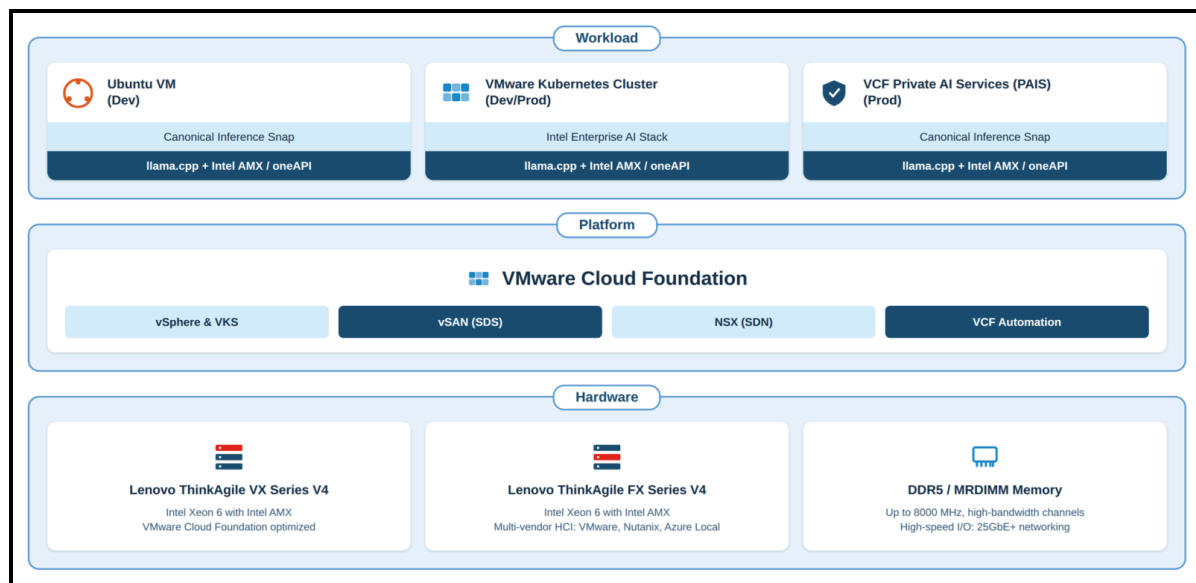


Figure 2. Private AI reference architecture on Lenovo ThinkAgile VX V4 and FX V4

Hardware: Lenovo ThinkAgile VX and FX V4

Lenovo offers two hyperconverged platforms built on Intel Xeon 6 processors for this architecture. Lenovo ThinkAgile VX V4 is purpose built and factory integrated for VMware Cloud Foundation and VMware vSphere Foundation. Lenovo ThinkAgile FX V4 shares the same Intel Xeon 6 hardware but supports VMware, Nutanix, and Microsoft Azure Local on a single appliance design, so organizations that haven't standardized on one HCI stack, or want the option to change later, can use common hardware across VMware and non-VMware environments.

Both ThinkAgile FX and ThinkAgile VX include Intel Advanced Matrix Extensions (Intel AMX) for accelerating the matrix multiplication at the core of transformer-based inference, supporting BF16 and INT8 precision without a dedicated accelerator for many enterprise inference workloads. Intel AMX is also available on 4th Gen Intel Xeon Scalable processors and newer, so organizations can begin running inference on servers already in their data center and move to ThinkAgile VX and FX V4 as part of a planned refresh.

The 2U ThinkAgile VX650 V4 and FX650 V4 and the 1U ThinkAgile VX630 V4 and FX630 V4 all support up to 86 cores per node and DDR5 memory, including Multiplexed Rank DIMMs (MRDIMMs) at up to 8000

MHz, to ease the memory bandwidth pressure that constrains large language model inference. The 2U VX650 V4 and FX650 V4 add optional GPU acceleration, managed within the same VCF environment, for organizations running larger models or higher-throughput inference workloads; the 1U VX630 V4 and FX630 V4 do not support GPUs and are sized for CPU-only inference.



Figure 3. ThinkAgile FX650 V4 (top) and FX630 V4 (bottom) designed for flexible hyperconverged infrastructure

Platform: VMware Cloud Foundation

VMware Cloud Foundation (VCF) provides the common operational layer across AI and traditional workloads: vSAN for software-defined storage, NSX for software-defined networking and micro-segmentation, and VCF Automation and Operations for lifecycle management. VCF Private AI Services extends this same operational model to AI specifically, supporting controlled model rollout, version management, Workload Domains for multi-tenancy, and role-based access control for AI resources, all governed by the same policies applied to the rest of the environment. vSphere Trust Authority and Virtual Trusted Platform Module (vTPM) help verify workload integrity before sensitive data or AI models are deployed.

Software: operating system and inference tooling

For organizations standardizing on Ubuntu, Canonical's operating system and Inference Snaps simplify model deployment: Inference Snaps detect the underlying hardware and install a matched runtime and model with a single command, and Ubuntu Pro extends security maintenance and compliance support across the environment. As of VCF 9.1, Ubuntu is a first-class citizen within VCF, with streamlined deployment and lifecycle management. For a standardized serving layer across teams, Intel AI for Enterprise Inference provides Kubernetes-based model serving with an API gateway, identity and access management, observability, and OpenAI-compatible APIs, complementing the lifecycle management in VCF Private AI Services.

Management: Lenovo XClarity

Lenovo XClarity Controller and Lenovo XClarity One give administrators a single view of firmware, health, and lifecycle management across the ThinkAgile VX and FX estate, consistent with how the rest of the Lenovo infrastructure is already managed.

Use Cases

The use cases below share a common thread: each depends on sensitive data that should stay inside the organization, and each benefits from running on the same governed VMware Cloud Foundation platform as the surrounding applications. All three can run CPU-only on Intel Xeon 6 processors with Intel AMX, with optional GPU acceleration on the 2U ThinkAgile VX650 V4 and FX650 V4 where larger models or higher throughput call for it.

- **Retrieval-augmented generation and intelligent assistants**

Organizations building internal chat assistants and knowledge search tools use retrieval-augmented generation to ground responses in proprietary documents, wikis, and databases. Running this inference on Lenovo ThinkAgile VX or FX V4 keeps that proprietary content inside the data center rather than sending it to an external API, while Intel AMX acceleration on Intel Xeon 6 processors delivers the token throughput these interactive workloads need without provisioning a separate GPU cluster.

- **Document processing and summarization**

Legal, compliance, and operations teams increasingly use generative AI to summarize contracts, reports, and case files. Because this content is often sensitive and high volume, private inference on infrastructure already covered by existing compliance controls avoids exposing documents to third-party services, while VCF Private AI Services manages model versions as summarization requirements evolve.

- **Multi-tenant AI services across business units**

As more teams across the enterprise adopt AI, they need access to different models and datasets under different policies. VCF Workload Domains and NSX micro-segmentation let organizations isolate AI services by business unit or project on shared ThinkAgile VX or FX V4 infrastructure, with role-based access control and audit capabilities governing who can deploy, update, or query each model.

Workloads and Sizing

Private AI inference on Lenovo ThinkAgile VX and FX V4 generally falls into two workload profiles, and two considerations determine how to size for them. Because the VX630 V4, VX650 V4, FX630 V4, and FX650 V4 support the same Intel Xeon 6 processor options and DDR5 and MRDIMM memory configurations, this guidance applies across the family. It describes how CPU-based inference behaves in general, not results measured on this specific hardware:

- **Interactive workloads:** chat assistants and retrieval-augmented generation queries, which typically serve a modest number of concurrent users per node at low-latency response times.
- **Departmental and API-driven workloads:** document summarization and batch processing services, which trade higher per-request latency for greater concurrency and throughput per node.
- **Memory bandwidth:** for transformer-based inference, memory bandwidth governs token generation throughput more than core count does. Intel Xeon 6 processors with MRDIMMs at up to 8000 MHz help sustain throughput as context length and concurrency grow.
- **Validation before rollout:** Lenovo recommends benchmarking the actual model and workload with tools such as vLLM bench serve and llama-bench to confirm service levels and node count before moving to production.

Business Outcome

Running private AI inference on Lenovo ThinkAgile VX and FX V4 delivers outcomes that come from how the stack is built, not from a published benchmark. In this architecture, memory bandwidth and Intel AMX acceleration, not GPU-class parallelism, govern throughput. The outcomes below describe how Intel Xeon 6-based inference and this software stack behave in general, not figures Lenovo measured on ThinkAgile VX or FX V4:

- **Predictable scaling:** CPU-based inference capacity grows as nodes are added to a cluster, so organizations can size toward a target number of concurrent users in node increments and confirm each step with benchmarking.
- **No parallel AI environment:** running AI workloads on the same VMware Cloud Foundation platform as everything else avoids the cost of a separate AI stack, with shared hardware, operational tooling,

and staff.

- **Faster path to production:** Canonical Inference Snaps remove the manual tuning work a from-scratch deployment normally takes, shortening the move from pilot to production.
- **Lower platform risk:** ThinkAgile FX V4 supports multiple HCI software platforms, so the organization can change its HCI software later without a hardware refresh.

Conclusion

Private AI works best when it extends the infrastructure, operations, and governance an organization already has, instead of replacing them. Lenovo ThinkAgile VX and FX V4, built on Intel Xeon 6 processors with Intel AMX acceleration and VMware Cloud Foundation, give IT organizations a practical way to run retrieval-augmented generation, document processing, and multi-tenant AI services alongside their existing applications, on infrastructure they already operate. From there, Lenovo can help size and benchmark the architecture for each organization's specific workload.

For More Information

To learn more about private AI inference on Lenovo ThinkAgile VX and FX V4, contact your Lenovo representative or Lenovo Business Partner, or visit the resources below.

References:

- VMware Cloud Foundation on Lenovo ThinkAgile VX and FX Reference Design:
<https://lenovopress.lenovo.com/lp1533-vmware-cloud-foundation-on-lenovo-thinkagile-vx-fx-reference-design>
- Lenovo ThinkAgile VX650 V4 Hyperconverged System Product Guide:
<https://lenovopress.lenovo.com/lp2135-lenovo-thinkagile-vx650-v4-hyperconverged-system>
- Lenovo ThinkAgile FX650 V4 Hyperconverged System Product Guide:
<https://lenovopress.lenovo.com/lp2338-lenovo-thinkagile-fx650-v4-hyperconverged-system>
- Multi-Vendor Hyperconverged Infrastructure with Lenovo ThinkAgile FX V4:
<https://lenovopress.lenovo.com/lp2520-multi-vendor-hyperconverged-infrastructure-with-lenovo-thinkagile-fx-v4>
- Lenovo ThinkAgile VX630 V4 Hyperconverged System Product Guide:
<https://lenovopress.lenovo.com/lp2134-lenovo-thinkagile-vx630-v4-hyperconverged-system>

Author

Chris Honoré is a Solutions Product Manager at Lenovo with deep expertise in datacenter products and solution offerings. He has a strong background in consulting and solution development, helping customers design and support on-premises and hybrid environments. Chris has spent the past 15 years with IBM and Lenovo, specializing in x86 server and data center solutions. Prior to that, he built two decades of experience in the telecommunications industry, serving in both technical and business leadership roles.

Related product families

Product families related to this document are the following:

- [Intel Alliance](#)
- [ThinkAgile FX Series](#)
- [ThinkAgile VX Series for VMware](#)
- [VMware Alliance](#)

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

© Copyright Lenovo 2026. All rights reserved.

This document, LP2546, was created or updated on October 2, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
<https://lenovopress.lenovo.com/LP2546>
- Send your comments in an e-mail to:
comments@lenovopress.com

This document is available online at <https://lenovopress.lenovo.com/LP2546>.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at <https://www.lenovo.com/us/en/legal/copytrade/>.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Lenovo®

ThinkAgile®

XClarity®

The following terms are trademarks of other companies:

Intel®, the Intel logo and Xeon® are trademarks of Intel Corporation or its subsidiaries.

Microsoft® and Azure® are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.