Production AI Where Your Data Lives: Lenovo ThinkAgile MX650a V4 and Azure Arc-Enabled Foundry Local

Solution Brief

Enterprises want the productivity gains of generative AI without exposing sensitive data, absorbing unpredictable cloud costs, or sacrificing latency and reliability. Lenovo ThinkAgile MX650a V4, equipped with NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, provides the validated infrastructure foundation for running enterprise AI inference on-premises and at the edge. Paired with Azure Arc-enabled Foundry Local, a localized AI inference runtime that deploys as an Azure Arc extension, the platform lets models run on customer-managed infrastructure, including environments with limited or disconnected connectivity, while retaining Azure-grade identity, governance, and lifecycle management through Microsoft Entra ID and Azure RBAC.

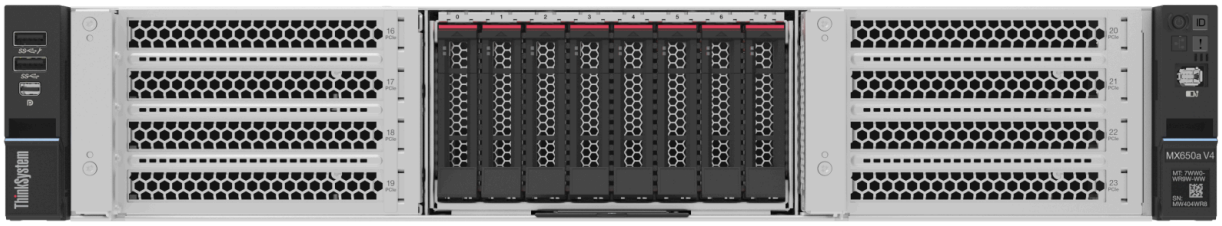

Figure 1. Lenovo ThinkAgile MX650a V4

This brief outlines the business challenge driving on-premises AI adoption, the Lenovo and Microsoft solution that addresses it, the use cases and workloads it enables, and the business outcomes joint customers can expect across regulated, edge, and enterprise environments.

Business Challenge

As organizations move generative AI from pilot to production, several constraints stand in the way of using public cloud AI endpoints alone:

- **Unpredictable economics**: Per-token cloud AI pricing creates volatile, hard-to-forecast costs for high-throughput, agent-driven, or batch workloads such as document processing and code generation.
- **Latency and reliability**: Interactive copilots, voice agents, and real-time analytics need consistent low latency that WAN round-trips to public endpoints cannot guarantee.
- **Edge and disconnected operations**: Branch offices, factories, ships, and remote sites need AI capability despite limited or intermittent connectivity.
- **Governance fragmentation**: Running AI outside the cloud often means abandoning enterprise identity and access controls in favor of a separate, harder-to-govern security model.
- **Vendor lock-in and complexity**: Building a bespoke AI operations stack, or committing to a single model provider, adds cost, integration burden, and long-term risk.

Solution

Lenovo ThinkAgile MX650a V4 and Azure Arc-enabled Foundry Local combine to deliver a validated, on-premises inference plane that behaves like Azure without leaving the customer's data center.

- **Lenovo ThinkAgile MX650a V4**: A 2U, dual-socket server powered by Intel Xeon 6 processors, engineered for Microsoft Azure Local and Azure Arc-managed hybrid cloud environments, with support for up to four double-wide GPUs, high-capacity DDR5 memory, and high-performance NVMe storage.
- **GPU foundation**: Validated with NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, each providing 96 GB of GDDR7 memory, enough headroom for 7B-13B parameter models in fp16 and 20B-class models with quantization or tensor parallelism.
- **Azure Arc-enabled Foundry Local**: An Azure-integrated, currently-in-preview platform that runs AI inference locally and deploys as an Azure Arc extension on Arc-enabled Kubernetes clusters, so the same Azure control plane manages identity, lifecycle, and policy.
- **Identity and governance**: Hybrid authentication (API keys plus Microsoft Entra ID), Azure RBAC authorization matching the Azure OpenAI operating model, and a consistent sidecar-based inference pod architecture: an nginx sidecar for TLS termination, an entra-sidecar for token validation, and an msi-adapter identity broker for disconnected nodes.
- **Flexible model deployment**: Deployment supports GPU-backed models via vLLM for maximum throughput and CPU-backed models (including the Phi family of small language models) for budget-sensitive, edge-style, or low-to-moderate traffic scenarios, using the same catalog, tooling, and application code in both cases.

The deployment path is validated end to end: cluster and GPU prerequisites, Entra ID application registration and token configuration, Azure CLI authorization, Azure RBAC role assignment, and installation of certificate management components and the Foundry Local inference operator, followed by model deployment and inference validation on the reference MX650a V4 platform.

Use Cases

Across industries, ThinkAgile MX650a V4 with Azure Arc-enabled Foundry Local powers high-value AI use cases while keeping regulated data on-premises.

Table 1. Use Cases

Industry / Use Case	How ThinkAgile MX650a V4 Delivers Value
Financial Services - Automated Contract & Document Review	Banks, insurers, and asset managers review, summarize, and flag risk clauses across thousands of contracts and filings daily, reducing review cycles from days to minutes while sensitive filings stay on-premises.
Healthcare - Clinical Documentation & Records Summarization	Synthesize patient records, surface relevant clinical history, and generate structured care-team summaries at scale, on-premises to meet data sovereignty and HIPAA requirements.
Legal - Intelligent Discovery & Case Preparation	Rapidly surface relevant documents, generate case summaries, and identify precedents across large document sets, scaling to many concurrent users under deadline pressure with data kept behind the firewall.
Manufacturing & Supply Chain - Operational Intelligence	Automate extraction of key terms from supplier contracts, technical specs, and regulatory documents; flag supply-chain risks and generate operational summaries while keeping sensitive IP on-premises.
Public Sector - Policy Research & Regulatory Compliance	Deploy agentic AI within air-gapped or controlled network environments to summarize policy documents, procurement records, and regulatory filings without exposing government data to the public cloud.
Cross-Industry - Enterprise Knowledge Management	Create intelligent interfaces over internal knowledge repositories such as product docs, HR policies, and runbooks, giving employees instant, authoritative summaries and reducing time spent searching.

Workloads

The platform's combination of on-premises data control, Azure-grade governance, and GPU density supports a broad range of enterprise AI workloads.

Table 2. Workloads

Workload	Description
Regulated-Industry Copilots	Banking, insurance, healthcare, and public-sector/defense copilots that keep policies, claims, PHI, and controlled content entirely on-premises.
Enterprise Knowledge & RAG	Retrieval-augmented search over SharePoint, Confluence, ERP/CRM data, and code repositories for legal, R&D, and engineering teams.
Developer & Engineering Productivity	Self-hosted code completion, refactor, and review assistants; bulk code generation and migration pipelines behind the firewall.
Document & Content Processing at Scale	Contract analysis, claims processing, KYC extraction, invoice understanding, and multilingual summarization as flat-cost batch jobs.
Agentic Workflows on Private Systems	Entra ID-authenticated agents that call deployed models to integrate with on-prem ERP, CRM, ITSM, and OT systems.
Edge and Branch AI	In-store assistants and loss-prevention video AI, shop-floor copilots, ward-level clinical assistants, and telco/MEC GenAI services.
Disconnected & Intermittent Environments	Ships, oil rigs, mining sites, and forward operating bases where inference and authentication continue through WAN outages.
Voice, Video & Real-Time Interaction	Contact-center copilots and live transcription where round-trip latency to a public endpoint is unacceptable.
Multi-Tenant Internal AI Platform	A shared on-prem inference service with Azure RBAC governing which business units can call which models.
Cost-Sensitive Batch Evaluation & Fine-Tuning Prep	Offline evaluation harnesses, synthetic data generation, and embedding back-fills run continuously on owned hardware.
Hybrid Bursting and Tiered Routing	Local infrastructure handles default and sensitive traffic; overflow or frontier-model requests route to Azure OpenAI under the same identity model.

Business Outcomes

Running Foundry Local on ThinkAgile MX650a V4 delivers measurable outcomes for joint Lenovo and Microsoft customers.

Table 3. Business Outcomes

Outcome	Business Impact
Data sovereignty & compliance	Models and prompts never leave customer premises; only control-plane metadata (Arc, RBAC, telemetry) reaches Azure, simplifying GDPR, HIPAA, FINMA, and ITAR requirements.
Azure-grade governance, on your hardware	Identity via Microsoft Entra ID and authorization via Azure RBAC mirror the Azure OpenAI operating model, unified across cloud, data center, and edge.
Predictable cost at scale	After hardware amortization, inference cost is flat regardless of token volume, with no per-call billing spikes for high-throughput or batch workloads.
Low, deterministic latency	Local PCIe-attached GPUs deliver single-digit to low-tens-of-ms time-to-first-token, removing the variability of public cloud endpoints.
Edge and disconnected resilience	Cached Entra signing keys and RBAC decisions keep authenticated inference running through short WAN outages at branch and edge sites.
High GPU efficiency and density	96 GB of GDDR7 memory per GPU and up to four GPUs per node enable several concurrent model deployments on one server.
Open model choice, no vendor lock-in	Any supported open-weights model (Llama, Phi, Mistral, gpt-oss, embeddings) served via vLLM or ONNX-GenAI, swappable without application changes.
Operational simplicity	Standard Kubernetes and Azure CLI tooling and a uniform sidecar pod architecture mean no bespoke AI operations stack.
Strong security posture	TLS termination, operator-managed certificates, and identity-aware ingress, with data-plane network egress tightly scoped or eliminated.
Sustainability and locality	Inference traffic stays on the local LAN, reducing WAN bandwidth, cloud egress cost, and the failure impact radius.

Conclusion

Lenovo ThinkAgile MX650a V4 and Azure Arc-enabled Foundry Local give enterprises a validated path to production AI without forcing a trade-off between data control and cloud-grade governance. Sensitive workloads stay on customer-owned infrastructure, while Microsoft Entra ID and Azure RBAC extend the same identity and access model used in Azure OpenAI, and the sidecar-based inference architecture keeps every model deployment consistent, auditable, and easy to operate.

For regulated industries, edge sites, and cost-sensitive, high-volume workloads alike, the platform delivers predictable economics, low and deterministic latency, and resilience through disconnected periods, all without giving up choice of model or runtime. Lenovo and Microsoft customers can move from pilot to production with confidence, wherever their data resides.

For More Information

To learn more, see the following resources:

- Planning and Implementation Guide
  https://lenovopress.lenovo.com/lp2471-next-gen-ai-with-thinkagile-mx650a-v4-and-azure-arc-enabled-foundry-local
- Lenovo ThinkAgile MX650a V4 Product Guide
  https://lenovopress.lenovo.com/lp2258-lenovo-thinkagile-mx650a-v4-hyperconverged-system

To learn more about running enterprise AI on Lenovo ThinkAgile MX650a V4 with Azure Arc-enabled Foundry Local, contact your Lenovo Business Partner or Lenovo representative.

Authors

Saleem Al Bouri is a Lenovo Solution Engineer based in Bucharest, Romania. He holds a Bachelor of Science in Computer Engineering and is currently pursuing a master's degree in security of Complex Information Networks. His expertise includes IT engineering, software development, containerized infrastructure, orchestration technologies, and Microsoft hybrid cloud and on-premises solutions.

Victor Talpeanu is a Microsoft Solutions Engineer with over 8 years of experience in the IT industry, specializing in Azure Local deployments and Azure Local cluster testing. He contributes to Lenovo's MX portfolio through Software Builder Extension (SBE) validation, ensuring seamless integration and performance across hybrid and edge infrastructures. Victor focuses on delivering secure, modern, and reliable solutions built on Azure Local and enterprise cloud technologies.

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:

- Artificial Intelligence
- Microsoft Alliance
- ThinkAgile MX Series for Microsoft Azure Local

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.

This document, LP2481, was created or updated on July 24, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2481
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2481.

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®

The following terms are trademarks of other companies:

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

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

NVIDIA® and NVIDIA RTX® are trademarks of NVIDIA Corporation.

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