# Nutanix Agentic AI on Lenovo ThinkAgile HX and FX V4
## Solution Brief

Agentic applications extend the request path well beyond a single prompt and response. A typical agent selects a model, retrieves enterprise data, calls one or more tools, evaluates intermediate results, and repeats those steps before returning an answer or taking an approved action. Each user request can therefore generate many model calls and infrastructure interactions.

Nutanix Agentic AI addresses this pattern by combining Nutanix Enterprise AI, Nutanix Kubernetes Platform, and Nutanix Cloud Infrastructure into a single platform to run and govern agentic applications. Because agent applications can run anywhere and use any model for inference, Nutanix Enterprise AI and Nutanix Agent Gateway support deployment across both private data centers and public clouds, with one control plane to deploy, manage, route, and fail over agents and large language model (LLM) inference across those environments.


Figure 1. ThinkAgile HX650 V4 (top) and HX630 V4 (bottom) for Nutanix hyperconverged infrastructure

For data, models, and workloads that must stay on premises, Lenovo ThinkAgile HX V4 and FX V4 systems with Intel Xeon 6 processors provide the integrated infrastructure. Intel and Lenovo testing on these platforms shows that 8-billion-parameter (8B) models can be served on CPU with up to 2x the throughput of the prior generation and sub-100 ms second-token latency, giving organizations a practical, CPU-first starting point for private inference.

This brief describes the business challenges, the solution architecture, representative use cases and industries, the test results, business outcomes, and a reference bill of materials.

## Business Challenge

Most organizations have moved past asking whether agentic AI can help the business and are now asking how to run it safely and predictably at scale. Early pilots often succeed with a single model, a small team, and a handful of users. Production is different: agents act on real enterprise data and systems, many teams depend on the same platform, and response times and costs must hold up under real demand. At the same time, sensitive data, regulated workloads, and data residency rules keep much of this work in the private data center, even when some models or applications run in the public cloud.

As a result, organizations moving agentic AI from pilot to production typically run into three challenges:

- **Control of data and model access:**

  Agents combine user instructions, retrieved documents, model outputs, and tool results. Authorization decisions must stay consistent across all of those paths, and keeping inference on premises does not remove the need for identity, endpoint policy, logging, and approval controls.

- **Operational complexity:**

  An agent platform can include inference servers, Kubernetes clusters, model registries, vector databases, object and file storage, network policy, and monitoring. Treating each as a separate project multiplies lifecycles and support boundaries across AI, platform, infrastructure, data, and security teams.

- **Concurrency and service levels:**

  A multi-step agent can issue several model requests for one business transaction. Capacity planning must account for concurrent prompts, token distributions, time to first token, time per output token, tool latency, and step count. An average token rate alone cannot predict application response time.

Left unaddressed, these challenges slow time to value, raise the risk of data exposure or unapproved actions, and lead to either overbuilt infrastructure or missed service levels. Organizations need a platform that brings governance, operations, and capacity planning together from the start.

## Solution

The Lenovo and Nutanix solution separates application concerns from shared infrastructure, so that each team owns a clear layer on a single operational foundation. The figure below shows the architecture: the Nutanix Enterprise AI control plane and Nutanix Agent Gateway span hybrid environments, while ThinkAgile HX V4 and FX V4 systems host the on-premises layers.

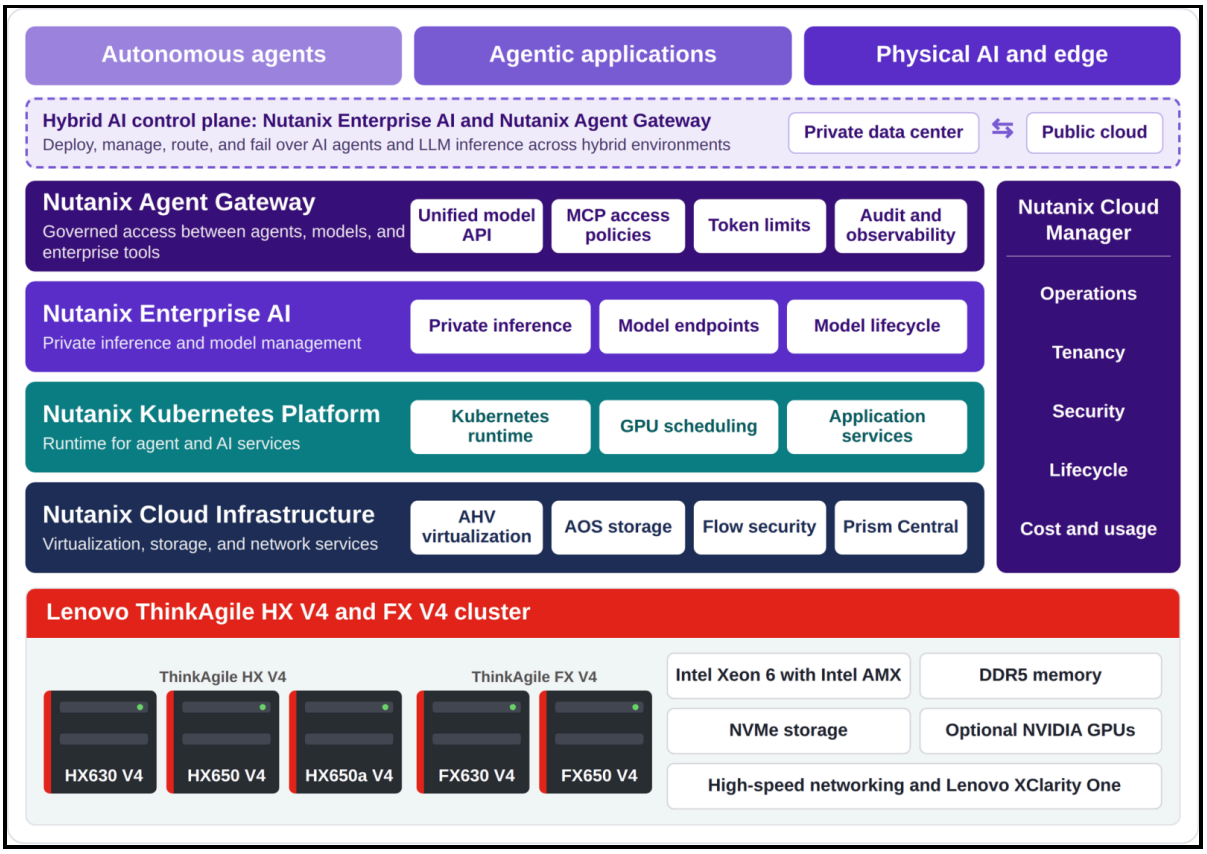


Figure 2. Nutanix Agentic AI architecture with hybrid control plane on Lenovo ThinkAgile HX V4 and FX V4

As the figure above illustrates, the solution is built as a stack of layers, each with a clear role. At the top, Agent Gateway and Nutanix Enterprise AI govern how agents reach models and tools and how inference endpoints are managed. Beneath them, Nutanix Kubernetes Platform provides the runtime for agent and AI services, while Nutanix Cloud Infrastructure and data services supply the virtualization, storage, networking, and management foundation. The private layers of this stack run in the organization's data center on Lenovo ThinkAgile HX V4 and FX V4 systems, where Intel Xeon 6 processors with built-in Intel Advanced Matrix Extensions (Intel AMX) accelerate AI inference directly on the CPU, so private models can be served without dedicated accelerators.

### Agent Gateway and inference management

Nutanix Enterprise AI serves as the centralized AI control plane for Agent Gateway and inference management. Agent Gateway standardizes how agents reach language models, Model Context Protocol (MCP) servers, and enterprise applications, with consistent authentication, observability, token-based rate limiting, usage accountability, and support for private and cloud-hosted model endpoints.

### Nutanix Enterprise AI

Nutanix Enterprise AI provides model endpoint deployment and lifecycle functions so that platform teams can offer models as a shared service. This separates model operation from each application team's code and lets endpoint, model, and capacity changes be managed centrally. The Intel testing in this brief used Nutanix Enterprise AI 2.5 with vLLM 0.13.0 as the serving stack.

### Nutanix Kubernetes Platform

Nutanix Kubernetes Platform supplies the Kubernetes environment for agent services and adjacent AI tools, with catalog categories that include notebooks, vector databases, MLOps engines, and agent frameworks. For production use, platform teams should standardize a supported subset, control images and dependencies, and separate development and production namespaces and credentials.

### Infrastructure and data services

Nutanix Cloud Infrastructure provides the foundation: AHV supplies the virtualization boundary, AOS provides distributed storage, Prism provides cluster operations, and Flow supplies networking and security controls. Nutanix Unified Storage and Nutanix Data Services for Kubernetes support shared file, object, block, and persistent Kubernetes data, depending on the application design. CPU-only and accelerator-backed model servers use the same operational foundation, although the bill of materials and performance profile differ.

Nutanix Cloud Manager (NCM), shown alongside the stack in the figure above, adds management capabilities for operations, tenancy, security, lifecycle, and cost and usage across the environment.

### On-premises deployment scope

In this design, private models, the Agent Gateway, Kubernetes services, application services, and the related data paths are all hosted in the organization's data center. The default reference path is private inference on ThinkAgile HX V4 and FX V4 clusters, so prompts, retrieved documents, and model outputs for sensitive workloads remain within the organization's control boundary.

Agent Gateway can also present governed access to public model services. Using those services changes the data flow and residency assessment, so a hosted model should be routed as a separate endpoint with explicit data classification, logging, rate limits, and cost controls. Gateway policy complements, rather than replaces, authorization inside the called application or MCP server.

### Lenovo ThinkAgile HX V4 and FX V4 infrastructure

Lenovo ThinkAgile HX V4 systems are integrated Nutanix hyperconverged infrastructure (HCI) appliances. ThinkAgile FX V4 systems are factory-integrated multi-vendor HCI platforms that can run Nutanix AOS and AHV, giving organizations the option to standardize on one hardware platform while keeping software stack choice. Both families run the same Nutanix software layers.


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

The table below summarizes the role of each model in this solution.

Table 1. ThinkAgile V4 deployment roles

| System | Form factor | Role in this solution | Important note |
|---|---|---|---|
| ThinkAgile HX630 V4 | 1U, 2-socket | Space-efficient Nutanix HCI nodes and smaller CPU inference footprints | HX product integrated with Nutanix software |
| ThinkAgile HX650 V4 | 2U, 2-socket | Higher memory and storage flexibility; the platform used in Lenovo testing (LP2362) | Built on ThinkSystem SR650 V4 |
| ThinkAgile HX650a V4 | 2U, accelerated | Configurations that require greater accelerator density | Not used in the CPU tests in this brief |

| System | Form factor | Role in this solution | Important note |
|---|---|---|---|
| ThinkAgile FX630 V4 | 1U, 2-socket | Sites that want a compact platform and the FX multi-vendor HCI model | Nutanix AOS and AHV are supported software-stack options |
| ThinkAgile FX650 V4 | 2U, 2-socket | Sites that want more storage or accelerator options and FX software-stack flexibility | Nutanix AOS and AHV are supported software-stack options |

Refer to the Lenovo product guides for the supported processor, memory, storage, networking, and accelerator combinations. All models are managed with Lenovo XClarity One and backed by Lenovo support.

**Intel Xeon 6 and CPU inference**

Intel Xeon 6 processors build AI acceleration into every core through Intel Advanced Matrix Extensions (Intel AMX), which accelerate the matrix multiplication at the heart of transformer inference. Intel AMX supports bfloat16 for training and inference and int8 for inference, and frameworks use it through Intel Extension for PyTorch and other optimized libraries, so applications benefit from AMX acceleration through standard AI frameworks.

Running inference on the CPU lets model endpoints share the same ThinkAgile nodes, virtualization layer, and operational model as the rest of the Nutanix cluster. Higher core counts, such as the 64-core Xeon 6767P, and DDR5-6400 memory increase the concurrency and token rate each node can serve, as the Workloads section shows.

CPU inference is a strong fit when:

- Models are in the 8B class or smaller, such as retrieval, summarization, classification, and tool-selection steps in an agent pipeline.
- Concurrency is moderate and latency targets are in the range of about 100 ms per output token.
- Data must stay local and operational simplicity matters more than maximum token throughput.

Larger models, long contexts, or high concurrency may require accelerator-backed configurations such as the ThinkAgile HX650a V4. Because both run on the same Nutanix foundation, teams can start with CPU inference and add GPU capacity without changing how endpoints are governed or operated.

## Use Cases

Agentic AI on this platform supports a broad range of enterprise workflows. The table below summarizes representative on-premises use cases, the agent behavior involved, and the controls each one requires.

Table 2. Representative on-premises use cases

| Use case | Agent behavior | Required controls and components |
|---|---|---|
| **Enterprise knowledge assistant** | Retrieves approved internal content and generates a cited answer | Private model endpoint, retrieval service, document permissions, response grounding checks |
| **Service operations assistant** | Summarizes incidents, proposes diagnostic steps, and opens or updates tickets after approval | Agent Gateway, MCP or API connectors, tool-specific authorization, action logging |
| **Document review workflow** | Extracts fields, compares documents to policy, and routes exceptions to a reviewer | Private inference, durable document storage, workflow state, human approval |
| **Application and platform support** | Queries telemetry, correlates alerts, and proposes runbook steps | Nutanix Kubernetes Platform services, observability connectors, read-only defaults, change approval |

| Use case | Agent behavior | Required controls and components |
|---|---|---|
| **Operational planning** | Combines structured data and internal documents to draft plans or recommendations | Governed data access, model endpoint, audit trail, versioned inputs |

Start with use cases where actions can remain read-only or require human approval. Grant write access only after the team has measured tool selection accuracy, failure handling, audit completeness, and rollback procedures.

The following examples show how three of these use cases apply in practice:

- **Enterprise knowledge assistant**

  Employees need answers grounded in internal policies, procedures, and technical documentation that cannot leave the data center. The agent retrieves approved content and generates a cited answer using a private model endpoint on Nutanix Enterprise AI, with document permissions and grounding checks enforced along the retrieval path. Financial services, healthcare, and professional services firms gain assistant capabilities without sending confidential content to a hosted model.

- **Service and platform operations assistant**

  IT operations teams face alert volume that outpaces staff. The agent summarizes incidents, queries telemetry, correlates alerts, and proposes diagnostic or runbook steps, then opens or updates tickets after human approval. Agent Gateway governs access to MCP and API connectors, services run on Nutanix Kubernetes Platform, and read-only defaults with action logging keep the blast radius small while the team measures accuracy.

- **Document review workflow**

  Manufacturing, public sector, and legal teams process large volumes of contracts, quality records, and case files. The agent extracts fields, compares documents to policy, and routes exceptions to a reviewer. Private inference, durable document storage on Nutanix data services, and workflow state on the same cluster keep sensitive records local and traceable.

## Market Verticals

Organizations in regulated and data-sensitive industries are among the strongest candidates for on-premises agentic AI, because they need AI capabilities while keeping sensitive data, models, and audit records under their own control. The following table shows the primary design concern and representative applications for each industry.

Table 3. Examples by industry

| Industry | Primary design concern | Representative applications |
|---|---|---|
| **Financial services** | Customer and transaction data, model auditability, controlled access | Policy and procedure assistants, analyst research, case preparation |
| **Healthcare and life sciences** | Protected information, traceability, data residency | Clinical document preparation, research retrieval, operations support |
| **Manufacturing** | Plant continuity, technical documentation, local data paths | Maintenance guidance, quality review, supply planning |
| **Government and public sector** | Sovereignty, disconnected operation, records control | Case summarization, policy retrieval, citizen-service support |
| **Professional services** | Client confidentiality, repeatable document workflows | Contract review support, research synthesis, project knowledge assistants |

Industry labels do not determine compliance. Each deployment must map its actual data, users, model behavior, tool permissions, record retention rules, and control evidence to the organization's legal and

security requirements.

## Workloads

Two separate test programs characterize CPU inference for 8-billion-parameter Llama models on Intel Xeon 6, a common model size for retrieval, summarization, and tool-selection steps in agentic applications. The two result sets use different models, runtimes, software versions, and latency definitions, so they are presented separately and should not be combined.

**Intel testing: Nutanix Enterprise AI 2.5 on Xeon 6**

Intel tested Nutanix Enterprise AI 2.5 with vLLM 0.13.0 serving Llama 3.1 8B Instruct at an approximately 100 ms time per output token (TPOT) service level. The configurations were:

- **Xeon 6 system:** Lenovo ThinkSystem SR650 V4 with 2 x Intel Xeon 6745P (32 cores, 300 W) and 512 GB DDR5-6400. The SR650 V4 is the base server for ThinkAgile HX650 V4, which supports the Xeon 6745P.
- **Comparison system:** Lenovo ThinkSystem SR650 V3 with 2 x Intel Xeon Platinum 8562Y+ (32 cores, 300 W) and 512 GB DDR5-5600.

The following chart shows the number of concurrent user prompts each system supported while meeting the TPOT service level, across six combinations of input and output length.

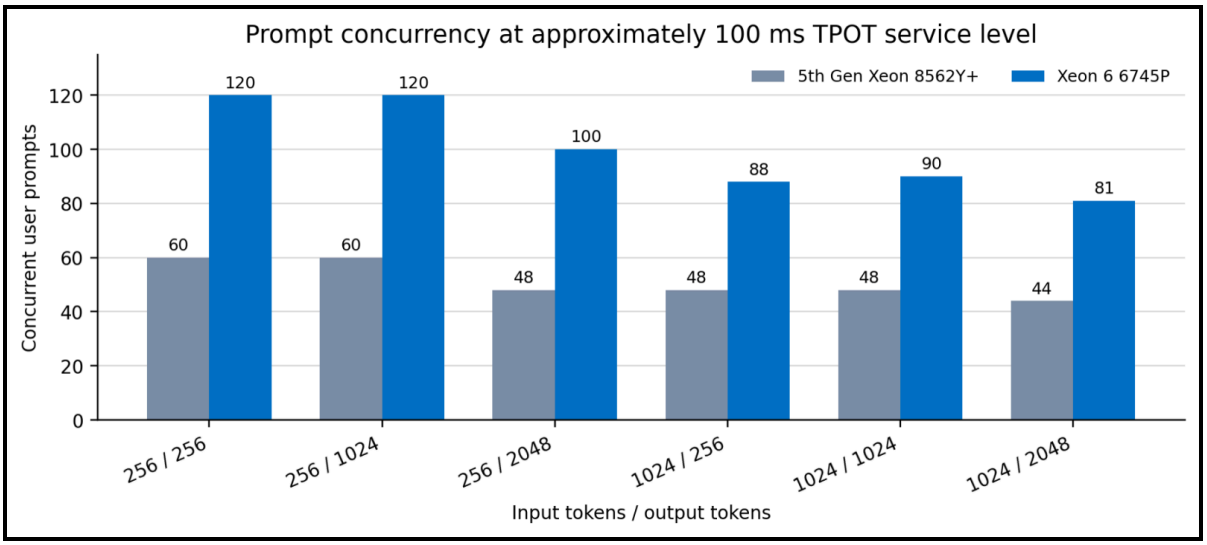


Figure 4. Intel testing: concurrent prompts meeting the approximately 100 ms TPOT service level (higher is better)

The Xeon 6 system supported 120 concurrent prompts versus 60 with 256 input tokens and 256 or 1024 output tokens, and 81 versus 44 with 1024 input and 2048 output tokens, while meeting the TPOT target. Memory speed, kernel, and firmware differed between systems, so this is a system comparison rather than a processor-only measurement.

**Lenovo testing: ThinkAgile HX650 V4 cluster**

Lenovo ran Llama 3 8B in bfloat16 with Intel Extension for PyTorch 2.3 and DeepSpeed 0.18.2 in a single Ubuntu 24.04 virtual machine (240 vCPUs, 384 GB RAM) on a four-node ThinkAgile HX650 V4 cluster running Nutanix 7.3 and AHV 10.3. Each node used 2 x Intel Xeon 6767P (64 cores, 2.4 GHz) with Intel AMX and 1 TB DDR5-6400. Tests used 32 to 2048 input tokens, 256 output tokens, and batch sizes 1 to 16.

The following chart second-token average latency for each input length as batch size increases, with the 100 ms target marked for reference.

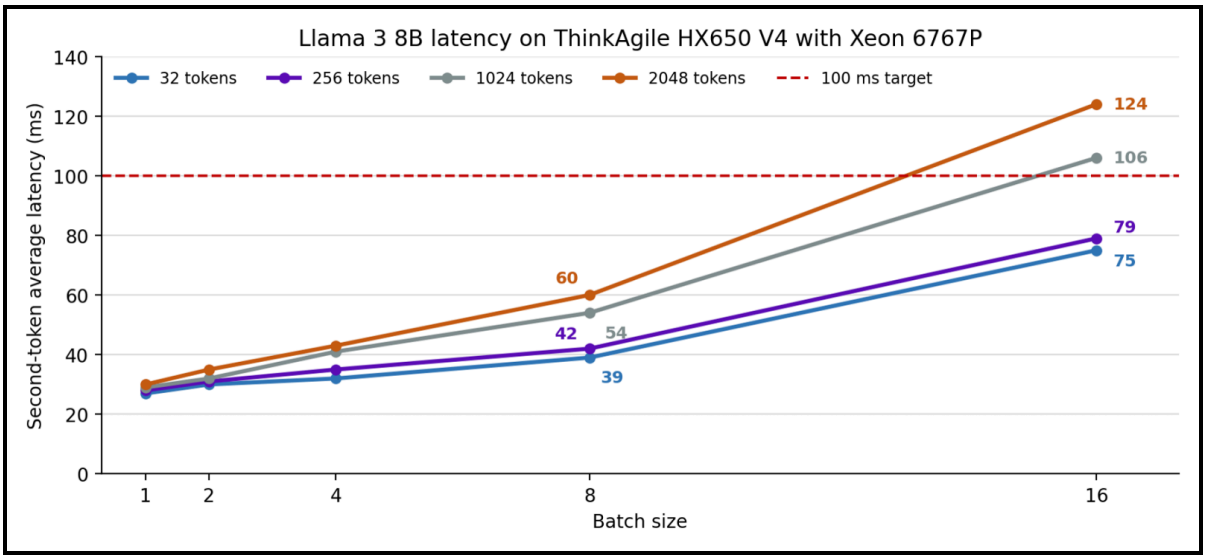


Figure 5. Lenovo testing: Llama 3 8B second-token average latency on ThinkAgile HX650 V4 with Xeon 6767P (lower is better)

Second-token average latency stayed below 100 ms through batch size 8 for every input length, making batch size 8 a practical steady-state point for prompts of about 2,000 tokens. At batch size 16, 1024-token and 2048-token inputs rose to 106 ms and 124 ms.

Notes about the tests:

- **Test dates and performance**:

  Intel results are based on testing by Intel as of April 2, 2026. Lenovo results are based on Lenovo testing on ThinkAgile HX650 V4. Performance varies by use, configuration, and other factors. Learn more on the Intel Performance Index site. Your costs and results may vary.

- **What the tests do not cover**:

  Both tests measure model serving, not a complete agent request. Agent Gateway policy, retrieval, MCP tool calls, and multi-step orchestration must be added in a proof of concept, with the cluster sized to carry the required load with one node unavailable.

## Business Outcome

The solution delivers value on three fronts: measurable inference performance on Intel Xeon 6, simpler and more controlled operations, and a lower-risk path from pilot to production:

- Performance and efficiency
- Operational and business value
- A lower-risk path to production

### Performance and efficiency

The published results show that Xeon 6-based systems give a credible CPU-first starting point for private inference:

- **Up to 2x output throughput:** in Intel testing of Nutanix Enterprise AI 2.5 on ThinkSystem SR650 V4 versus SR650 V3, with up to 120 concurrent prompts meeting an approximately 100 ms TPOT service level.
- **Up to approximately 2.3x prompts per watt:** normalized to the 5th Gen Xeon system in the same Intel test.

- **Sub-100 ms latency through batch size 8:** in Lenovo testing on ThinkAgile HX650 V4, from 39 ms to 60 ms across 32-token to 2048-token inputs.

Serving 8B-class models on CPU also lets organizations start agentic AI projects on the same ThinkAgile nodes they use for other Nutanix workloads, and reserve GPU investment for the models and concurrency levels that measurably need it.

**Operational and business value**

Beyond raw performance, the architecture addresses the control, complexity, and capacity challenges described earlier. The following table summarizes how.

Table 4. Business outcomes and how the solution contributes

| Outcome | How the solution contributes |
|---|---|
| Data location and access control | Private endpoints and data services remain in the data center, with Agent Gateway and application policies controlling access |
| Shared platform operations | AI endpoints and Kubernetes services use the same Nutanix infrastructure and management foundation, reducing the number of separate lifecycles and support boundaries |
| Capacity based on service levels | CPU and accelerator choices can be matched to model size, prompt mix, latency, and concurrency rather than sized for peak GPU demand up front |
| Controlled tool use | Agent Gateway and application controls provide policy points for model and tool access, with usage accountability and audit data |
| Incremental deployment | Teams can begin with CPU-served 8B-class models and add nodes or accelerators when measured demand requires it |

**A lower-risk path to production**

Because model serving is only one part of an agent request, a structured validation sequence before production rollout is recommended:

- **Define service levels:** end-to-end latency, model latency, completion rate, concurrency, and failure behavior.
- **Replay real traffic:** a representative mix of prompt lengths, output lengths, retrieval sizes, models, and agent step counts.
- **Test the full path:** include Agent Gateway policy, identity, retrieval, MCP or API calls, and the target enterprise systems.
- **Test resilience:** run a node failure and a maintenance event while retaining the capacity reserve the availability policy requires.
- **Measure cost and power:** measure power at one stated boundary and period before calculating prompts per watt or cost per successful business task.

Following this sequence turns the published benchmarks into a sizing decision grounded in the organization's own workload, which reduces the risk of over-provisioning or missing service levels after launch.

## Bill of Materials

The table below lists the core items from the all-flash ThinkAgile HX650 V4 bill of materials used in Lenovo testing. It is a reference for the tested hardware, not a final production quote. The list includes Nutanix Cloud Platform but not Nutanix Enterprise AI or Nutanix Kubernetes Platform entitlements. Add those software subscriptions, support, network optics and switches, rack power, and any backup or disaster recovery components the final design requires.

Table 5. Core bill of materials for the tested ThinkAgile HX650 V4 all-flash configuration

| Feature | Description | Qty |
|---|---|---|
| 7DG4CTO1WW | Lenovo ThinkAgile HX650 V4 Hyperconverged System | 1 |
| C6TQ | ThinkAgile HX650 V4 Base | 1 |
| B15S | Nutanix Software Stack on Nutanix AHV | 1 |
| BVKV | Nutanix Cloud Platform Pro license with Mission Critical Support | 1 |
| C5QY | Intel Xeon 6767P 64C 350W 2.4 GHz processor | 2 |
| C0TQ | ThinkSystem 64 GB TruDDR5 6400 MHz RDIMM | 16 |
| C26V | ThinkSystem M.2 RAID B545i-2i SATA or NVMe adapter | 1 |
| C46P | ThinkSystem 2U V4 8 x 2.5-inch NVMe backplane | 2 |
| B0SW | Nutanix flash node configuration | 1 |
| C2BR | ThinkSystem 2.5-inch U.3 7500 PRO 1.92 TB NVMe PCIe 4.0 SSD | 6 |
| BKSR | ThinkSystem M.2 7450 PRO 960 GB NVMe PCIe 4.0 SSD | 2 |
| BN2T | ThinkSystem Broadcom 57414 10 or 25GbE SFP28 2-port OCP adapter | 1 |
| C0U3 | ThinkSystem 2000 W 230 V Titanium hot-swap power supply | 2 |
| 6400 | 2.8 m C13 to C14 jumper cord | 2 |
| C2DJ | ThinkSystem Advanced Toolless Slide Rail Kit V4 | 1 |
| C3RG | ThinkSystem SR650 V4 left rack latch with USB and MiniDP | 1 |
| C3RD | ThinkSystem 2U 6056 20K performance fan module | 6 |
| SCJC | XClarity One managed device with one year software support | 1 |

Confirm exact Nutanix software versions, entitlements, firmware, and hardware compatibility with your Lenovo representative before ordering, and size the cluster for the required workload with one node unavailable.

## Conclusion

Nutanix Enterprise AI governs model access and inference endpoints, Nutanix Kubernetes Platform hosts agent services, and Nutanix Cloud Infrastructure supplies the virtualized foundation. A single control plane governs agents and inference across private data centers and public clouds, while sensitive data and models stay on premises. Lenovo ThinkAgile HX V4 and FX V4 systems with Intel Xeon 6 processors deliver that on-premises foundation with integrated hardware, management, and support. Organizations can begin with read-only, human-approved agents on CPU inference, validate end to end with their own models, data, tools, and concurrency, and scale capacity as the workload proves itself.

## For More Information

To learn more about Nutanix Agentic AI on Lenovo ThinkAgile HX V4 and FX V4 systems, contact your Lenovo representative or Lenovo Business Partner, or visit the resources below.

References:

- Intel Xeon 6 Processors and Nutanix Enterprise AI Deliver 2x Higher Throughput on LLM Inferencing, Intel Community blog: https://community.intel.com/t5/Blogs/Tech-Innovation/Artificial-Intelligence-AI/Intel-Xeon-6-Processors-and-Nutanix-Enterprise-AI-Deliver-2x/post/1747665
- Deploy and Scale Generative AI in Enterprises with Intel AMX and Lenovo ThinkAgile HX V4 and FX V4 (LP2362): https://lenovopress.lenovo.com/lp2362
- Intel Performance Index: https://edc.intel.com/content/www/us/en/products/performance/benchmarks/overview/
- Multi-Vendor Hyperconverged Infrastructure with Lenovo ThinkAgile FX V4 (LP2520): https://lenovopress.com/lp2520
- Lenovo ThinkAgile HX650 V4 Product Guide: https://lenovopress.lenovo.com/lp2133
- Lenovo ThinkAgile HX630 V4 Product Guide: https://lenovopress.lenovo.com/lp2132
- Lenovo ThinkAgile FX650 V4 Product Guide: https://lenovopress.lenovo.com/lp2338
- Lenovo ThinkAgile FX630 V4 Product Guide: https://lenovopress.lenovo.com/lp2337
- Nutanix Agentic AI solution: https://www.nutanix.com/solutions/ai
- Lenovo ThinkAgile HX and FX reference architecture: https://lenovopress.lenovo.com/lp0665

## Authors

**Chandrakandh Mouleeswaran** is a Solution Architect with 18+ years of experience in software development, performance testing and engineering, having worked on designing and architecting many scalable enterprise applications. He has spent a decade in technical enablement and partner solution development for VMware, Nutanix, Oracle and other ISVs across industries and technologies. He specializes in architecting infrastructure solutions for virtualization, VDI, database, cloud, data science, AI/ML solutions and various enterprise workloads.

**Cristian Ghetau** is an Advisory Engineer for Lenovo in Romania and has experience in Cloud Infrastructure technologies. He has had more than 13 years of experience working with virtual environments from VMware, Microsoft, Oracle, and Linux.

**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:

- Nutanix Alliance
- ThinkAgile FX Series
- ThinkAgile HX Series for Nutanix

## 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, LP2533, was created or updated on September 25, 2026.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2533
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2533.

## 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®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

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

DeepSpeed® is a trademark 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.