Running Edge AI Workloads with Lenovo ThinkAgile HX360 V2 Edge Servers

Last update: **25 June 2025**

Version 1.0

Lenovo ThinkAgile Hyperconverged infrastructure solution for edge deployment with Nutanix

Validated design for Lenovo ThinkAgile HX360 V2 Servers for Edge use cases

Includes AI Inference benchmark results and Bill of Material

Chandrakandh Mouleeswaran

LENOVO PRESS

Running Edge AI Workloads with Lenovo ThinkAgile HX360 V2 Edge Servers

Edge AI Infrastructure

Organizations are increasingly moving their AI workloads to edge where data is generated and to achieve real time inference for making instantaneous decisions and to avoid challenges with latency, bandwidth and privacy requirements in cloud-based architectures. High-definition video streams, sensor data and other data intensive resources consume more bandwidth and processing time and processing this data locally at Edge reduces complexities in data movement and privacy and security requirements.

Lenovo ThinkAgile HX360 V2 Edge is a hyperconverged infrastructure (HCI) solution powered by purpose-built, compact Lenovo SE360 V2 server and Nutanix software stack provides simplified management, ensuring optimal performance and availability for AI workloads at the edge. These small form-factor nodes accelerate AI rollouts at the edge or at ROBO locations to rapidly drive business and reduces the complexity of adopting AI at the edge by providing prescriptive steps for deploying the infrastructure. It supports uninterrupted downtime to achieve operational efficiency and enables more processing at the Edge than in datacenter or cloud and thousands of edge nodes can be managed and monitored centrally. With support for capabilities to process richer data and diverse application scenarios, HX360 V2 enables AI adoption at the Edge across industries.

Lenovo ThinkAgile HX360 V2 Edge

Lenovo ThinkAgile HX360 V2 Edge is a 2U high and half width making it significantly smaller than a traditional server, ideal for deployment in tight spaces and fully validated and integrated hardware and firmware that is certified and preloaded with Nutanix software.

Lenovo ThinkAgile HX360 V2 Edge server

The server supports a single processor from the Intel Xeon D-2700 Processors family, processors up to 16 cores, core speeds of up to 2.1 GHz, and TDP ratings of up to 100W. The maximum memory supported is 256 GB using 4 TruDDR4 64 GB DIMMs and 10GbE/25GbE SFP28 I/O board to support low and high-speed networking to back-end servers. The Intel Xeon D-2700 processors support Intel AVX-512, VNNI and OpenVINO features to accelerate AI workloads, and the server supports one NVIDIA A2 or L4 GPU to enable to enable AI inferencing and data visualization scenarios.

The ThinkAgile HX360 V2 Edge is available in a pre-validated bundle with Nutanix Cloud Platform and leading open-source AI frameworks to run AI inferencing workloads. Featuring an edge and ROBO friendly form factor, the ThinkAgile HX360 V2 Edge leverages Nutanix Validated Design to accelerate AI Inferencing deployments at the edge and it can be securely claimed and activated remotely through the ThinkShield Key Vault Portal, available on the web and mobile app. It combines hardware, software, and professional services to provide an end-to-end solution from edge to cloud.

Nutanix Cloud Platform is a comprehensive hybrid cloud infrastructure solution comprises of Nutanix HCI technology with additional cloud services, automation, and integrations, enabling organizations to build, manage, and optimize their IT infrastructure across on-premises, edge, and public cloud environments.

Nutanix Kubernetes Platform simplifies AI application deployment at the edge and containers can be managed and scaled across edge and on-premises.

Nutanix Flow Network Security providing a software-defined firewall and microsegmentation capabilities for securing applications and data at the edge.

Nutanix Validated Design

Nutanix validated design uses pod architecture with management cluster running on core datacenter (on-premises) with ThinkAgile HX V3 servers and the edge clusters for each edge location with one, two or three node ThinkAgile HX360 V2 Edge servers. Each pod can manage maximum of 2000 edge locations distributed across 5 building blocks each of which supporting maximum 400 edge nodes.

Three node cluster provides maximum redundancy and processing capacity to run AI inference workloads and AI applications.

Nutanix Cloud Platform supports Zero-Touch Deployment of edge clusters though management clusters and the service enable automated deployment workflow with GitHub Runner.

Table 1 shows the recommended virtual machine sizing for deploying AI inference workloads on HX360 V2 Edge nodes. Nutanix CVM shares compute with other workloads running at the edge and it is recommended to size and consider enough headroom for peak scenarios.

Table 1. Edge AI Workload virtual machine sizing for ThinkAgile HX360 V2 Edge server

Processor	**Intel Xeon D-2733NT 8C 2.1 GHz**	**Intel Xeon D-2752TER 12C 1.8 GHz**	**Intel Xeon D-2775TE 16C 2.0 GHz**
Virtual CPU	8	8	8
Virtual memory	16 GB	16 GB	16 GB
Virtual storage	100 GB	100 GB	100 GB
Virtual NIC	1	1	1

GPU passthrough	1	1	1
Maximum VM instances (1 Node Cluster)	1	1	2
CVM Configuration	8 VCPU + 48GB Memory	8 VCPU + 48GB Memory	8 VCPU + 48GB Memory
Maximum VM instances (2 Node Cluster)	2	2	4
Maximum VM instances (3 Node Cluster)	3	3	6

Performance Testing Details and Results

Geekbench AI

GeekbenchAI is a benchmark to test machine learning inference performance for many real world AI applications. The benchmark contains ten workloads to support Computer Vison and Natural Language Processing tasks with large datasets and the tests are performed with single precision, half precision, and quantized to measure performance. The benchmark stresses the edge device compute to achieve the maximum throughput. Inferences Per Second (IPS) metric quantifies how many AI inferences a device can perform per second and a higher IPS indicates faster processing of AI tasks. Geekbench AI workload scores are calibrated using results from a baseline system (a Lenovo ThinkStation P340 with a Core i7-10700 processor)

Table 2. Geekbench AI performance testing hardware configuration

Component	**Details**
Server	1 x Lenovo ThinkAgile HX665 V3 Appliance
CPU	Intel(R) Xeon(R) D-2733NT CPU @ 2.10GHz, 8C
RAM	256 GB (4 x Lenovo ThinkSystem 64GB TruDDR4 3200MHz)
Network adapter	ThinkSystem Mellanox ConnectX-6 Dx 100GbE QSFP56 2-port PCIe Ethernet Adapter)
Disks	4 x ThinkSystem M.2 7450 PRO 1.92TB Read Intensive NVMe PCIe 4.0 x4 NHS SSD 1 x ThinkSystem M.2 7450 PRO 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD (Boot drive)
Hypervisor	AHV v.10.0.1
Nutanix AOS	7.0.1
Guest VM	Microsoft Windows Server 2022 Standard (64-bit)

VM Configuration	1 virtual machine, 8 VCPU, 64GB memory, 200 GB disk
AI Framework	ONNX, Clang 16 compiler
Geekbench AI	1.3

Table 3 shows inferences per second for different AI workloads used in the benchmark. These models are used across manufacturing, hospitality, retail and other industries which require computer vision and natural language processing capabilities for AI use cases. The workloads achieved different IPS and accuracy also a crucial factor to consider for each workload. Refer published results ThinkAgile HX360 V2 Geekbench AI Results

The virtual machine is utilized at 90% CPU during the test for many of the use cases.

Table 3. Geekbench AI inference testing results on HX360 V2 Edge

Workload	**Model**	**FP32**	**FP16**	**INT8**
Image Classification	MobileNetV1	211.6 IPS	50.5 IPS	740.4 IPS
Image Segmentation	DeepLabV3+, MobileNetV2	12.6 IPS	8.28 IPS	21.8 IPS
Pose Estimation	OpenPoseV2, VGG19	4.60 IPS	3.40 IPS	14.0 IPS
Object Detection	SSD, MobileNetV1	141.9 IPS	31.5 IPS	231.1 IPS
Face Detection	RetinaFace, MobileNetV2	49.3 IPS	11.1 IPS	59.7 IPS
Depth Estimation	ConvNets, EfficientNet-lite3	39.8 IPS	10.1 IPS	64.5 IPS
Style Transfer	Image Transform Net	12.8 IPS	7.10 IPS	17.5 IPS
Image Super-Resolution	Residual Feature Distillation Network (RFDN)	66.4 IPS	42.1 IPS	79.5 IPS
Text Classification	BERT-Tiny	1250 IPS	553.4 IPS	376.3 IPS
Machine Translation	Transformer	24.1 IPS	6.27 IPS	28.0 IPS
Image Classification	MobileNetV1	211.6 IPS	50.5 IPS	740.4 IPS
Image Segmentation	DeepLabV3+, MobileNetV2	12.6 IPS	8.28 IPS	21.8 IPS

Bill of Materials: ThinkAgile HX360 V2 Edge

Part number	**Product Description**	**Qty**
7DJ7CTO2WW	Chassis : ThinkAgile HX360 V2 Edge 2U2N Enclosure	1
C3ZP	ThinkAgile HX360 V2 Edge 2U2N Enclosure	1
5WS7C03039	3Yr Premier 24x7 4Hr Resp ThinkAgile HX360 V2 Enclosure	1
7DJDCTO1WW	Node : ThinkAgile HX360 V2 Edge	1
C3V4	ThinkAgile HX360 V2 Edge Chassis	1

BS57	ThinkEdge SE360 V2 4x 10/25Gb, 2x 2.5Gb(TSN) I/O Module	1
B0W3	XClarity Pro	1
B15S	Nutanix SW Stack on Nutanix AHV	1
BVKV	Nutanix Cloud Platform (NCP) Pro Software License with Mission Critical Support	1
BS41	ThinkEdge SE350 V2/SE360 V2 Planar with Intel Xeon D-2733NT 8C 80W 2.1 GHz	1
B966	ThinkSystem 64GB TruDDR4 3200 MHz (2Rx4 1.2V) RDIMM	2
5977	Select Storage devices - no configured RAID required	1
BC4V	Non RAID NVMe	1
B0SW	Nutanix Flash Node Config	1
BQUK	ThinkSystem M.2 7450 PRO 1.92TB Read Intensive NVMe PCIe 4.0 x4 NHS SSD (with Heatsink)	4
BS5M	ThinkEdge SE360 V2 M.2 Cabled Adapter Module	1
BS46	ThinkSystem M.2 7450 PRO 480GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD (with Heatsink)	1
BS5F	ThinkEdge SE360 V2 Riser Assembly (M.2 Riser + PCIe Riser)	1
BUGM	Internal AC Power Supply (230V/115V)	1
BUGU	ThinkEdge SE360 V2 AC Power Input Board	1
BW8U	ThinkEdge SE360 V2 500W 230V/115V Non-Hot Swap Power Supply	1
BPHZ	0.5m, 10A/100-250V, C13 to C14 Jumper Cord	1
BS5W	ThinkEdge SE360 V2 Fan Assembly (Front to Rear)	1
BS4E	ThinkEdge 130mm USB-C to VGA Display Cable	1
BS6S	ThinkEdge SE360 V2 MB to M.2 Cabled Adapter Power Cable1	1
BS6N	ThinkEdge SE360 V2 MB to M.2 Gen3 Cabled Adapter Signal Cable	1
BS6L	ThinkEdge SE360 V2 PCIe Riser Card Signal Cable	1
BS6U	ThinkEdge SE360 V2 PCIe Riser Card Power Cable	1
BTPS	ThinkEdge SE360 V2 Chassis Intrusion Cable	2
BS65	ThinkEdge SE360 V2 IO Cover Assembly for 10/25GbE I/O Module	1
BS64	ThinkEdge SE360 V2 Rear Operational Panel Module	1
BS6H	ThinkEdge SE360 V2 MB IO Board Power Cable	1
BS6G	ThinkEdge SE360 V2 IO Board to MB Cable	1
BS6E	ThinkEdge SE360 V2 Operational Panel to Rear Operational Cable	1
7S0PCTO3WW	Nutanix P&P Software for ThinkAgile HX	1
SAPU	Nutanix Cloud Platform Pro, Mission Critical Support Per Core, 3Yr	8
5WS7C03051	3Yr Premier 24x7 4Hr Resp ThinkAgile HX360 V2 Edge	1
5MS7B00045	ThinkAgile HX Remote Deployment (up to 3 node cluster)	1

References:

Lenovo ThinkAgile HX360 V2 Edge
https://lenovopress.lenovo.com/lp1963-thinkagile-hx360-v2-edge

Lenovo Validated Design for Retail: Edge Deployment and Management
https://lenovopress.lenovo.com/lp2037.pdf

Geekbench AI Workloads
https://www.geekbench.com/doc/geekbench-ai-workloads.pdf

Nutanix for Enterprise Edge Computing
https://portal.nutanix.com/page/documents/solutions/details?targetId=RA-2147-Nutanix-for-Enterprise-Edge:RA-2147-Nutanix-for-Enterprise-Edge

Trademarks and special notices

References in this document to Lenovo products or services do not imply that Lenovo intends to make them available in every country.

Lenovo, the Lenovo logo, ThinkSystem, ThinkAgile, ThinkCentre, ThinkVision, ThinkVantage, ThinkPlus and Rescue and Recovery are trademarks of Lenovo.

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

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

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used Lenovo products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-Lenovo products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by Lenovo. Sources for non-Lenovo list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. Lenovo has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-Lenovo products. Questions on the capability of non-Lenovo products should be addressed to the supplier of those products.

All statements regarding Lenovo future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only. Contact your local Lenovo office or Lenovo authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in Lenovo product announcements. The information is presented here to communicate Lenovo’s current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard Lenovo benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-Lenovo websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this Lenovo product and use of those websites is at your own risk.