VMware Private AI with Intel on Lenovo ThinkAgile VX V3 and ThinkSystem V3

Solution Brief

Lenovo V3 Systems with Intel 4th and 5th Gen Scalable Processors

Lenovo ThinkSystem SR650 V3 2U and SR630 V3 1U servers and ThinkAgile VX650 V3 2U and VX630 V3 2U hyperconverged solutions with VMware vSAN powered by 4th and 5th gen Intel Xeon Scalable processors are optimized for AI workloads and **Accelerated by Intel** offerings. Lenovo V3 systems support up to 64 cores per socket with 5th Gen Intel Xeon processors and up to 60 cores per socket with 4 th Gen processors.

ThinkAgile VX V3 systems are factory-integrated, pre-configured ready-to-go integrated systems built on proven and reliable Lenovo ThinkSystem servers that provide compute power for a variety of workloads and applications and are powered by industry-leading hyperconverged infrastructure software from VMware. It provides quick and convenient path to implement a hyperconverged solution powered by VMware Cloud Foundation (VCF) or VMware vSphere Foundation (VVF) software stacks with "one-stop shop" and a single point of contact provided by Lenovo for purchasing, deploying, and supporting the solution.

Intel Optimized AI Libraries & Frameworks

Intel provides a comprehensive portfolio of AI development software including data preparation, model development, training, inference, deployment, and scaling. Using optimized AI software and developer tools can significantly improve AI workload performance, and developer productivity, and reduce compute resource usage costs. Intel® oneAPI libraries enable the AI ecosystem with optimized software, libraries, and frameworks. Software optimizations include leveraging accelerators, parallelizing operations, and maximizing core usage.

Intel® Advanced Matrix Extensions (Intel® AMX)

Intel® AMX is a new set of instructions designed to work on matrices and it enables AI fine-tuning and inference workloads to run on the CPU. Its architecture supports bfloat16 (training/inference) and int8 (inference) data types and Intel provides tools and guides to implement and deploy Intel AMX. The Intel AMX architecture is designed with two components,

1. **Tiles**: These consist of eight two-dimensional registers, each 1 kilobyte in size, that store large chunks of data.
2. **Tile Matrix Multiplication (TMUL)**: TMUL is an accelerator engine attached to the tiles that performs matrix-multiply computations for AI.

Refer more information about Intel AMX here.

With integrated Intel AMX on 4th and 5th gen Intel Xeon Scalable processors, many AI inferencing and fine-tuning workloads, including many Generative AI use cases, can run optimally.

Intel AI software and optimization libraries provide scalable performance using Intel CPUs and GPUs. Many of the libraries and framework extensions are designed to leverage the CPU to provide optimal performance for machine learning and inference workloads. Developers looking to leverage these tools can download the AI Tools from AI Tools Selector.

Table 1: Intel AI optimization software and development tools

Software/Solution	Details
Intel® oneAPI Library	· Intel® oneAPI Deep Neural Network Library (oneDNN) · Intel® oneAPI Data Analytics Library (oneDAL) · Intel® oneAPI Math Kernel Library (oneMKL) · Intel® oneAPI Collective Communications Library (oneCCL)
MLOPs	Cnvrg.io is a platform to build and deploy AI models at scale
AI Experimentation	SigOpt is a guided platform to design experiments, explore parameter space, and optimize hyperparameters and metrics
Intel® Extension for PyTorch	Intel Extension for PyTorch extends PyTorch with the latest performance optimizations for Intel hardware, also taking advantage of Intel AMX
Intel Distribution for Python	· Optimized core python libraries (scikit-learn, Pandas, XGBoost) · Data Parallel Extensions for Python. · Extensions for TensorFlow, PyTorch, PaddlePaddle, DGL, Apache Spark, and for machine learning · NumPy, SciPy, Numba, and numba-dpex.
Intel® Neural Compressor	This open-source library provides a framework-independent API to perform model compression techniques such as quantization, pruning, and knowledge distillation, to reduce model size and speed up inference.

VMware Cloud Foundation

VMware Cloud Foundation (VCF) is a multi-cloud platform supporting virtual machines and containerization of workloads on common virtualized infrastructure built on top of vSphere, vSAN, and NSX. The suite includes VMware Aria Suite for private/hybrid cloud management and VMware Tanzu for Kubernetes workloads. Refer more details in VMware Cloud Foundation reference design here.

VMware Private AI with Intel

VMware Private AI with Intel solution enables enterprises to develop and deploy classical machine learning models and generative AI applications on the infrastructure powered by Intel AI software and built-in accelerators and managed by VMware Cloud Foundation. VCF provides integrated security capabilities to secure AI, and it is an ideal platform for training and running private LLMs across business functions in an enterprise. The Intel AI software suite and VMware Cloud Foundation are validated on Lenovo ThinkSystem and ThinkAgile servers with 4th and 5th gen Intel Xeon Scalable processors and private LLMs or generative AI models can be deployed at scale along with other AI use cases.

Intel AMX instructions are supported on vSphere 8.0 and above with VMs using virtual HW version 20 and above. The guest OS running Linux should use kernel 5.16 or later and if Tanzu Kubernetes is used, the worker nodes should use Linux kernel 5.16 or later.

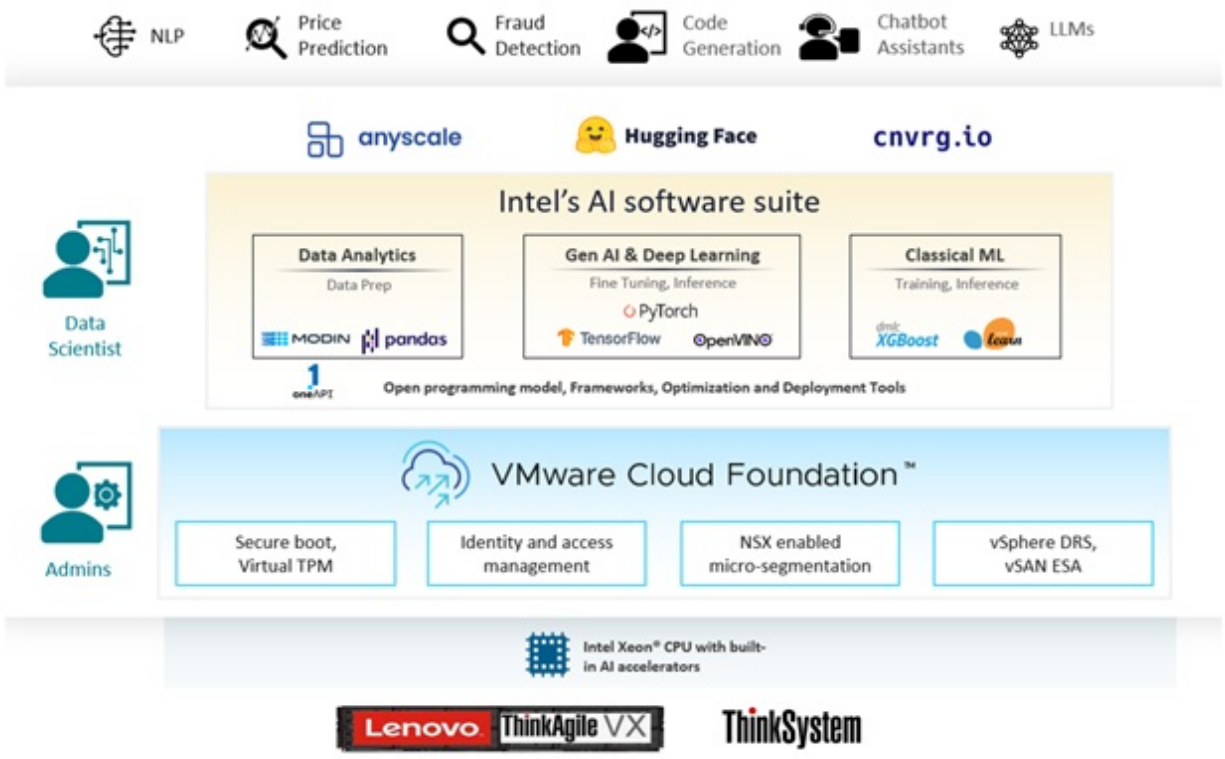

Figure 1. VMware Private AI with Intel on Lenovo ThinkAgile VX and ThinkSystem Servers

Llama2 LLM Inference Performance with 4th Gen Intel Xeon Scalable Processors

The Generative AI inference testing with Llama2 7B and 13B model was done on ThinkAgile VX650 V3 server with 4th Gen Intel Xeon Scalable processors by Intel on May 14, 2024. The test was carried out with different input token sizes 32/256/1024/2048 with varying batch sizes of 1-16 to simulate concurrent requests with static output token size 256. The objective of the testing is to validate different scenario's performance with acceptable latency of less than 100ms latency and to compare the results with/without Intel AMX.

The test and inference serving is targeted on a single node with local storage running ESXi 8.0 U2 and two Ubuntu 22.04.4 guest virtual machines. The model performance can be scaled out by using multiple nodes, but it is not in scope of the current version.

Table 2. Test Hardware Configuration

Server	Lenovo ThinkAgile VX650 V3 CN
Processor	2x Intel Xeon Gold 6448H processors, 2x32C, 2.4 GHz
Memory	1024GB (16x64GB DDR5 4800 MT/s [4800 MT/s])
NIC	1x ThinkSystem Mellanox ConnectX-6 Lx 10/25GbE SFP28 2-Port PCIe Ethernet Adapter
Disk	1x ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD 8x ThinkSystem 2.5" U.3 7450 MAX 6.4TB Mixed Use NVMe PCIe 4.0 x4 HS SSD
Hyperthreading	Intel® Hyper-Threading Technology Enabled
Turbo	Intel® Turbo Boost Technology Enabled
NUMA Nodes	2
BIOS	2.14
Microcode	0x2b000461
Hypervisor	VMware ESXi 8.0 U2 22380479
BIOS Settings	Performance (BIOS and ESXi profile), Max C-State =C0/C1
Guest VM	Ubuntu 22.04.4 LTS, 5.15.0-105-generic
VM HW Version	· VM vHardware gen 21 - Intel AMX available for guest OS · VM vHardware gen 17 - Intel AMX is not available for guest OS; Intel® Advanced Vector Extensions 512 (Intel® AVX-512) VNNI is available
VM Configuration	60vCPU (reservation) 400GB RAM (reservation) vmxnet3 Latency sensitivity mode:high multi socket scenario (30 cores per AI instance)

Table 3. Test Configuration

Workload	LLM Inference
Application	Intel Extension for PyTorch (IPEX) with DeepSpeed
Libraries	IPEX 2.2 with DeepSpeed 0.13; Pytorch 2.2 (public releases)
Script	https://github.com/intel/intel-extension-for-pytorch/tree/v2.2.0%2Bcpu/examples/cpu/inference/python/llm
Test Run settings	· warm up steps = 5 · steps = 50 · -a flag (Max number of threads (this should align with · OMP_NUM_Threads)) = 60 · e (Number of inter threads: e=1: run 1 thread per core; e=2: run two threads per physical core) = 1
Model	Llama2 7B & 13B
Dataset	IPEX.LLM prompt.json (subset of pile-10k)
Batch Size	1/2/4/8/16
Precision	bfloat16
Framework	IPEX 2.2 (public release)
# of instances	2

Llama2 7B Performance Results with/without Intel AMX

Figure 2 shows the 2nd token average latency performance with Intel AMX on 4th gen Intel Xeon Scalable processors for Llama 7B model and Figure 3 shows the results without Intel AMX. The test with Intel AMX shows up to 42% in 2nd token latency for the scenario with input/output token size 32/256. The 2nd token latency for different concurrent requests scenarios (batch sizes 1/2/4/8/16) with input/output token size of 32/256, and 256/256 are within an acceptable threshold of 100 milliseconds and it shows significant throughput increase can be achieved with Intel AMX. The results without Intel AMX show all the scenarios with batch size 8/16 exceeded the 100 milliseconds threshold.

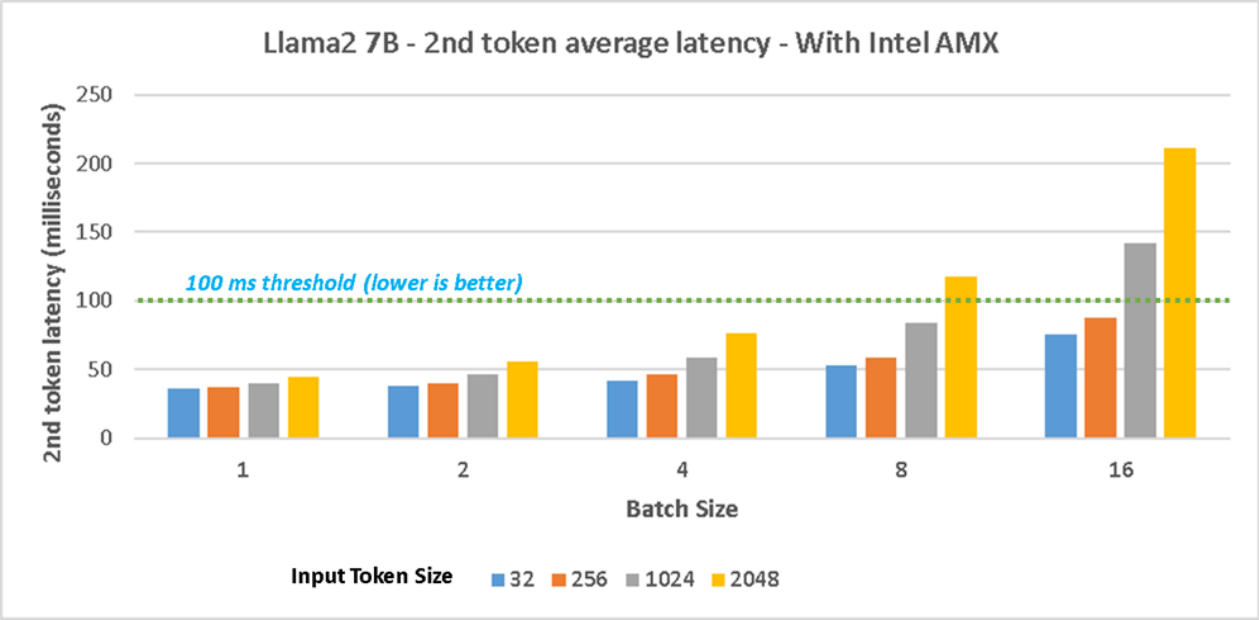

Figure 2. Llama2 7B testing with 4th Gen Intel Xeon CPUs with Intel AMX - 2nd token average latency

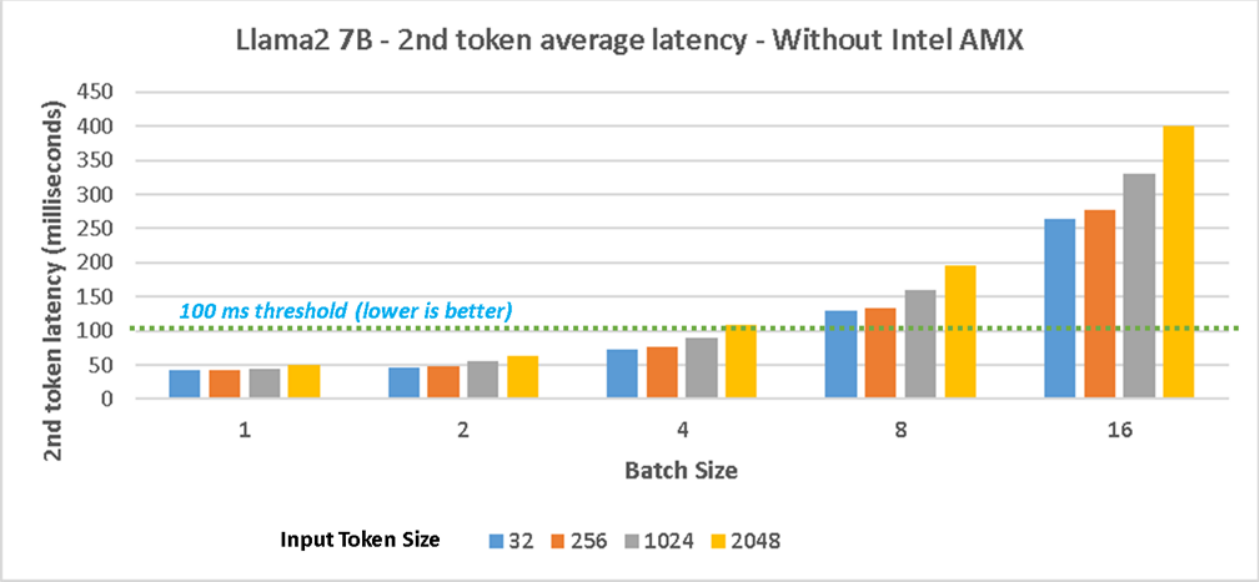

Figure 3. Llama2 7B testing with 4th Gen Intel Xeon CPUs without Intel AMX - 2nd token average latency

Llama2 13B Performance Results with/without Intel AMX

Figure 4 shows the 2nd token average latency performance with Intel AMX on 4th gen Intel Xeon Scalable processors for Llama 13B model and Figure 5 shows the results without Intel AMX. The test with Intel AMX shows up to 18% decrease in 2nd token latency for the scenario with input/output token size 32/256. The 2nd token latency for different concurrent user scenarios (batch sizes 1/2/4/8) with input token size of 32/256 are within an acceptable threshold of 100 milliseconds and it shows considerable throughput increase can be achieved with Intel AMX. The results without Intel AMX shows most of the scenarios with batch size 4/8/16 exceeded the 100ms next token latency threshold.

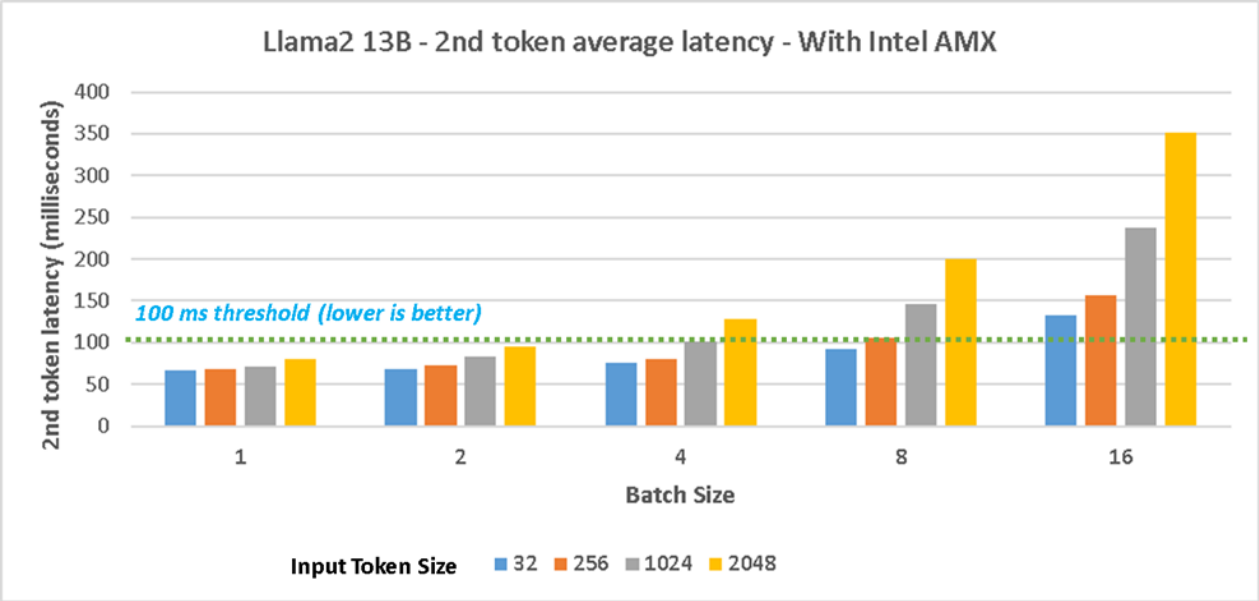

Figure 4. Llama2 13B testing with 4th Gen Intel Xeon CPUs with Intel AMX - 2nd token average latency

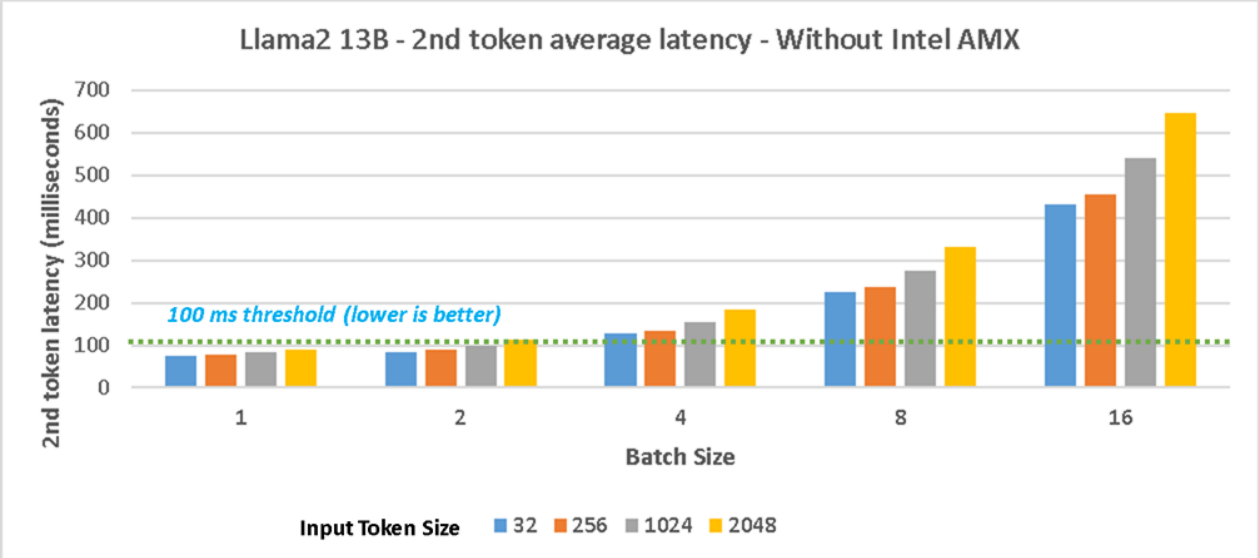

Figure 5. Llama2 13B testing with 4th Gen Intel Xeon CPUs without Intel AMX - 2nd token average latency

Bill of Materials for ThinkAgile VX650 V3

Table 1. Bill of Materials

Part number	Product Description	Quantity
7D6WCTO1WW	Server: Lenovo ThinkAgile VX650 V3 Integrated System	1
BRY9	ThinkAgile VX V3 2U 24x2.5" Chassis	1
B0W3	XClarity Pro	1
BZAK	Customer has VMware by Broadcom Software License	1
BN8K	ThinkAgile VX Remote Deployment	1
BPQD	Intel Xeon Gold 6448Y 32C 225W 2.1GHz Processor	2
BNF9	ThinkSystem 64GB TruDDR5 4800MHz (2Rx4) 10x4 RDIMM	16
5977	Select Storage devices - no configured RAID required	1
B8P1	ThinkSystem 440-16i SAS/SATA PCIe Gen4 12Gb Internal HBA	1
BT2G	vSAN ESA	1
BYRN	AF-2	1
BNEH	ThinkSystem 2.5" U.2 P5620 3.2TB Mixed Use NVMe PCIe 4.0 x4 HS SSD	6
B8LU	ThinkSystem 2U 8x2.5" SAS/SATA Backplane	1
BH8B	ThinkSystem 2U/4U 8x2.5" AnyBay Backplane	1
B8P9	ThinkSystem M.2 NVMe 2-Bay RAID Adapter	1
BTTY	M.2 NVMe	1
BKSR	ThinkSystem M.2 7450 PRO 960GB Read Intensive NVMe PCIe 4.0 x4 NHS SSD	2
BLA3	SW stack for ThinkAgile VX Appliance	1
BN2T	ThinkSystem Broadcom 57414 10/25GbE SFP28 2-Port OCP Ethernet Adapter	2
BPK9	ThinkSystem 1800W 230V Titanium Hot-Swap Gen2 Power Supply	2
6400	2.8m, 13A/100-250V, C13 to C14 Jumper Cord	2
BLL6	ThinkSystem 2U V3 Performance Fan Module	6
BRPJ	XCC Platinum	1
BTSL	ThinkAgile VX650 V3 IS	1
BQQ6	ThinkSystem 2U V3 EIA right with FIO	1
BM8T	ThinkSystem SR650 V3 Firmware and Root of Trust Security Module	1
BP46	ThinkSystem 2U Main Air Duct	1
BLL3	ThinkSystem SR650 V3 PSU Duct	1
BSWK	ThinkAgile SR650 V3 Agency Label - No CCC	1
BPDR	ThinkSystem SR850 V3/SR650 V3 Standard Heatsink w/ Heatpipes	2
BMPF	ThinkSystem V3 2U Power Cable from MB to Front 2.5" BP v2	2
BS6Y	ThinkSystem 2U V3 M.2 Signal & Power Cable, SLx4 with 2X10/1X6 Sideband, 330/267/267mm	1
BACB	ThinkSystem V3 2U SAS/SATA Y Cable from CFF C0,C1/ C2,C3 to Front 8x2.5" BP	2
BSYM	ThinkSystem SR650 V3,PCIe4 Cable,Swift8x-SL8x,2in1,PCIe 6/5(MB) to BP1/BP2	1
BMP2	ThinkSystem V3 2U Power Cable from MB to CFF / Exp v2	1
BRPV	ThinkSystem SR650 V3,PCIe Gen4 CBL,SLx8-Swift,CFF IN-PCIe4	1
BPE3	ThinkSystem SR650 V3 MCIO8x to SL8x CBL, PCIe4, 8x2.5AnyBay, 200mm	2

Part number	Product Description	Quantity
BE0E	N+N Redundancy With Over-Subscription	1
BK15	High voltage (200V+)	1
BQ11	G4 x16/x8/x8 PCIe Riser BLKL for Riser 1 Placement	1
BLKL	ThinkSystem V3 2U x16/x8/x8 PCIe Gen4 Riser1 or 2	1
5641PX3	XClarity Pro, Per Endpoint w/3 Yr SW S&S	1
1340	Lenovo XClarity Pro, Per Managed Endpoint w/3 Yr SW S&S	1
B8Q8	ThinkSystem 440-16i SAS/SATA PCIe Gen4 12Gb Internal HBA Placement	1
5PS7B73066	Premier Advanced ThinkAgile IS - 3Yr 24x7 6Hr CSR + YDYD VX650 V3	1
5AS7B15971	Hardware Installation (Business Hours) for VX650 V3	1
5MS7A87711	ThinkAgile VX Remote Deployment (up to 4 node cluster)	

Accelerated by Intel

To deliver the best experience possible, Lenovo and Intel have optimized this solution to leverage Intel capabilities like processor accelerators not available in other systems. Accelerated by Intel means enhanced performance to help you achieve new innovations and insight that can give your company an edge.

For More Information

To learn more about this Lenovo solution contact your Lenovo Business Partner or visit:
https://www.lenovo.com/us/en/servers-storage/solutions/database/

References:

Lenovo ThinkAgile VX650 V3 2U Integrated System and VX650 V3 2U Certified Node

ThinkAgile VX630 V3 1U Integrated System and Certified Node

Intel AI Development Software

Related product families

Product families related to this document are the following:

- ThinkAgile VX Series for VMware
- ThinkSystem SR630 V3 Server
- ThinkSystem SR650 V3 Server
- 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.

This document, LP1988, was created or updated on July 16, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1988
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1988.

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®
AnyBay®
ThinkAgile®
ThinkSystem®
XClarity®

The following terms are trademarks of other companies:

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

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

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