Microsoft Azure AI services on Lenovo ThinkAgile MX

Solution Brief

Leveraging Azure AI Services On-Premises for Real-World Solutions

In today’s digital landscape, organizations are constantly seeking ways to harness the power of AI to drive innovation, optimize operations, and deliver personalized experiences. While cloud-based AI solutions offer significant benefits, there are scenarios where organizations require on-premises deployment to meet stringent data residency, security, latency, or compliance requirements. Azure AI provides the flexibility to deploy its robust AI services in hybrid and on-premises environments, empowering organizations to use the same advanced capabilities they would find in the cloud, but within the confines of their own infrastructure.

This document introduces how Azure AI's on-premises deployment options enable businesses to unlock new value while addressing their unique needs. By utilizing solutions such as Azure Arc, Azure Machine Learning, and containerized AI services, companies can bring AI closer to their data, maintain control over sensitive information, and ensure compliance with regulatory standards—all while benefiting from the rich toolset and continuous innovation that Azure provides.

In the following sections, we will explore how Azure AI services can be leveraged on-premises to deliver a real-world solution, tackling industry-specific challenges while gaining a competitive edge. This journey demonstrates the seamless integration of AI services into existing workflows and illustrates the transformative potential of Azure's hybrid capabilities to create a scalable, secure, and efficient AI-powered solution.

Azure AI services description

Azure AI services allow developers and organizations access a suite of cloud based artificial intelligence tools without having extensive expertise on the topic or the massive infrastructure behind it.

Azure AI services are a set of artificial-intelligence tools that are provided by Microsoft to help developers and organizations bult AI applications without in depth knowledge on this topic. Some of the services offered are natural language processing, speech recognition, image analysis, and others.

The main available Azure AI services, enabled by Azure Arc, are split into three offerings:

Azure Machine Learning Studio:

- It is a GUI based IDE that helps in constructing ML workflows on azure that can be used by data scientist and engineers across different skill levels and gives the option to have a *no code* or *code-first*

Edge-local retrieval augmented generation (RAG)

- Enhancing the capabilities of LLM`s with local data is essential for companies that want to get the most of their investment in AI capabilities. And, as in certain cases the data governance rules do not allow coping sensitive data off premises.

Azure Video Indexer

- Allows o extract insights and key information from videos. It can categorize objects, scenes, detect faces, recognize speech or even identifying certain actions.

Azure AI services on-premises

One way to access the Azure AI services is to use Docker containers. This flexible approach comes to help customers comply with regulatory requirements, security, or other specific scenarios where processing the raw data in the cloud is not feasible.

Running Azure AI services locally in containers offers customers high throughput and low latency without transaction per second limitations. Additionally, containerizing the application allows for easy scaling up or down using existing Docker infrastructure.

Currently, not all Azure AI services are available in containers. As the this list of services is continuously updated, please refer to the official page for the most up-to-date information:

https://learn.microsoft.com/en-us/azure/ai-services/cognitive-services-container-support

Another option for running Azure AI services is to use them in a disconnected mode through the Docker containers. This allows the usage of certain Azure AI API`s without an internet connection. This requirement may occur in certain cases where the security requirements dictate this or in remote locations where an internet connection may not be available. Some of the services available at this time are natural language processing, speech recognition, image analysis, and more. Please check the latest list at the following link:

https://learn.microsoft.com/en-us/azure/ai-services/cognitive-services-container-support#containers-in-azure-ai-services

Azure AI services on Lenovo ThinkAgile MX455 V3 Edge Premier Solution

The Lenovo ThinkAgile MX455 V3 Edge Premier Solution offers one 4th Generation AMD EPYC processor, up to 576GB memory, up to eight drive bays, and up to 100Gb network connectivity and with a 2U height and short depth case that can go almost anywhere. It comes with the Azure Local 23H2 operating system and Lenovo ThinkAgile support, offering a single point of contact for hardware issues and L1, L2 software support from Lenovo. The hardware configurations can be tailored to meet specific task requirements, including options for up to six NVIDIA A2 GPUs or increased storage. This solution is validated to operate continuously in temperatures from 5°C to 45°C. Some configurations meet NEBS Level-3 and ETSI standards, enabling them to function for 96 hours in temperatures ranging from -5°C to 55°C, and to withstand high-dust and vibration environments. This makes it an excellent choice for locations without optimal climate control.

Deploying a containerized service on the ThinkAgile MX455 V3 Edge Premier Solution and running it on Azure Kubernetes Services (AKS) locally offers several advantages. By positioning the MX455 solution where data is generated, you can process data locally, which reduces latency and improves response times. This approach also lowers bandwidth requirements, leading to cost savings on data transfer. Additionally, using AKS allows for easy scaling of applications based on demand, ensuring optimal performance. Containerized services provide the flexibility to deploy, manage, and update applications seamlessly.

Figure 1.ThinkAgile MX455 V3 Edge Premier Solution

For more detailed information about the ThinkAgile MX455 V3 Edge Premier Solution please access the following link:
https://lenovopress.lenovo.com/lp1889-thinkagile-mx455-v3-edge-premier-solution

Azure AI Services on ThinkAgile MX455 V3 Premier Solution Use Cases

Possible use cases for the Azure AI Services on ThinkAgile MX455 V3 Premier Solution with are:

Edge-local Retrieval Augmented Generation (RAG)

- More and more companies are starting to use LLM models to optimize their operations. The next evolution involves ingesting internal data, enabling the models to provide information based on this data. With the increasing attention given to where internal data is being placed, one option is to use Edge-local RAG. One scenario is offering a chatbot that can assist employees in choosing various benefits or even accessing certain documents usually provided by the human resources department. Additionally, as chat history has been recently integrated, it helps maintain a more natural conversation by preserving continuity and context.

Azure Video Indexer

- A company that needs to monitor their parking lot for and identify who accessed the premises by using the face detection.
- A grocery store that has surveillance cameras can search in their recording by specific event using natural language rather than scrubbing in all the recording and switching from one camera to another.
- A website that allows users to upload videos on their website can use the visual content moderation feature to prevent explicit content being uploaded.

Azure Machine Learning Studio

- By using it`s complete web interface a manufacturing company can use the data collected on their machines (environmental information, wear information and also maintenance historical data) to predict the best maintenance interval and avoid downtime.

Azure AI services on ThinkAgile MX650 V3 and MX630 V3

Both the MX650 V3 and MX630 V3 solutions feature the 4th generation Intel Xeon Scalable processors and 5th Generation Intel Xeon Scalable processors, up to 8TB of system memory, multiple storage configuration options and GPUs for accelerating AI workloads. MX Integrated Systems deliver fully validated and integrated Lenovo hardware and firmware, certified and preloaded with licensed Microsoft software. They also include ThinkAgile Premier support with one single point of contact for support of the hardware and software.

Utilizing Azure AI services on the MX650 V3 or MX630 V3 can leverage the enhanced performance provided by chassis that support higher specifications for compute and storage. Additionally, placing these systems in the same datacenter where all the necessary data is collected can significantly reduce processing and access times. For cases where it is necessary, certain Azure AI services can be used in a disconnected environment without internet access, thus complying with some of the most stringent security requirements.

For a detailed list of ThinkAgile MX650 V3 2U Integrated System and Certified node please access the following link:
https://lenovopress.lenovo.com/lp1675-thinkagile-mx650-v3-2u-integrated-system-and-certified-node-4th-generation

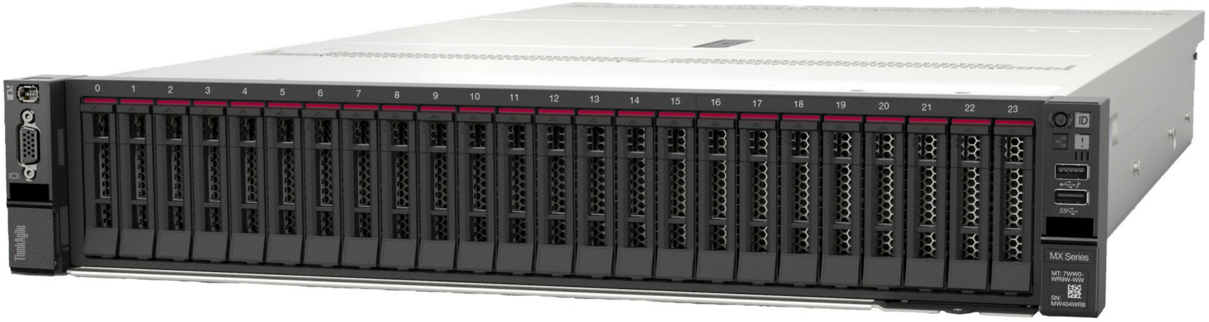

Figure 2. Lenovo ThinkAgile MX650 V3

For a detailed list of ThinkAgile MX630 V3 1U Integrated System and Certified node please access the following link:

https://lenovopress.lenovo.com/lp1674-thinkagile-mx630-v3-1u-integrated-system-and-certified-node-4th-generation

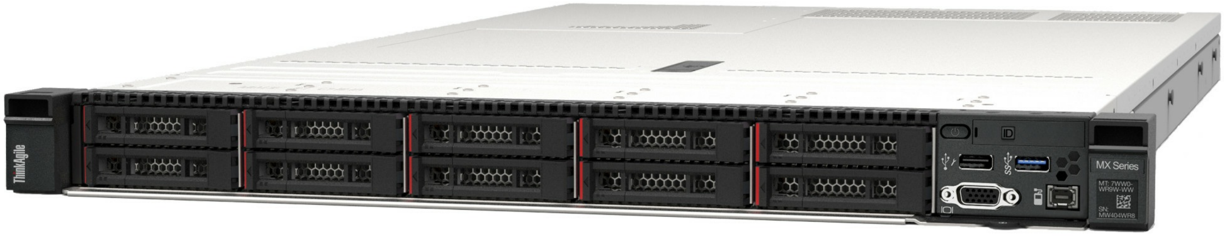

Figure 3. Lenovo ThinkAgile MX650 V3

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, LP2104, was created or updated on December 6, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2104
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2104.

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:

AMD and AMD EPYC™ are trademarks of Advanced Micro Devices, Inc.

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

Microsoft®, Arc®, 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.