Accelerated Artificial Intelligence on Lenovo ThinkAgile MX1021

Vinay Kulkarni
Solutions Architect, Lenovo

In today's world, as a record amount of data is being generated, there is a need to process this data at the edge and gain insights in short order. For performance reasons, the remote data-generating devices must be close to computing and storage resources. Lenovo ThinkSystem SE350 Edge server which is the building block for ThinkAgile MX1021 with its small footprint and power efficiency make it ideal for reliable server-class performance at many Edge locations.

The rugged SE350 can handle temperatures from 0° to 55°C, as well as tolerate locations with high-dust and vibration—such as construction site trailers and manufacturing floors. It can be deployed equally well in a traditional office or branch location due to its office-friendly acoustics. The half-width, short-depth, 1U SE350 can be installed almost anywhere: hung on a wall, stacked on a shelf, or mounted in a rack.

This high-performance server, using the Intel Xeon-D processor, features up to 16 cores, 256GB of RAM, and 16TB of internal solid-state storage. The SE350 also supports the NVIDIA Tesla T4 for workloads such as Edge Inferencing, making it an ideal solution for Artificial Intelligence deployments.

ThinkSystem SE350 server is validated for the Microsoft Azure Stack HCI program and will be available as ThinkAgile MX1021 in the coming weeks. Lenovo ThinkAgile MX1021 solutions are designed to run virtualized applications on-premises in a familiar way, with simplified access to Azure for hybrid cloud scenarios with all of the features you expect like Live Migration, High Availability and VM Load Balancing all built-in. In addition, Azure Stack HCI provides simplified access to Azure for hybrid cloud scenarios like Azure Security Center, Azure Monitor and Azure File Sync for virtually bottomless file storage. This is a perfect solution for IT to leverage existing skills to run virtualized applications on new hyperconverged infrastructure while taking advantage of cloud services and building cloud skills.

The following figures shows the SE350 server and the PCIe Riser Cage, when viewed from the underside. The figure shows an M.2 SATA/NVMe adapter installed in the left wing and an NVIDIA T4 GPU installed in the right wing. The left wing supports SATA or NVMe M.2 drives and the right wing offers a PCIe 3.0 x16 Low Profile slot for supported adapters.

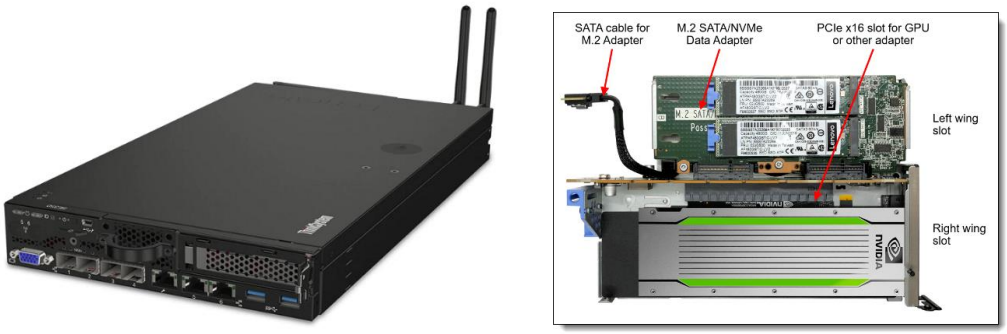

Figure 1. ThinkSystem SE350 and PCIe Riser Cage with NVIDIA T4

Lenovo has worked with Nvidia and Microsoft to test Accelerated AI and Machine Language (ML) scenarios on a couple of SE350 servers in an Lenovo ThinkAgile MX1021 solution.

Artificial Intelligence is the capability of a machine to imitate human intelligence. An artificially intelligent machine can understand speech, analyze images, interact in a natural fashion and make predictions based on data analysis.

Microsoft Azure Machine Learning can be used to quickly and easily build, train and deploy models in the cloud. If you need to run AI/ML scenarios on-premises for regulatory and latency reasons, you can use Azure Notebooks to develop a machine learning module and deploy it to a Linux device running Azure IoT Edge. You can use IoT Edge modules to deploy code that implements your business logic directly to your IoT Edge devices.

Microsoft and a community of partners created ONNX as an open standard for representing machine learning models. Models from many frameworks including TensorFlow, PyTorch, SciKit-Learn, Keras, Chainer, MXNet, and MATLAB can be exported or converted to the standard ONNX format. Once the models are in the ONNX format, they can be run on a variety of platforms and devices.

ONNX Runtime is a high-performance inference engine for deploying ONNX models to production. It's optimized for both cloud and edge and works on Linux, Windows, and Mac. Written in C++, it also has C, Python, and C# APIs. ONNX Runtime provides support for all of the ONNX-ML specification and also integrates with accelerators on different hardware such as TensorRT on NVidia GPUs.

A 2-node Azure Stack HCI solution based on Lenovo ThinkSystem SE350 has been tested using Azure ML modules for Azure IoT Edge on Linux VMs. The testing was done use the Nvidia deepstream stack. Nvidia Discrete Device Assignment (DDA) capable drivers were used to mount vGPU to Hyper-V Linux guest OS. DDA limits exposure of a GPU to a single VM and does not allow for VM failover. Nvidia's container platform was used for this testing.

Summary:
With the large install base of Windows Server, a Lenovo ThinkAgile MX1021 for Microsoft Azure Stack HCI solution built on ThinkSystem SE350 server with NVIDIA T4 GPU is the best platform to run AI/ML scenarios hosted in Linux VMs on Windows Server 2019 host operating system.

Lenovo ThinkSystem SE350 product guide:
https://lenovopress.com/lp1168-thinksystem-se350-edge-server
https://docs.microsoft.com/en-us/azure/machine-learning/concept-onnx#deploy-onnx-models-in-azure