Hybrid Cloud on the Edge: Optimizing for AI with Nutanix and ThinkAgile HX

Article

According to a recent survey by S&P Global Market Intelligence, 77% of IT leaders plan to invest in generative AI, and 96% of that group are looking to extend AI capabilities to edge locations to capitalize on real-time data processing and decision-making.[1] The intense interest in AI at edge locations is not surprising given the potential to enhance customer experiences, streamline operations, and provide competitive advantage — especially leveraging inference AI.

AI inference leverages trained AI models to make predictions or conclusions from net new data. It enables immediate action wherever data is consumed, and it's changing the way organizations operate at edge locations. There are virtually unlimited use cases for this capability. For instance, AI inference might warn of imminent equipment failure at a remote factory. It can help medical staff monitor patients and improve the quality of healthcare decision-making. AI inference can help banks monitor financial transactions at edge locations and flag suspicious activity in real time.

While AI inference unlocks new possibilities, it also creates a unique set of hybrid cloud challenges, any of which can undermine the success of a new AI project before it gains traction. Without a robust hybrid cloud infrastructure, organizations face delays, cost overruns, and the potential for slow adoption when launching AI at the edge.

(1) S&P Global Market Intelligence, "2024 Trends in Data, AI, and Analytics," November 2023

Key considerations for AI inference at edge and ROBO locations

There are a few things to keep top of mind when planning a hybrid cloud for AI inference at the edge:

1. **Scalability and simplicity are critical**

   Solutions with preconfigured hardware and software stacks are essential for edge and ROBO use cases. Deploying new nodes should be a plug-and-play capability. Centralized management tools are also important. Search for solutions with a single pane of glass to monitor and manage both the hardware and software elements of your cloud.

2. **The form factor**

   Scaling to ROBO locations often means putting IT infrastructure in unusual locations. In remote offices, high-end IT appliances may find a home in a broom closet, a small conference room, and even under someone's desk. In situations like this, it helps to have appliances that are purpose-built for ROBO environments, including being smaller and easier to fit into tight spaces.

3. **Reliability matters even more than usual**

   Those edge appliances under a desk or in a broom closet at remote offices? They are expensive to get to, expensive to fix, and expensive to replace — more so than equivalent data center assets. Given the downsides of repeatedly deploying IT resources to edge locations, look for the most reliable appliances possible and best-in-class high availability features.

4. **Choose a partner with the reach you need**

   The fastest way to scale AI at the edge is with a partner who can provide a single point of support for hardware and software, wherever your company does business. Keep your IT team at their desks and focused on more strategic work while a partner enables remote locations. Seek vendors who offer capabilities including 24/7 technical support, regular maintenance schedules, and fast response times. Additionally, consider solutions that provide robust remote monitoring and management capabilities to further minimize the need for on-site interventions.

5. **It may help to engage a design partner too**

   An experienced partner who works side by side with your team to design a hybrid cloud for AI that is tailored to your needs can accelerate the project and improve outcomes. From the speed of the initial rollout to the long-term security of your edge devices, augmenting your team's capabilities during the planning stages can make a significant impact.

6. **Pay close attention to bandwidth and latency**

   Minimize latency for real-time processing and decision-making — or fail to take advantage of the full potential of AI inference at the edge, as well as the security and compliance benefits of processing data at the edge. This is especially important for applications that require immediate data analysis and response. AI is an appropriate application to invest in best-in-class infrastructure up and down the technology stack.

AI at the edge generates large amounts of traffic between remote locations and data centers. Bandwidth management is a critical success factor for this data-hungry application.

Optimize AI inference performance with Lenovo and Nutanix

Organizations seeking to optimize hybrid cloud performance for AI inference should consider the comprehensive solutions offered through the partnership between Lenovo and Nutanix. This collaboration brings together advanced hardware, software, and global services to deliver scalable, easy-to-manage, and secure AI infrastructure.

The **Lenovo ThinkAgile™ HX series**, running Nutanix Cloud Platform, consolidates compute, storage, and virtualization software into plug-and-play building blocks, easily managed in scale-out clusters through a single interface to simplify fleet management of large-scale edge deployments. Maximize high availability with zero-touch deployment and uninterrupted updates, data redundancy features, and cloud backup for maximum uptime.

Scale as you grow from a single node to a multi-node cluster with near-limitless edge nodes at remote locations, including the purpose-built **Lenovo ThinkAgile™ HX360 V2 Edge**, in a pre-validated bundle with Nutanix software and leading open-source AI frameworks to run AI inferencing workloads. Featuring an edge- and ROBO-friendly form factor, the ThinkAgile™ HX360 V2 Edge can take advantage of Nutanix Validated Design for Enterprise Edge with AI, enabling go-live within weeks.

According to a 2024 study by ESG [2], ThinkAgile HX solutions with Nutanix Cloud Platform provide up to 61% reduced TCO and up to 418% ROI.

(2) Enterprise Strategy Group, “Economic Validation: The Economic Benefits of Lenovo ThinkAgile HX Series with Nutanix Cloud Platform,” May 2024

More information

For more information on how Lenovo and Nutanix can optimize your hybrid cloud for AI at the edge, visit https://www.lenovo.com/nutanix-infrastructure.

Authors

Ritu Jain is a Senior Product Manager in Lenovo and she is currently the worldwide product manager for the Lenovo ThinkAgile HX family of Software Defined Infrastructure (SDI) systems. She brings more than 10 years of experience in SDI, Converged and Hyperconverged solutions.

Amalu Susan Santhosh is the Worldwide Technical Product Manager for Lenovo’s ThinkAgile HX and MX/SXM Series of Hyperconverged Infrastructure (HCI) solutions. Amalu is responsible for showcasing the business value and differentiation of Lenovo’s hybrid cloud solutions and contributing to the product lifecycle process.

Related product families

Product families related to this document are the following:

- Nutanix Alliance
- 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, LP1990, was created or updated on July 15, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1990
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1990.

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®

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