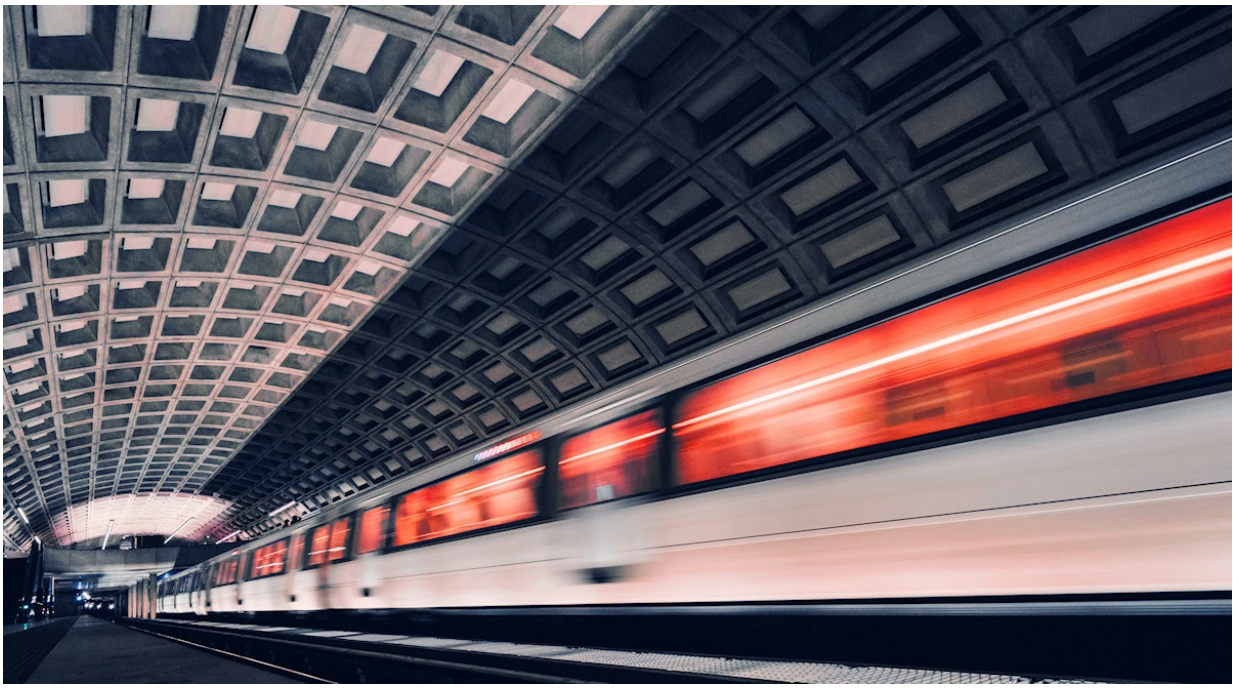
Modernizing Your Hybrid Cloud for AI Inference Workloads for Nutanix and ThinkAgile HX

Article

In today's rapidly evolving tech landscape, organizations are increasingly turning to hybrid cloud solutions to gain agility, scalability, and efficiency. The rising costs associated with public cloud, as well as the performance and security advantages of hybrid cloud have contributed to this shift, while the growing importance of AI inference has accelerated it.

In the 2024 Lenovo Global CIO Report, AI/ML was tied with security as the top priority of the 750 IT leaders surveyed. Supporting AI applications has become one of the fundamental responsibilities of IT, and AI inference is a major part of that for many enterprises. Modern hybrid cloud infrastructure makes it simple, straightforward, and cost-effective to leverage AI inference, while providing transformative performance benefits.



The benefits of hybrid cloud environments in AI

Implementing a modern hybrid cloud for AI inference provides numerous advantages, especially for applications at the edge and remote office/back office (ROBO) locations. AI at the edge allows for real-time processing of data, reducing latency and improving the speed of decision-making. This creates opportunities for applications that leverage real-time insights — at the locations where those insights are most valuable.

AI inferencing on premise also provides greater control over data, helping to improve bandwidth management, reduce costs, and enhance security. But hybrid clouds with aging infrastructure or suboptimal design may not be suited for AI use cases and could become an obstacle for the business.

Considerations in hybrid cloud modernization

If you are examining the readiness of your hybrid cloud for current and future requirements, consider the following:

1. **Scalability and simplicity**

   Scalability — and the ease with which you can scale — is a primary concern. A modern hybrid cloud solution should make the addition of new nodes a seamless and efficient process. Simplicity in scaling ensures that as your business grows, your infrastructure can effortlessly keep pace without significant overhauls or disruptions.

2. **Ease of management**

   Ease of management is related to this. Modern hybrid cloud solutions should offer a unified management interface to simplify the administration of both hardware and software elements. A unified interface reduces the complexity of managing multiple systems and minimizes the risk of errors. Likewise, the interface should have a level of usability that enables basic tasks to be accomplished quickly. Ease of management translates to lower operational costs and reduces the need for specialized IT staff, allowing the organization to focus its resources on strategic initiatives.

3. **Reliability and performance**

   Hybrid cloud infrastructure must be robust enough to handle AI workloads without interruptions. This involves having redundant systems, effective backup solutions, and proactive maintenance protocols. It also means selecting providers with a clear track record of building reliable solutions.

   AI applications, including AI inference, require greater computing resources than most other applications, which means infrastructure should be equipped with the most advanced processors, ample memory, and high-speed storage. This ensures AI models can be trained and deployed quickly, providing timely insights and accelerating business value.

4. **Security and compliance**

   Maintaining data integrity and preventing unauthorized access is nonnegotiable. Security solutions must integrate advanced security measures such as data encryption, secure access controls, and real-time threat monitoring. These features safeguard sensitive data and ensure compliance with industry standards and regulations.

5. **Cost efficiency and sustainability**

   Modern hybrid cloud solutions should reduce operational expenses through efficient resource utilization and automation. Moreover, adopting solutions that offer a consumption-based pricing model can help organizations scale their infrastructure as needed without financial strain.

   Additionally, sustainability has become a priority for businesses worldwide. Infrastructure solutions that promote eco-friendly operations and reduced power consumption can help organizations reduce costs and meet their environmental goals. Features such as CO_2 offsets and asset recovery services are also valuable for achieving sustainability objectives.

6. **Integration and interoperability**

   A well-designed hybrid cloud should integrate seamlessly with existing IT infrastructure and enterprise applications. This ensures a smooth transition and enhances overall operational effectiveness. Compatibility with solutions from a wide variety of vendors provides another layer of future-proofing against unforeseen requirements, and enables workloads to be balanced across different environments for optimal performance and cost management.

7. **Support and professional services**

   Engaging with a single trusted, experienced provider who offers comprehensive support spanning hardware and software, from initial setup to ongoing management, significantly reduces the complexity and risk of hybrid cloud adoption. Conversely, working with consultants and support specialists from multiple technology providers is painstaking and time consuming. Finding a turnkey, one-stop partner is invaluable.

Lenovo and Nutanix simplify and accelerate AI inference deployments

The partnership between Lenovo and Nutanix offers compelling solutions to bring hybrid cloud AI to any organization. Lenovo's **ThinkAgile HX665 V3** and **ThinkAgile HX650 V3** with GPT-in-a-Box™ Nutanix Validated Design (NVD) provide turnkey AI solutions for organizations wanting to implement Generative Pre-trained Transformer (GPT) capabilities while maintaining control over data and applications.

These NVDs are architected and fully tested bundled solutions, including hardware, software, and services which are pre-validated and can be pre-configured to accelerate the deployment of AI initiatives. The solution enables customers to quickly launch every layer of the stack, delivering consistent and verified results. Rather than starting from scratch, customers are provided a simple, proven recipe for success. These solutions include support for several popular large language models, including Llama and Falcon.

Lenovo AI solutions are supported by the expertise of **Lenovo Professional Services**, which has helped enterprises worldwide turn their hybrid cloud vision into reality. With Lenovo and Nutanix, ensure your hybrid cloud is ready to support business growth and transformation in the AI era. To learn more, visit https://www.lenovo.com/nutanix-infrastructure.

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:

- 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, LP1989, 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/LP1989
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1989.

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.