WHITE PAPER

How On-Premises Deployment Can Overcome Six Critical AI Challenges

Certified, interoperable hardware and software bundles and expert support enable enterprises to start small and easily scale AI applications.

IDG COMMUNICATIONS INC.

ACROSS INDUSTRIES, businesses need a fast, cost-effective onramp to apply artificial intelligence (AI) and machine learning technologies to their most pressing business problems. For example:

Financial service providers: Text and speech analysis with AI can identify trading signals and analyze historical information to guide high-frequency trading.

Healthcare providers: AI-enabled image processing helps find diseases in medical images more quickly and efficiently than humans, while AI analysis of past diagnoses can help predict diseases.

Retailers: AI can help better forecast demand; optimize pricing to maximize revenue or profits; determine the optimal placement and display of products; and guide product ordering and placement in advance of adverse weather events.

Manufacturers: AI can predict and prevent failures in production equipment; increase the efficiency of production processes; find and prevent manufacturing defects; and create optimized and lower-cost products.

However, barely 50% of AI projects reach production, according to a 2020 Gartner report, AI in Organizations by Claudia Ramos and Erick Brethenoux. Deploying pre-certified bundles of hardware and software on premises—supported by enterprise-grade professional services—can help to overcome challenges and reduce the time, cost, and risk of implementing and scaling an AI workflow.

Public cloud providers offer scalable AI infrastructure, as well as tools for collecting data, creating, training, and managing AI applications. However, on-premises AI can deliver the greatest business benefits most quickly and cost effectively.

According to Gartner, barely 50% of AI projects reach production.

Here are some examples of when to consider using on-prem AI

✓ **For AI applications that use sensitive or proprietary data, which for regulatory or competitive purposes, must remain on premises.**

✓ **When a business can leverage a virtualized, on-premises AI infrastructure for other uses, such as virtual desktop interface (VDI) that is not required for resource-intensive work like training AI applications.**

✓ **In situations where cloud costs (such as those incurred for moving very large quantities of data) rise to unacceptable levels as AI processing needs increase.**

✓ **An organization needs specific configurations of AI hardware that are not available from cloud providers, or for which performance cannot be assessed and assured beforehand.**

✓ **Enterprise-grade support is needed to supplement an organization's own staff and expertise.**

Six Hurdles to AI Success

There are six common challenges that can delay the implementation of AI applications, or prevent companies from achieving competitive advantage from AI.

1 Tools and frameworks

Many AI tools are available from hyperscalers and the open-source community. However, enterprises that lack extensive in-house, AI-trained staff often need help evaluating, choosing, deploying, and integrating the tools needed for the AI lifecycle that spans from data preparation to training of applications to inferring insights.

2 Lack of skills and support

Organizations require support for virtualized and containerized server hardware and software, as well as for the associated storage and networks. While such support is available from hyperscalers, it may be dependent on purchasing large amounts of cloud resources from them. Open-source support relies on the community to maintain the software and fix security vulnerabilities, which can be challenging when integrating multiple open-source tools. Also, open-source developers may not provide the immediate levels of quality support that can be demanded of a paid vendor.

Successful AI projects also require specialized architects and operations staff skilled in the use of virtual machines (VMs) and containers—and knowledgeable about when to use which technology. Successful AI projects also require data managers who can cost-effectively gather, cleanse, and verify data necessary for effective training in the AI workflow. The need for such skills will grow along with the internal use of AI, making the availability of trusted, scalable support even more vital.

3 Performance

Maintaining performance levels as the organization's AI needs increase requires more than just choosing the highest performance GPUs, CPUs, storage, and networking. It also necessitates optimized infrastructure to run AI software, and to tune that infrastructure properly. This includes provisioning the ideal mix of GPUs and CPUs to provide the best performance at the lowest cost, while maintaining service-level agreements and achieving response times. Performance from a business perspective also means not just the speed of an AI inference process, for example, but also how easily and flexibly that high-performance hardware can be repurposed for other uses when AI requirements are less intense, or service-level agreements are not as stringent.

4 Manageability

This refers to the ability to optimize and secure the entire AI hardware and software stack—from proof of concept through to production deployments, as well as through AI project stages including data prep, training, inference, and scaling. Ideally, existing IT operations staff can use virtualization and containerization technologies to dynamically optimize the infrastructure needed for AI.

5 Risk

In AI implementations, technical risks include outright incompatibilities between different components in the hardware and software stack, and the inability to assure specific levels of performance for each combination of CPUs, GPUs, storage, and networking. These factors can also pose a risk when trying to determine the right levels of cloud resources and their associated costs, especially as projects scale.

6 Scaling deployment

This is the ability to maintain performance as AI objectives, the number and size of jobs, the amount of AI data, and associated management requirements increase. It can be especially difficult to manage and scale all these factors when IT teams are lean.

Supermicro and NVIDIA: Speeding AI Success

Supermicro and NVIDIA have combined their strengths in cutting-edge AI servers with leading performance hardware and the associated software stack to meet these six common AI challenges and speed customers to success. Whether GPU servers that incorporate the NVIDIA HGX™ GPUs or the NVIDIA PCI-E GPUs, Supermicro and NVIDIA deliver the maximum performance for organizations that require a complete and integrated system.

Tools and frameworks.

NVIDIA AI Enterprise is a software suite running in the VMware vSphere environment on Supermicro NVIDIA-Certified Systems™. It includes AI tools and frameworks, cloud-native deployment, and infrastructure optimization software to enable rapid AI development and deployment in familiar VMware infrastructures. Its best-in-class tools and frameworks address AI requirements for data preparation (NVIDIA RAPIDS™), training neural networks (TensorFlow and PyTorch), inference (TensorRT™), and scaling inference operations (NVIDIA Triton Inference Server). The suite allows customers to begin with one or more GPUs within a server, and efficient multi-node scaling as datasets grow.

Lack of skills and support.

Supermicro and NVIDIA provide enterprise-level support for the hardware, OS, NVIDIA AI Enterprise suite, and the server, storage, and network infrastructure, as well as Kubernetes consulting services for multi-server containerized deployment. Consulting services include support for specific technical questions, architecting and implementing the complete hardware and software stack to deliver value from AI initiatives, and data preparation and management. The Supermicro and NVIDIA solution also enables IT operations staff to deploy AI workloads in the familiar VMware vSphere environment.

Performance.

NVIDIA AI Enterprise is optimized, certified, and supported on VMware to achieve near bare-metal performance with virtualization of AI workloads on Supermicro's wide range of NVIDIA-Certified Systems. These systems support PCI-E Gen 4-based NVIDIA A30, A40, and A100, as well as the NVIDIA HGX-A100™ 8 and 4-GPU systems that allow customers to optimize performance, energy usage, and data center cooling through innovative server design. Supermicro servers that can house NVIDIA GPUs range from 1U to 4U in size, support 1 to 10 GPUs, and can be configured with NVIDIA network accelerators, ConnectX, and Blue-Field® Data Processing Units for fast, low-latency network connectivity.

Manageability.

The tightly integrated and pretested hardware and software solution, combined with end-to-end support, helps customers achieve their AI goals with a centralized infrastructure more efficiently and effectively than is possible in the cloud. NVIDIA AI Enterprise offers flexibility in partitioning GPU resources among VMs to allow for applications that need more CPU and GPU cores for compute-intensive operations, such as matrix multiplications for AI applications.

The NVIDIA A40 enables organizations to use GPUs flexibly—allowing, for instance, deployment for virtual 3D graphics applications during the day and AI workloads at night.

Pre-certified systems and enterprise-class support help organizations scale and optimize their environments and budgets for increased AI requirements.

Risk.

The solution minimizes AI deployment risks because all systems are pretested through NVIDIA's rigorous certification process. Its turnkey solution assures not only compatibility among the hardware and software components, but also assures levels of performance for various configurations. This makes it easier to predict infrastructure costs as AI requirements scale and reduces the cost and effort of coordinating testing of various hardware configurations for performance.

Scaling deployment.

Supermicro GPU servers allow customers to begin with small AI investments and easily grow implementations over time, scaling by adding server clusters to their existing infrastructure implementations.

These solutions can accommodate AI workloads repatriated from the cloud or can run in conjunction with them. NVIDIA AI software containers can run on Supermicro systems in the data center, in edge micro datacenters, on edge servers, and in the cloud if cloud bursting is needed. Kubernetes automates the deployment, scaling, and management of containers in the production environment. The combination of pre-certified systems and enterprise-class support helps organizations scale and optimize their environments and budgets for increased AI requirements.

Summary

The use of AI is not just a technical exercise. It is driven by the need to understand more about the business's operations, customers, and external environment. The choice of AI hardware, software, and hosting location should be driven not only by short-term cost minimization or promises of unlimited hardware scalability, but also by the usability and scalability of the overall AI environment.

While major public cloud providers offer many benefits for various AI use cases, on-premises AI infrastructure—provided by pre-certified hardware and software solutions backed by enterprise-grade support—can dramatically reduce the time to value for AI applications. This is especially true when businesses must protect sensitive data, need highly customized hardware and software configurations, require flexibility in redeploying their AI infrastructure, and require higher levels of support than either hyperscalers or the open-source community can provide.

Power your AI from edge to cloud.

For more information, visit:
supermicro.com/en/products/ampere

SUPERMICRO

NVIDIA