AT A GLANCE

Simplify AI with FlashStack

AI holds the promise of revolutionizing the way we work and the strategies that organizations, regardless of their size, employ to fulfill their missions. Key sectors such as healthcare, finance, and manufacturing have already embraced AI as an integral component of their operations, offering the potential to unlock valuable insights and operational efficiencies. To harness the unique capabilities of AI, IT teams face the formidable challenge of deploying a comprehensive AI infrastructure. According to the Cisco AI Readiness Index, 95 percent of IT leaders believe that AI will increase their infrastructure workloads.

FlashStack for AI, a holistic solution jointly developed by Cisco and Pure Storage, provides organizations with a validated architecture and automation playbook that reduces the risks in building a compute, network, and storage infrastructure for AI workloads at scale.

Mainstream your AI infrastructure

FlashStack is a proven architecture adopted by thousands of customers worldwide. Customers with operating models built on FlashStack can now bring AI workloads into that same domain of simplicity, scalability, security, and control, without introducing a new infrastructure silo. FlashStack offers an AI-ready infrastructure that combines:

Modern AI infrastructure

- The Cisco UCS® X-Series Modular System with X-Fabric technology allows for flexible CPU/GPU ratios and cloud-based management for computing distributed anywhere across the core and edge.
- Cisco Data Center Networking (DCN) AI/ML blueprint with Cisco Nexus® 9000 Series switches for enterprise networking delivers the high performance, throughput, and lossless Ethernet fabrics needed for AI/ML workloads.
- High-performance storage systems from Pure Storage provide wide range of storage options with the scalability and efficiency that large training data sets and model serving require.

Cisco Intersight® and Pure1 intelligent management solutions help simplify both complex and mundane tasks associated with AI infrastructure. Intersight provides single-pane-of-glass observability of your entire infrastructure stack and, along with Pure1, simplifies end-to-end infrastructure lifecycle management and operations for AI projects.

Automate AI deployments

FlashStack for AI uses Cisco Validated Designs (CVDs) to provide detailed blueprints for foolproof deployment of AI platforms. CVDs can reduce deployment time by up to 60 percent and ensure that installation is done correctly and reliably, reducing the risks often associated with running complex AI infrastructure. Additionally, the CVDs are accompanied with automation playbooks (available in the Cisco UCS Solutions GitHub repository) that greatly simplify the deployment of the entire FlashStack for AI infrastructure stack.

Secure your AI platforms

FlashStack for AI incorporates Cisco's and Pure Storage's best-in-class practices. Cisco UCS servers are designed to help prevent attackers from gaining access to systems, installing malicious code, and exploring data. The servers are secured from the firmware up, and a secure boot process helps ensure that the software customers intend to run is what runs. Plus, Pure Storage's SafeModeTM immutable snapshots ensure that AI data cannot be compromised. Additionally, customers can leverage Cisco's cloud security and Extended Detection and Response (XDR) cybersecurity solutions for protection from the data-center core to the network edge.

Sustain your AI architecture

IT data centers typically consume vast amounts of energy, and new AI implementations significantly add to that climate strain. However, FlashStack was redesigned from the ground up to be the most sustainable and energy-efficient infrastructure available. Users have found that FlashStack typically reduces their data-center footprint and energy usage by 85 percent, easily addressing the challenges of power-hungry AI environments.

Full stack AI for enterprises

Integrating new infrastructure for AI workloads can be a daunting task, marked by significant expenses and ongoing management commitments. The demanding prerequisites of vast and diverse AI workloads can strain conventional IT compute, network, and storage architectures, pushing them to their limits. Furthermore, AI applications frequently deal with sensitive enterprise data, intensifying the imperative for robust security and strict adherence to regulatory compliance standards. Consequently, it comes as no surprise that numerous organizations grapple with the formidable demands that emerging AI workloads place on their IT infrastructure.

FlashStack®

AT A GLANCE

FlashStack's high-density, high-performance architecture, designed to power demanding applications, is ideally suited to satisfy the need for efficient AI deployments at any scale. Together with a full partner ecosystem, new Cisco Validated Designs for FlashStack for AI will help IT teams deploy accelerated compute and high-performance storage on a proven converged-infrastructure solution.

The new FlashStack for AI validated designs offer a full stack of integrated hardware and software solutions that enterprises can use to accelerate their AI/ML efforts. The validated designs include:

- An inferencing blueprint for Generative AI, coupled with NVIDIA AI Enterprise software. This solution provides a scalable, modular, and high-performance architecture that enables enterprises to design and deploy an inferencing sever and install Large Language Models (LLMs) to support a wide range of Generative AI use cases such as text and image generation, chatbots, virtual assistants, etc.
- An infrastructure solution for MLOps using Red Hat OpenShift AI for rapidly orchestrating and operationalizing models into production. The solution enables enterprises to manage multiple AI model initiatives simultaneously, with ease, consistency, and at scale, by providing capabilities such as complete life-cycle management and pipeline automation, integrated AI tooling (for example, Jupyter Notebooks with Python, TensorFlow, PyTorch, etc.), and flexible model-serving options (for example, Intel® OpenVINO or NVIDIA Triton).

Visit Cisco design zone to see all the FlashStack design guides.

AI deployments now commonly use Kubernetes containerization. FlashStack for AI validated designs are built on the industry-leading container platform – Red Hat OpenShift. The solutions support Portworx by Pure Storage, the leading container storage and data management platform. Portworx can be easily integrated into FlashStack infrastructure to simplify management of persistent storage for AI containerized workloads.

The validated designs use popular AI models such as Stable Diffusion and Llama 2 LLMs, to demonstrate how customers can operationalize AI applications with FlashStack for AI.

Learn more

FlashStack from Cisco and Pure Storage

flashstack@purestorage.com | www.cisco.com/go/flashstack | www.flashstack.com

C45-4111851-00 01/24