AHEAD

How FlashStack Supports the Latest *Generative AI Workloads*

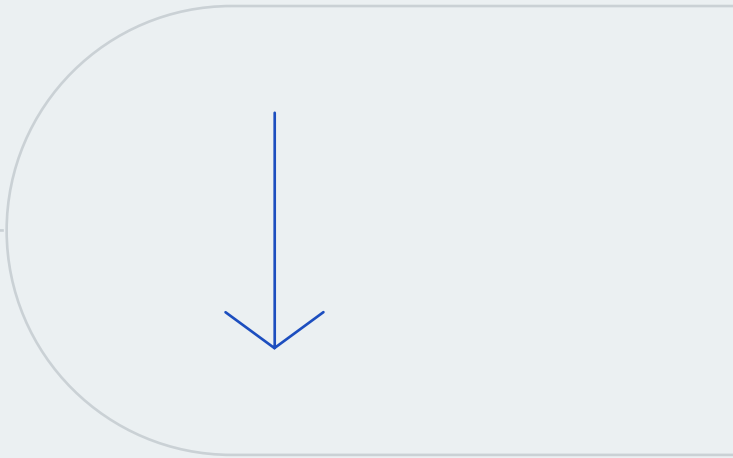
FlashStack from Pure Storage and Cisco along with AHEAD are delivering AI-ready infrastructure

Although generative AI adoption offers strategic opportunities for most enterprises, it also introduces new challenges. The large and diverse data sets required for generative AI model training and inferencing are pushing the performance limits and capabilities of traditional compute and storage architectures. This means many data centers can no longer handle the physical size, power, and cooling requirements of modern AI infrastructure.

At the same time, building full-stack hybrid infrastructure for generative AI from scratch can be complicated and costly. For example, retrieval augmented generation (RAG) is a common technique to enhance LLMs, but it's challenging to build scalable and reliable RAG pipelines.

That's why organizations need to consider a modern software defined infrastructure approach to simplify and optimize their AI infrastructure—and Cisco and Pure Storage have the solution, offering an iteration of their FlashStack reference architecture to address the demands of AI today. FlashStack is a proven AI solution platform with a scalable reference architecture design from application-specific pods to full scale data center virtualization.

In this whitepaper, we'll discuss how FlashStack can handle modern generative AI workloads, along with the validated designs from Cisco and AHEAD Foundry™ to streamline infrastructure deployment.



Introducing FlashStack for AI

FlashStack for AI is a hybrid infrastructure solution from Cisco and Pure Storage that reduces the complexity of setting up AI environments and makes it easier to scale as business needs grow.

Performance

FlashStack can handle data-heavy AI workloads with its low-latency, all-flash storage platform from Pure Storage along with Cisco networking and UCSTM compute resources. This includes the latest FlashArray and FlashBlade storage technologies, along with GPU enabled UCS compute resources.

Scalability

The software-defined hybrid infrastructure enables organizations to selectively add more compute, storage, and networking capacity as needed to scale components independently without over-provisioning.

Simplicity

Cisco Intersight and Pure1 provide intelligent insights and enable the automation of simple and complex tasks across the entire hybrid AI infrastructure stack.

Security

FlashStack supports Cisco's Security Cloud and XDR cybersecurity solutions for layered protection from the data center core to the network edge. Pure Storage's SafeMode also protects data by employing immutable snapshots.

Energy Savings

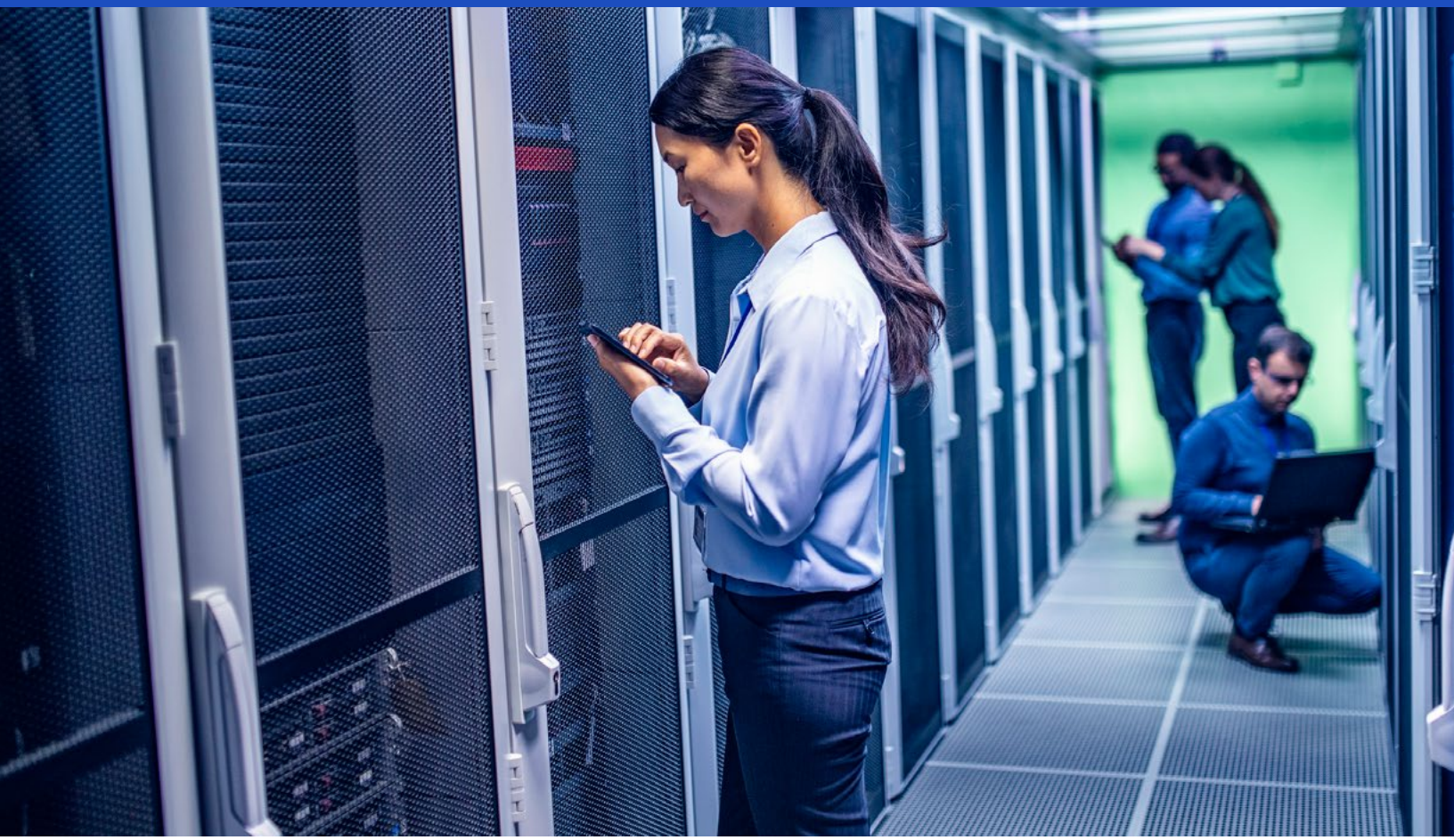
FlashStack was redesigned to be sustainable and energy-efficient infrastructure by consolidating inefficient workloads, using denser designs, and up to 10:1 data reduction.

Data Management

FlashStack makes it easier to move, back up, and access data across hybrid cloud environments with Purity ActiveCluster, Snapshots, and other features specifically designed to maximize the performance of solid state media.

By partnering with a leading provider of enterprise technology solutions, organizations can further reduce the complexity of implementing modern AI infrastructure. For example, AHEAD's Foundry engineers can optimize everything from engineering and integration to deployment so that the configuration of FlashStack meets current and future requirements.

AHEAD PURESTORAGE® cisco Partner



Cisco Validated Designs for FlashStack

Cisco Validated Designs provide detailed blueprints for deploying FlashStack for dozens of use cases. These validated designs offer pre-tested reference architectures that can accelerate delivery time by up to 60% and reduce the deployment and operation risks of complex AI infrastructure.

The Generative AI Inferencing Cisco Validated Design provides a reference architecture for a scalable, high-performance solution aimed at deploying Large Language Models (LLM) and other generative AI models in enterprises. This allows organizations to quickly and economically implement optimized inferencing for various AI models and applications.

The Enterprise RAG Pipeline Cisco Validated Design presents a robust and meticulously tested solution for deploying enterprise-grade RAG pipelines with FlashStack infrastructure and NVIDIA AI Enterprise software. This enables organizations to harness their proprietary data to provide accurate, context-aware responses to generative AI prompts.

Along with validated designs specifically for generative AI, Cisco also provides reference architectures for data pipelines, MLOps, deep learning, and other data-intensive use cases beyond AI. This means FlashStack can be easily deployed to meet a diverse range of business requirements.

Operationalizing FlashStack with AHEAD Foundry

AHEAD is an enterprise technology expert that can help you more easily procure, deploy, orchestrate, and manage large-scale IT systems for generative AI. This includes pre-integrated solutions through AHEAD Foundry, simplified asset management for clients with our AHEAD Hatch® software, and both professional and managed services to support full platform operations.

AHEAD Foundry provides comprehensive services for building and integrating all hardware and networking infrastructure at scale in our dedicated facility rather than on-site. We can pre-configure, integrate, and kit racks in AHEAD Foundry before shipping them to data centers globally.

In addition, AHEAD engineers – working together with Cisco and Pure teams – have created a reference architecture approach to FlashStack solutions with appropriate sizing and high availability designs. These building blocks can be further tailored to the size and scale necessary to meet your use case requirements.

Our plug-and-play approach with AHEAD Foundry accelerates the rollout of new hardware infrastructure like FlashStack for AI and reduces the time to value. In addition, our full-stack integration facilities with liquid-cooling capacity are coming online this year to support future AI infrastructure requiring increased density.

AHEAD Hatch® is our proprietary IT intelligence platform that transforms the way operations teams plan, execute, and manage the lifecycle of their IT infrastructure. Here are some of the key features of AHEAD Hatch:

- **Inventory Management:** Track orders in real-time from procurement and integration through to global delivery. Hatch allows clients to get up to the minute information on orders in production or en route to their destination.
- **Asset Management:** Obtain detailed asset data, such as serial numbers, license keys, mac addresses, and more. This information is harvested during configuration and stored in Hatch so clients can track assets from the site-level down to individual components.
- **Contract Management:** Hatch maintains up-to-date service entitlements and expiration dates to help clients make educated decisions about renewals, cancellations, and when to sunset equipment.
- **Real-time Data Sharing:** Hatch can also integrate with leading enterprise platforms like ServiceNow to share relevant asset data in real-time. And with Hatch's open API, creating custom integrations is straightforward and easy to execute.

AHEAD FlashStack Professional and Managed Services

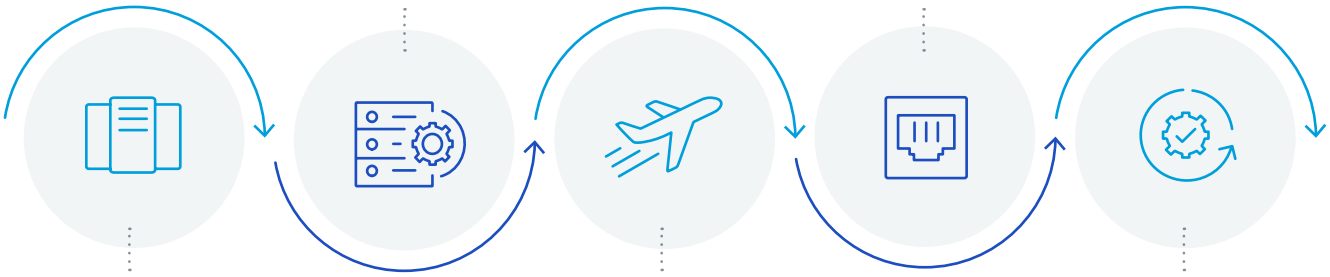
Beyond building, integrating, and deploying FlashStack, AHEAD can provide professional and managed services to reduce the burden on your IT team. This includes on-site final mile configuration services, Intersight and Purity setup to enable unified management across on-premises and cloud environments, and additional managed services for full stack infrastructure.

Rack Build and OS Image Load

AHEAD integration teams install the Pure and Cisco systems, switches, and other equipment into racks, followed by power-on, burn in, and base configuration load.

On-Site Deployment/ FlashStack Configuration

AHEAD's Modern Data Center Team can come on-site to handle connectivity, configuration, storage allocations, and application migrates anywhere these services are required.



Engineering Consultation

AHEAD specialty engineers consult with clients to right-size the rack configuration for the user case. Servers and switches are procured and delivered to AHEAD Foundry to build, configure, and test the integrated solution.

Global Logistics

AHEAD's in-house global logistics team provides end-to-end shipment oversight services into 185+ countries, including the appropriate documentation and certifications to ensure racks arrive on time to locations.

Managed Services for Day 2 Operations

AHEAD's Managed Services Team can be engaged to oversee ongoing FlashStack Infrastructure, hypervisor, storage and backup operations, support, and maintenance.

In short, AHEAD accelerates the impact of FlashStack and other technology investments. Our team of data scientists, architects, and engineers have already helped leading enterprises implement world-class computing solutions.

> **Contact AHEAD for your on-premise AI infrastructure and to learn more about deploying FlashStack for AI.**

AHEAD

Combining cloud-native capabilities in software and data engineering with an unparalleled track record of modernizing infrastructure, we're uniquely positioned to help accelerate the promise of digital transformation.

Visit us at ahead.com.

National Hubs

CHICAGO
444 W. Lake Street
Suite 3000
Chicago, IL 60606

NEW YORK
500 5th Avenue
Floor 17
New York NY 10010

ATLANTA
1117 Perimeter Center
W406
Atlanta, GA 30338

SAN FRANCISCO
2000 Crow Canyon Place
Suite 250
San Ramon, CA 94583