FlexPod®

FlexPod AI for accelerated RAG pipeline

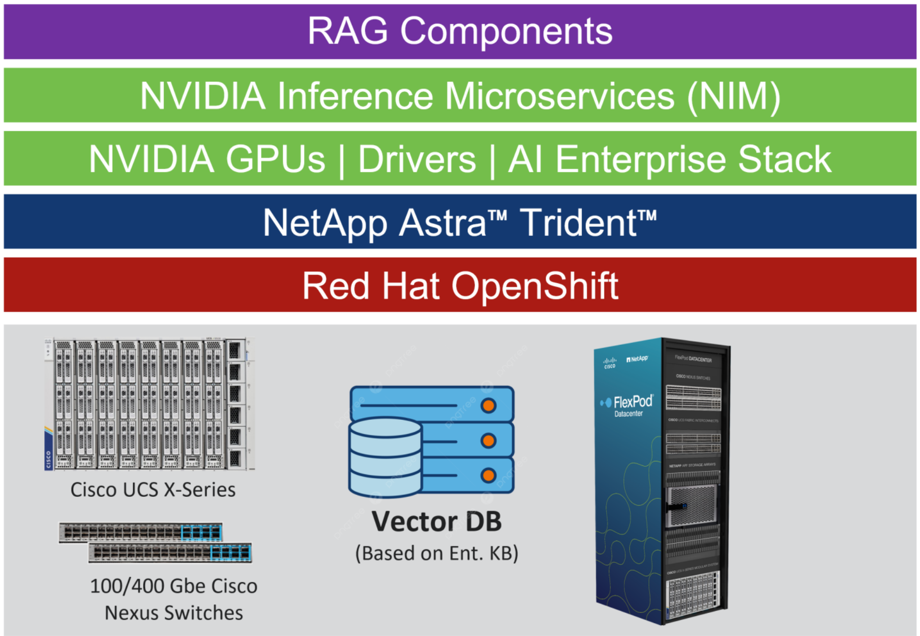
Solution Brief

FlexPod AI brings retrieval-augmented generation to the data center

Generative AI is one of the fastest growing markets in the technology industry. Cisco, NetApp, and NVIDIA have spent years optimizing FlexPod® designs to meet AI challenges. The FlexPod team previously published Cisco Validated Designs (CVD) covering the deployment of generative AI applications in a virtualized environment, as well as a bare-metal deployment of Red Hat OpenShift Container Platform (OCP). NetApp engineers have combined these designs to implement an end-to-end retrieval-augmented generation (RAG) pipeline with NVIDIA Inference Microservices (NIM) and Cisco Webex Chat Bot to provide a real-world example of how RAG can benefit your business.

RAG unlocks the best of large language models

Creating a large language model is no easy task. Some of the most popular models on the market have taken hundreds of days and thousands of GPUs to build a generalized LLM. Although these models are great for high-level inquiries, RAG can help improve and modify them with more specific and up-to-date information by referencing knowledge outside of the LLM's training data. For instance, if you ask a generalized LLM to show the lyrics of your favorite song, it might get a verse or two correct, but then it starts making up the lyrics. This is referred to as a "hallucination." RAG can help by validating the accuracy of the answer against other knowledge bases. Another issue that arises from LLM use is the knowledge cutoff date. RAG can help incorporate data that becomes available after the cutoff, without having to retrain the model.



RAG Components

NVIDIA Inference Microservices (NIM)

NVIDIA GPUs | Drivers | AI Enterprise Stack

NetApp Astra™ Trident™

Red Hat OpenShift

Infrastructure to meet the demand of AI and RAG

This FlexPod CVD uses the principles of RAG to be as efficient as possible.

Performance: The Cisco UCS X-Series Direct tackles some of the largest LLMs with ease. It houses six UCS X210c M7 compute nodes, combined with two UCS X440p PCIe nodes, for a total of four NVIDIA L40S GPUs.

Form factor: With the X-Series Direct chassis, the S9108 Fabric Interconnect fits within the UCS X-Series chassis, reducing form factor and preserving precious rack units.

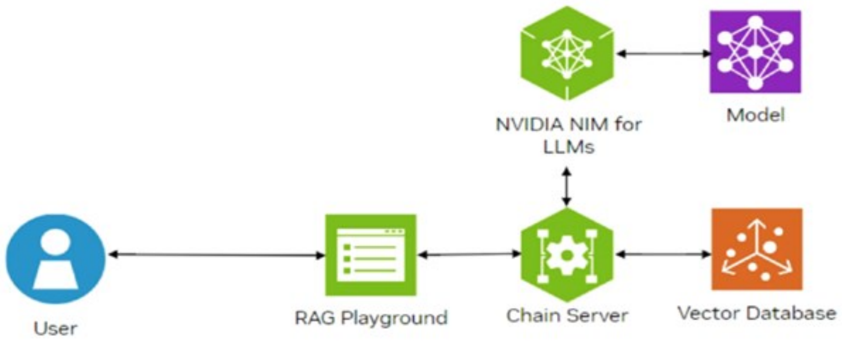
Sustainability: As vector databases grow in size, GPU memory becomes exhausted. To prevent this from happening, the NetApp® ONTAP® AFF C800 provides cost-effective all-flash storage through iSCSI, NVMe over TCP, and NFS while reducing energy consumption.

RAG evaluated through benchmarks and real-world use cases

The CVD covers a bare-metal deployment of Red Hat OCP with NetApp Astra™ Trident™ and DataOps Toolkit, which lays the foundation for the microservices to demonstrate RAG evaluation.

Several benchmarks are used, such as GenAI-Perf, Massive Text Embedding Benchmark, synthetic datasets measured against Ragas, and Milvus against VectorDBBench , to measure inferencing performance and accuracy. These benchmarks set the stage for the real-world use case test: inferencing with Webex Chat Bot and NVIDIA NIM integration. NVIDIA Chain-Server was used to perform RAG testing, along with the ChatNVIDIA Library. Python was used to integrate these with the Webex Chat Bot, which proved to answer the inquiries accurately.

For more information, check out the CVD here.


NVIDIA NIM for LLMs
Model
User
RAG Playground
Chain Server
Vector Database