Scalable conversational AI with HPE GreenLake and NVIDIA AI Enterprise



Businesses across all verticals are evaluating the benefits of incorporating artificial intelligence (AI) into their operations. Conversational AI is a popular use case for increasing efficiency and improving customer experience to gain a competitive advantage. It uses natural language processing and machine learning to understand and respond to human language.

This conversational AI solution is built on a strategic partnership between NVIDIA® and Hewlett Packard Enterprise, combining the capabilities of NVIDIA AI Enterprise and HPE GreenLake. It is designed in a scalable way, helping ensure a seamless transition from pilot phase to full-scale production. This end-to-end scalable solution lays a robust and secure foundation for on-premises speech AI applications and use cases such as intelligent virtual assistants.

The solution

Conversational AI use cases demand an end-to-end optimized solution stack to achieve real-time machine conversing experience indistinguishable from talking to a person. This solution enables on-premises customized cloud experience tailored for both standard and custom GPU compute. It is designed for speech and vision inferencing coupled with high-performance storage choices for audio, text, and image data. The integrated generative AI (GenAI) framework for rapid no-code / low-code development of speech AI applications has the flexibility to integrate seamlessly into modern AI environments. It also integrates into external speech models such as OpenAI.

HPE GreenLake

The solution stack consists of multiple components to deploy an enterprise-ready and secure AI environment on-premises. It builds the foundation for multiple AI applications and is designed to scale and handle increasing requirements for compute and storage systems.

- **NVIDIA AI Enterprise** is a robust, cloud-native software platform that accelerates data science workflows and simplifies the creation and rollout of high-quality AI applications. This platform expedites data science workflows and facilitates the creation and deployment of AI applications. NVIDIA AI Enterprise includes NVIDIA Riva, a GPU-accelerated speech AI application framework and model that facilitates real-time conversational AI pipelines.
- **HPE GreenLake** offers agility and speed to the solution with a simplified on-prem cloud experience. It also allows for rapid deployment, which is a critical factor for businesses looking to integrate AI into their operations. HPE GreenLake is a scalable IT infrastructure service that provides a usage-based IT platform and vertical-based workloads that are aligned to capacity usage, allowing IT organizations to easily scale up and down to handle fluctuations in demand.
- **HPE ProLiant servers** are designed to enable high-performance computing applications. These servers are optimized and certified with NVIDIA L40S GPUs to accelerate model inferencing. The HPE systems are optimized configurations for AI that are validated for performance, manageability, security, and scalability.
- **HPE GreenLake for File Storage** provides an all-flash scale-out solution to handle hundreds of parallel streams. This integration of different technologies processes and responds to user input instantaneously, enhancing the overall user experience. HPE GreenLake for File Storage is designed to meet the demands of emerging file storage workloads, particularly those related to AI, including conversational AI.

Additional components include the Red Hat® OpenShift Container Platform (RHOCP). RHOCP provides a Kubernetes distribution to build, deploy, and manage AI applications. It includes GPU support and enables scale-up training and inference environment seamlessly. For customization and fine-tuning, HPE Machine Learning Development Environment and HPE Machine Learning Data Management can be integrated for optimizing the AI pipeline. HPE Machine Learning Development Environment provides a machine learning development platform. For pipelines and version control, HPE Machine Learning Data Management can be integrated to provide a cloud-native GenAI workflow engine. Figure 1 describes the conversational AI solution architecture.

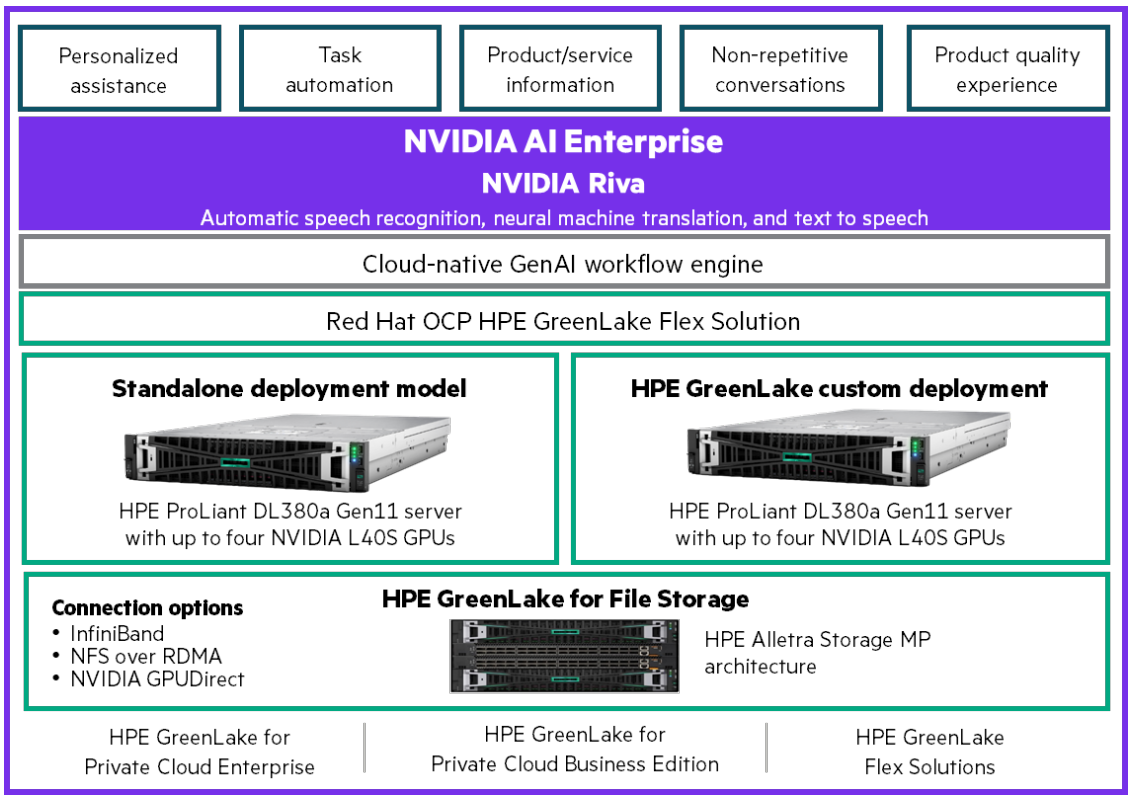

Figure 1. Conversational AI solution architecture

Key benefits

- **Scalability:** The solution scales linearly as load increases, helping ensure robust and reliable performance, even as the number of users grows. The infrastructure can facilitate the rapid scaling of resources to accommodate fluctuations in demand, helping ensure consistent performance regardless of the number of active users.
- **Auto scaling:** The environment can automatically adjust its resources based on the demand, providing an efficient, cost-effective solution that can smoothly handle peak loads and then scale back down during quieter periods. This helps ensure more efficient performance at all times, regardless of fluctuations in user numbers.
- **Enhanced user experience:** Together, these technologies can process and respond to user inputs instantaneously, enhancing the overall user experience. The proposed solution is built with cloud-native data pipeline on RHOCP. These components boosted with NVIDIA AI Enterprise enable integration with varied external data sources to fetch data from and provide streaming capabilities for conversational AI.

Modern storage and optimized compute

With its multiprotocol support and connectivity improvements, HPE GreenLake for File Storage accelerates AI workloads. Its compatibility with InfiniBand, remote direct memory access (RDMA), and NVIDIA GPUDirect technology facilitates quicker data aggregation and preparation for AI, thereby accelerating the fine-tuning and inference of AI models. The all-flash disaggregated HPE Alletra Storage MP architecture provides performant data access to the AI-optimized compute with HPE ProLiant and NVIDIA GPUs to support the entire AI lifecycle. The architecture enables multiple AI pipelines to be run with different sets of requirements and stages. This allows enterprises to adopt AI and keep up with the fast pace of innovation. Figure 2 describes the scalable storage integration to accelerate conversational AI applications.

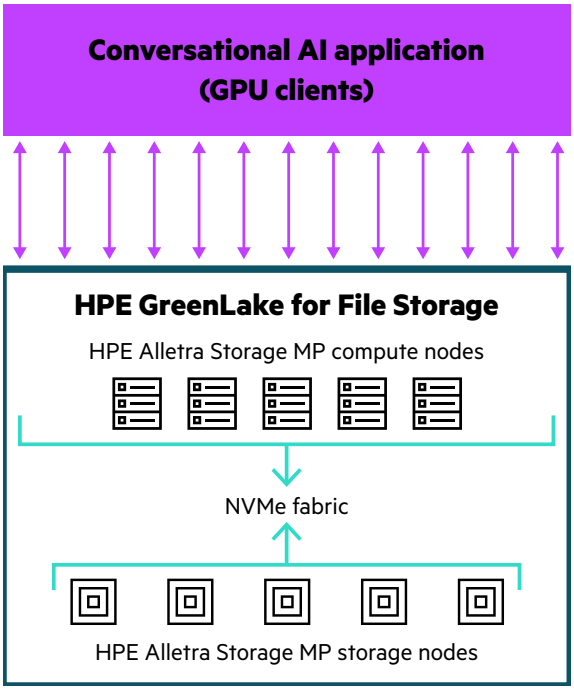

Figure 2. High-performance and scalable storage for AI workloads

These components collectively form the backbone of applications such as intelligent virtual chat assistants, which are becoming increasingly important in enhancing customer interactions across various industries. Virtual assistants can help businesses provide superior, personalized customer service, uncover valuable insights, and drive efficiencies in a scalable way, helping ensure a seamless transition from pilot phase to full-scale production.

Conclusion

The conversational AI solution from HPE presents an end-to-end, high-performing, and scalable solution to run a modern AI environment on-premises. HPE GreenLake provides simple management and a cloud-like experience. Deploying conversational AI on-premises helps ensure regulatory compliance and safeguards sensitive information and data privacy. Its adaptability, robustness, and scalability are designed to meet the demands of a comprehensive and optimized solution to achieve a near real-time machine conversing experience. The integration of these technologies offers an optimized and streamlined solution for businesses, enabling efficient and effective deployment of conversational AI applications and use cases.

Learn more at

HPE.com/us/en/HPE-GreenLake-File-Storage.html

Explore **HPE GreenLake**

Chat now (sales)

a50010218ENW

Hewlett Packard Enterprise