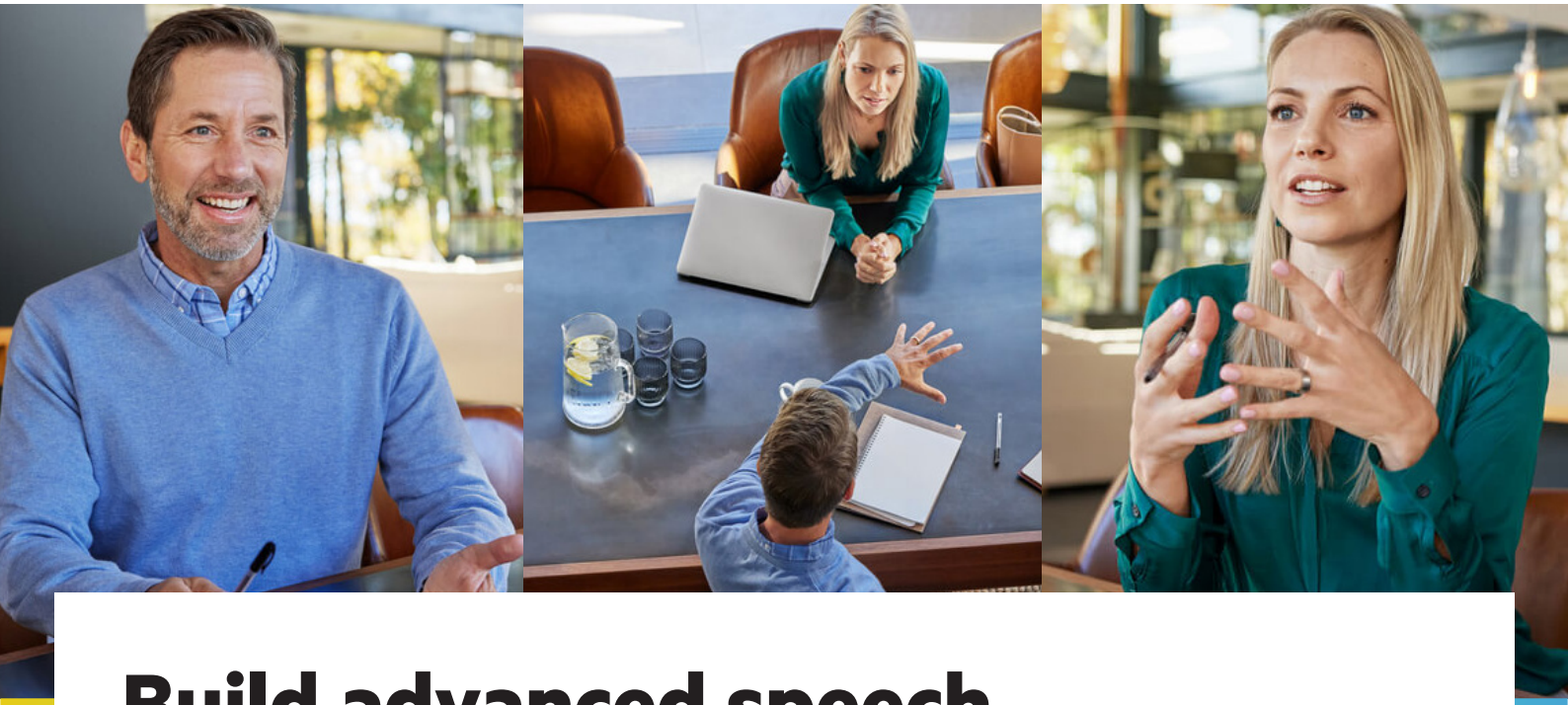
HPE GreenLake



Build advanced speech AI systems with HPE GreenLake and NVIDIA Riva

Enhance your business operations with speech AI-based solutions

Speech AI is powering conversational AI applications

Companies across all industries interact with customers for billions of minutes every day. They deploy conversational applications to use insights from these conversations and build better products, such as customer care agent assists, virtual assistants, and digital avatars. These applications need to enable interactions with users in many languages, understand industry-specific jargon, and respond in real time.

Delivering world-class speech AI skills

NVIDIA® Riva is a GPU-accelerated speech artificial intelligence (AI) SDK with automatic speech recognition (ASR) and text-to-speech (TTS) skills for conversational applications in the cloud, on-premises, at the edge, and in embedded devices. Riva offers out-of-the-box (OOTB), modern speech models that are trained for millions of hours on thousands of hours of audio data. The ASR and TTS pipelines are optimized for real-time performance, with inference running far below the natural conversation threshold of 300 milliseconds. To achieve the best possible accuracy, developers can further fine-tune Riva models on their domain-specific data.

Offering high-quality automatic speech recognition

Every speech AI application relies on converting human voice into readable text (ASR). Often the first step of the speech pipeline, ASR quality influences the effectiveness of downstream conversational AI tasks. Riva offers world-class OOTB ASR models in several languages—including English, Spanish, Mandarin, Hindi, Russian, German, French, Japanese, Arabic, Korean, Portuguese, and Italian—empowering enterprises to deploy high-quality speech AI applications globally. You can customize these models further through word boosting, customized punctuation, and inverse text normalization. Models can also be fine-tuned for custom jargon, domain-specific words, and noisy environments.

Creating expressive, human-like text to speech

Generating human voices from text is essential for customer-facing services in conversational applications. However, producing an expressive and engaging human-like voice requires state-of-the-art AI models, is compute-intensive, and calls for a mature pipeline to fine-tune and express voices. NVIDIA Riva provides professional female and male OOTB voices and enables users to easily customize the models and pipeline for voice pitch, volume, and speed.

Key challenges

Complexity

Building speech AI applications requires hundreds of thousands of hours of audio data, tools, and expertise to build and customize models

Accuracy

Low accuracy leads to a poor customer experience

Performance

Natural conversation requires responses under 300 milliseconds

NVIDIA Riva on HPE GreenLake benefits

- **High speech accuracy**—Riva's pretrained models enable immediate world-class accuracy
- **Easy customization**—Users can fine-tune Riva speech AI models to achieve the best possible accuracy
- **Scale resources easily**—Use continuous monitoring with HPE GreenLake to right-size capacity while capacity-bursting on-site and on demand
- **High throughput**—Riva's GPU acceleration on HPE GreenLake can provide a better throughput than CPU-based applications
- **High scalability**—Riva is fully containerized and can easily scale to hundreds and thousands of real-time streams on HPE GreenLake
- **Flexible deployment**—Riva speech AI skills can run as a service on HPE GreenLake cloud or via standalone deployment on HPE servers

Figure 1 showcases the various deployment models such as HPE Ezmeral, HPE ProLiant servers, and HPE GreenLake that are part of HPE and NVIDIA Riva joint speech AI solution.

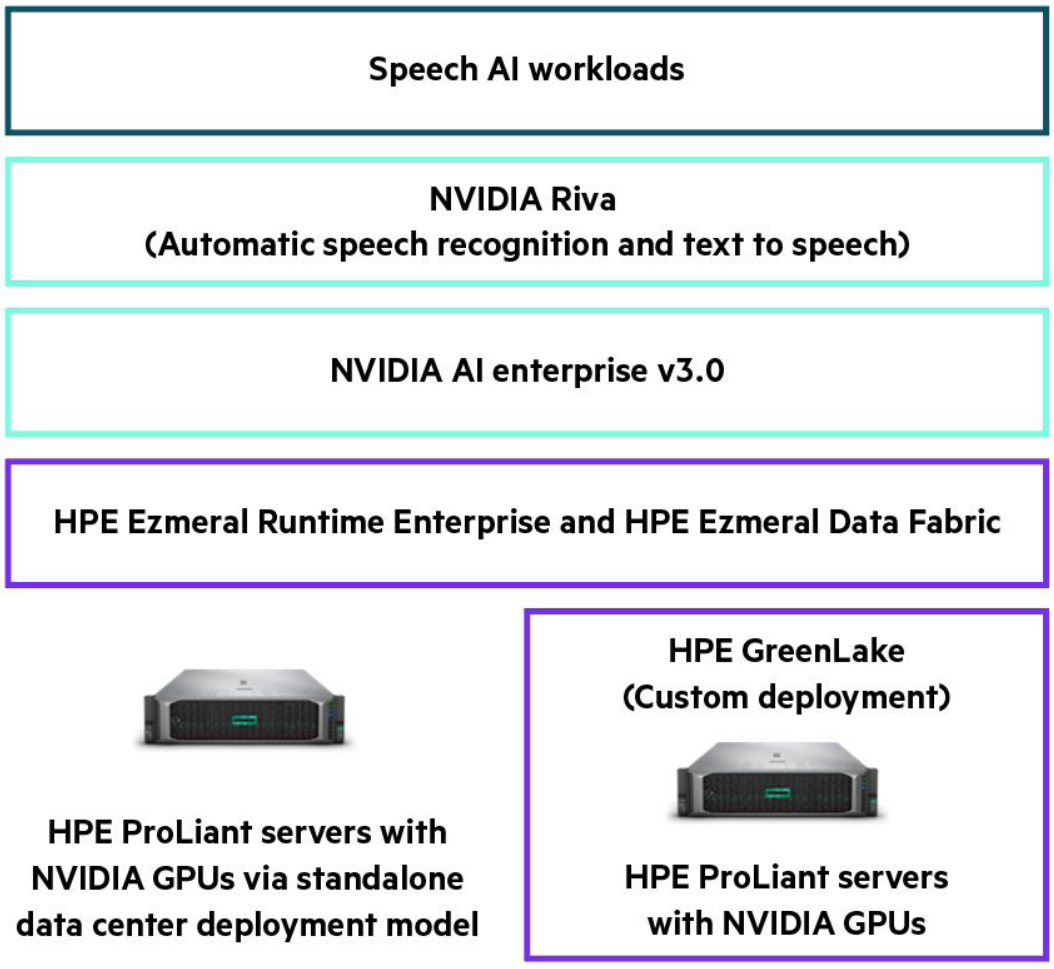


Figure 1. Deployment models of HPE and NVIDIA Riva joint speech AI solution

Ease of deployment and management

HPE GreenLake offers a private cloud service that can accelerate business outcomes for NVIDIA Riva-based workloads by providing easy access to on-demand compute, GPU, and storage resources, pay-per-use* flexibility, and simplified IT operations. HPE GreenLake delivers a cloud-like experience for NVIDIA Riva developers to build, deploy, and monitor their speech AI workloads. Developers can provision development, test, and production environments in minutes, gaining targeted and rapid innovation that can evolve quickly in response to changing data.

Meanwhile, HPE Ezmeral Data Fabric supports IT analytics and data science initiatives by providing a fully integrated data storage and analytics platform. It supports a wide variety of data types and formats, reducing the time-consuming task of negotiating access to geographic speech data silos.

HPE Ezmeral Data Fabric provides a unified view and data access point, simplifies speech and on-speech data management, and reduces the overhead necessary to access RIVA data that might be stored in hybrid and multiclouds. Enterprise customers can use the service to create a unified data repository for data scientists, developers, and IT to access and use, with control of how it is used and shared.

Speech AI has become an integral and essential part of consumers' everyday lives. Enterprises are discovering new ways of bringing great value to their products by incorporating speech AI services. NVIDIA and HPE can help enterprises adopt speech AI with an end-to-end solution. The combined offering of NVIDIA Riva, HPE Ezmeral, HPE certified NVIDIA servers, and HPE GreenLake enable enterprises to quickly build and deploy automatic speech recognition and speech synthesis services-based applications.

Learn more at

hpe.com/us/en/greenlake.html

hpe.com/us/en/software/marketplace.html

nvidia.com/en-us/ai-data-science/products/riva/

Visit **HPE GreenLake**

* May be subject to minimums or reserve capacity may apply

Make the right purchase decision.
Contact our presales specialists.

Chat now (sales)

Call now

Get updates

Hewlett Packard Enterprise

a50007920ENW