

AI Driven Speech Recognition and Synthesis on Dell APEX Cloud Platform for Red Hat OpenShift

A Dell Technologies reference design for conversational AI applications powered by NVIDIA Riva and OpenShift AI

March 2024

H19983

Technical White Paper

Abstract

This document introduces a solution that achieves automated speech recognition (ASR) and text-to-speech (TTS) capabilities using NVIDIA GPUs on the Dell APEX Cloud Platform for Red Hat OpenShift, with Dell software-defined storage and Dell switches.

Dell Technologies AI Solutions

Dell

Reference Design

Copyright

The information in this publication is provided as is. Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

Copyright © 2024 Dell Inc. or its subsidiaries. Published in the USA March 2024 H19983.

Dell Inc. believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Contents

Executive summary.....4

Solution overview.....6

Solution design10

Testing and Application.....15

Conclusion.....22

References.....24

Executive summary

AI has revolutionized human interaction with devices, machines, and computers. Speech recognition has become ubiquitous across various domains, including wearable devices, smart homes, driving assistants, and beyond. This paper provides an overview of a conversational AI solution that leverages Dell Technologies' infrastructure and the NVIDIA Riva software development kit. Dive into the building blocks of this solution and explore how the Dell APEX Cloud Platform for Red Hat OpenShift can streamline the deployment of your own Natural Language Processing (NLP) solution.

Overview

Deep learning (DL), machine learning (ML), and artificial intelligence (AI) are pivotal forces driving business success today. These cutting-edge technologies use models to learn from existing data and generate valuable insights based on new inputs. Their success stems from refined algorithms, access to large datasets, and the increased computational power of CPUs and GPUs.

Natural language processing (NLP) exemplifies AI-powered technology that streamlines operations and simplifies customer journeys. In this context, speech recognition and synthesis have emerged as valuable features. The ability to develop conversational applications with high accuracy and efficiency becomes a competitive advantage across sectors such as retail, manufacturing, IT, banking, telecom, and healthcare.

To help businesses understand the benefits of conversational AI applications and how to implement them, Dell Technologies developed a reference design based on the Dell APEX Cloud Platform for Red Hat OpenShift. The infrastructure comprises a jointly engineered solution containing Dell servers, the Red Hat OpenShift platform, and a management layer integrated into the Red Hat OpenShift console. The speech models and services are based on the NVIDIA Riva solution, which is part of the NVIDIA AI Enterprise platform. Finally, Red Hat OpenShift AI and Red Hat OpenShift Developer are proposed for easy testing and deployment.

About this document

In this paper, we introduce a solution that achieves automated speech recognition (ASR) and text-to-speech (TTS) capabilities using NVIDIA GPUs on the Dell APEX Cloud Platform for Red Hat OpenShift, in conjunction with Dell software-defined storage and Dell switches.

This paper offers an overview of the journey an IT administrator or developer would follow, starting by deploying NVIDIA Riva and performing basic tests, then going through more elaborate testing of its AI speech models with Red Hat OpenShift AI, and ending by building an application to illustrate the capabilities of the solution.

Note: The contents of this document are valid for the described software and hardware versions. For information about updated configurations for newer software and hardware versions, contact your Dell Technologies sales representative.

Audience

This document is intended for decision-makers, managers, IT administrators, architects, field consultants, sales engineers, and anyone else who is interested in developing

conversational AI applications and configuring and deploying OpenShift AI on APEX Cloud Platform for Red Hat OpenShift with NVIDIA GPUs.

Readers should be familiar with the basics of AI/ML, Red Hat OpenShift, containerized applications, and developer/data science platforms.

Revisions

Date	Part number/ revision	Description
March 2024	H19983	Initial release

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on this document. Contact the Dell Technologies team by [email](#).

Author: Fabio Souza and Bryan McFeeters (AI Technical Marketing Engineering)

Contributors: Tiffany Fahmy (AI Technical Marketing Engineering) and Josh Sugarman (Product Marketing)

Note: For links to other documentation for this topic, see [Workload Solutions AI Info Hub](#).

Solution overview

Imagine how much insight a company could extract from its support calls. Frequently asked questions, product usage patterns, customer behavior, and even market trends are examples of invaluable insights that might be available to many businesses in the form of audio. An accurate, high-performance, and scalable speech recognition solution is required to unlock the potential of audio data.

To help clients add conversational capabilities to their modern containerized applications, the solution described in this white paper uses NVIDIA Riva Speech services running on Dell APEX Cloud Platform for Red Hat OpenShift. This platform is well-suited for AI applications, because it offers a turnkey solution for containerized applications, bringing the cloud experience to customers' data centers. It simplifies deployment and alleviates management burdens through a ready-to-use platform, combined with the leading Kubernetes orchestration solution.

Dell APEX Cloud Platform for Red Hat OpenShift integrates with Red Hat OpenShift AI, an open-source ML platform for the hybrid cloud, to build and deploy AI applications. OpenShift AI combines Red Hat components, open-source software, and technology partner offerings with the flexibility to develop and serve models on-premises or in public clouds. The platform makes it simple to embrace hardware acceleration without requiring users to perform daily management of Kubernetes.

NVIDIA Riva is a Kubernetes-based software development kit (SDK) that builds GPU-accelerated speech AI applications. Its pre-trained models empower the creation and deployment of fully customizable, real-time AI pipelines, delivering world-class accuracy across diverse environments—whether in the cloud, on-premises, at the edge, or on embedded devices.

Although, as mentioned in the [Executive summary](#), conversational AI applications are relevant in several use cases and various industries, this document frames the application of speech recognition and natural language processing in the context of call center operations requirements.

Business challenges

Have you heard the famous message “this call might be monitored for quality purposes”? Although this phrase is very commonly used, businesses struggle to leverage their archived call data to achieve perceptible gains in quality and efficiency in their services over time. Meanwhile, customer expectations for high-quality service interactions are growing more than ever. IT teams can play a vital role in overcoming some of the main challenges in this space by tapping into AI solutions, including automatic speech recognition, text-to-speech, natural language processing, and multilingual translation in a multi-cloud environment.

Pressure for quality and efficiency

Call centers constitute an environment that is relentlessly under external and internal pressure. Customers are used to seamless interactions with technology, elevating expectations about service quality, including conversational tools. On the internal side, teams are expected to implement the available AI technologies to run more efficiently, reducing costs and delivering more with less.

Talent retention and training

Because of this challenging environment, it is not uncommon to see an elevated level of turnover in the workforce, making training a constant effort. Business leaders rely on written content to transfer knowledge to new hires but often lose the expertise accumulated by departing workers.

Limited resources for development

Call centers typically handle mostly low-value, high-volume interactions with customers, employees, or other stakeholders. For this reason, despite mounting expectations and challenges with people management, low investments might be available for improvements. Also, rigorous evaluation processes and long approval cycles are required when higher investment initiatives are presented.

Data security concerns

Because of policies, regulations, and other strategic concerns, customer data and business data security are top of mind. Call center teams must assess and ensure data security before adopting any solution to their technical stack. When breaches or data leakages occur, customers' trust is lost and difficult to regain.

Solution approach

An architecture based on the Dell APEX Cloud Platform for Red Hat OpenShift is proposed to address these [business challenges](#). This turnkey solution platform includes Dell hardware with integrated Red Hat software components.

The following figure shows an overview of the main elements.

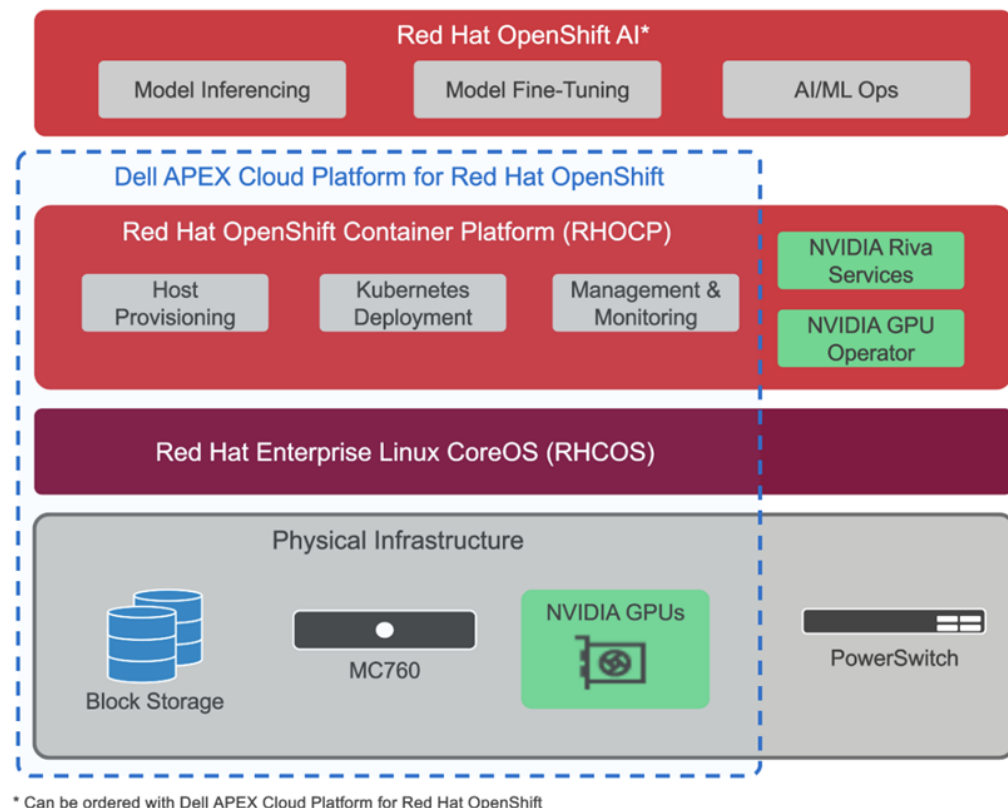


Figure 1. NVIDIA Riva on Dell APEX Cloud Platform for Red Hat OpenShift diagram

While most components in the diagram can be part of the standard Dell APEX Cloud Platform for Red Hat OpenShift delivery, some components, such as the top-of-rack switches, are acquired separately. Dell Technologies offers a family of switches and network products for customers who need a network solution.

A summary of the required installation process for the software components, such as the Red Hat OpenShift AI, NVIDIA GPU Operator, and NVIDIA Riva, are highlighted in the section [Implementation guidance](#). Key elements of this solution are described in the following subsections.

Dell APEX Cloud Platform for Red Hat OpenShift

APEX Cloud Platform for Red Hat OpenShift encompasses most of the three bottom layers in [Figure 1](#). The servers with NVIDIA GPUs are preconfigured with Red Hat Enterprise Linux CoreOS and the Red Hat OpenShift Container Platform, which are responsible for host provisioning, Kubernetes deployment, management, monitoring, and more. They also include the APEX Cloud Platform Foundation Software, which integrates the infrastructure management into the OpenShift Web Console. The storage is built on Dell software-defined storage.

NVIDIA Modules

The two main NVIDIA components in the solution are the NVIDIA GPU operator and the NVIDIA Riva services. Both are installed in the OpenShift container platform, as shown in [Figure 1](#). The GPU operator¹ uses the operator framework in Kubernetes to automate the management of all NVIDIA software components needed to provision the GPU. NVIDIA Riva deployment consists of two containers: an init container to download and deploy the desired models and a second for the Riva API microservices. NVIDIA GPU operator and NVIDIA Riva services use Dell software-defined storage through the CSI (Container Storage Interface) driver to store the models and, optionally, audio and text files, depending on the developed application. Depending on workload and performance requirements, alternative storage solutions, such as ObjectScale or PowerScale, could also be considered.

Red Hat OpenShift AI

Red Hat OpenShift AI, shown as the top layer in [Figure 1](#), offers organizations an efficient way to deploy an integrated set of common open-source and third-party tools to perform ML modeling. The ML models developed using Red Hat OpenShift AI are portable to deploy in production, on containers, on-premises, at the edge, or in the public cloud. It requires a subscription to be installed as an add-on. It is then available in the OpenShift console to create and configure specific data science projects, defining storage in the OpenShift cluster, models, and data sources.

By choosing this solution, IT teams not only overcome the barriers of AI deployment but also optimize business outcomes, as described in the section [Solution benefits](#).

Solution benefits

¹ See the [NVIDIA GPU Operator documentation](#) page for specific details about versions and upgrade.

This solution stands out in the current technological landscape because of the practical and measurable benefits it can provide to any business, particularly in a call center scenario.

Accuracy and speed

The AI-driven speech recognition and synthesis solution is based on NVIDIA Riva's world-class accuracy. The efficient models, running on high-performance hardware, can beat the natural conversation threshold of 300 milliseconds, delivering real-time performance²³. This allows call center teams to use customer voice input instantly to launch queries in the knowledge base and quickly bring solution options to the agent, reducing the time associated with authentication and initial screening and eliminating the need for representatives to physically type the call content. The solution could also enable real-time customer satisfaction monitoring and agent performance. With these features, management can identify root causes and take effective remediation measures faster.

Low learning curve and knowledge preservation

Conversational AI applications for a call center can help shorten the learning cycle for new employees, considering that they have the searchable content of previous calls at their disposal. This agent augmentation approach reduces training costs and preserves accumulated knowledge, even when an employee decides to depart. This will translate into higher customer satisfaction.

Fast implementation

Given the pressure on call center teams, they cannot afford to wait months and years to see their improvement initiatives come to life. Leveraging a fully managed infrastructure that uses the leading Red Hat OpenShift interface significantly reduces the deployment effort and removes the need for new skills in the IT team. Also, NVIDIA Riva includes pre-trained language models for inference, resulting in a shorter time to value.

Reliability and scalability

With the proliferation of technologies that come and go, it is essential to choose a solid option that is both innovative and stable in the market. Dell APEX Cloud Platform for Red Hat OpenShift is built upon trusted solutions that were submitted to thousands of hours of tests. Dell APEX Cloud Platform for Red Hat OpenShift offers a single contact with the world-class Dell Technologies support team for any needs related to the entire infrastructure. The solution also allows IT teams implementing it to have a cloud-like experience to build anything and deploy anywhere, scaling as needed.

Security and flexibility

By deploying an AI-enabled speech solution on-premises, companies can quickly and safely use their internal data to customize the models by introducing specific industry jargon or proprietary product terminology. A call center that owns the infrastructure with end-to-end cyber-resilient architecture has more control over how customer and operations data are handled.

² Consult [NVIDIA Riva developer blog](#) to learn more about NVIDIA real-time performance.

³ See the paper [NVIDIA Riva on Red Hat OpenShift with Dell PowerFlex](#) for a methodology to evaluate NVIDIA Riva performance.

Solution design

This section details the components proposed for the AI conversational solution. It also outlines high-level steps for implementation.

Hardware design The architecture for this solution follows the two-layer integrated deployment option for the Dell APEX Cloud Platform, consisting of compute and storage MC (multicloud) nodes, including 4th Generation Intel® Xeon® CPUs.

Compute

The compute nodes run Red Hat OpenShift software on bare metal, meaning that OpenShift runs directly on the hardware, allowing for virtualization and serverless computing without needing a hypervisor. The currently available NVIDIA Ampere architecture-based GPUs are compatible with [NVIDIA Riva support matrix](#). For the latest product configurations, including GPU options, consult the [Dell APEX Cloud Platform for Red Hat OpenShift specification sheet](#).

Storage

The storage nodes run Dell software-defined storage (SDS) software on Red Hat Enterprise Linux (RHEL) directly on the storage nodes. The management components related to Dell SDS are co-resident on the storage nodes and do not consume resources on the compute nodes. Essential integration between OpenShift and the Dell SDS Element Manager is available. For more information, see the [Dell APEX Cloud Platform for Red Hat OpenShift specification sheet](#).

Networking

The Dell APEX Cloud Platform for Red Hat OpenShift uses two dual-port network interface cards (NICs) to connect to top-of-rack switches and a single gigabit management switch for the iDRACs (Integrated Dell Remote Access Controller). Although the NIC options include 25Gb and 100Gb, the 25Gb NICs can be connected to 10Gb switches if needed. Three VLANs (virtual local area network) are used in the solution: one Management VLAN for OpenShift application traffic and two Data Path VLANs for connectivity to the storage cluster.

Software design This solution is built upon the Red Hat OpenShift Platform. A more detailed description of the components and interfaces is provided in the following subsections.

Red Hat OpenShift web console

Red Hat OpenShift is an enterprise-grade Kubernetes platform for orchestration of containerized applications. It integrates tested and trusted services to reduce the friction of developing, modernizing, deploying, running, and managing applications. It also incorporates Kubernetes enhancements, enabling users to configure and use GPU resources easily. One of the key features of Red Hat OpenShift is its user-friendly graphical interface or the web console. This console enables access to settings and management of all cluster resources. Two main perspectives are available for distinct functions. From the administrator's perspective, the user will find access to workloads, operators, storage, computing, networking, and more. The web console for the Dell APEX Cloud Platform for Red Hat OpenShift also contains a menu to monitor and manage the

platform cluster. From the developer's perspective, the user can build applications and pipelines, and manage helm releases and repositories.

Red Hat OpenShift Operators

Red Hat OpenShift Operators are a powerful mechanism for automating the creation, configuration, and management of Kubernetes-native applications. The validated operators in the OperatorHub provide automation and easy updates on several levels, such as managing the underlying platform components and handling applications as managed services. The following operators are used in this solution:

- **Web Terminal Operator:** enables you to launch a complete terminal session directly in your browser from the OpenShift console, allowing you to interact with the cluster without needing local tool installations. This web terminal comes preloaded with essential CLI (command line interface) tools, including `oc`, for comprehensive OpenShift management.
- **Node Feature Discovery Operator (NFD):** manages the detection of hardware features and configuration in the OpenShift cluster by labeling the nodes with specific attributes. It is a prerequisite for the NVIDIA GPU Operator.
- **[NVIDIA GPU Operator](#):** automates the management of NVIDIA components needed to provision GPU, including the NVIDIA drivers, Kubernetes device plugin for GPUs, the NVIDIA Container Toolkit, automatic node labeling, and others.
- **Red Hat OpenShift AI Operator:** prepares the resources and settings required to run data science projects in the Red Hat OpenShift AI platform.

Red Hat OpenShift AI

Red Hat [OpenShift AI](#) is an open-source offering based on the Open Data Hub project that allows rapid development, training, and testing of AI/ML models. It also makes hardware acceleration access easy and supports integration with popular open-source tools such as Jupyter, TensorFlow, and PyTorch.

Red Hat OpenShift AI provides an AI sandbox platform for ML workloads in the cloud or on-premises. The platform is available as an add-on cloud service or self-managed software product. It can be ordered, installed, and deployed together with APEX Cloud Platform for Red Hat OpenShift, allowing customers to get started with their AI projects quickly.

NVIDIA Riva

Riva offers pre-trained speech models based on BERT (Bidirectional Encoder Representations from Transformers), Transformer-based Seq2Seq, Conformer-CTC, and others. They can be used out-of-the-box for automatic speech recognition (ASR) in multiple languages and speech synthesis (TTS) with expressive human-like voices. The models can also be retrained or fine-tuned using the NVIDIA NeMo framework to include custom datasets such as domain-specific knowledge and custom voice. Models can be deployed as a speech service on-premises or in the cloud using helm charts. Riva's inference, powered by NVIDIA TensorRT optimizations, can deliver the real-time performance required for a natural, human-like interaction. Riva is served using the NVIDIA Triton Inference Server, also part of the NVIDIA AI Enterprise platform. See the different NVIDIA Riva pipelines represented in [Figure 2](#).

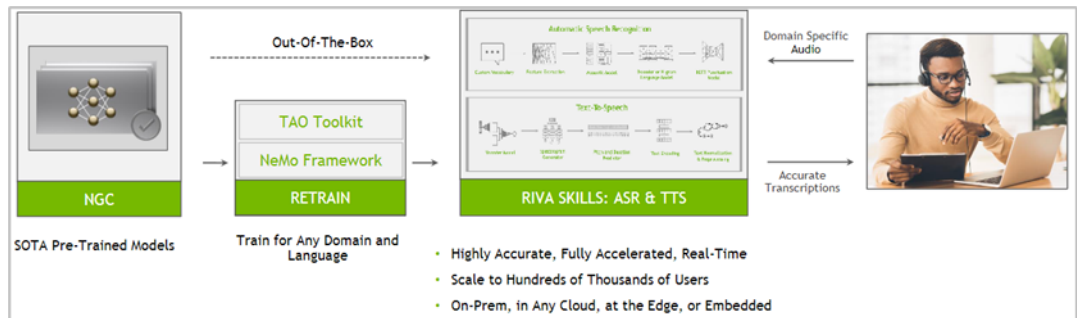


Figure 2. NVIDIA Riva pipelines⁴

NVIDIA Riva API server is available through the NVIDIA NGC portal. By downloading this service, using [Riva Helm chart](#), two containers will be installed: riva-model-init and riva-speech-api. The API is available as a gRPC-based (general purpose remote procedure call) microservice for low latency streaming and high-throughput offline use cases. NVIDIA Riva supports Linux x86_64 and ARM64 architectures and is available for local Docker or Kubernetes deployment.

The following NVIDIA components are also mentioned in this paper for reference, although interacting directly with them is not required for the proposed solution.

- NVIDIA [NeMo](#) is a framework for building, customizing, and deploying AI models. It includes data curation (such as cleaning, filtering, deduplication) and guardrails to assure model accuracy and safety. NVIDIA NeMo can fine-tune NVIDIA Riva pre-trained models.
- NVIDIA [Triton Inference Server](#) is intended to deploy AI models from multiple frameworks in a multicloud environment. NVIDIA Riva is served using the NVIDIA Triton Inference Server.
- NVIDIA [TensorRT](#) optimizes large language models (LLM) for high inference performance on GPUs. NVIDIA TensorRT optimizes NVIDIA Riva.

Riva Contact

The sample application [Riva Contact](#) is a peer-to-peer video chat with streaming ASR and NLP. Two web clients send audio streams from the video chat participants to the Riva Contact server, which submits a gRPC call to the Riva API server, which returns the ASR transcripts. These transcripts are submitted for the NLP service for named entity recognition (NER). The annotated transcripts are then returned to the web clients. This sample application is a single example of a nearly infinite variety of software programs that adopters of this paper's solution can develop to meet their business needs. See [Figure 3](#) for a graphical summary of the process.

⁴ Image source: [NVIDIA Riva user guide](#).

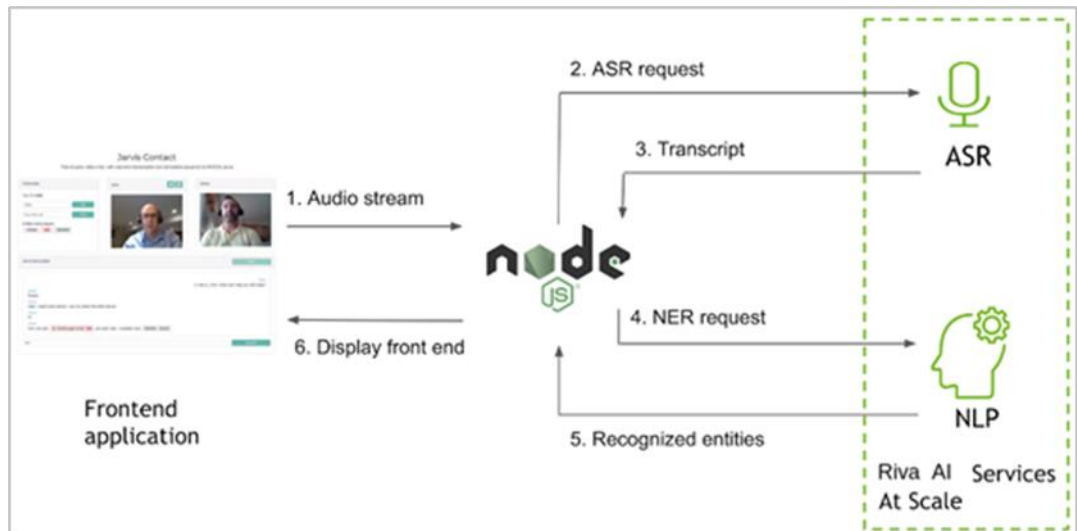


Figure 3. NVIDIA Riva Contact architecture overview⁵

Implementation guidance

This section describes the main steps for implementing the conversational AI application. It includes how to install all components needed, how to enable basic tests using container-based CLI, how to perform more advanced testing using Red Hat OpenShift AI, and how to deploy an application using Red Hat OpenShift builder. The Dell APEX Cloud Platform for Red Hat OpenShift makes installing components easy through the operator process. The OperatorHub marketplace is available in the user interface.

Before you start, verify that you have an available worker node with compatible GPU resources in your Dell APEX Cloud Platform for Red Hat OpenShift cluster. See [NVIDIA RIVA support matrix](#) for more details. Also, confirm your access as a cluster-admin.

Your first step is to install the operators required for this solution. From the administrator perspective in the Red Hat OpenShift console, go to **OperatorHub** and search for the following operators:

- Web Terminal operator
- Node Feature Discovery (NFD) operator
- NVIDIA GPU operator
- Red Hat OpenShift AI operator

Follow the on-screen installation process for each of them until you confirm that they are available in the installed operators' menu, as in [Figure 4](#).

Note: The NFD operator is a prerequisite for the NVIDIA GPU operator.

⁵ Image source: [NVIDIA Riva Contact technical blog](#).

Installed Operators

Installed Operators are represented by ClusterServiceVersions within this Namespace. For more information, see the [Understanding Operators documentation](#). Or create an Operator and ClusterServiceVersion using the [Operator SDK](#).

Name ▾ Search by name... /







Name	Namespace	Managed Namespaces	Status	Provided APIs
 NVIDIA GPU Operator 23.9.1 provided by NVIDIA Corporation	 nvidia-gpu-operator	 nvidia-gpu-operator	✓ Succeeded Up to date	ClusterPolicy NVIDIADriver
 Node Feature Discovery Operator 4.13.0-202402011837 provided by Red Hat	 openshift-nfd	 openshift-nfd	✓ Succeeded Up to date	NodeFeatureDiscovery NodeFeatureRule

Figure 4. Web console showing installed operators

Next, install NVIDIA Riva. To proceed, make sure you have access to the NVIDIA NGC portal and generate an API key, following the steps indicated in the [NGC user guide](#).

The NGC Helm Repository hosts a chart designed for the automated deployment of NVIDIA Riva to Kubernetes clusters, which needs to be adapted for OpenShift. The helm chart pulls container images and model artifacts from NGC, generates the Triton Inference Server model repository, starts the Riva Speech AI server, and exposes it as a service.

You can perform the deployment process in the Red Hat OpenShift web console terminal, as described in the detailed [deployment guide](#) prepared by the “AI on OpenShift” website. The high-level steps to perform the deployment of NVIDIA Riva in Red Hat OpenShift are:

- Download the helm chart from the NGC portal using your API key.
- Edit the **deployment.yaml** file for your Riva container with specific security settings for OpenShift.
- Modify the **values.yaml** file by adding your NGC credentials and specifying PVC (persistence volume claim) and storage class name.
- If applicable, customize your **values.yaml** by choosing specific models that you want to enable/disable in the deployment (commented-out models are disabled). For example, Conformer-CTC for automatic speech recognition, megatron for natural language processing, and vocoder HiFiGAN for speech synthesis are some model categories available.
- Create your project and install Riva with the helm chart.

Testing and Application

This section describes the tests conducted to demonstrate the straightforward setup of NVIDIA Riva on the Dell APEX Cloud Platform for Red Hat OpenShift. The process simulates a hypothetical AI application development cycle, from getting familiar with using AI to delivering an AI-powered application. The approach begins with basic CLI tests to verify server functionality within the pod. Subsequently, integration with Jupyter Notebooks using OpenShift AI is showcased. Finally, a web application operationalized by the Riva-API server, deployed on the same OpenShift cluster, is presented.

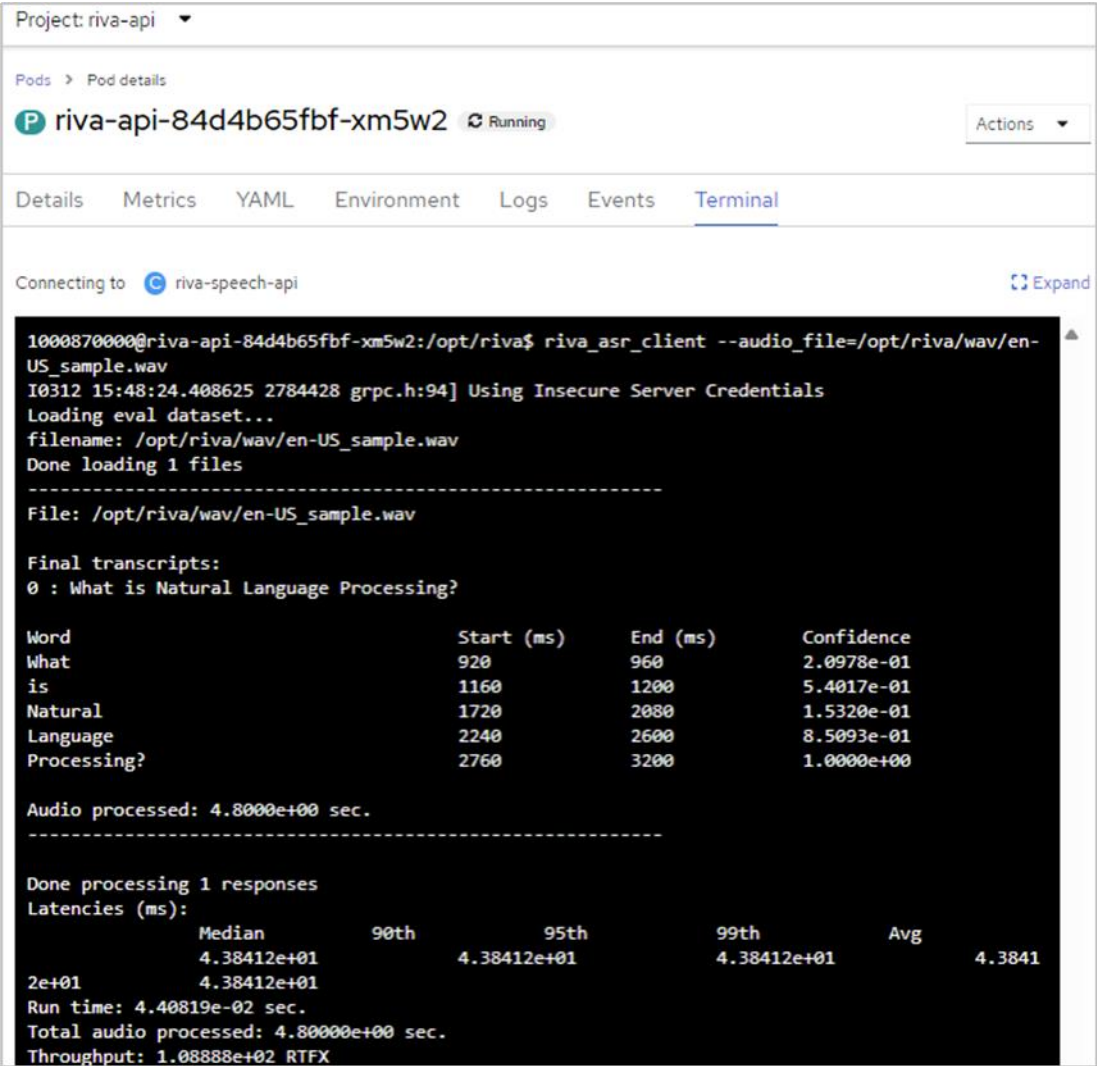
Riva API Pod internal tests

NVIDIA makes it easy to verify that the deployment worked correctly by providing a Riva client and test data within the Riva API server container. The first test performed is a simple audio transcription from a terminal connected to the Riva API container.

To perform this test, in the Red Hat OpenShift console, navigate to pods within the **workloads** menu. Locate your Riva API pod and start a terminal session to the Riva API container. In the terminal, run the following command to test the service:

```
riva_asr_client --audio_file=/opt/riva/wav/en-US_sample.wav
```

The en-US_sample.wav audio file has the question “What is natural language processing?” recorded. When you call the ASR service, it processes your audio file and returns the associated phrase. Notice (in [Figure 5](#)) that the server will also output some performance statistics.



```
1000870000@riva-api-84d4b65fbf-xm5w2:/opt/riva$ riva_asr_client --audio_file=/opt/riva/wav/en-US_sample.wav
I0312 15:48:24.408625 2784428 grpc.h:94] Using Insecure Server Credentials
Loading eval dataset...
filename: /opt/riva/wav/en-US_sample.wav
Done loading 1 files
-----
File: /opt/riva/wav/en-US_sample.wav

Final transcripts:
0 : What is Natural Language Processing?

Word                Start (ms)    End (ms)      Confidence
What                920          960           2.0978e-01
is                  1160         1200           5.4017e-01
Natural             1720         2080           1.5320e-01
Language            2240         2600           8.5093e-01
Processing?         2760         3200           1.0000e+00

Audio processed: 4.8000e+00 sec.
-----

Done processing 1 responses
Latencies (ms):
              Median      90th      95th      99th      Avg
2e+01      4.38412e+01      4.38412e+01      4.38412e+01      4.3841
Run time: 4.40819e-02 sec.
Total audio processed: 4.80000e+00 sec.
Throughput: 1.08888e+02 RTFX
```

Figure 5. Using OpenShift container terminal for initial tests

Additional test codes and data are available for other NVIDIA Riva services. See [NVIDIA Riva quick start guide](#) for more examples.

**Red Hat
OpenShift AI
tests**

Having confirmed that the service is functioning as expected, the next step involves verifying that you can connect to the service externally but within the same OpenShift cluster. Considering this hypothetical AI application development journey, in this phase you learn how to write your own code using the various Riva speech services. You can do this by using the Red Hat OpenShift AI platform to leverage Jupyter notebooks.

Then, in the console application launcher (the black-and-white icon that resembles a grid), navigate to OpenShift Self-Managed Services to open the Red Hat OpenShift AI environment. Under the menu **Data Science Projects**, create a new project.

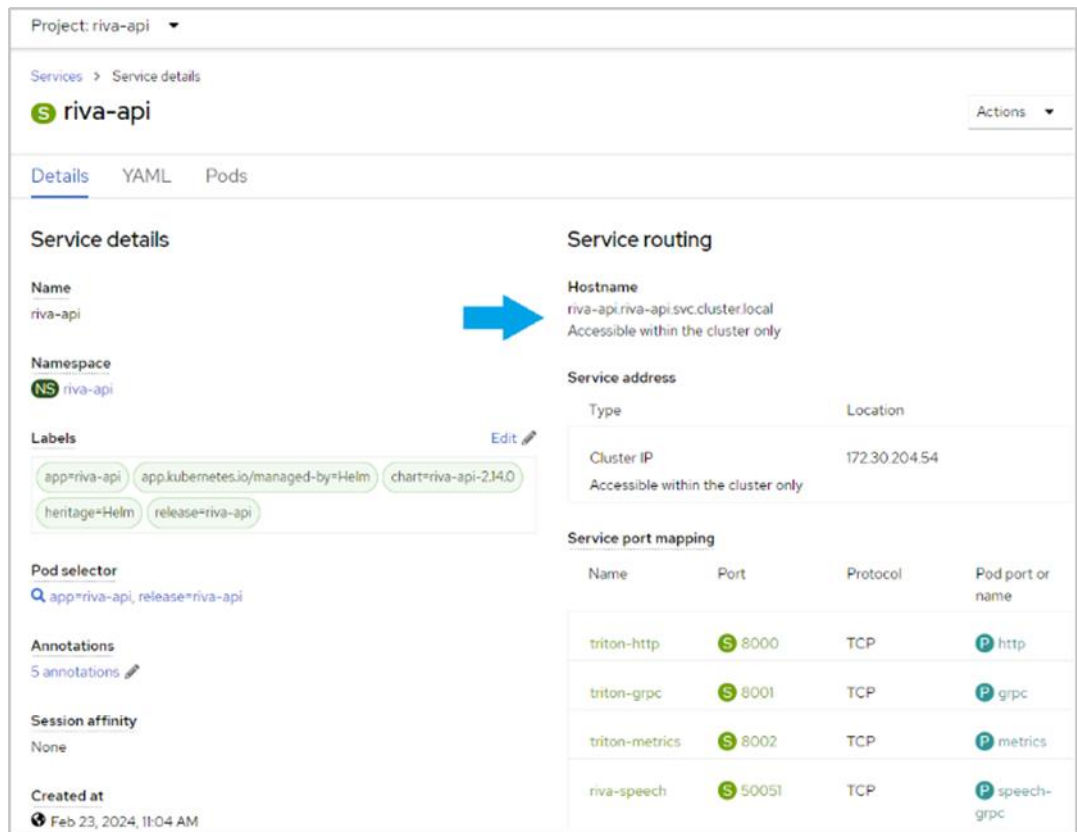
Follow the on-screen steps to create a workbench using your preferred notebook image (minimal Python, for example) and select NVIDIA GPU as the accelerator. You are then required to define your persistent storage.

After the project is created, launch the workbench to see the Jupyter Lab Launcher. Then, in the left navigation bar, go to Git and click **clone repository**. Add the GitHub URL <https://github.com/nvidia-riva/tutorials> to use NVIDIA-created sample notebooks.

Still within the OpenShift AI project workbench, in the Jupyter Lab Launcher, open a terminal and use the following command to install the NVIDIA Riva Python client:

```
pip install nvidia-riva-client
```

The initial test will use the notebook **asr-basics.ipynb**. Notice that after opening a notebook for the first time, you must modify the address for the API server, as shown in [Figure 7](#). You can find this internal hostname under **Services** in the **Networking** menu in the Red Hat OpenShift console. See [Figure 6](#) to identify where the hostname is located.



The screenshot shows the OpenShift console interface for the 'riva-api' service. The 'Service routing' section is highlighted, showing the 'Hostname' as 'riva-api.riva-api.svc.cluster.local' and the 'Service address' as 'Cluster IP' with location '172.30.204.54'. A blue arrow points from the 'Name' field 'riva-api' to the 'Hostname' field.

Type	Location
Cluster IP	172.30.204.54

Name	Port	Protocol	Pod port or name
triton-http	8000	TCP	http
triton-grpc	8001	TCP	grpc
triton-metrics	8002	TCP	metrics
riva-speech	50051	TCP	speech-grpc

Figure 6. Service routing hostname for Riva API server

```
auth = riva.client.Auth(uri='riva-api.riva-api.svc.cluster.local:50051')
riva_asr = riva.client.ASRService(auth)
```

Figure 7. ASR Jupyter notebook with URI configuration for Riva API server

This notebook processes the same **en-US_sample.wav** file used in the section [Riva API Pod internal tests](#). The transcription output is shown in [Figure 8](#).

```

response = riva_asr.offline_recognize(content, config)
asr_best_transcript = response.results[0].alternatives[0].transcript
print("ASR Transcript:", asr_best_transcript)

print("\n\nFull Response Message:")
print(response)

ASR Transcript: What is Natural Language Processing?

Full Response Message:
results {
  alternatives {
    transcript: "What is Natural Language Processing? "
    confidence: 0.430416673
  }
  channel_tag: 1
  audio_processed: 4.8
}
id {
  value: "e8d674d6-8d15-47eb-869d-7824d501d283"
}

```

Figure 8. ASR Jupyter notebook example running with Red Hat OpenShift AI

To demonstrate another example of available functionality, navigate to the notebook **tts-basics-customize-ssml.ipynb**. Again, change the URI for the server hostname as described for the ASR notebook. The first example in this notebook shows how to generate synthetic speech for a given text entry. After running the notebook, you will see a play button in the output cells. This is an audio file with the AI-generated response. Click the play button to hear the results.

Make a gRPC request to the Riva server

For batch inference mode, use `synthesize`. Results are returned when the entire audio is synthesized.

```

[11]: req["text"] = "Is it recognize speech or wreck a nice beach?"
      resp = riva_tts.synthesize(**req)
      audio_samples = np.frombuffer(resp.audio, dtype=np.int16)
      ipd.Audio(audio_samples, rate=sample_rate_hz)

```


[11]: 

Figure 9. TTS Jupyter notebook example running with Red Hat OpenShift AI

This notebook also shows several options to customize the speech output, such as rate, pitch, emotion, emphasis, and even pronunciation.

Note: Parts of these notebooks will only run if the associated models are enabled when deploying Riva. If applicable, go back to your `values.yaml` file to confirm that the specific models you need are included. For instance, a megatron model is required for multilingual neural machine translation (NMT).

Web application sample

Now that you are familiar with the Riva API service, the next step is to deploy an application that can leverage these services. In this example, a video chat doing automatic transcription and named entity recognition (NER) will be deployed using the publicly available code for the Riva Contact application as a proof of concept. This Riva Contact Center Video Conference is a lightweight Node.js sample application. A similar application could be used in a call center, where the call transcription and named entry recognition could launch additional queries to reduce response time, assess agent performance or customer satisfaction metrics, and ingest data for AI model training or fine-tuning.

Before you start to build your application, confirm that you have the integrated OpenShift Container Registry (OCR) deployed to manage your container images. The container registry will be responsible for storing the output from the 'source to image' build, which will be used for deployment, as explained in [understanding image builds](#). See the [internal registry overview](#) page to learn more about deploying a registry in your OpenShift cluster.

Switch to Developer Mode in your Red Hat OpenShift console to start the application deployment. Click the **+Add** option to create your new project. Alternatively, you can select an existing project from the drop-down menu. Find the tile **Git Repository** and select it to import from Git. In the URL field, enter the address <https://github.com/nvidia-riva/sample-apps.git> for the NVIDIA Riva samples apps repository.

In the advanced Git options, add **/riva-contact** in the **Context dir** field to specify the correct application to be imported. A Node.js image will be automatically suggested, as shown in [Figure 10](#). Finalize the remaining steps by giving your application a name and defining a target port. Be sure to select the **Create Route** check box to enable exposure of this application at a public URL.

Project: riva-contact Application: All applications

Git Repo URL *

✓

Validated

▼ Hide advanced Git options

Git reference

Optional branch, tag, or commit.

Context dir


Optional subdirectory for the source code, used as a context directory for build.

Source Secret

▼

Secret with credentials for pulling your source code.

✓ **Builder Image detected.**
A Builder Image is recommended.

 **Node.js 16 (UBI 8)** [Edit Import Strategy](#)

BUILDER NODEJS

Figure 10. Red Hat OpenShift developer tab with application builder image detected

When creating your own application, you will be required to set the address of your Riva API using your internal hostname. For the purposes of this example, a secret was created to overwrite the env.txt file with another one containing the correct Riva API server hostname, because the application code was imported directly from the NVIDIA sample GitHub repository without edits. You can also clone the application and edit the env.txt file to point to the Riva API server and service port deployed in your Red Hat OpenShift cluster.

When you click Create, OpenShift builds the application pod and makes a service and a route to expose it. After the creation process is concluded, you can find the URL link for the application by switching back to the administrator view under Routes in the networking menu.

The application sample has a video chat for two participants. Users can connect using their auto-assigned ID. Start the application to see the speech recognition and natural language processing features in action, with the AI-generated transcript and the live tagging that captures, in this example, persons, locations, organizations, time/date, and others.

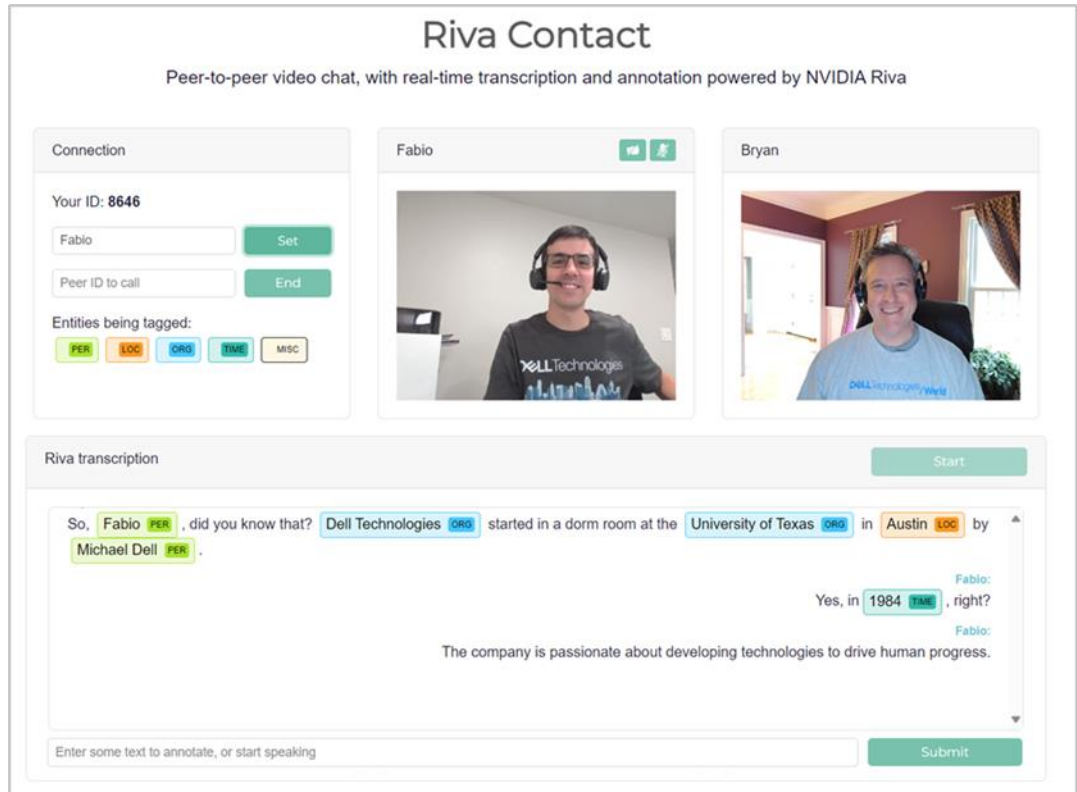


Figure 11. Riva Contact web application demo

This example shows the simplicity of deploying an AI application that can be customized to any business needs. The charts in Figure 12 show that the NVIDIA A2 Tensor GPU usage considering the web application running in the OpenShift cluster was minimal. This proof of concept used basic resources available in the APEX Cloud platform for Red Hat OpenShift. Customers who want to tailor a solution for their business needs and performance requirements, including scaling up applications for concurrent streaming calls, can consider other options of GPUs or storage.

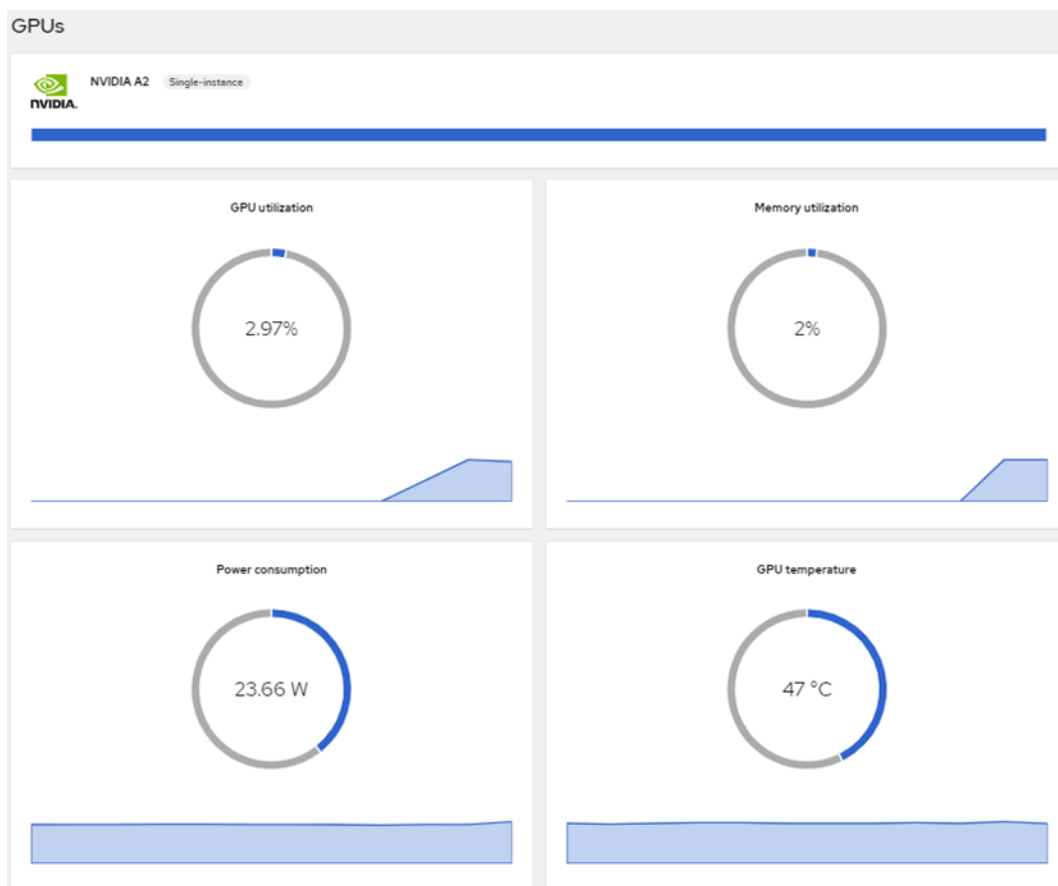


Figure 12. GPU metrics

Conclusion

Conversational AI applications can play a pivotal role in driving business success. Their versatility extends from internal processes through customer-facing interactions, making them indispensable tools for modern enterprises. Adopting speech AI solutions becomes paramount as businesses grapple with quality, efficiency, and resource constraints.

Implementing AI-driven conversational applications on the Dell APEX Cloud Platform for Red Hat OpenShift with NVIDIA Riva speech services offers a transformative path for remote customer service teams. These solutions empower call center staff to deliver exceptional service by shortening learning cycles and preserving institutional knowledge.

Support centers, concierge systems, self-service terminals, and other use cases across various domains can also reap the benefits. The solution's robust ecosystem, combining Dell Technologies hardware with NVIDIA GPUs, operates seamlessly within Red Hat OpenShift's enterprise-grade Kubernetes orchestration, leading to minimized investment and accelerated deployment.

In Red Hat OpenShift AI, developers and users find a sandbox for testing, fine-tuning, and deploying AI speech models. Meanwhile, applications interacting with NVIDIA Riva speech services scale effortlessly, thanks to Red Hat OpenShift's automatic deployment and management tools.

This solution empowers call center agents and managers with AI-driven tools, elevating processes, and the overall customer experience. Teams can have real-time transcription and live tagging during their customer calls, unlocking the power within their audio data and extracting actionable insights from it. This means that the customer will find faster clarity, assistance, or resolution with more accuracy, resulting in higher efficiency and higher customer satisfaction.

For more information, see the resources in the [References](#) section and contact your Dell Technologies sales representative. Although this paper does not provide specific guidance regarding sizing, Dell Technologies sales teams are primed to discuss and support your unique business needs.

References

Dell Technologies documentation

The following Dell Technologies documentation provides additional information related to this solution.

- [Dell Technologies Info Hub for AI Solutions](#)
- [Dell APEX Cloud Platform for Red Hat OpenShift](#)
- [Dell APEX Cloud Platform for Red Hat OpenShift Specification Sheet](#)

Red Hat documentation

The Red Hat documentation below contains more details about the OpenShift enterprise Kubernetes container platform.

- [Red Hat OpenShift Container Platform](#)
- [Red Hat OpenShift AI](#)
- [OpenShift Container Platform Registry](#)

NVIDIA documentation

The following NVIDIA resources expand on the NVIDIA Riva solution and its components.

- [NGC portal user guide](#)
- [NVIDIA Riva user guide](#)
- [NVIDIA Riva technical blog](#)
- [NVIDIA GPU Operator](#)
- [NVIDIA Riva deployment Helm chart](#)
- [Riva Contact user guide](#)
- [NVIDIA Riva Contact technical blog](#)
- [NVIDIA NeMo](#)
- [NVIDIA Triton Inference Server](#)
- [NVIDIA TensorRT](#)

Repositories and deployment guides

Refer to the following links for the main deployment guides and repositories needed to implement the solutions presented in this document.

- [NVIDIA Riva deployment on Red Hat OpenShift](#)
- [NVIDIA Riva tutorials GitHub repository](#)
- [NVIDIA Riva sample apps GitHub repository](#)