

5. Riva Server

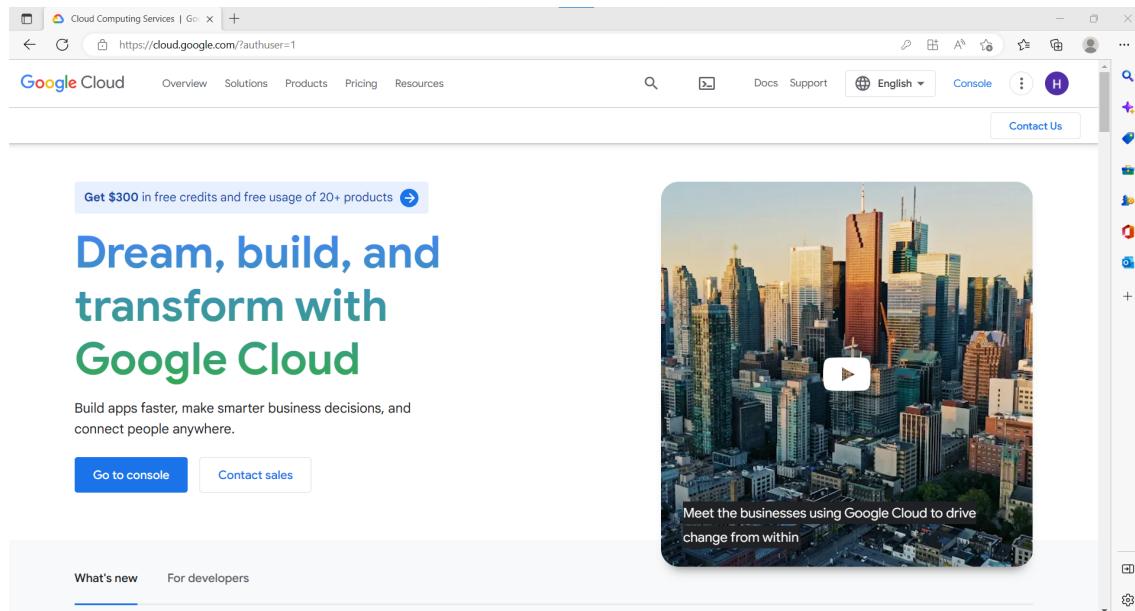
5.1 Background

There are many choices of Riva Server Deployment, some examples are AWS and Google Cloud Service. In this tutorial, we will be focusing on deploying Riva Server using the Compute Engine on Google Cloud Service.

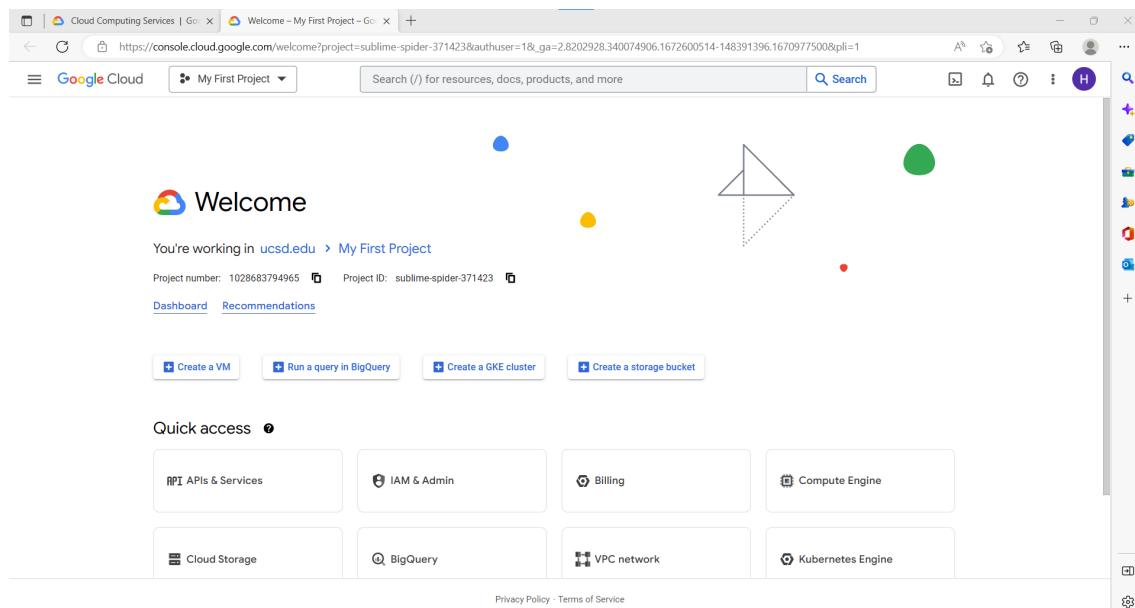
5.2 Server Setup

5.2.1 Google Cloud Server

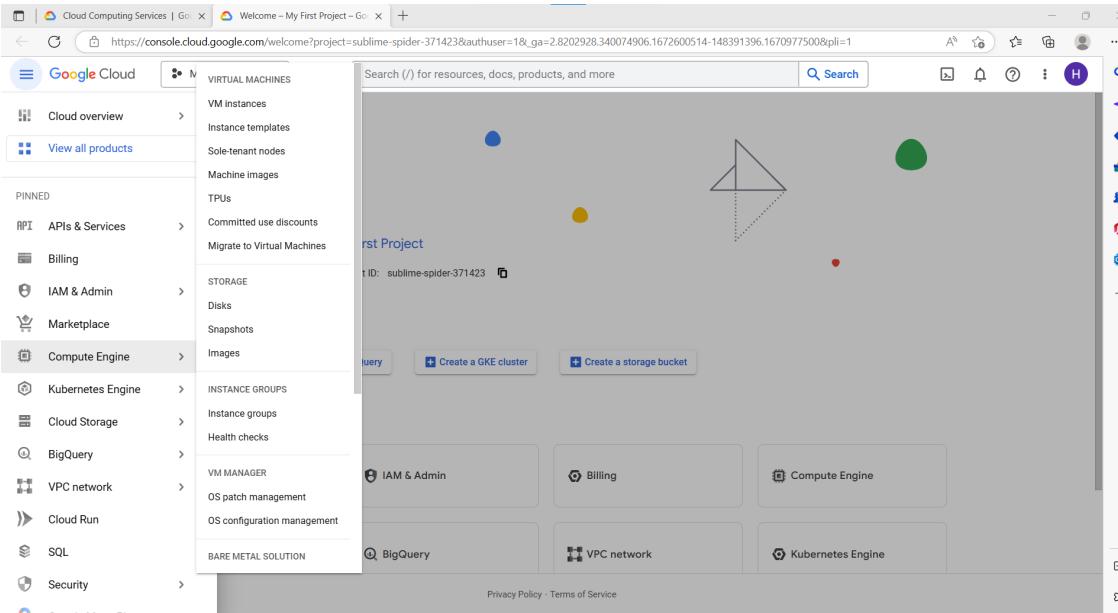
1. Go to <https://console.cloud.google.com/>.
2. Click on the **Go to console** symbol (if applicable).



3. Click on the **3-hyphen symbol**.

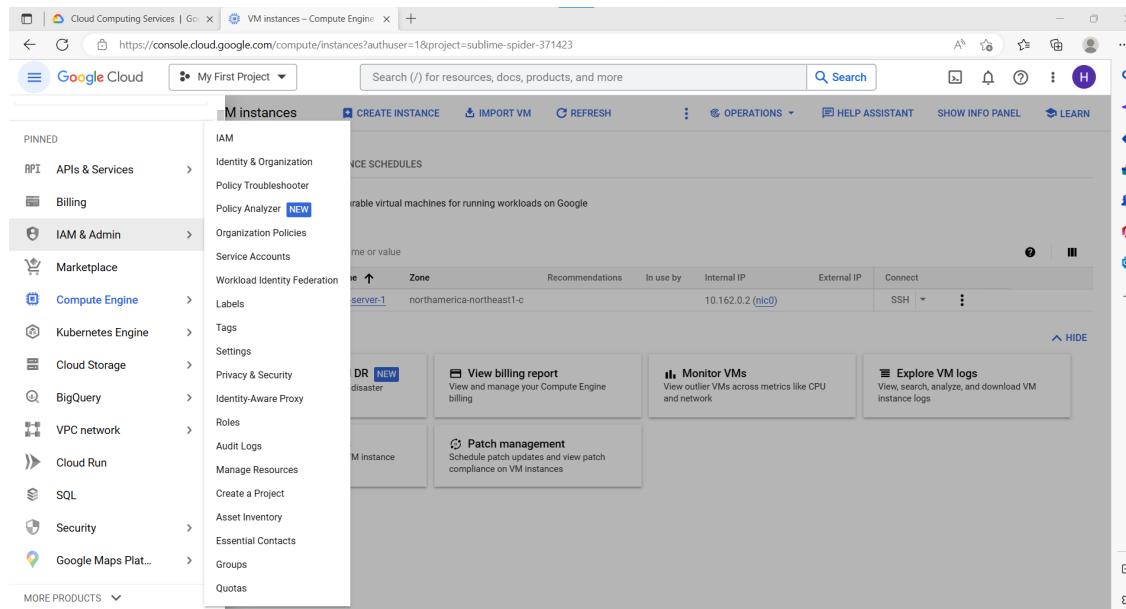


4. Go to the **Compute Engine** section.



5. Activate the Compute Engine.

6. Click on the *3-hyphen symbol* and then click on **IAM & Admin**.



7. Within **IAM & Admin**, click on **Quotas** (since we want to have GPU access).

Permissions for project "My First Project"

These permissions affect this project and all of its resources. [Learn more](#)

Type	Principal	Name	Role	Security Insights	Inheritance
Compute Engine default service account	1028683794965-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor	6167/6170 excess permissions	
Google APIs Service Agent	1028683794965@cloudservices.gserviceaccount.com	Google APIs Service Agent	Editor	6170/6170 excess permissions	
User	homiao@ucsd.edu	Hongbin Miao	Owner	6946/7095 excess permissions	

8. Inside of **Quotas**, filter the result using **GPUs (all regions)**.
9. You will then get 1 result as **Compute Engine API**; click on it and click on the **EDIT QUOTAS** on the top right side.

Quotas for project "My First Project"

QUOTAS INCREASE REQUESTS

Service	Quota	Dimensions (e.g. location)	Limit	Current usage percentage	Current usage	7 day peak usage percentage	7 day peak usage
Compute Engine API	GPUs (all regions)		1	0%	0	0%	0

10. Inside of the **EDIT QUOTAS** tab, set a new limit for your GPU quota (for my case: 1).

The screenshot shows the Google Cloud IAM & Admin Quotas interface for the project "My First Project". The left sidebar is titled "IAM & Admin" and includes options like Service Accounts, Workload Identity Federation, Labels, Tags, Settings, Privacy & Security, Identity-Aware Proxy, Roles, Audit Logs, Asset Inventory, Essential Contacts, Groups, and Quotas. The "Quotas" option is selected. The main area displays "Quotas for project 'My First Project'". It shows three categories: "Near the limit" (0), "Low usage" (8,291), and "All quotas" (8,527). A filter bar allows searching by service, quota, dimensions, limit, and current usage percentage. A modal window titled "Compute Engine API" is open, showing a quota for "GPUs (all regions)" with a current limit of 1. A text input field is set to "New limit *" and a "DONE" button is visible. A "SUBMIT REQUEST" button is located at the bottom of the main quota table.

11. After you **received email about GPU quota increase**, you can proceed with the steps below.
12. Go back to **Compute Engine** section and select on **CREATE INSTANCE** (the blue tab).
13. It is time to configured your VM instance, below will be my sample configuration.

The screenshot shows three stacked screenshots of the Google Cloud Compute Engine instance details page for 'riva-server-1'.

Screenshot 1: Basic information

Name	riva-server-1
Instance Id	3907374263379417148
Description	None
Type	Instance
Status	Stopped
Creation time	Dec 11, 2022, 3:25:08 PM UTC-08:00
Zone	northamerica-northeast1-c
Instance template	None
In use by	None
Reservations	Automatically choose
Labels	None
Tags	None
Deletion protection	Disabled
Confidential VM service	Disabled
Preserved state size	0 GB

Screenshot 2: Machine configuration

Machine type	n1-highmem-2
CPU platform	Unknown CPU Platform
Architecture	x86/64
vCPUs to core ratio	—

Screenshot 3: Network interfaces

Name	Network	Subnetwork	Primary internal IP address	Alias IP ranges	Stack Type	External IP address	Network
nic0	default	default	10.162.0.2		IPv4	Ephemeral	Premium

Storage

Name	Image	Interface type	Size (GB)	Device name	Type	Architecture	Encryption	Mode	Wb
riva-server-1	ubuntu-2004-focal-v20221206	SCSI	200	riva-server	Standard persistent disk	x86/64	Google-managed	Boot, read/write	Del

14. Configure your VM with similar setup, notice that you will need at least 1 GPU with the equal or greater performance to a Nvidia T4 to meet the configuration requirement of the Riva Server.
15. After the VM has been configured and started, install the Nvidia CUDA and related drivers by following the instruction on this link: <https://cloud.google.com/compute/docs/gpus/install-drivers-gpu>

5.2.2 Docker

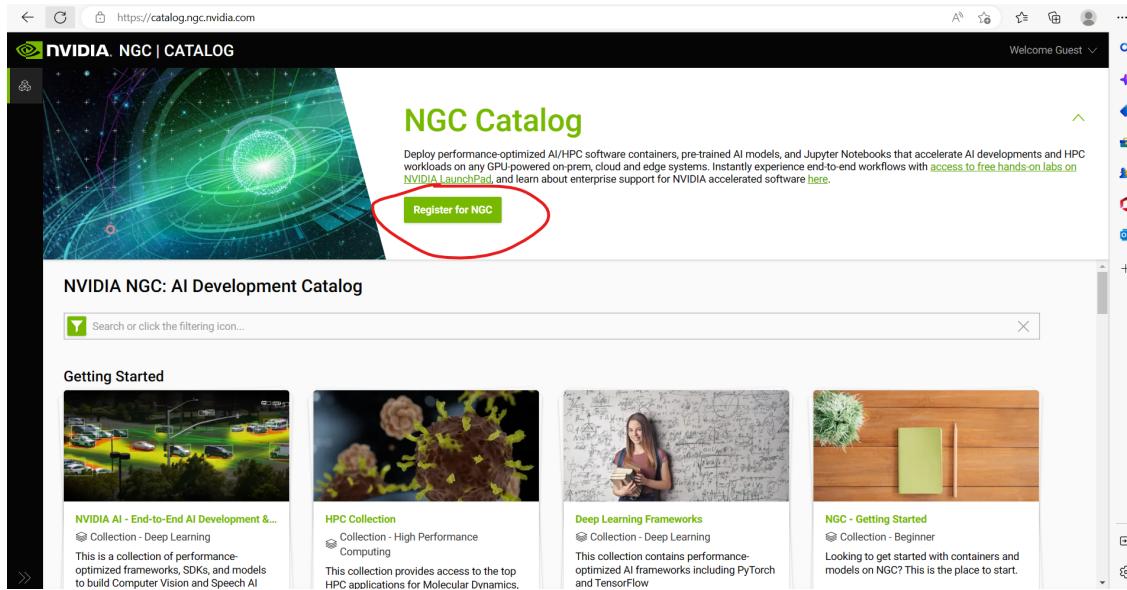
1. Install docker by following the official tutorial: <https://docs.docker.com/engine/install/ubuntu/>.

The screenshot shows a web browser displaying the Docker Engine installation guide for Ubuntu. The page title is "Install Docker Engine on Ubuntu". The left sidebar contains a navigation menu with sections like SLES, Ubuntu (which is selected), Binaries, Post-installation steps, Troubleshoot installation, Storage, Networking, Working with Docker Engine, Logging, Security, Advanced concepts, Deprecated features, Release notes, Docker Build, Docker Compose, Docker Hub, and Docker subscription. The main content area starts with a section titled "Prerequisites" which lists required Ubuntu versions (Kinetic, Jammy, Focal, Bionic) and architectures (x86_64, armhf, arm64, s390x). It also includes sections for "OS requirements", "Uninstall old versions", and instructions for removing Docker packages. On the right side, there is a "Page details" panel showing a 10-minute read time, edit and request change buttons, and a "Tags" section with links to requirements, apt, installation, ubuntu, install, uninstall, upgrade, and update. Below that is a "Contents" sidebar with links to Prerequisites, OS requirements, Uninstall old versions, Installation methods, and various sub-sections for installing and upgrading Docker.

2. Install the Nvidia Docker Container by following the instruction in this link for Ubuntu 20.04: <https://docs.nvidia.com/datacenter/cloud-native/container-toolkit/install-guide.html#docker>.

5.2.3 Nvidia NGC CLI

1. Register an account in <https://catalog.ngc.nvidia.com/>



2. SSH into the Google Cloud Server with root user access (type `sudo -i` in the SSH terminal), then type the command by the picture below:

```
wget -o ngccli_cat_linux.zip https://ngc.nvidia.com/downloads/ngccli_cat_linux.zip
unzip -o ngccli_cat_linux.zip
find ngc-cl/ -type f -exec md5sum {} + | LC_ALL=C sort | md5sum -c ngc-cl.md5
echo "export PATH=\"$PATH;$(pwd)/ngc-cl\\"" >> ~/.bash_profile && source ~/.bash_profile
ngc config set
```

5.2.4 Riva SDK

Prerequisites

- Set up Google Cloud Server
- Set up Docker
- Set up Nvidia NGC CLI

1. SSH into your Riva Server, and make sure you log in the terminal as root user by typing `sudo -i`.

```

https://ssh.cloud.google.com/v2/ssh/projects/sublime-spider-371423/zones/northamerica-northeast1-c/instances/riva-server-1?authuser=1&hl=en_US&projectNumber=1028683794965&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot4005Ena... □ X
https://ssh.cloud.google.com/v2/ssh/projects/sublime-spider-371423/zones/northamerica-northeast1-c/instances/riva-server-1?authuser=1&hl=en_US&projectNumber=1028683794965&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot4005Ena... A
SSH-in-browser
UPLOAD FILE DOWNLOAD FILE
System information as of Sun Jan 1 20:18:12 UTC 2023
* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/advantage

System load: 0.06 Processes: 169
Usage of /: 21.6% of 193.65GB Users logged in: 0
Memory usage: 5% IPv4 address for docker0: 172.17.0.1
Swap usage: 0% IPv4 address for ens5: 10.162.0.2

* Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s just raised the bar for easy, resilient and secure K8s cluster deployment.
  https://ubuntu.com/engage/secure-kubernetes-at-the-edge

2 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

Last login: Fri Dec 16 07:21:03 2022 from 35.235.242.34
homiao@riva-server-1:~$ sudo -i
root@riva-server-1:~# 
```

2. With the NGC CLI already set up, download the Riva Skills Quick Start through this link: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/riva/resources/riva_quickstart/files (will need to log in to your ngc account).
3. From step 2, you will get CLI download command similar to this: **ngc registry resource download-version"nvidia/riva/riva_quickstart:2.8.1"**, type the command into the terminal of the Riva Server to download the Riva Quick Start Toolkit.

```

https://ssh.cloud.google.com/v2/ssh/projects/sublime-spider-371423/zones/northamerica-northeast1-c/instances/riva-server-1?authuser=1&hl=en_US&projectNumber=1028683794965&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot4005Ena... □ X
https://ssh.cloud.google.com/v2/ssh/projects/sublime-spider-371423/zones/northamerica-northeast1-c/instances/riva-server-1?authuser=1&hl=en_US&projectNumber=1028683794965&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot4005Ena... A
SSH-in-browser
UPLOAD FILE DOWNLOAD FILE
root@riva-server-1:/home/Riva/riva_quickstart_v2.8.0# ls
aer_im_tools config.sh nemoRiva-2.8.0-py3-none-any.whl protos riva_clean.sh riva_init.sh riva_start.sh riva_start_client.sh riva_stop.sh
root@riva-server-1:/home/Riva/riva_quickstart_v2.8.0# 
```

4. After the toolkit has been downloaded, run **riva_init.sh** to configure the Riva Docker environment.
5. After the Riva Docker environment has been configured, run **riva_start.sh** to start the Riva Server.

```
Welcome to Ubuntu 20.04.5 LTS (GNU/Linux 5.15.0-1025-gcp x86_64)

 * Documentation: https://help.ubuntu.com
 * Management: https://landscape.canonical.com
 * Support: https://ubuntu.com/advantage

System information as of Sun Jan 1 20:18:12 UTC 2023

System load: 0.06 Processes: 169
Usage of /: 21.6% of 193.65GB Users logged in: 0
Memory usage: 5% IPv4 address for docker0: 172.17.0.1
Swap usage: 0% IPv4 address for ens5: 10.162.0.2

* Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
just raised the bar for easy, resilient and secure K8s cluster deployment.
  https://ubuntu.com/engage/secure-kubernetes-at-the-edge

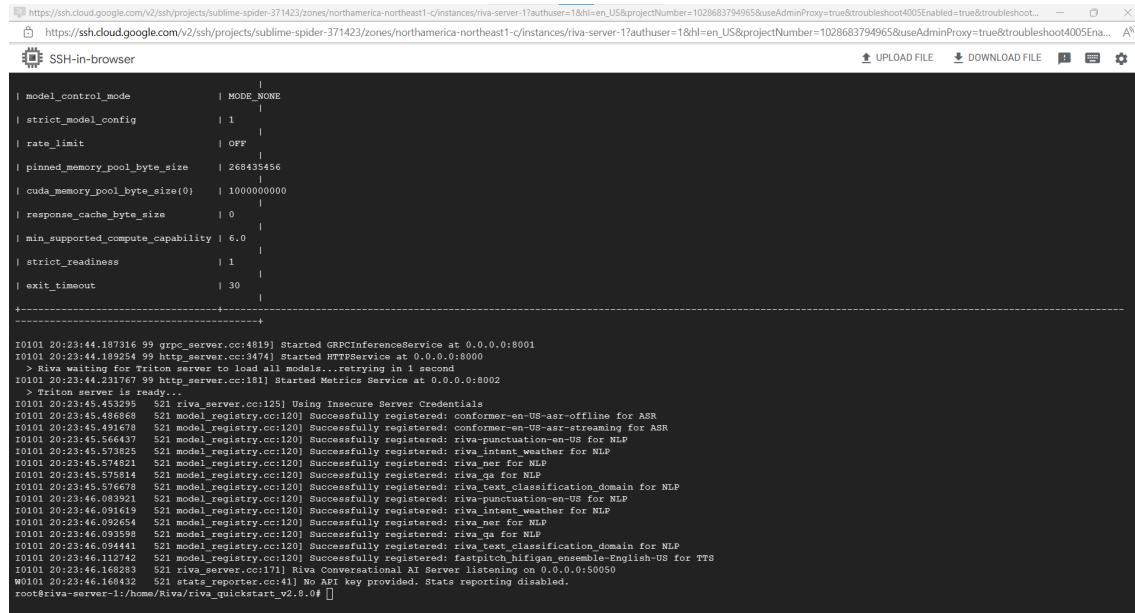
2 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

Last login: Fri Dec 16 07:21:03 2022 from 35.235.242.34
homiao@riva-server-1:~$ sudo -i
root@riva-server-1:~# 
```

6. The Riva Server will take a few seconds to start, once it starts, the Riva Server will be launched as a localhost at port 50050 as default. You can change the port number and relative config inside **config.sh**. (You will need to run **riva_init.sh** every time you changed the configuration inside of **config.sh**).

```
root@riva-server-1:/home/Riva/riva_quickstart_v2.8.0# ls
asr_lm_tools config.sh nemo_riva-2.8.0-py3-none-any.whl protos riva_clean.sh riva_init.sh riva_start.sh riva_start_client.sh riva_stop.sh
root@riva-server-1:/home/Riva/riva_quickstart_v2.8.0# riva_start.sh
riva_start.sh: command not found
root@riva-server-1:/home/Riva/riva_quickstart_v2.8.0# ./riva_start.sh
Starting Riva Speech Server... This might take several minutes depending on the number of models deployed.
7fbab8276a9891b36d559c4c53d60f0eeab1d81d90d3eabbcc61499
Waiting for Riva server to load all models... retrying in 10 seconds
Waiting for Riva server to load all models... retrying in 10 seconds
Waiting for Riva server to load all models... retrying in 10 seconds
Waiting for Riva server to load all models... retrying in 10 seconds
Waiting for Riva server to load all models... retrying in 10 seconds
Waiting for Riva server to load all models... retrying in 10 seconds
Waiting for Riva server to load all models... retrying in 10 seconds
Riva server is ready.
root@riva-server-1:/home/Riva/riva_quickstart_v2.8.0# 
```

7. To get the log message of the running Riva Server, type: docker logs riva-speech.



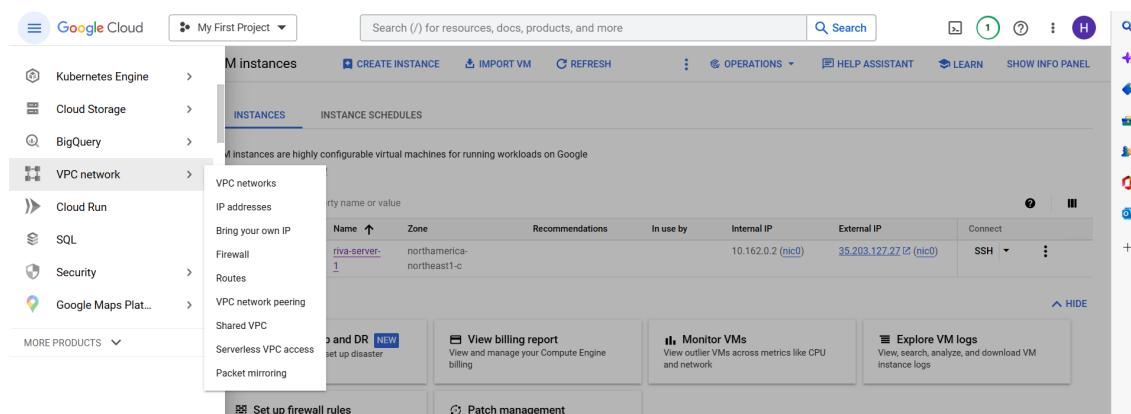
```

model_control_mode | MODE_NONE
strict_model_config | 1
rate_limit | OFF
pinned_memory_pool_byte_size | 268435456
cuda_memory_pool_byte_size(0) | 1000000000
response_cache_byte_size | 0
min_supported_compute_capability | 6.0
strict_readiness | 1
exit_timeout | 30
+-----+
T0101 20:23:44.187316 99 grpc_server.cc:4819] Started GRPCInferenceService at 0.0.0.0:8001
T0101 20:23:44.189254 99 http_server.cc:3474] Started HTTPService at 0.0.0.0:8000
> Riva waiting for Triton server to load all models...retrying in 1 second
T0101 20:23:44.231767 99 http_server.cc:151] Started Metric Service at 0.0.0.0:8002
Riva server is ready.
T0101 20:23:45.453025 521 riva_server.cc:125] Using Insecure Server Credentials
T0101 20:23:45.488688 521 model_registry.cc:120] Successfully registered: conformer-en-US-asr-offline for ASR
T0101 20:23:45.491678 521 model_registry.cc:120] Successfully registered: conformer-en-US-asr-streaming for ASR
T0101 20:23:45.566437 521 model_registry.cc:120] Successfully registered: riva-punctuation-en-US for NLP
T0101 20:23:45.573025 521 model_registry.cc:120] Successfully registered: riva_intent_weather for NLP
T0101 20:23:45.573025 521 model_registry.cc:120] Successfully registered: riva_ner for NLP
T0101 20:23:45.573025 521 model_registry.cc:120] Successfully registered: riva_text_classification_domain for NLP
T0101 20:23:45.576678 521 model_registry.cc:120] Successfully registered: riva_tctx_classification_domain for NLP
T0101 20:23:46.089392 521 model_registry.cc:120] Successfully registered: riva-punctuation-en-US for NLP
T0101 20:23:46.091619 521 model_registry.cc:120] Successfully registered: riva_intent_weather for NLP
T0101 20:23:46.092655 521 model_registry.cc:120] Successfully registered: riva_ner for NLP
T0101 20:23:46.093598 521 model_registry.cc:120] Successfully registered: riva_qa for NLP
T0101 20:23:46.094445 521 model_registry.cc:120] Successfully registered: riva_text_classification_domain for NLP
T0101 20:23:46.168283 521 riva_server.cc:171] Riva Conversational AI Server listening on 0.0.0.0:50050
W0101 20:23:46.169432 521 stats_reporter.cc:41] No API key provided. Stats reporting disabled.
root@riva-server-1:/home/Riva/riva_quickstart_v2.8.0# 

```

5.3 Access to the local host of the server

1. Make sure your Riva Server has already started.
2. Click on the *3-hyphen symbol*.
3. Click on **VPC network**.



4. Find the **Firewall** on the left hand side and click on it.
5. Find **CREATE FIREWALL RULE** and click on it.

VPC network Firewall

Get real-time analytics with Network Intelligence Center

Use Network Intelligence Center for comprehensive monitoring and troubleshooting. [Learn more](#)

- ✓ Visualize your network resources
- ✓ Diagnose and prevent connectivity issues
- ✓ View packet loss and latency metrics
- ✓ Keep your firewall rules strict and efficient

[GO TO NETWORK INTELLIGENCE CENTER](#) [REMIND ME LATER](#)

⚠ You don't have required permissions:

- * compute.organizations.listAssociations

VPC firewall rules

Firewall rules control incoming or outgoing traffic to an instance. By default, incoming traffic from outside your network is blocked. [Learn more](#)

Note: App Engine firewalls are managed in the [App Engine Firewall rules section](#).

SMTP port 25 disallowed in this project

[REFRESH](#) [CONFIGURE LOGS](#) [DELETE](#)

Filter Enter property name or value

6. We will set up a firewall rule, below is my setup.

VPC network Firewall rule details

[EDIT](#) [DELETE](#)

riva-client

Logs [view in Logs Explorer](#)

SHOW LOGS DETAILS

Network
default

Priority
1000

Direction
Ingress

Action on match
Allow

Source filters
IP ranges 0.0.0.0/0

Protocols and ports
tcp:50050

Enforcement
Enabled

Insights
None

Google Cloud My First Project Search (/) for resources, docs, products, and more [Search](#)

VPC network Firewall

Insights
None

Hit count monitoring [sum RESENT 0h0m0s \(1 hour\)](#) 6 hours 12 hours 1 day 2 days 4 days 7 days 14 days 30 days 6 weeks 2 months 3 months 6 months 12 months 24 months

No data is available for the selected time frame.

Applicable to instances

The following table shows only the VM instances that you have permission to view. It also does not show any App Engine flexible environment instances.

Name	Subnetwork	Internal IP ranges	External IP ranges	Tags	Service accounts	Project	Labels	Next
riva-server-1	default	10.162.0.2	35.203.127.27	http-ser...	1028683794965-compute@developer.g...	sublime-spider-371423		

EQUIVALENT REST

7. After the firewall rule has been setup, you can access the local host server through the following syntax: external IP of your Server:Port of your Riva Server. For example, if the external IP of my Riva Server is: "12.34.56.789" and the port of my Riva Server is "50050", then the link to my Riva Server will be "12.34.56.789:50050". Notice that if you do not have a static ip address, the external IP address of your server will change everytime you restart your server.
8. The above link is useful to this python file: <https://github.com/zslrmhb/Omniverse-Virtual-Assisstant/blob/main/config.py>

6. Audio2Face

6.1 Background

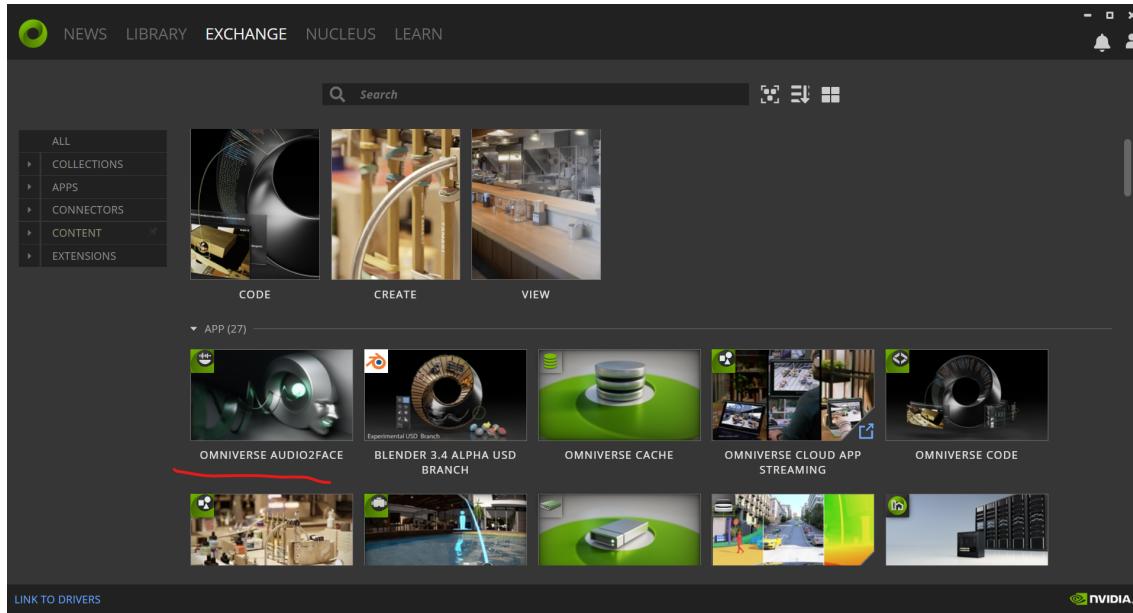
We will be using Audio2Face to leverage the power of Riva Server and animate the virtual assistant.

6.2 Audio2Face Setup

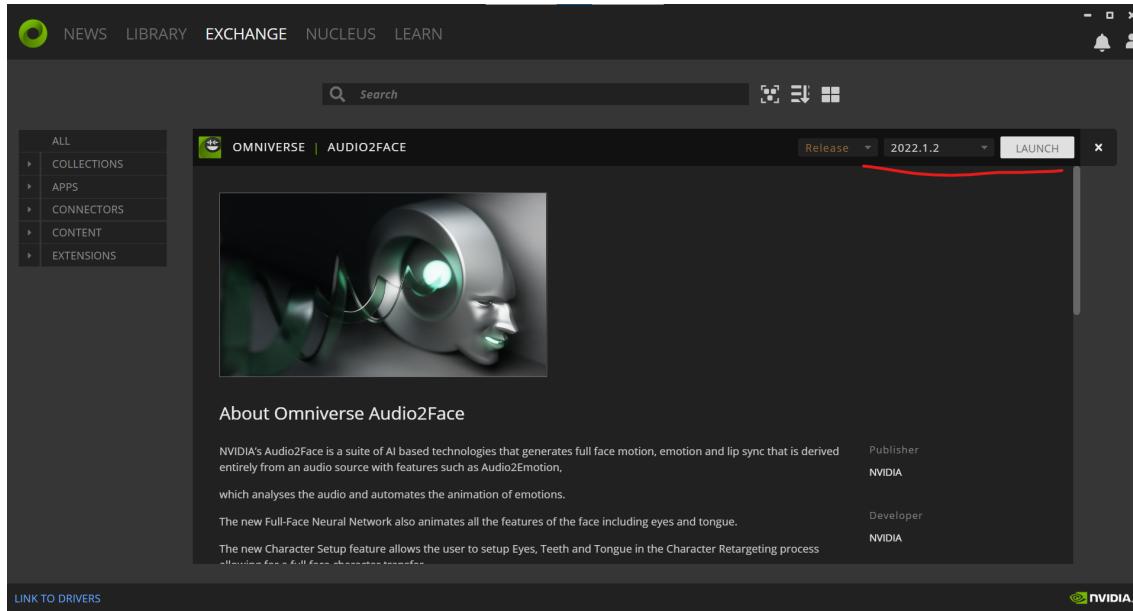
6.2.1 Download Audio2Face

Prerequisites

- Windows 10, Version 1903 and Above (At the time of this documentation, only Windows is the supported platform.)
 - Have Nvidia Omniverse installed. If not, install through the following link: <https://www.nvidia.com/en-us/omniverse/download/>
1. Open the Omniverse Launcher.
 2. Go to the **EXCHANGE** section.
 3. Find **OMNIVERSE AUDIO2FACE**.

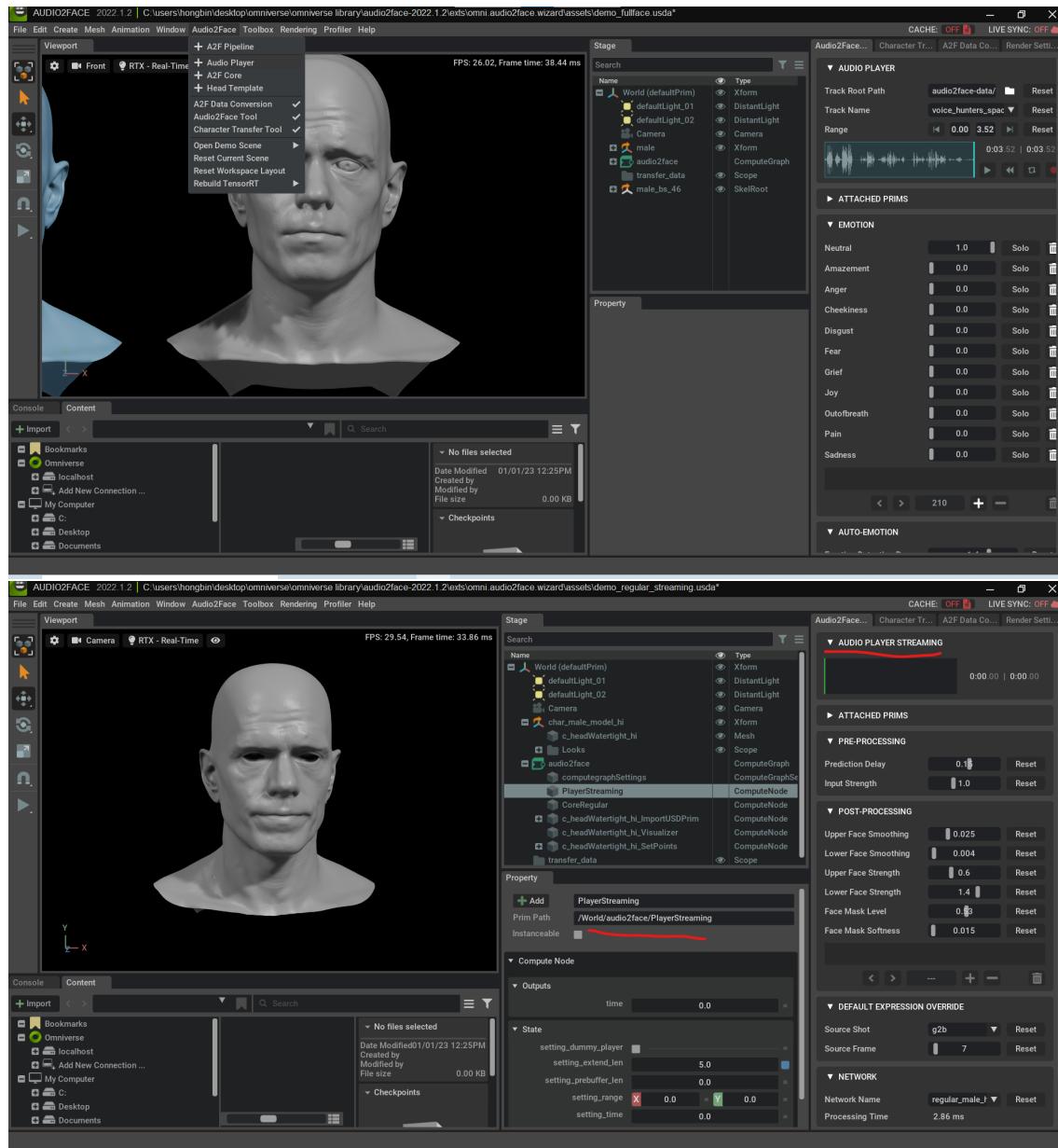


4. Select a version to download (I choose the latest one)



6.2.2 Integrate Audio2Face with Riva SDK

1. Launch the Riva Server.
2. Launch the Audio2Face and choose either the **Regular Core + Streaming Player** or the **Full Core + Streaming Player** Demo Scene.



3. Clone the following repository: <https://github.com/zslrmhb/Omniverse-Virtual-Assisstant.git>
4. Go to the config.py of the cloned repo and type the URI in the following format: external IP address of your Riva Server:Port of your River Server. For example, if the external IP address is 12.34.56.789 and the port is 50050, then the URI in the config.py will be 12.34.56.789:50050.

Figure 6.1: External IP address example

Filter Enter property name or value						
Status	Name	Zone	Recommendations	In use by	Internal IP	External IP
<input checked="" type="checkbox"/>	riva-server-1	northamerica-northeast1-c			10.162.0.2 (nic0)	35.203.127.27 (nic0)

5. Run main.py from the cloned repository.

