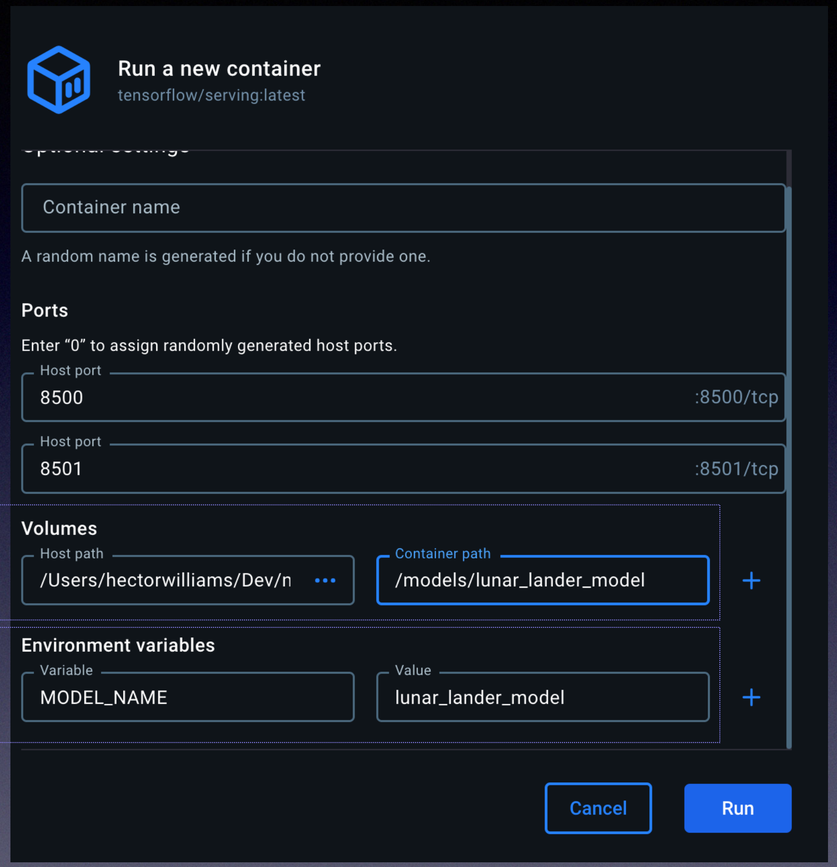
Docker


Run a new container
tensorflow/serving:latest
Container name
A random name is generated if you do not provide one.
Ports
Enter “0” to assign randomly generated host ports.
Host port
8500
:8500/tcp
Host port
8501
:8501/tcp
Volumes
Host path
/Users/hectorwilliams/Dev/n
Container path
/models/lunar_lander_model
Environment variables
Variable
MODEL_NAME
Value
lunar_lander_model
Cancel
Run

Google Cloud Service

Prediction Service

IAM Roles (Legacy Basic Roles)

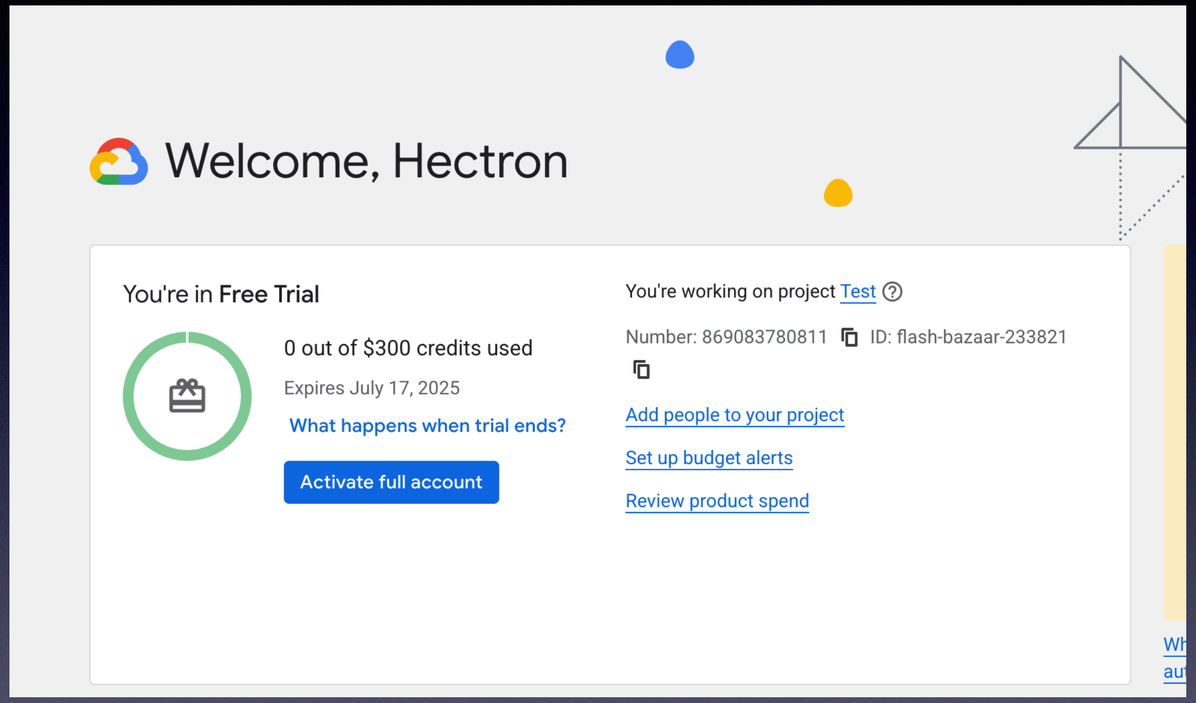
Browser	Read access to browse hierarchy for project (folder, org, allow policies). No access to resources in project
Owner	Handles Billing Manages roles and permissions for project(s) Managing tag binding for Compute Engine
Editor	Modify data state for most Google Cloud services
Viewer	Read-Only, not affecting state of data

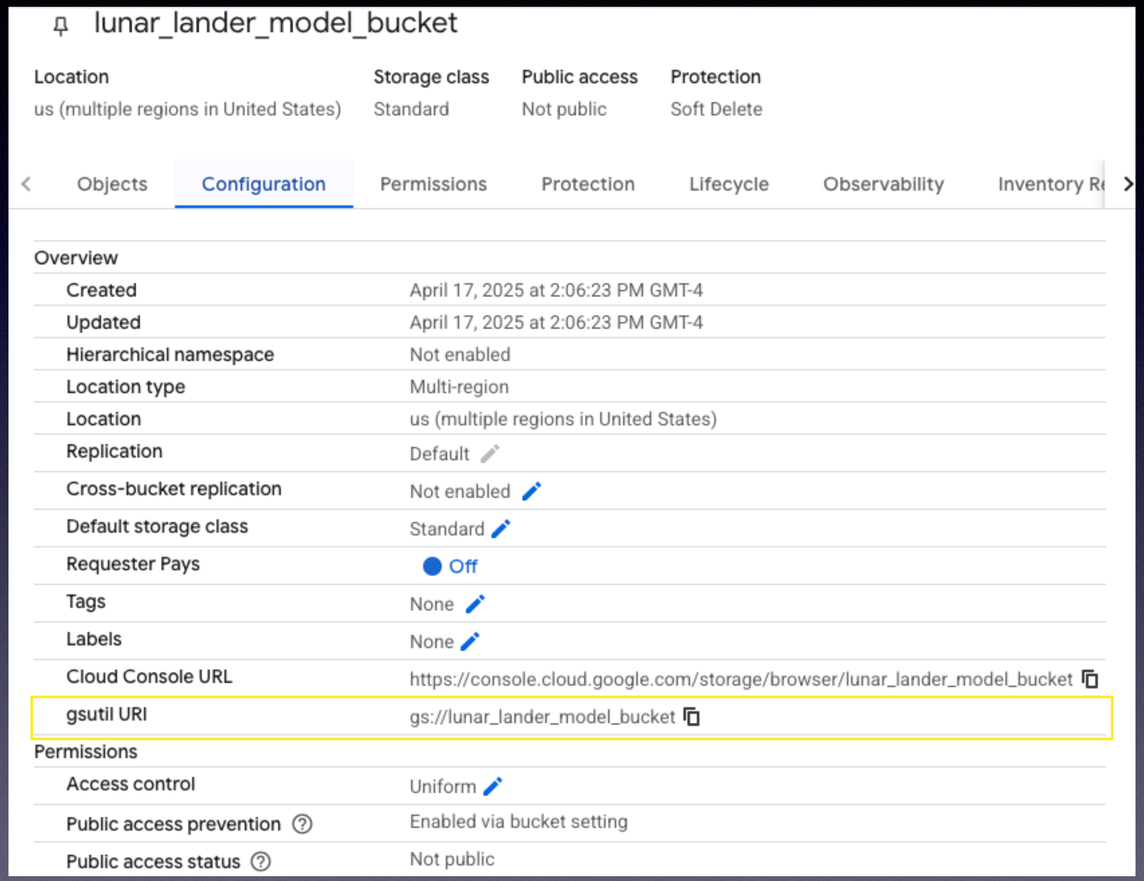
IAM Roles

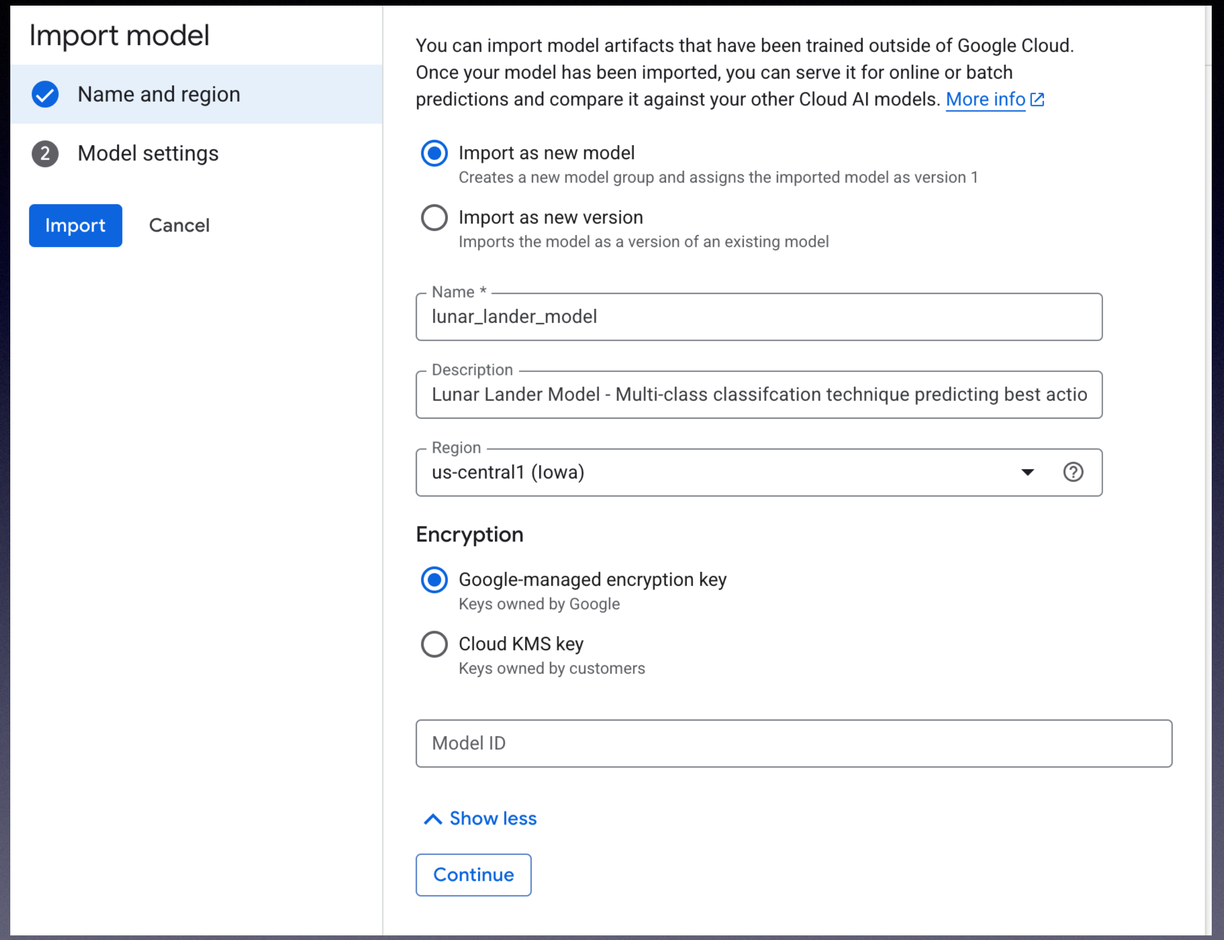
Admin	Handles Billing Manages roles and permissions for project(s) Managing tag binding for Compute Engine
Writer	Modify data state for most Google Cloud services
Reader	Read-Only, not affecting state of data

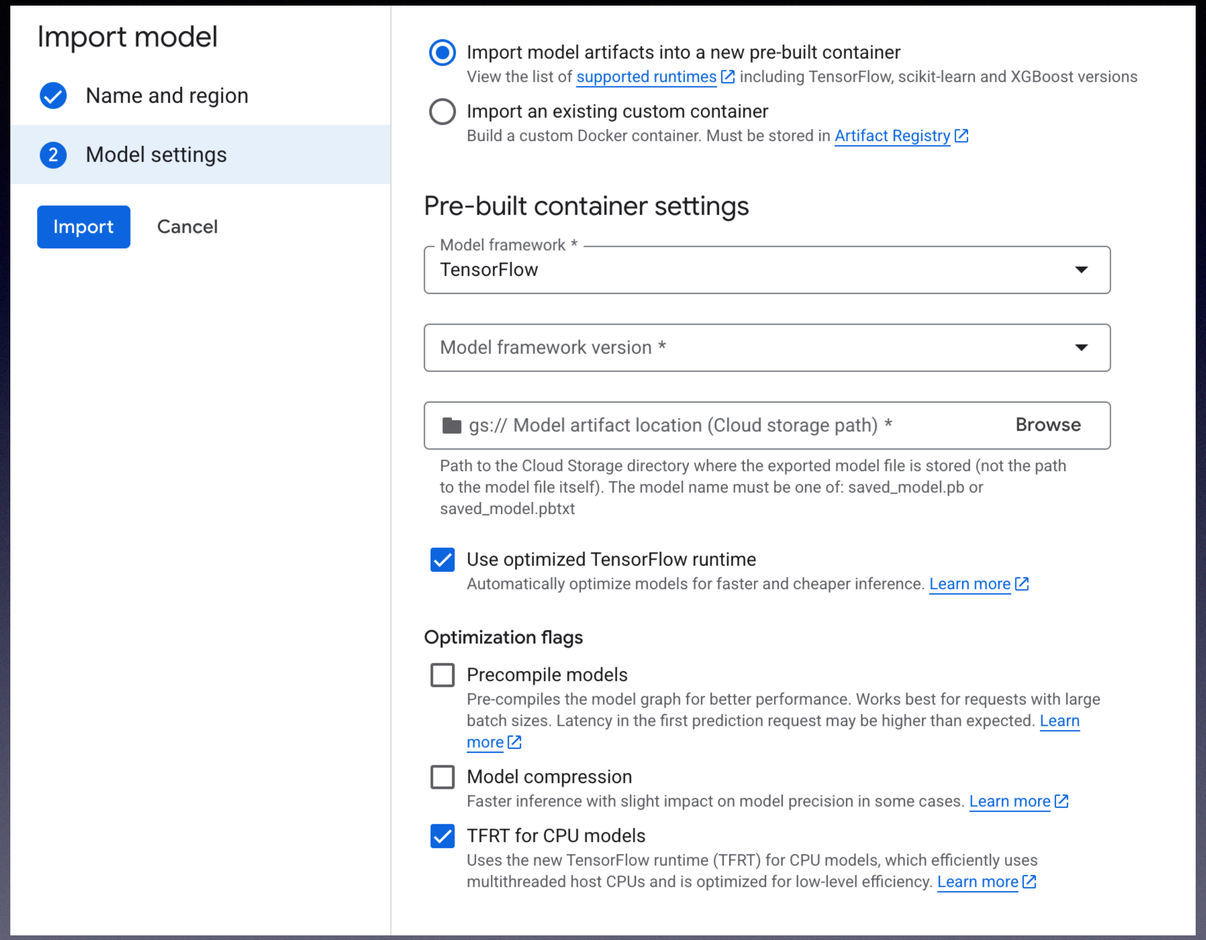
Import Model

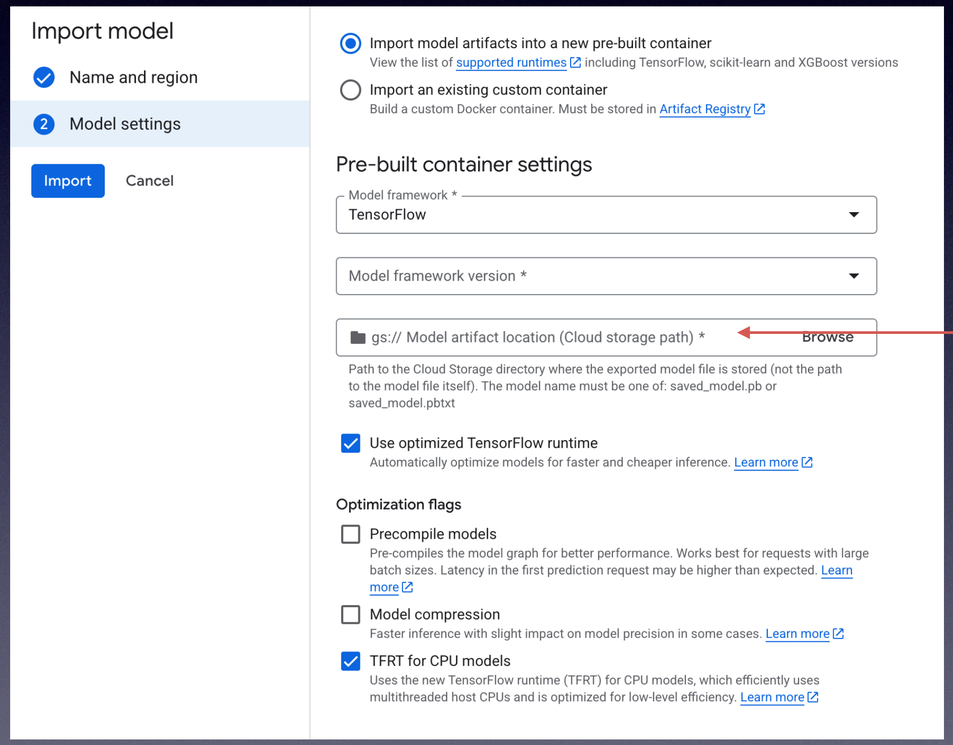
Prediction Service (deployment)

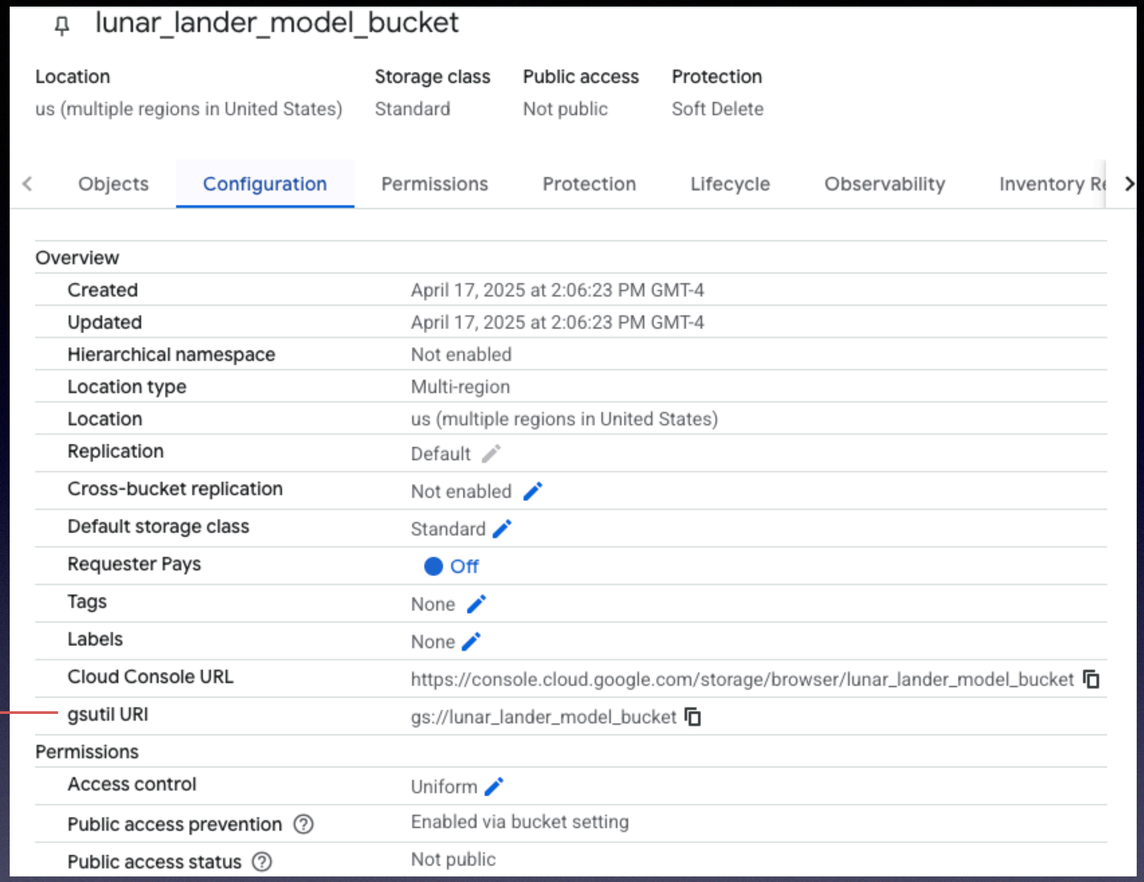

Welcome, Hectron
You're in Free Trial
0 out of $300 credits used
Expires July 17, 2025
What happens when trial ends?
Activate full account
You're working on project Test
Number: 869083780811
ID: flash-bazaar-233821
Add people to your project
Set up budget alerts
Review product spend

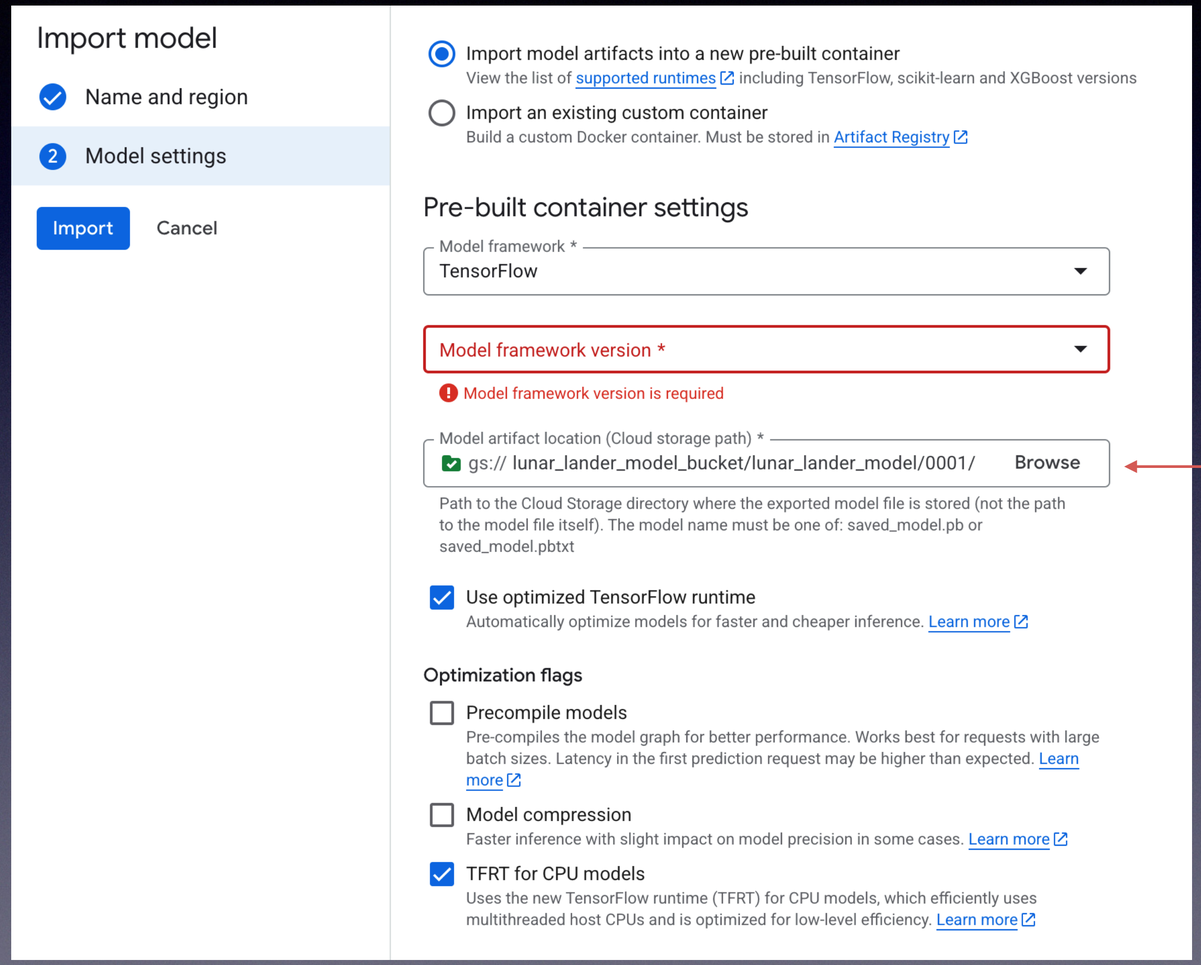

lunar_lander_model_bucket
Location
us (multiple regions in United States)
Storage class
Standard
Public access
Not public
Protection
Soft Delete
Objects
Configuration
Permissions
Protection
Lifecycle
Observability
Inventory R
Overview
Created
April 17, 2025 at 2:06:23 PM GMT-4
Updated
April 17, 2025 at 2:06:23 PM GMT-4
Hierarchical namespace
Not enabled
Location type
Multi-region
Location
us (multiple regions in United States)
Replication
Default
Cross-bucket replication
Not enabled
Default storage class
Standard
Requester Pays
Off
Tags
None
Labels
None
Cloud Console URL
https://console.cloud.google.com/storage/browser/lunar_lander_model_bucket
gsutil URI
gs://lunar_lander_model_bucket
Permissions
Access control
Uniform
Public access prevention
Enabled via bucket setting
Public access status
Not public

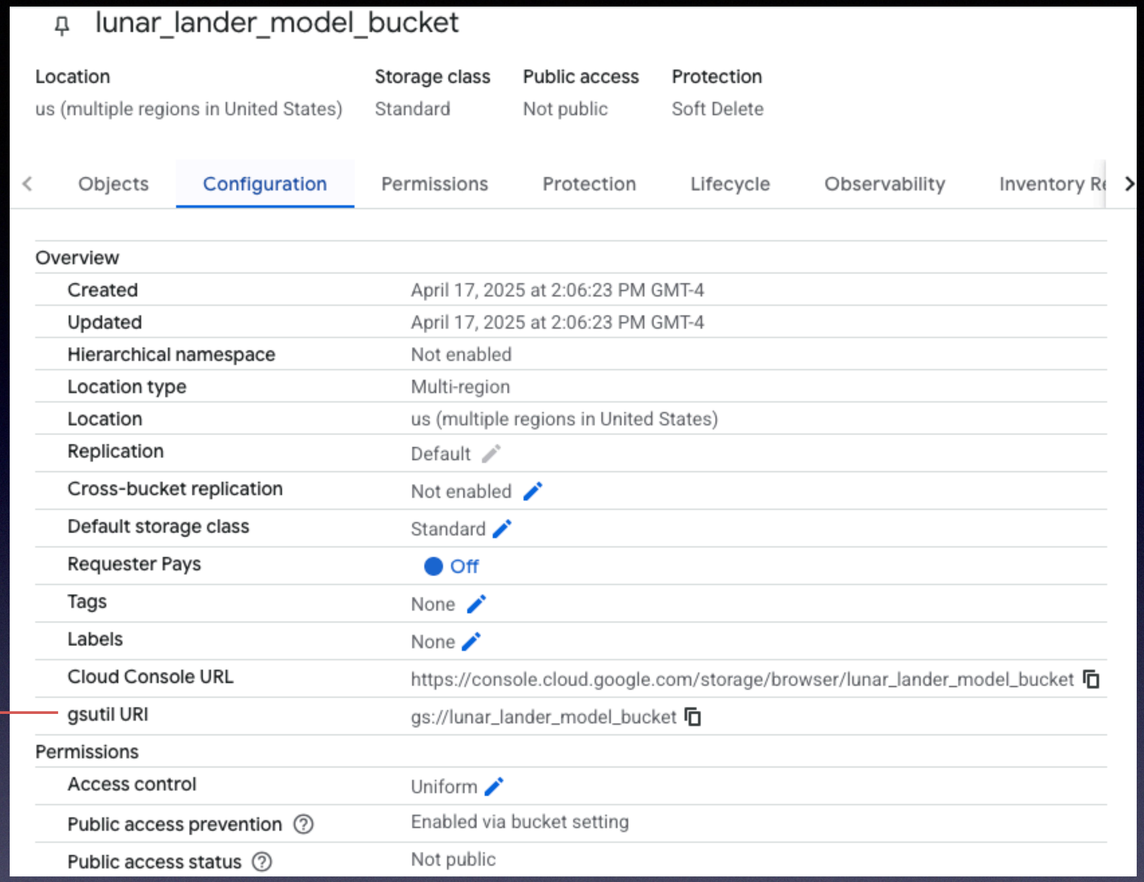

Import model
Name and region
2
Model settings
Import
Cancel
You can import model artifacts that have been trained outside of Google Cloud. Once your model has been imported, you can serve it for online or batch predictions and compare it against your other Cloud AI models. More info
Import as new model
Creates a new model group and assigns the imported model as version 1
Import as new version
Imports the model as a version of an existing model
Name *
lunar_lander_model
Description
Lunar Lander Model - Multi-class classifcation technique predicting best actio
Region
us-central1 (Iowa)
Encryption
Google-managed encryption key
Keys owned by Google
Cloud KMS key
Keys owned by customers
Model ID
Show less
Continue


Import model
Name and region
2
Model settings
Import
Cancel
Import model artifacts into a new pre-built container
View the list of supported runtimes including TensorFlow, scikit-learn and XGBoost versions
Import an existing custom container
Build a custom Docker container. Must be stored in Artifact Registry
Pre-built container settings
Model framework *
TensorFlow
Model framework version *
gs:// Model artifact location (Cloud storage path) *
Browse
Path to the Cloud Storage directory where the exported model file is stored (not the path to the model file itself). The model name must be one of: saved_model.pb or saved_model.pbtxt
Use optimized TensorFlow runtime
Automatically optimize models for faster and cheaper inference. Learn more
Optimization flags
Precompile models
Pre-compiles the model graph for better performance. Works best for requests with large batch sizes. Latency in the first prediction request may be higher than expected. Learn more
Model compression
Faster inference with slight impact on model precision in some cases. Learn more
TFRT for CPU models
Uses the new TensorFlow runtime (TFRT) for CPU models, which efficiently uses multithreaded host CPUs and is optimized for low-level efficiency. Learn more


Import model
Name and region
2 Model settings
Import
Cancel
Import model artifacts into a new pre-built container
View the list of supported runtimes including TensorFlow, scikit-learn and XGBoost versions
Import an existing custom container
Build a custom Docker container. Must be stored in Artifact Registry
Pre-built container settings
Model framework *
TensorFlow
Model framework version *
gs:// Model artifact location (Cloud storage path) *
Browse
Path to the Cloud Storage directory where the exported model file is stored (not the path to the model file itself). The model name must be one of: saved_model.pb or saved_model.pbtxt
Use optimized TensorFlow runtime
Automatically optimize models for faster and cheaper inference. Learn more
Optimization flags
Precompile models
Pre-compiles the model graph for better performance. Works best for requests with large batch sizes. Latency in the first prediction request may be higher than expected. Learn more
Model compression
Faster inference with slight impact on model precision in some cases. Learn more
TFRT for CPU models
Uses the new TensorFlow runtime (TFRT) for CPU models, which efficiently uses multithreaded host CPUs and is optimized for low-level efficiency. Learn more


lunar_lander_model_bucket
Location
us (multiple regions in United States)
Storage class
Standard
Public access
Not public
Protection
Soft Delete
Objects
Configuration
Permissions
Protection
Lifecycle
Observability
Overview
Created
April 17, 2025 at 2:06:23 PM GMT-4
Updated
April 17, 2025 at 2:06:23 PM GMT-4
Hierarchical namespace
Not enabled
Location type
Multi-region
Location
us (multiple regions in United States)
Replication
Default
Cross-bucket replication
Not enabled
Default storage class
Standard
Requester Pays
Off
Tags
None
Labels
None
Cloud Console URL
https://console.cloud.google.com/storage/browser/lunar_lander_model_bucket
gsutil URI
gs://lunar_lander_model_bucket
Permissions
Access control
Uniform
Public access prevention
Enabled via bucket setting
Public access status
Not public


Import model
Name and region
2 Model settings
Import
Cancel
Import model artifacts into a new pre-built container
View the list of supported runtimes including TensorFlow, scikit-learn and XGBoost versions
Import an existing custom container
Build a custom Docker container. Must be stored in Artifact Registry
Pre-built container settings
Model framework *
TensorFlow
Model framework version *
Model framework version is required
Model artifact location (Cloud storage path) *
gs:// lunar_lander_model_bucket/lunar_lander_model/0001/
Browse
Path to the Cloud Storage directory where the exported model file is stored (not the path to the model file itself). The model name must be one of: saved_model.pb or saved_model.pbtxt
Use optimized TensorFlow runtime
Automatically optimize models for faster and cheaper inference. Learn more
Optimization flags
Precompile models
Pre-compiles the model graph for better performance. Works best for requests with large batch sizes. Latency in the first prediction request may be higher than expected. Learn more
Model compression
Faster inference with slight impact on model precision in some cases. Learn more
TFRT for CPU models
Uses the new TensorFlow runtime (TFRT) for CPU models, which efficiently uses multithreaded host CPUs and is optimized for low-level efficiency. Learn more


lunar_lander_model_bucket
Location
us (multiple regions in United States)
Storage class
Standard
Public access
Not public
Protection
Soft Delete
Objects
Configuration
Permissions
Protection
Lifecycle
Observability
Overview
Created
April 17, 2025 at 2:06:23 PM GMT-4
Updated
April 17, 2025 at 2:06:23 PM GMT-4
Hierarchical namespace
Not enabled
Location type
Multi-region
Location
us (multiple regions in United States)
Replication
Default
Cross-bucket replication
Not enabled
Default storage class
Standard
Requester Pays
Off
Tags
None
Labels
None
Cloud Console URL
https://console.cloud.google.com/storage/browser/lunar_lander_model_bucket
gsutil URI
gs://lunar_lander_model_bucket
Permissions
Access control
Uniform
Public access prevention
Enabled via bucket setting
Public access status
Not public

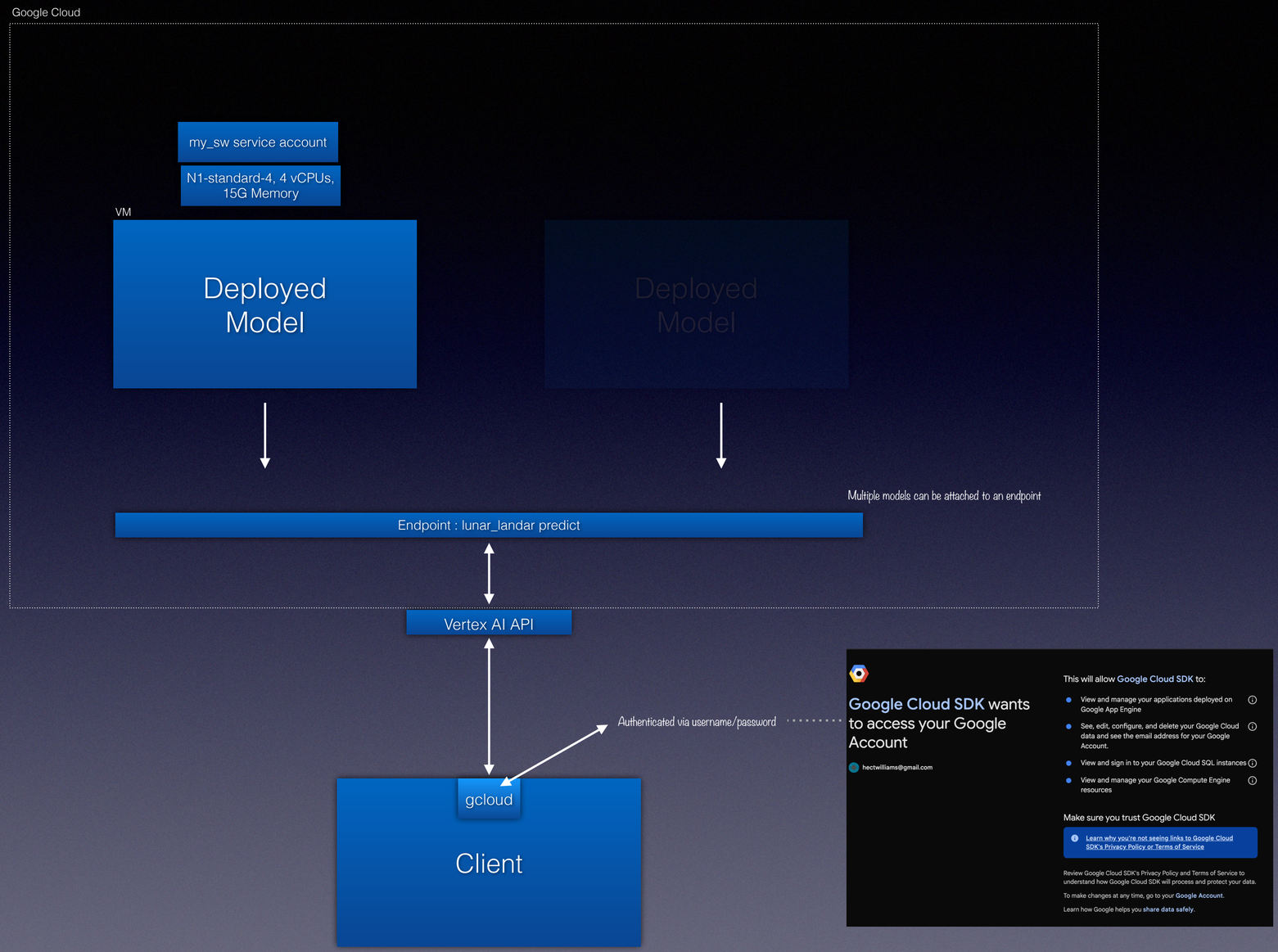
Pip install

google-cloud-storage

google-cloud-aiplatform

Install gcloud

```
tar -vxf Downloads/google-cloud-cli-darwin-x86_64.tar.gz --directory ./
```


Google Cloud
my_sw service account
N1-standard-4, 4 vCPUs, 15G Memory
VM
Deployed Model
Deployed Model
Multiple models can be attached to an endpoint
Endpoint : lunar_landar predict
Vertex AI API
Authenticated via username/password
gcloud
Client

Google Cloud SDK wants to access your Google Account
hectwilliams@gmail.com
This will allow Google Cloud SDK to:
View and manage your applications deployed on Google App Engine
See, edit, configure, and delete your Google Cloud data and see the email address for your Google Account.
View and sign in to your Google Cloud SQL instances
View and manage your Google Compute Engine resources
Make sure you trust Google Cloud SDK
Learn why you're not seeing links to Google Cloud SDK's Privacy Policy or Terms of Service
Review Google Cloud SDK's Privacy Policy and Terms of Service to understand how Google Cloud SDK will process and protect your data.
To make changes at any time, go to your Google Account.
Learn how Google helps you share data safely.

Google CLI

Request

```
hectorwilliams@Hectors-Air ~ % ./google-cloud-sdk/bin/gcloud ai endpoints predict 1941215266629222400  --region=us-central1 --json-request=request.json

Using endpoint [https://us-central1-prediction-aiplatform.googleapis.com/]
```

Response

```
[[0.312848389, 0.236094132, 0.21771571, 0.233341753]]
```

CURL

request.json

```
{ "instances": [[ 0.30, 0.20, 0.20, 0.10, 0.20, 0.10, 0.20, 0.110]]  }
```

Request

```
% curl \
-X POST \
-H "Authorization: Bearer $(./google-cloud-sdk/bin/gcloud auth print-access-token)" \
-H "Content-Type: application/json" \
"https://us-central1-aiplatform.googleapis.com/v1/projects/${PROJECT_ID}/locations/us-central1/endpoints/${ENDPOINT_ID}:predict" \
-d "@${INPUT_DATA_FILE}"
```

Response


```
{
  "predictions": [
    [
      0.312848389,
      0.236094132,
      0.21771571,
      0.233341753
    ]
  ],
  "deployedModelId": "8489429333616623616",
  "model": "projects/869083780811/locations/us-central1/models/814655652809932800",
  "modelDisplayName": "lunar_lander_model",
  "modelVersionId": "1"
}
```

Resource

Project Principals

Owner

Email: hectwilliams@gmail.com

Resource

Service Agents

Compute Engine

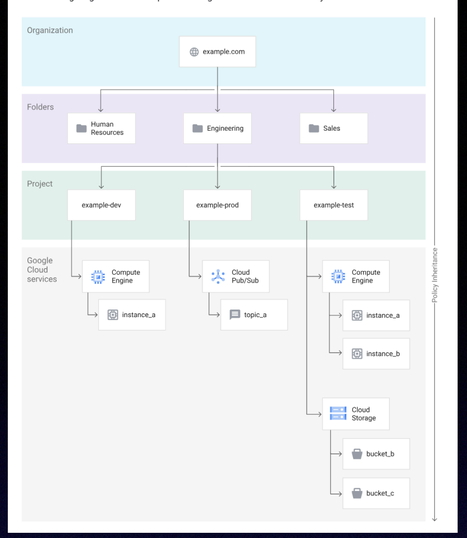
Email: my-sw-821@flash-bazaar-233821.iam.gserviceaccount.com

Compute Engine

Email: 869083780811-compute@developer.gserviceaccount.com

Keys

Processes - (Resource Performs hidden processes)



Resource

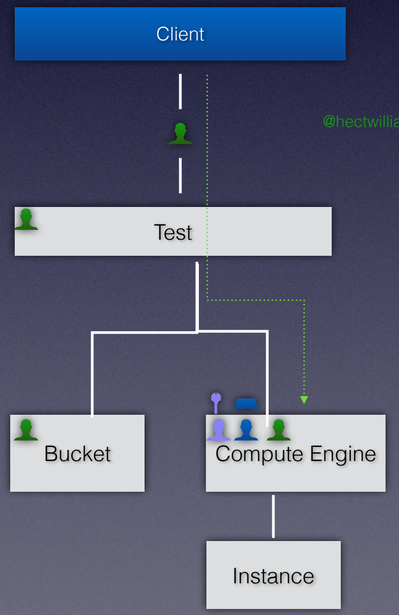
Project Principals

Owner, Storage Viewer, Storage

Email: hectwilliams@gmail.com

Reader

Email: my-sw-821@flash-bazaar-233821.iam.gserviceaccount.com



@hectwilliams principal account permits access to query instance to previous two example slides

AI Vertex SDK

```python
from google.cloud import aiplatform
import os

PROJECT_ID = '869083780811'
ENDPOINT_ID = '1941215266629222400'
LOCATION_ID = 'us-central1'

os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = os.path.join(os.getcwd(), 'cred.json')

def endpoint_predict_sample( project: str, location: str, instances: list, endpoint: str):
    aiplatform.init(project=project, location=location, staging_bucket='gs://my_staging_bucket')

    endpoint = aiplatform.Endpoint(endpoint)

    prediction = endpoint.predict(instances=instances)
    print(prediction)
    return prediction
```

Request

```python
instances =  [[ 0.30, 0.20, 0.20, 0.10, 0.20, 0.10, 0.20, 0.110], [0.30, 0.20, 0.20, 0.10, 0.20, 0.10, 0.20, 0.110]]

if __name__ == "__main__":
    ret = endpoint_predict_sample(project=PROJECT_ID, location= LOCATION_ID, instances=instances, endpoint=ENDPOINT_ID)
    print(ret)
```

Response

Prediction(predictions=[[0.312848389, 0.236094147, 0.21771571, 0.233341753], [0.312848389, 0.236094147, 0.21771571, 0.233341753]], deployed_model_id='8489429333616623616', metadata=None, model_version_id='1', model_resource_name='projects/869083780811/locations/us-central1/models/814655652809932800', explanations=None)

Prediction(predictions=[[0.312848389, 0.236094147, 0.21771571, 0.233341753], [0.312848389, 0.236094147, 0.21771571, 0.233341753]], deployed_model_id='8489429333616623616', metadata=None, model_version_id='1', model_resource_name='projects/869083780811/locations/us-central1/models/814655652809932800', explanations=None)

Resource

Project Principals

Owner

Email: hectwilliams@gmail.com

Resource

Service Agents

Compute Engine

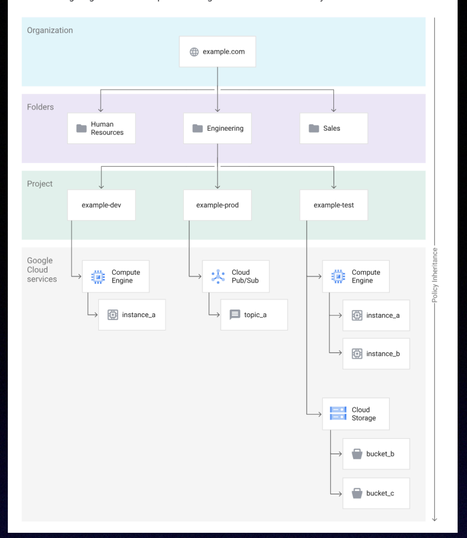
Email: my-sw-821@flash-bazaar-233821.iam.gserviceaccount.com

Compute Engine

Email: 869083780811-compute@developer.gserviceaccount.com

Keys

Processes - (Resource Performs hidden processes)



Resource

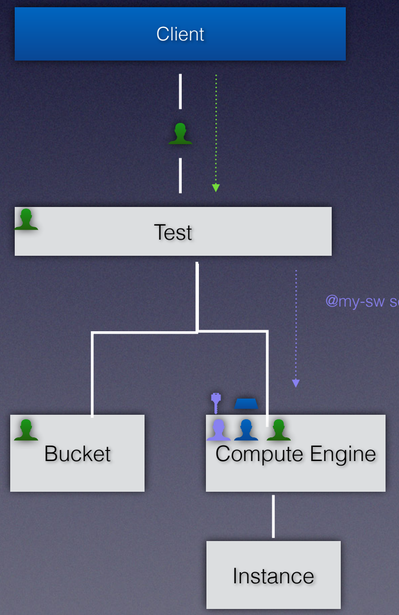
Project Principals

Owner, Storage Viewer, Storage

Email: hectwilliams@gmail.com

Reader

Email: my-sw-821@flash-bazaar-233821.iam.gserviceaccount.com



@my-sw service agent that was elevated to principal, permits access to query instance because the private key required for authentication was generated by my-sw agent

Resource

Owner

Compute Engine Service Agent

AI Platform Custom Code Service Agent

AI Platform Service Agent

Bucket Principals

Email: hectwilliams@gmail.com

Email: service-869083780811@compute-system.iam.gserviceaccount.com

Email: service-869083780811@gcp-sa-aiplatform-cc.iam.gserviceaccount.com

Email service-869083780811@gcp-sa-aiplatform.iam.gserviceaccount.com

Service Agents

Email: service-869083780811@compute-system.iam.gserviceaccount.com

Email: service-869083780811@gcp-sa-aiplatform-cc.iam.gserviceaccount.com

Email service-869083780811@gcp-sa-aiplatform.iam.gserviceaccount.com

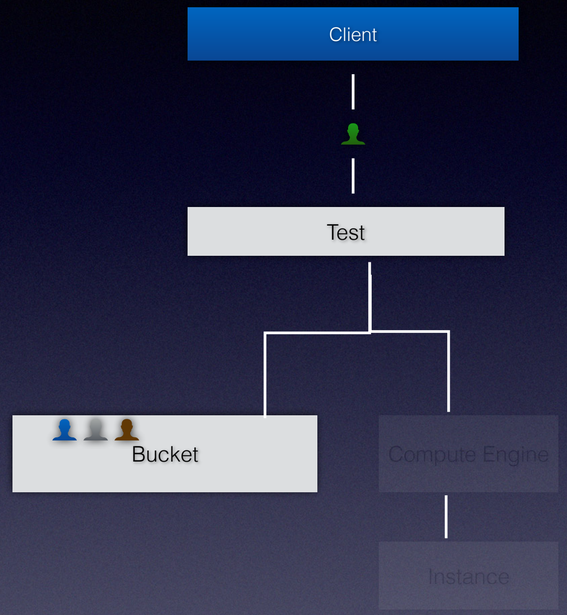
Processes - (Resource Performs hidden processes)

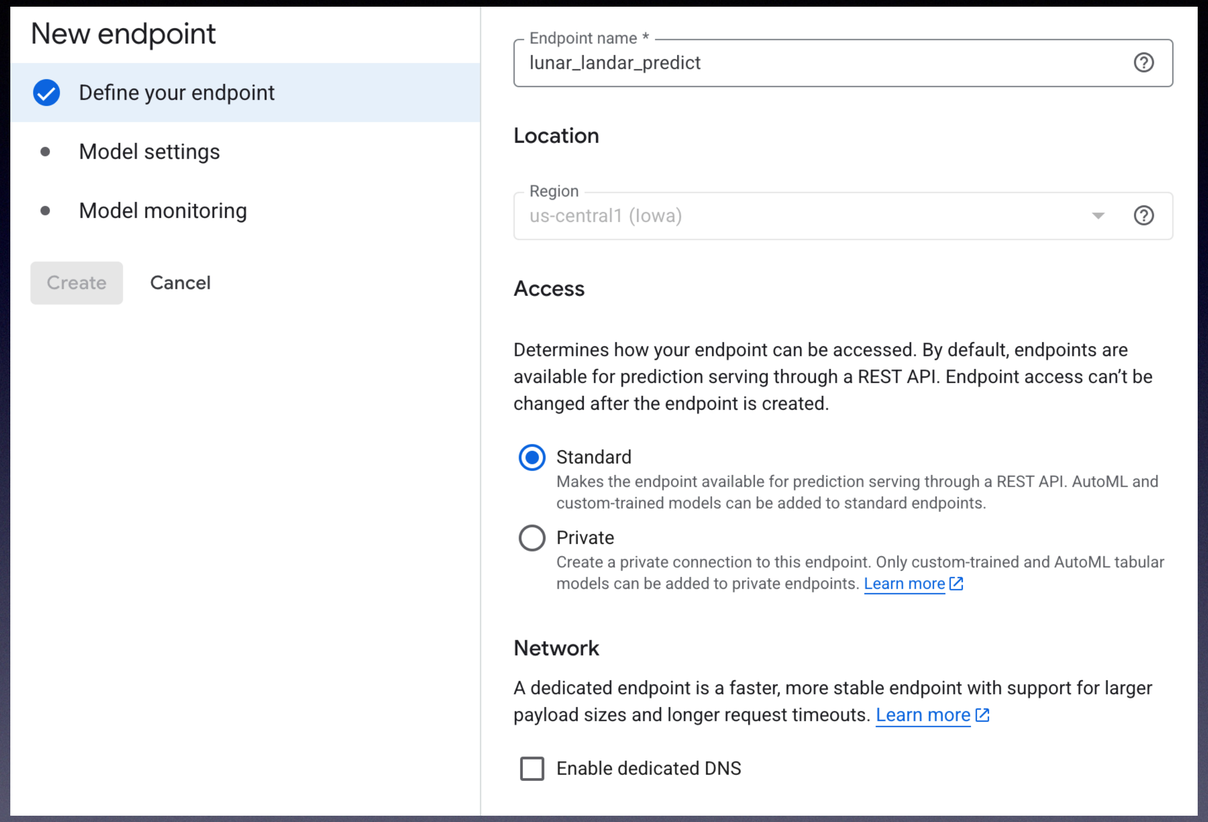
GET Request

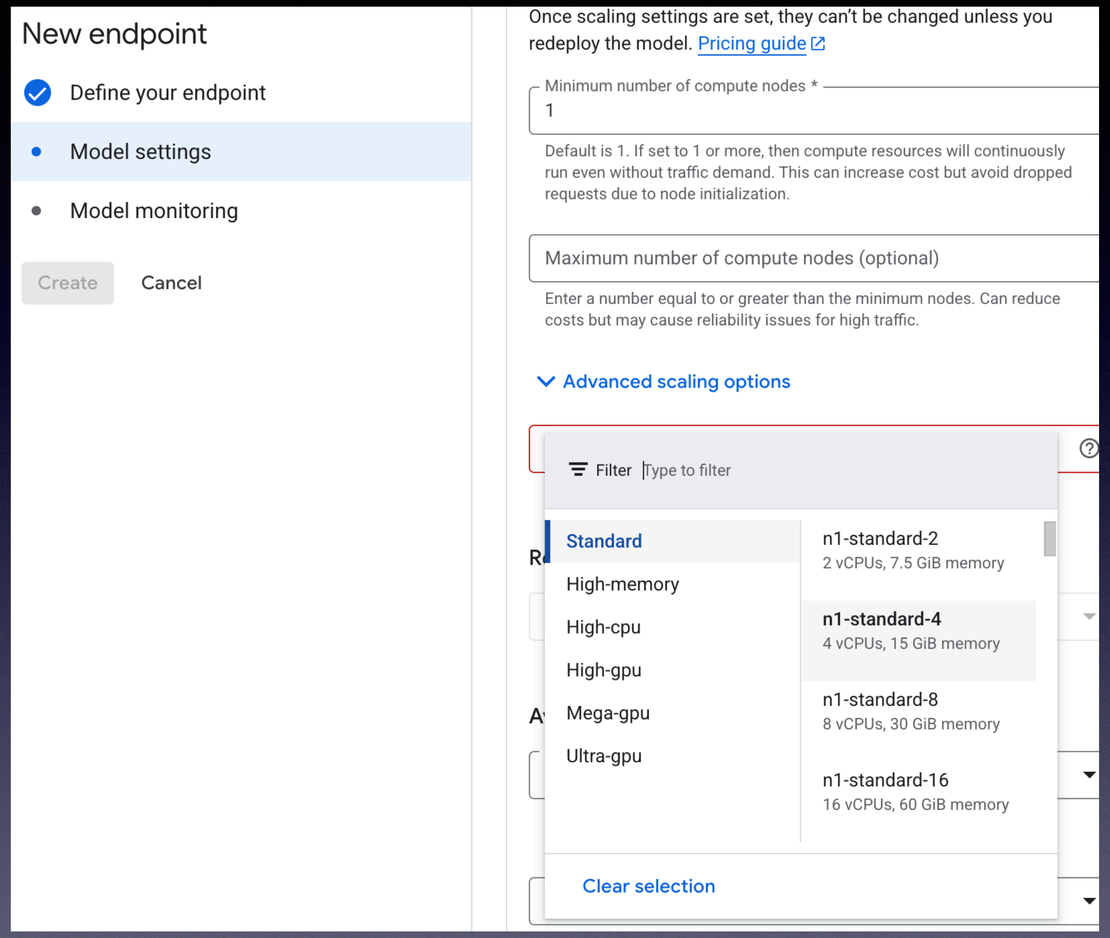
```
hectorwilliams@Hectors-Air ~ % curl \
-X GET \
-H "Authorization: Bearer $(./google-cloud-sdk/bin/gcloud auth print-access-token)" \
-H "Content-Type: application/json" \
"https://storage.googleapis.com/storage/v1/b/lunar_lander_model_bucket/iam"
```

Response

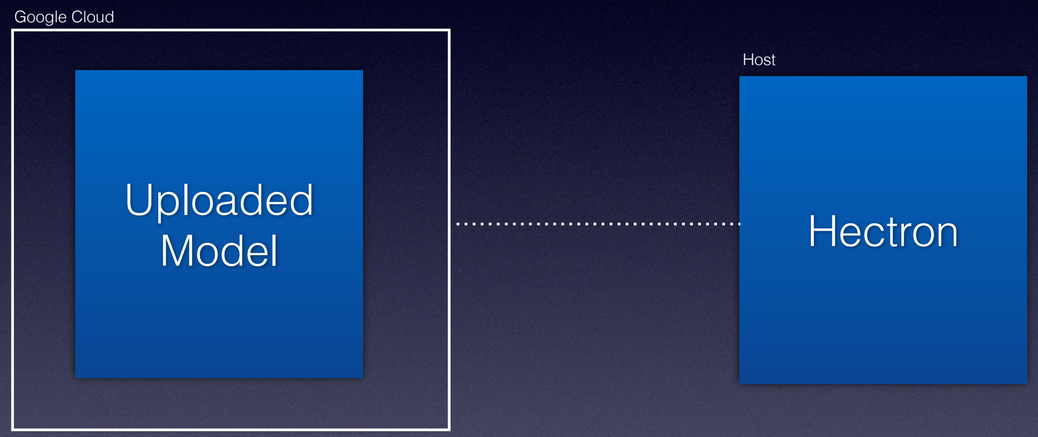

```
{
  "kind": "storage#policy",
  "resourceId": "projects/_/buckets/lunar_lander_model_bucket",
  "version": 1,
  "etag": "CAY=",
  "bindings": [
    {
      "role": "roles/storage.legacyBucketOwner",
      "members": [
        "projectEditor:flash-bazaar-233821",
        "projectOwner:flash-bazaar-233821"
      ]
    },
    {
      "role": "roles/storage.legacyBucketReader",
      "members": [
        "projectViewer:flash-bazaar-233821"
      ]
    },
    {
      "role": "roles/storage.legacyObjectOwner",
      "members": [
        "projectEditor:flash-bazaar-233821",
        "projectOwner:flash-bazaar-233821"
      ]
    },
    {
      "role": "roles/storage.legacyObjectReader",
      "members": [
        "projectViewer:flash-bazaar-233821"
      ]
    }
  ]
}
```


New endpoint
Define your endpoint
Model settings
Model monitoring
Create
Cancel
Endpoint name *
lunar_landar_predict
Location
Region
us-central1 (Iowa)
Access
Determines how your endpoint can be accessed. By default, endpoints are available for prediction serving through a REST API. Endpoint access can't be changed after the endpoint is created.
Standard
Makes the endpoint available for prediction serving through a REST API. AutoML and custom-trained models can be added to standard endpoints.
Private
Create a private connection to this endpoint. Only custom-trained and AutoML tabular models can be added to private endpoints. Learn more
Network
A dedicated endpoint is a faster, more stable endpoint with support for larger payload sizes and longer request timeouts. Learn more
Enable dedicated DNS


New endpoint
Define your endpoint
Model settings
Model monitoring
Create
Cancel
Once scaling settings are set, they can't be changed unless you redeploy the model. Pricing guide
Minimum number of compute nodes *
1
Default is 1. If set to 1 or more, then compute resources will continuously run even without traffic demand. This can increase cost but avoid dropped requests due to node initialization.
Maximum number of compute nodes (optional)
Enter a number equal to or greater than the minimum nodes. Can reduce costs but may cause reliability issues for high traffic.
Advanced scaling options
Filter Type to filter
Standard
High-memory
High-cpu
High-gpu
Mega-gpu
Ultra-gpu
n1-standard-2
2 vCPUs, 7.5 GiB memory
n1-standard-4
4 vCPUs, 15 GiB memory
n1-standard-8
8 vCPUs, 30 GiB memory
n1-standard-16
16 vCPUs, 60 GiB memory
Clear selection

Query a prediction Model (Tensorflow model)



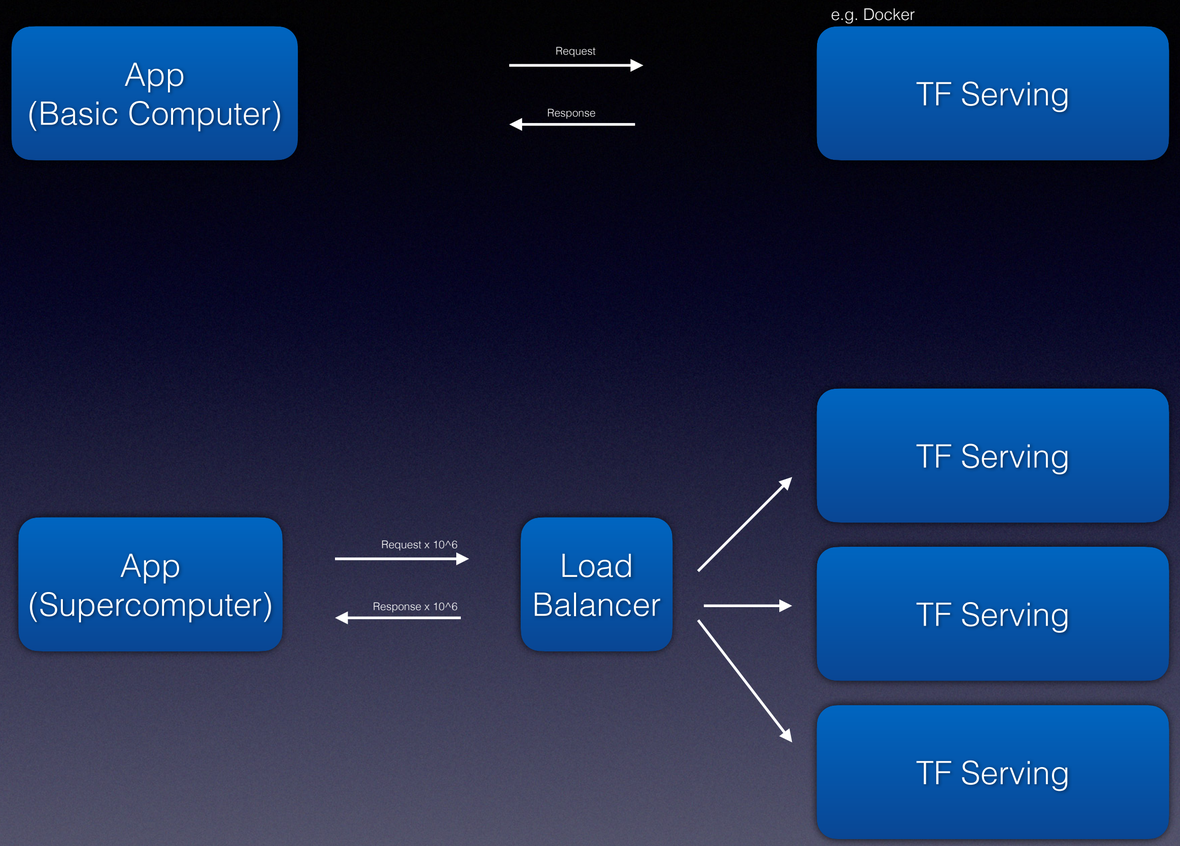
Query a prediction Model

```
#! /usr/bin/env python3

"""
    Send an online prediction request to a shared public endpoint

"""

from google.cloud import aiplatform
import os

PROJECT_ID = '869083780811'
ENDPOINT_ID = '1941215266629222400'
ENDPOINT_ID2 = '8611890739695058944'
LOCATION_ID = 'us-central1'
os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = os.path.join(os.getcwd(), 'cred.json')

def endpoint_predict_sample( project: str, location: str, instances: list, endpoint: str):
    aiplatform.init(project=project, location=location, staging_bucket='gs://my_staging_bucket')
    endpoint = aiplatform.Endpoint(endpoint)
    prediction = endpoint.predict(instances=instances)
    return prediction

instances = [

    [  0.30, 0.20, 0.20, 0.10, 0.20, 0.10, 0.20, 0.110],

    [-1.371715  ,  0.58048093,  0.14552322,  0.30510637,  1.4215977 ,0.57935536, -0.30329934,  1.4869808 ],

    [ 1.8956385 , -0.45161435,  0.46339658, -0.46701705,  0.24996527,-0.8164176 ,  1.0037286 ,  1.355299  ]

]


if __name__ == "__main__":
    ret = endpoint_predict_sample(project=PROJECT_ID, location= LOCATION_ID, instances=instances, endpoint=ENDPOINT_ID2)
    print(ret)

"""
Google Cloud Model

Model ID:
  814655652809932800

Name:
  lunar_lander_model

Version:
  1

ENDPOINT_ID:
  1941215266629222400

SDK Output:

  Prediction(predictions=[[0.312848359, 0.236094132, 0.217715696, 0.233341739], [0.414523602, 0.112857334, 0.306751847, 0.16586718], [0.319322735, 0.398763537, 0.159842864, 0.122070901]], deployed_model_id='8489429333616623616', metadata=None, model_version_id='1',
model_resource_name='projects/869083780811/locations/us-central1/models/814655652809932800', explanations=None)

"""

"""
Google Cloud Model

Model ID:
  814655652809932800

Name:
  lunar_lander_model

Version:
  6

ENDPOINT_ID2:
  8611890739695058944

SDK Output:

Prediction(predictions=[[0.262676597, 0.271419078, 0.255823731, 0.210080639], [0.269151241, 0.29415518, 0.262208253, 0.174485326], [0.26882872, 0.209912121, 0.406539112, 0.114720076]], deployed_model_id='214346493300047872', metadata=None, model_version_id='6',
model_resource_name='projects/869083780811/locations/us-central1/models/814655652809932800', explanations=None)


"""
```

Saved model exported using model.export(<path>), not tf.saved_model.save()

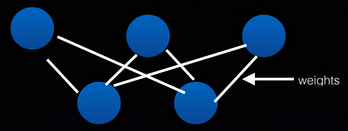
Deploying

TF Serving Containers



Load Balancer and multiple Servings distributed across machines are required to handler large request loads

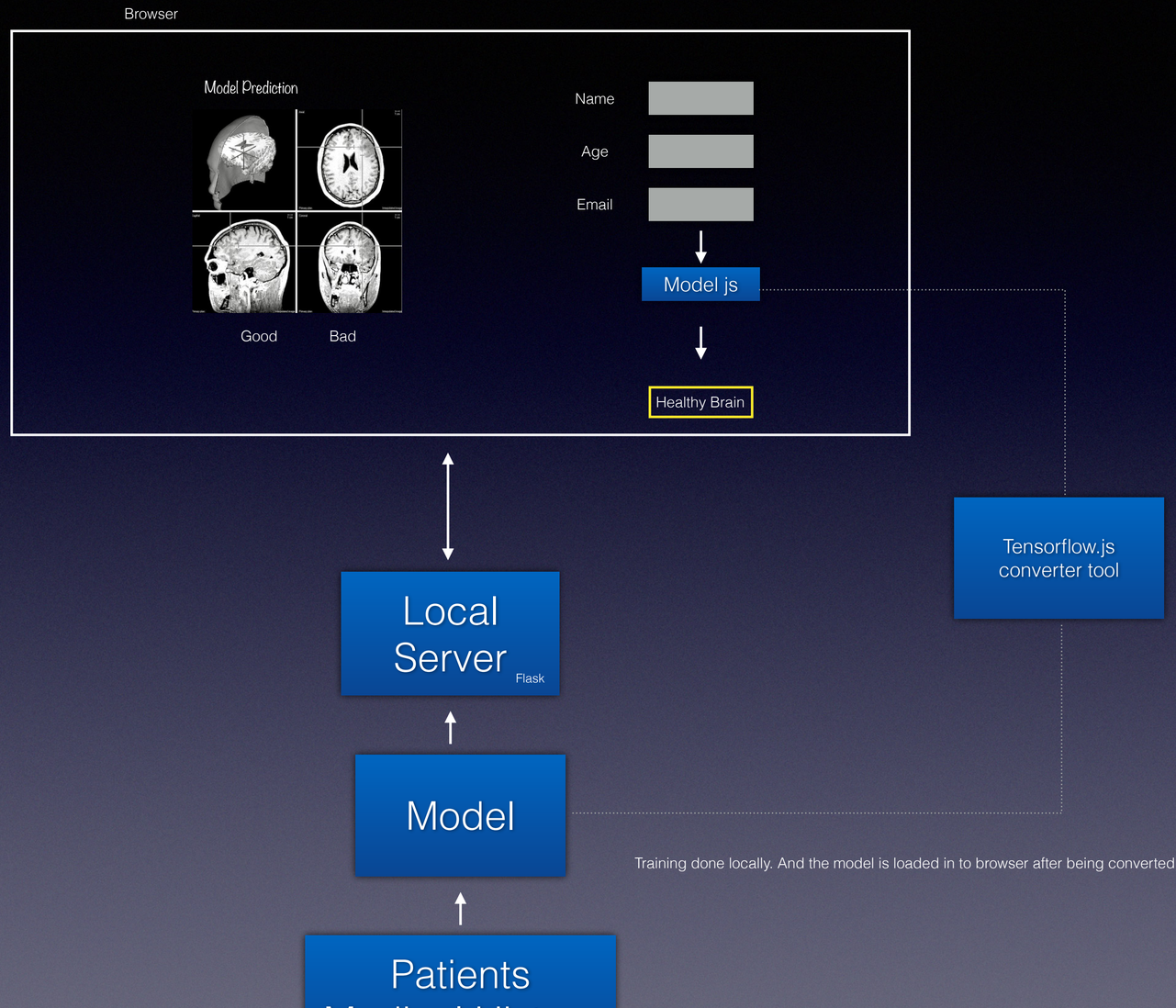
Post-Training Quantization



$$\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}$$

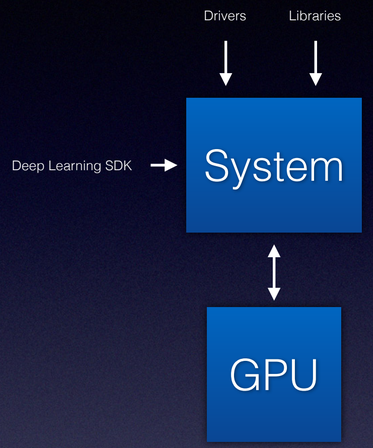
weights: float → Post Train Quantization → weights: bytes → Download/Storage → weights: bytes → Process { Undo Post Train Quantization (Recover) → weights: float → Inference }

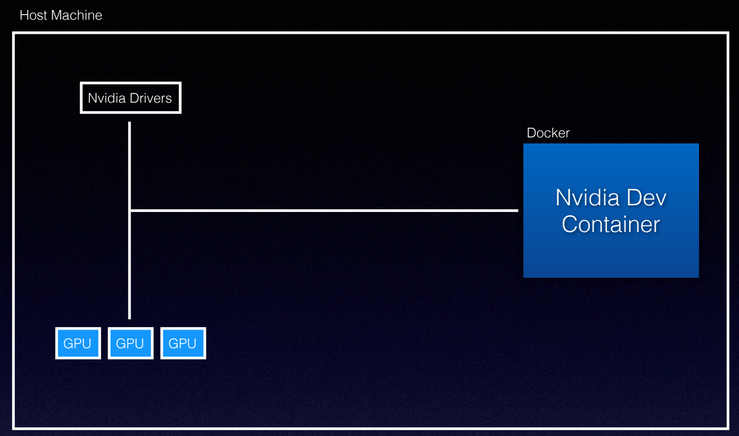
Browser Tensorflow



Training done locally. And the model is loaded in to browser after being converted to a Tensorflow.js Layers format.

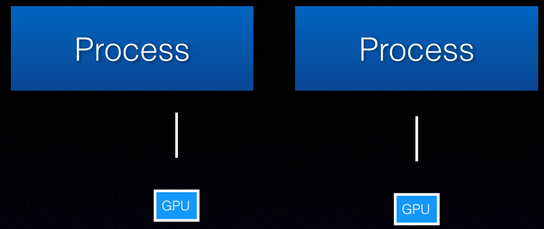
GPU



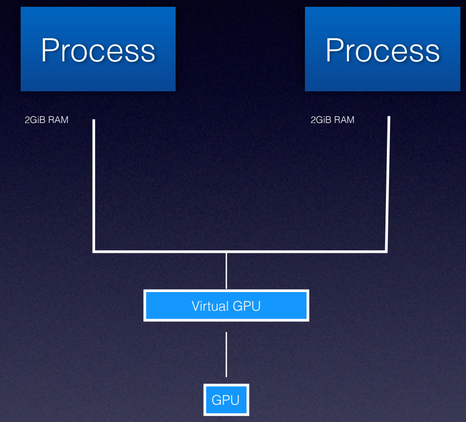


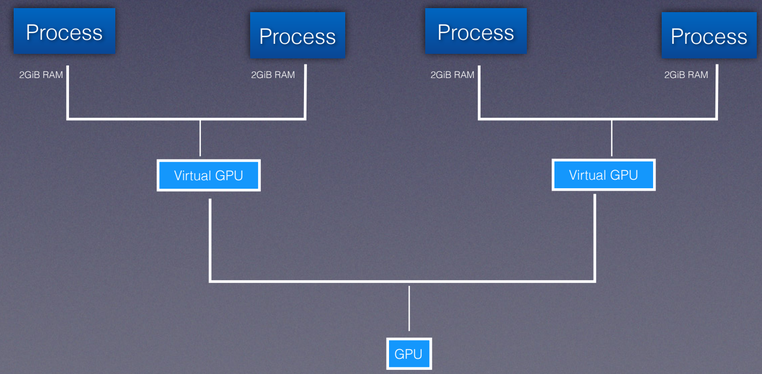
Drivers installed on host machine to use GPU development libraries

Tensorflow programs require independent GPUs

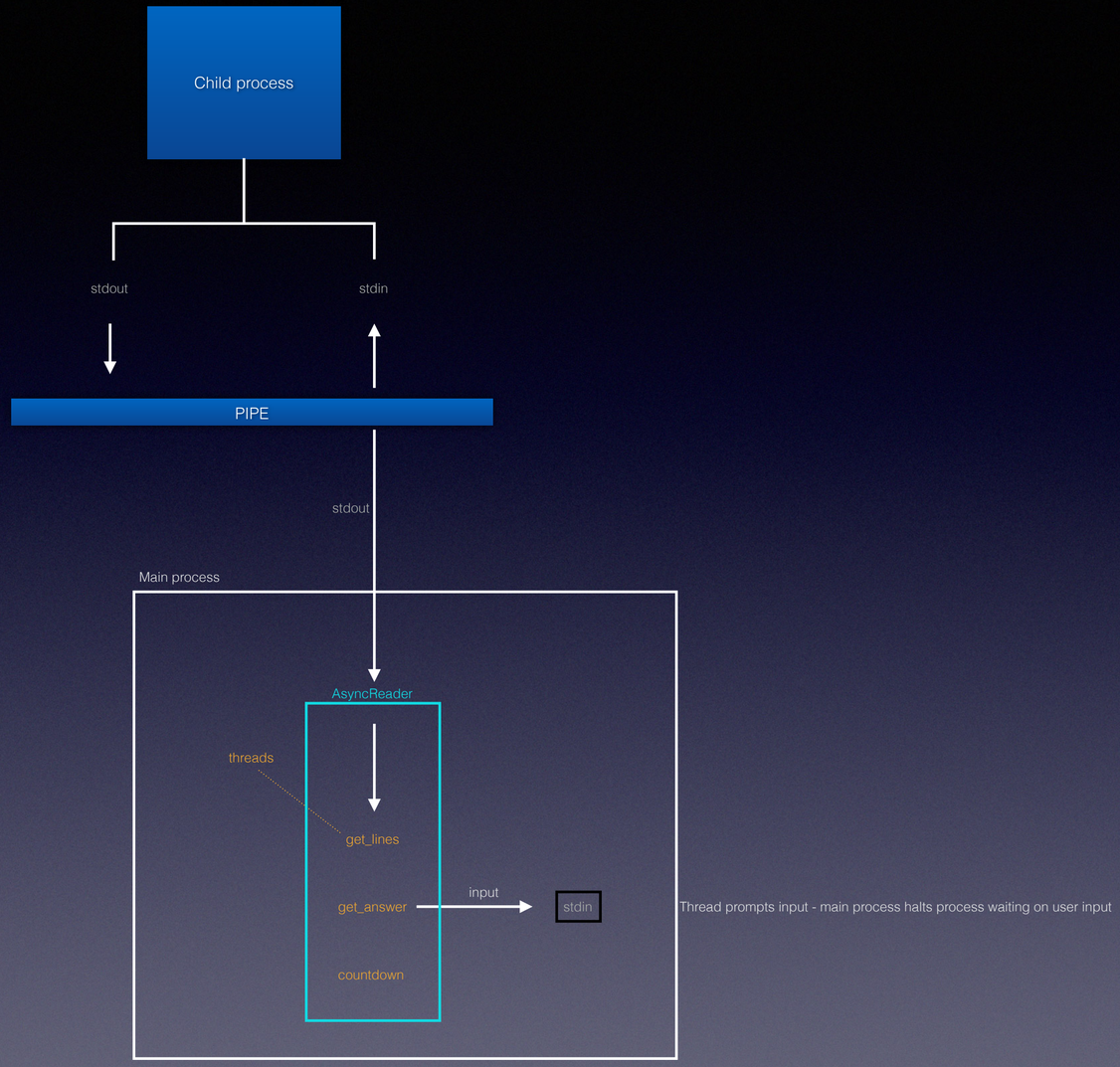


Programs limit memory demand on single GPU (2GiB memory per program)



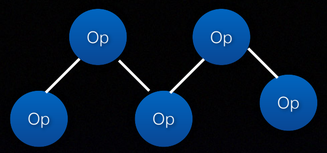


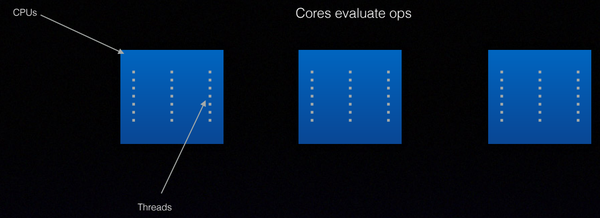
AsyncReader



Scrapped because input routine is required. Inputs halts process, which is not good for automation.

CPU

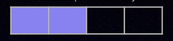




Cores evaluate ops

Queued ops when ready

T0



Multi-core operation (multithreaded is operations permit)

T1

Single-core operation (multithreaded is operations permit)

T2

GPU

Sequential ops by GPU

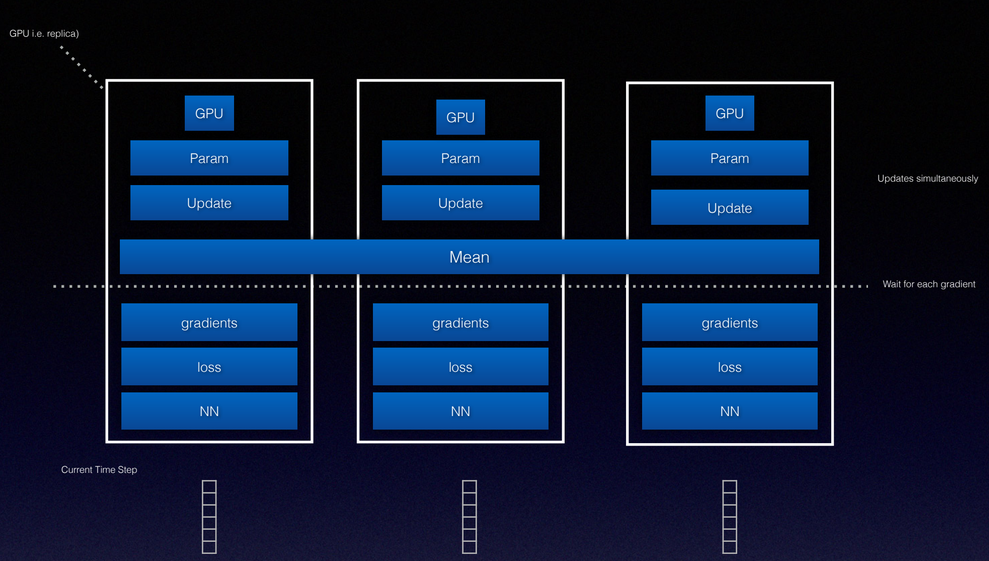
T0

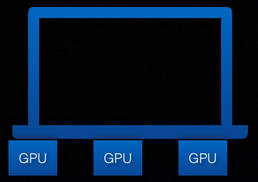
T1

T2

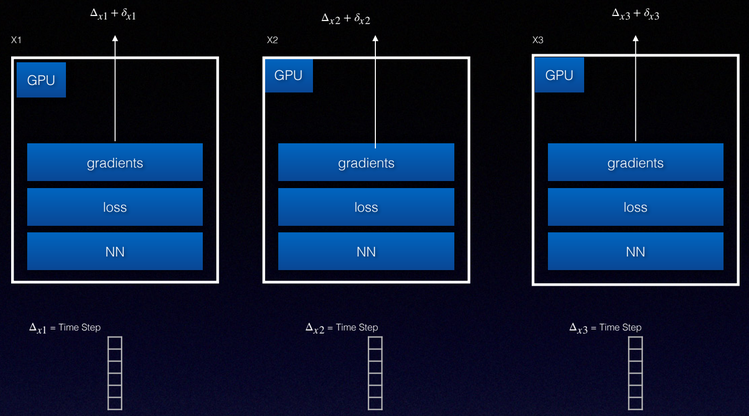
T3

Data Parallelism (Synchronous)

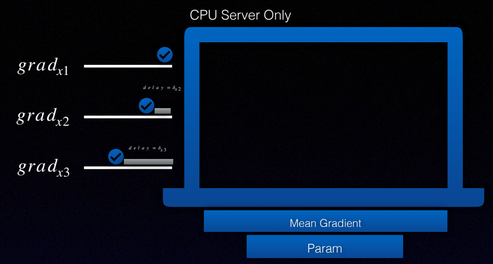




Data Parallelism (Asynchronous)



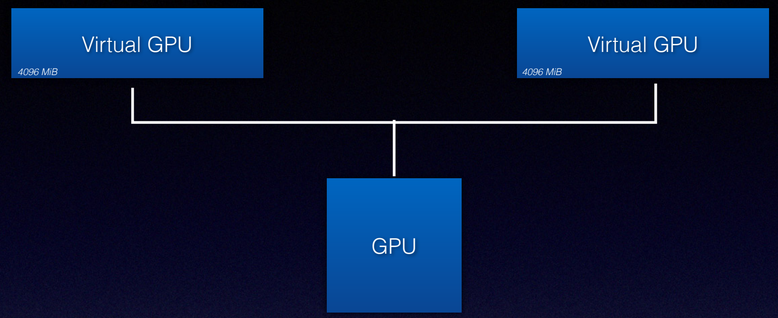
Gradients status received by parameter server.
These gradients are handled by event queue to update parameters

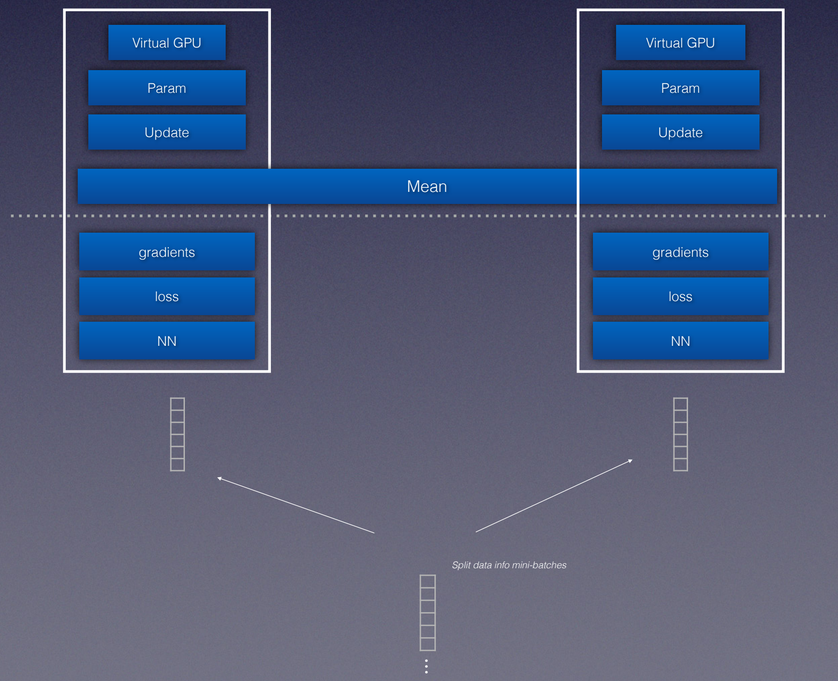


Recommended Design Path:

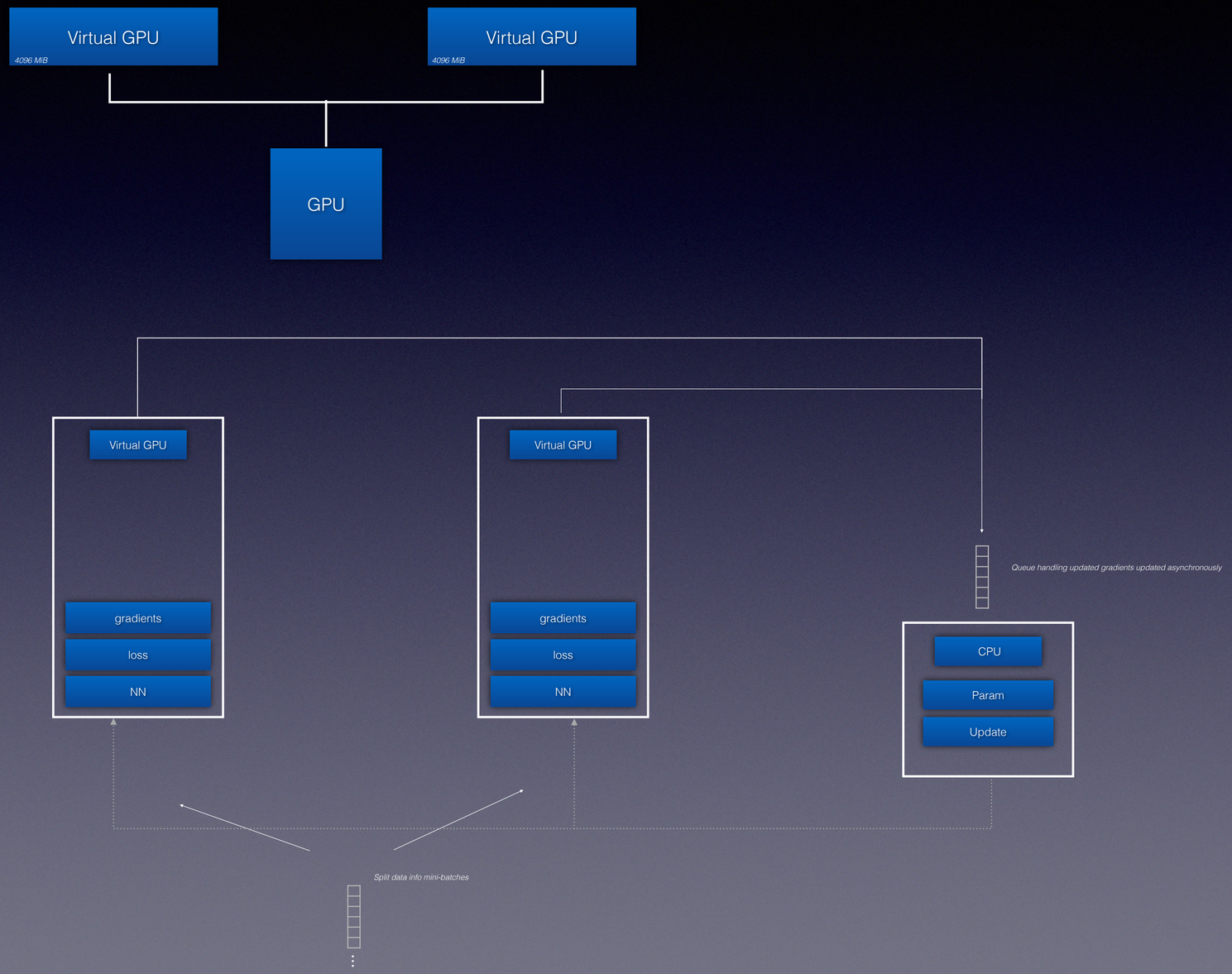
First 20 epochs on on replica or one GPU → Epoch > 20 apply training to all replicas. Use a reduced learning rate

Data Parallelism (Mirrored Example)

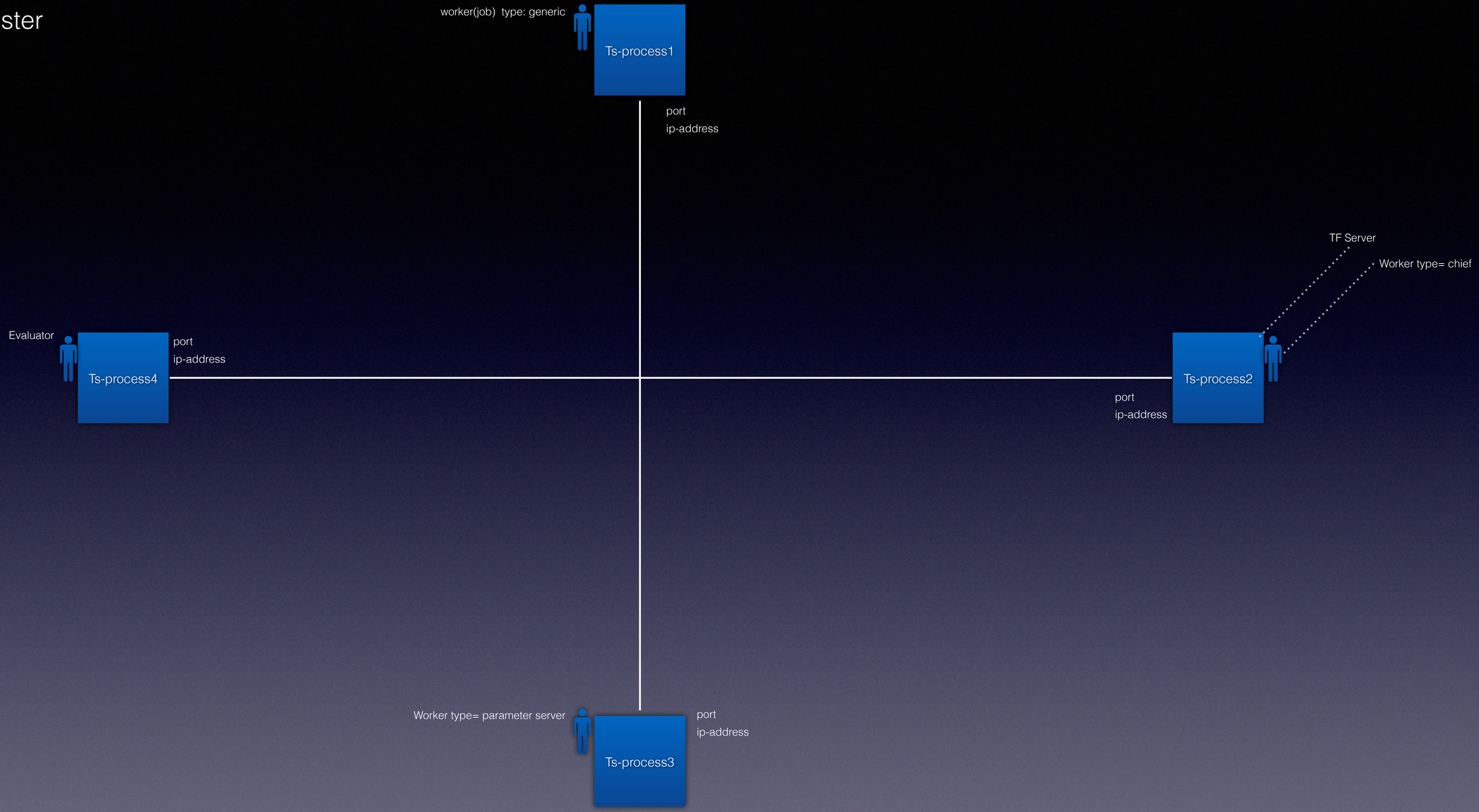




Data Parallelism (Centralized Example)

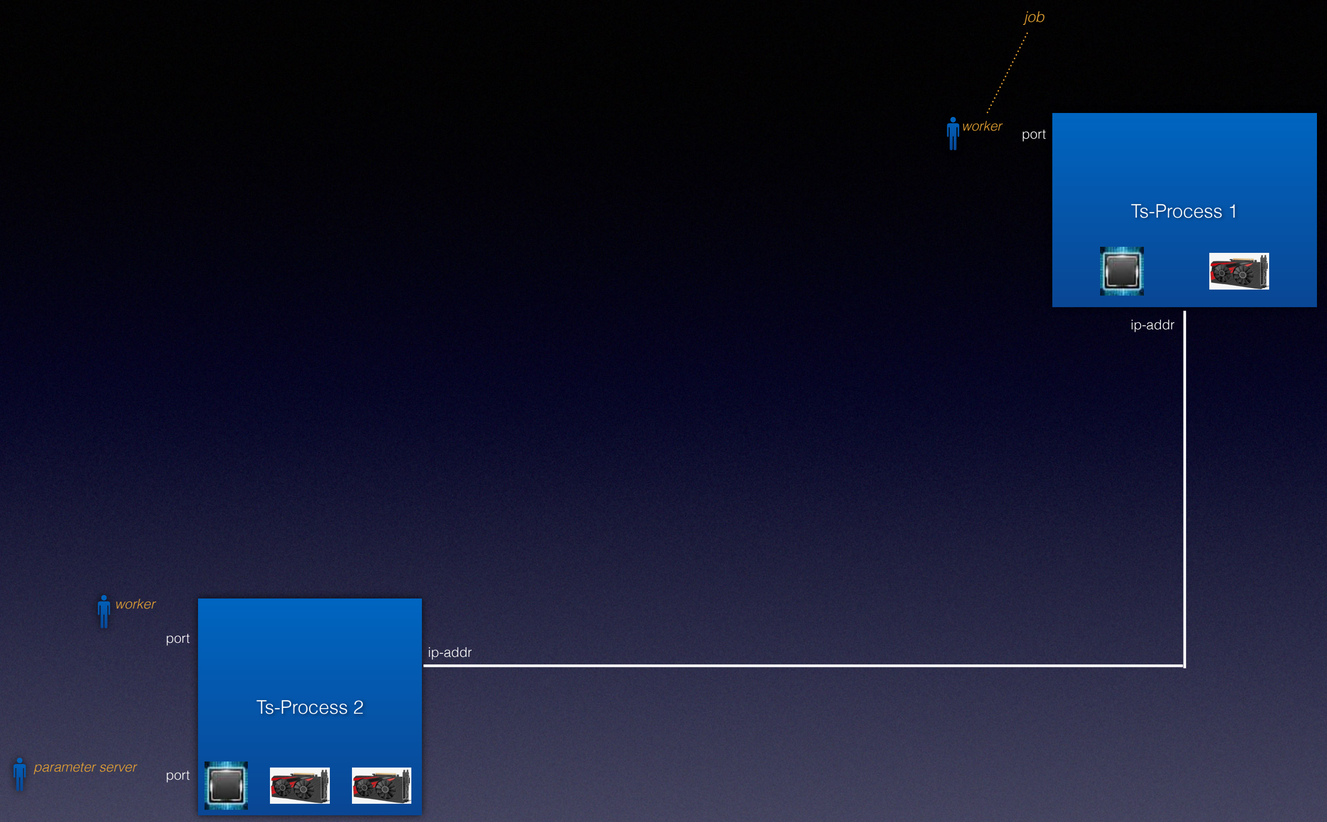


Tensorflow Cluster



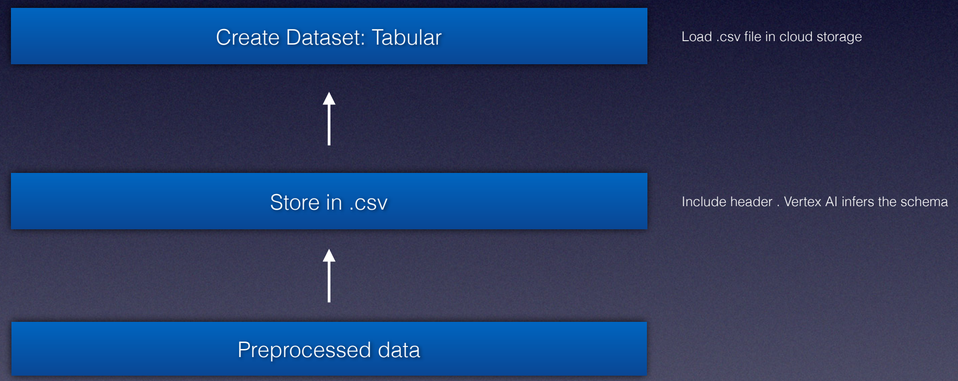
Working together to complete task: training or execute neural network

Tensorflow Cluster

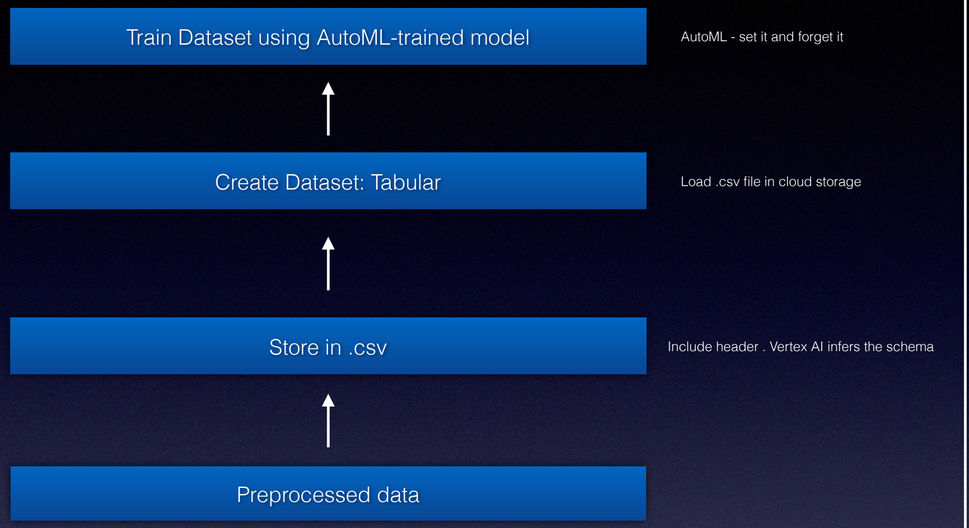


Working together to complete task: training or execute neural network

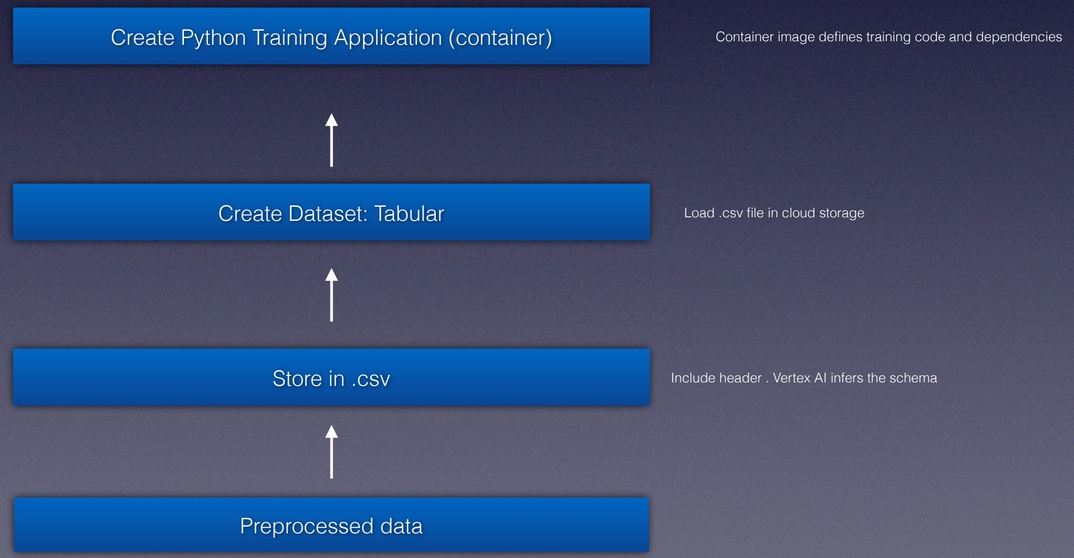
Data Parallelism (Mirror Example)



Train Job Google Cloud



https://cloud.google.com/vertex-ai/docs/training/create-custom-job#create_custom_job-gcloud



Python Packaging

Package Installer

Package Format

Wheel

Ships libraries with compiled artifacts (C, C++,etc

Package Format

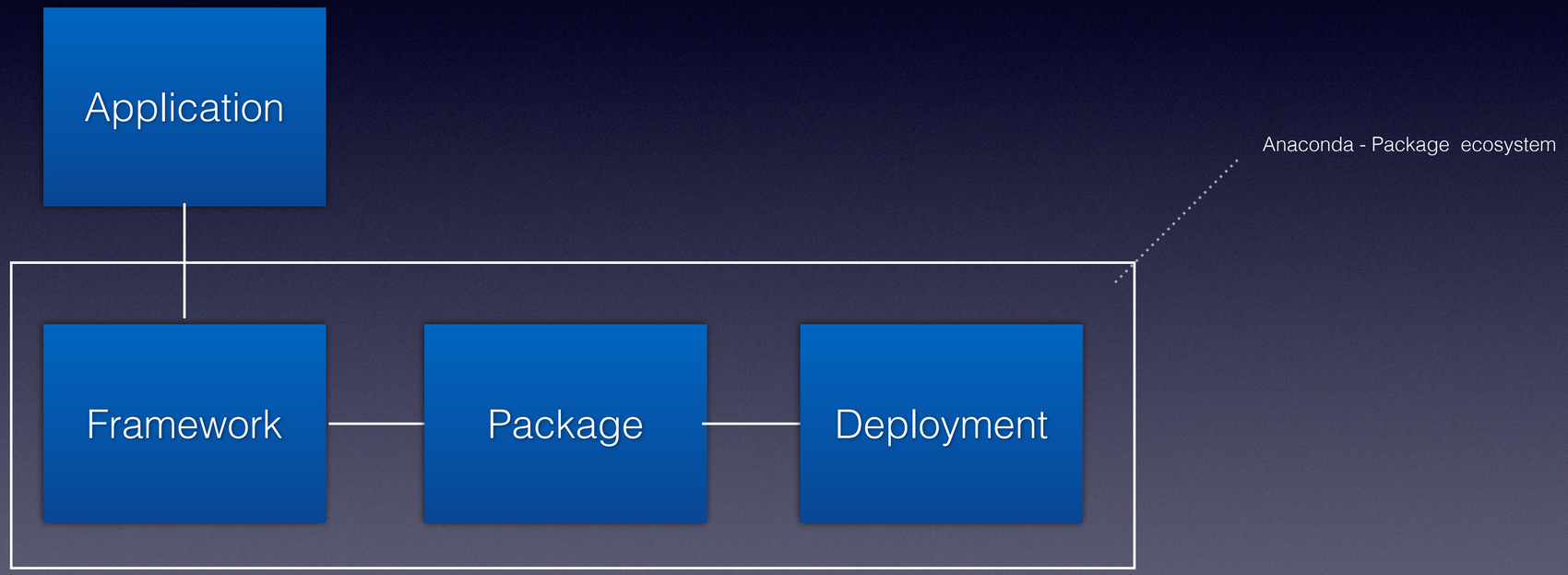
sdist

Users build binaries from source distributions

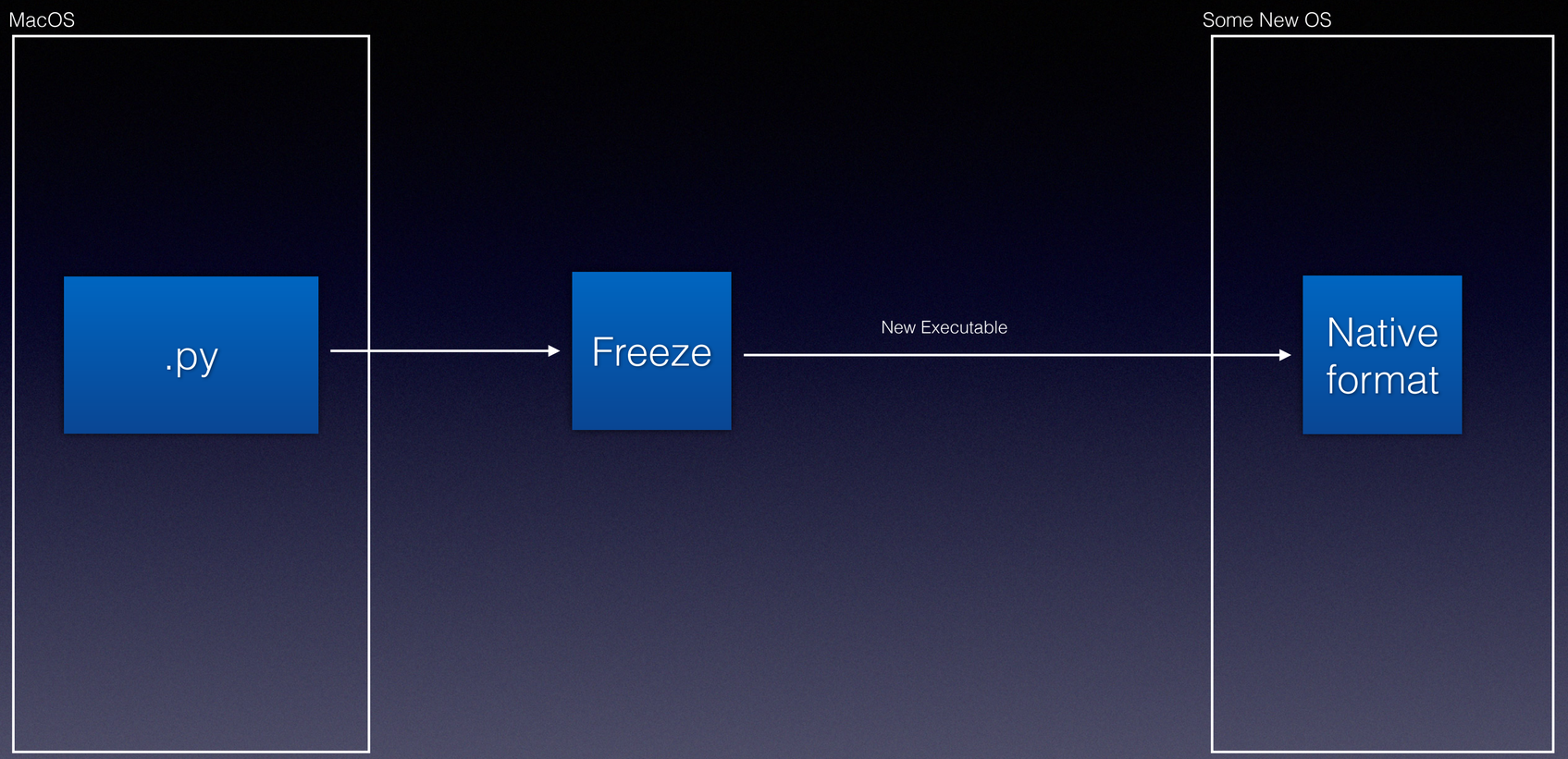
Package Format

.py

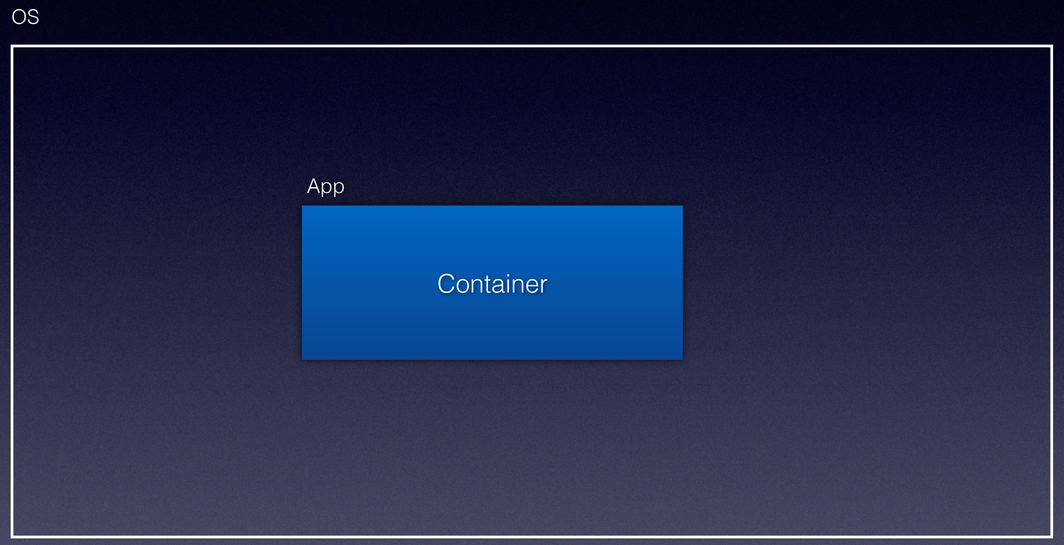
Standalone modules


Application
Framework
Package
Deployment
Anaconda - Package ecosystem

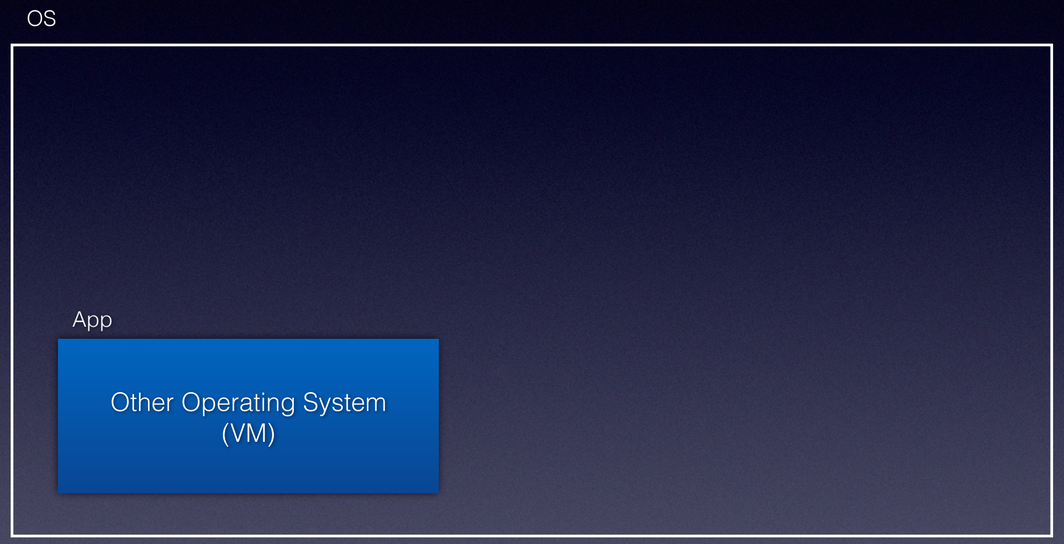
Python Executable to Native Format via Freezing



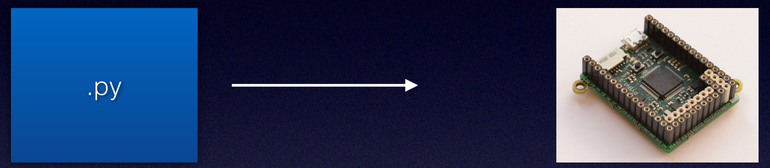
OS Level Virtualization



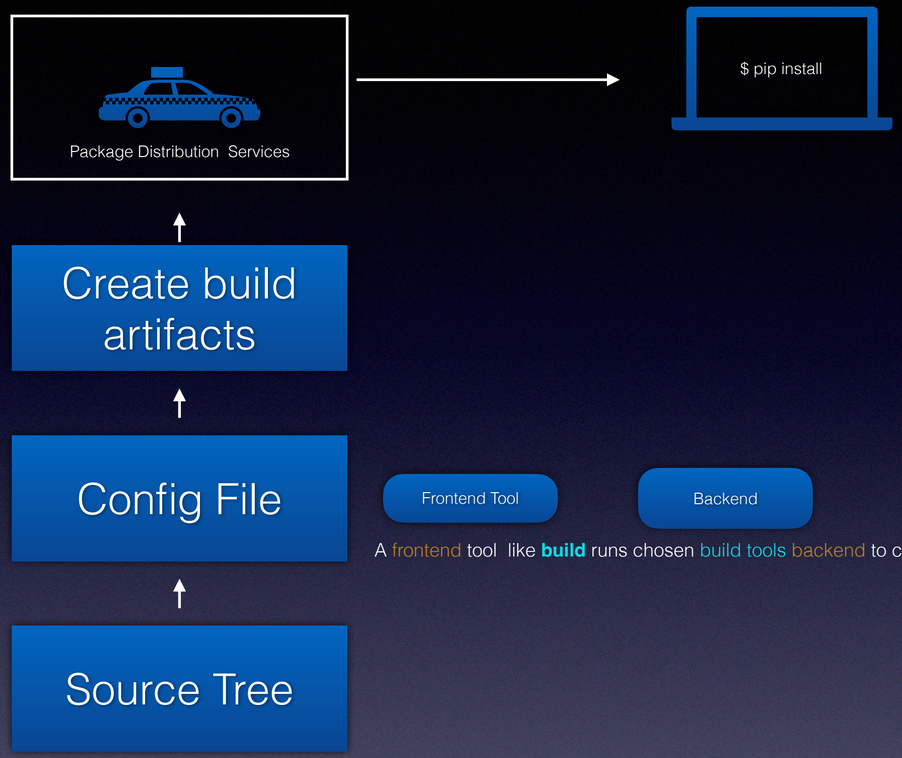
Classical Virtualization



Embed onto Mini Microcontrollers

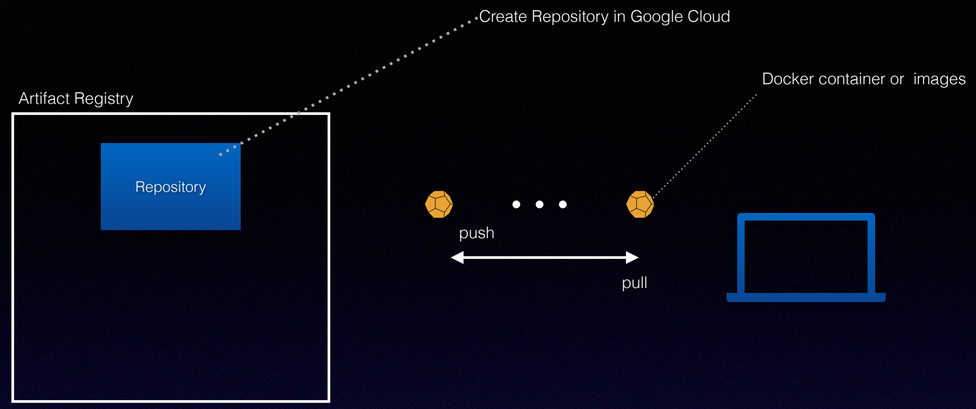


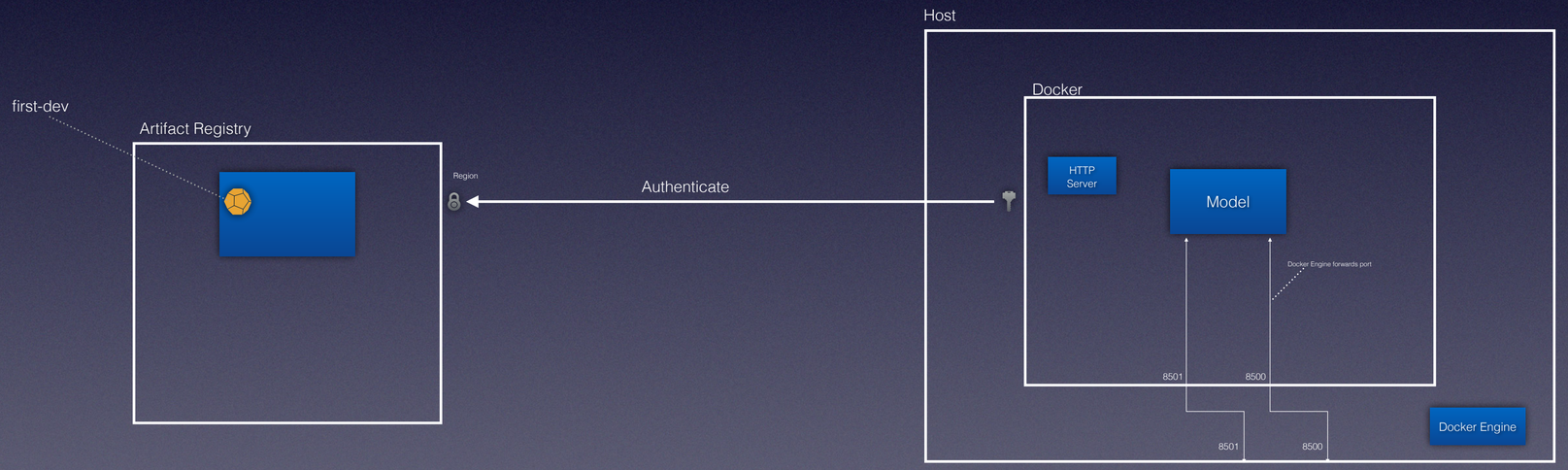
Packaging Flow

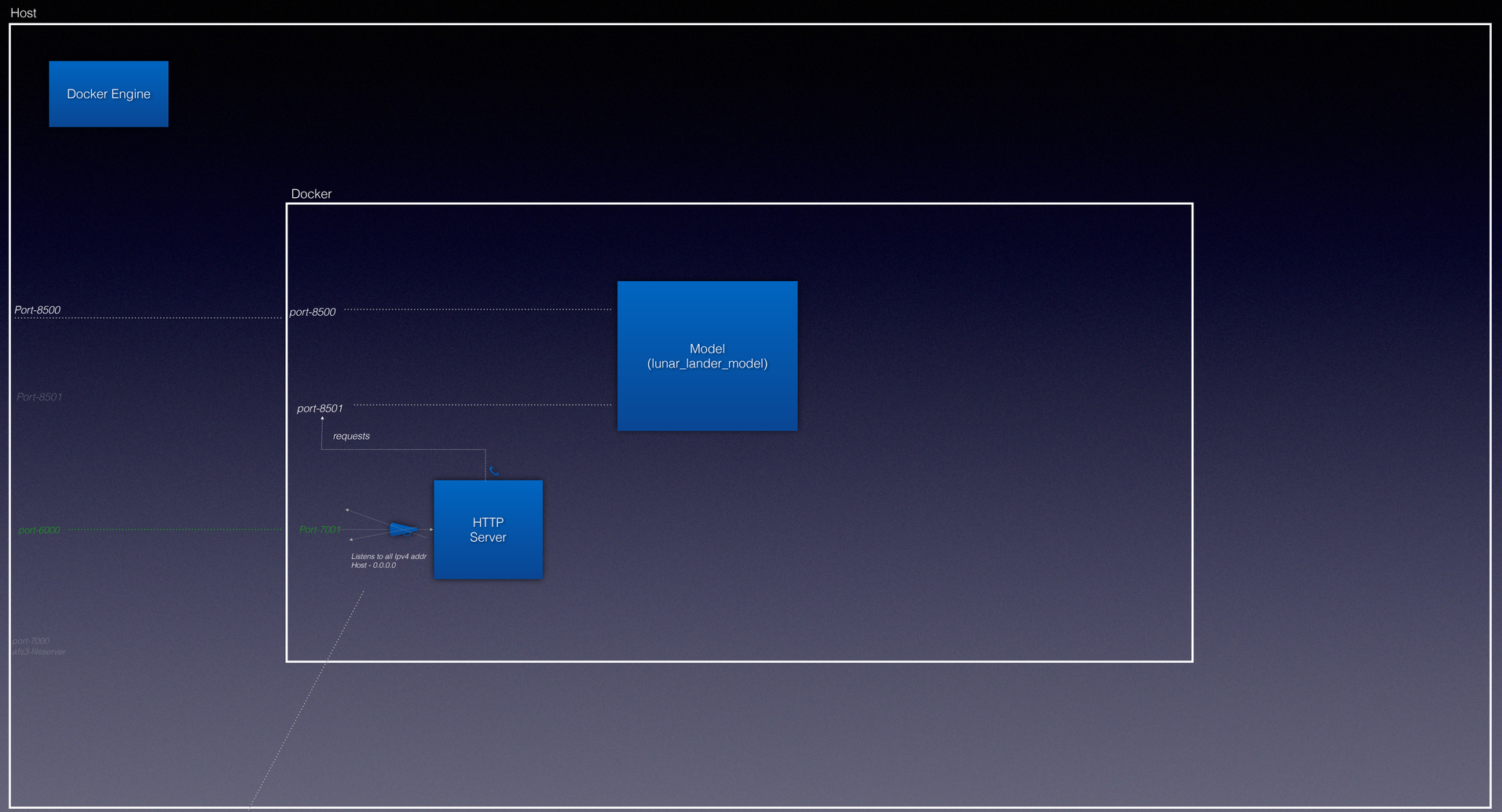


Create Python Training Application

https://console.cloud.google.com/artifacts?project=flash-bazaar-233821

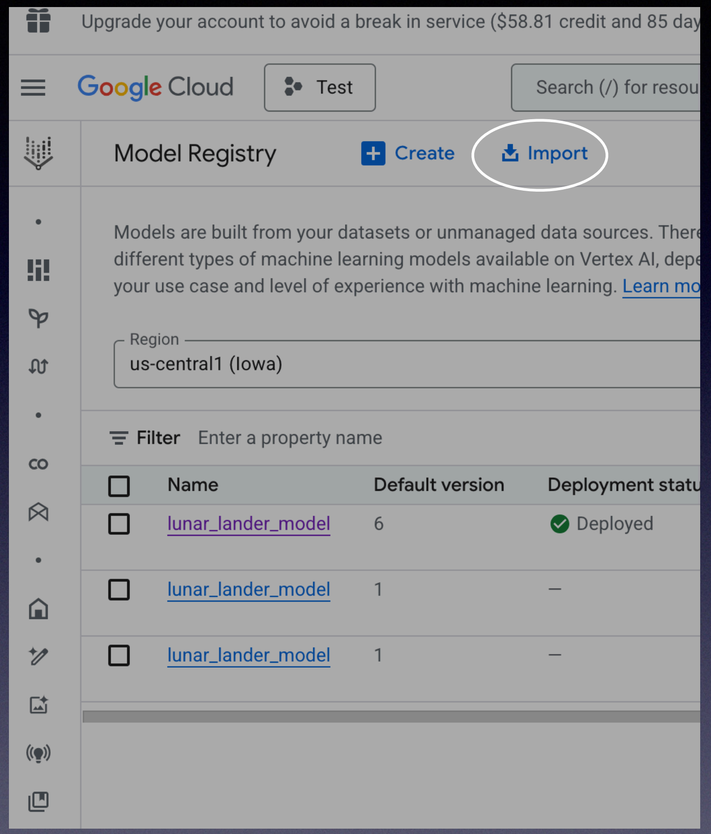






Host request to model POST http://localhost:8501/v1/models/lunar_lander_model:predict

All address can predict the server.
Set to a single address instead (i.e. localhost)


Upgrade your account to avoid a break in service ($58.81 credit and 85 day
Google Cloud
Test
Search (/) for resou
Model Registry
Create
Import
Models are built from your datasets or unmanaged data sources. There
different types of machine learning models available on Vertex AI, depe
your use case and level of experience with machine learning. Learn mo
Region
us-central1 (Iowa)
Filter
Enter a property name
Name
Default version
Deployment statu
lunar_lander_model
6
Deployed
lunar_lander_model
1
–
lunar_lander_model
1
–

Import model

1. Name and region
2. Model settings
3. Explainability (optional)

Import | Cancel

You can import model artifacts that have been trained outside of Google Cloud. Once your model has been imported, you can serve it for online or batch predictions and compare it against your other Cloud AI models. More info

- (•) Import as new model
  Creates a new model group and assigns the imported model as version 1
- ( ) Import as new version
  Imports the model as a version of an existing model

Name *
classify_custom_model

Description
Model uses docker container

Region
us-central1 (Iowa)

Encryption

- (•) Google-managed encryption key
  Keys owned by Google
- ( ) Cloud KMS key
  Keys owned by customers

Model ID

Show less

Continue

Import model

- ✓ Name and region
- 2 Model settings
- 3 Explainability (optional)

Import | Cancel

○ Import model artifacts into a new pre-built container
View the list of supported runtimes including TensorFlow, scikit-learn and XGBoost versions

◉ Import an existing custom container
Build a custom Docker container. Must be stored in Artifact Registry

Custom container settings

Container image
us-docker.pkg.dev/flash-bazaar-233821/secondrepo/mytest@sh Browse

The URI for the Docker image in Artifact Registry. You must have permission to access the image. Learn more

Command

Optional. Specifies the command that runs when the container starts. Overrides the container's ENTRYPOINT instructions.

gs:// Model artifact location (Cloud storage path) Browse

Path to the Cloud Storage directory where the exported model file is stored (not the path to the model file itself).

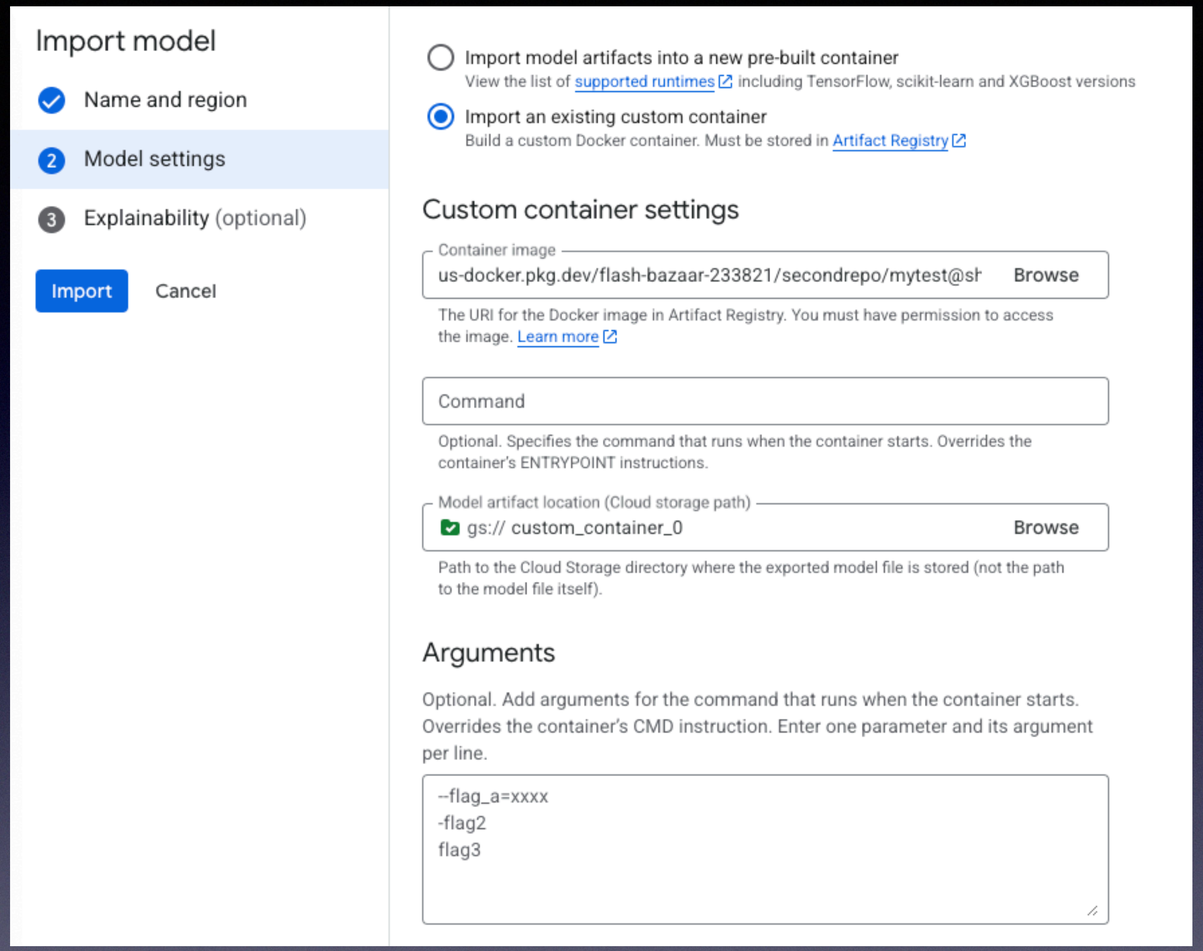
Arguments

Optional. Add arguments for the command that runs when the container starts. Overrides the container's CMD instruction. Enter one parameter and its argument per line.

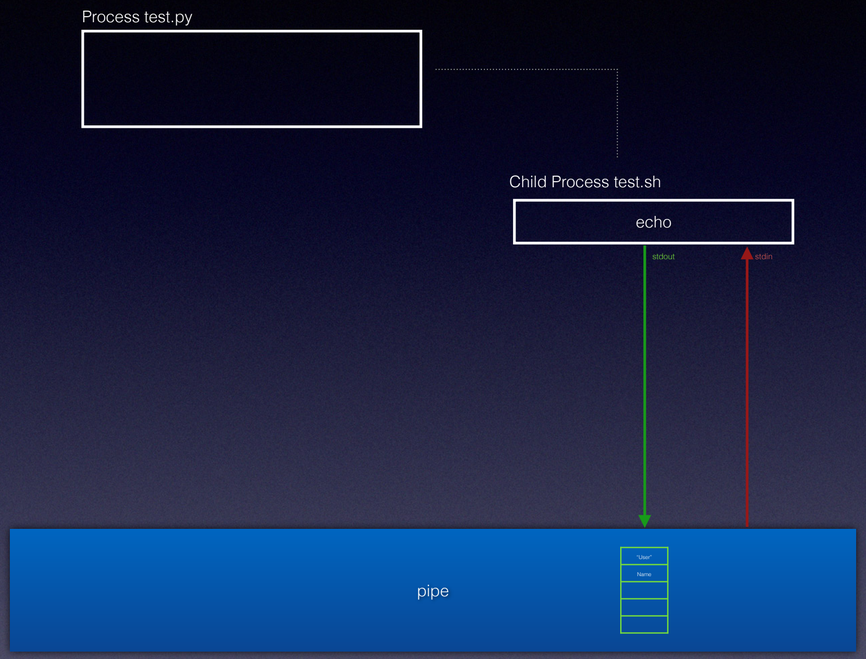
```
--flag_a=xxxx
-flag2
flag3
```

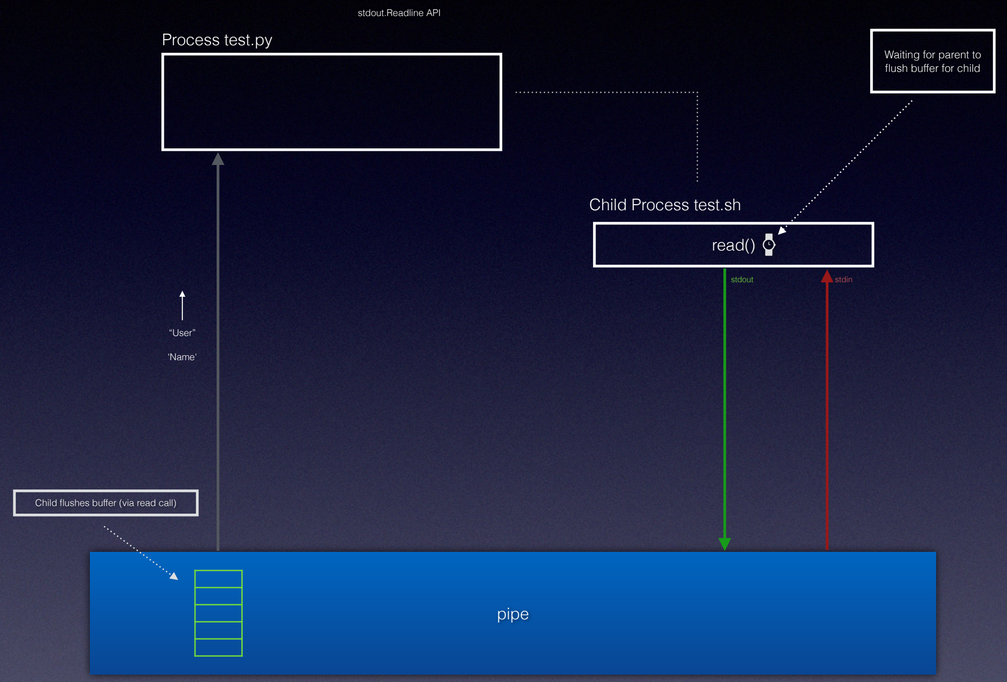
Environment variables

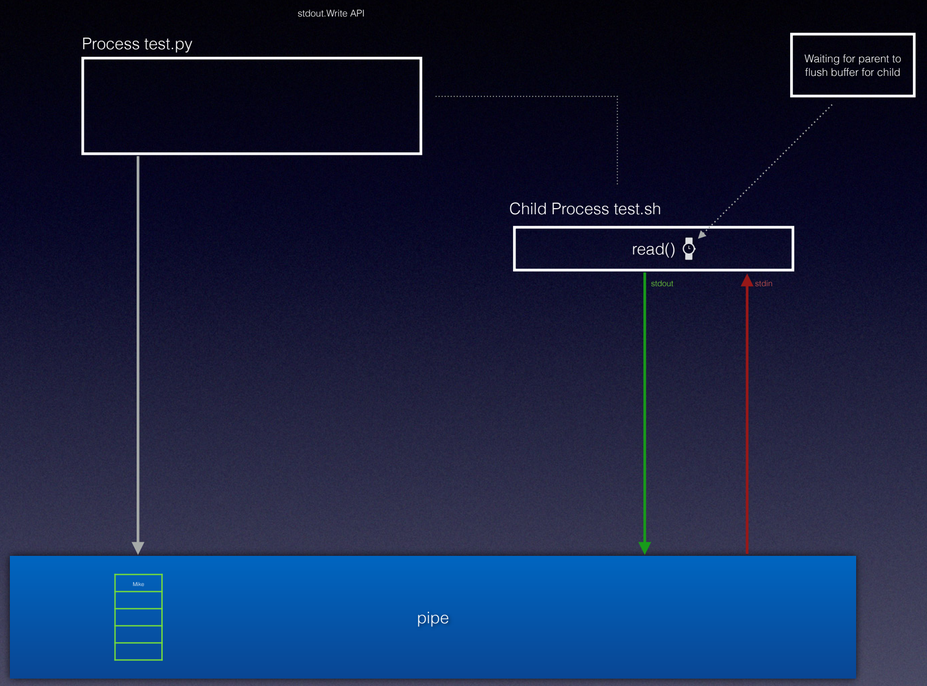
Optional. Enter one variable per line, separating the key and value with an equal (=) sign.

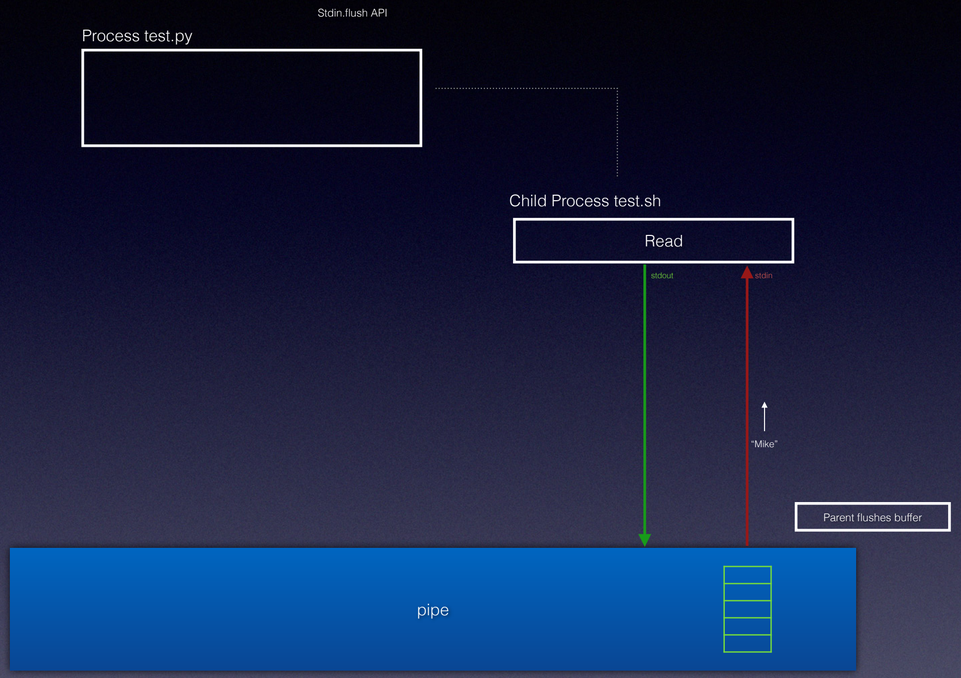

Import model
Name and region
2 Model settings
3 Explainability (optional)
Import
Cancel
Import model artifacts into a new pre-built container
View the list of supported runtimes including TensorFlow, scikit-learn and XGBoost versions
Import an existing custom container
Build a custom Docker container. Must be stored in Artifact Registry
Custom container settings
Container image
us-docker.pkg.dev/flash-bazaar-233821/secondrepo/mytest@sh
Browse
The URI for the Docker image in Artifact Registry. You must have permission to access the image. Learn more
Command
Optional. Specifies the command that runs when the container starts. Overrides the container's ENTRYPOINT instructions.
Model artifact location (Cloud storage path)
gs:// custom_container_0
Browse
Path to the Cloud Storage directory where the exported model file is stored (not the path to the model file itself).
Arguments
Optional. Add arguments for the command that runs when the container starts. Overrides the container's CMD instruction. Enter one parameter and its argument per line.
--flag_a=xxxx
-flag2
flag3

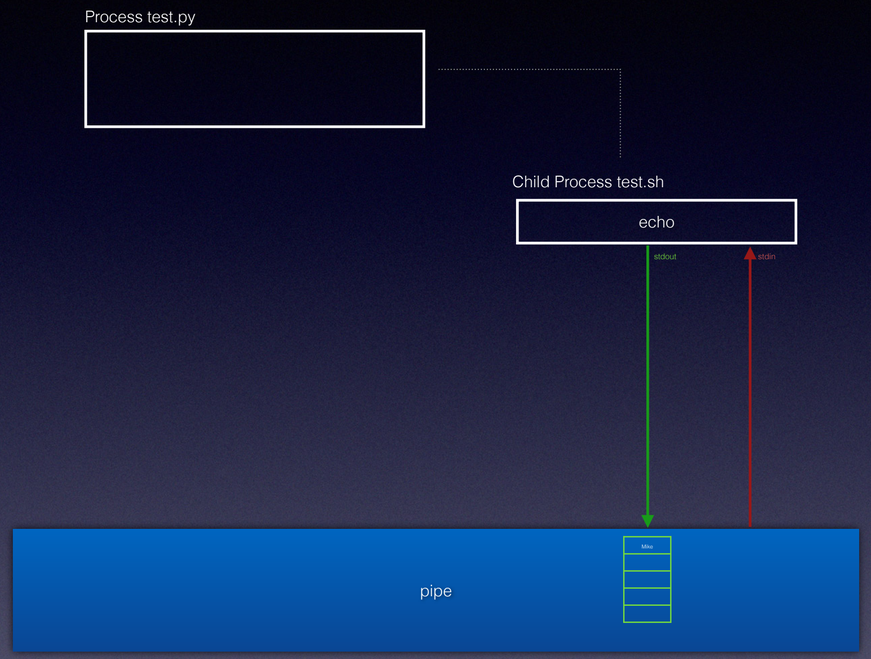
Async Reader

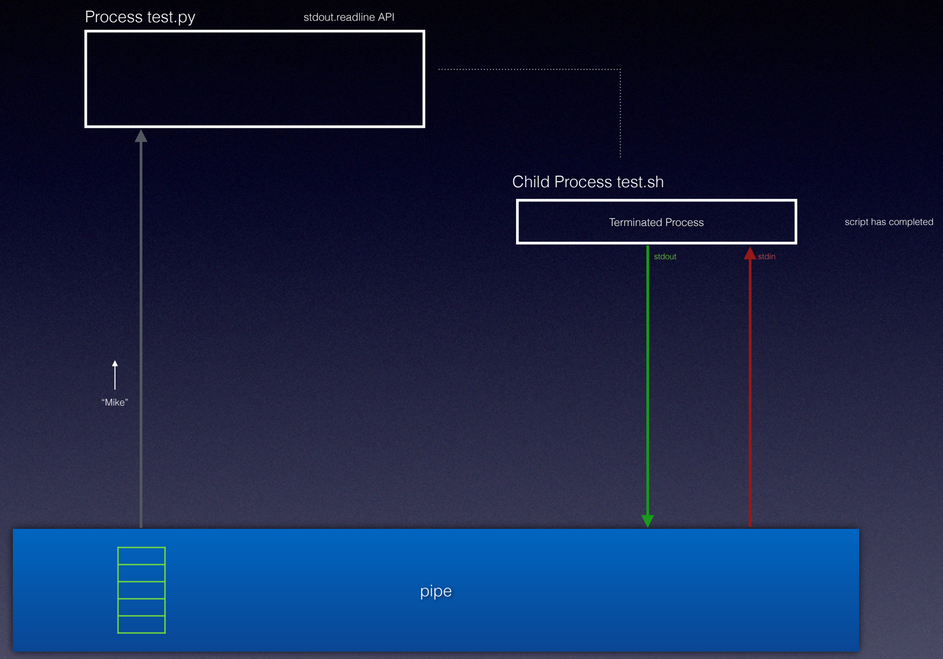

Process test.py
Child Process test.sh
echo
stdout
stdin
pipe
"User"
Name


stdout.Readline API
Process test.py
Waiting for parent to flush buffer for child
Child Process test.sh
read()
stdout
stdin
"User"
'Name'
Child flushes buffer (via read call)
pipe

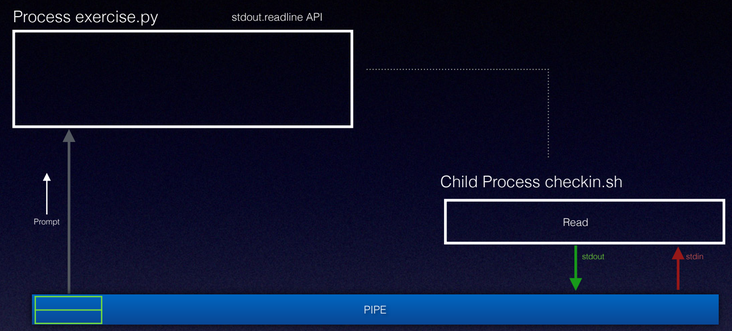

stdout.Write API
Process test.py
Waiting for parent to flush buffer for child
Child Process test.sh
read()
stdout
stdin
Mike
pipe

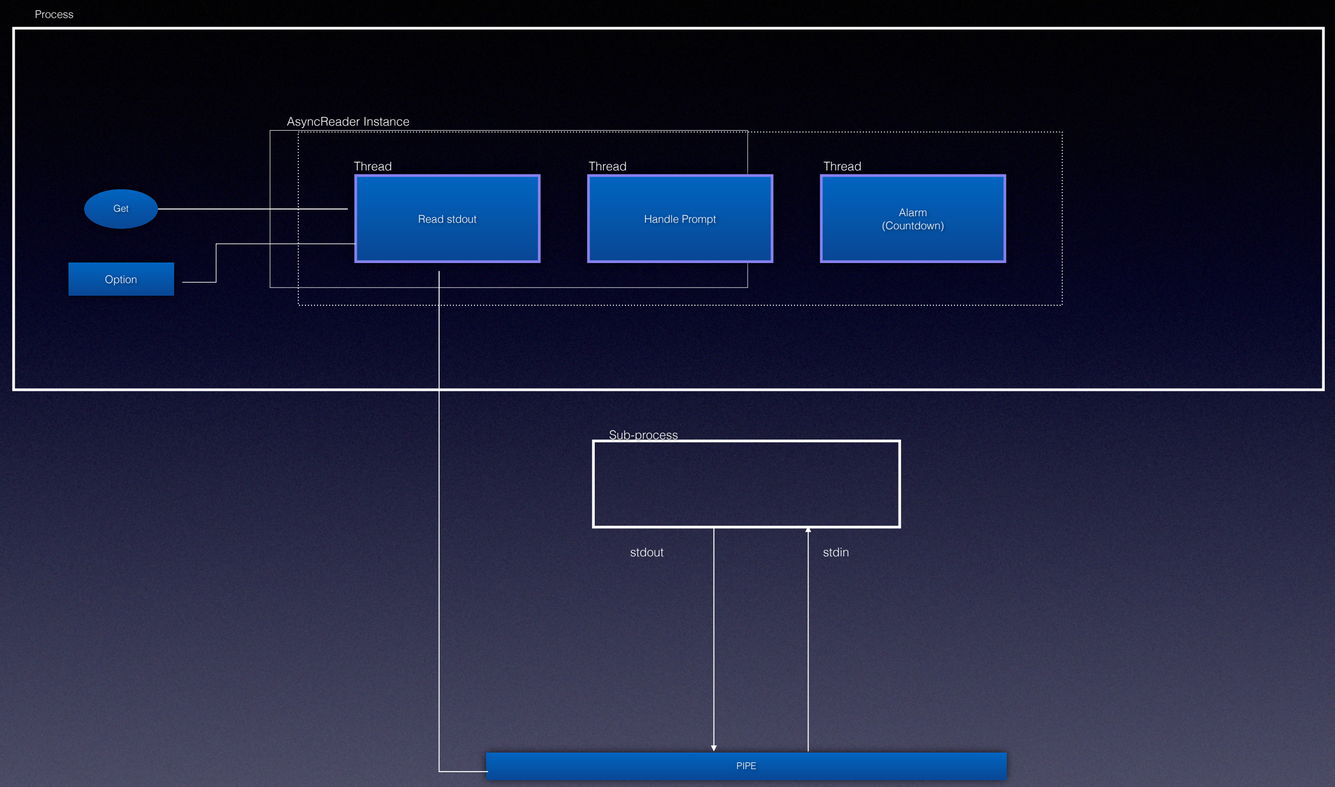

Stdin.flush API
Process test.py
Child Process test.sh
Read
stdout
stdin
"Mike"
Parent flushes buffer
pipe


Process test.py
Child Process test.sh
echo
stdout
stdin
Miku
pipe


Process test.py
stdout.readline API
Child Process test.sh
Terminated Process
script has completed
stdout
stdin
"Mike"
pipe

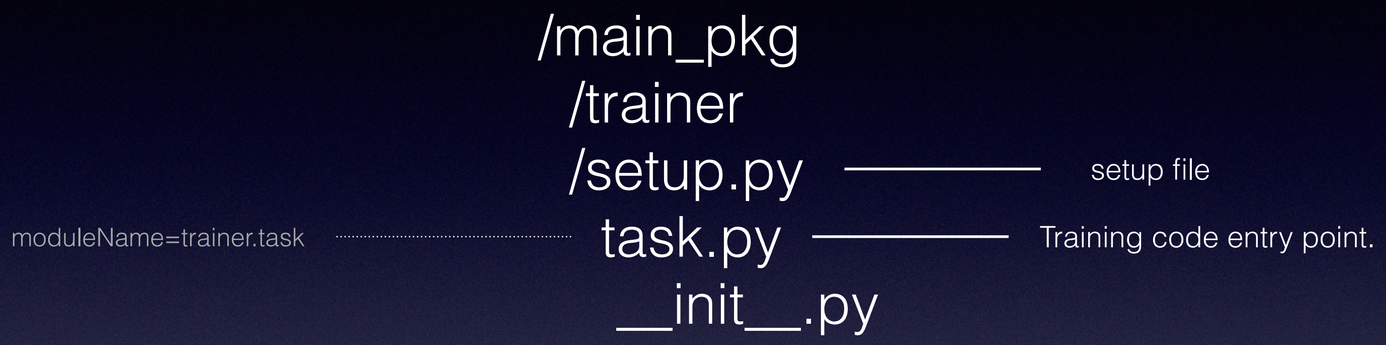

Process exercise_8.py
Child Process checkin.sh
echo
stdout
stdin
PIPE
Prompt

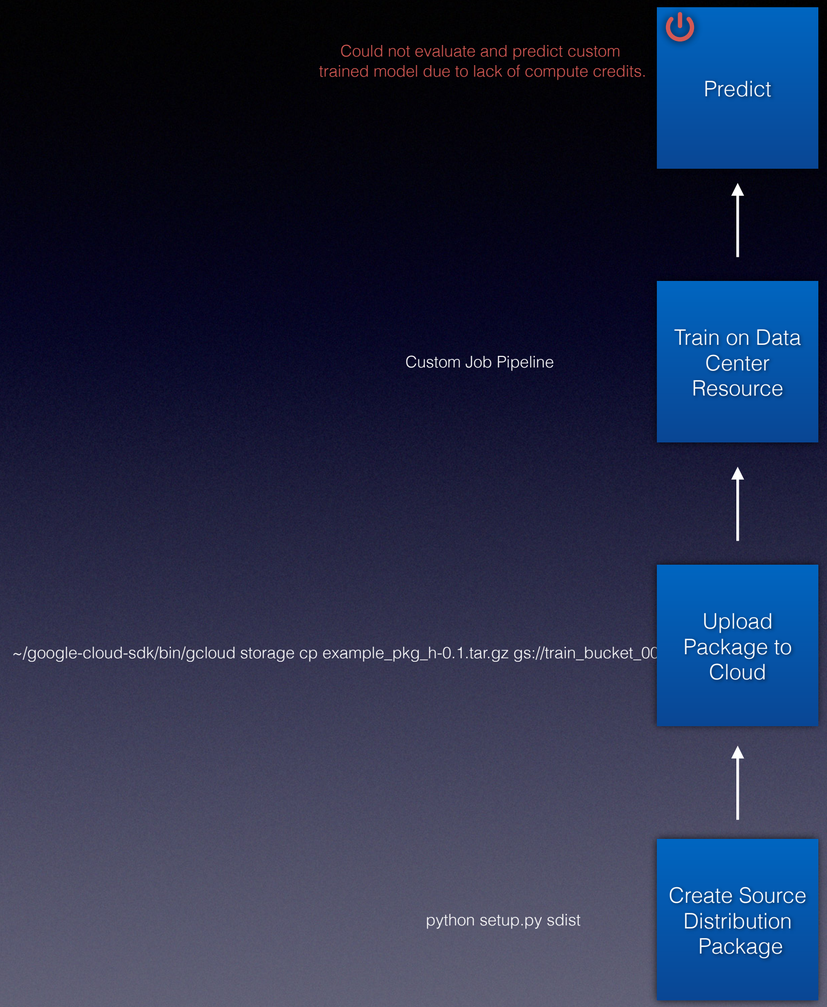

Process exercise.py
stdout.readline API
Prompt
Child Process checkin.sh
Read
stdout
stdin
PIPE


Process
AsyncReader Instance
Thread
Thread
Thread
Get
Read stdout
Handle Prompt
Alarm
(Countdown)
Option
Sub-process
stdout
stdin
PIPE

Package for Vertex API

Tree


/main_pkg
/trainer
/setup.py
setup file
moduleName=trainer.task
task.py
Training code entry point.
__init__.py


Could not evaluate and predict custom trained model due to lack of compute credits.
Predict
Custom Job Pipeline
Train on Data Center Resource
~/google-cloud-sdk/bin/gcloud storage cp example_pkg_h-0.1.tar.gz gs://train_bucket_00
Upload Package to Cloud
python setup.py sdist
Create Source Distribution Package