

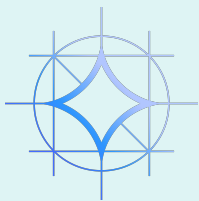
Inside Open LLM Kubernetes Deployment

Olga Mirensky - Platform Engineer - ANZx

May 2024

Agenda

1. Model



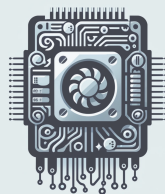
2. Inference
and serving



3. k8s
Deployment
Requirements

requests:
cpu: 8
memory: 29Gi
ephemeral: 80Gi
nvidia.com/gpu: 1

4. Cluster
Hardware



5. Demo



6. What's next
and Q&A



LLM

Meet The Star



google/gemma-2b-it like 488

Text Generation Transformers Safetensors GGUF gemma conversational Inference Endpoints text-generation-inference 23 papers License: gemma

Model card Files and versions Community 41

Edit model card

Gated model You have been granted access to this model

Gemma Model Card

Model Page: [Gemma](#)

This model card corresponds to the 2B instruct version of the Gemma model. You can also visit the model card of the [2B base model](#), [7B base model](#), and [7B instruct model](#).

Resources and Technical Documentation:

- [Responsible Generative AI Toolkit](#)
- [Gemma on Kaggle](#)
- [Gemma on Vertex Model Garden](#)

Downloads last month
451,501

Safetensors Model size 2.51B params Te

Inference API

Text Generation

Input a message to start chatting with

Sure, here's a simplified overview of how the brain processes information: The brain receives information from the senses (eyes, ears, nose, mouth) and the nervous system. This information is then processed and interpreted by different areas of the brain.

Inference API (serverless)
Serverless inference for prototyping

Inference Endpoints (dedicated)
Inference deployments for production


Amazon SageMaker
Deploy with SageMaker


Azure ML
Deploy with AzureML



Google Cloud
Deploy with Google Cloud

Spaces
Deploy as a Gradio app in one click

What's inside?


 **Hugging Face**



















[Models](#) [Datasets](#) [Spaces](#) [Posts](#) [Docs](#) [Solutions](#) [Pricing](#) [⌵](#) 

google / **gemma-2b-it**   like 489

[Text Generation](#) [Transformers](#) [Safetensors](#) [GGUF](#) [gemma](#) [conversational](#) [Inference Endpoints](#) [text-generation-inference](#) [23 papers](#) [License: gemma](#)

[Model card](#) [Files and versions](#) [Community 41](#) [⋮](#) [Train](#) [Deploy](#) [Use in Transformers](#)

[main](#) [gemma-2b-it](#)  13 contributors [History: 16 commits](#) [+ Contribute](#)

 pcuenq HF STAFF Fix tokenizer (#40) de144fb VERIFIED			4 days ago
.gitattributes	1.62 kB		Squashing commit about 2 months ago
README.md	23.6 kB		Update benchmark scores and averages for 1.1 9 days ago
config.json	627 Bytes		Squashing commit about 2 months ago
gemma-2b-it.gguf	10 GB  LFS		Squashing commit about 2 months ago
generation_config.json	137 Bytes		Squashing commit about 2 months ago
model-00001-of-00002.safetensors	4.95 GB  LFS		Squashing commit about 2 months ago
model-00002-of-00002.safetensors	67.1 MB  LFS		Squashing commit about 2 months ago
model.safetensors.index.json	13.5 kB		Squashing commit about 2 months ago
special_tokens_map.json	636 Bytes		Fix tokenizer (#40) 4 days ago
tokenizer.json	17.5 MB  LFS		Fix tokenizer (#40) 4 days ago
tokenizer.model	4.24 MB  LFS		Upload tokenizer.model about 2 months ago
tokenizer_config.json	34.2 kB		Fix tokenizer (#40) 4 days ago

We've got the model, now what?

Inference and serving

And a million other platform-y things

- Gateway
- Load Balancing
- HA
- Resiliency
- Cost efficiency
- Resource management
- Caching
- Components upgrade
- Model updates
- Monitoring
- Logging
- Security & compliance
- Performance optimization

Minimal k8s Deployment

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: vllm-gemma-deployment
spec:
  replicas: 1
  selector:
    matchLabels:
      app: gemma-server
  template:
    metadata:
      labels:
        app: gemma-server
        ai.gke.io/model: gemma-2b-it
        ai.gke.io/inference-server: vllm
        examples.ai.gke.io/source: user-guide
    spec:
      containers:
        - name: inference-server
          image: us-docker.pkg.dev/vertex-ai/vertex-vision-model-garden-dockers/pytorch-vllm-serve:20240220_0936_RC01
          resources:
            requests:
              cpu: "2"
```

// demo deploy,
// bare minimum

// more advanced
// options, comments

Kubernetes Deployment


```
ai.gke.io/model: gemma-2b-it
ai.gke.io/inference-server: vllm
examples.ai.gke.io/source: user-guide
spec:
  containers:
    - name: inference-server
      image:
us-docker.pkg.dev/vertex-ai/vertex-vision-model-garden-d
ockers/pytorch-vllm-serve:20240220_0936_RC01
      resources:
        requests:
          cpu: "2"
          memory: "7Gi"
          ephemeral-storage: "10Gi"
          nvidia.com/gpu: 1
        limits:
          cpu: "2"
          memory: "7Gi"
          ephemeral-storage: "10Gi"
          nvidia.com/gpu: 1
      command: ["python3", "-m",
"vllm.entrypoints.api_server"]
      args:
        - --model=$(MODEL_ID)
        - --tensor-parallel-size=1
        - --dtype=half
```

Kubernetes Deployment

```

    ephemeral-storage: "10Gi"
    nvidia.com/gpu: 1
    command: ["python3", "-m",
"vllm.entrypoints.api_server"]
    args:
    - --model=$(MODEL_ID)
    - --tensor-parallel-size=1
    - --dtype=half
    env:
    - name: MODEL_ID
      value: google/gemma-2b-it
    - name: HUGGING_FACE_HUB_TOKEN
      valueFrom:
        secretKeyRef:
          name: hf-secret
          key: hf_api_token
    volumeMounts:
    - mountPath: /dev/shm
      name: dshm
    volumes:
    - name: dshm
      emptyDir:
        medium: Memory
    nodeSelector:

```

Kubernetes Deployment

initContainers:

- command: [aws s3 cp s3://mdl/ ...]

initContainers:

- command: [gsutil cp gs://mdl/ ...]

Cloud storage, OCI-compatible
registry => better integration,
IAM, etc

Kubernetes Deployment

```
- name: MODEL_ID
  value: google/gemma-2b-it
- name: HUGGING_FACE_HUB_TOKEN
  valueFrom:
    secretKeyRef:
      name: hf-secret
      key: hf_api_token
volumeMounts:
- mountPath: /dev/shm
  name: dshm
volumes:
- name: dshm
  emptyDir:
    medium: Memory
nodeSelector:
  cloud.google.com/gke-accelerator: nvidia-tesla-t4
```

PVC

```
spec:
  accessModes:
    - ReadWriteMany
  resources:
    requests:
      storage: 100Gi
```

Kubernetes Cluster

```
gcloud container node-pools create gpupool \  
... \  
  --accelerator \  
"type=nvidia-tesla-t4,gpu-driver-version=latest, \  
count=1" \  
  --machine-type "n1-standard-8" \  
  --
```

```
apiVersion: eksctl.io/v1alpha5 \  
kind: ClusterConfig \  
... \  
nodeGroups: \  
  - name: gpu-nodegroup \  
    instanceType: p3.2xlarge
```

Attachable GPU

GCP: General purpose - n1, (GPU T4, P4, V100)

Integrated GPU

GCP: Accelerator-optimized. a2, a3, g2 (GPU: L4, A100, H100)

AWS: P- and G- Series

Demo

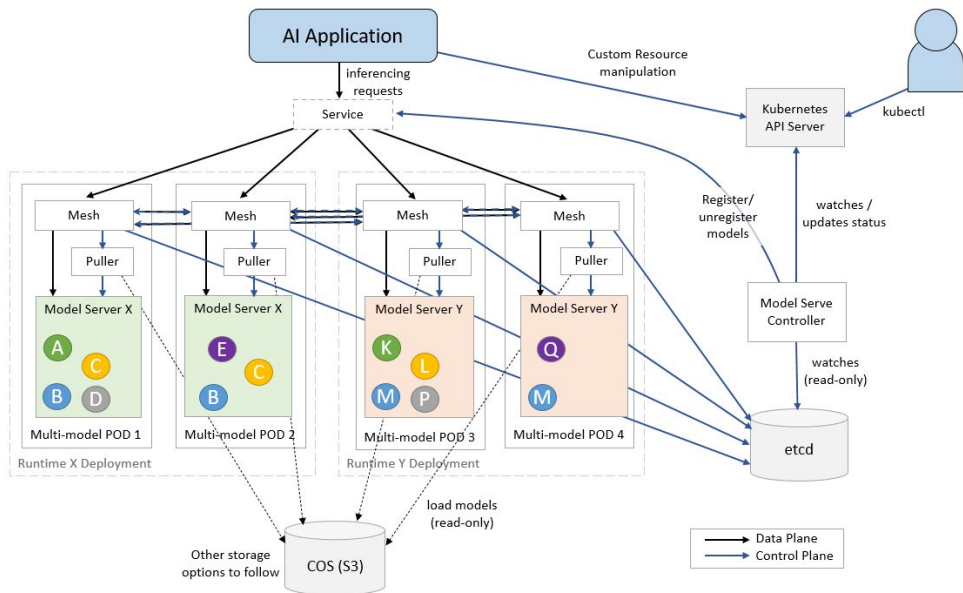
Demo - Interacting with the Model

```
+ USER_PROMPT='Share some mind-blowing statistic related to running an OSS LLM on a
kubernetes cluster'

+ curl -X POST http://localhost:8000/generate -H 'Content-Type: application/json' -d @-
{"predictions":["Prompt:\n<start_of_turn>user\nShare some mind-blowing statistic related
to running an OSS LLM on a kubernetes cluster<end_of_turn>\nOutput:\n**The numbers don't
lie:**\n\n* **12.6 petabytes of data** was processed and trained by the LLM.\n* **5.7
trillion parameters** were learned by the LLM.\n* **384,000 training steps** were run to
build the model.\n* **1,000+ cores** were used to train the LLM.\n* **60+ different
languages** were supported by the LLM.\n* **40% of the world's knowledge** could fit on
the LLM's hard drive." ]}%
```

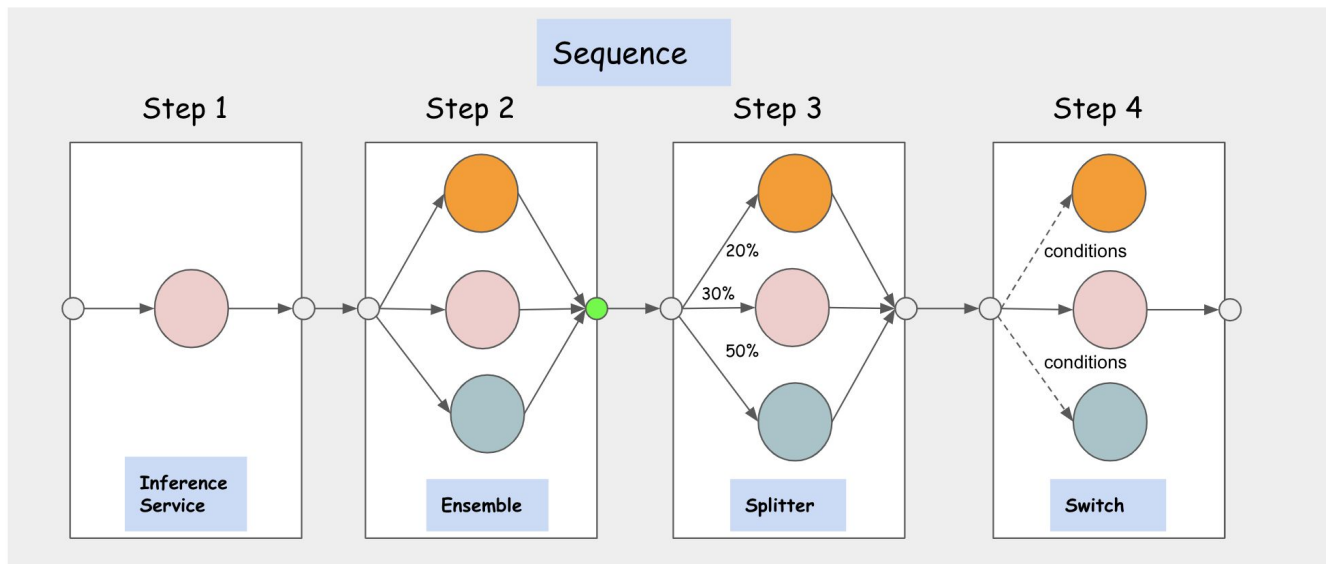
KServe

KServe - Model Mesh



<https://kserve.github.io/website/latest/model-serving/mms/modelmesh/overview/>

KServe



https://kserve.github.io/website/latest/modelserving/inference_graph

Do I need a Mortgage to Experiment Using AI?

Disclaimer: This is not endorsement/recommendation.

- Quotas
- Billing budgets alerts
- Spot provisioning
- Estimate with GenAI chat and walk through official docs

Can  you provide estimated Compute Engine costs for 1 hour running following GKE nodepool

```
...  
container node-pools create gpupool-t4 \  
    --accelerator type=nvidia-tesla-t4,count=1 \  
    --machine-type=n2-standard-32 \  
    --num-nodes=1 --spot  
...
```

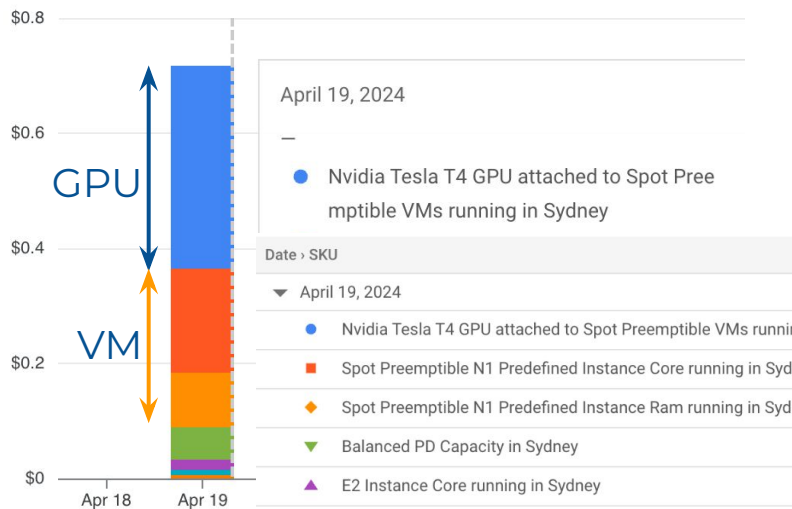
. . .



Total Estimated Cost per Hour: \$0.4387 (VM) + \$0.105 (GPU) = ~~\$0.5437~~ per hour

After SPOT discount

Actual Cost - 90 min 2B model



GKE cluster with 1 `n1-standard-8` node + Nvidia T4 GPU, spot.

Date > SKU	Service	SKU ID	Usage	Cost
▼ April 19, 2024				
● Nvidia Tesla T4 GPU attached to Spot Preemptible VMs running in Sydney	Compute Engine	C932-A161-E489	1.54 hour	\$0.35
■ Spot Preemptible N1 Predefined Instance Core running in Sydney	Compute Engine	1BE3-7C1B-4E3E	12.33 hour	\$0.18
◆ Spot Preemptible N1 Predefined Instance Ram running in Sydney	Compute Engine	2E28-1B91-432D	46.23 gibibyte hour	\$0.09
▼ Balanced PD Capacity in Sydney	Compute Engine	B046-9075-75BA	0.27 gibibyte month	\$0.06
▲ E2 Instance Core running in Sydney	Compute Engine	9819-4AB8-1A87	0.38 hour	\$0.02
■ E2 Instance Ram running in Sydney	Compute Engine	F9E1-E6EF-86F3	1.53 gibibyte hour	\$0.01
✦ External IP Charge on a Spot Preemptible VM	Compute Engine	4AF8-7C1F-39C4	1.54 hour	\$0.01
✱ Network Internet Data Transfer Out from Sydney to Australia	Compute Engine	65C5-505D-97C1	0.01 gibibyte	\$0.00
▼ Network Inter Region Data Transfer In from Finland to Sydney	Compute Engine	2B3A-0796-17C1	0 gibibyte	\$0.00
◆ Network Internet Data Transfer In from China to Sydney	Compute Engine	231A-EB04-2720	0 gibibyte	\$0.00

[MORE RESULTS](#)

What's next

Advanced Serving And Deployment

KServe
Skypilot

Newsletter

TLDR +

Cloud ML/AI Hub

<https://aws.amazon.com/machine-learning/learn>

Explore 🤗

- Browse models*
- Read blogs
- Explore

(*model access is gated)

Run locally + RAG

<https://github.com/binchenX/rag-up>

Q & A