How Cookpad Leverages Triton Inference Server To Boost Their Model Serving

cookpad

100m +

6m +

30 +

70+

In season now

Spinach

Potato

Purple Sprouting Broccoli

Rhubarb

Rosemary

Thyme

Pollock

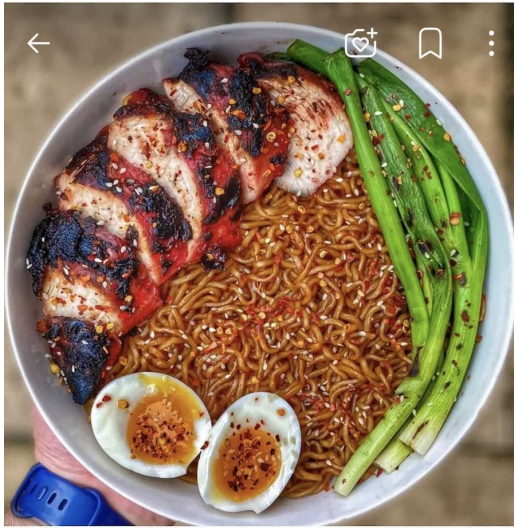
Wild Garlic

Spring Onion

Nettle

Spring Greens

Crab



Char Siu Chicken w/ Hong Kong Style Noodles

Craig Stokes

@whatcraigcooks

Manchester, England, United King...

A Chinese inspired chicken with Hong Kong style noodles, great for any evening

Origin: China

40 minutes

1 cooksnap

Search recipes, ingredients or tips

Today's popular searches

Side dishes

Ukraine

Hong kong

Pasta bake

Spicy food

Fish recipes

Cookies

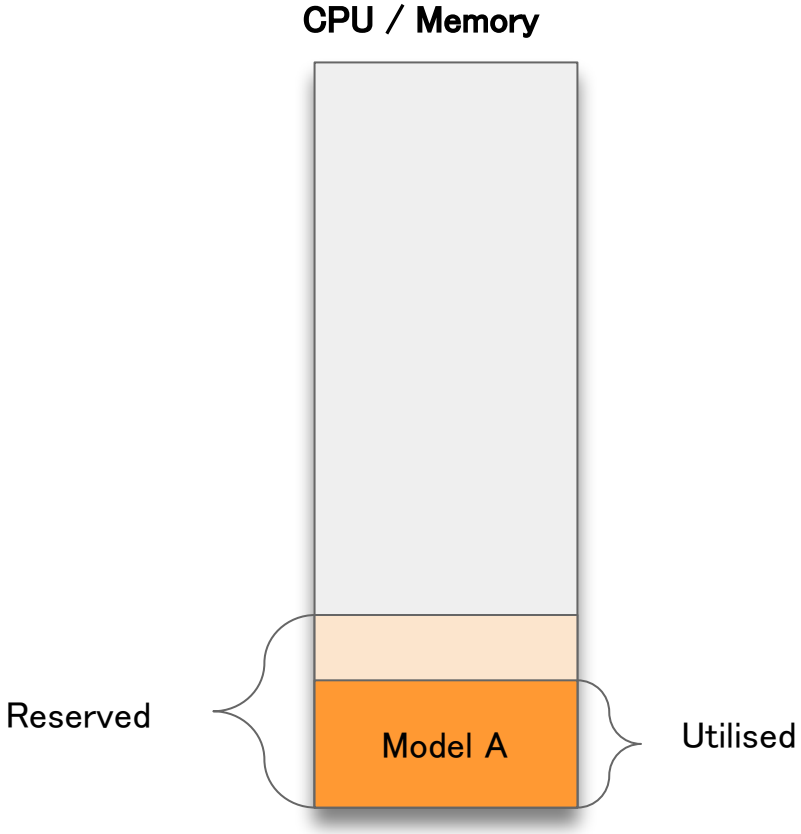
Puff pastry

Chicken breast

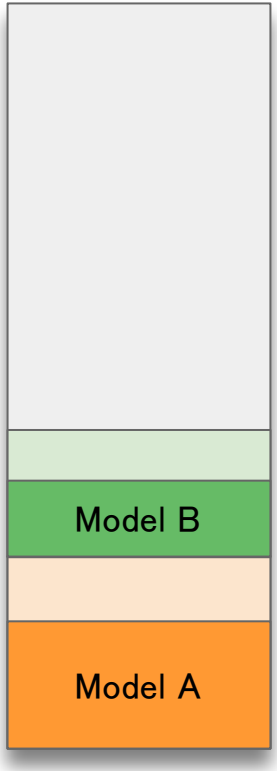
Cauliflower

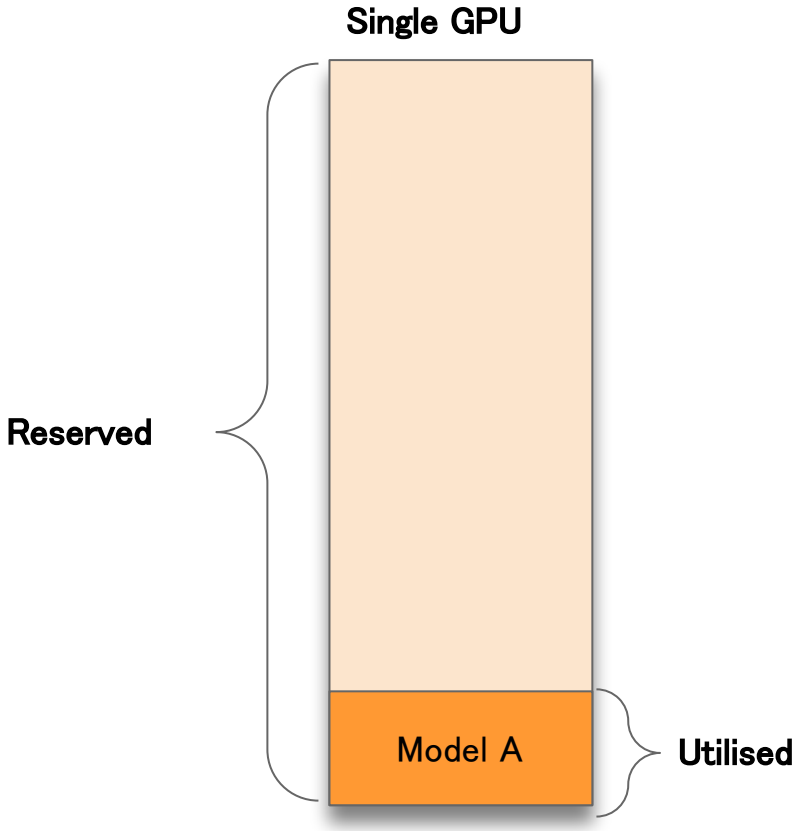
Improving your model deployment with Triton Inference Server

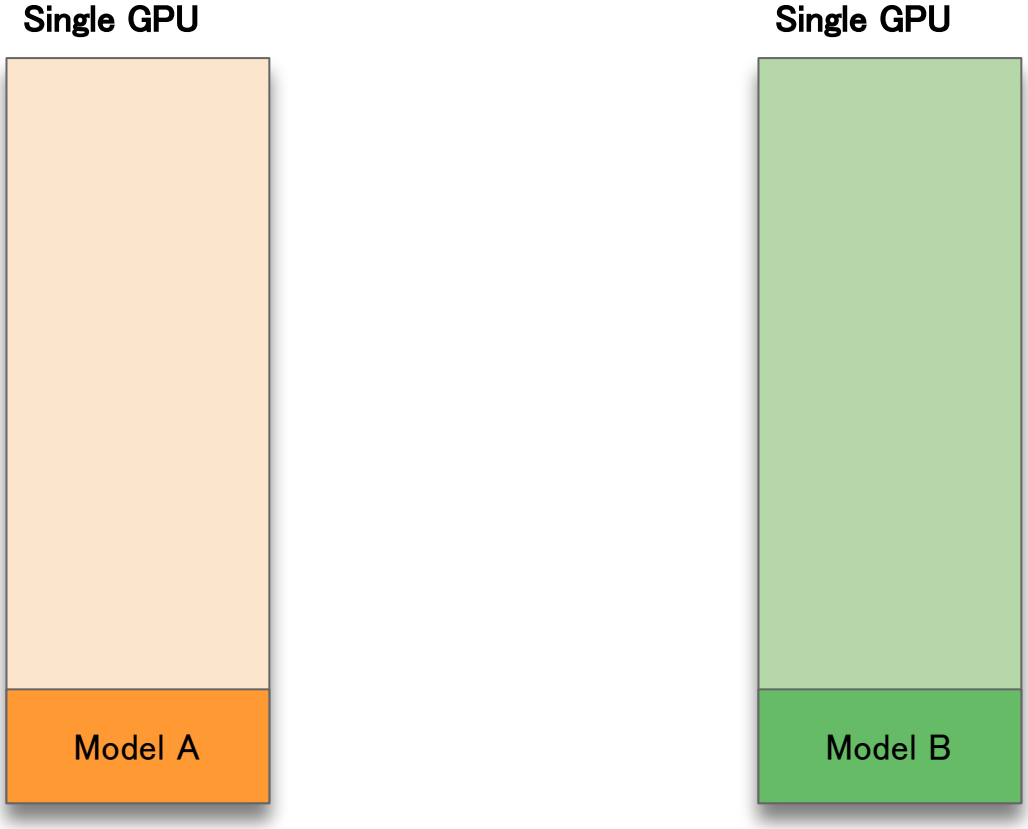
cookpad



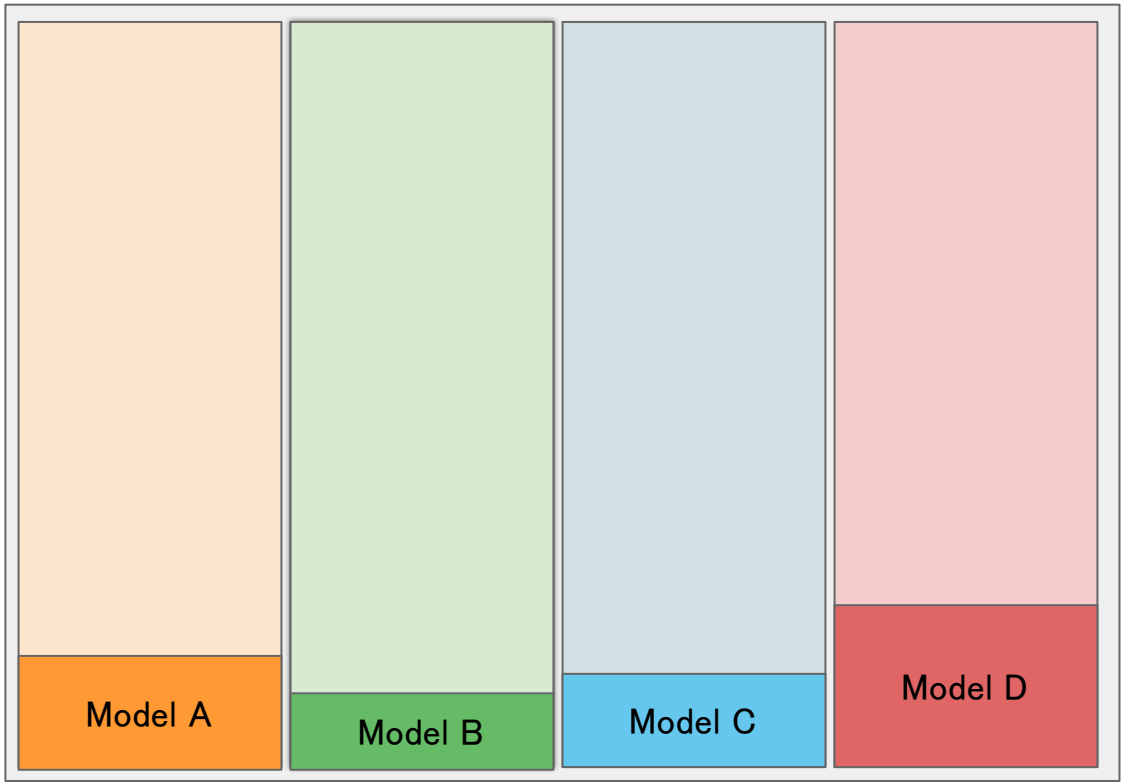
CPU / Memory



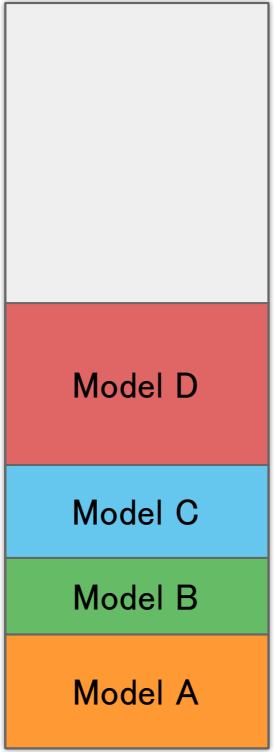




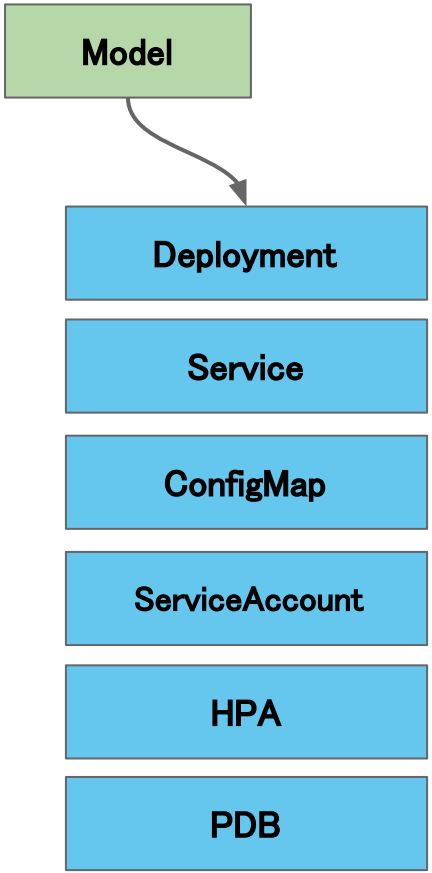
Multiple GPUs

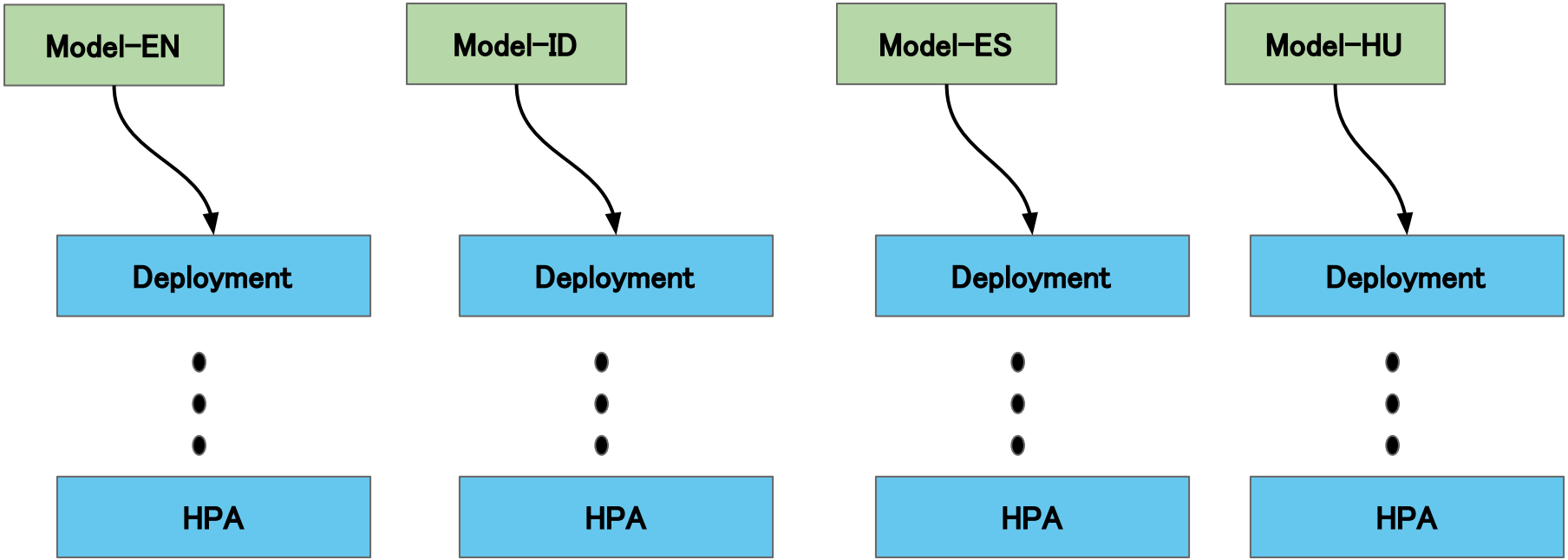


Triton on single GPU

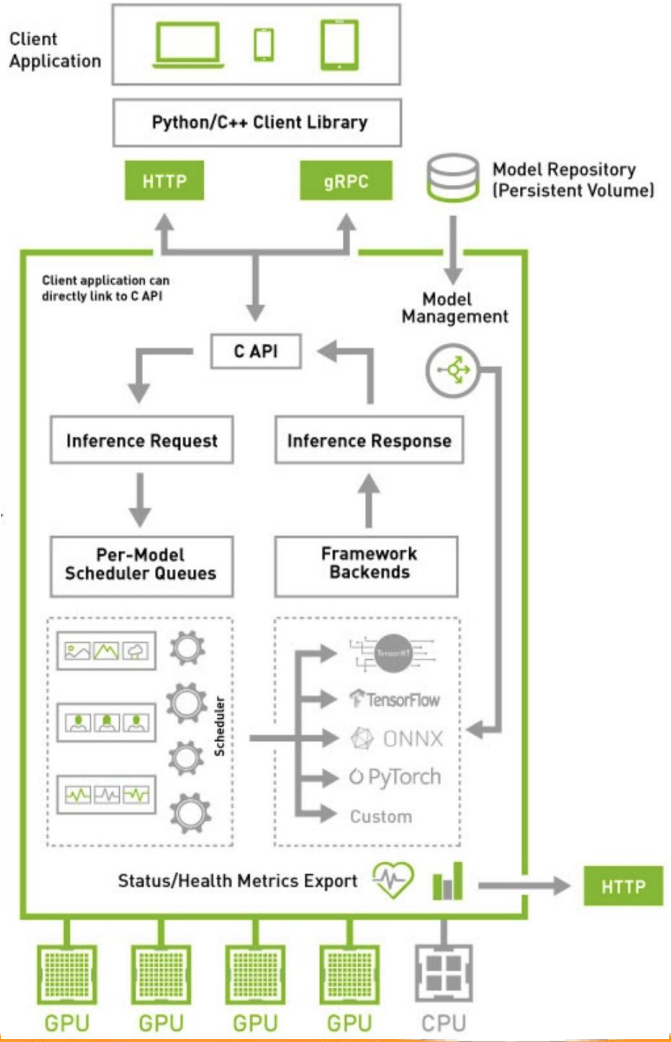


Triton Multiple GPUs

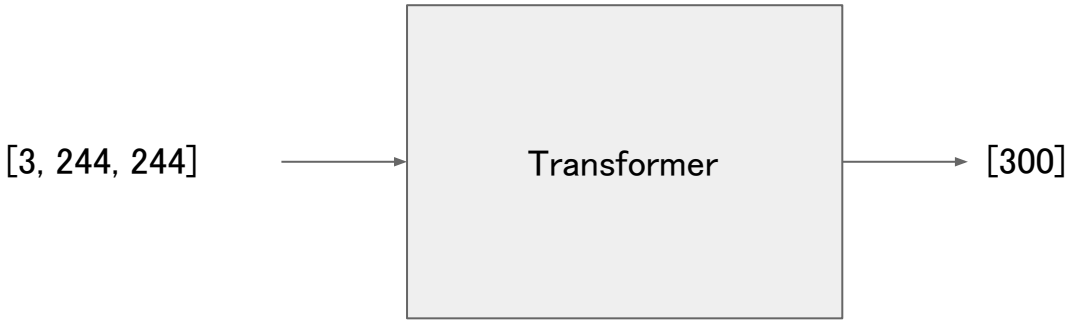


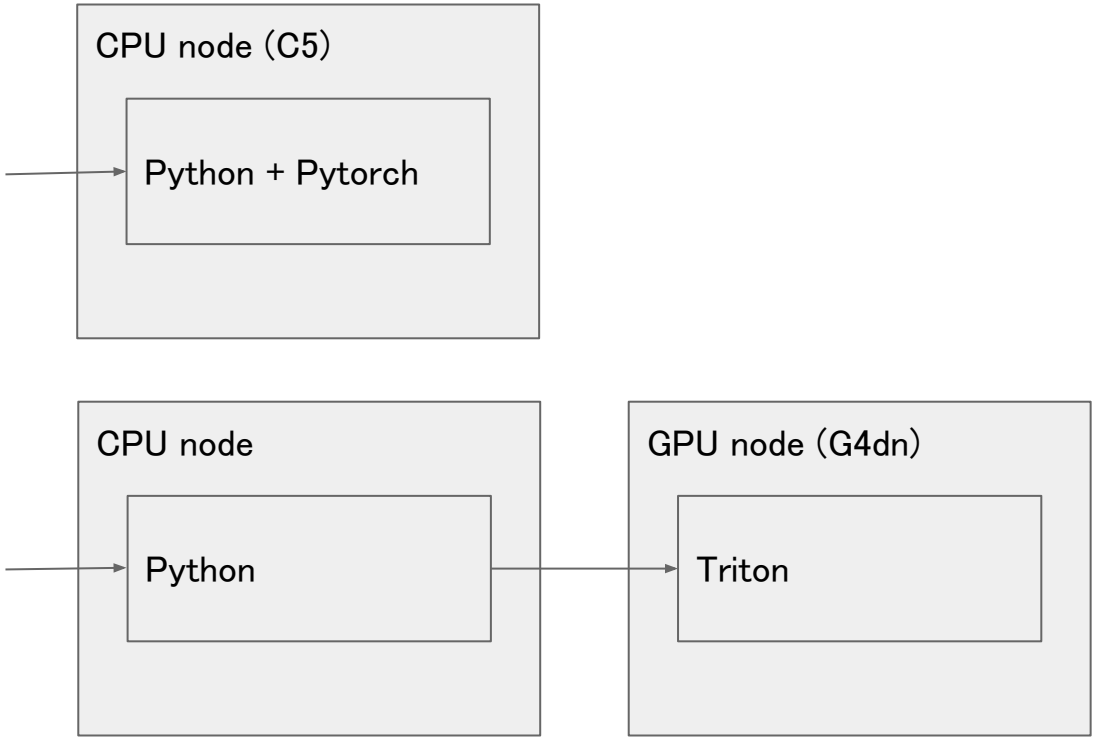


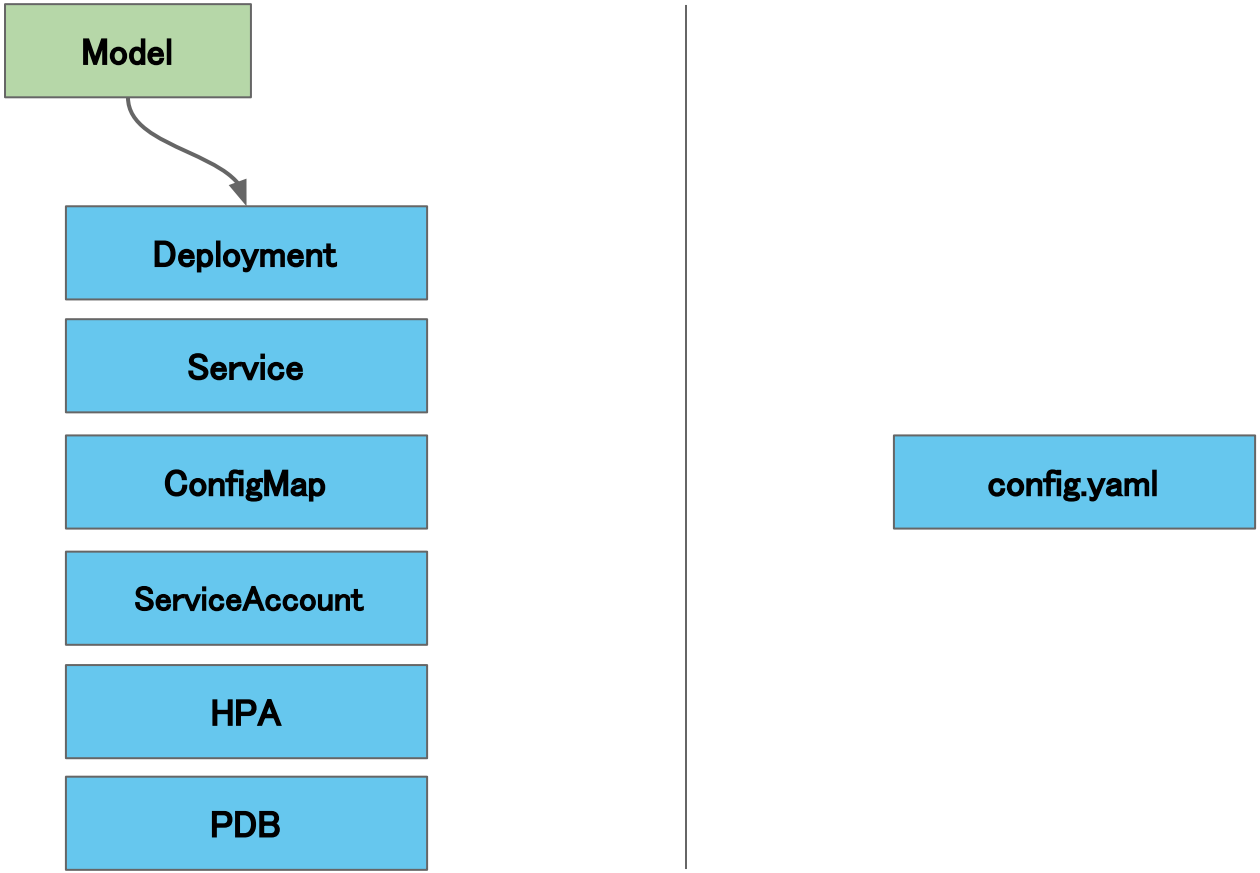

```
local_model_dir: /models
model_files:
  - text-embeddings/en/model.tar.gz
  - text-embeddings/es/model.tar.gz
  - text-embeddings/id/model.tar.gz
  - text-embeddings/hu/model.tar.gz
  - image-embeddings/model.tar.gz
  - categorisations/optimized/model.tar.gz
  - onnx_xlm_roberta_base.tar.gz
s3_bucket: <bucket_name>
s3_models_dir: deployment-artifacts
```

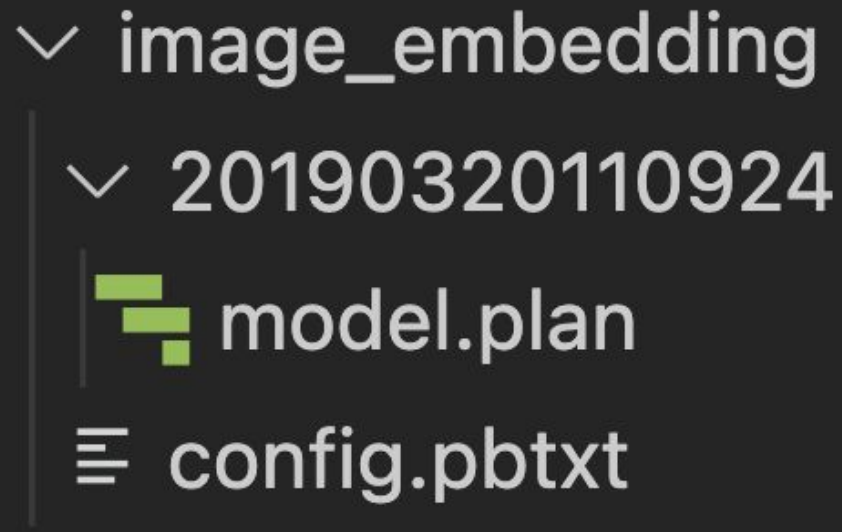



cookpad








image_embedding
20190320110924
model.plan
config.pbtxt

config.pbtxt

image_embedding > config.pbtxt

```
platform: "tensorrt_plan"
max_batch_size: 1
input [
  {
    name: "tensor"
    data_type: TYPE_FP32
    dims: [ 3, 224, 224 ]
  }
]
output [
  {
    name: "1930"
    data_type: TYPE_FP32
    dims: [ 300 ]
  }
]
```

config.yaml

```yaml
local_model_dir: /models
model_files:
  - image_embedding.tar.gz
s3_bucket: ember.us-east-1
s3_models_dir: packages

```

root@ubuntu: / (kubectl)

```
Processing:   257  272  14.1    269     337
Waiting:      254  272  14.1    269     337
Total:        257  272  14.1    269     337

Percentage of the requests served within a certain time (ms)
  50%    269
  66%    273
  75%    277
  80%    279
  90%    286
  95%    297
  98%    327
  99%    337
 100%    337 (longest request)
root@ubuntu:/#
```

root@ubuntu: / (kubectl)

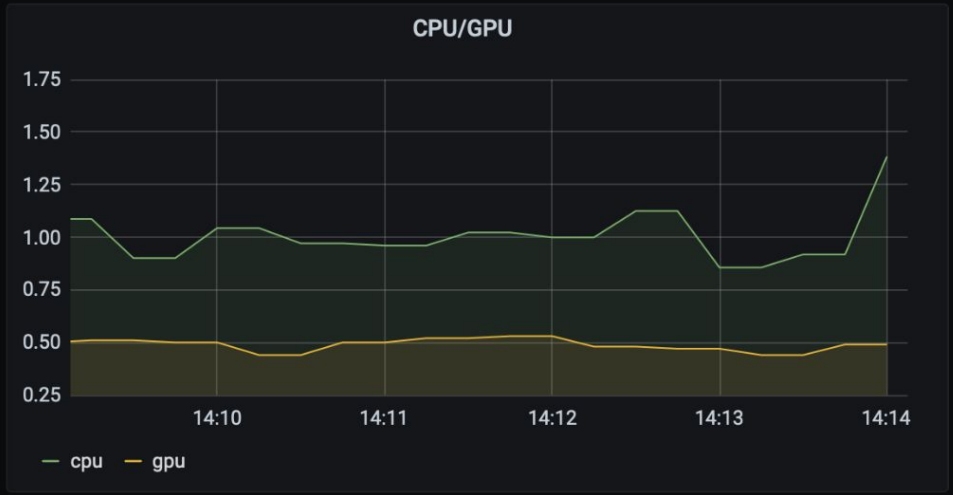
```
Processing:    17   18   1.0     18      22
Waiting:       17   18   0.9     18      22
Total:         17   19   1.0     19      23

Percentage of the requests served within a certain time (ms)
  50%     19
  66%     19
  75%     19
  80%     20
  90%     20
  95%     21
  98%     23
  99%     23
 100%     23 (longest request)
root@ubuntu:/#
```

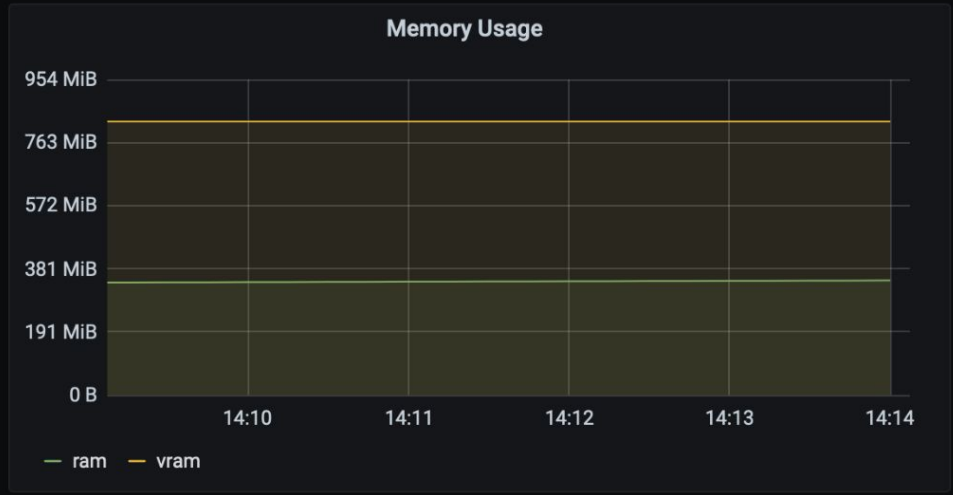
root@ubuntu: / (kubectl)

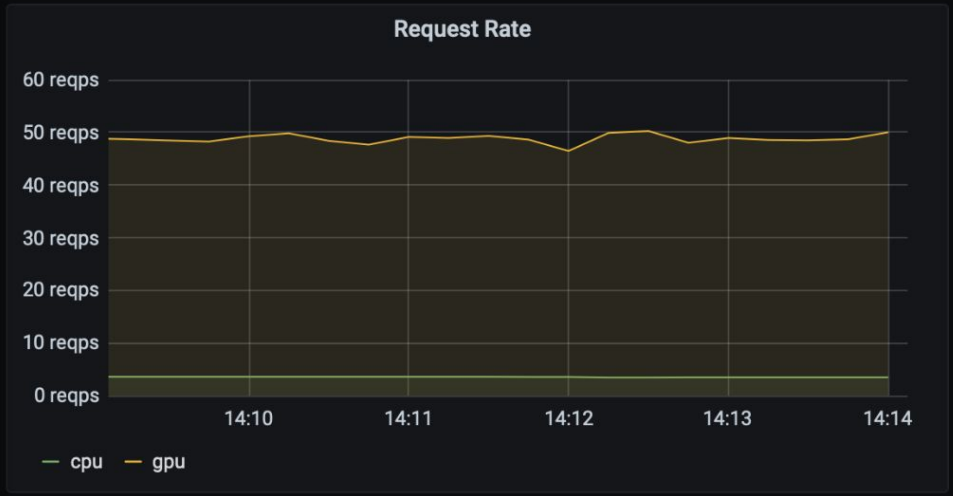
```
Processing:   269  538  43.0    538     631
Waiting:      268  537  42.7    538     631
Total:        269  538  43.0    538     631

Percentage of the requests served within a certain time (ms)
  50%    538
  66%    549
  75%    552
  80%    553
  90%    568
  95%    576
  98%    631
  99%    631
 100%    631 (longest request)
root@ubuntu:/#
```

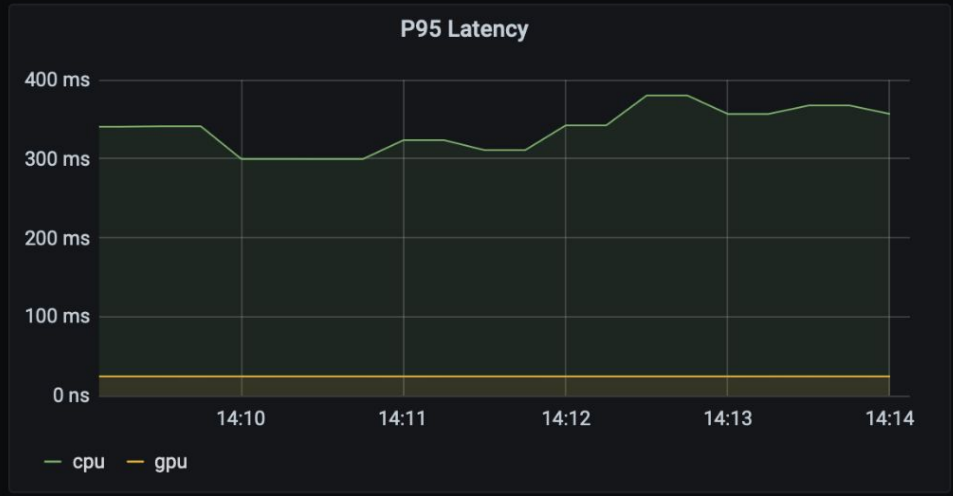
root@ubuntu: / (kubectl)

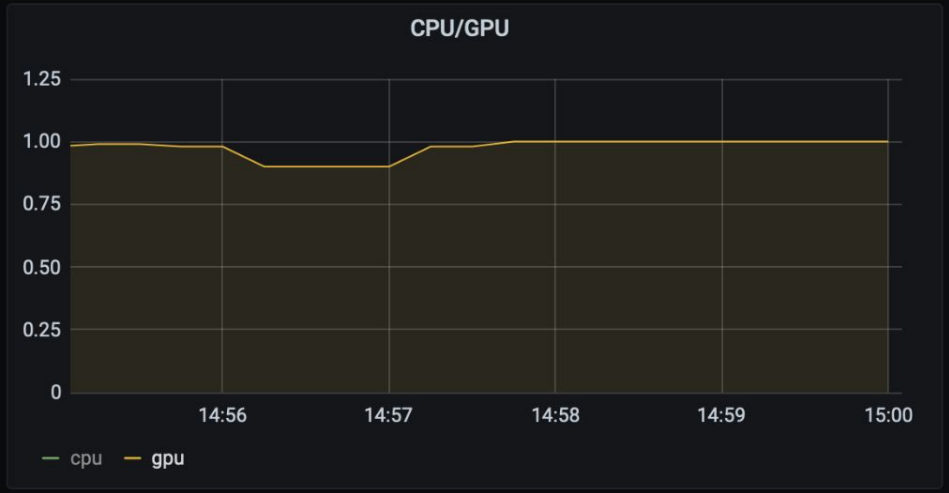
```
Total:        18   20   1.6     19      29
ERROR: The median and mean for the initial connection time are more than twice the standard
       deviation apart. These results are NOT reliable.

Percentage of the requests served within a certain time (ms)
  50%     19
  66%     20
  75%     20
  80%     20
  90%     22
  95%     23
  98%     27
  99%     29
 100%     29 (longest request)
root@ubuntu:/#
```

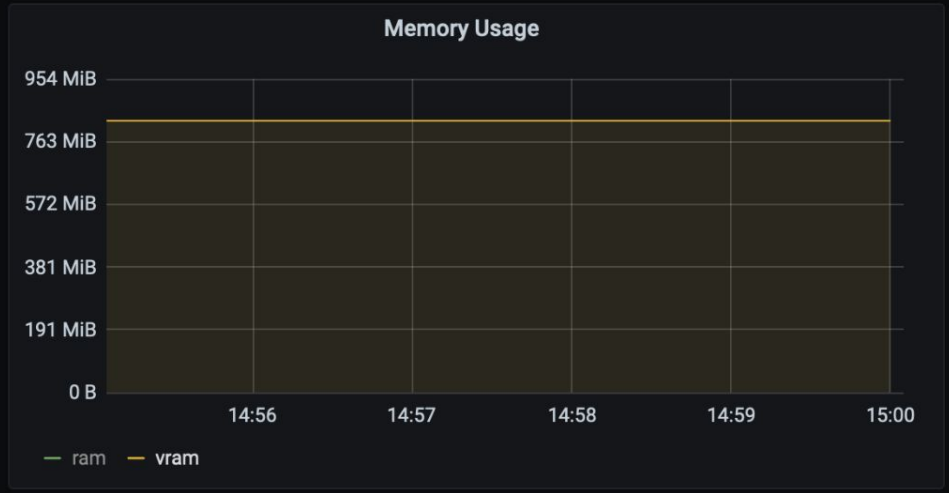

CPU/GPU
1.75
1.50
1.25
1.00
0.75
0.50
0.25
14:10
14:11
14:12
14:13
14:14
cpu
gpu

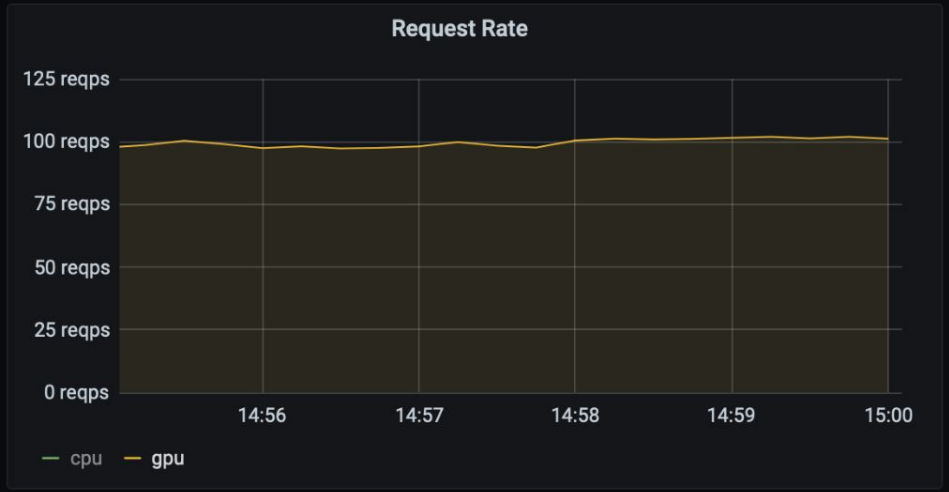

Memory Usage
954 MiB
763 MiB
572 MiB
381 MiB
191 MiB
0 B
14:10
14:11
14:12
14:13
14:14
ram
vram

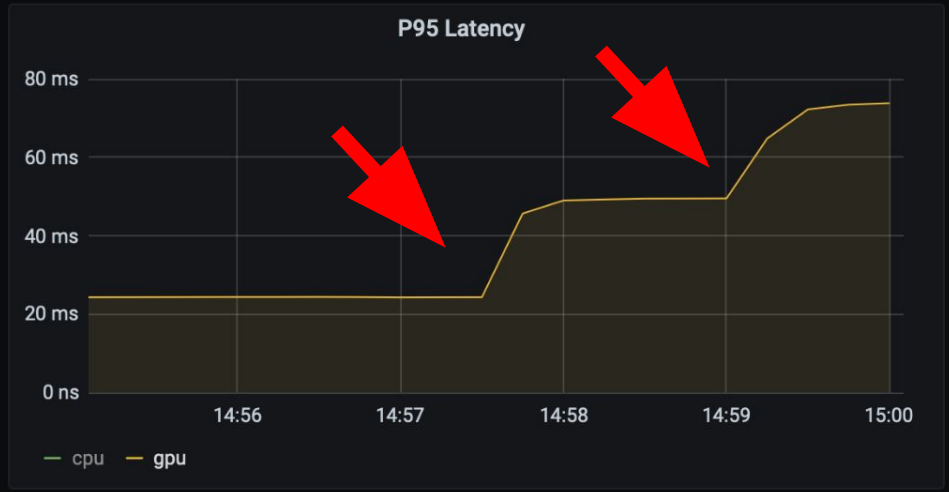

Request Rate
60 reqps
50 reqps
40 reqps
30 reqps
20 reqps
10 reqps
0 reqps
14:10
14:11
14:12
14:13
14:14
cpu
gpu


P95 Latency
400 ms
300 ms
200 ms
100 ms
0 ns
14:10
14:11
14:12
14:13
14:14
cpu
gpu


CPU/GPU
1.25
1.00
0.75
0.50
0.25
0
14:56
14:57
14:58
14:59
15:00
cpu
gpu


Memory Usage
954 MiB
763 MiB
572 MiB
381 MiB
191 MiB
0 B
14:56
14:57
14:58
14:59
15:00
ram
vram


Request Rate
125 reqps
100 reqps
75 reqps
50 reqps
25 reqps
0 reqps
14:56
14:57
14:58
14:59
15:00
cpu
gpu


P95 Latency
80 ms
60 ms
40 ms
20 ms
0 ns
14:56
14:57
14:58
14:59
15:00
cpu
gpu

config.pbtxt

models > image_embedding > config.pbtxt

```
platform: "tensorrt_plan"
max_batch_size: 8
input [...
]
output [...
]
instance_group [
  {
    count: 2
    kind: KIND_GPU
    gpus: [ 0 ]
  }
]
```

config.pbtxt

models > image_embedding > config.pbtxt

```
platform: "tensorrt_plan"
max_batch_size: 8
input [...
]
output [...
]
dynamic_batching {
  preferred_batch_size: [ 4, 8 ],
  max_queue_delay_microseconds: 100
}
```

Q & A

We are hiring

https://careers.cookpad.com

cookpad