* 1. Given example

**Server**

model=/data/models--meta-llama--Llama-2-7b-hf

volume=/opt/models/

docker run -p 8080:80 -v $volume:/data --runtime=habana -e HABANA\_VISIBLE\_DEVICES=all -e OMPI\_MCA\_btl\_vader\_single\_copy\_mechan

ism=none -e BATCH\_BUCKET\_SIZE=16 --cap-add=sys\_nice --ipc=host ghcr.io/huggingface/tgi-gaudi:1.2.1 --model-id $model

MODEL\_ID=/opt/models/models--meta-llama--Llama-2-7b-hf/

INFO text\_generation\_launcher: Args { model\_id: "/data/models--meta-llama--Llama-2-7b-hf", revision: None, validation\_workers: 2, sharded: None, num\_shard

: None, quantize: None, dtype: None, trust\_remote\_code: false, max\_concurrent\_requests: 128, max\_best\_of: 2, max\_stop\_sequences: 4, max\_top\_n\_tokens: 5, max\_input\_length: 1024, max\_to

tal\_tokens: 2048, waiting\_served\_ratio: 1.2, max\_batch\_prefill\_tokens: 4096, max\_batch\_total\_tokens: None, max\_waiting\_tokens: 20, hostname: "91ffcc3ac849", port: 80, shard\_uds\_path:

"/tmp/text-generation-server", master\_addr: "localhost", master\_port: 29500, huggingface\_hub\_cache: Some("/data"), weights\_cache\_override: None, disable\_custom\_kernels: false, cuda\_me

mory\_fraction: 1.0, rope\_scaling: None, rope\_factor: None, json\_output: false, otlp\_endpoint: None, cors\_allow\_origin: [], watermark\_gamma: None, watermark\_delta: None, ngrok: false,

ngrok\_authtoken: None, ngrok\_edge: None, env: false }

Parameter explanation: <https://github.com/huggingface/tgi-gaudi/blob/habana-main/docs/source/basic_tutorials/launcher.md>

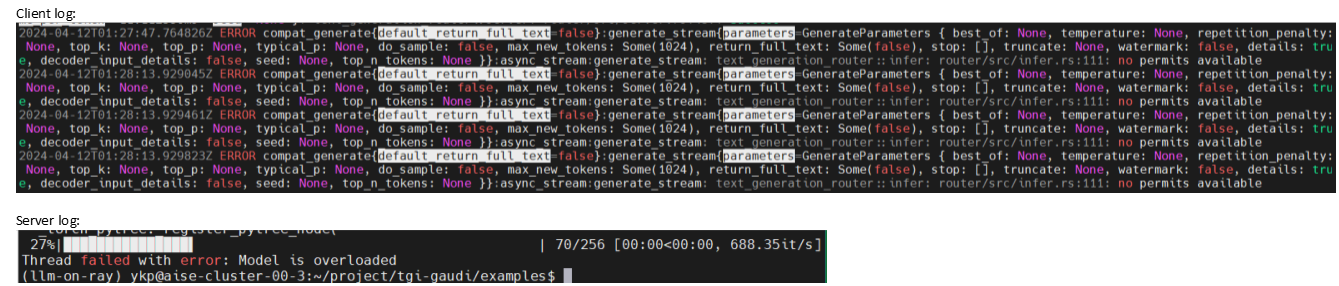
**Client**

python run\_generation.py --model\_id $MODEL\_ID --total\_sample\_count 128

文本

描述已自动生成

>128 will fail



* 1. Adjustment parameter BATCH\_BUCKET\_SIZE=4 (Parameter explanation: <https://github.com/huggingface/tgi-gaudi/blob/habana-main/README.md#running-tgi-on-gaudi>)

**Server**

docker run -p 8080:80 -v $volume:/data --runtime=habana -e HABANA\_VISIBLE\_DEVICES=all -e OMPI\_MCA\_btl\_vader\_single\_copy\_mechanism=none -e BATCH\_BUCKET\_SIZE=4 --cap-add=sys\_nice --ipc=host ghcr.io/huggingface/tgi-gaudi:1.2.1 --model-id $model

**Client**

python run\_generation.py --model\_id $MODEL\_ID --total\_sample\_count 128

电脑萤幕画面

描述已自动生成

* 1. Adjustment parameter BATCH\_BUCKET\_SIZE=8

**Server**

docker run -p 8080:80 -v $volume:/data --runtime=habana -e HABANA\_VISIBLE\_DEVICES=all -e OMPI\_MCA\_btl\_vader\_single\_copy\_mechan

ism=none -e BATCH\_BUCKET\_SIZE=8 --cap-add=sys\_nice --ipc=host ghcr.io/huggingface/tgi-gaudi:1.2.1 --model-id $model

**Client**

python run\_generation.py --model\_id $MODEL\_ID --total\_sample\_count 128

文本

描述已自动生成

>128 will also fail

* 1. Adjustment parameter BATCH\_BUCKET\_SIZE=32

**Server**

docker run -p 8080:80 -v $volume:/data --runtime=habana -e HABANA\_VISIBLE\_DEVICES=all -e OMPI\_MCA\_btl\_vader\_single\_copy\_mechan

ism=none -e BATCH\_BUCKET\_SIZE=32 --cap-add=sys\_nice --ipc=host ghcr.io/huggingface/tgi-gaudi:1.2.1 --model-id $model

**Client**

python run\_generation.py --model\_id $MODEL\_ID --total\_sample\_count 128

文本

描述已自动生成

* 1. Adjustment parameter MAX\_CONCURRENT\_REQUESTS=256

**Server**

docker run -p 8080:80 -v $volume:/data --runtime=habana -e HABANA\_VISIBLE\_DEVICES=all -e OMPI\_MCA\_btl\_vader\_single\_copy\_mechan

ism=none -e BATCH\_BUCKET\_SIZE=16 -e MAX\_CONCURRENT\_REQUESTS=256 --cap-add=sys\_nice --ipc=host ghcr.io/huggingface/tgi-gaudi:1.2.1 --model-id $model

INFO text\_generation\_launcher: Args { model\_id: "/data/models--meta-llama--Llama-2-7b-hf", revision: None, validation\_workers: 2, sharded: None, num\_shard: None, quantize: None, dtype: None, trust\_remote\_code: false, **max\_concurrent\_requests: 256**, max\_best\_of: 2, max\_stop\_sequences: 4, max\_top\_n\_tokens: 5, max\_input\_length: 1024, max\_total\_tokens: 2048, waiting\_served\_ratio: 1.2, max\_batch\_prefill\_tokens: 4096, max\_batch\_total\_tokens: None, max\_waiting\_tokens: 20, hostname: "c1f4144e518e", port: 80, shard\_uds\_path: "/tmp/text-generation-server", master\_addr: "localhost", master\_port: 29500, huggingface\_hub\_cache: Some("/data"), weights\_cache\_override: None, disable\_custom\_kernels: false, cuda\_memory\_fraction: 1.0, rope\_scaling: None, rope\_factor: None, json\_output: false, otlp\_endpoint: None, cors\_allow\_origin: [], watermark\_gamma: None, watermark\_delta: None, ngrok: false, ngrok\_authtoken: None, ngrok\_edge: None, env: false }

**Client**

python run\_generation.py --model\_id $MODEL\_ID --total\_sample\_count 256

文本

描述已自动生成

* 1. Adjustment parameter MAX\_CONCURRENT\_REQUESTS=256 BATCH\_BUCKET\_SIZE=4

**Server**

docker run -p 8080:80 -v $volume:/data --runtime=habana -e HABANA\_VISIBLE\_DEVICES=all -e OMPI\_MCA\_btl\_vader\_single\_copy\_mechanism=none -e BATCH\_BUCKET\_SIZE=4 -e MAX\_CONCURRENT\_REQUESTS=256 --cap-add=sys\_nice --ipc=host ghcr.io/huggingface/tgi-gaudi:1.2.1 --model-id $model

**Client**

python run\_generation.py --model\_id $MODEL\_ID --total\_sample\_count 256

文本

描述已自动生成

BATCH\_BUCKET\_SIZE: Batch size for decode operation will be rounded to the nearest multiple of this number. This limits the number of cached graphs.

MAX\_CONCURRENT\_REQUESTS: The maximum amount of concurrent requests for this particular deployment.

llm on ray(dynamic batching) on guadi 1hpu bf16

|  |  |  |  |
| --- | --- | --- | --- |
|  | input/output 32/4064 | input/output 32/3072 | input/output 1024/3072 |
| bs=256 |  |  |  |
| bs=8 |  |  |  |
| bs=4 | Total time: 185.142 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 0.108 requests/s  Input Token Throughput: 3.457 tokens/s  output Token Throughput: 439.015 tokens/s  Average latency per Request: 111.105 s  Average latency per Token: 0.027 s | Total time: 131.250 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 0.152 requests/s  Input Token Throughput: 4.876 tokens/s  output Token Throughput: 468.113 tokens/s  Average latency per Request: 78.771 s  Average latency per Token: 0.026 s | Total time: 148.557 s  Prompt Length (Min/Med/Max): 1024 / 1024 / 1024  Request Throughput (QPS): 0.135 requests/s  Input Token Throughput: 137.860 tokens/s  output Token Throughput: **413.579 tokens/s**  Average latency per Request: 89.154 s  Average latency per Token: 0.029 s |
| bs=1 |  |  |  |

**256 reqs 32/64**

**IPEX dataset**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **llm on ray with vllm** | **llm on ray(dynamic batching)** | **llm on ray with ipex(dynamic batching)** | **llm on ray(dynamic batching) on guadi 1hpu** |  |
| **bs=256** | Total time: 45.306 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 5.651 requests/s  Input Token Throughput: 180.817 tokens/s  output Token Throughput: 359.691 tokens/s  Average latency per Request: 45.129 s  Average latency per Token: 0.711 s      **numactl -N 0 -m 0 -C 0-55 ray start --head**  Total time: 42.353 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 6.044 requests/s  Input Token Throughput: 193.420 tokens/s  output Token Throughput: 386.840 tokens/s  Average latency per Request: 42.283 s  Average latency per Token: 0.661 s | **transformers 4.38.0**  Total time: 573.882 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 0.446 requests/s  Input Token Throughput: 14.275 tokens/s  output Token Throughput: 28.549 tokens/s  Average latency per Request: 554.848 s  Average latency per Token: 8.669 s      **transformers 4.35.0**  Total time: 219.779 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 1.165 requests/s  Input Token Throughput: 37.274 tokens/s  output Token Throughput: 74.548 tokens/s  Average latency per Request: 214.349 s  Average latency per Token: 3.349 s | Total time: 91.661 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 2.793 requests/s  Input Token Throughput: 89.373 tokens/s  output Token Throughput: 178.745 tokens/s  Average latency per Request: 77.021 s  Average latency per Token: 1.203 s        model\_config.pop("torch\_dtype") | Total time: 38.190 s Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 6.703 requests/s  Input Token Throughput: 214.505 tokens/s  output Token Throughput: 429.010 tokens/s  Average latency per Request: 35.718 s  Average latency per Token: 0.558 s |  |
| **bs=8** | Total time: 207.085 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 1.236 requests/s  Input Token Throughput: 39.559 tokens/s  output Token Throughput: 79.117 tokens/s  Average latency per Request: 107.096 s  Average latency per Token: 1.673 s | Total time: 712.173 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 0.359 requests/s  Input Token Throughput: 11.503 tokens/s  output Token Throughput: 23.006 tokens/s  Average latency per Request: 368.889 s  Average latency per Token: 5.764 s | Total time: 170.494 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 1.502 requests/s  Input Token Throughput: 48.049 tokens/s  output Token Throughput: 96.097 tokens/s  Average latency per Request: 86.428 s  Average latency per Token: 1.350 s | Total time: 53.401 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 4.794 requests/s  Input Token Throughput: 153.405 tokens/s  output Token Throughput: 306.811 tokens/s  Average latency per Request: 29.213 s  Average latency per Token: 0.456 s |  |
| **bs=1** | Total time: 4473.013 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 0.057 requests/s  Input Token Throughput: 1.831 tokens/s  output Token Throughput: 3.663 tokens/s  Average latency per Request: 2237.513 s  Average latency per Token: 34.961 s | Total time: 2151.076 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 0.119 requests/s  Input Token Throughput: 3.808 tokens/s  output Token Throughput: 7.617 tokens/s  Average latency per Request: 1079.180 s  Average latency per Token: 16.862 s | Total time: 1154.480 s  Prompt Length (Min/Med/Max): 32 / 32 / 32  Request Throughput (QPS): 0.222 requests/s  Input Token Throughput: 7.096 tokens/s  output Token Throughput: 14.192 tokens/s  Average latency per Request: 579.645 s  Average latency per Token: 9.057 s |  |  |

debug:

Total time: 241.215 s

Prompt Length (Min/Med/Max): 32 / 32 / 32

Request Throughput (QPS): 1.061 requests/s

Input Token Throughput: 33.961 tokens/s

output Token Throughput: 67.923 tokens/s

Average latency per Request: 227.352 s

Average latency per Token: 3.552 s

python benchmarks/benchmark\_serving.py --model-endpoint-base <http://localhost:8000/llama-2-7b-chat-hf> --model-name llama-2-7b-chat-hf --dataset ./dataset/prompt.json --num-prompts 32 --dataset-format IPEX --input-tokens 32 --max-new-tokens 3072 --simple

Total time: 746.586 s

Prompt Length (Min/Med/Max): 32 / 32 / 32

Request Throughput (QPS): 0.043 requests/s

Input Token Throughput: 1.372 tokens/s

output Token Throughput: 131.666 tokens/s

Average latency per Request: 419.990 s

Average latency per Token: 0.137 s

after set parameters:

Total time: 594.705 s

Prompt Length (Min/Med/Max): 32 / 32 / 32

Request Throughput (QPS): 0.054 requests/s

Input Token Throughput: 1.722 tokens/s

output Token Throughput: 165.259 tokens/s

Average latency per Request: 334.556 s

Average latency per Token: 0.109 s

python benchmarks/benchmark\_serving.py --model-endpoint-base <http://localhost:8000/llama-2-7b-chat-hf> --model-name llama-2-7b-chat-hf --dataset ./dataset/prompt.json --num-prompts 20 --dataset-format IPEX --input-tokens 32 --max-new-tokens 3072 --simple

Total time: 371.714 s

Prompt Length (Min/Med/Max): 32 / 32 / 32

Request Throughput (QPS): 0.054 requests/s

Input Token Throughput: 1.722 tokens/s

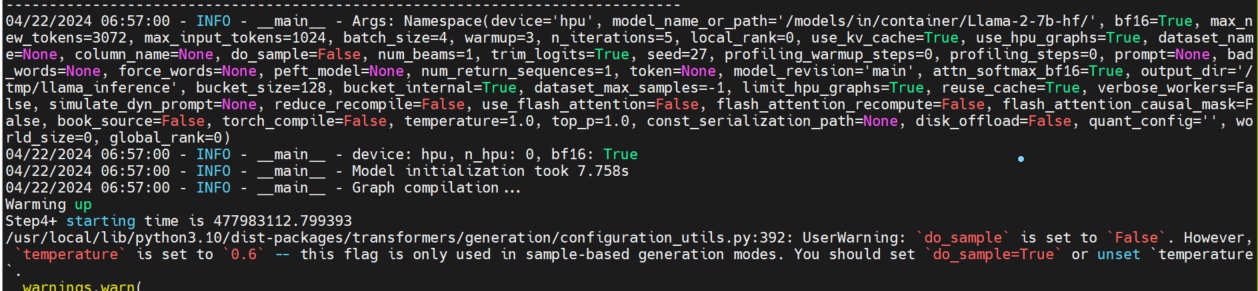
output Token Throughput: 165.283 tokens/s

Average latency per Request: 223.048 s

Average latency per Token: 0.073 s

[optimum-habana-fork/examples/text-generation at habana-main · HabanaAI/optimum-habana-fork (github.com)](https://github.com/HabanaAI/optimum-habana-fork/tree/habana-main/examples/text-generation)

python3 run\_generation.py --num\_beams=1 --model\_name\_or\_path=/mnt/weka/data/pytorch/llama2/Llama-2-7b-hf/ --bucket\_size=128 --max\_new\_tokens=3072 --batch\_size=4 --max\_input\_tokens=1024 --output\_dir=/tmp/llama\_inference --attn\_softmax\_bf16 --use\_hpu\_graphs --use\_kv\_cache --bf16 --bucket\_internal --reuse\_cache --trim\_logits --limit\_hpu\_graphs



Stats:

--------------------------------------------------------------------------------------------------------------

Throughput (including tokenization) = 413.8225152286256 tokens/second

Number of HPU graphs = 61

Memory allocated = 20.95 GB

Max memory allocated = 20.99 GB

Total memory available = 94.62 GB

Graph compilation duration = 95.14851417898899 seconds

--------------------------------------------------------------------------------------------------------------

greedy search

python benchmarks/benchmark\_serving.py --model-endpoint-base <http://localhost:8000/llama-2-7b-chat-hf> --model-name llama-2-7b-chat-hf --dataset ./dataset/prompt.json --num-prompts 20 --dataset-format IPEX --input-tokens 32 --max-new-tokens 3072 --simple

Total time: 359.257 s

Prompt Length (Min/Med/Max): 32 / 32 / 32

Request Throughput (QPS): 0.056 requests/s

Input Token Throughput: 1.781 tokens/s

output Token Throughput: 171.020 tokens/s

Average latency per Request: 215.574 s

Average latency per Token: 0.070 s

bf16 fixed:

warm up

python benchmarks/benchmark\_serving.py --model-endpoint-base <http://localhost:8000/llama-2-7b-chat-hf> --model-name llama-2-7b-chat-hf --dataset ./dataset/prompt.json --num-prompts 12 --dataset-format IPEX --input-tokens 32 --max-new-tokens 3072 --simple

Total time: 85.223 s

Prompt Length (Min/Med/Max): 32 / 32 / 32

Request Throughput (QPS): 0.141 requests/s

Input Token Throughput: 4.506 tokens/s

output Token Throughput: 432.559 tokens/s

Average latency per Request: 58.978 s

Average latency per Token: 0.019 s

benchmark

python benchmarks/benchmark\_serving.py --model-endpoint-base <http://localhost:8000/llama-2-7b-chat-hf> --model-name llama-2-7b-chat-hf --dataset ./dataset/prompt.json --num-prompts 20 --dataset-format IPEX --input-tokens 32 --max-new-tokens 3072 --simple

Total time: 131.243 s

Prompt Length (Min/Med/Max): 32 / 32 / 32

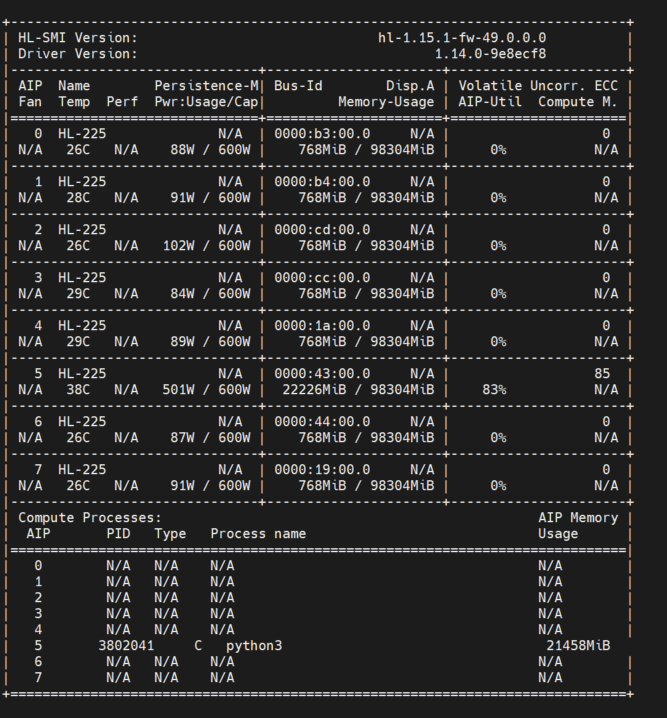
Request Throughput (QPS): 0.152 requests/s

Input Token Throughput: 4.876 tokens/s

output Token Throughput: 468.139 tokens/s

Average latency per Request: 78.767 s

Average latency per Token: 0.026 s



Total time: 148.557 s

Prompt Length (Min/Med/Max): **1024 / 1024 / 1024**

Request Throughput (QPS): 0.135 requests/s

Input Token Throughput: 137.860 tokens/s

output Token Throughput: 413.579 tokens/s

Average latency per Request: 89.154 s

Average latency per Token: 0.029 s

the prompt token length is padding to 1024

print("max\_length: ", args.max\_input\_tokens) # 1024

if args.max\_input\_tokens > 0:

input\_tokens = tokenizer.batch\_encode\_plus(

input\_sentences,

return\_tensors="pt",

padding="max\_length",

max\_length=args.max\_input\_tokens,

truncation=True,

)

vllm engine best performance(OMP\_PROC\_BIND=true)

Successful requests: **1000**

Benchmark duration: 1096.264717 s

Total input tokens: 233094

Total generated tokens: 246059

Request throughput: 0.91 requests/s

Input token throughput: 212.63 tokens/s

Output token throughput: 224.45 tokens/s

Mean TTFT: 524667.43 ms

Median TTFT: 510300.43 ms

P99 TTFT: 1034158.82 ms

Mean TPOT: 9323.75 ms

Median TPOT: 2841.03 ms

P99 TPOT: 95161.80 ms

-------------------------------------------------------

Successful requests: **3000**

Benchmark duration: 3081.517546 s

Total input tokens: 727798

Total generated tokens: 727052

Request throughput: 0.97 requests/s

Input token throughput: 236.18 tokens/s

Output token throughput: 235.94 tokens/s

Mean TTFT: 1550532.22 ms

Median TTFT: 1541576.98 ms

P99 TTFT: 3006719.80 ms

Mean TPOT: 33879.00 ms

Median TPOT: 7883.37 ms

P99 TPOT: 322671.32 ms

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 socket | 1 socket | 2 sockets | 2 sockets | 2 nodes 3 sockets | 2 nodes 3 sockets | 2 nodes 4 sockets | 2 nodes 4 sockets |
| core\*replica | completion | track token latency | completion | track token latency | completion | track token latency | completion | track token latency |
| 27\*2  req1000 | request from dr11s05:  Total time: 955.390 s  Prompt Length (Min/Med/Max): 4 / 95 / 1024  Request Throughput (QPS): 1.047 requests/s  Input Token Throughput: 243.978 tokens/s  output Token Throughput: 256.569 tokens/s  Average latency per Request: 440.329 s  Average latency per Token: 6.631 s | Total time: 992.931 s  Prompt Length (Min/Med/Max): 4 / 95 / 1024  Request Throughput (QPS): 1.007 requests/s  Input Token Throughput: 234.753 tokens/s  output Token Throughput: 247.825 tokens/s  Average latency per Request: 459.584 s  Average latency per Token: 6.268 s  Average latency for First Tokens: 170.659 s  Average latency for Next Tokens: 2.309 s |  |  |  |  |  |  |
| 27\*2  req3000 | Total time: 2583.592 s  Prompt Length (Min/Med/Max): 4 / 99 / 1024  Request Throughput (QPS): 1.161 requests/s  Input Token Throughput: 281.700 tokens/s  output Token Throughput: **280.368** tokens/s  Average latency per Request: 1270.713 s  Average latency per Token: 28.254 s | Total time: 2725.957 s  Prompt Length (Min/Med/Max): 4 / 99 / 1024  Request Throughput (QPS): 1.101 requests/s  Input Token Throughput: 266.988 tokens/s  output Token Throughput: 266.703 tokens/s  Average latency per Request: 1347.492 s  Average latency per Token: 26.473 s  Average latency for First Tokens: 858.171 s  Average latency for Next Tokens: 1.752 s | Total time: 1832.773 s  Prompt Length (Min/Med/Max): 4 / 99 / 1024  Request Throughput (QPS): 1.637 requests/s  Input Token Throughput: 397.102 tokens/s  output Token Throughput: **395.225** tokens/s  Average latency per Request: 846.161 s  Average latency per Token: 15.272 s | Total time: 1871.001 s  Prompt Length (Min/Med/Max): 4 / 99 / 1024  Request Throughput (QPS): 1.603 requests/s  Input Token Throughput: 388.989 tokens/s  output Token Throughput: 388.542 tokens/s  Average latency per Request: 878.161 s  Average latency per Token: 14.231 s  Average latency for First Tokens: 467.231 s  Average latency for Next Tokens: 2.781 s | numactl -N 0,1 -m 0,1 -C 0-55,56-111 ray start --head --include-dashboard False --num-cpus 112  numactl -N 0 -m 0 -C 0-55 ray start --address='10.0.11.2:6379' --num-cpus 56  Total time: 1238.467 s  Prompt Length (Min/Med/Max): 4 / 99 / 1024  Request Throughput (QPS): 2.422 requests/s  Input Token Throughput: 587.660 tokens/s  output Token Throughput: 584.882 tokens/s  Average latency per Request: 525.786 s  Average latency per Token: 8.174 s | Total time: 1326.402 s  Prompt Length (Min/Med/Max): 4 / 99 / 1024  Request Throughput (QPS): 2.262 requests/s  Input Token Throughput: 548.701 tokens/s  output Token Throughput: 548.123 tokens/s  Average latency per Request: 546.226 s  Average latency per Token: 7.713 s  Average latency for First Tokens: 200.486 s  Average latency for Next Tokens: 2.884 s | numactl -N 0,1 -m 0,1 -C 0-55,56-111 ray start --head --include-dashboard False --num-cpus 112  numactl -N 0,1 -m 0,1 -C 0-55,56-111 ray start --address='10.0.11.2:6379' --num-cpus 112  Total time: 1058.798 s  Prompt Length (Min/Med/Max): 4 / 99 / 1024  Request Throughput (QPS): 2.833 requests/s  Input Token Throughput: 687.381 tokens/s  output Token Throughput: 684.131 tokens/s  Average latency per Request: 461.273 s  Average latency per Token: 7.008 s | Total time: 1160.074 s  Prompt Length (Min/Med/Max): 4 / 99 / 1024  Request Throughput (QPS): 2.586 requests/s  Input Token Throughput: 627.372 tokens/s  output Token Throughput: **626.685** tokens/s  Average latency per Request: 490.081 s  Average latency per Token: 6.897 s  Average latency for First Tokens: 151.122 s  Average latency for Next Tokens: 3.544 s |
|  |  |  | 395/280 = 1.4107 |  | 584/280 = 2.0857  584/395 = 1.4785 |  | 684/584 = 1.1712  684/395 = 1.7316  684/280 = 2.4429 |  |

|  |  |
| --- | --- |
| 32 cores wo tcmalloc | 32 cores wi tcmalloc |
| OMP\_PROC\_BIND=close numactl --physcpubind=0-55 --membind=0 ray start --head --num-cpus=56 --num-gpus=0  llm\_on\_ray-serve --config\_file llm\_on\_ray/inference/models/vllm/llama-2-7b-chat-hf-vllm\_27\_2.yaml --simple --max\_concurrent\_queries $VALUE\_INF --vllm\_max\_num\_seqs 256 --keep\_serve\_terminal | LD\_PRELOAD="/home/ykp/miniconda3/pkgs/gperftools-2.10-h09c0d1c\_0/lib/libtcmalloc.so" OMP\_PROC\_BIND=close numactl --physcpubind=0-55 --membind=0 ray start --head --num-cpus=56 --num-gpus=0 |
| Total time: 1284.481 s  Prompt Length (Min/Med/Max): 4 / 95 / 1024  Request Throughput (QPS): 0.779 requests/s  Input Token Throughput: 181.469 tokens/s  output Token Throughput: 190.835 tokens/s  Average latency per Request: 598.520 s  Average latency per Token: 11.722 s | Total time: 1006.545 s  Prompt Length (Min/Med/Max): 4 / 95 / 1024  Request Throughput (QPS): 0.993 requests/s  Input Token Throughput: 231.578 tokens/s  **output Token Throughput: 243.530 tokens/s**  Average latency per Request: 461.659 s  Average latency per Token: 8.920 s  (243.530-190.835)/190.835 = 0.276 |

(243.530-190.835)/190.835 = 0.2761

|  |  |
| --- | --- |
| 53 cores wo tcmalloc | 53 cores wi tcmalloc |
| OMP\_PROC\_BIND=close numactl --physcpubind=0-55 --membind=0 ray start --head --num-cpus=56 --num-gpus=0 | LD\_PRELOAD="/home/ykp/miniconda3/pkgs/gperftools-2.10-h09c0d1c\_0/lib/libtcmalloc.so" OMP\_PROC\_BIND=close numactl --physcpubind=0-55 --membind=0 ray start --head --num-cpus=56 --num-gpus=0 |
| Total time: 1514.727 s  Prompt Length (Min/Med/Max): 4 / 95 / 1024  Request Throughput (QPS): 0.660 requests/s  Input Token Throughput: 153.885 tokens/s  output Token Throughput: 161.827 tokens/s  Average latency per Request: 692.899 s  Average latency per Token: 13.396 s | Total time: 911.117 s  Prompt Length (Min/Med/Max): 4 / 95 / 1024  Request Throughput (QPS): 1.098 requests/s  Input Token Throughput: 255.833 tokens/s  **output Token Throughput: 269.037 tokens/s**  Average latency per Request: 410.956 s  Average latency per Token: 7.961 s |

(269.037-161.827)/161.827=0.6624976054675671

|  |  |
| --- | --- |
| 27 cores\*2 wo tcmalloc | 27 cores\*2 wi tcmalloc |
| numactl -N 0 -m 0 -C 0-55 ray start --head --include-dashboard False | LD\_PRELOAD="/home/ykp/miniconda3/pkgs/gperftools-2.10-h09c0d1c\_0/lib/libtcmalloc.so" numactl --physcpubind=0-55 --membind=0 ray start --head --include-dashboard False --num-cpus=56 --num-gpus=0 |
| Total time: 875.856 s  Prompt Length (Min/Med/Max): 4 / 95 / 1024  output\_num\_tokens: 245124, benchmark\_time: 875.8559360094368  Request Throughput (QPS): 1.142 requests/s  Input Token Throughput: 266.133 tokens/s  output Token Throughput: 279.868 tokens/s  Average latency per Request: 379.244 s  Average latency per Token: 5.335 s  replica end time: 1715011380.1538506 | Total time: 733.818 s  Prompt Length (Min/Med/Max): 4 / 95 / 1024  Request Throughput (QPS): 1.363 requests/s  Input Token Throughput: 317.646 tokens/s  **output Token Throughput: 334.039 tokens/s**  Average latency per Request: 316.536 s  Average latency per Token: 4.312 s |

|  |  |
| --- | --- |
| 27 cores \*4 wo tcmalloc | 27 cores\*4 wi tcmalloc |
| head: sr233 socket0 worker: spr02 client: sr233 socket1  sr233: numactl -N 0 -m 0 -C 0-23 ray start --head --include-dashboard False --num-cpus 0  spr02 worker1: numactl -N 0 -m 0 -C 0-55 ray start --address='10.0.0.133:6379'  spr02 worker2: numactl -N 1 -m 1 -C 56-111 ray start --address='10.0.0.133:6379'  spr02 server: llm\_on\_ray-serve --config\_file llm\_on\_ray/inference/models/vllm/llama-2-7b-chat-hf-vllm\_27\_4.yaml --simple --max\_concurrent\_queries $VALUE\_INF --vllm\_max\_num\_seqs 256 --keep\_serve\_terminal > scaling\_test/test\_2/serve\_27\_4\_req2000\_v4.out  client: numactl -N 1 -m 1 python benchmarks/benchmark\_serving.py --model-endpoint-base <http://localhost:8000/llama-2-7b-chat-hf> --model-name llama-2-7b-chat-hf --dataset ./dataset/ShareGPT\_V3\_unfiltered\_cleaned\_split.json --num-prompts 2000 --dataset-format ShareGPT --simple --vllm-engine > scaling\_test/test\_2/client\_27\_4\_req2000\_v4.out | sr233:  LD\_PRELOAD="/home/ykp/miniconda3/envs/llmonray\_master\_vllmup\_3/lib/python3.9/site-packages/bigdl/llm/libs/libtcmalloc.so" numactl --physcpubind=0-23 --membind=0 ray start --head --include-dashboard False --num-cpus=0 --num-gpus=0  spr02:  LD\_PRELOAD="/home/ykp/miniconda3/pkgs/gperftools-2.10-h09c0d1c\_0/lib/libtcmalloc.so" numactl --physcpubind=0-55 --membind=0 ray start --address='10.0.0.133:6379' --num-cpus=56 --num-gpus=0  LD\_PRELOAD="/home/ykp/miniconda3/pkgs/gperftools-2.10-h09c0d1c\_0/lib/libtcmalloc.so" numactl --physcpubind=56-111 --membind=1 ray start --address='10.0.0.133:6379' --num-cpus=56 --num-gpus=0  sr233:  numactl -N 1 -m 1 python benchmarks/benchmark\_serving.py --model-endpoint-base <http://localhost:8000/llama-2-7b-chat-hf> --model-name llama-2-7b-chat-hf --dataset ./dataset/ShareGPT\_V3\_unfiltered\_cleaned\_split.json --num-prompts 2000 --dataset-format ShareGPT --simple --vllm-engine |
| Total time: 851.841 s  Prompt Length (Min/Med/Max): 4 / 97 / 1024  output\_num\_tokens: 487783, benchmark\_time: 851.8406078591943  Request Throughput (QPS): 2.348 requests/s  Input Token Throughput: 563.360 tokens/s  output Token Throughput: 572.622 tokens/s  Average latency per Request: 382.017 s  Average latency per Token: 5.350 s | Total time: 735.316 s  Prompt Length (Min/Med/Max): 4 / 95 / 1024  Request Throughput (QPS): 2.720 requests/s  Input Token Throughput: 652.635 tokens/s  output Token Throughput: **663.365 tokens/s**  Average latency per Request: 317.963 s  Average latency per Token: 4.310 s |

jemalloc

图表, 折线图

描述已自动生成