Counsel GPT
A Study on Efficient Inference serving using llama.cpp on various Quantized LoRA Legal fine-tuned LLMs on Cloud.

Vasavi Doddamani
Mathesh Thirumalai

UC SANTA CRUZ | Baskin Engineering

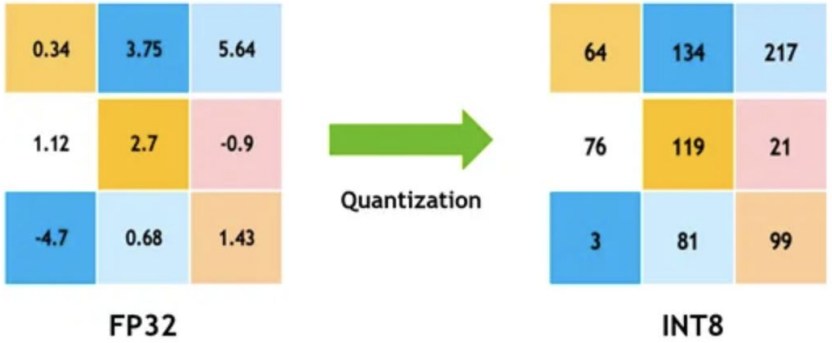
AGENDA

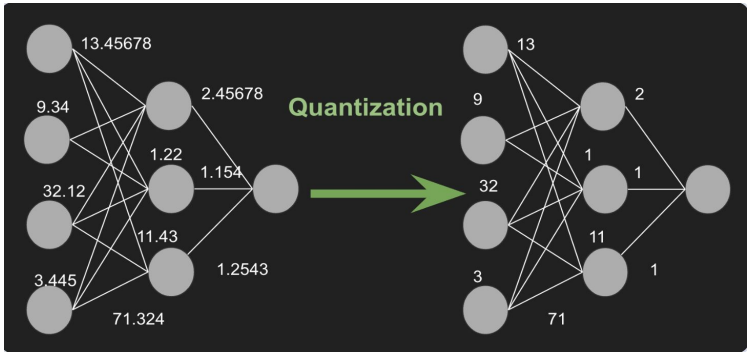
1. PROBLEM STATEMENT
2. SYSTEM ARCHITECTURE
3. COST ANALYSIS
4. BENCHMARKING
5. RESULTS
6. SUMMARY

What is Counsel GPT?

- Legal based fine-tuned LLM
- Fine tuning done using PEFT with LoRA.
- Models used:
  - LLAMA 7B - 4 Bit (llama-2-7b-chat.Q4_K_M)
  - Qwen 7B - 8 Bit (Qwen2.5-7B-Instruct-Q8_0)
- Deployed on GCP & Nautilus

Quantization





Credits:
1.https://www.digitalocean.com/community/tutorials/model-quantization-large-language-models
2. https://medium.com/@florian_algo/model-quantization-1-basic-concepts-860547ec6aa9

Baskin Engineering

Llama.cpp and GPU Acceleration

- Llama.cpp enables efficient inference on consumer GPU's or edge devices or even on CPU's.
- The models that we use has:
  - **Llama 2 7B -> 32 layers**
  - **Qwen 2.5 7B -> 28 layers**
- Allows entire GPU acceleration.
- Run on specific images for CPU and GPU(CUDA).

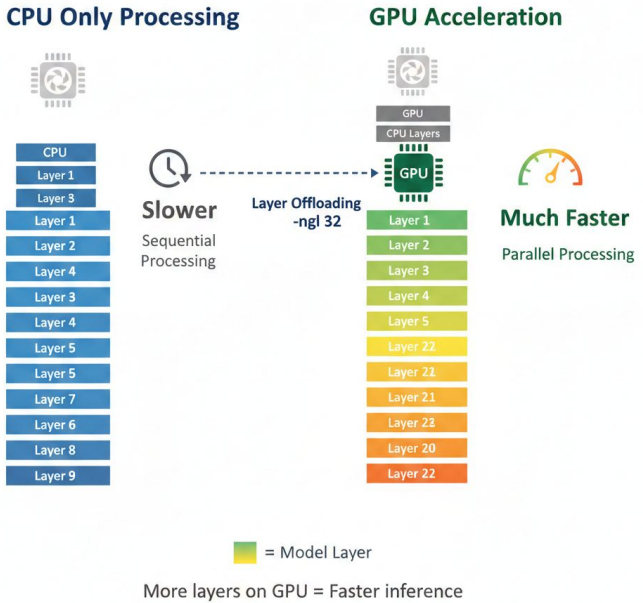


Image Credits: Nano Banana (Google)

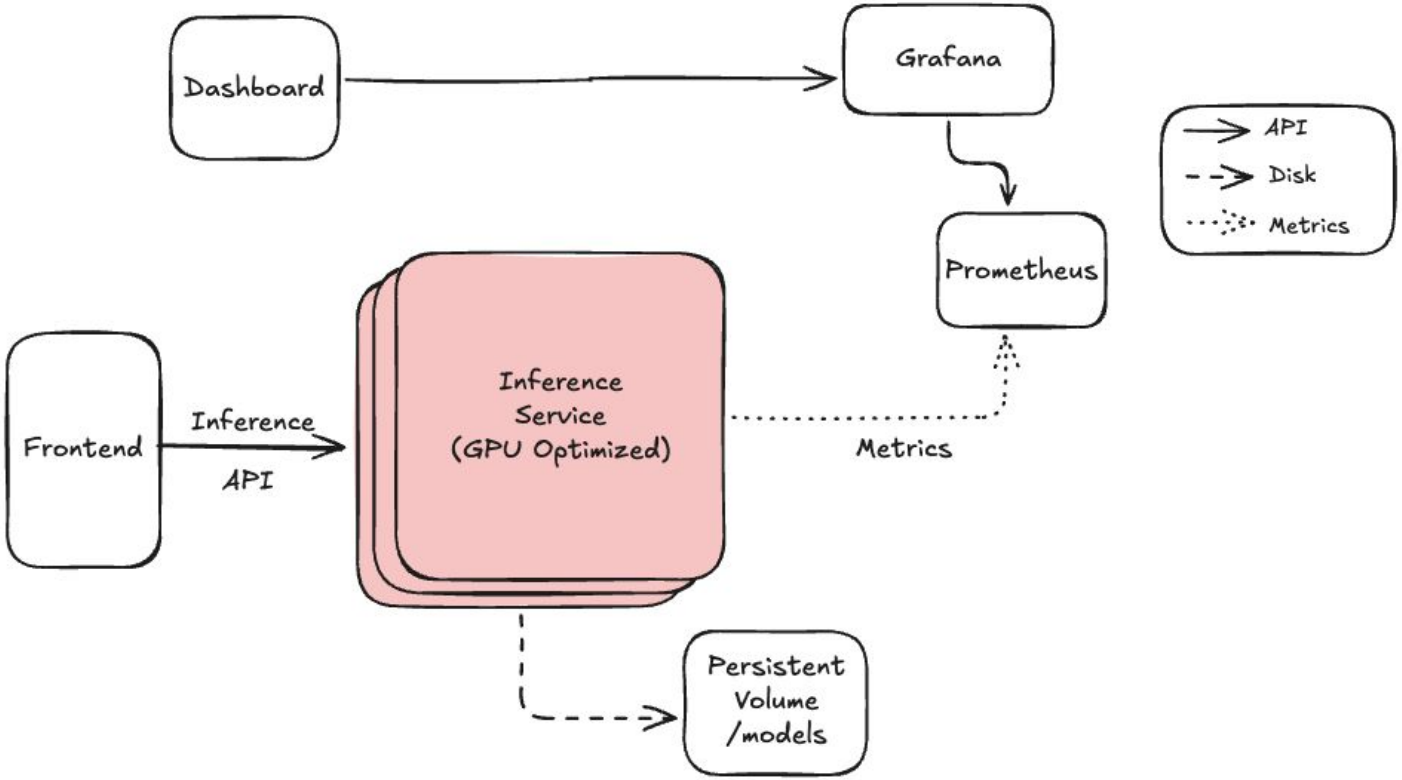
SYSTEM ARCHITECTURE

Approach - I

Objectives: Make the system work

Baskin Engineering

Basic Architecture



Challenges with GPU-Only Setup in GCP

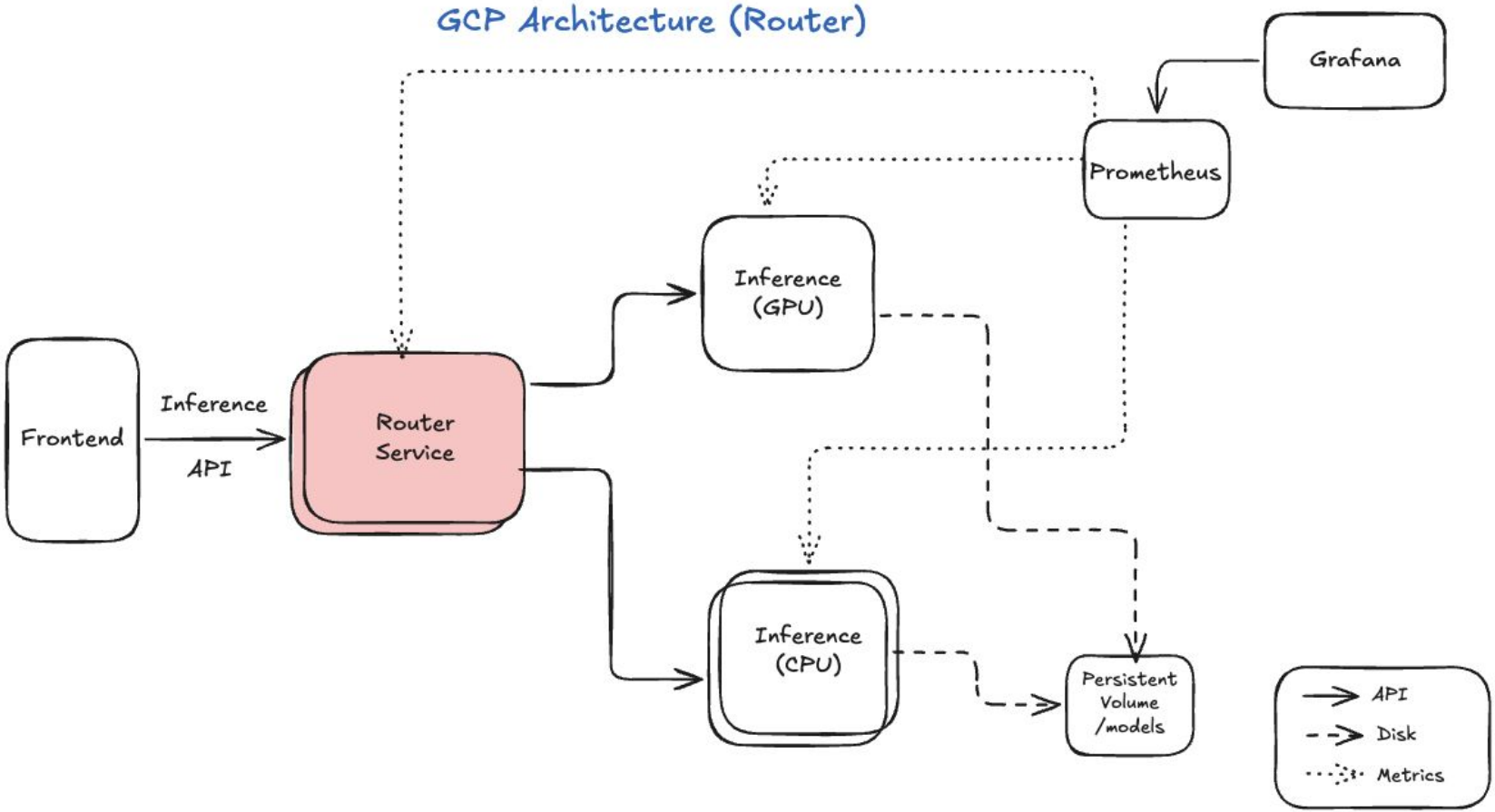
- We had very limited GPU quota on GCP (only one L4).
- GPU nodes are expensive, so we couldn't run everything on GPU.
- When the GPU got busy, requests would fail or slow down.

SYSTEM ARCHITECTURE

Approach - II

New Objective: Make the system efficient and available as much as possible with given constraints

Baskin Engineering

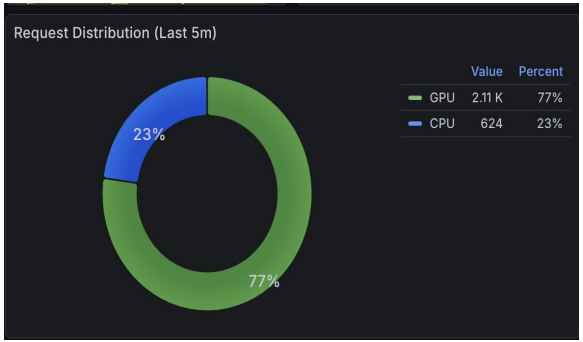

GCP Architecture (Router)
Grafana
Prometheus
Inference
(GPU)
Frontend
Inference
API
Router
Service
Inference
(CPU)
Persistent
Volume
/models
API
Disk
Metrics

Router Service

- Lightweight Python-based router.
- Route requests to GPU backend when healthy and not overloaded.
- Automatic fallback to CPU when GPU is down or queue is full.
- Exposes request latency histograms and request counts.

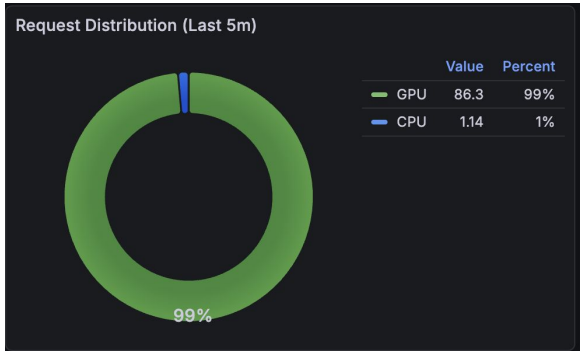
```
function handle_request(request):

    if request.use_cpu:
        return call_cpu_backend(request)

    # gatekeepers
    if not circuit_allows("gpu") or
       not backend_healthy("gpu") or
       not gpu_queue.has_capacity():
        return call_cpu_backend(request)

    # enqueue on GPU or fall back
    if not gpu_queue.enqueue(request):
        return call_cpu_backend(request)

    return
```

Router v/s No Router



2025-12-06 17:26:30	
GPU p50	27.0 s
GPU p95	55.8 s
GPU p99	59.2 s
CPU p50	1.25 mins
CPU p95	1.48 mins
CPU p99	1.50 mins

```
http_req_failed
✓ 'rate < 0.05' rate=3.33%
```

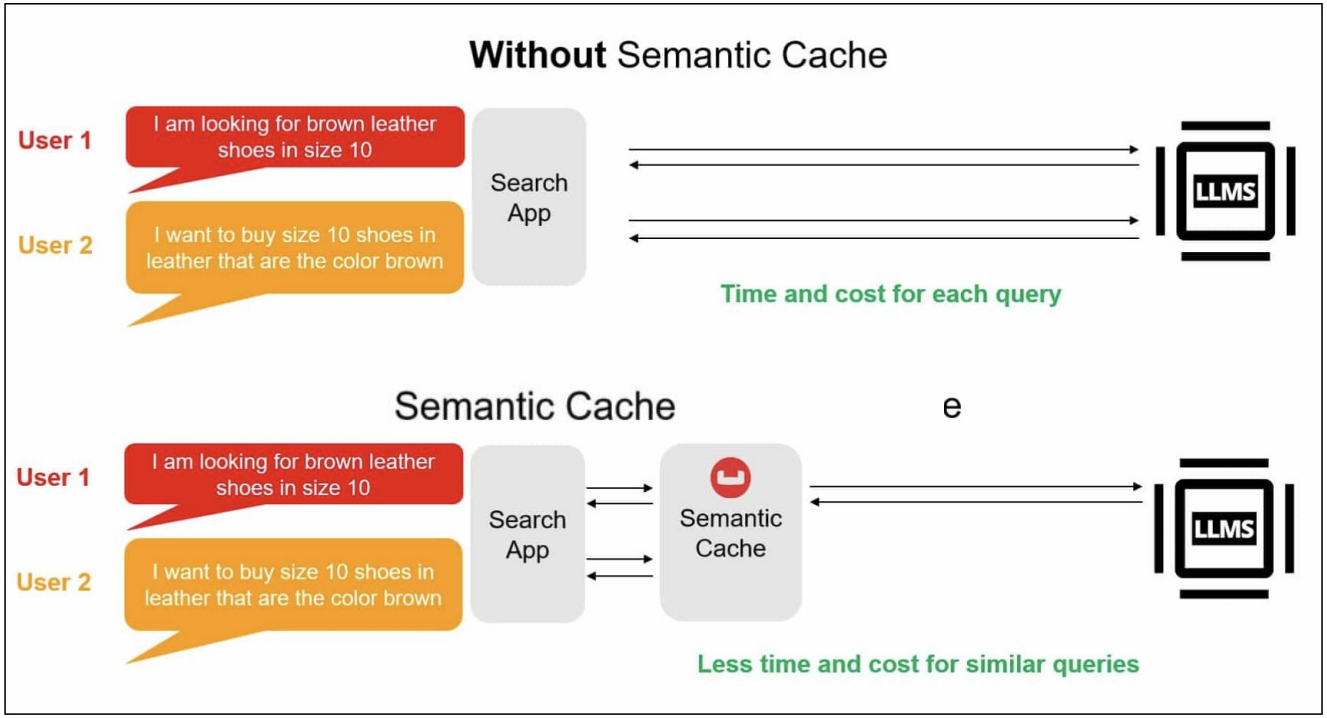



```
http_req_failed
✗ 'rate < 0.05' rate=15.38%
```

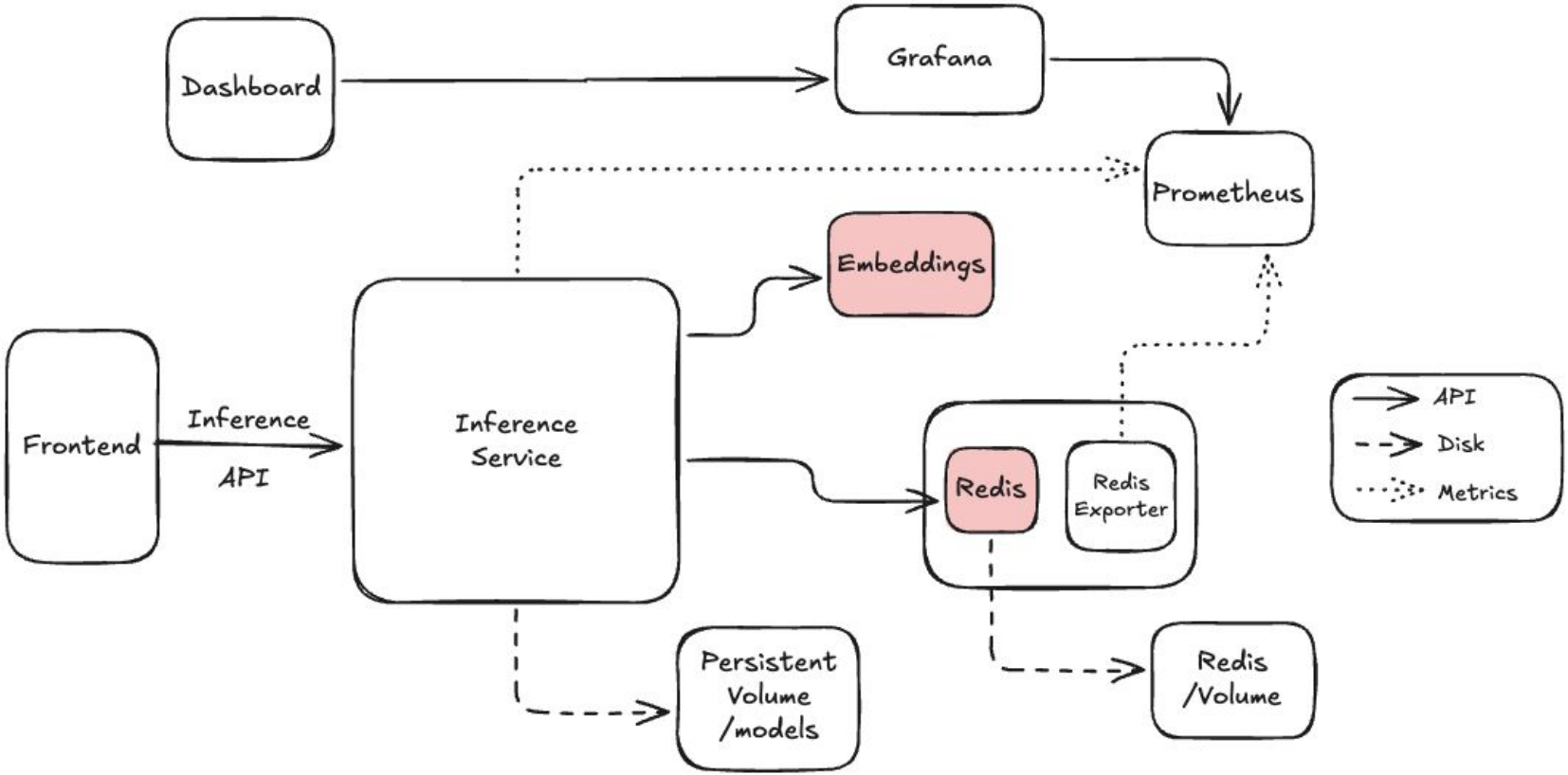
2025-12-06 17:45:00	
GPU Inference p50	52.5 s
GPU Inference p95	1.37 mins

Baskin Engineering

Semantic Caching



Architecture (Semantic Cache)

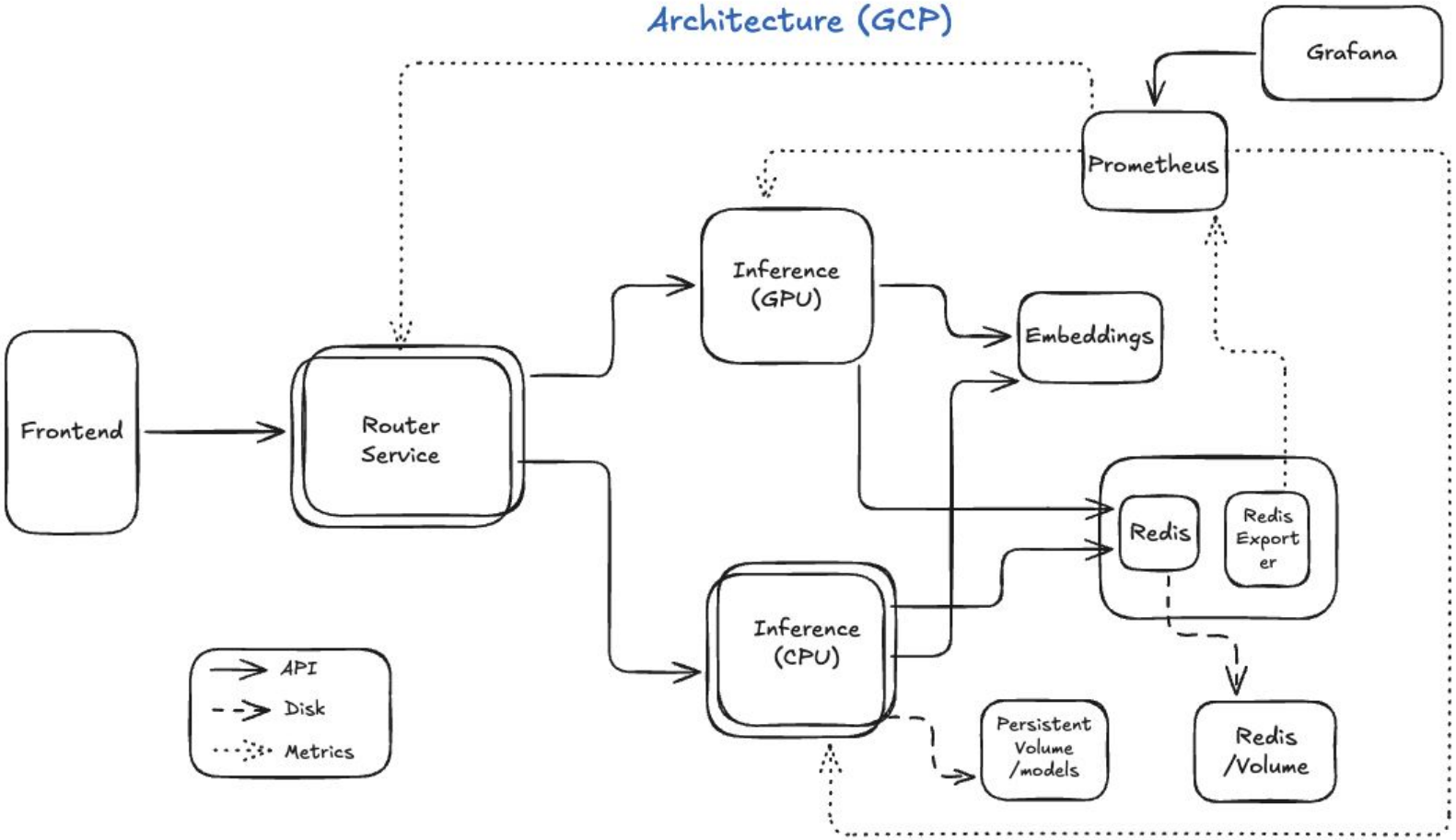


SYSTEM ARCHITECTURE

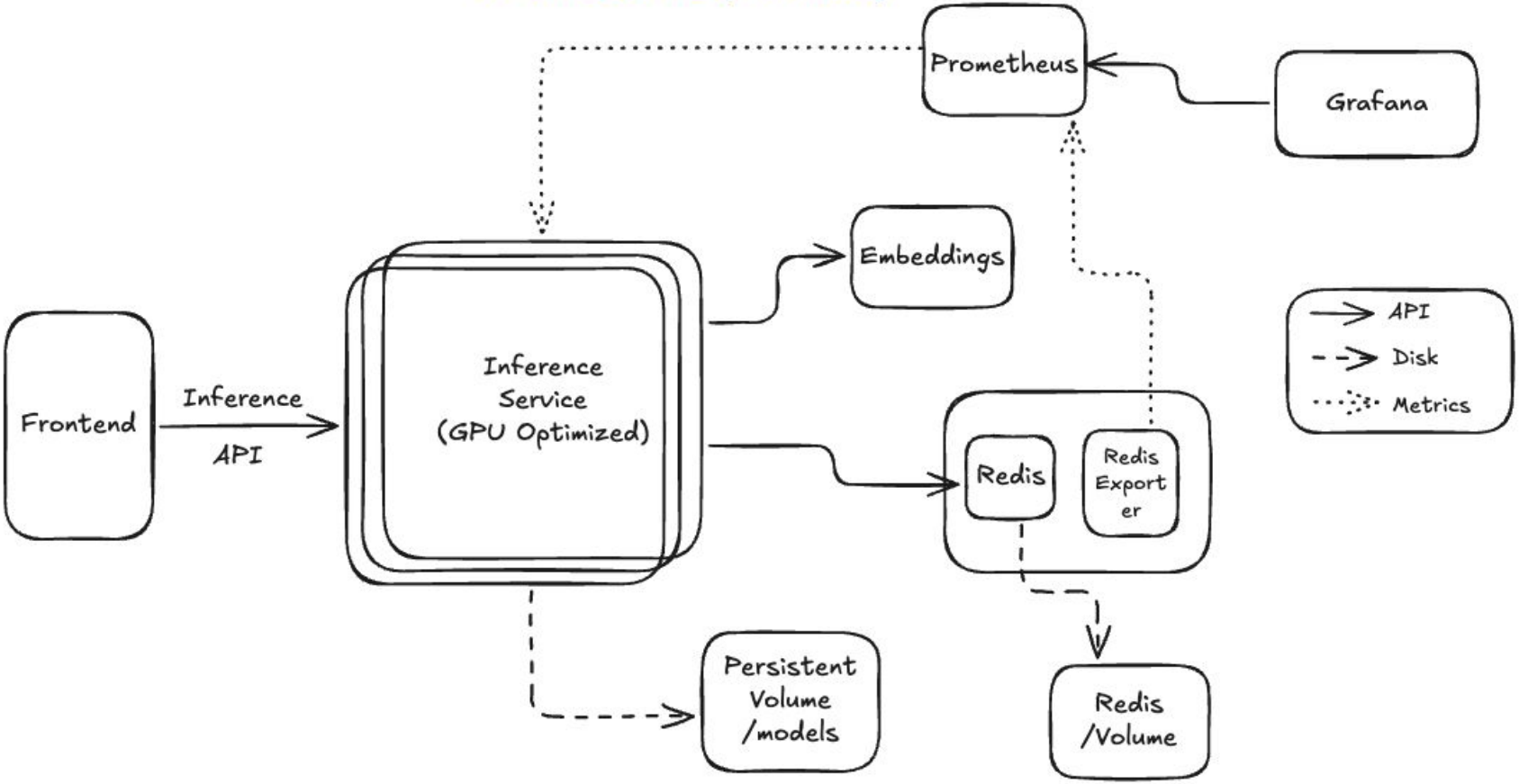
Final Approach

Objective: Combine all optimizations like router service, semantic cache into the existing architecture

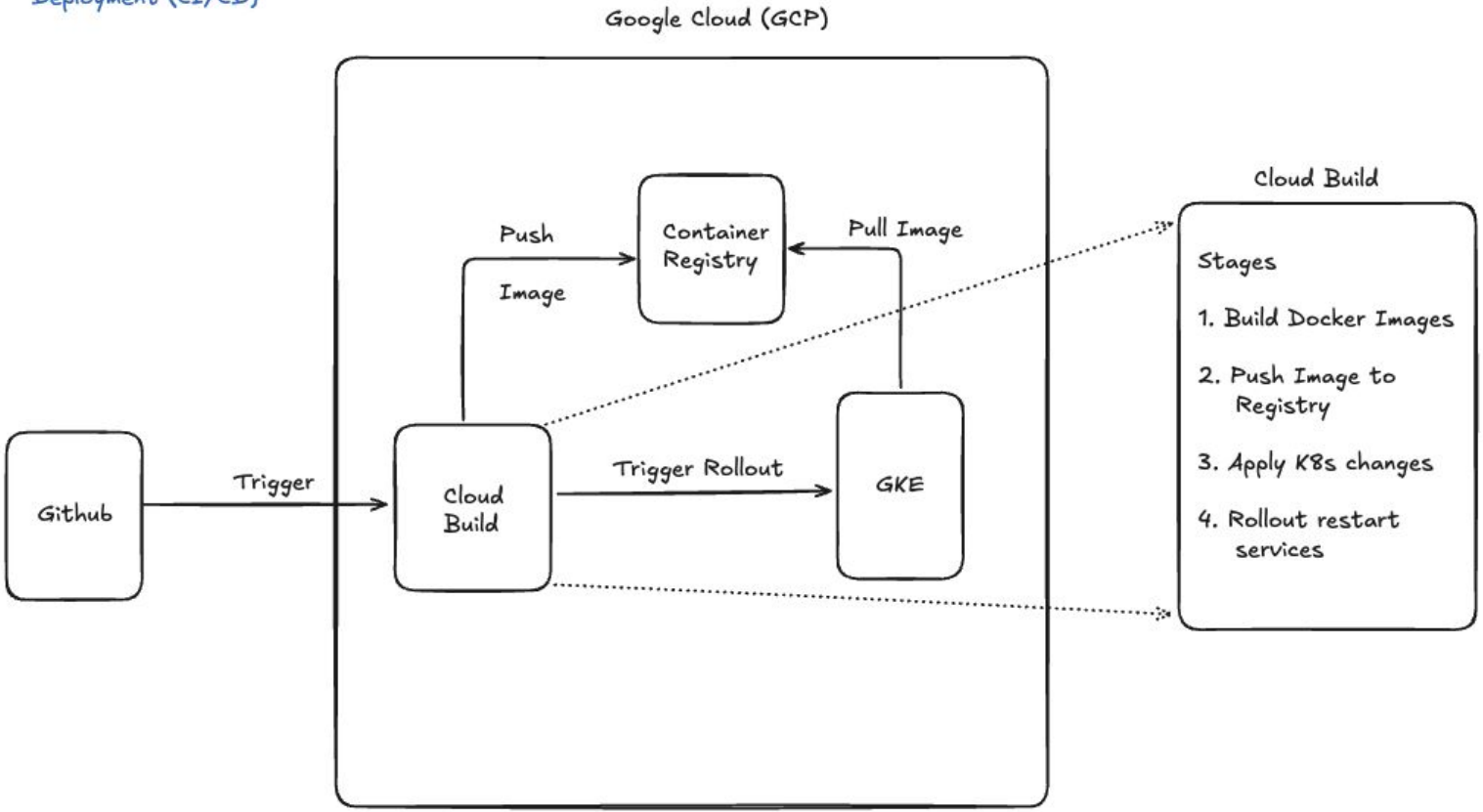
Baskin Engineering


Architecture (GCP)
Grafana
Prometheus
Inference (GPU)
Embeddings
Frontend
Router Service
Redis
Redis Exporter
Inference (CPU)
Persistent Volume /models
Redis /Volume
API
Disk
Metrics

Architecture (Nautilus)



Deployment (CI/CD)



Kubernetes Deployments

GKE - Standard (GCP)

- Separate deployments for cpu-based and gpu-based backends.
- 1 GPU node (NVIDIA L4) (Quota Limitation)
- CPU backend autoscaled via HPA based on CPU.

Nautilus

- Single deployment - gpu-based backends only.
- GPU-based scheduling preferences
  - NVIDIA-L40 (P1)
  - Tesla-V100-SXM2-32GB/16GB (P2)
  - NVIDIA-L4 (P3)
- Autoscaled via HPA based on CPU

GCP Cost

Forecasted cost

Download CSV

Service	Usage cost	Savings programs	Other savings	Subtotal	% Change
Cloud Monitoring	$0.21	–	-$0.21	$0.00	–
Cloud Build	$0.00	–	$0.00	$0.00	–
Cloud Storage	$0.05	–	-$0.05	$0.00	–
Kubernetes Engine	$2.37	–	-$2.37	$0.00	–
Networking	$0.98	–	-$0.98	$0.00	–
Artifact Registry	$0.18	–	-$0.18	$0.00	–
Compute Engine	$23.00	–	-$23.00	$0.00	–

Monthly Forecast

Monthly estimate

$516.99

That's about $0.71 hourly

Pay for what you use: no upfront costs and per second billing

Item	Monthly estimate
4 vCPU + 16 GB memory	$107.16
1 NVIDIA L4	$408.83
10 GB balanced persistent disk	$1.00
Logging	Cost varies
Monitoring	Cost varies
Snapshot schedule	Cost varies
Total	$516.99

Monthly estimate

$1,292.55

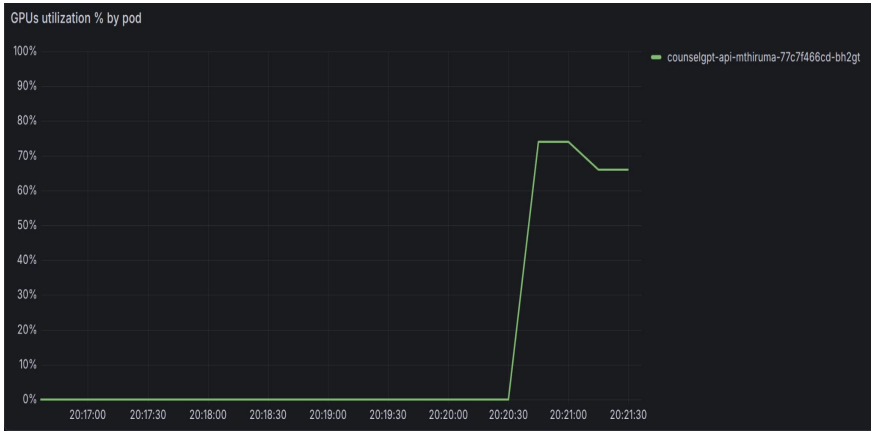
That's about $1.77 hourly

Pay for what you use: no upfront costs and per second billing

Item	Monthly estimate
1 vCPU + 3.75 GB memory	$34.67
1 NVIDIA V100	$1,810.40
10 GB balanced persistent disk	$1.00
Use discount	-$553.52
Logging	Cost varies
Monitoring	Cost varies
Snapshot schedule	Cost varies
Total	$1,292.55

Nautilus Cost

Baskin Engineering


GPUs utilization % by pod
counselgpt-api-mthiruma-77c7f466cd-bh2gt
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
20:17:00
20:17:30
20:18:00
20:18:30
20:19:00
20:19:30
20:20:00
20:20:30
20:21:00
20:21:30

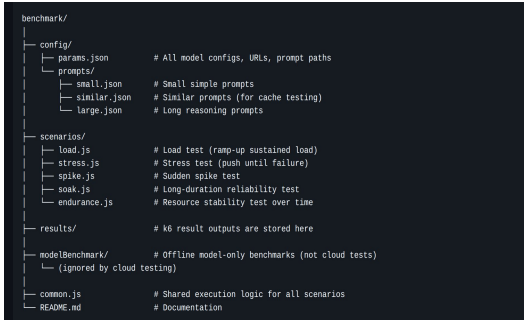

GPUs utilization % by pod
counselgpt-api-mthiruma-77c7f466cd-bh2gt
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
20:21:30
20:22:00
20:22:30
20:23:00
20:23:30
20:24:00
20:24:30
20:25:00
20:25:30
20:26:00

BENCHMARKING

Baskin Engineering

How did we benchmark and test

- Benchmarking
  - Load, Stress, Spike, Soak (both short and very long duration)
- Scenario based testing
  - Normal, middle, deep reasoning on Legal questions
- GCP vs Nautilus
- Architecture testing
  - Similar prompts - For Semantic Caching
  - GPU acceleration vs normal CPU
- 8 bit - 4 bit and Model Performance (Quality of output) -> LLM As Judge and Sentence Transformer.



https://github.com/Mati02K/CounselGPT/tree/main/benchmark

Baskin Engineering

Parameter Tuning for testing

```
{
  "qwen_gpu_on_nautilus": {
    "model": "qwen",
    "use_gpu": true,
    "use_cache": false,
    "max_tokens": 150,
    "api_url": "https://counselgpt-mathesh.nrp-nautilus.io/infer",
    "prompts": "./config/prompts/normal.json"
  },

  "qwen_gpu_on_gcp": {
    "model": "qwen",
    "use_gpu": true,
    "use_cache": false,
    "max_tokens": 150,
    "api_url": "https://34.111.194.27.nip.io/infer",
    "prompts": "./config/prompts/normal.json"
  },

  "llama_gpu_on_nautilus": {
    "model": "llama",
    "use_gpu": true,
    "use_cache": false,
    "max_tokens": 150,
    "api_url": "https://counselgpt-mathesh.nrp-nautilus.io/infer",
    "prompts": "./config/prompts/normal.json"
  },

  "qwen_gpu_off_nautilus": {
    "model": "qwen",
    "use_gpu": false,
    "use_cache": false,
    "max_tokens": 150,
    "api_url": "https://counselgpt-mathesh.nrp-nautilus.io/infer",
    "prompts": "./config/prompts/normal.json"
  },

  "llama_gpu_off_nautilus": {
    "model": "llama",
    "use_gpu": false,
    "use_cache": false,
    "max_tokens": 150,
    "api_url": "https://counselgpt-mathesh.nrp-nautilus.io/infer",
    "prompts": "./config/prompts/normal.json"
  },
```

See more: https://github.com/Mati02K/CounselGPT/tree/main/benchmark

Baskin Engineering

Assumptions and Trials

- Be within limits provided.
- Try for Availability sacrificing a little of performance to understand tradeoffs.
- Try different kind of test and prompts as possible.
- SOTA models serve >1,000,000 RPS and at least 40 - 120 RPS per pod because of using different versions of vLLM’s and strong GPU’s like H100/A100 or even powerful ones.

RESULTS

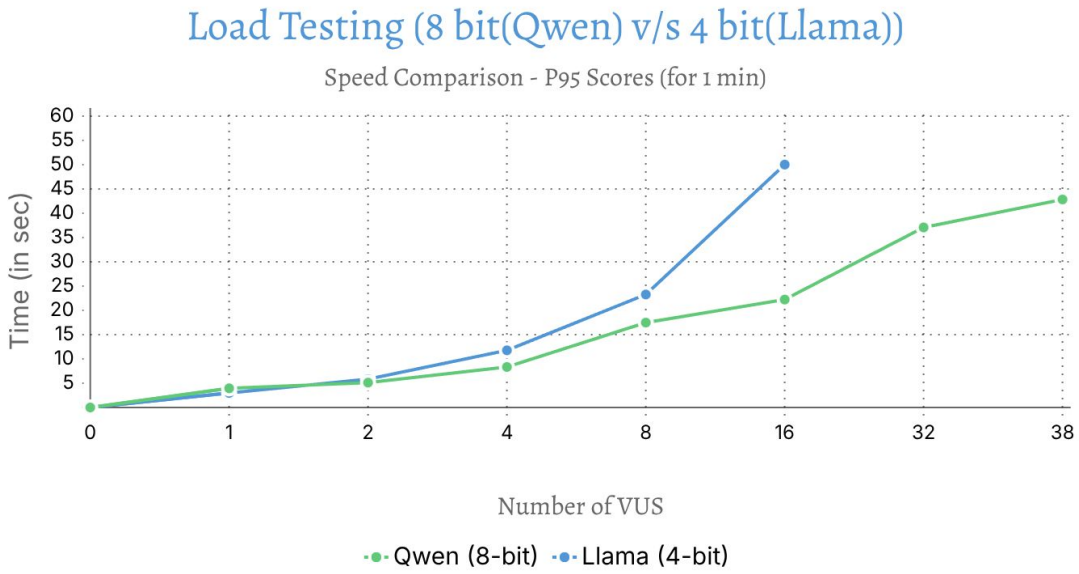
Baskin Engineering

Test dataset

```
"Does this section change the legal status or rights of a human being before they are born alive?",
"What happens when a bill, order, resolution, or vote becomes a law or takes effect?",
"What occurs when the President returns a bill with his objections, but it is reconsidered and passed with a two-thirds majority in both Houses of Co
"What is the role of the Archivist of the United States in this process?",
"What is the responsibility of the Secretary of State regarding treaties and international agreements?",
"Under what circumstances may the publication of certain categories of agreements not be required?",
"How does the Secretary of State ensure public access to treaties and international agreements?",
"How can duplication and waste in publication be avoided?",
"Can the number of copies published be reduced?",
"Is it possible to publish more than one Supplement for each Congress?",
"What is the requirement for the President at the beginning of each regular session of Congress?",
"How are the number of Representatives determined for each state until a reapportionment takes place?",
"What happens when a state has not been redistricted after an apportionment in terms of electing Representatives?",
"Who has the authority to determine the qualifications of employees before their appointment under the supervision of an officer of the Congress?",
"What actions can be taken by an officer of the Congress with regards to employees under their supervision?",
"How is the term officer of the Congress defined in this section?",
"What is the responsibility of the Committees on the Budget of the House of Representatives and the Senate?",
" What are some of the proposals that will be studied by the Committees on the Budget?",
"Can other committees or joint committees of the Congress also conduct studies to improve the budgetary process?",
"What is the responsibility of the Select Committee on Ethics?",
"Who is required to complete the ethics training program conducted by the Select Committee on Ethics?",
"When should new Senators or staff complete the ethics training program?",
"How much compensation does the President receive for their services during their elected term?",
"What is the purpose of the expense allowance given to the President?",
"What privileges does the President have in terms of the use of furniture and other effects belonging to the United States?",
"What are the factors that personnel actions affecting covered employees should be free from discrimination based on?",
"What is the condition under which modifications to the regulations may be made by the President or their designee?",
"What is the remedy for a violation of discriminatory practices based on race, color, religion, sex, or national origin?",
"What can happen to someone who puts something on the American flag or shows it with something added?",
```

Baskin Engineering

Outcomes



Test conditions

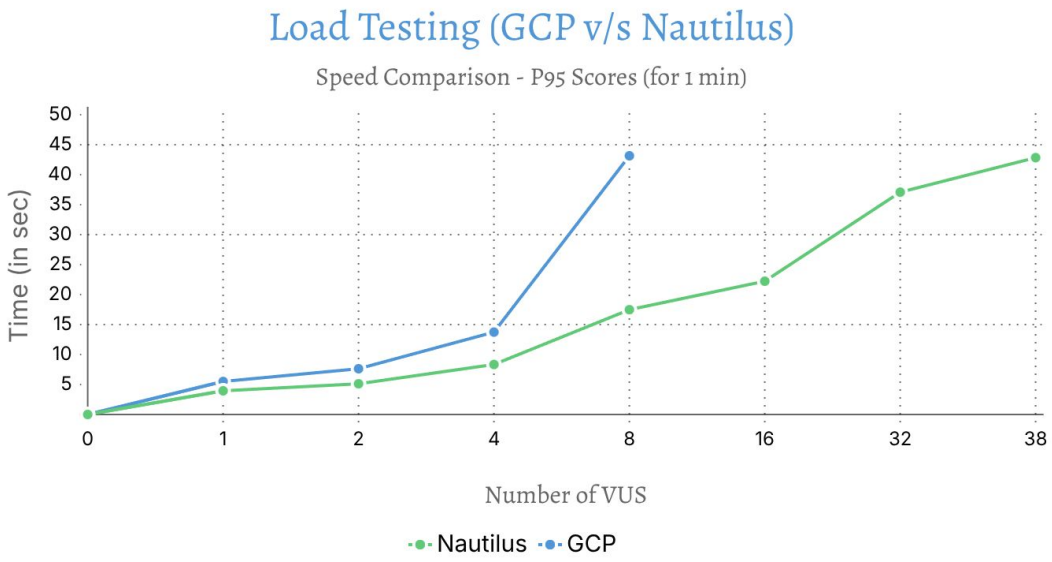
- Failure rate < 1%
- Run for 1 min at 0.2 sec sleep

Few Observations

- RPS Supported by llama ~ 0.5
- RPS supported by Qwen ~ 1.3
- Qwen2.5 is an **efficient architecture also optimized for reasoning** than Llama-2 and has less layers.

Baskin Engineering

Outcomes



Test conditions

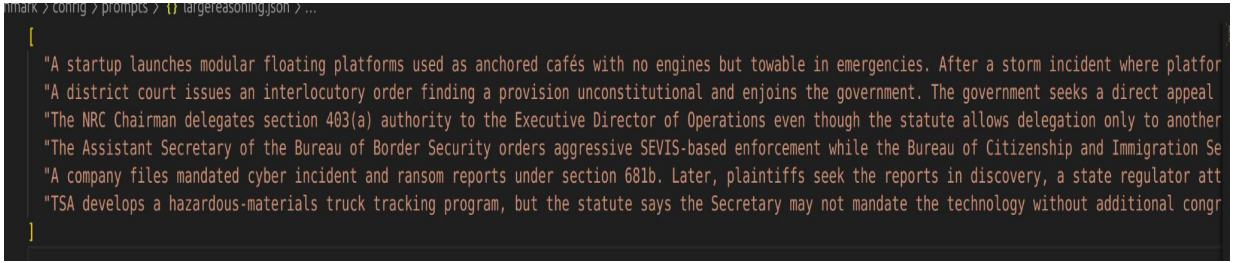
- Failure rate < 1%
- Run for 1 min at 0.2 sec sleep

Few Observations

- RPS Supported by GPS ~ 0.3
- RPS supported by Nautilus ~ 1.3
- Max VUS Supported by GPS ~ 8
- Max VUS supported by Nautilus ~ 38

Baskin Engineering

Test dataset

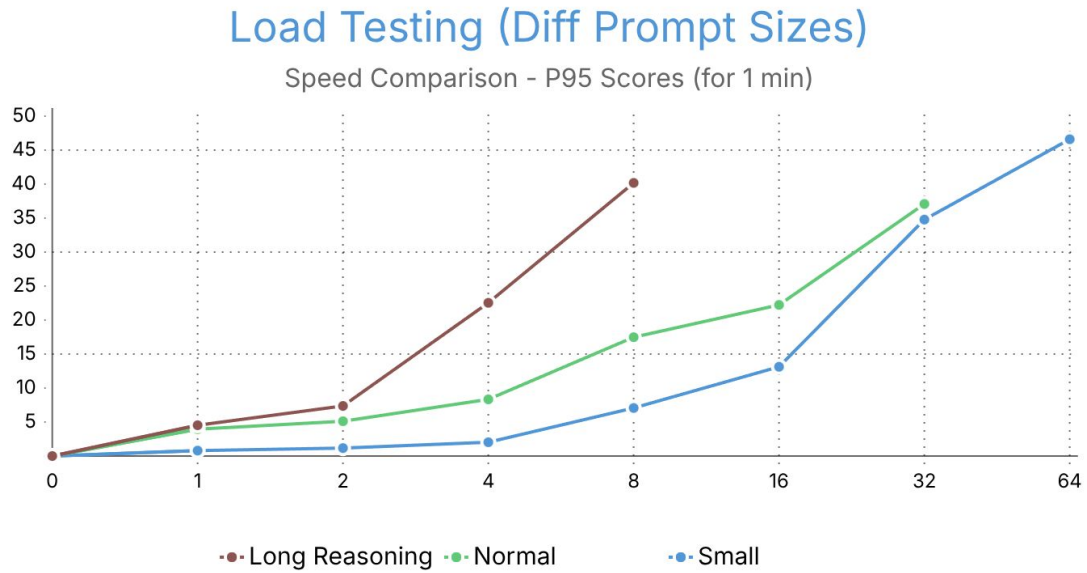


large-reasoning.json

small.json

Baskin Engineering

Outcomes



Test conditions

- Failure rate < 1%
- Run for 1 min at 0.2 sec sleep

Few Observations

- RPS Supported by large ~ 0.3
- RPS supported by normal ~ 1.3
- RPS supported by small ~ 2.4

Baskin Engineering

Test dataset

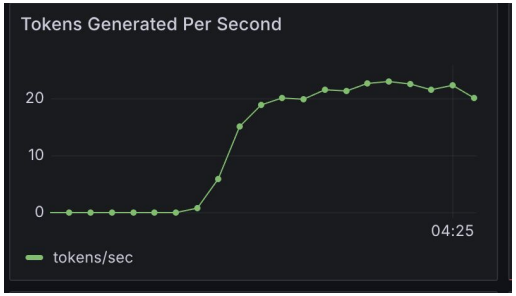
```
[
  "Explain contract law basics.",
  "Give me the basic principles of contract law.",
  "What are the fundamentals of contract law?",

  "What are key elements of a valid contract?",
  "List the essential elements required for a valid contract.",
  "What conditions must be met for a contract to be valid?",

  "Describe enforceability of contracts.",
  "How do you determine if a contract is enforceable?",
  "Explain what makes a contract legally enforceable.",

  "Define breach of contract.",
  "What does it mean when a contract is breached?",
  "Explain the concept of contract breach.",

  "What is consideration in a contract?",
  "Explain the role of consideration in contracts.",
  "Why is consideration required for contractual validity?",

  "Describe remedies for breach of contract.",
  "What legal remedies exist for contract breaches?",
  "How can courts remedy a breach of contract?",

  "What is the difference between void and voidable contracts?",
  "Explain void vs voidable contracts.",
  "How do void and voidable contracts differ legally?",

  "Define offer and acceptance in contract formation.",
  "Explain how offer and acceptance work in forming contracts.",
  "What is meant by offer and acceptance?",

  "What is contractual capacity?",
  "Explain capacity requirements in contract law.",
  "Who is legally capable of entering into a contract?",

  "What is the statute of frauds?",
  "Explain the purpose of the statute of frauds.",
  "Which contracts fall under the statute of frauds?"
]
```

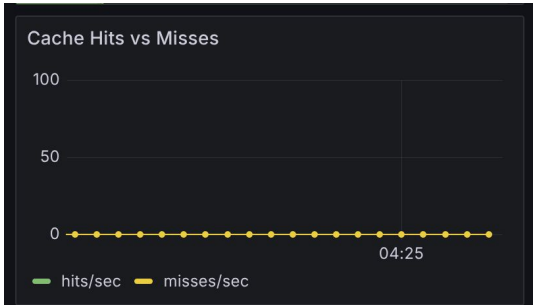
similar.json

Baskin Engineering

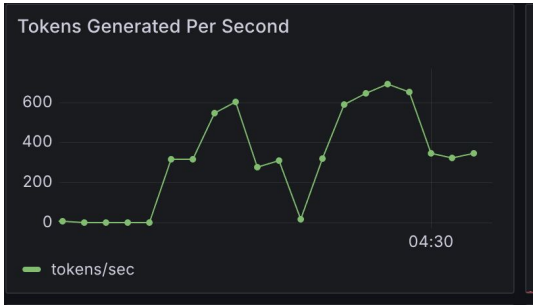
Cache

Pre- Cache
(19 sec)





Post- Cache
(13 sec)



*can't directly say tokens generated is proportional to cache rate but if operated on a similar queries then yes.

Few observations (Consolidated)

Test	Throughput
Without GPU Acceleration (5 min)	1
With GPU Acceleration (5 min)	52

Please find full report here: https://github.com/Mati02K/CounselGPT/tree/main/benchmark/results

Few observations (Consolidated)

Test	Max VUS	P95 Latency (sec)
Load	4	8.09
Stress	25	41.47*
Spike	25	29.39
Soak (10 min)	15	39.73
Soak (2 hour)	15	36.2 (1% failure)

- Significant loss - 20 %

Please find full report here: https://github.com/Mati02K/CounselGPT/tree/main/benchmark/results

Few observations (Consolidated)

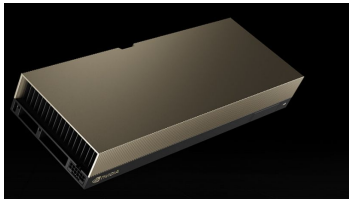
Model	Model Performance Score
Llama - 4 Bit	75%
Qwen - 8 Bit	90%

Used LLM-AS_Judge and Sentence Transformer

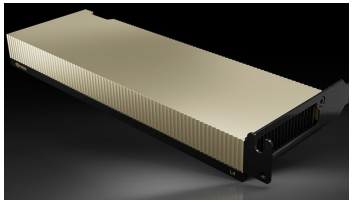
Please find full report here: https://github.com/Mati02K/CounselGPT/tree/main/benchmark/results

GPU's

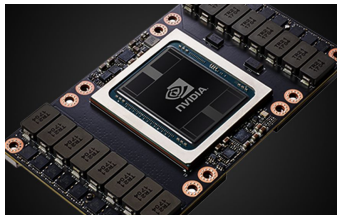
Metric / GPU	NVIDIA L4	NVIDIA L40	Tesla V100 (16–32GB)
Architecture	Ada Lovelace (low-power)	Ada Lovelace (high-end)	Volta (older gen)
FP16 TFLOPs	~30 TFLOPs	~90 TFLOPs	~15.7 TFLOPs
INT8 TOPS	~242 TOPS	~725 TOPS	~125 TOPS
Power Draw (TDP)	72W	300W	250W
Tokens/sec (7B model, llama.cpp)	~15–25 t/s	~40–60 t/s	~25–35 t/s
Memory Bandwidth	~300 GB/s	~650 GB/s	~900 GB/s



Nvidia L40



Nvidia L4

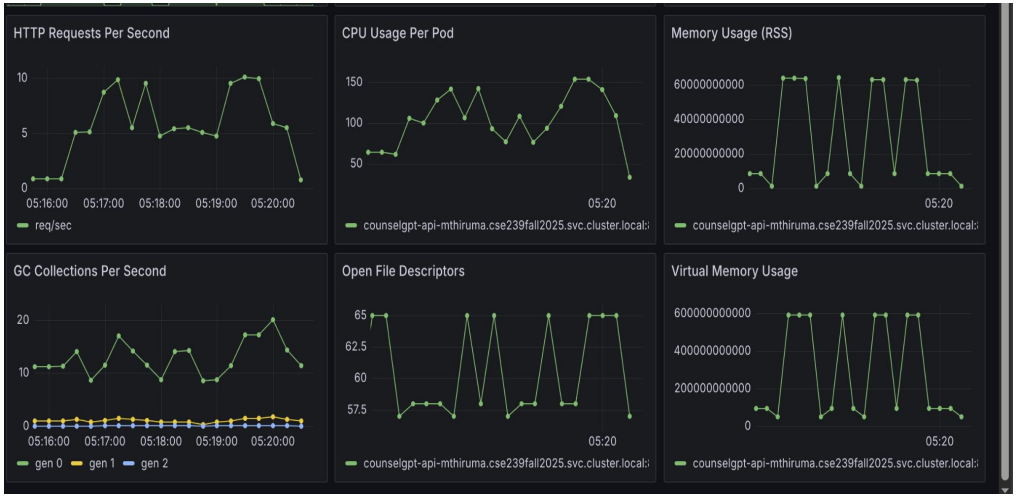


Nvidia Tesla V100

- **L4 is weaker → each token takes longer**, so the GPU stays busy more → **higher utilization %.**
- **V100/L40 are much faster** → they finish the same workload quickly → often sit idle → **lower utilization %.**

Baskin Engineering

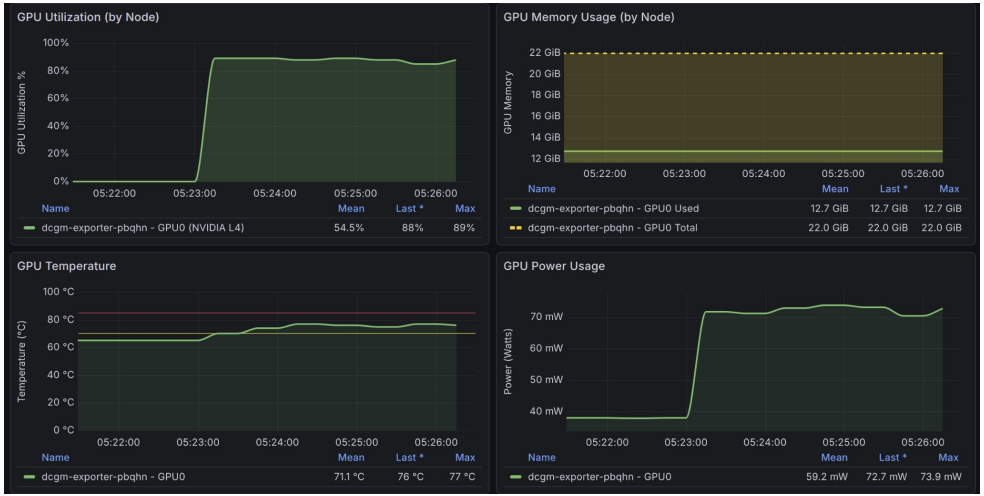
Few Snapshots from our Grafana Metrics



Nautilus Grafana (GPU Utilization): Nautilus GPU
Nautilus Grafana (LLM and Other Metrics): Nautilus LLM and Other metrics
GCP Grafana: GCP Grafana

Baskin Engineering

Few Snapshots from our Grafana Metrics(cont)



Nautilus Grafana (GPU Utilization): Nautilus GPU
Nautilus Grafana (LLM and Other Metrics): Nautilus LLM and Other metrics
GCP Grafana: GCP Grafana

Baskin Engineering

Outcomes and Tech Debt

- Nautilus is more more more speed than GCP, which is expected.
- We observed 38 VUS for nautilus (without cache) and 8 VUS for GCP without any throttle for normal queries. GPU matters performance wise, efficient wise.
- Apart from availability, not a good idea to use CPU.
- 8 bit performed better than 4 bit in all factors. (Due to model architecture). Neural Nets are black box.
- Semantic Caching comes at cost initially but if done properly and when serving large users will be a huge hit since the drop is 10x.
- Cold Start - Model loaded on new HPA instance was affecting a bit, but there is no enough resources for warm start.

Takeaways / Summary / Future Work

- We started with getting one ‘Hi’ msg at 2 mins inference time to getting the same in milliseconds and scaling it with the resource constraint.
- Utilization, router, availability, Horizontal scaling, monitoring etc was done with optimal practices.
- Llama.cpp (edge) v/s vLLM (since we have Data Center GPU’s and our fine tuned model relies on System prompt and KV caching using paged attention should be efficient and achieve high throughput on paper)

Efficient Memory Management for Large Language Model Serving with *PagedAttention*

Woosuk Kwon[1,*] Zhuohan Li[1,*] Siyuan Zhuang[1] Ying Sheng[1,2] Lianmin Zheng[1] Cody Hao Yu[3]
Joseph E. Gonzalez[1] Hao Zhang[4] Ion Stoica[1]
[1]UC Berkeley [2]Stanford University [3]Independent Researcher [4]UC San Diego

Abstract

High throughput serving of large language models (LLMs) requires batching sufficiently many requests at a time. However, existing systems struggle because the key-value cache (KV cache) memory for each request is huge and grows and shrinks dynamically. When managed inefficiently, this memory can be significantly wasted by fragmentation and redundant duplication, limiting the batch size. To address this problem, we propose PagedAttention, an attention algorithm inspired by the classical virtual memory and paging techniques in operating systems. On top of it, we build vLLM, an LLM serving system that achieves (1) near-zero waste in KV cache memory and (2) flexible sharing of KV cache within and across requests to further reduce memory usage. Our evaluations show that vLLM improves the throughput of popular LLMs by 2-4× with the same level of latency compared to the state-of-the-art systems, such as FasterTransformer and Orca. The improvement is more pronounced with longer sequences, larger models, and more complex decoding algorithms. vLLM's source code is publicly available at https://github.com/vllm-project/vllm.

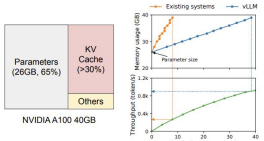


Figure 1. *Left:* Memory layout when serving an LLM with 13B parameters on NVIDIA A100. The parameters (gray) persist in GPU memory throughout serving. The memory for the KV cache (red) is (de)allocated per serving request. A small amount of memory (yellow) is used ephemerally for activation. *Right:* vLLM smooths out the rapid growth curve of KV cache memory seen in existing systems [31, 60], leading to a notable boost in serving throughput.

the cost per request—of *LLM serving* systems is becoming more important.

Baskin Engineering

Project Link

- Counsel GPT
- Code Link: GitHub Link
- LoRA Training: GitHub Link

Baskin Engineering

Thank you!