GenAI Nantes

“Intègre un assistant vocal dans ta webapp”

Raconte

Godefroy de Compreignac

CEO @Lonestone
Founder @Raconte.ai

11 juin 2025 à 19h

10 Rue Magdeleine, 44200 Nantes

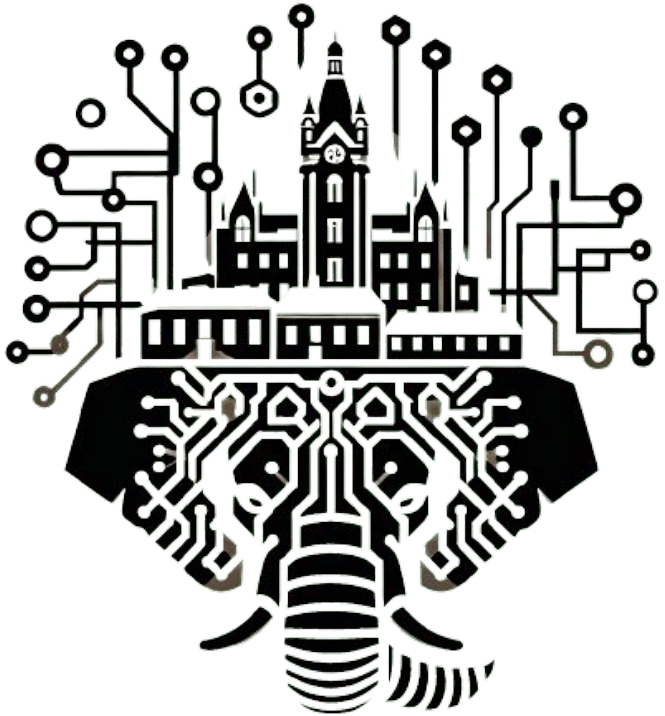
chez Sii

RCA zenika [sfeir] lonestone

Qui n’est jamais venu au meetup Gen AI Nantes ?

GenAI Nantes

🎤 15 événements / an
🏴‍☠️ 1 hackathon (Shift)
👷 1 workshops
❤️ 1 communauté de 900p*



* 9.000p selon le syndicat des llamas 🦙

Qui vient pour la première fois? 👔

Qui est tech ? 

MLOps, dénoncez-vous ! 💪

Qui héberge un LLM sur sa propre infra ? 🤑

Qui intègre la voix dans des apps ? 📣

Qui cherche un job dans la GenAI ? 🏋️

Qui recrute dans la GenAI ? 🚀

Qui souhaite s’associer dans la GenAI ? 💑

Schedule

👀 1- News

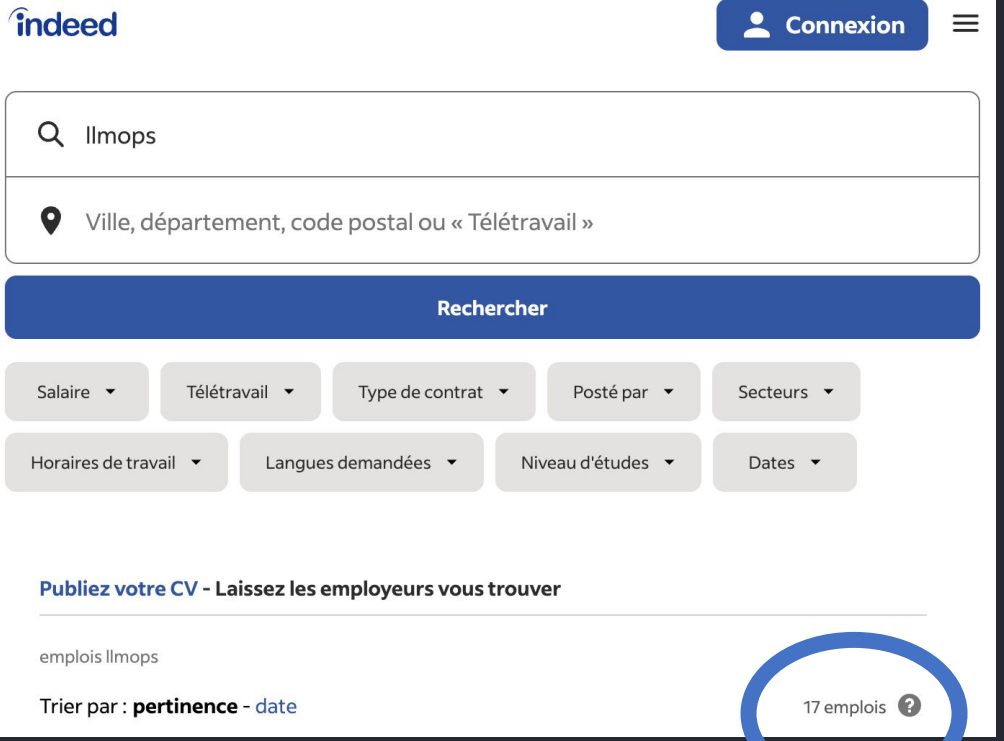
🔍 2- Raconte.ai

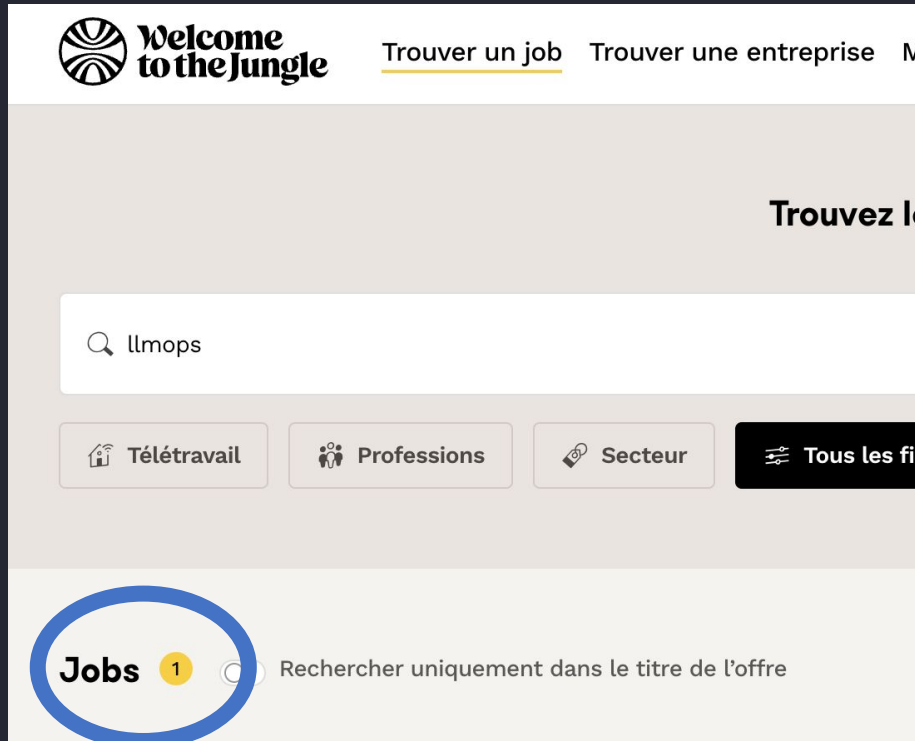
🍻 3- Enjoy

News tech

Serving LLMs at scale
LLMOps 101

A technical deep dive into the challenges and solutions for serving large language models at scale across datacenter infrastructure.


indeed
Connexion
llmops
Ville, département, code postal ou « Télétravail »
Rechercher
Salaire
Télétravail
Type de contrat
Posté par
Secteurs
Horaires de travail
Langues demandées
Niveau d'études
Dates
Publiez votre CV - Laissez les employeurs vous trouver
emplois llmops
Trier par : pertinence - date
17 emplois


Welcome to the Jungle
Trouver un job
Trouver une entreprise
Trouvez l
llmops
Télétravail
Professions
Secteur
Tous les fi
Jobs 1
Rechercher uniquement dans le titre de l'offre

There is a golden rule in IT:

"Limit everything in space and time"

Once upon a time, a very big prompt

```
curl https://api.openai.com/v1/chat/completions \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer YOUR_API_KEY" \
  -d "{
    \"model\": \"gpt-4o\",
    \"messages\": [
      {
        \"role\": \"system\",
        \"content\": \"$(cat bible.txt | sed 's/\"/\\\"/g')\"
      },
      {
        \"role\": \"user\",
        \"content\": \"What is the answer to the universe and everything?\"
      }
    ]
  }"
```

Once upon a time, a very big prompt

Gemini 2.5 Pro – Extremely High Latency on Large Prompts (100K–500K Tokens)

Posted on 05-05-2025 03:29 PM

Y

speedy

Bronze 1

Post Options

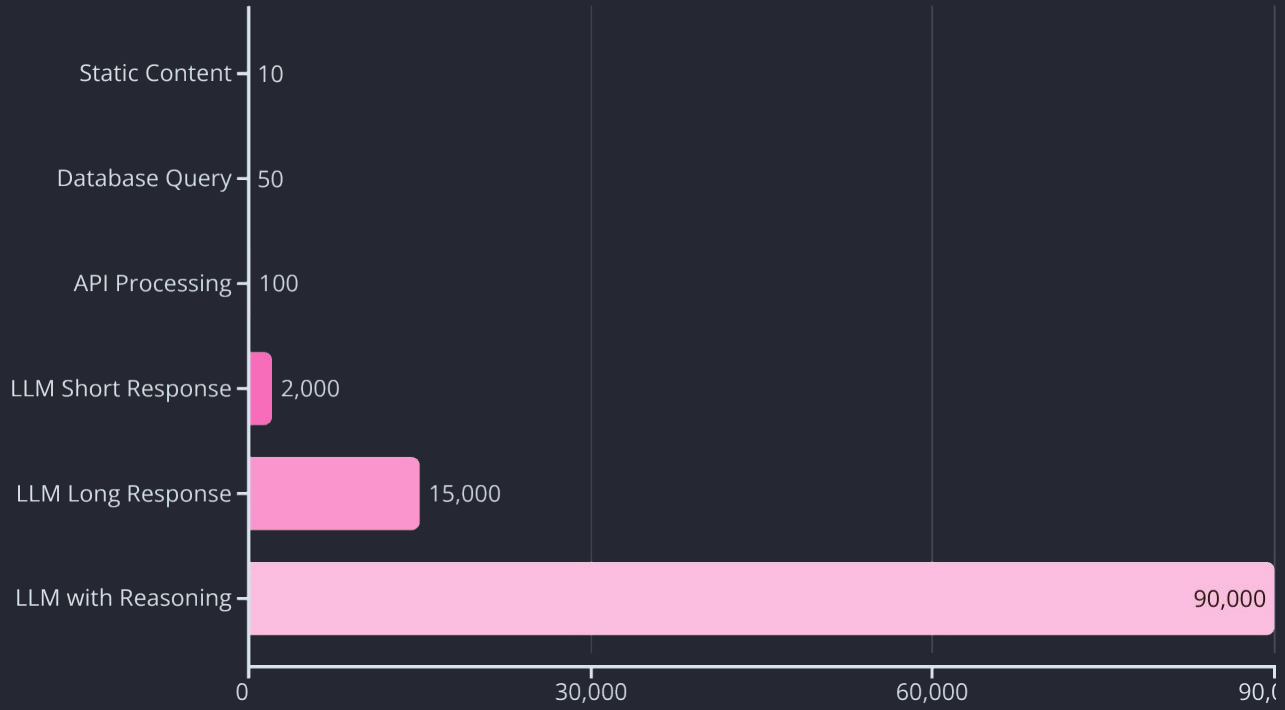
Hi all,

I'm using the model `gemini-2.5-pro-preview-03-25` through Vertex AI's `generateContent()` API, and facing very high response latency even on one-shot prompts.

Current Latency Behavior:
- Prompt with 100K tokens → ~2 minutes
- Prompt with 500K tokens → 10 minutes+
- Tried other Gemini models too — similar results

This makes real-time or near-real-time processing impossible.

Request Duration Distribution (ms)



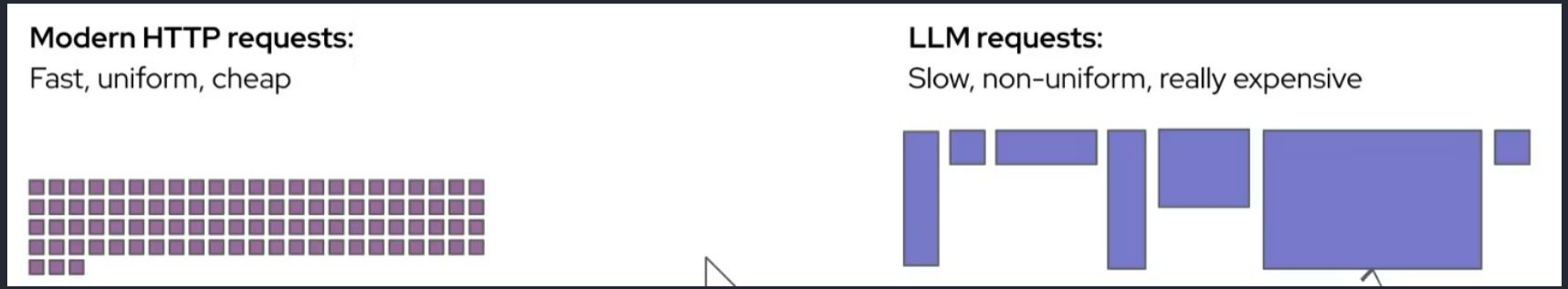
Traditional vs. LLM Hosting

Traditional Apps

- Predictable request patterns
- Low processing times
- Graceful stop with short grace period

LLM

- Highly variable request duration
- Long-running jobs
- Costly retry
- Long-lived connections



Requests to LLMs are not Idempotent

- Temperature=0
- Seed=4242424242
- Floating point: non-deterministic
- No efficient retry

```
curl https://api.openai.com/v1/responses \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  -d '{
    "model": "gpt-4.1",
    "temperature": 0,
    "seed": 4944116822809979520,
    "input": "Tell me a story."
  }'
```

Anthropic Status page

Conclusion:

We need a new architecture

Disclaimer:

I do not work for a cloud provider.
I’ve never served LLM in production.
Just some readings.

1 - LLM Fundamentals

LLM Fundamentals: Tokens

What are Tokens?

Atomic units of text processing

Words split into subword pieces

Tokenizer

Converts raw text to token IDs

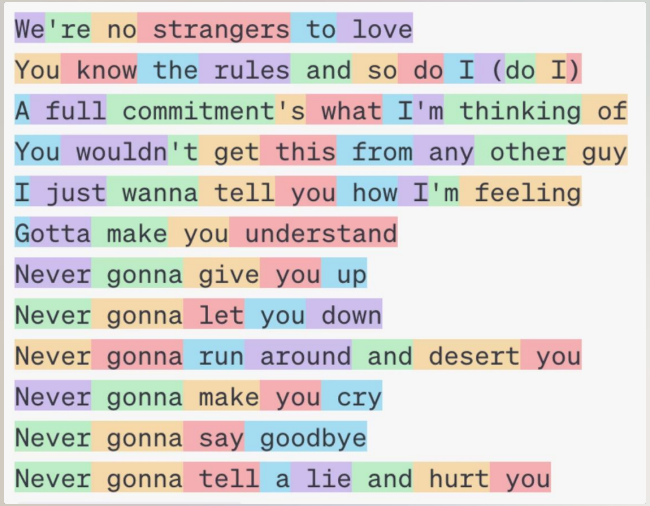
BPE algorithm

Example:

"She enjoys playing tennis"

→ ["She", "enjoy", "##s", "playing", "tennis"]

→ [215, 4523, 987, 7632, 3241]



LLM Fundamentals: Vocabulary

Vocabulary

Fixed set of subword pieces

Typically 50K-200K entries

Coding

Vocabulary must contain:

[{ ; , \t …

World Languages

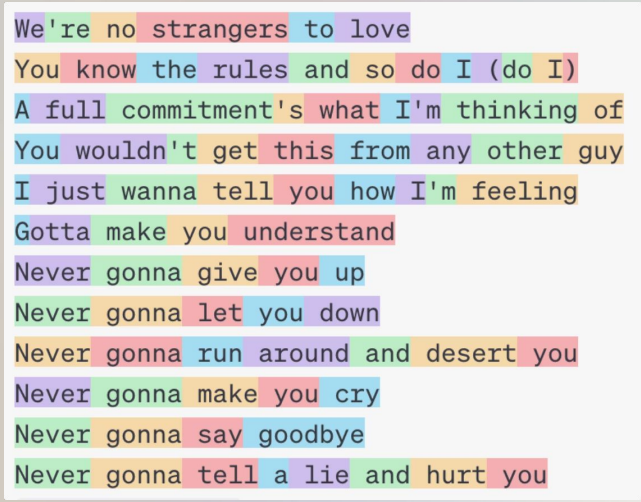
Thai: พยัญชนะ

French: éèàù

LLM features

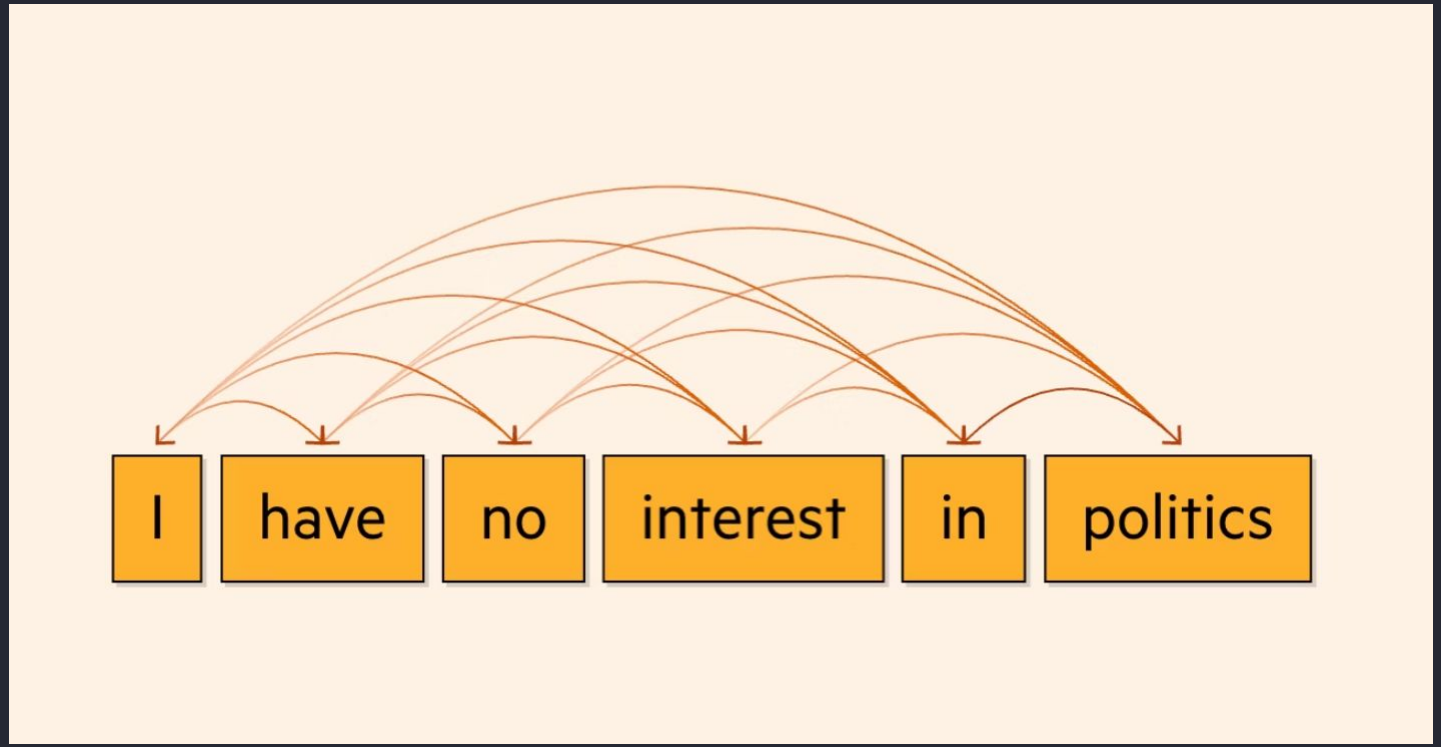
<!StartOfReasoning!>

<!EndOfSequence!>



Attention is all you need

Self-attention mechanism: find correlations between tokens



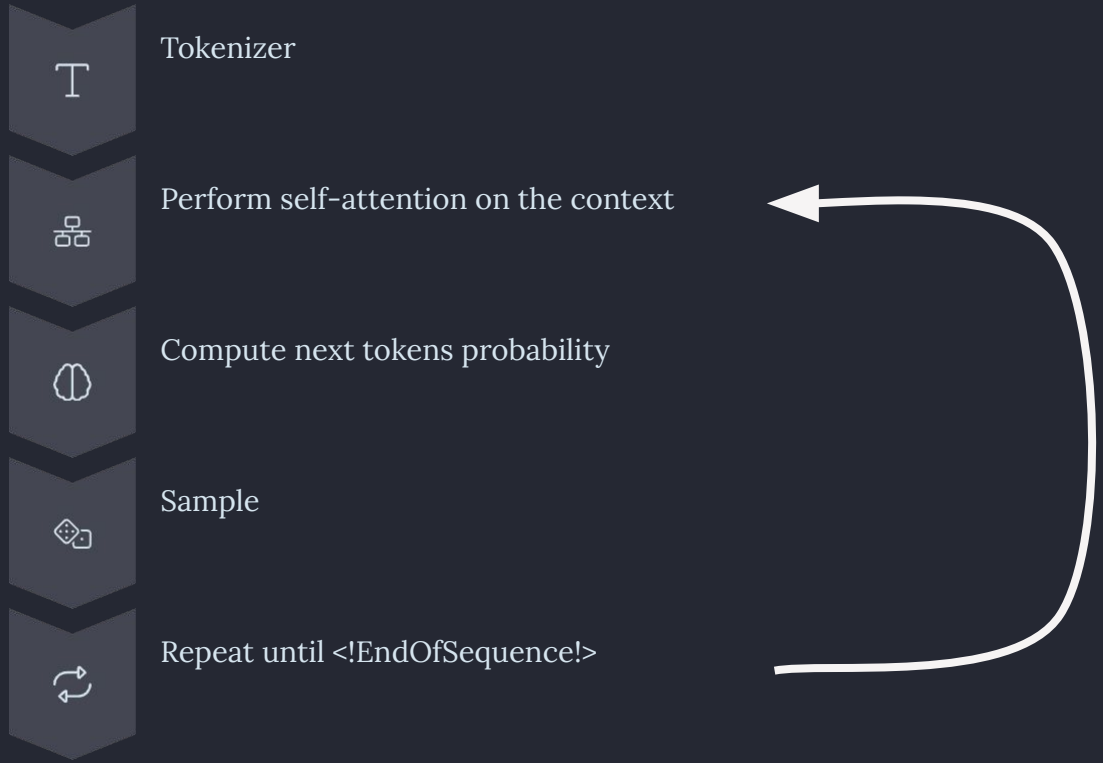
Complexity: $O(n^2)$

100 tokens vs. 1000 tokens

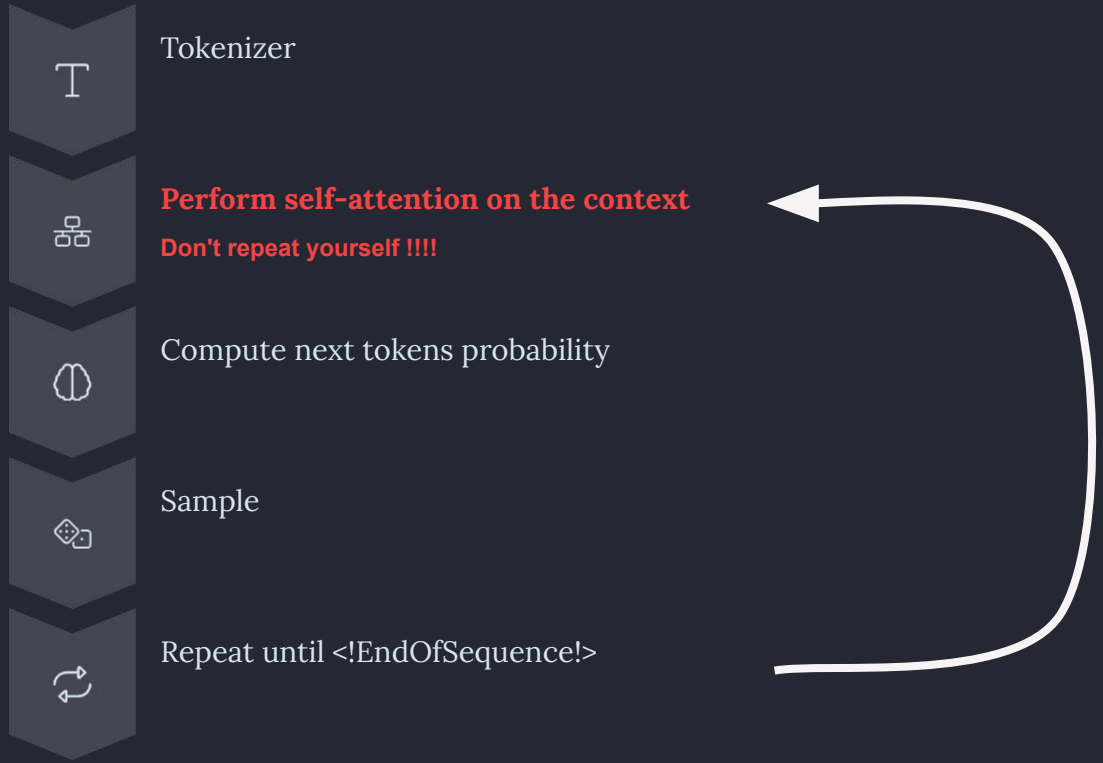
Cost: **x100**

K/V Cache

Token Generation Process



Token Generation Process



K/V Cache

What is K/V Cache?

Storage of self-attention from previous tokens

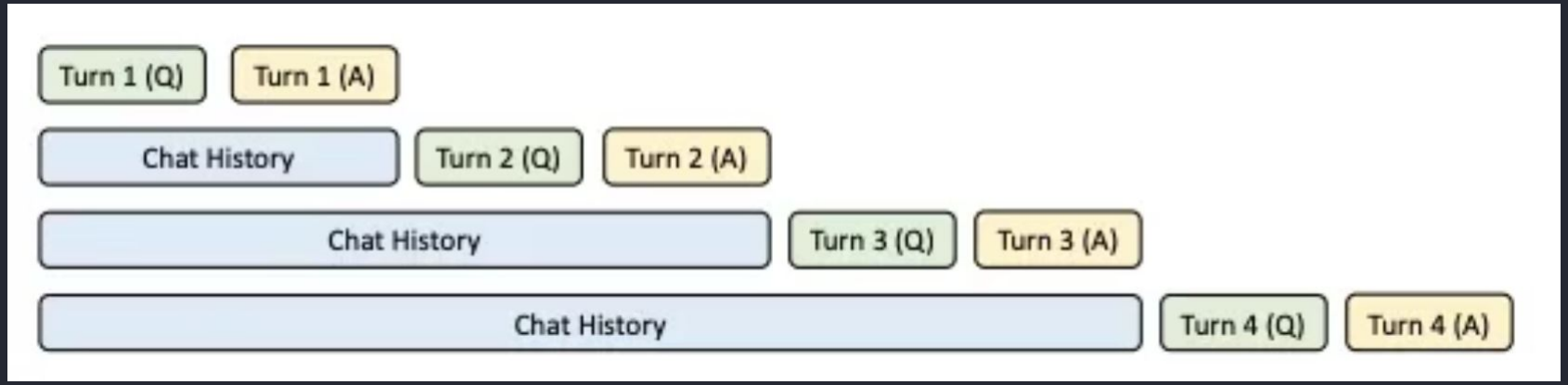
Performance Benefit

Perform masked self-attention on the last generated token only

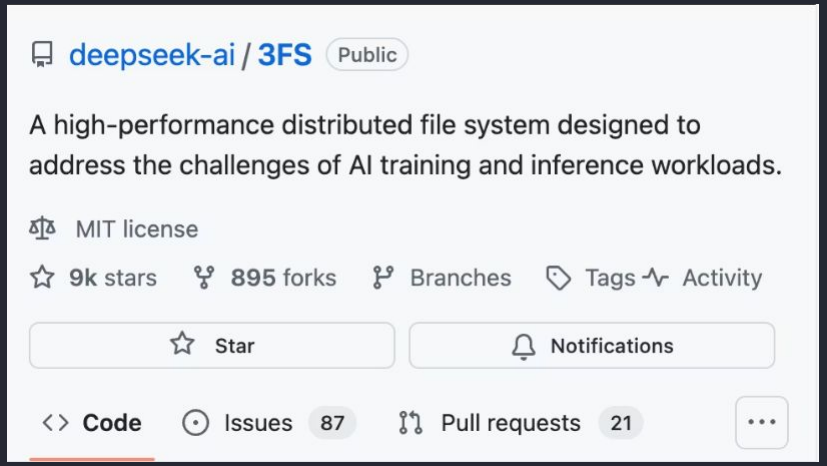
K/V Cache grow linearly with sequence length

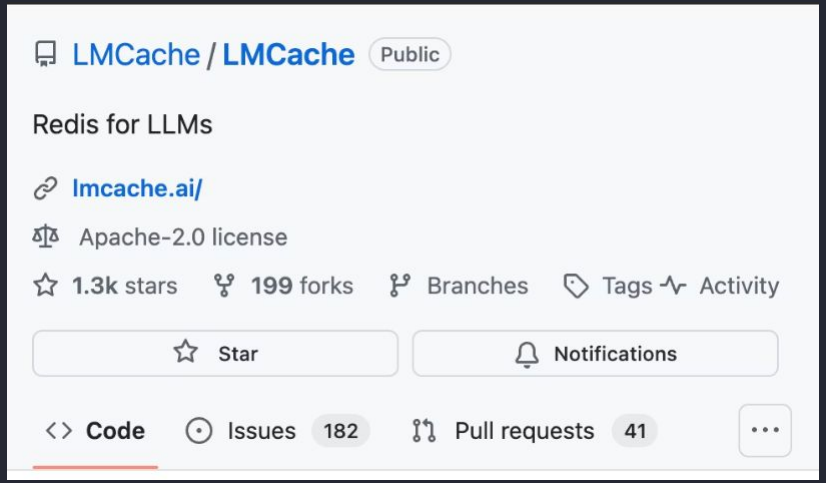
Prefix caching for multiples queries

Prefix caching in ChatGPT

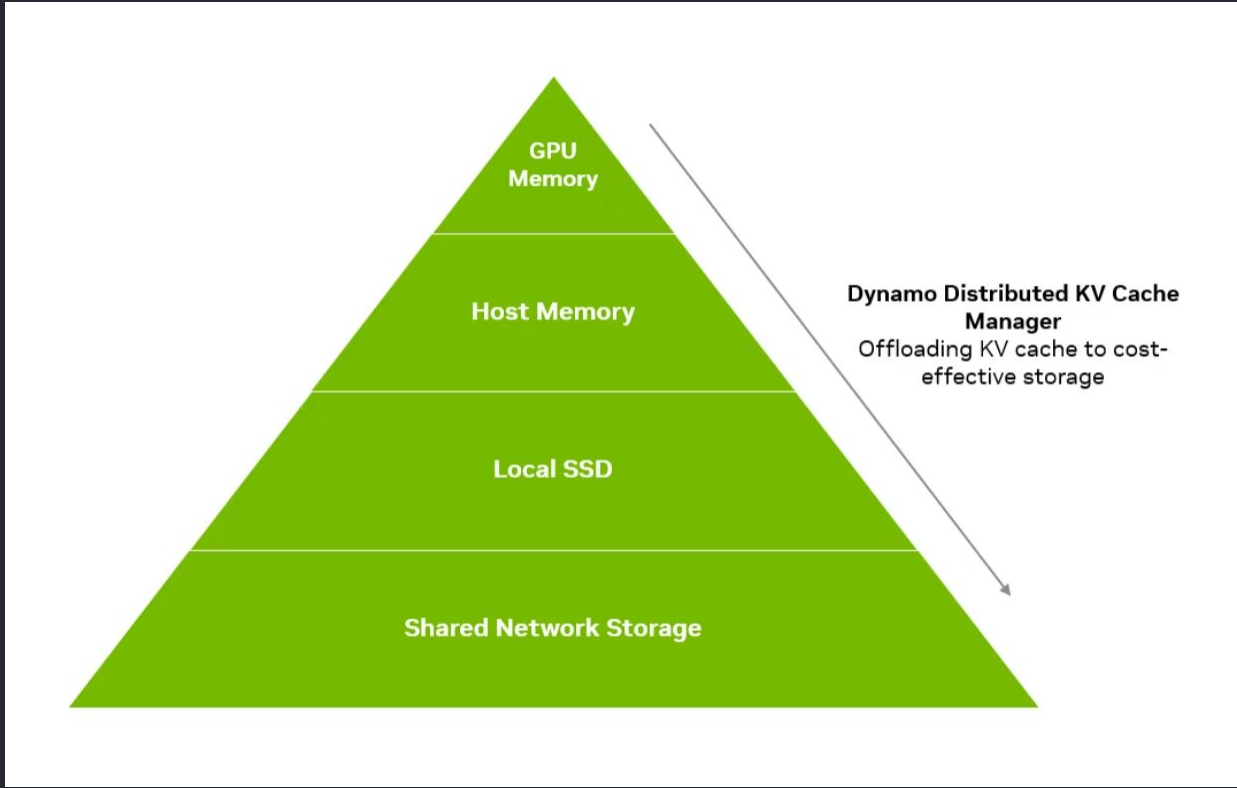


Distributed K/V Cache





Distributed K/V Cache



Takeaway:

A 3 times longer prompt cost:

- 9x more computing
- 3x more memory

K/V Cache Size Economics

0.3-1.5 MB

Per Tokens

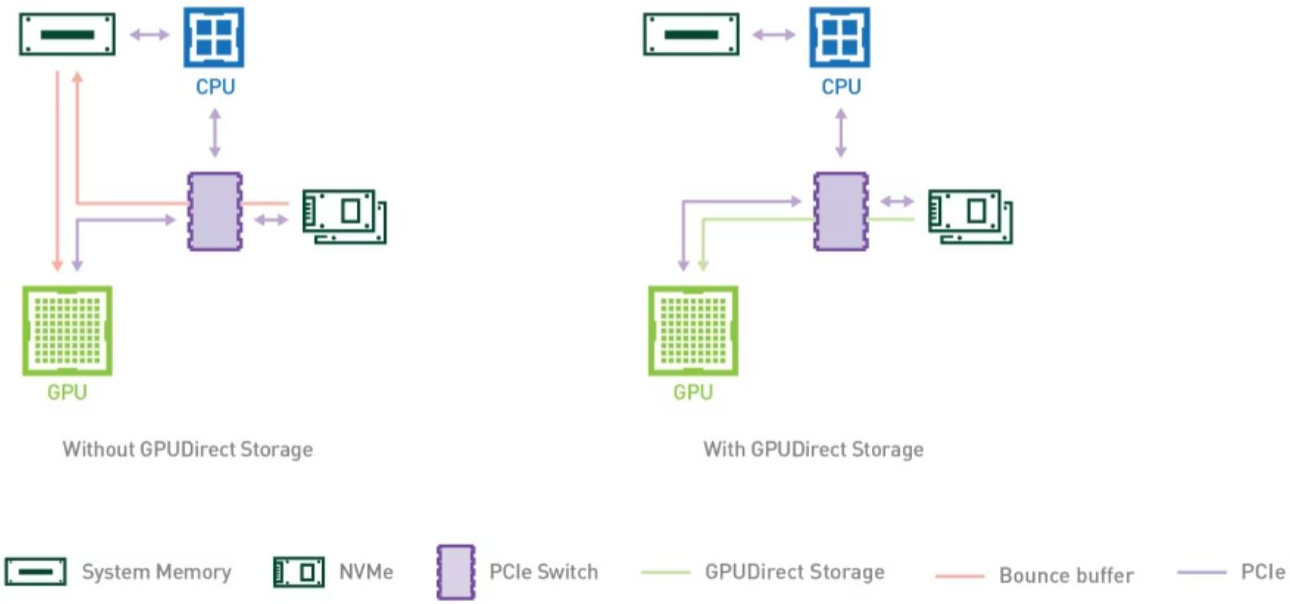
0.3-1.5 GB

For 1K Context

10k Tokens

System prompt of Claude

Load K/V Cache via RDMA (Infiniband, GPUDirect...)



Hardware

GPU Architecture

High Bandwidth Memory (HBM)

Large capacity

Stores model weights

High throughput

On-PCB

H100:

- 80 GB
- 3.3 TB/s

On-chip Cache (L2)

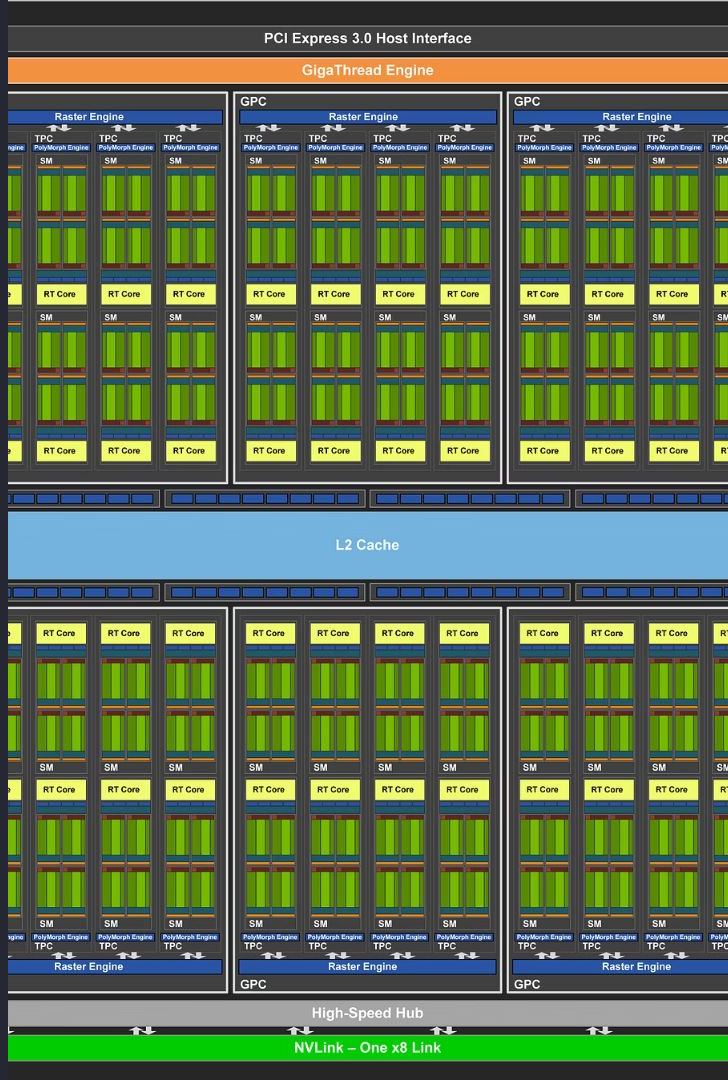
Very low latency

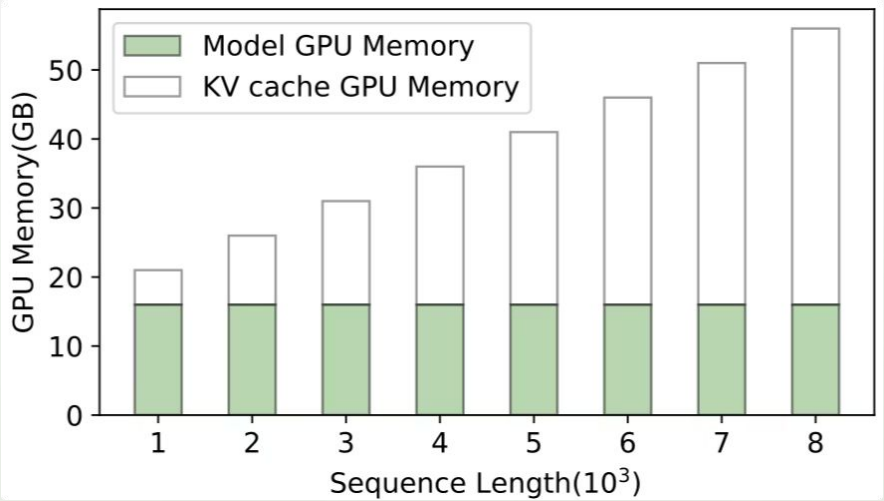
Limited size

On-chip

H100:

- 60 MB
- 4 TB/s





Memory requirements

Model weights

Llama3.3 70B

BF16

Total: 140 GB

K/V cache

10k token input

1k token output

0.35 MB per cache entry

Total: ~4 GB

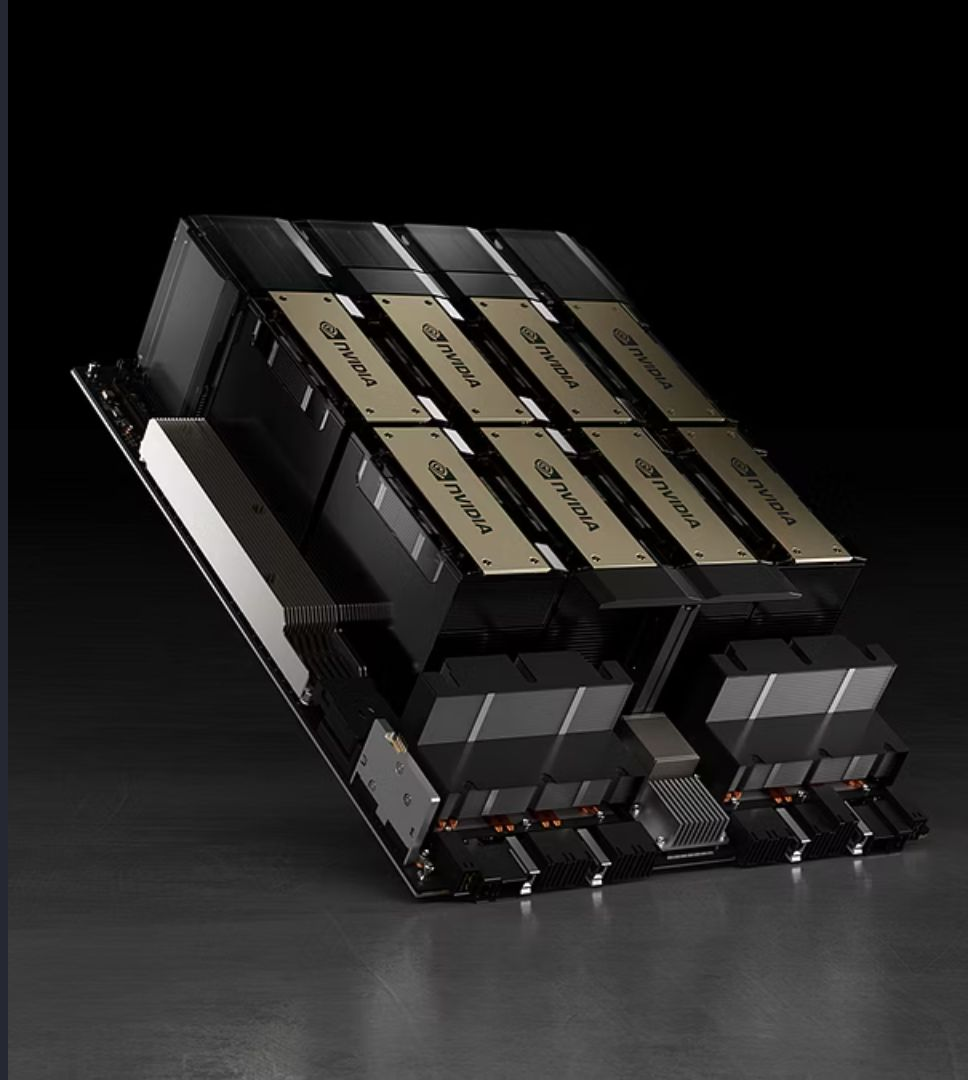
Multi GPU inference

Because 140GB + 4GB would OOM a single H100 HBM

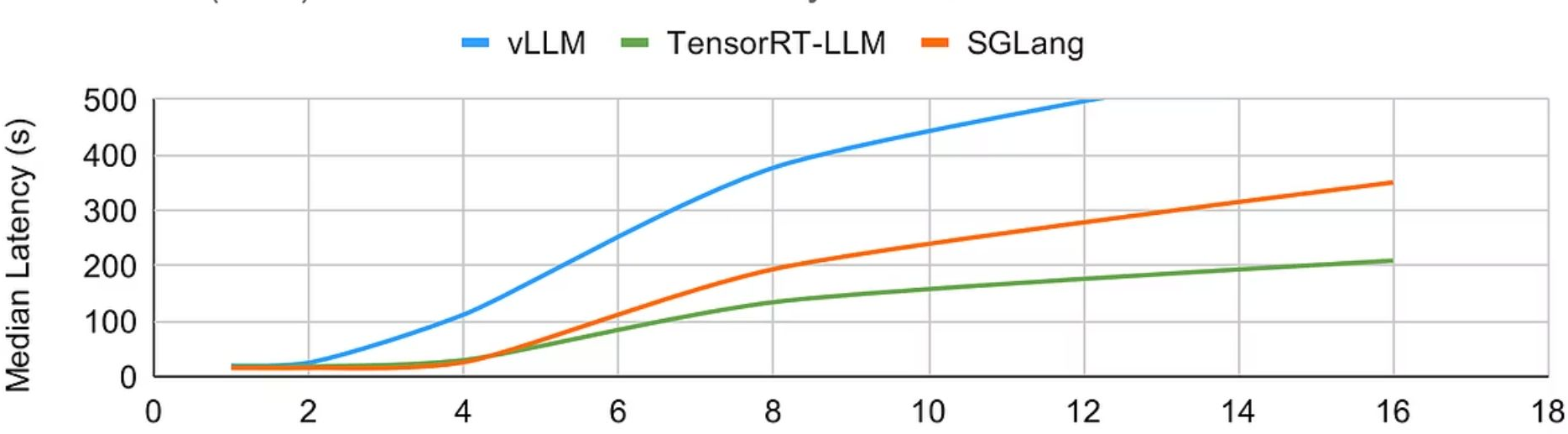
4x H100 → 320 GB HBM

Llama3.3:

35+1 GB per GPU



Inference



LLM Inference Engine

Popular Engines	Operates GPU(s)	API
vLLM, TensorRT-LLM, SGLang, llama.cpp	K/V cache, batching, quantization, MoE, tokenizer, speculative decoding	OpenAI-compatible REST endpoints HDW: CUDA, Triton

2 stages: Prefill vs. Decoding Different Workloads

Prefill Phase

Load K/V prefix from distributed cache.

Process entire prompt in parallel.

Output a K/V cache.

Decoding Phase

Generate one token at a time.

Generate K/V cache for the last token.

DistServe: Disaggregating Prefill and Decoding

Paper: https://arxiv.org/abs/2401.09670

	Context	Max Output	Input	Output	Latency	Throughput	Uptime
Mistral FR	33K	33K	$0.10	$0.30	0.38 s	143.1 tps	
Ubicloud bf16	33K	33K	$0.30	$0.30	0.98 s	32.87 tps	

Performance Metrics: TTFT & Tokens/Second

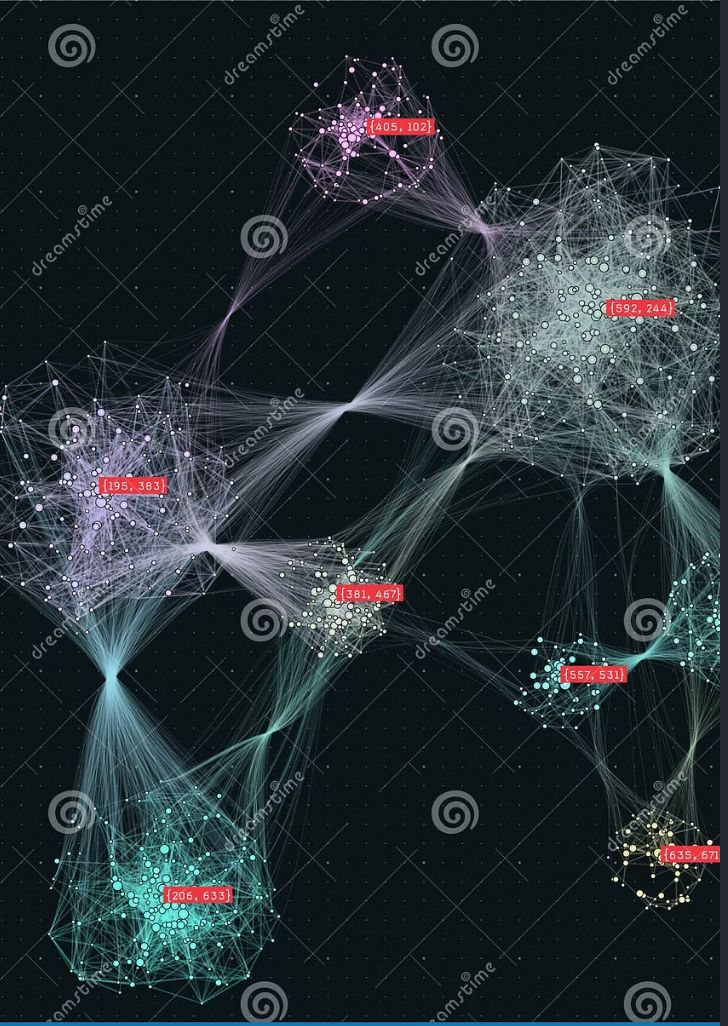
Time To First Token (TTFT)

Latency from request to first generated token.

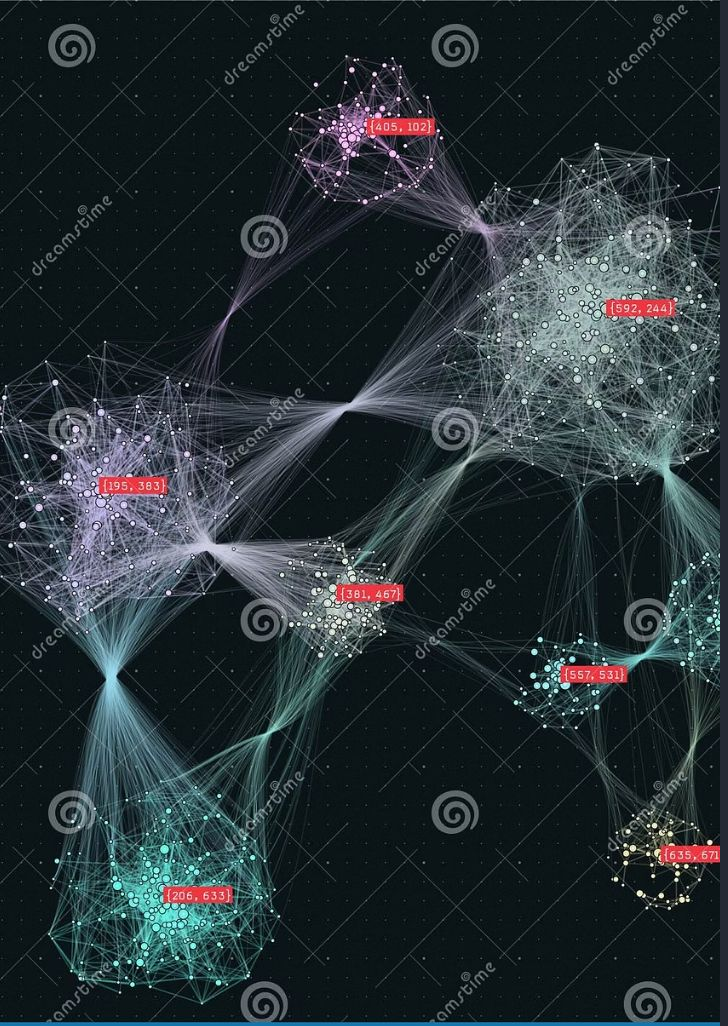
Critical for user experience.

Tokens Per Second

Throughput metric for generation speed. Higher is better for long outputs.

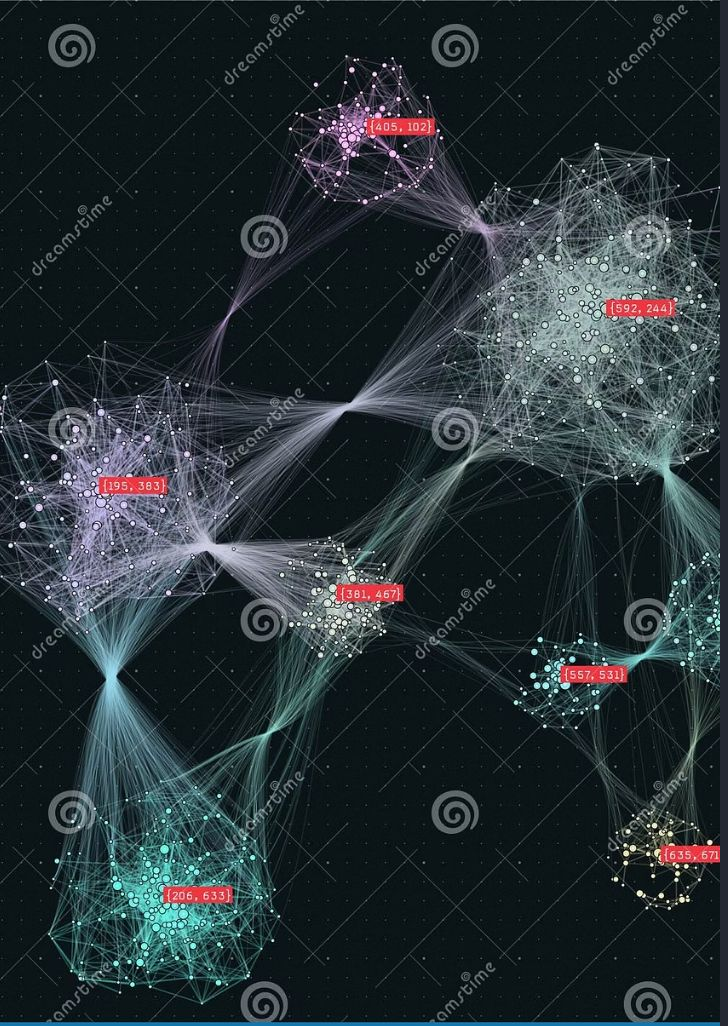


Prefill stage (prompt processing)



Prefill stage (prompt processing)

- Load weights of layer 0 from HBM to L2 cache
- Process self-attention for layer 0 and token 0..N
- Unload weights
- Load weights of layer 1 from HBM to L2 cache
- Process self-attention for layer 1 and token 0..N
- Unload weights
- ...



Prefill stage (prompt processing)

Load weights of layer 0 from HBM to L2 cache

Process self-attention for layer 0 and token 0..N
👆 **Compute bound** 👆

Unload weights

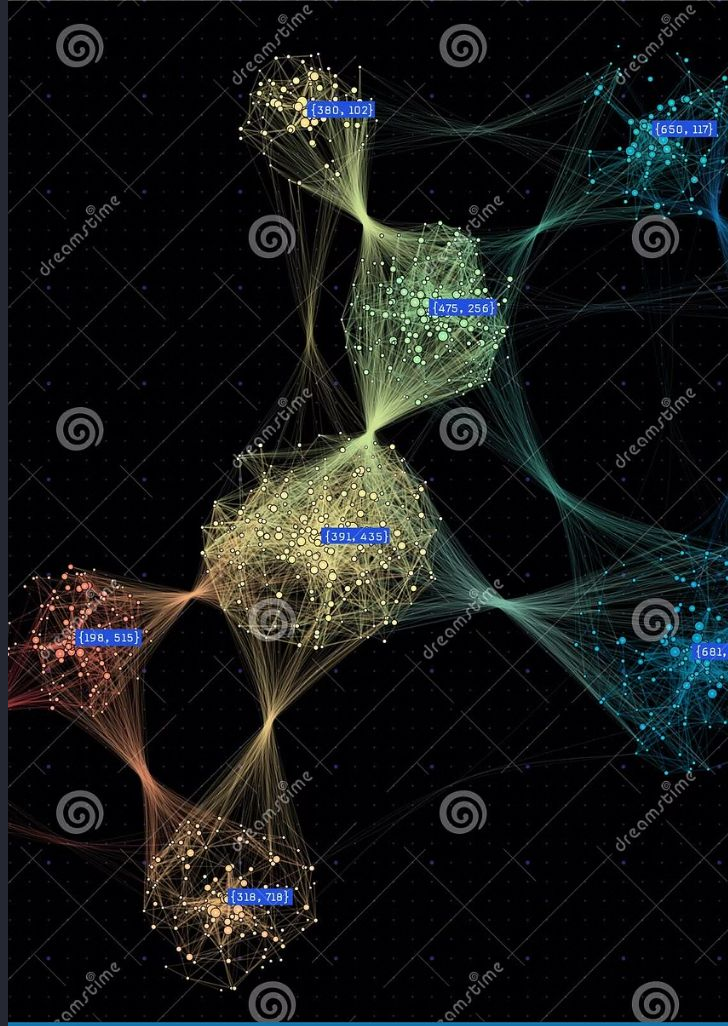
Load weights of layer 1 from HBM to L2 cache

Process self-attention for layer 1 and token 0..N
👆 **Compute bound** 👆

Unload weights

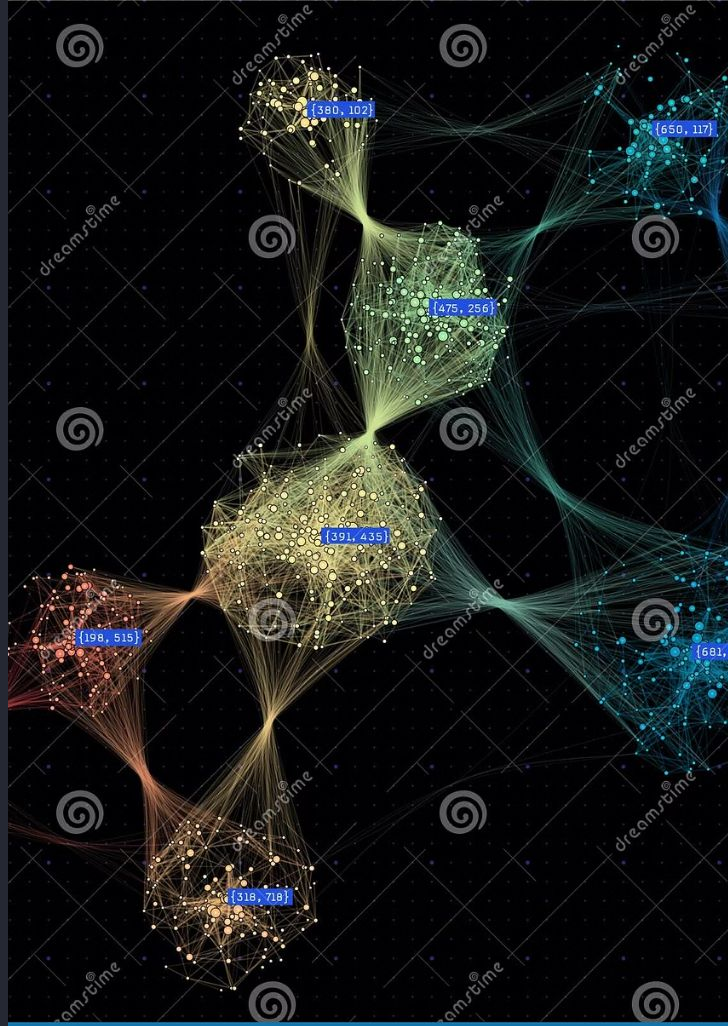
...

Decoding stage (output generation)



Decoding stage (output generation)

- Load K/V cache
- Load layer 0
- Process
- Load layer 1
- Process
- Load layer ..N
- Return the first token and repeat



Decoding stage (output generation)

Load K/V cache
👆 Grow linear with context size 👆

Load layer 0
👆 Grow linear with model size 👆

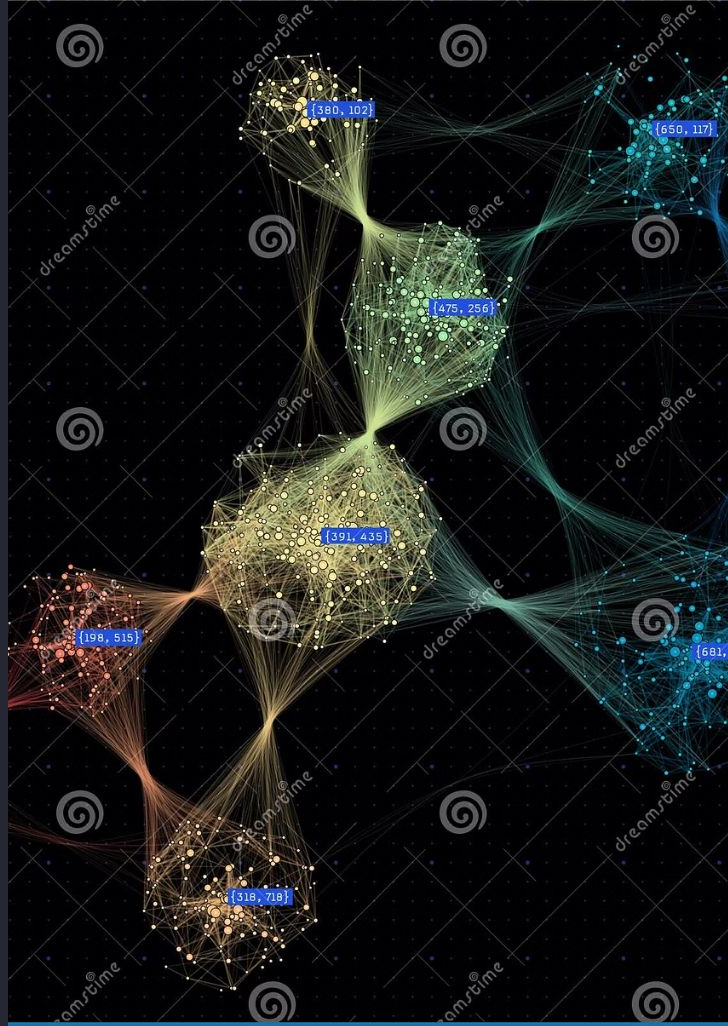
Process

Load layer 1
👆 Grow linear with model size 👆

Process

Load layer ..N
👆 Grow linear with model size 👆

Return the first token and repeat



Decoding stage (output generation)

Load K/V cache
👆 Bandwidth bound 👆

Load layer 0
👆 Bandwidth bound 👆

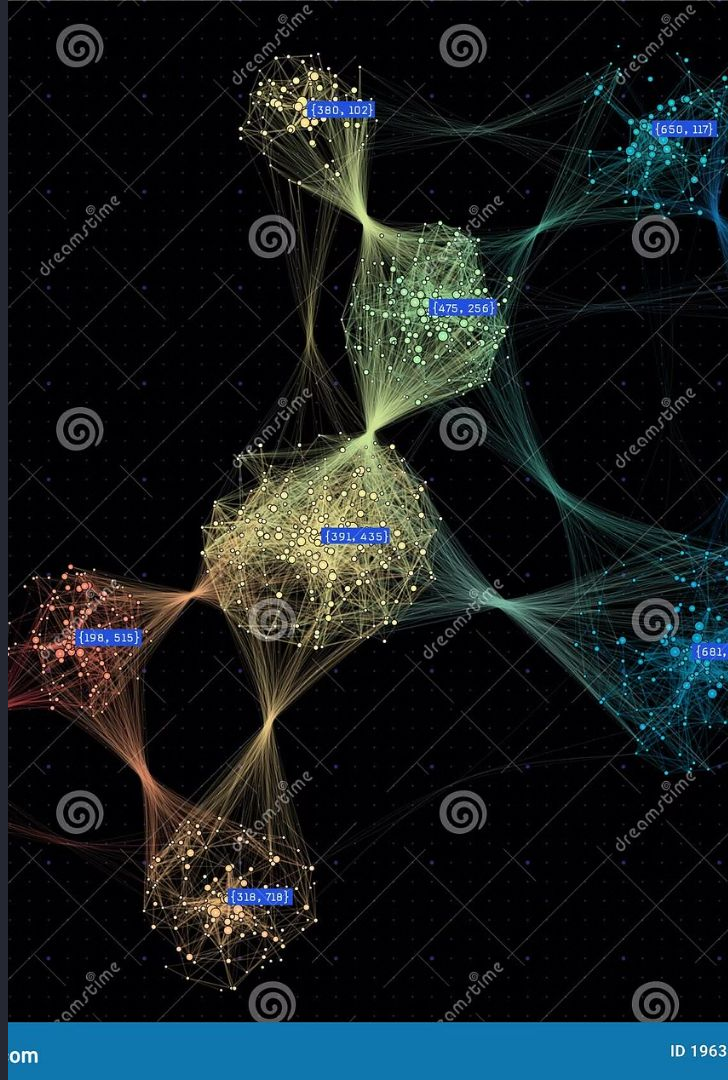
Process

Load layer 1
👆 Bandwidth bound 👆

Process

Load layer ..N
👆 Bandwidth bound 👆

Return the first token and repeat

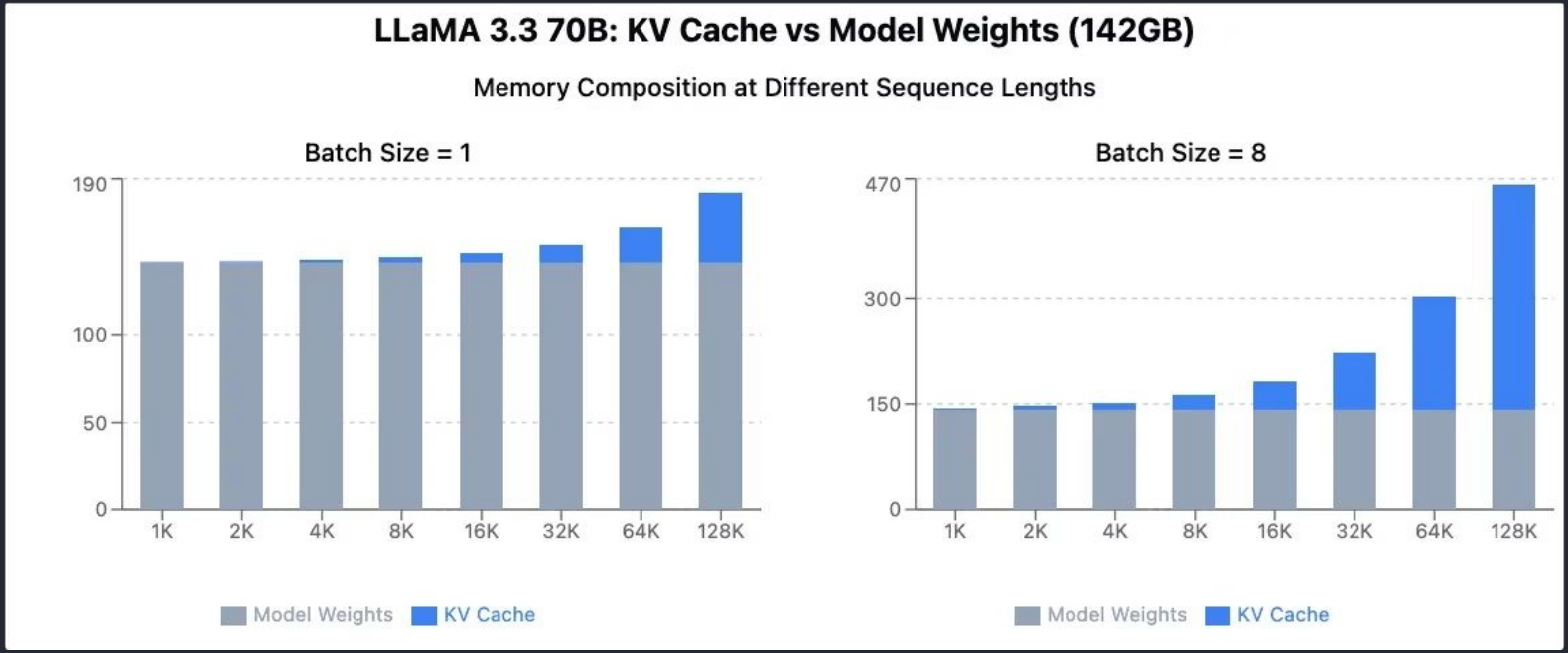


	Context	Max Output	Input	Output	Latency	Throughput	Uptime
Mistral FR	33K	33K	$0.10	$0.30	0.38 s	143.1 tps	
Ubicloud bf16	33K	33K	$0.30	$0.30	0.98 s	32.87 tps	

Takeaway

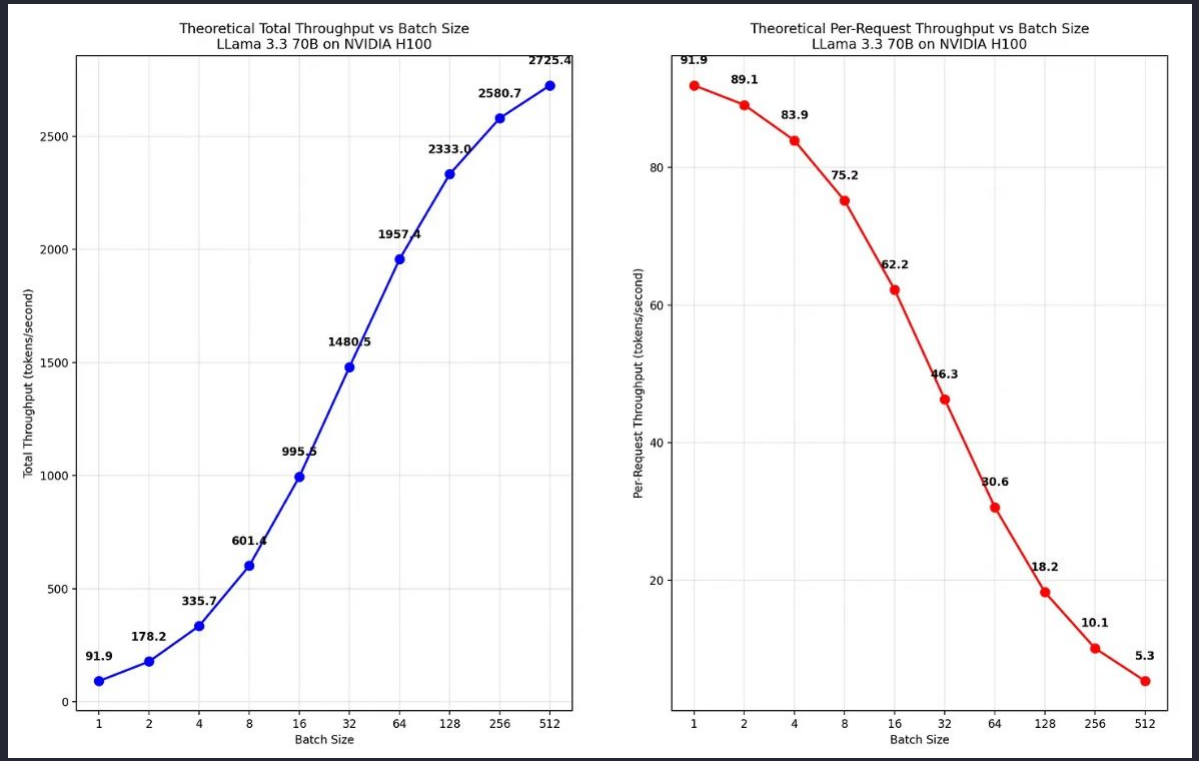
- Prefill is compute bound
- Decoding is bandwidth bound

Optimization: Batching



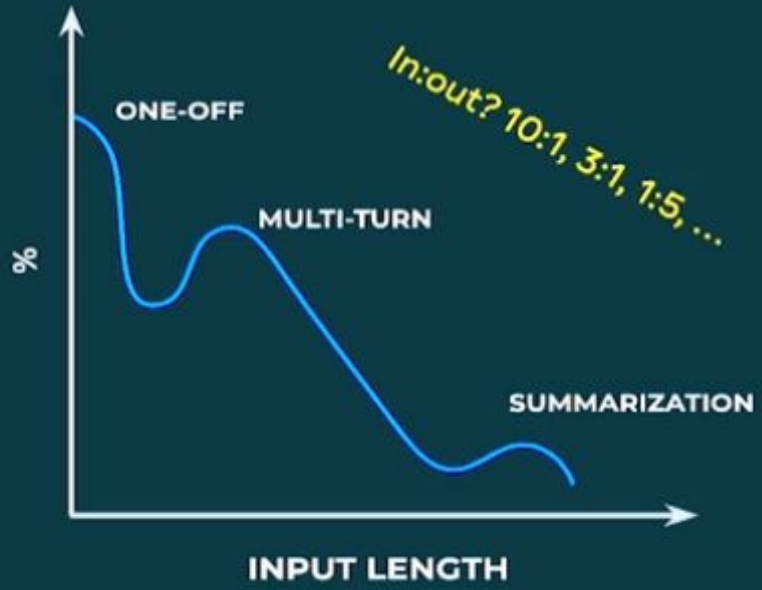
Decoding stage: Batching

Load layer once and compute more.



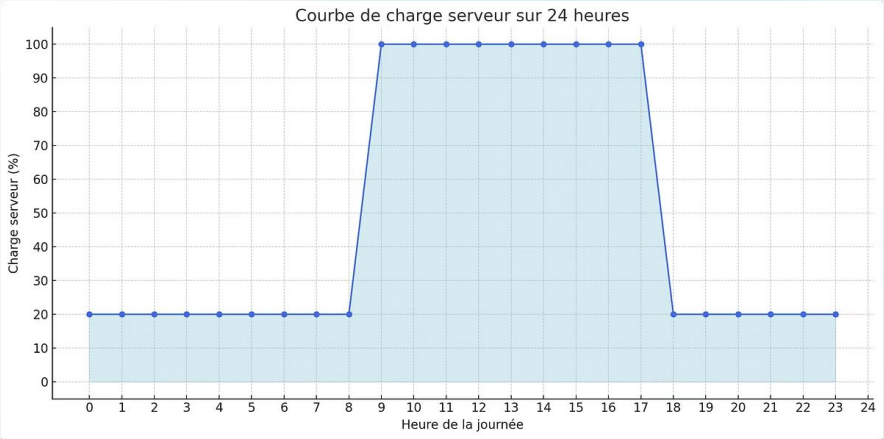
Large batch → more tok/s per GPU+, but less tok/s per request

A **trade-off** between cost and tok/s



Input length and traffic change over time

It requires careful scheduling and request routing.

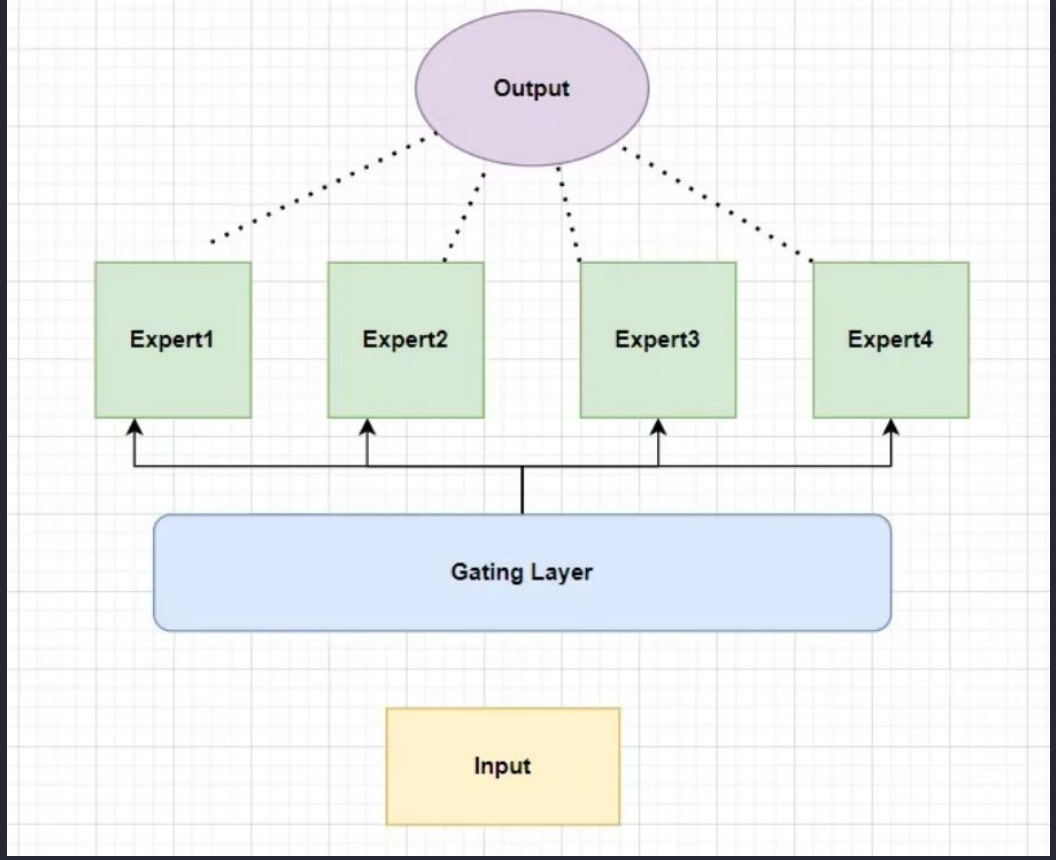


Capacity planning

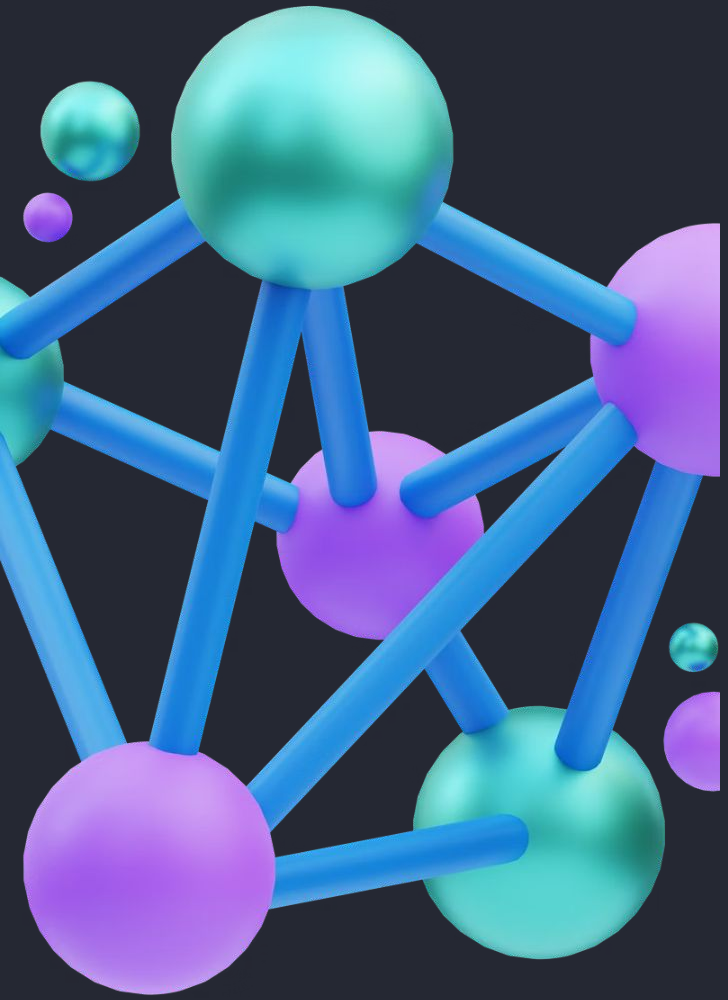
2 strategies during the night:

- scale-down (cheaper)
- or smaller batch (faster)

Optimization: parallelism



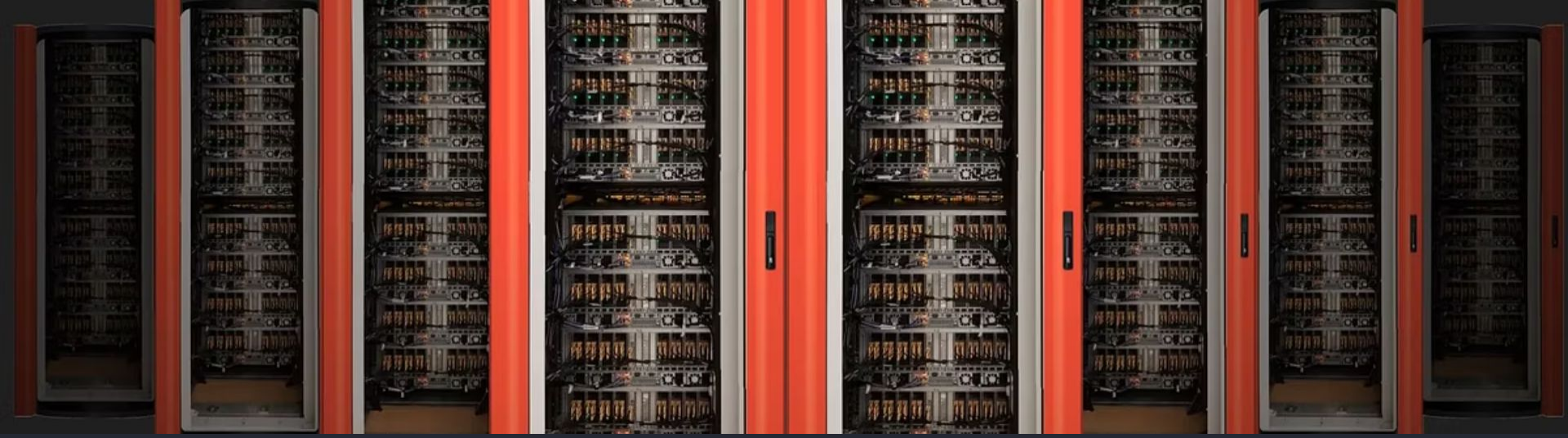
Mixture of Expert (MoE)



Tensor Parallelisms
Pipeline Parallelisms
Expert Parallelism

Pick one

Hardware:
LPU

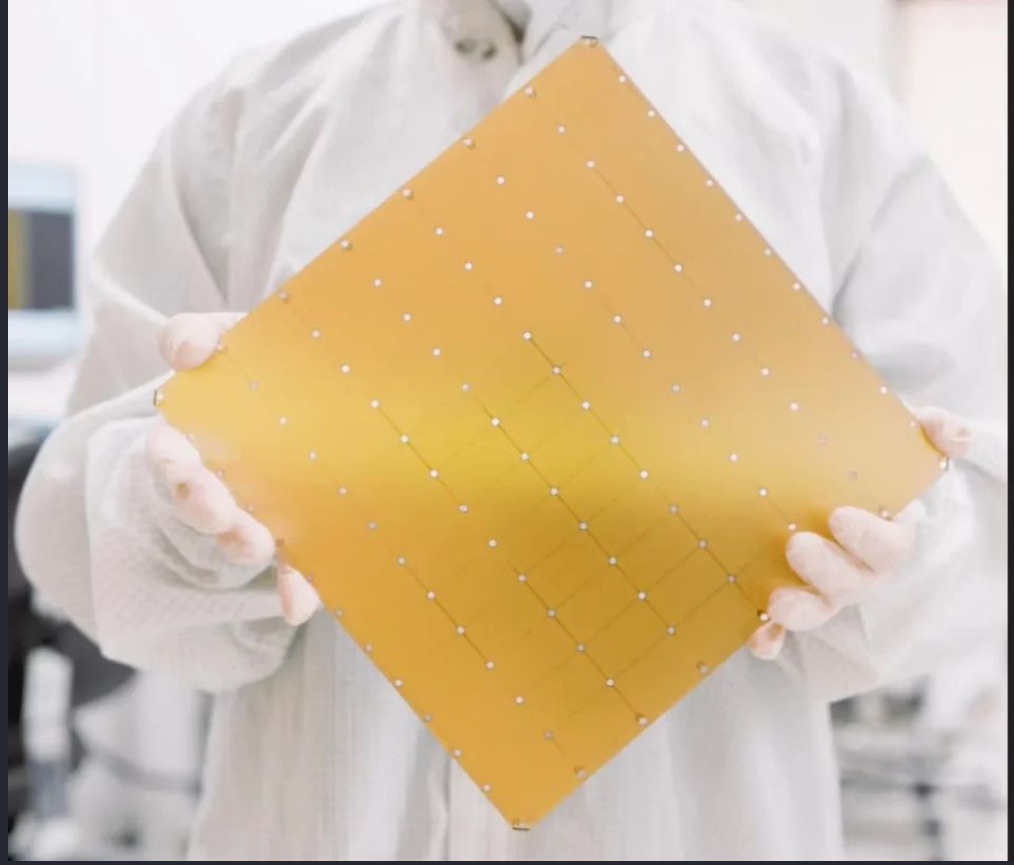


LPU - Groq

No HBM → only on-chip memory -> on-chip model weight

Less memory bandwidth bottleneck

~1000 LPU for running Llama3 70B



Cerebras Wafer-Scale Engine

The fastest AI chip on earth again

- 4 trillion transistors
- 46,225 mm2 silicon
- 900,000 cores optimized for sparse linear algebra
- 5nm TSMC process
- 125 Petaflops of AI compute
- 44 Gigabytes of on-chip memory
- 21 PByte/s memory bandwidth
- 214 Pbit/s fabric bandwidth

Cerebras & LeChat: YOLO (size: 21.5 cm x 21.5 cm)

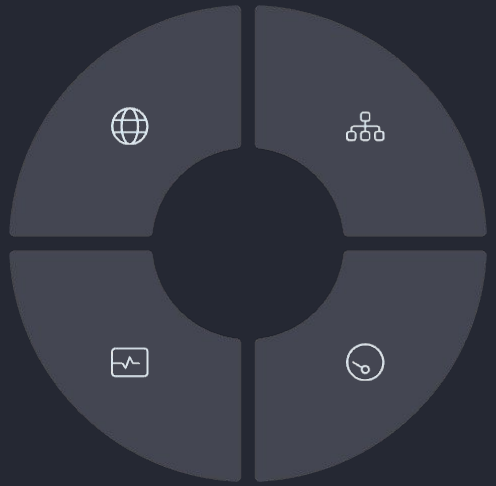
Software

Kubernetes: Inference Gateway

API Layer

OpenAI-compatible endpoints

Request validation and authentication



Router

Model selection and version routing

Load balancing across inference servers

Models & LoRA adapters

Rollout and serve multiple models

QoS

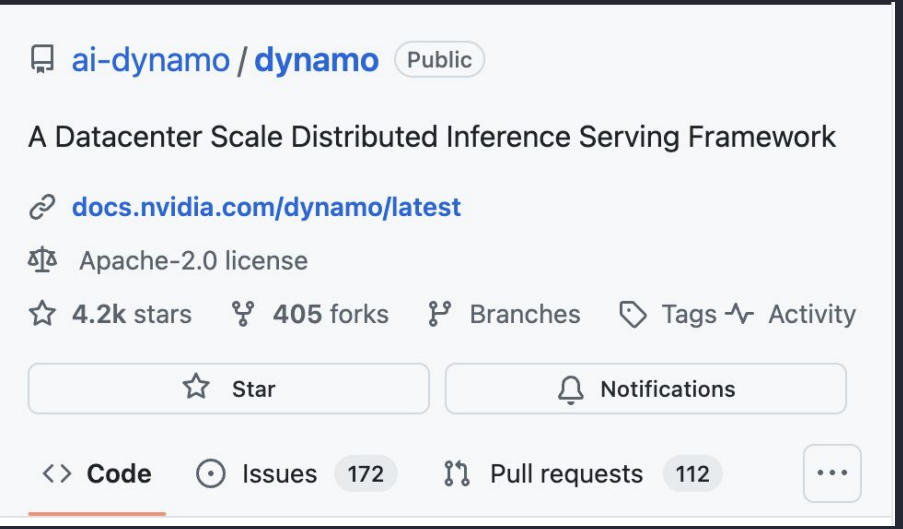
Priority and latency tolerance

gateway-api-inference-extension.sigs.k8s.io

https://gateway-api-inference-extension.sigs.k8s.io/

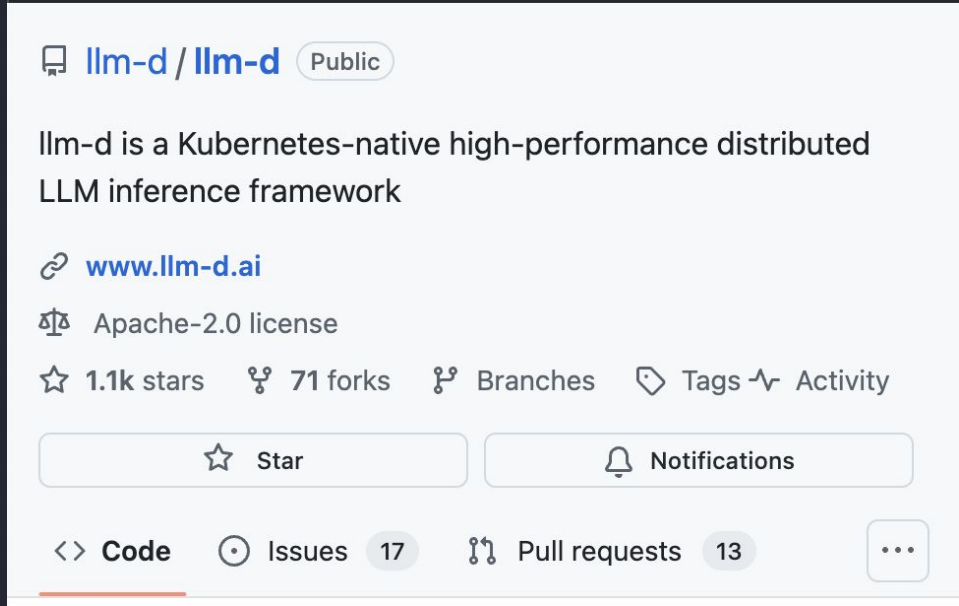
Gateway API Inference Extension is an official Kubernetes project that optimizes self-hosting Generative Models on Kubernetes.

2 frameworks on top of kubernetes



Nvidia Dynamo

llm-d (RedHat)



Missing feature: QoS

Batch

Async job
50% lower costs

```
import OpenAI from "openai";
const openai = new OpenAI();

const batch = await openai.batches.create({
  input_file_id: "file-abc123",
  endpoint: "/v1/chat/completions",
  completion_window: "24h"
});

console.log(batch);
```

```
import OpenAI from "openai";
const openai = new OpenAI();

const fileResponse = await openai.files.content("file-xyz123");
const fileContents = await fileResponse.text();

console.log(fileContents);
```

Missing feature: QoS

Flex processing

Sync job
50% lower costs

```python
from openai import OpenAI
client = OpenAI(
    # increase default timeout to 15 minutes (from 10 minutes)
    timeout=900.0
)

# you can override the max timeout per request as well
response = client.with_options(timeout=900.0).responses.create(
    model="o3",
    instructions="List and describe all the metaphors used in this book.",
    input="<very long text of book here>",
    service_tier="flex",
)

print(response.output_text)
```

Key takeaway

Key takeaway

1- Don’t do this at home,
call your lovely cloud provider.

Key takeaway

1- Don’t do this at home,
call your lovely cloud provider.

2- If you **really** need to self-host,
use dedicated framework.

Key takeaway

1- Don’t do this at home,
call your lovely cloud provider.

2- If you **really** need to self-host,
use dedicated framework.

3- If you insist, hire a good LLMOps

Ressources

https://www.tensoreconomics.com/

https://llm-d.ai/blog/llm-d-announce

https://docs.vllm.ai/en/latest/

https://docs.nvidia.com/dynamo/latest/

GenAI Nantes

“Intègre un assistant vocal dans ta webapp”

Raconte

Godefroy de Compreignac

CEO @Lonestone
Founder @Raconte.ai

11 juin 2025 à 19h

10 Rue Magdeleine, 44200 Nantes

chez Sii

RCA zenika [sfeir] lonestone

Slides dispo sur:
https://github.com/genai-nantes-meetup/meetups/

GenAi France

GenAI Day débarque à Rennes le 18 Septembre

- Pour la première fois, la conférence GenAI Day quitte Paris pour s'installer en région, et c'est Rennes qui a l'honneur de l'accueillir !
- GenAI Day, c'est une journée unique pour rencontrer les experts de l'IA Générative et explorer des usages concrets dans tous les secteurs.
- A vos agendas :

  📅 18 Septembre
  📍 Carrousel de la Courrouze (metro B)
  🌐 https://genai-rennes.lovable.app/

- Conférence "All Inclusive" : Petit déjeuner, pause(s) café, déjeuner et afterwork inclus !!
- Seulement 100 billets Early Bird disponibles, ca va partir vite !

Les Talks :

malt
Doctolib
slite
sncfconnect
Gladia
Groupama
Photoroom

GenAI Nantes

“Workshop: Ensorcelle ton IDE. De zéro à démo”

CURSOR

Agent mode, rules, MCP...

Maxime Courant

Consultant et Développeur en Gen AI

9 juillet 2025 à 18h30

2 Pl. Louis Daubenton, 44100 Nantes

chez EPITECH LEARN DIFFERENT*

RCA zenika [sfeir] lonestone