Own your AI

Tim Fischer

netlight

About Netlight

Netlight is an international digital consulting firm, helping leading companies to succeed in the digital landscape, from advice to implementation.

Our service contains the collective intelligence of 2000 consultants offering a comprehensive range of digital services, from strategy to technology. We support industries that are facing new challenges and opportunities based on new technology, to make better business.

Netlight has been awarded several times for profitable growth and management, as a top employer, and for our engagement in Diversity, Equity, and Inclusion. Located in Stockholm, Oslo, Helsinki, Copenhagen, Munich, Hamburg, Berlin, Frankfurt, Zurich, Cologne, Amsterdam and Toronto. Co-creating the future today, since 1999.

aws PARTNER Advanced Tier Services

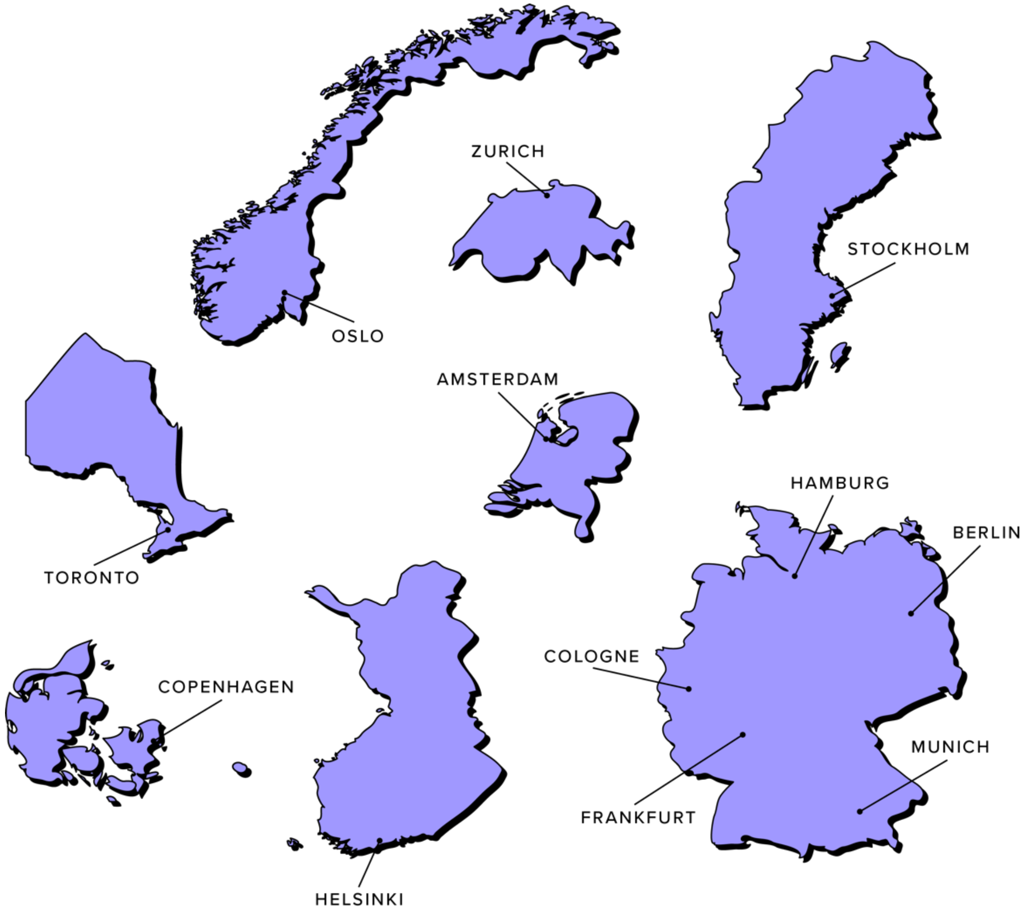
aws PARTNER Data & Analytics Services Competency

aws PARTNER AWS Glue Delivery

aws PARTNER AWS Lambda Delivery

Google Cloud Partner

Microsoft Silver Partner



OFFICES **12** EMPLOYEES **2000+** CLIENTS **350+** FOUNDED **1999** WOMEN AT NETLIGHT **37 %**

netlight

Agenda

Why?

How? (Architectural Setup)

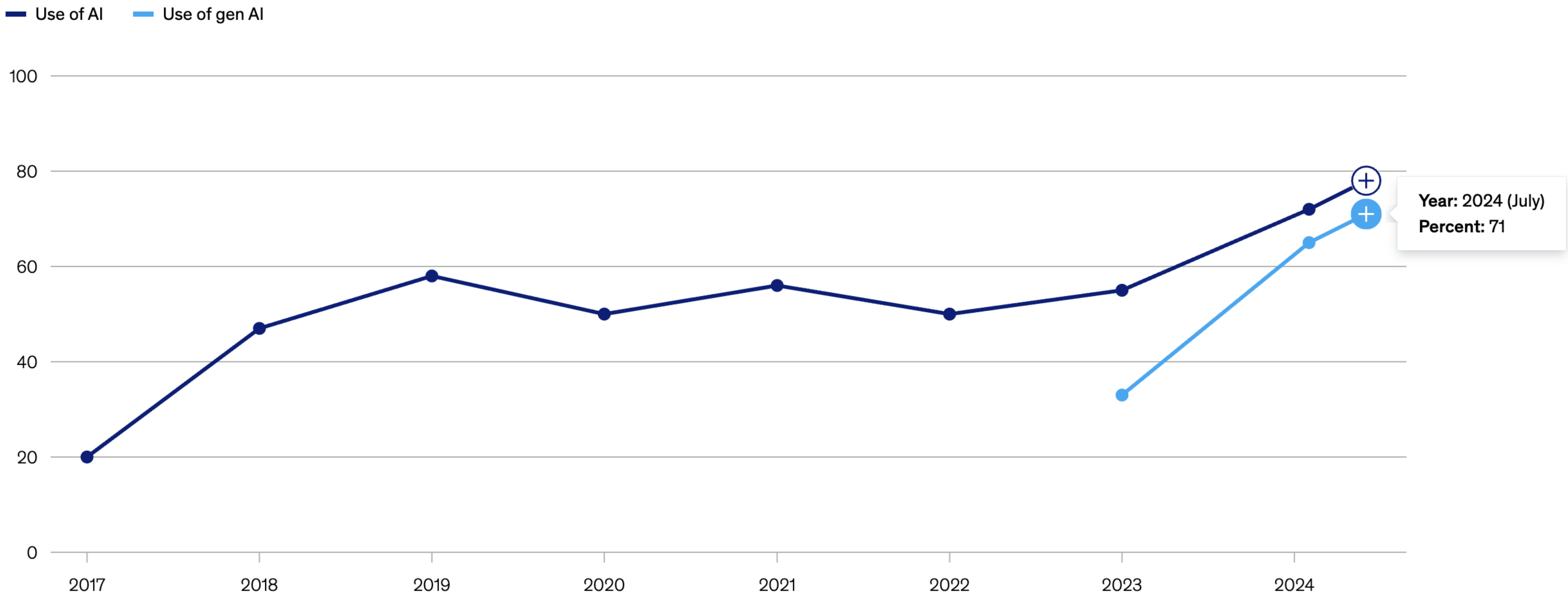
What I wish I would have known before

Key Take-aways

Q&A

Organizations’ use of AI has accelerated markedly in the past year, after years of little meaningful change.

Organizations that use AI in at least 1 business function,[1] % of respondents



*Source: McKinsey Global Surveys on the state of AI

What's the problem?

- **Data Privacy & Control**: Sensitive data may be exposed, used without consent, or fail to meet compliance requirements like GDPR or HIPAA.

- **Vendor Dependency**: Risk of outages, AI model changes, API updates when locked-in with a single provider.

- **Security Risks**: Limited control over encryption, and reliance on the provider's security practices increases breach risks.

- **Customization Limits**: Inability to access underlying mechanics for transparency and optimization.

- **Cost & Scalability:** Pricing changes are unpredictable, leaving you vulnerable to cost increases that can disrupt budgets

European tech industry coalition calls for ‘radical action’ on digital sovereignty — starting with buying local

Natasha Lomas
Techcrunch, March 16, 2025

Big Tech’s future in Europe at risk from Trump, tariffs and data concerns

Eugenia Perozo
Investment Monitor, 31 March, 2025

Legally compliant use of US cloud services: BDI warns against end of agreement

Martin Holland
heise, Mar 31, 2025

Is overreliance on US Big Tech a threat to Europe? The Netherlands may soon find out

Anna Desmarais
Euronews, Feb 27, 2025

Companies in the EU are starting to look for ways to ditch Amazon, Google, and Microsoft cloud services amid fears of rising security risks from the US.

Matt Burgess
WIRED, Mar 24, 2025

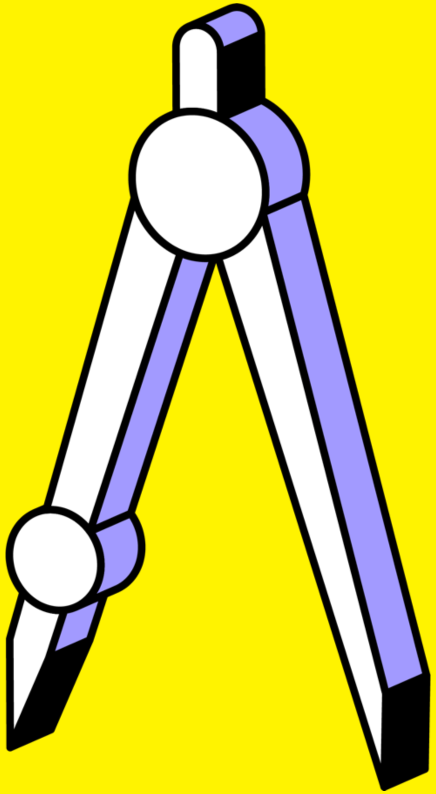
Will the US cloud soon be illegal in the EU?

by Tristan Fincken
CIO, Jan 27, 2025

Tim Fischer

Consultant @ Netlight

Cloud, Infrstructure, Kubernetes, GenAI



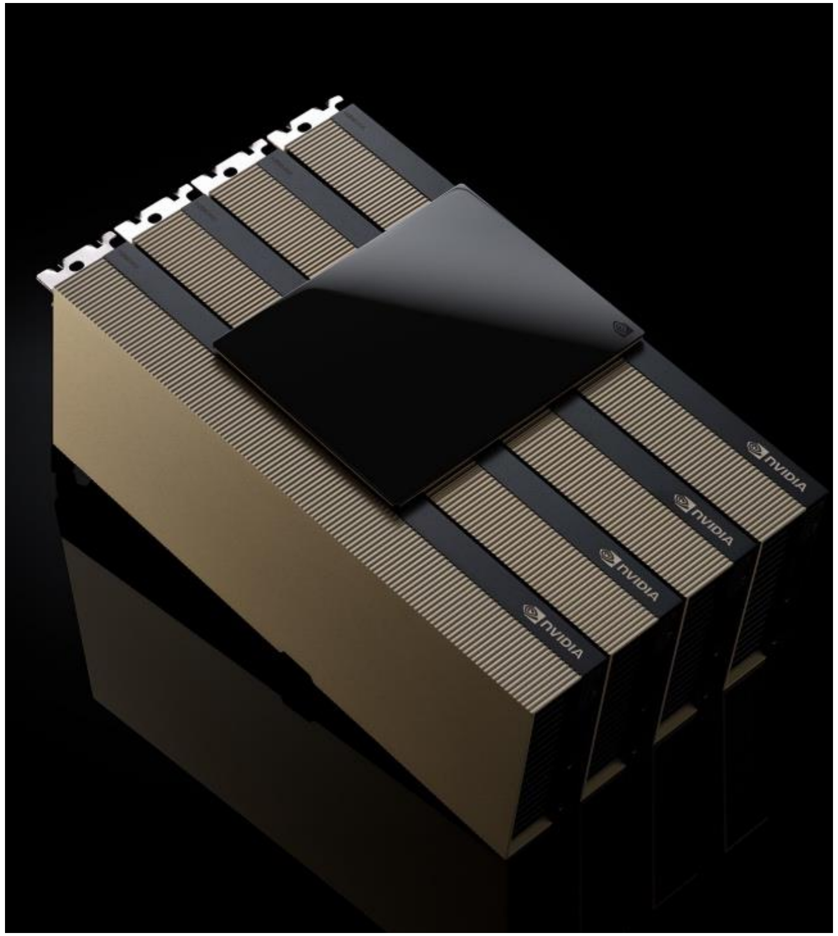
Project Challenge

Create a GenAI Platfom that…

- ensures the highest levels of compliance and security
- minimizes lock-in effects
- provides full control over infrastructure, data, and workflows
- is scalable, resilient, portable and cost-efficient

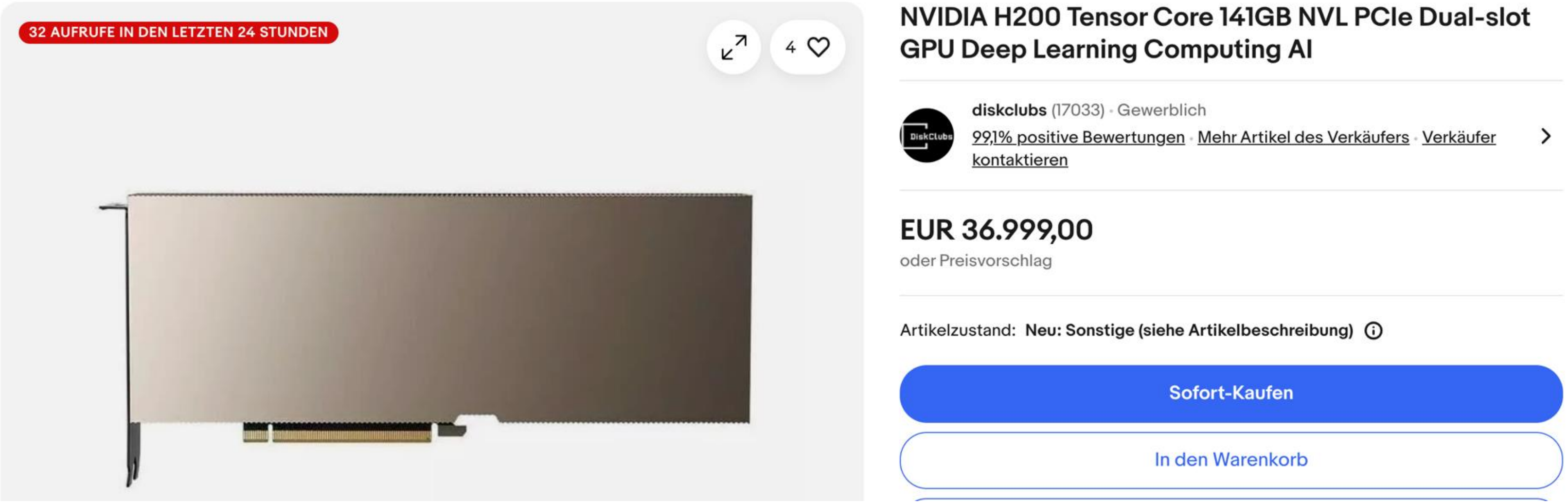
netlight

We need GPUs



Nvidias newest GPU Generation “Blackwell” is reportedly sold out for the next 12 months

Matthew Fox
Business Insider, Oct 11, 2024



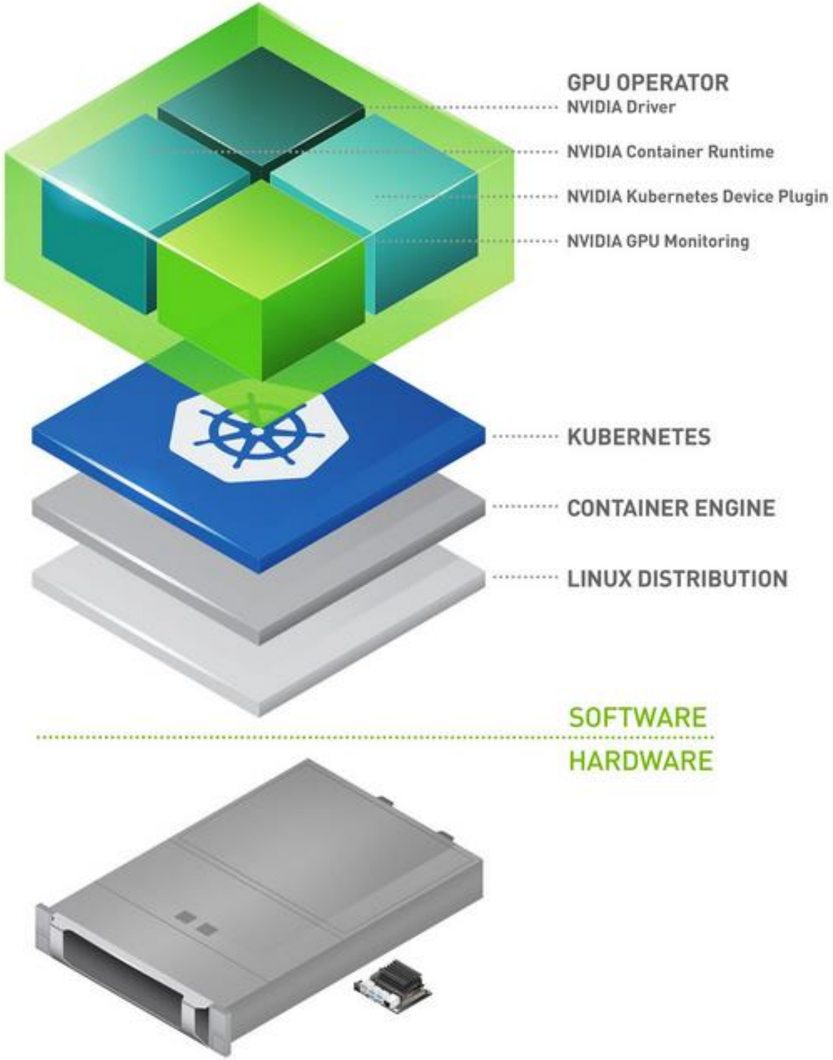
32 AUFRUFE IN DEN LETZTEN 24 STUNDEN

NVIDIA H200 Tensor Core 141GB NVL PCIe Dual-slot GPU Deep Learning Computing AI

diskclubs (17033) · Gewerblich
99,1% positive Bewertungen · Mehr Artikel des Verkäufers · Verkäufer kontaktieren

EUR 36.999,00
oder Preisvorschlag

Artikelzustand: Neu: Sonstige (siehe Artikelbeschreibung)



kubernetes

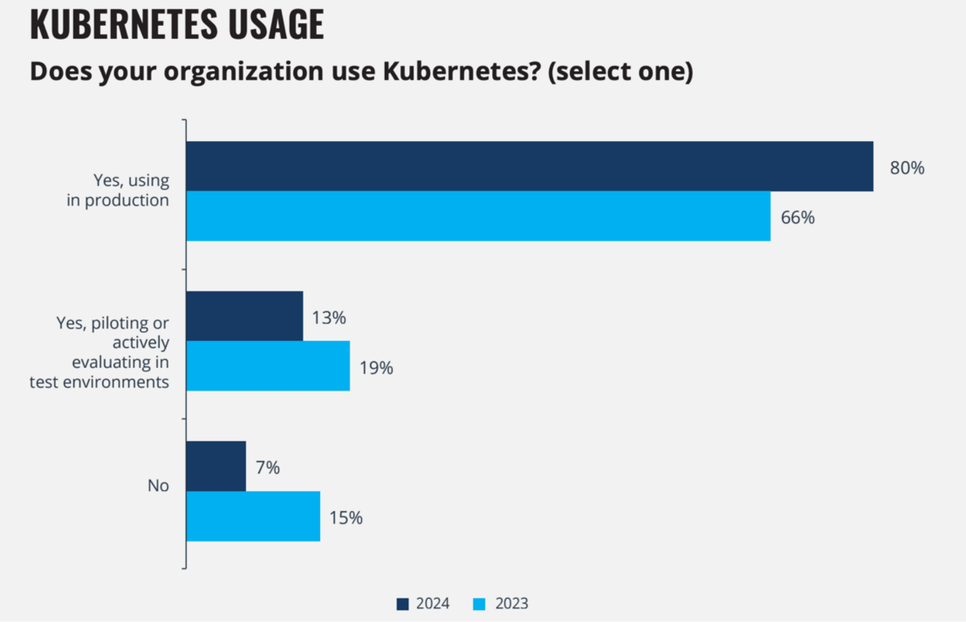
- **Cloud-Agnostic and GPU-Ready using GPU Operator**: Kubernetes provides portability across on-premises and cloud environments, enabling efficient use of GPU resources from any infrastructure provider.
- **Flexible and Open Platform**: As an open, free, and extensible SDK, Kubernetes empowers developers to build and deploy at scale without vendor lock-in.
- **Scalable and Reliable**: Kubernetes automates scaling, load balancing, and fault tolerance, making it ideal for serving AI/ML models with high availability and performance.
- **Ecosystem and Innovation**: A huge effort by the community (e.g., SIG Serving) positions Kubernetes as the AI serving layer, ensuring rapid innovation.

Setup

80% of companies use Kubernetes in production

**CNCF annual survey 2024*

Many companies already have platform teams for managing and provisioning the clusters



KUBERNETES USAGE

Does your organization use Kubernetes? (select one)

- **Cilium**: Provides deep network security with eBPF-based policies, identity-aware security, encryption, and transparent visibility into L3-L7 traffic flows.
- **Kyverno:** Strong policy enforcement tool with Kubernetes-native approach, covers validating, mutating, and generating resources based on security policies.
- **ArgoCD:** GitOps principles that provide complete audit trails, version control, and approval workflows while automatically detecting and remediating configuration drift. Enables advanced deployment strategies for smooth rollouts

Model selection

Selection of state-of-the-art Open-Weight-models

deepseek-ai/DeepSeek-R1

google/gemma-3

Qwen/Qwen2.5

meta-llama/Llama-4

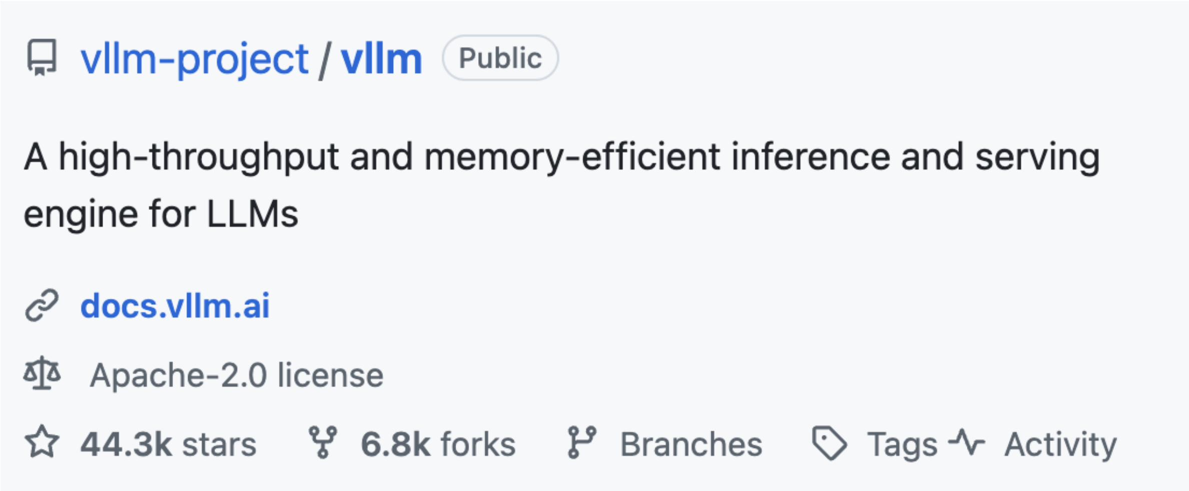
mistralai/Mistral-Small-3.1

- **Open-Weight model**: Trained model weights are publicly available for use, modification, and fine-tuning
- **Closed-Weight model**: Model weights are proprietary and inaccessible, usable only via APIs (e.g. OpenAI, Anthropic Claude)

→ closed models are still slightly better, but open models are catching up fast 🚀

How to run it?

vLLM



vllm-project / vllm Public

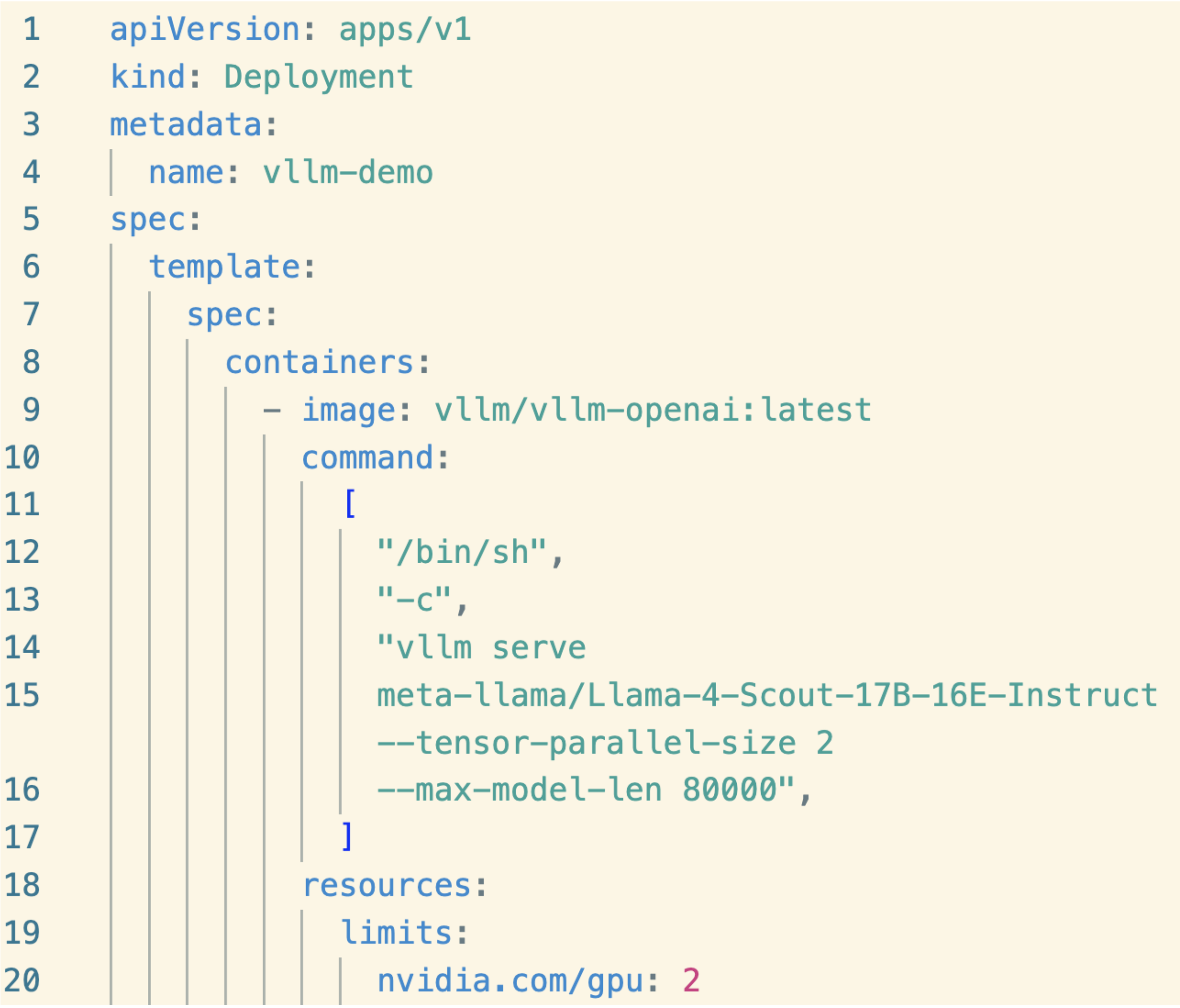
A high-throughput and memory-efficient inference and serving engine for LLMs

docs.vllm.ai

Apache-2.0 license

44.3k stars 6.8k forks Branches Tags Activity

- vLLM is a **fast and easy-to-use library for LLM inference** and serving.
- Offers OpenAI-compatible API server
- Supports NVIDIA GPUs, AMD CPUs and GPUs, Intel CPUs and GPUs, PowerPC CPUs, TPU, and AWS Neuron.
- Downloads models directly from huggingface

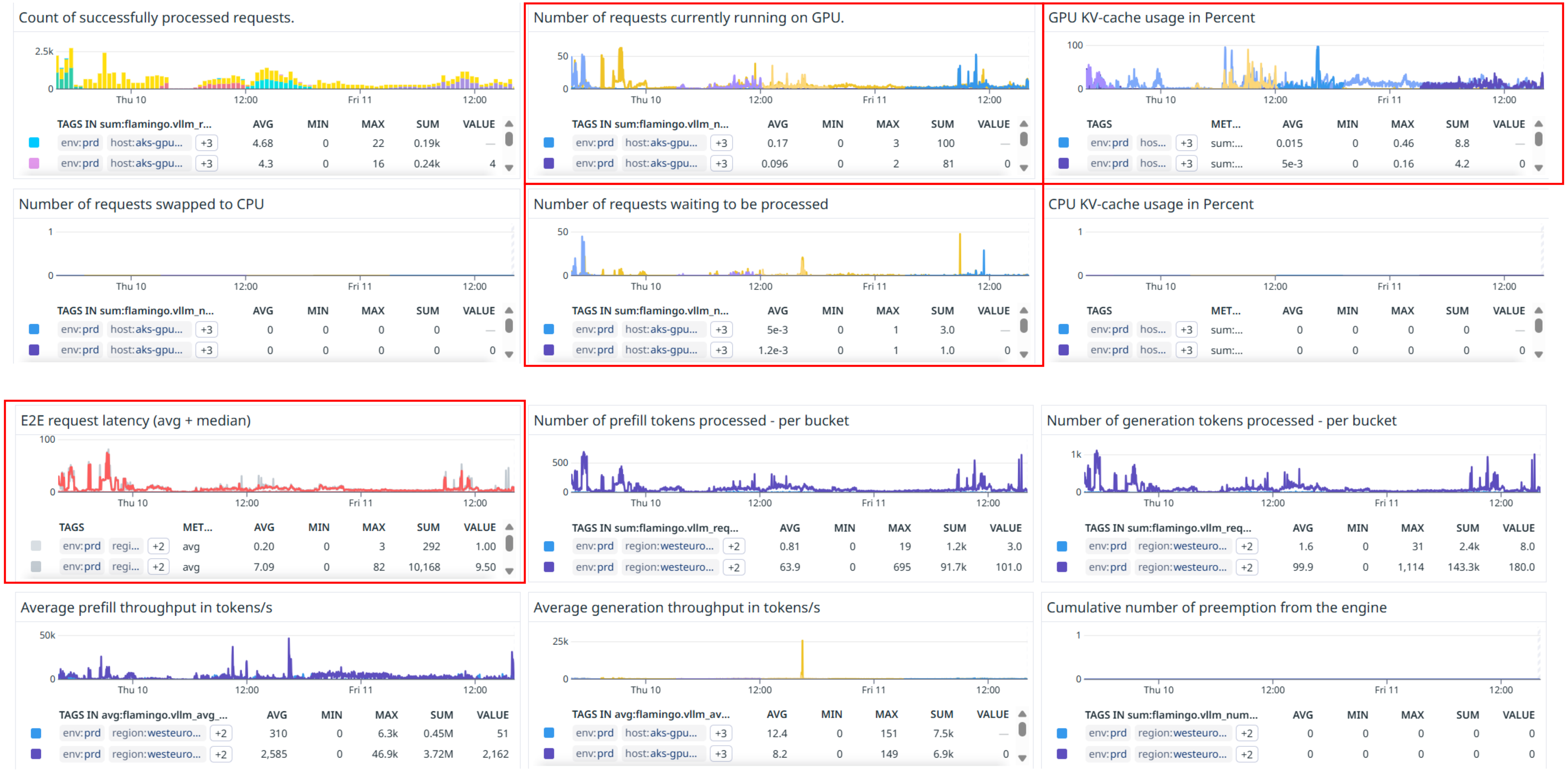


```yaml
apiVersion: apps/v1
kind: Deployment
metadata:
  name: vllm-demo
spec:
  template:
    spec:
      containers:
        - image: vllm/vllm-openai:latest
          command:
            [
              "/bin/sh",
              "-c",
              "vllm serve
              meta-llama/Llama-4-Scout-17B-16E-Instruct
              --tensor-parallel-size 2
              --max-model-len 80000",
            ]
          resources:
            limits:
              nvidia.com/gpu: 2
```

netlight

Observability

vLLM and Nvidia DCGM expose Prometheus metrics endpoint



Count of successfully processed requests.

	TAGS IN sum:flamingo.vllm_r...			AVG	MIN	MAX	SUM	VALUE
🟦	env:prd	host:aks-gpu...	+3	4.68	0	22	0.19k	–
■	env:prd	host:aks-gpu...	+3	4.3	0	16	0.24k	4

Number of requests currently running on GPU.

	TAGS IN sum:flamingo.vllm_n...			AVG	MIN	MAX	SUM	VALUE
🟦	env:prd	host:aks-gpu...	+3	0.17	0	3	100	–
🟪	env:prd	host:aks-gpu...	+3	0.096	0	2	81	0

GPU KV-cache usage in Percent

	TAGS			MET...	AVG	MIN	MAX	SUM	VALUE
🟦	env:prd	hos...	+3	sum:...	0.015	0	0.46	8.8	–
🟪	env:prd	hos...	+3	sum:...	5e-3	0	0.16	4.2	0

Number of requests swapped to CPU

	TAGS IN sum:flamingo.vllm_n...			AVG	MIN	MAX	SUM	VALUE
🟦	env:prd	host:aks-gpu...	+3	0	0	0	0	–
🟪	env:prd	host:aks-gpu...	+3	0	0	0	0	0

Number of requests waiting to be processed

	TAGS IN sum:flamingo.vllm_n...			AVG	MIN	MAX	SUM	VALUE
🟦	env:prd	host:aks-gpu...	+3	5e-3	0	1	3.0	–
🟪	env:prd	host:aks-gpu...	+3	1.2e-3	0	1	1.0	0

CPU KV-cache usage in Percent

	TAGS			MET...	AVG	MIN	MAX	SUM	VALUE
🟦	env:prd	hos...	+3	sum:...	0	0	0	0	–
🟪	env:prd	hos...	+3	sum:...	0	0	0	0	0

E2E request latency (avg + median)

	TAGS			MET...	AVG	MIN	MAX	SUM	VALUE
■	env:prd	regi...	+2	avg	0.20	0	3	292	1.00
■	env:prd	regi...	+2	avg	7.09	0	82	10,168	9.50

Number of prefill tokens processed - per bucket

	TAGS IN sum:flamingo.vllm_req...			AVG	MIN	MAX	SUM	VALUE
🟦	env:prd	region:westeuro...	+2	0.81	0	19	1.2k	3.0
🟪	env:prd	region:westeuro...	+2	63.9	0	695	91.7k	101.0

Number of generation tokens processed - per bucket

	TAGS IN sum:flamingo.vllm_req...			AVG	MIN	MAX	SUM	VALUE
🟦	env:prd	region:westeuro...	+2	1.6	0	31	2.4k	8.0
🟪	env:prd	region:westeuro...	+2	99.9	0	1,114	143.3k	180.0

Average prefill throughput in tokens/s

	TAGS IN avg:flamingo.vllm_avg_...			AVG	MIN	MAX	SUM	VALUE
🟦	env:prd	region:westeuro...	+2	310	0	6.3k	0.45M	51
🟪	env:prd	region:westeuro...	+2	2,585	0	46.9k	3.72M	2,162

Average generation throughput in tokens/s

	TAGS IN avg:flamingo.vllm_av...			AVG	MIN	MAX	SUM	VALUE
🟦	env:prd	host:aks-gpu...	+3	12.4	0	151	7.5k	–
🟪	env:prd	host:aks-gpu...	+3	8.2	0	149	6.9k	0

Cumulative number of preemption from the engine

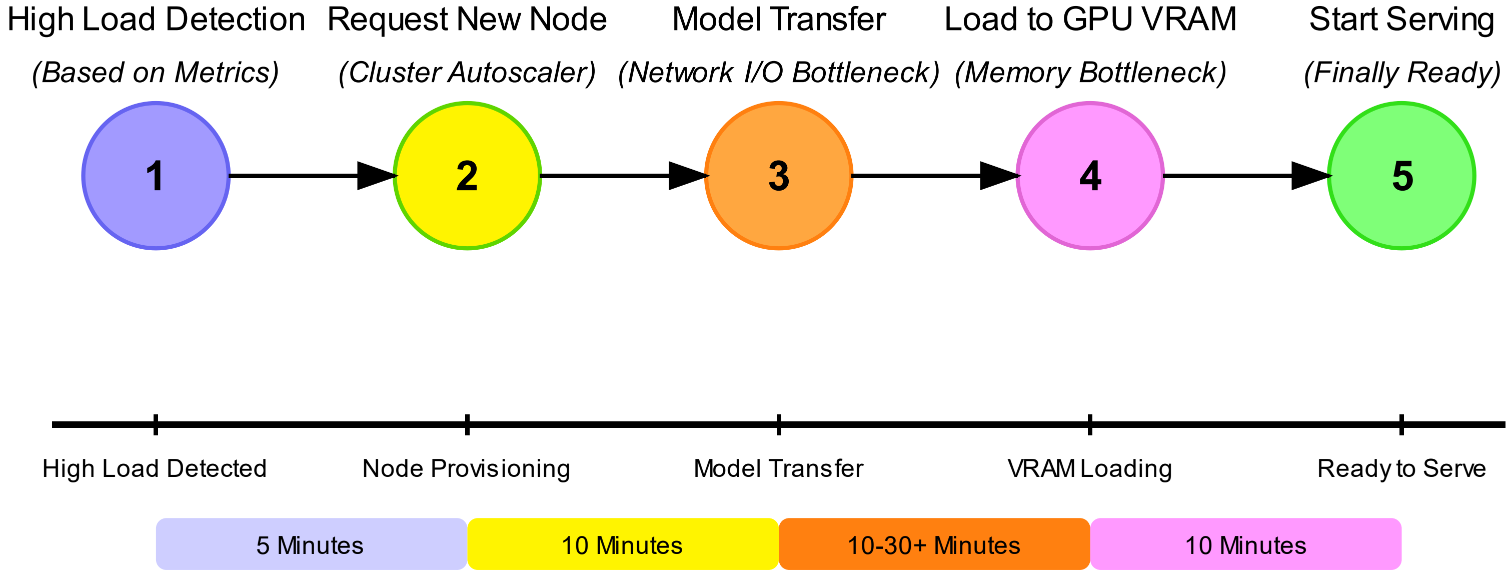
	TAGS IN sum:flamingo.vllm_num...			AVG	MIN	MAX	SUM	VALUE
🟦	env:prd	region:westeuro...	+2	0	0	0	0	0
🟪	env:prd	region:westeuro...	+2	0	0	0	0	0

netlight

The problem with autoscaling…

Problem: Autoscaling is triggered by immediate demand…

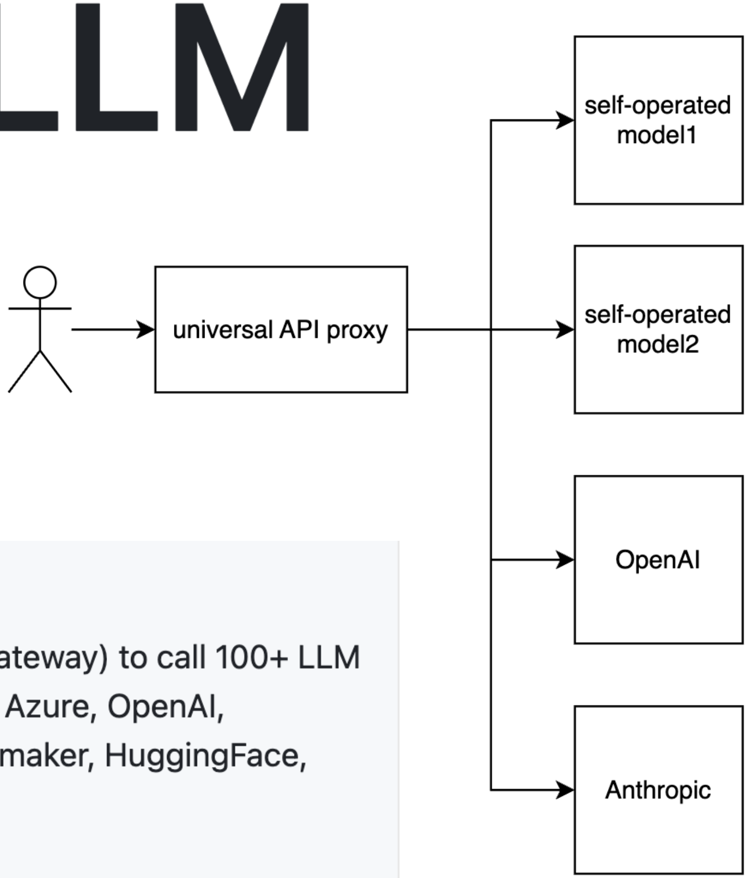
…but new capacity takes 20-60+ minutes to become available

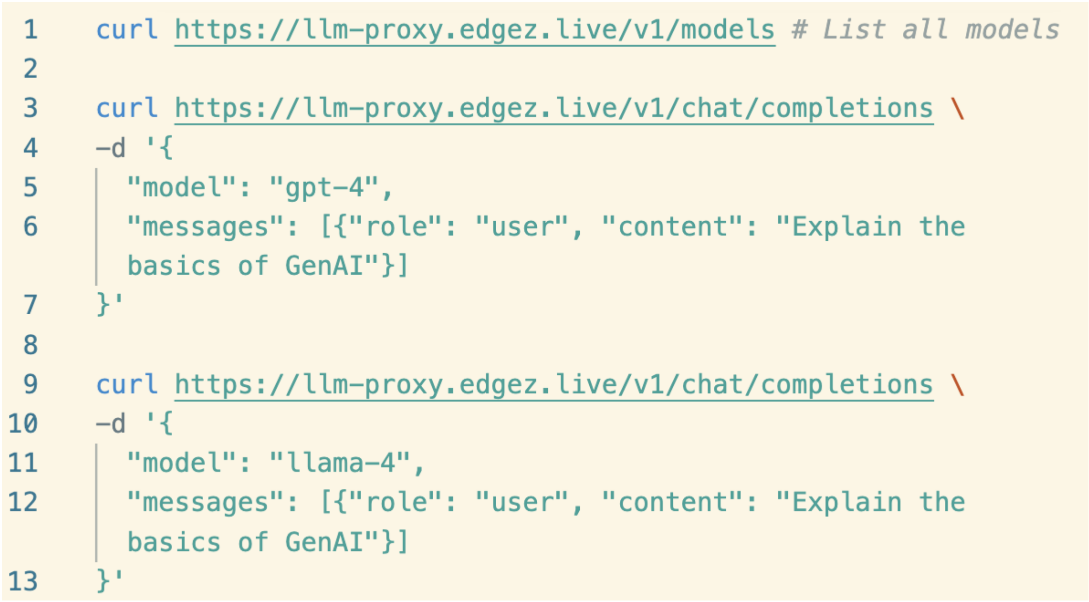


Predictive Scaling Required Instead of Reactive Scaling

API Management

LiteLLM





```
curl https://llm-proxy.edgez.live/v1/models # List all models

curl https://llm-proxy.edgez.live/v1/chat/completions \
-d '{
  "model": "gpt-4",
  "messages": [{"role": "user", "content": "Explain the basics of GenAI"}]
}'

curl https://llm-proxy.edgez.live/v1/chat/completions \
-d '{
  "model": "llama-4",
  "messages": [{"role": "user", "content": "Explain the basics of GenAI"}]
}'
```

BerriAI / litellm Public

Python SDK, Proxy Server (LLM Gateway) to call 100+ LLM APIs in OpenAI format - [Bedrock, Azure, OpenAI, VertexAI, Cohere, Anthropic, Sagemaker, HuggingFace, Replicate, Groq]

docs.litellm.ai/docs/

View license

20.7k stars 2.6k forks Branches Tags

Activity

- Custom routing logic
- Retry/fallback logic across multiple deployments
- Set Budgets & Rate limits per project, api key, model

Learnings 💡

What I wish I knew before building our first GenAI deployment.

- Even major hyperscalers occasionally face **GPU shortages** that can last for hours.
- The cost of requests to an LLM can vary by a factor > 100,000. Implement robust rate limiting and user budgeting mechanisms early on; **rate limiting solely by request counts is not helpful**.
- Leverage **predictive auto-scaling mechanisms by using forecasting** and closely monitor latency to maintain performance (e.g., KEDA ScaledObject).
- To tune and optimize model deployments effectively, it helps to **understand how model serving works**—but sometimes, the underlying mechanics are so complex that you just need to take them as they are.

Pro 👍

- **Compliance**: Operating in highly regulated industries with very sensitive data
- **Flexibility**: Need full control and customizability (quantization techniques; model finetuning; niche-models)
- **Trust issues using open APIs**
- **High Usage expected**: Cost-effective at scale
- If your company aims to build in-house expertise in generative AI that goes beyond simply using available APIs, **it can lead to exciting collaborations between data scientists and infrastructure engineers**. 🚀

Cons 👎

- **Low Usage**: API solutions are cheaper for low demand.
- **Risk of Over-Engineering**: If your use case is very straightforward or small-scale, self-operating models can be unnecessarily complex and resource-intensive.
- **Lack of knowledge building a platform within your company**
- **Open-weight models are not matching your requirements** and you need close—weight models (e.g. Anthropic Claude, OpenAI)

Q&A

Thank you!

Tim Fischer
Consultant @ Netlight
tim.fischer@netlight.com

netlight