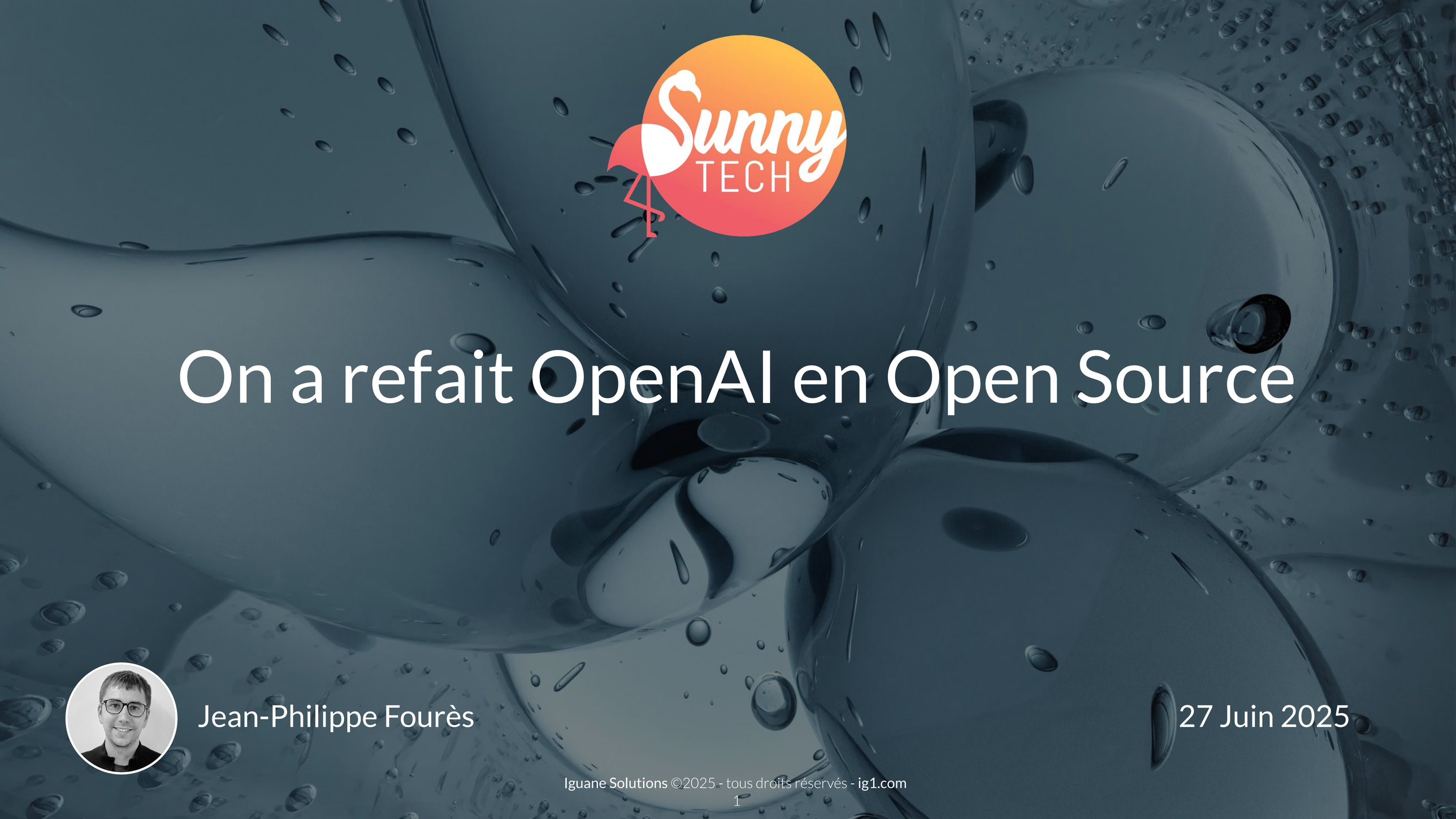


On a refait OpenAI en Open Source

Jean-Philippe Fourès

27 Juin 2025

Disclaimer

- Pas Data Scientist
- Pas Machine Learning Engineer
- Pas éditeur de modèle
- Retour d’expérience de l’IA sur nos usages infra

Parlons IA Générative

IA Génératives

Quelles options possibles ?

IA Cloud

OpenAI

Gemini

ANTHROP\C

deepseek

MISTRAL AI_

Grok

IA Managées

Amazon Bedrock

Scaleway Managed Inference

Azure AI Foundry

IA Open Sources

deepseek

MISTRAL AI_

Qwen

LLaMA by Meta

Black Forest Labs

IA Génératives

Comparons les options

	IA Cloud	IA Managées	IA Open Weights
PROS	- Simple à utiliser - Paiement à l'usage - Leadership - Abstraction de l'infra		
CONS	- (non) contrôle de la donnée - Modèles limités - Biais, conditions Service - Cher à l'échelle		

IA Génératives

Comparons les options

	IA Cloud	IA Managées	IA Open Weights
PROS	- Simple à utiliser - Paiement à l'usage - Leadership - Abstraction de l'infra	- Simple à utiliser - Paiement à l'usage - Infrastructure gérée - Modèles gérés	
CONS	- (non) contrôle de la donnée - Modèles limités - Biais, conditions Service - Cher à l'échelle	- Modèles limités - Compétences SRE / Tech - Peu / Pas d'unification des modèles	

IA Génératives

Comparons les options

	IA Cloud	IA Managées	IA Open Weights
PROS	- Easy to use - Pay as you go - Leadership - Invisible Infra	- Simple à utiliser - Paiement à l'usage - Infrastructure gérée - Modèles gérés	- Catalogue illimité - Accès Modèles non censuré - Liberté de choix - Modèles à poids ouverts
CONS	- Loss of data control - Limited models - Biases, Editor policies - Can be very expensive	- Modèles limités - Compétences SRE / Infra - Peu / Pas d'unification des modèles	- Déploiement / gestion complexe - Maturité des projets - Compétences SRE / Infra - Accès GPU pas simple

Le Dilemme

Notre Choix ?

Notre “OpenAI” en Open Source

Image: https://commons.wikimedia.org/wiki/File:We_Can_Do_It!_NARA_535413_-_Restoration_2.jpg

IA génératives

Pourquoi ce choix ?

Compétences Tech : ✅

Accès à l'infra: ✅

Expérience des projets Open Sources jeunes: ✅

Introduction

Pourquoi faire notre propre stack IA ?

Autonomie

Liberté dans le choix des modèles

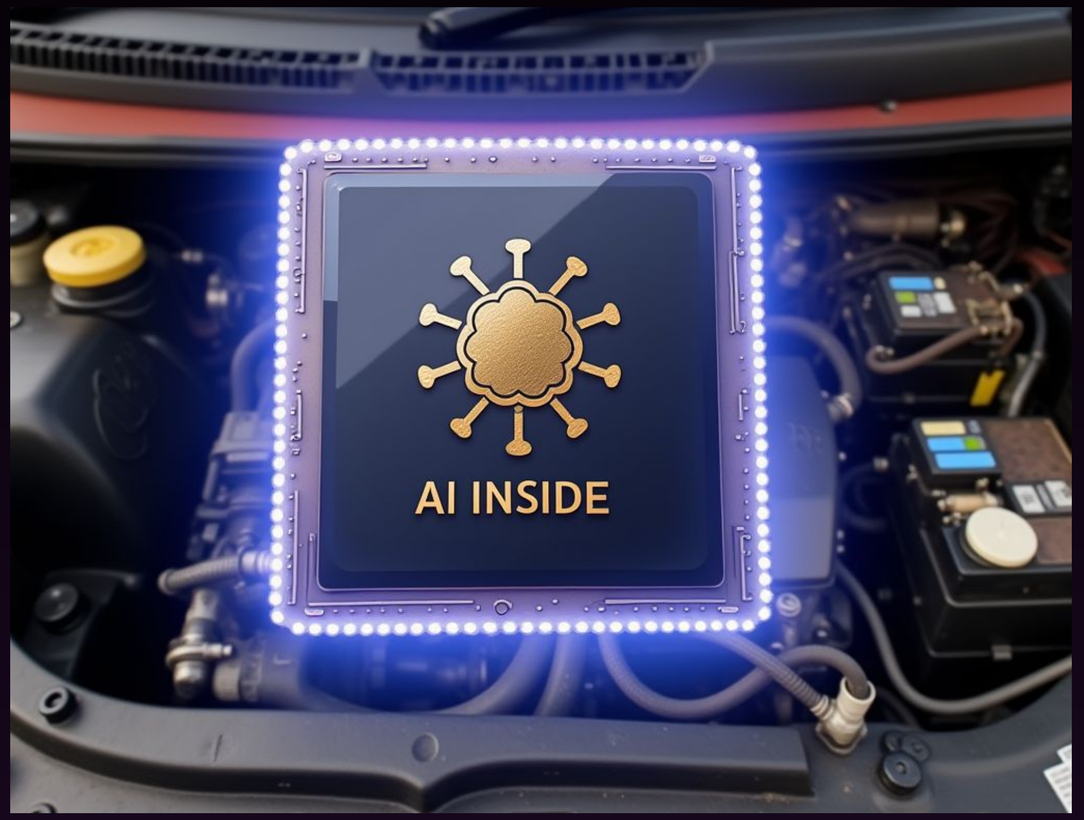
Gouvernance

Contrôle des données
Maîtrise du périmètre

Expertise

Maîtrise du savoir-faire
Augmentation compétences

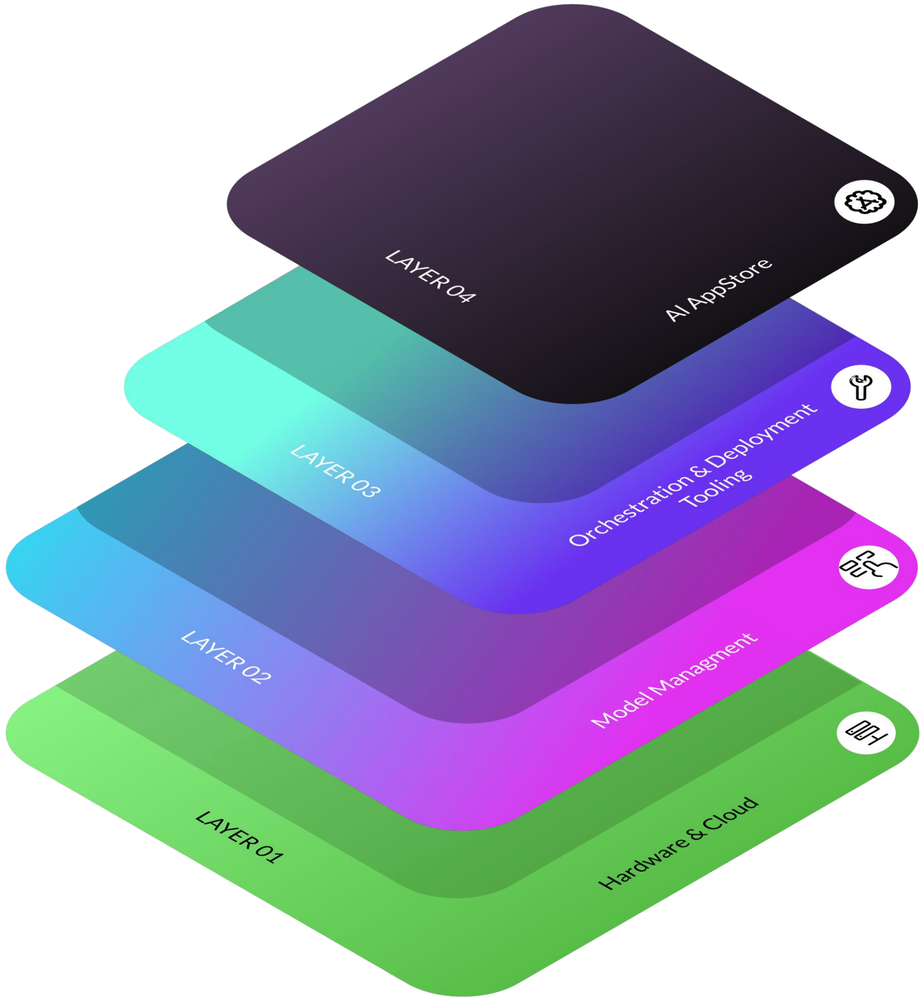
Le dessous des IA génératives



Point commun

Partagent le même concept de plateforme

Point commun



Mise en pratique

Construisons la stack IA ensemble

Besoin : Assistant IA pour accélérer vos tâches

On veut un assistant capable de:

- répondre à tout type de demande internes
- intégré dans notre messagerie
- répond dans la langue de la question
- s'appuyer sur nos docs et les docs externes

Mise en pratique

Construisons la stack IA ensemble

- Quelle infra minimale ?
- Quel Modèle LLM ? open source de préférence
- Quelle performance possible pour usage API ?

INFRA

Infrastructure

Il faut du GPU !

GPU vs CPU: Parallèle vs Séquentiel

VRAM : Plus rapide que RAM

Performances maximales (en théorie)

Infrastructure

Importance du GPU pour l'inférence

Performance Maximale: Inférence sur GPU seul

Important: Modèle + Contexte en VRAM

VRAM GPU: paramètre dimensionnant GPU

Infrastructure

Quel GPU pour mon IA ?

Prenons un LLM 72B - bon compromis :

QWEN 2.5 72B Instruct BF16	144 GB
QWEN 2.5 72B Instruct FP8	**72** GB

RAM pour contexte et KV cache: 10GB / 32k tokens

Infrastructure

Quelle Infra pour mon IA ?

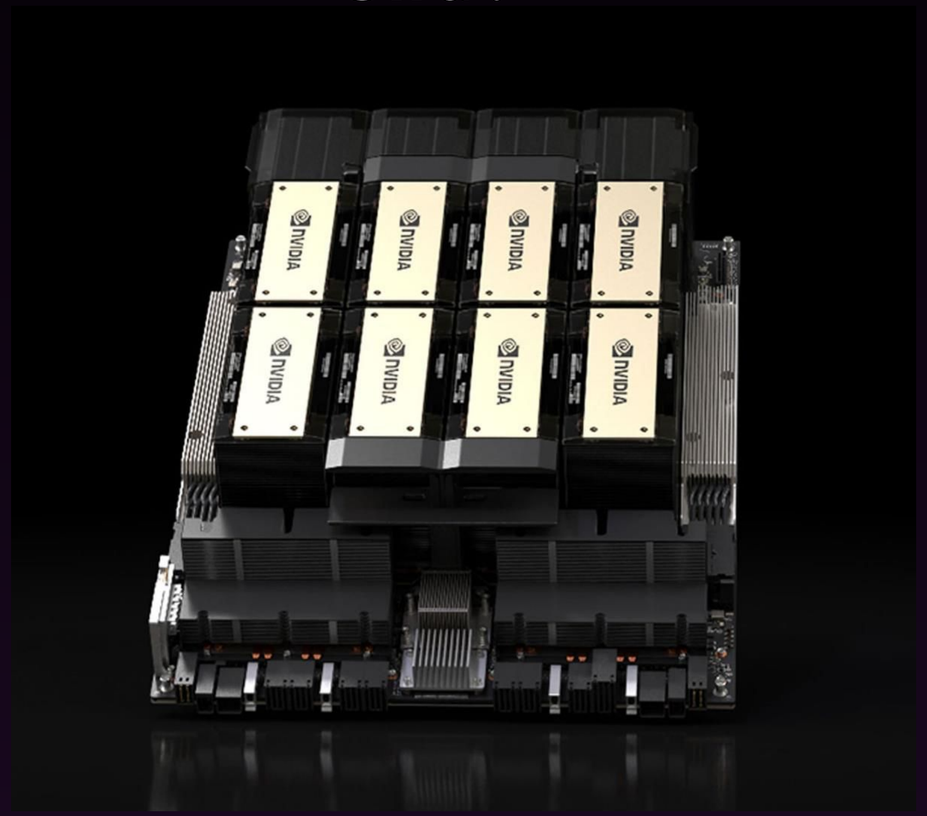
GPUs très performants pour production :

GPU Nvidia H100	80 GB VRAM
GPU Nvidia H200	141 GB VRAM

Notre choix : 1 H200 (moins cher que 2 H100)

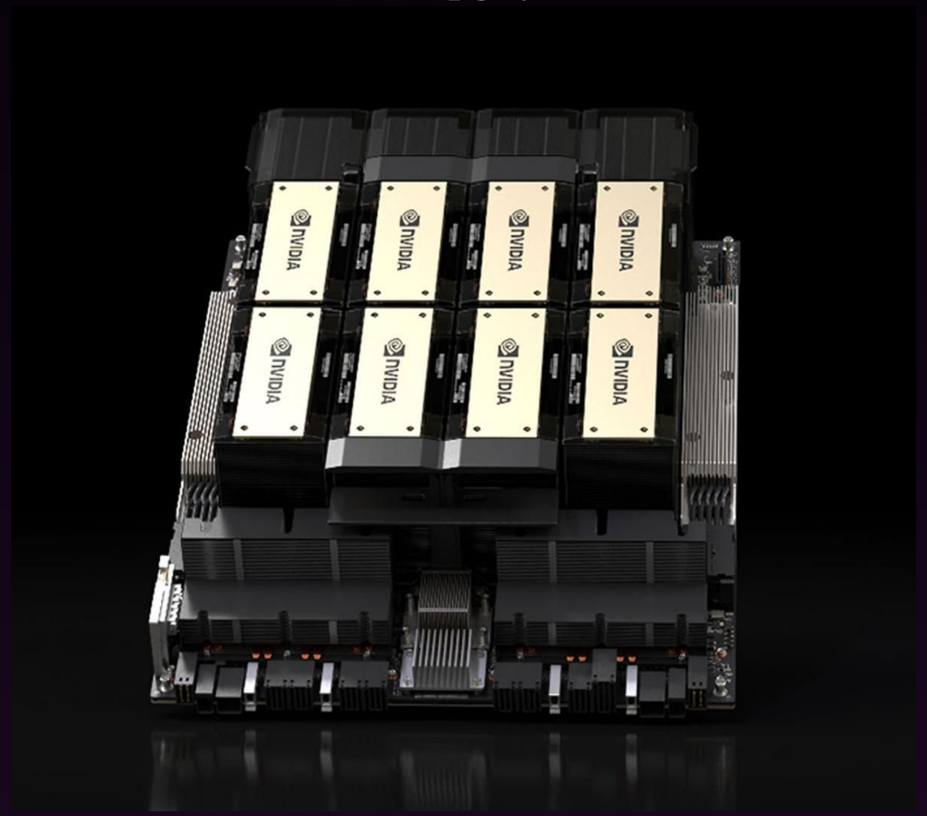
Infrastructure

On a !



Infrastructure

MAIS !



Infrastructure

HGX H200 SXM

HGX: Serveur avec 8 H200 SXM5

NVLink entre GPUs

Comment isoler les GPU pour usage dédié ?

Infrastructure

HGX H200 SXM

La Virtualisation

- OpenNebula 6.10
- Serveur GPU: hyperviseur (KVM)
- VMs avec 1, 2 ou 4 GPUs selon besoin

Infrastructure

HGX H200 SXM

Notre objectif :

- Passthrough 1 (ou plus) GPU(s) à 1 VM
- Autoriser NVLink entre GPUs sur une même VM
- Interdire NVLink entre GPUs sur VMs différentes

Infrastructure

HGX H200 SXM

Partitionnement GPU NVLink

- Shared NVSwitch Virtualization Model
- VM Fabric Manager
- Passthrough GPU(s) sur VMs

Infrastructure

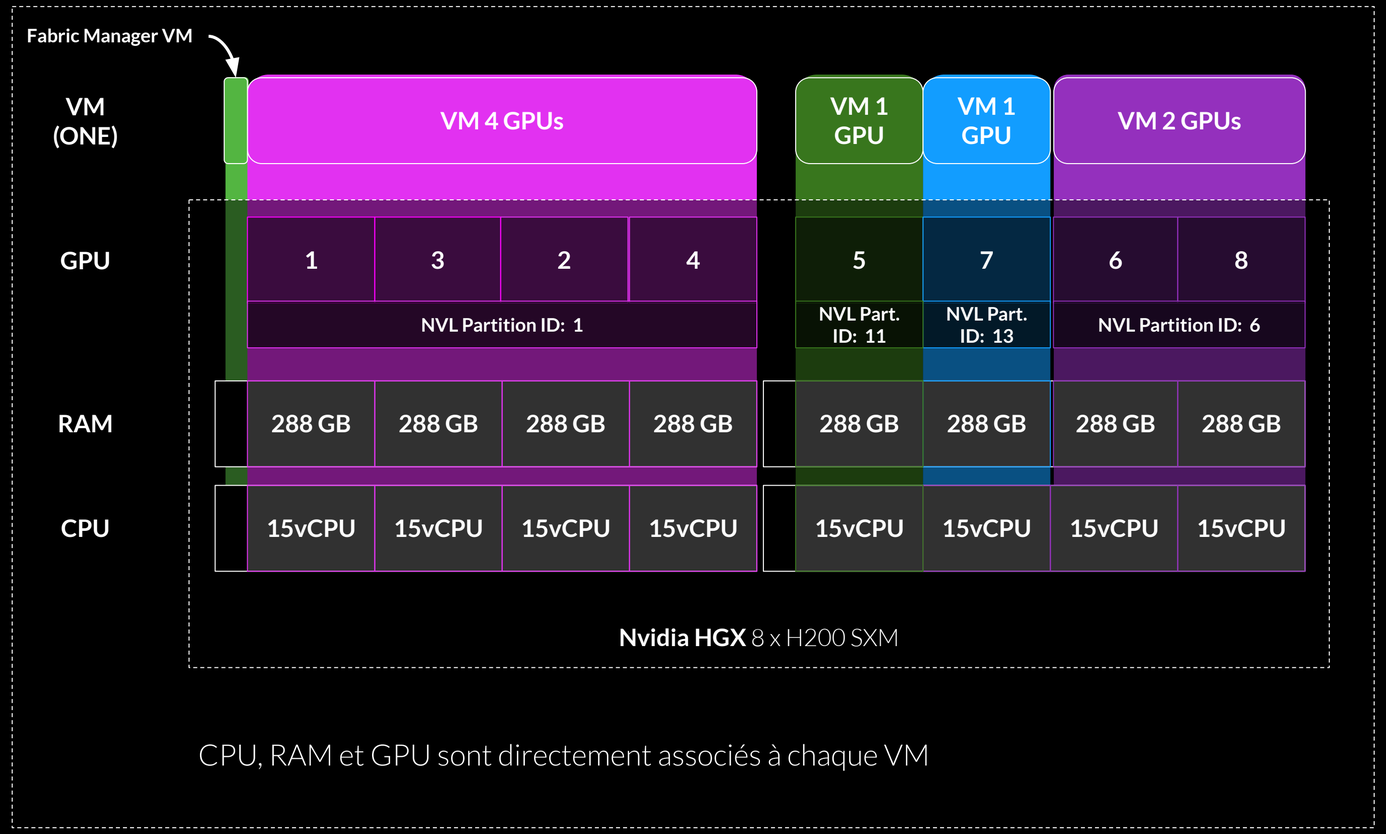
HGX H200 SXM

D’un serveur GPU à un hyperviseur avec GPUs :

- IOMMU & HW virtualisation support
- Chargement VFIO driver au boot kernel
- Eviter le binding GPU (blacklist drivers)
- Autoriser QEMU à accéder aux devices VFIO

Infrastructure

HGX H200 SXM - Exemple de configuration



CPU, RAM et GPU sont directement associés à chaque VM

Infrastructure

Une infra modulaire

Pourquoi plusieurs GPUs ?

- Le modèle est trop grand
- Contexte important
- Besoin de puissance de calcul (usage très intensif)

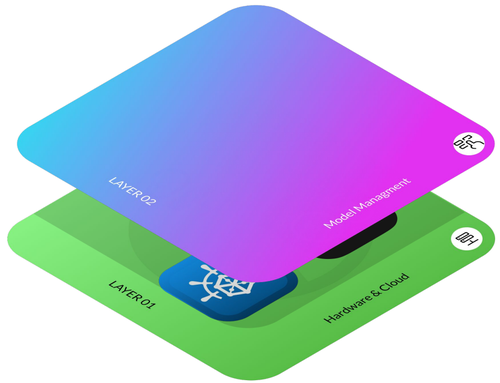
Infrastructure

On a l’infra pour notre IA

Modèles

Modèles Open Source

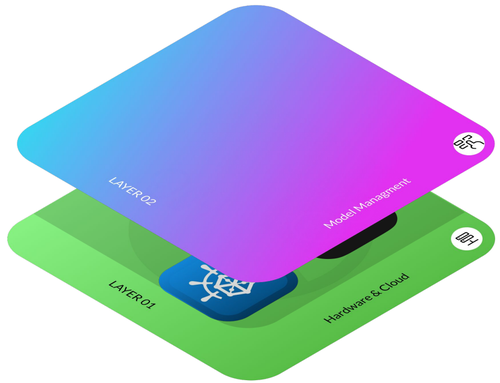
Un catalogue “illimité”

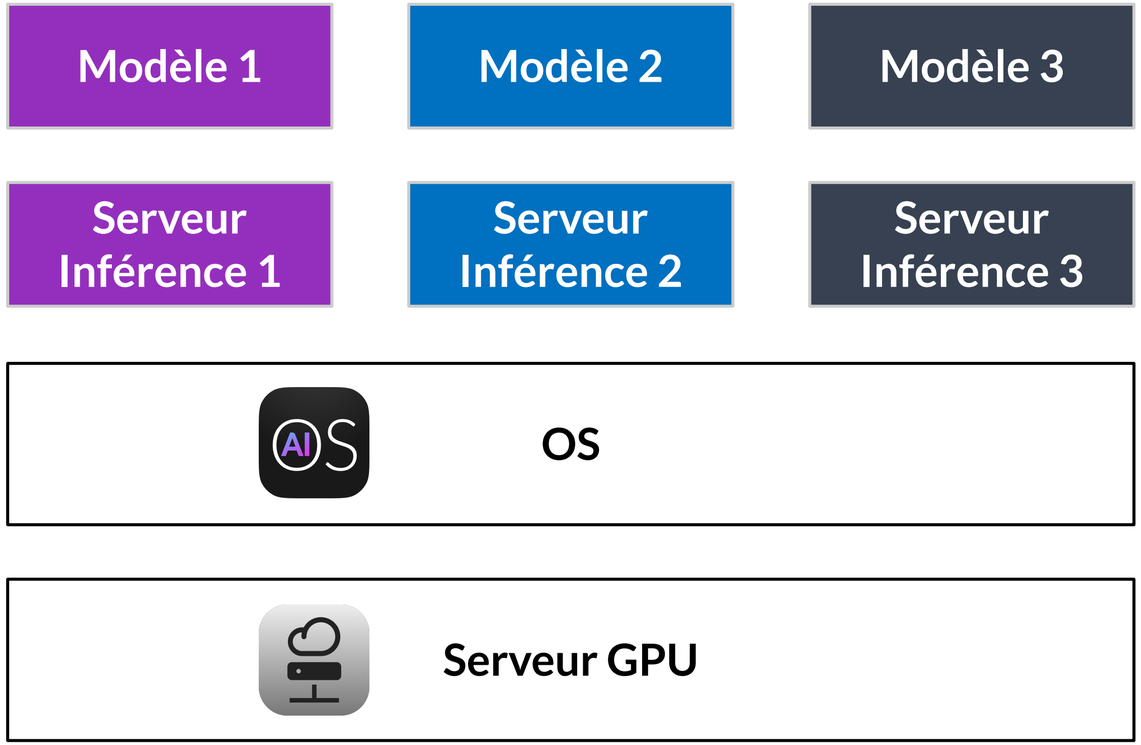


- Hugging Face: La référence
- Large choix de modèles
- Modèles adaptés à son marché

Modèles Open Source

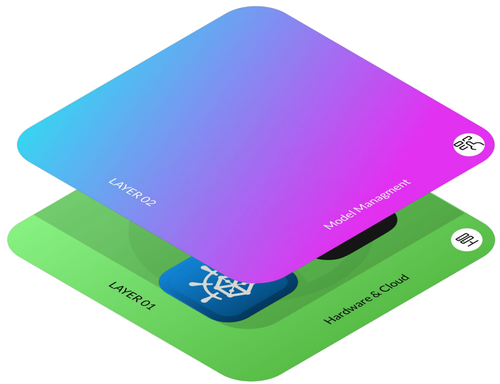
Comment déployer un modèle





Modèles Open Source

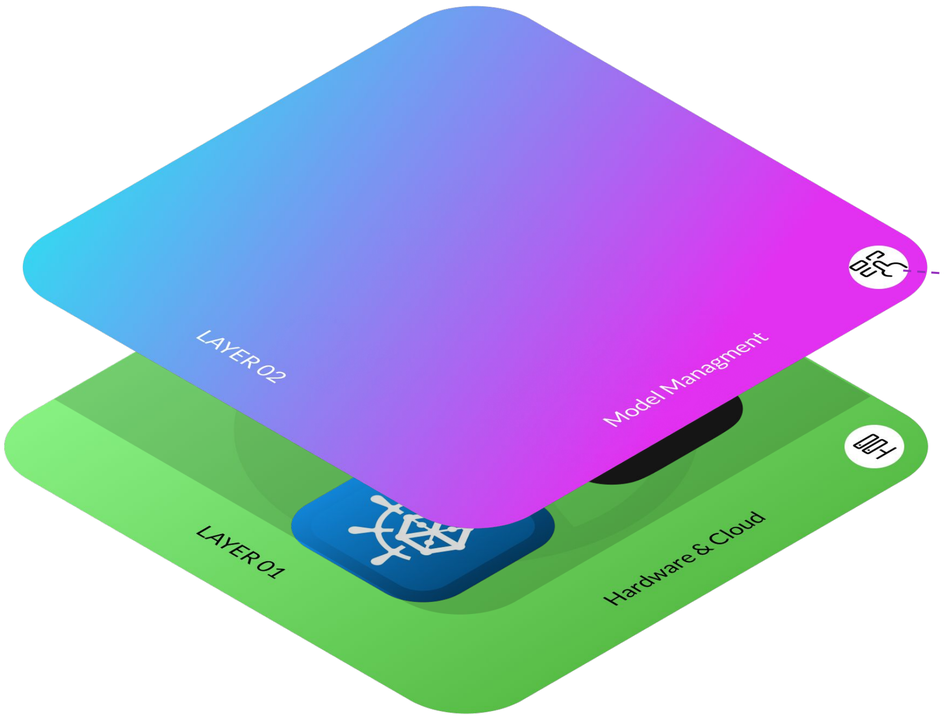
OS



- Base: Ubuntu 24.04 (LTS)
- Drivers Nvidia (CUDA, container-toolkit, smi)
- Dépendances : Docker, Conda, …
- Security: firewall rules, VPN

Modèles Open Source

Serveur Inférence - Plusieurs projets selon l'usage



2

Layer 02: Model Management

Text Gen. & Tools
LLM with tools support
vLLM

Code
Code Gen. &Chat
VLLM

RAG
embed
rerank
Infinity

Multimodal
Vision LM
vLLM

Image Gen.
Image Gen.
ComfyUI

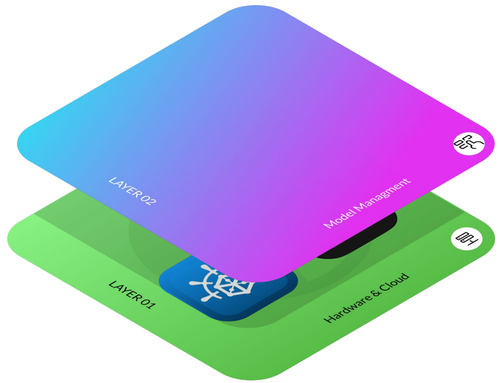
Reasoning
Reasoning LM
vLLM

Speech to Text
Speech to Text
Faster whisper

Text to Speech
Text to Speech
Kokoro

Modèles Open Source

Partage du GPU - Plusieurs Modèles



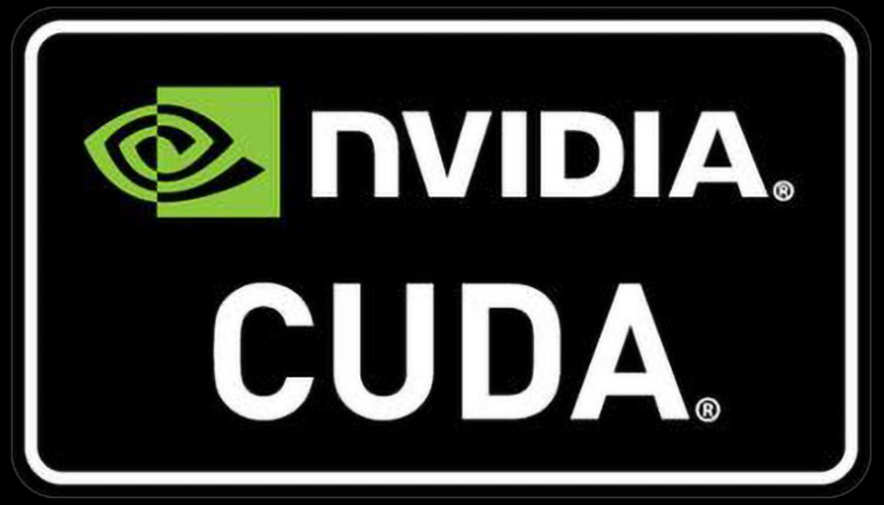
Dans le cas de plusieurs Modèles:

- Partage du temps d’accès au GPU
- Perception “saccadée” ou ralentie du traitement

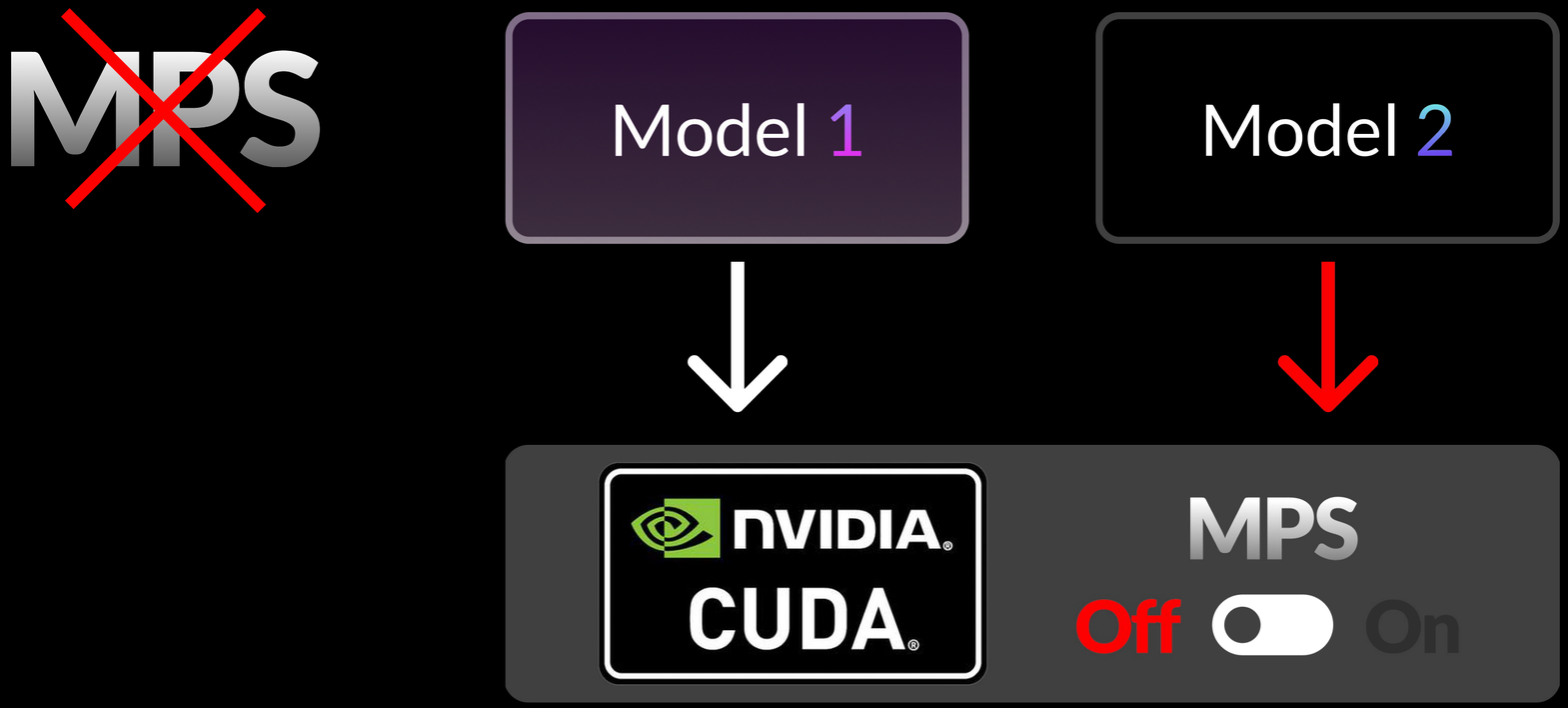
MPS

Multi-Process Service

https://docs.nvidia.com/deploy/mps/index.html



Compute Unified Device Architecture

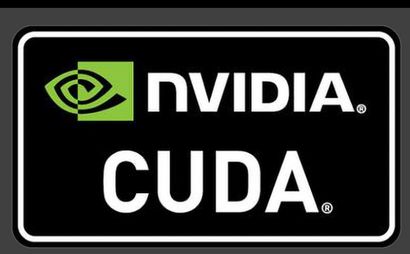

MPS
Model 1
Model 2
NVIDIA
CUDA
MPS
Off
On

Model 1

Model 2

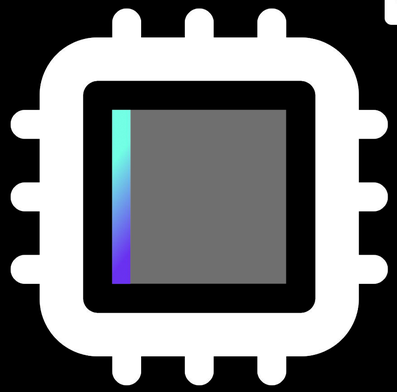
Model 1

Model 2



MPS

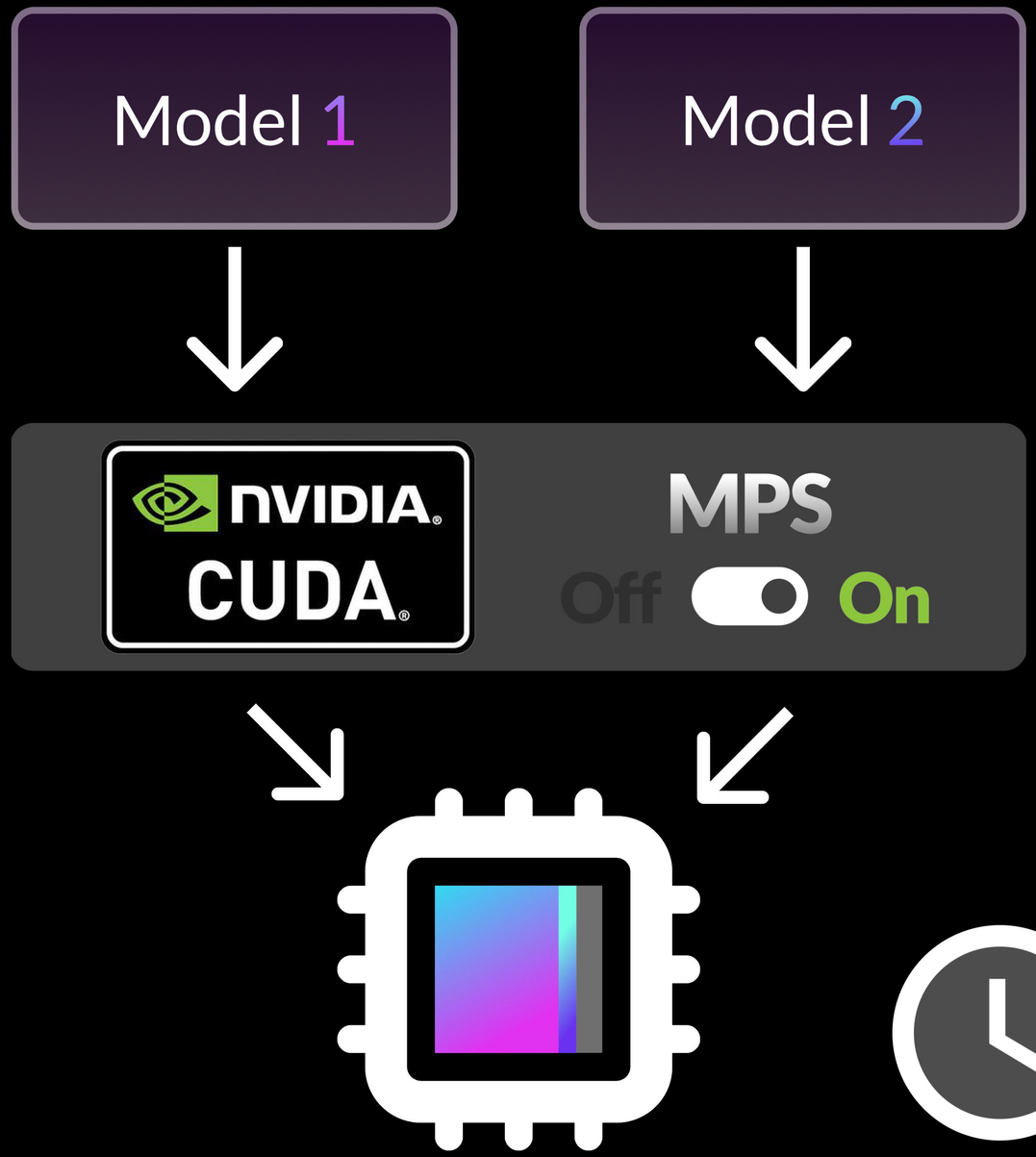
Off On



40%

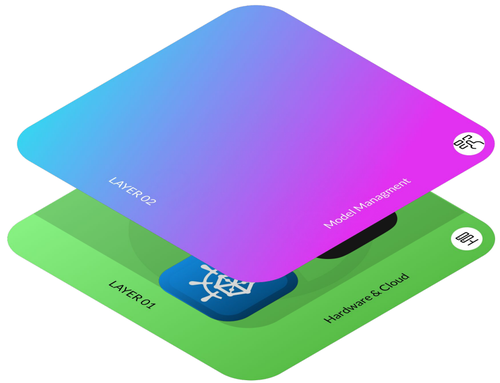

NVIDIA
CUDA
MPS
Off
On

MPS



Modèles Open Source

Partage du GPU - Plusieurs Modèles

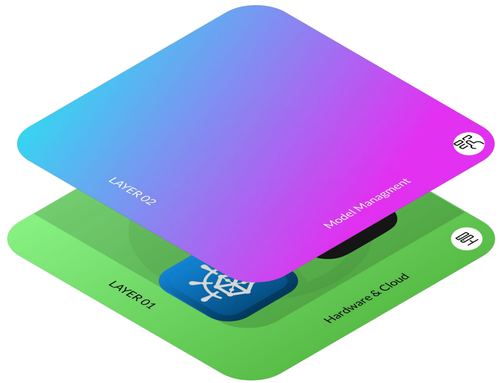


Grâce au MPS:

- Optimisation du GPU
- Plus efficient en Multi-modèle
- Fluidité des traitements

Modèles Open Source

Et notre cas d'usage alors ?

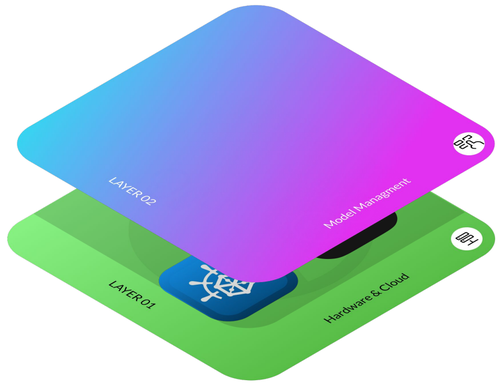


Notre LLM 72B "FP8" sur H200:

- Taille : 72GB
- Contexte 32 k tokens: 10GB
- TOTAL: 82 GB

Modèles Open Source

Et notre cas d'usage alors ?



Avec la place restante on peut pousser le cas:

- Modèles Embedding / Reranking: RAG
- Modèle de Reasoning 32B
- Modèle de code 32B pour les dev

Modèles Open Source

On a déployé nos modèles

Exploitation

Exploitation

AI stack Management operations

Orchestrateur (LiteLLM)

- Endpoint API “OpenAI”
- Gestion utilisateurs / Modèles
- Routage des modèles

Exploitation

On a notre équivalent “OpenAI”

…

Selon la vision utilisateur

Exploitation

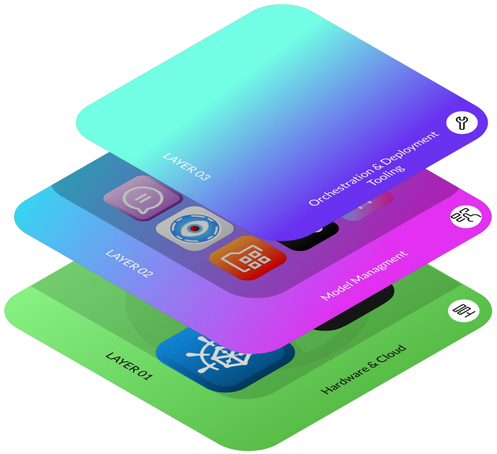
Gestion de la plateforme IA

Metrology (Prometheus / Grafana)

- Métriques HW et LLM
- Tracing Applications LLM
- Mesure Empreinte Carbone

Exploitation

Gestion de la plateforme IA



Outils (Continue.dev, IG1)

- Serveur Configuration Copilot Dev
- Traducteur API Ollama < > OpenAI

Exploitation

On a une plateforme IA privée et complète

Sympa ton API mais OpenAI ce n’est pas qu’une API.

Applications

Applications

Les App OpenAI

OpenAI propose plusieurs Applications :

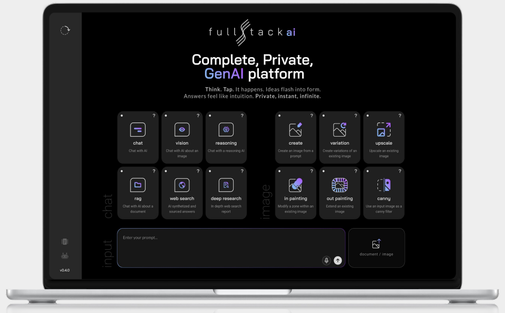
- ChatGPT
- Les GPTs,
- Operator
- Un SDK

Applications

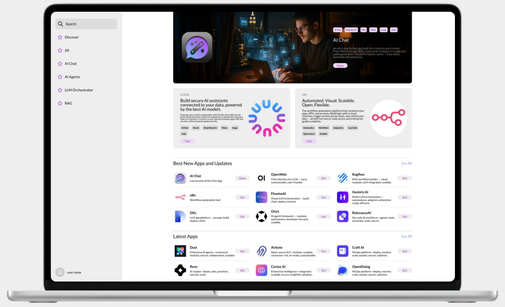
Layer 04: AI Applications

4

AI App Chat



AI AppStore App Custom (via API)



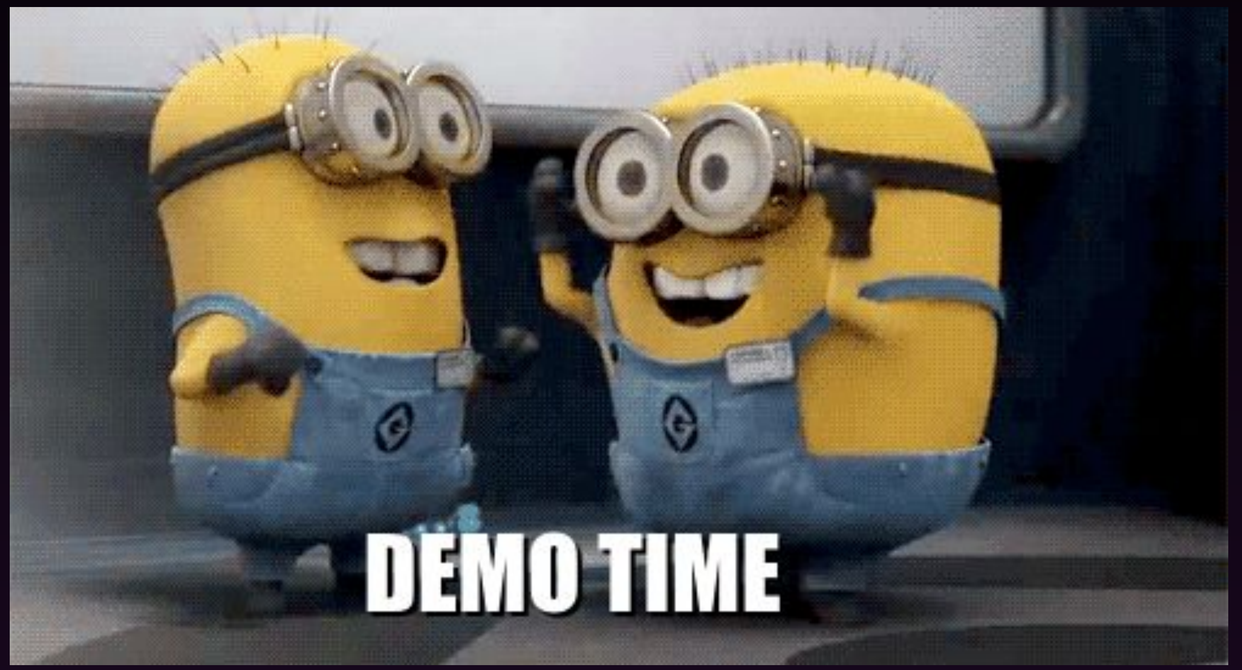
Applications

Les App Open Source pour sa Plateforme IA

Projets intéressants pour sa stack IA:

- OpenWebUI : Chat
- Perplexica: Web Search
- ComfyUI: Pipeline Génération image
- Onyx, Ragflow, Dify: Chatbot RAG
- N8N : Agents IA

Demo ?



Demo

Réalisons notre Assistant IA

Besoin : Assistant IA pour accélérer vos tâches

On veut un assistant capable de:

- répondre à tout type de demande internes
- intégré dans notre messagerie
- répond dans la langue de la question
- s’appuier sur nos docs et les docs externes

Demo

Assistant IA

Pour cet assistant : Onyx

- Outil OpenSource
- Nombreuses intégrations avec Bases de connaissances
- RAG puissant intégré

C’est simple non ?

MERCI !

Jean-Philippe Foures

VP Product

@jpfoures

@jaypif

Iguane Solutions

fullstackai

MERCI !

Jean-Philippe Foures

VP Product

@jpfoures

@jaypif

Vos feedbacks 🙏 :