KUBERNETES COMMUNITY DAYS BEIJING 2025

GenAI时代的开源: 多样性算力的机遇与挑战

姜逸坤 (@Yikun)，Huawei, Principal Engineer

@Yikun

Huawei, Principal Engineer

- Leads an “upstream first” R&D team
- vllm-project/vllm-ascend maintainer
- PyTorch TAC Member
- Apache Spark Committer / PMC member
- CNCF Volcano reviewer
- Ex-OpenStack Core reviewer

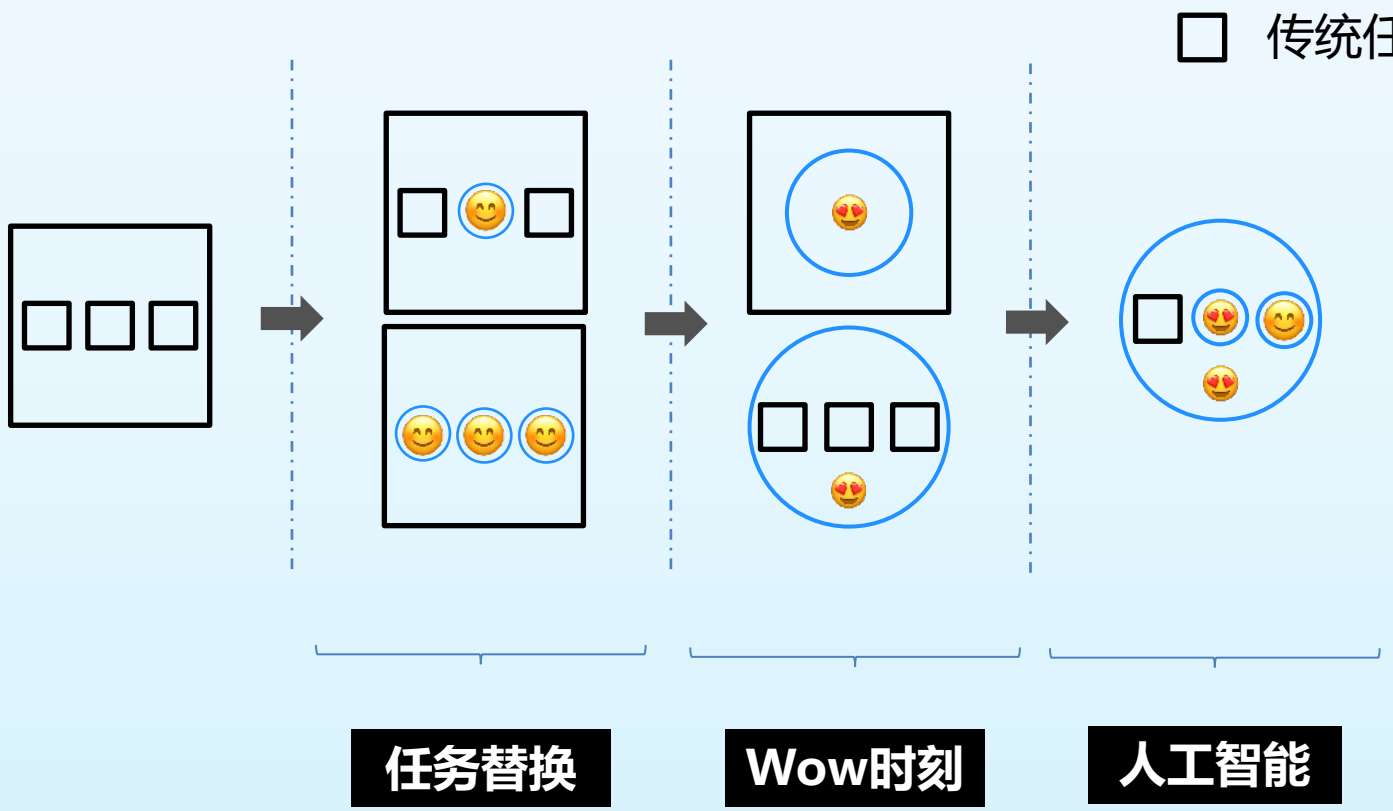
Content

目录

数据、算法、算力

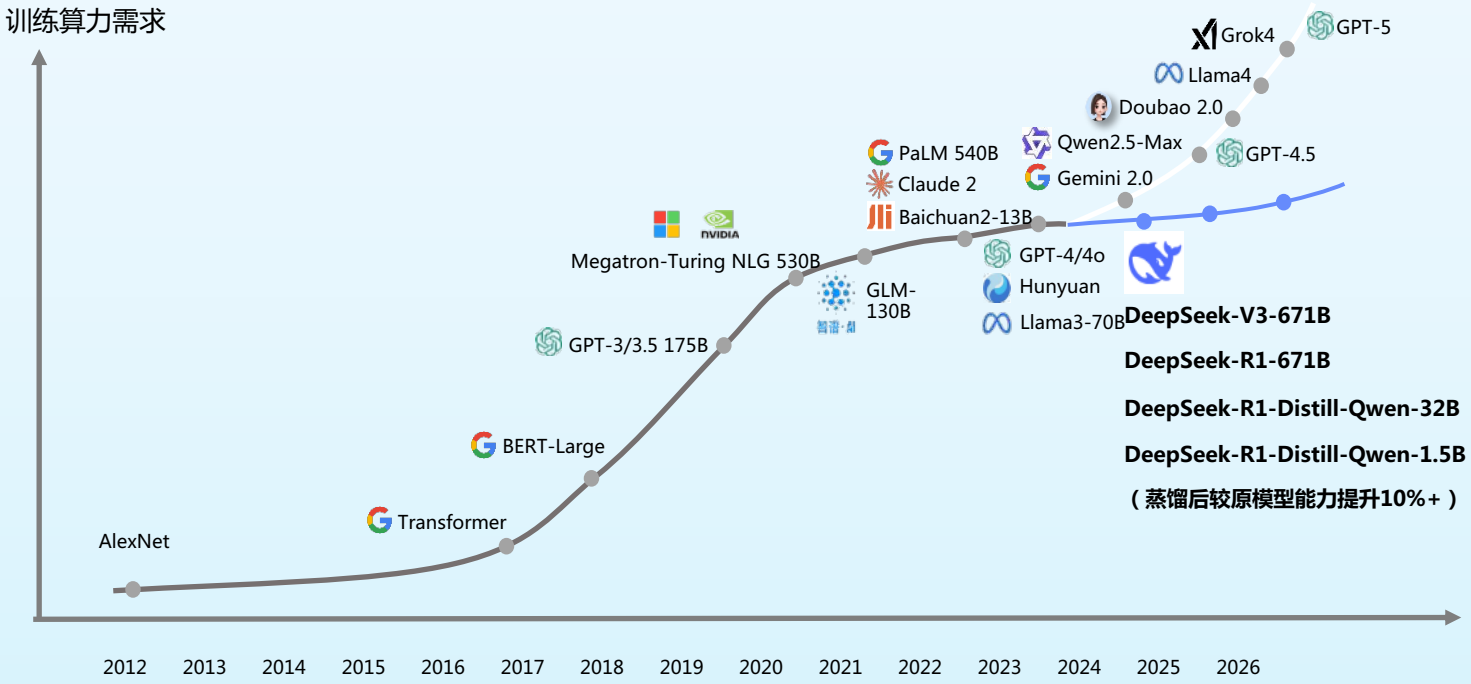
应用 + 模型 + 算力

趋势 #1 应用趋势：正在经历人工智能的Wow时刻



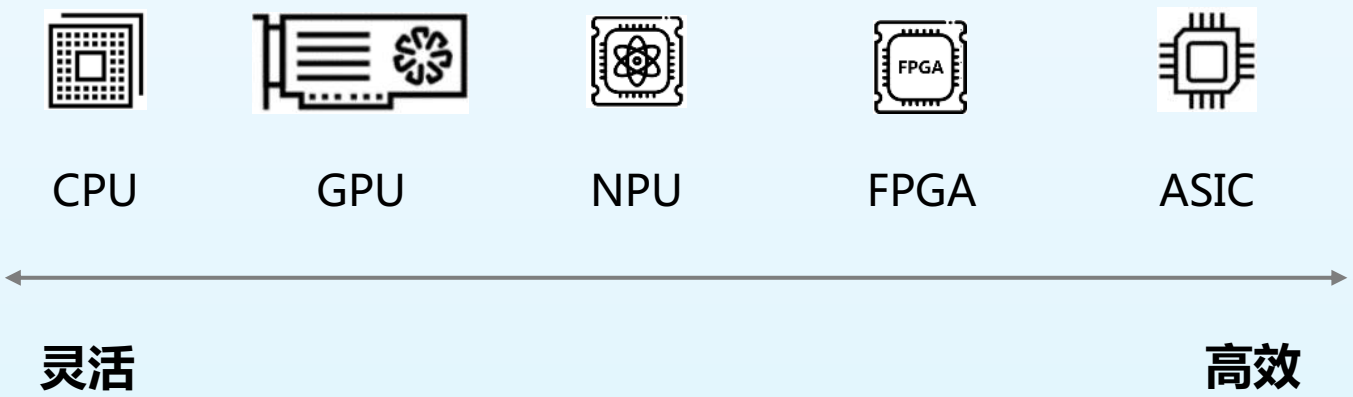
- 从**任务替换**到**人工智能**
- 从**简单任务**到**复杂编排**
- 从**自动化**到**Wow时刻**

趋势 #2 模型趋势：技术摸高与工程创新并行



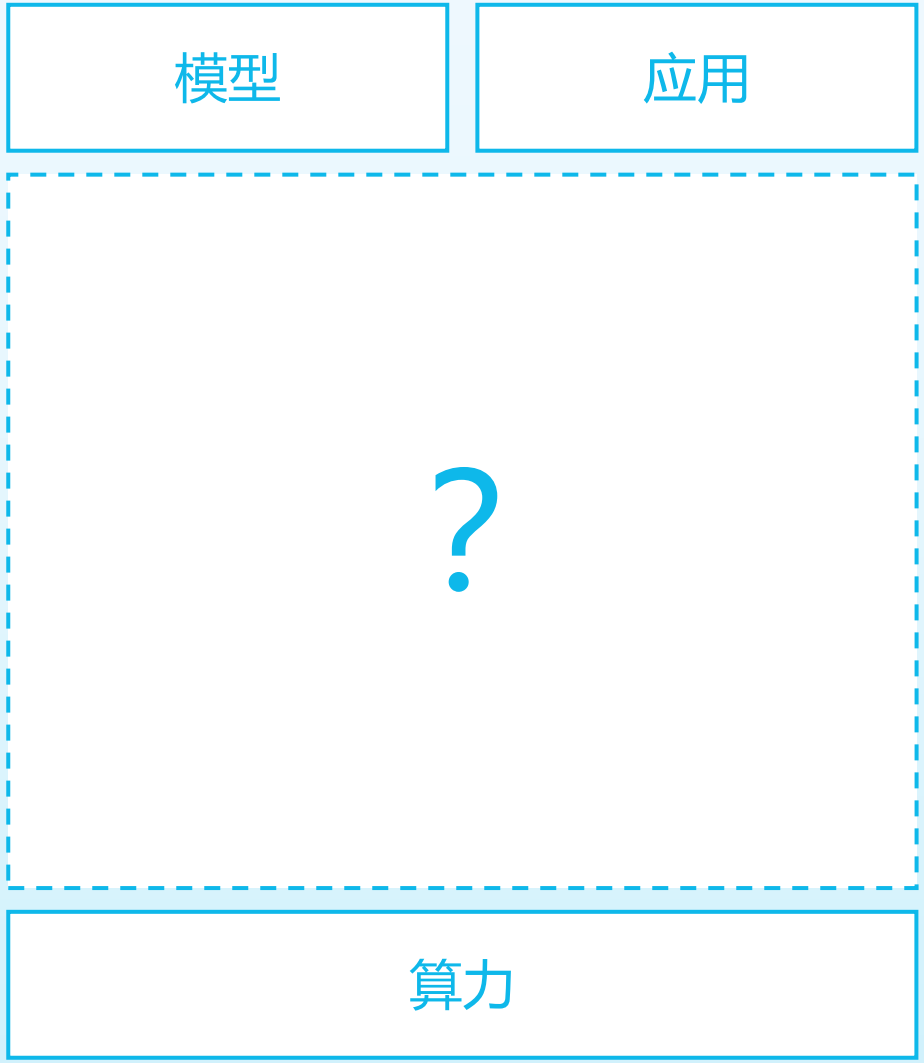
- **技术摸高，算法创新、突破**
- **工程创新，极致性能、成本**

趋势 #3 算力趋势：从单一到多样，在灵活与高效寻找平衡



- 从单一到多样
- 从同构到异构
- 从CPU到GPU、NPU

应对应用、模型、算力的变化：软件栈是什么样子的？



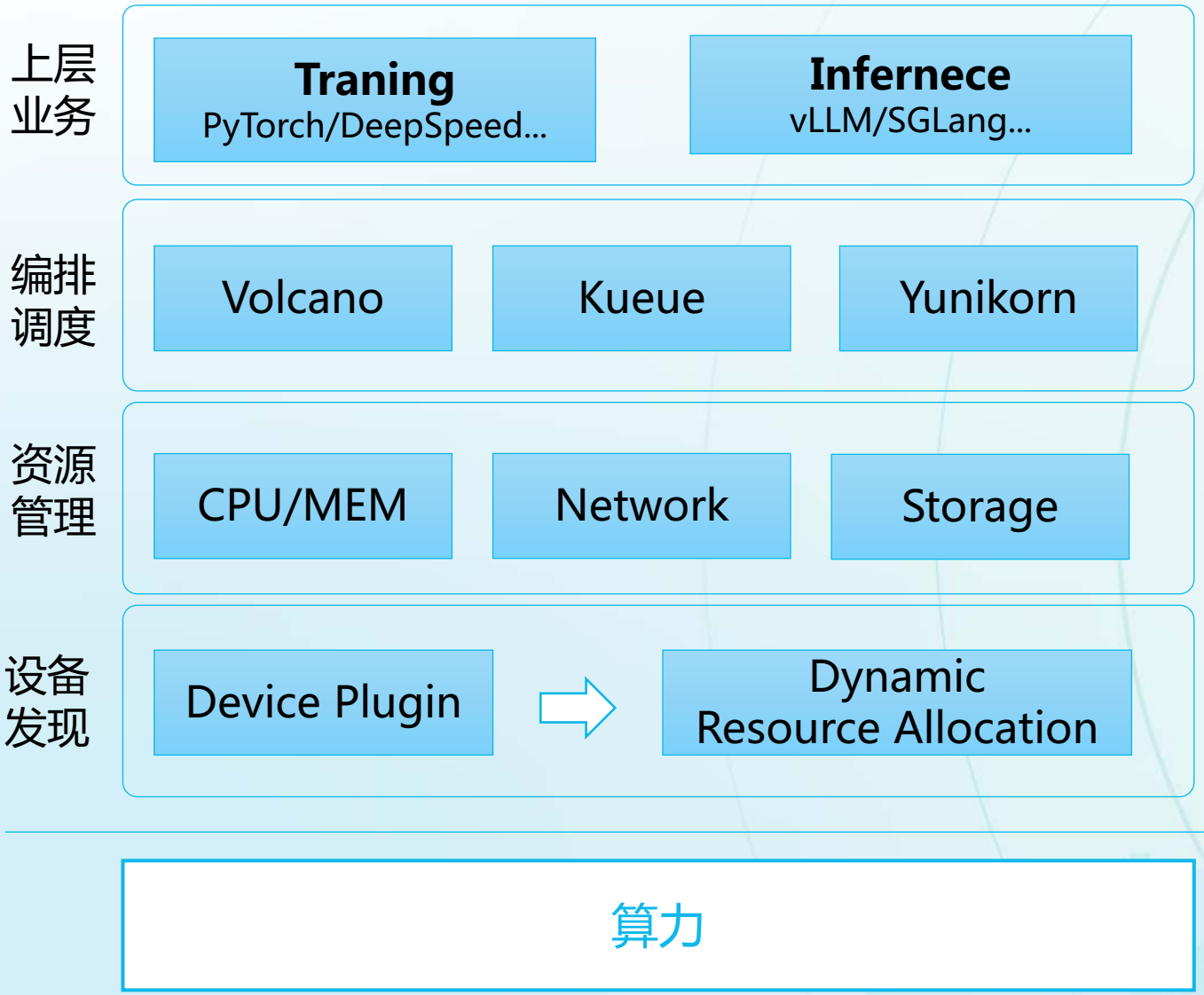
- **模型、应用**与**算力**的GAP

应对应用、模型、算力的变化：软件栈是什么样子的？

kubernetes

- **稳定性**：高可靠、稳定的底座

GenAI时代坚实的底座：通过Kubernetes轻松定义AI Infra



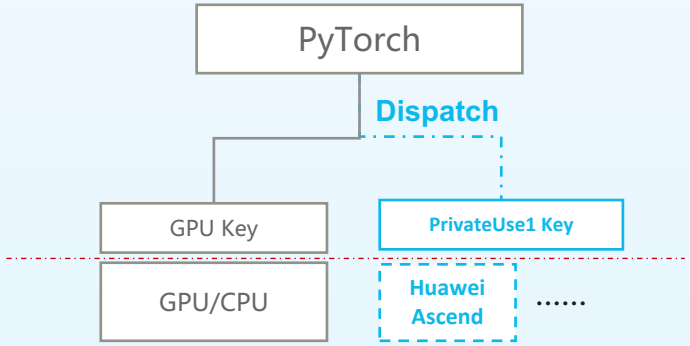

```
apiVersion: v1
kind: Pod
image: ascend/vllm-ascend:v0.7.3

spec:
    schedulerName: volcano
    resources:
      requests:
        cpu: "250m"
        huawei.com/Ascend910: "8"
        ephemeral-storage: "500Gi"
```

应对应用、模型、算力的变化：软件栈是什么样子的？

PyTorch

- **灵活性**：具备足够泛化的抽象能力
- **稳定性**：提供高可靠、稳定的底座

推动 PyTorch 多样性算力支持，官方支持昇腾



- v2.1: [RFC] Improved device support: PrivateUse1

- v2.4: The interoperability Standard of Third-party Backend Integartion Mechanism

- v2.5: [RFC] Autoload Device Extension

- v2.6: [RFC] Open Registration Extension

- [Working Group] [RFC] Accelerator test and CI

PyTorch 1.1　PyTorch 2.1　PyTorch 2.2　PyTorch 2.3　……

2023年起，PyTorch主流版本与官方社区同步发布

华为加入PyTorch基金会，成为Premier会员

95+%	80+%	400+
基础及高级功能覆盖	主流官方库支持	主流模型支持

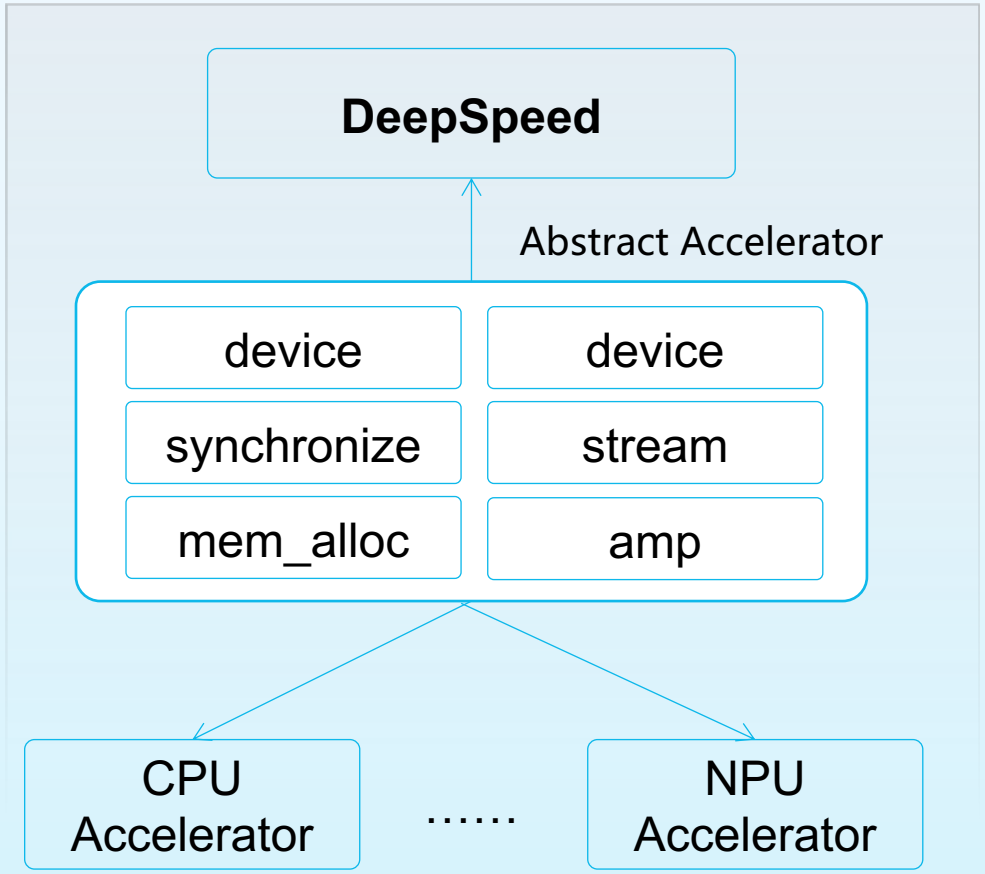
```
$ pip install torch torch-npu
>>> import torch
>>> torch.npu.is_available()
>>> torch.rand(2025, 03, 15).npu()
```

应对应用、模型、算力的变化：软件栈是什么样子的？

deepspeed vLLM

- **高性能**：充分释放多样性算力性能
- **灵活性**：具备足够泛化的抽象能力
- **稳定性**：提供高可靠、稳定的底座

故事1：DeepSpeed原生支持昇腾，深度加速大模型训练



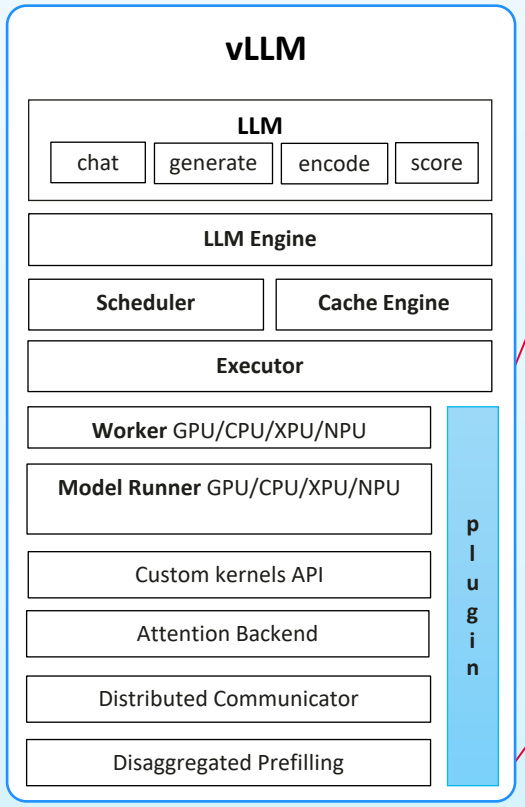
deepspeed × Ascend

```
$ pip install deepspeed
>>> from deepspeed.accelerator import get_accelerator
>>> print('accelerator:', get_accelerator()._name)
accelerator: npu
```

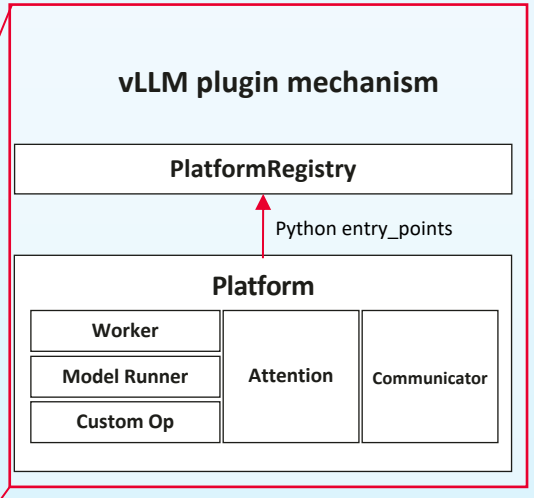
[1] Full feature support with Ascend NPU: https://github.com/deepspeedai/DeepSpeed/issues/4567 (2024.01)

故事2：vLLM原生支持昇腾，加速大模型推理创新

vllm-project/vllm

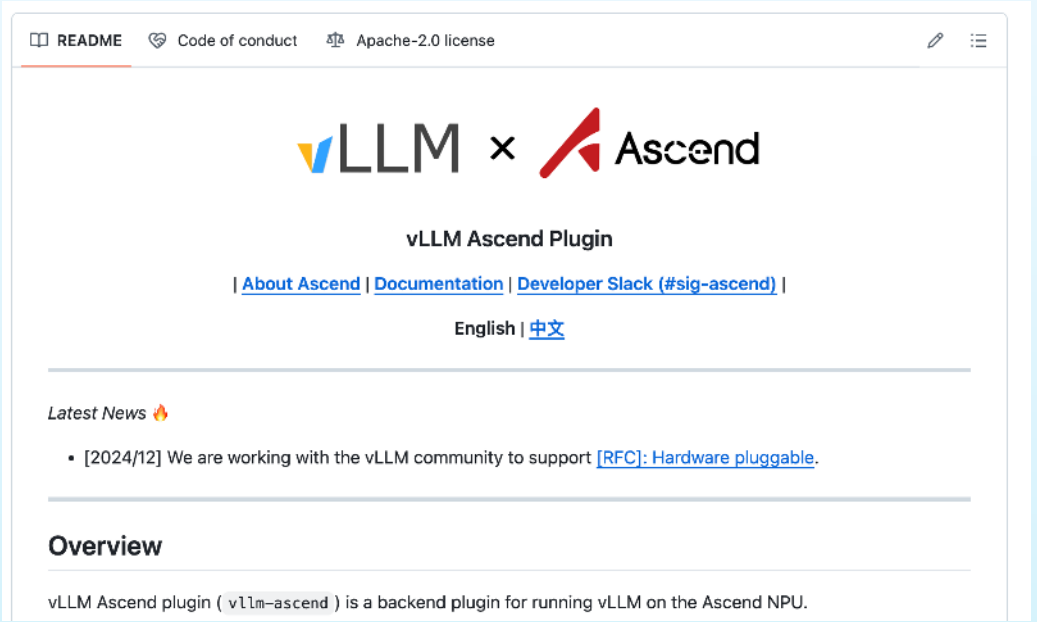


vllm-project/vllm-ascend



$ pip install vllm vllm-ascend

$ vllm serve deepseek-ai/deepseek-r1



[1] https://github.com/vllm-project/vllm-ascend
[2] https://vllm-ascend.readthedocs.io/en/latest

应对应用、模型、算力的变化：软件栈是什么样子的？

- **易用性**：足以应对快速变化的应用
- **高性能**：充分释放多样性算力性能
- **灵活性**：具备足够泛化的抽象能力
- **稳定性**：提供高可靠、稳定的底座

工具链：积极融入主流开源生态，加速昇腾原生支持与创新

主流开源模型

数百个开源模型开箱即用

DeepSeek	Qwen	Baichuan	Open-Sora
Llama	Mistral	GLM	.....

Hugging Face 语言、图文大模型

Transformers 原生支持100+大语言模型

Diffusers 原生支持10+图文模型

OpenMMLab 机器视觉主流模型

MMDetection等原生测试100%通过

业界主流工具链及加速库

原生支持业界主流20+主流工具链、加速库

Datasets	OpenCompass
deepspeed	TRL
LLaMA-Factory	SD Web UI

15+

工具链**原生支持**
覆盖语言/图文/语音
等主流场景

ONNX Runtime	PEFT
Whisper	LMDeploy
llama.cpp	FastChat

10+

训推场景主流加速库、
引擎**原生支持昇腾**

Take away

模型

应用

Hugging Face 工具链

deepspeed vLLM 加速库/引擎

PyTorch 框架

kubernetes 底座

算力

目前正在、已经昇腾原生支持开源软件：

模型社区：

Gitee AI： https://ai.gitee.com/apps 魔乐社区：https://modelers.cn/

微调、工具链：

•Huggingface transformers(since v4.32, 2023): huggingface/transformers/pull/24879

•Huggingface peft (since 0.5.0 , 2023) : huggingface/peft/pull/772

•Huggingface accelerate(since 0.22.0 , 2023): huggingface/accelerate/pull/1676

•LLaMA-Factory (since v0.7.1, 2024) : hiyouga/LLaMA-Factory/pull/975

•FastChat (since v0.2.29, 2023): lm-sys/FastChat/pull/2422

•stable-diffusion-webui (since v1.8.0 , 2024): stable-diffusion-webui/pull/14801

•text-generation-webui (since v1.8 , 2024): text-generation-webui/pull/5541

•OpenCompass (since v0.3.4, 2024): opencompass/pull/1250 & 1618

•lm-evaluation-harness (since v0.4.4, 2024): lm-evaluation-harness/pull/1886

•ComfyUI (since Dec.2024): ComfyUI/pull/5436

•DeepSpeed (since 2024.01)

推理引擎：

vLLM vllm-project/vllm-ascend

ONNX Runtime (since v1.13.1) microsoft/onnxruntime/pull/12416

llama.cpp (since July.2024) llama.cpp/pull/6035

Whisper.cpp (since Aug, 2024) whisper.cpp/pull/2336

AI框架：

PyTorch (since 2.1, 2023) pytorch/releases/tag/v2.1.0

MindSpore(since 1.0, 2020)

KCD GROWING CLOUD NATIVE TOGETHER BEIJING

Thanks.