

面向RISC-V CPU大模型推理引擎 PerfXLM移植与优化

张先轶
澎峰科技

xianyi@perfxlab.com

关于澎峰科技

- 2016年，澎峰科技 (PerfXLab)成立，核心团队来自中科院
- 公司一直致力于研发**算力基础软件及AI解决方案**（高性能计算库、异构计算框架以及软硬融合解决方案等），为算力芯片和算力应用行业加速计算解决方案（华为，燧原，平头哥，华大九天，中船等）

公司主要获奖：

- 2016年，中国计算机学会科技进步二等奖
- 2017年，中国科学院杰出科技成就奖
- 2018年，北京雏鹰人才计划，国家高新企业
- 2021年，数字中国·集成电路赛道特等奖
- 2021年，创芯中国·决赛一等奖
- 2021年，CRVA联盟，软件工作组副组长单位
- 2022年，OpenCAX SIG10组长单位
- 2022年11月15日，ChinaSC中国超级算力大会荣获“算力软件基建领军企业”和“中国智能计算卓越贡献奖”双项荣誉
- 2023年，入选北京市“专精特新”中小企业
- 2023年，入选中国互联网协会算网云协同系统工作委员会成员单位
- 2023年，OpenBLAS获得全球开源贡献Open100
- 2023年，北京市自然科学二等奖



可用

好用

高效使用

行业用户：

- * 加速计算解决方案
- * AI Infra解决方案
- * AI 解决方案(合作伙伴)

PerfXCloud/智算中心：

- 提供成熟的软件产品和技术服务，赋能智算中心升级建设
- 提供私有化部署服务

PerfXAPI® 异构计算软件栈
架起
算力与应用的桥梁

服务

合作

共赢

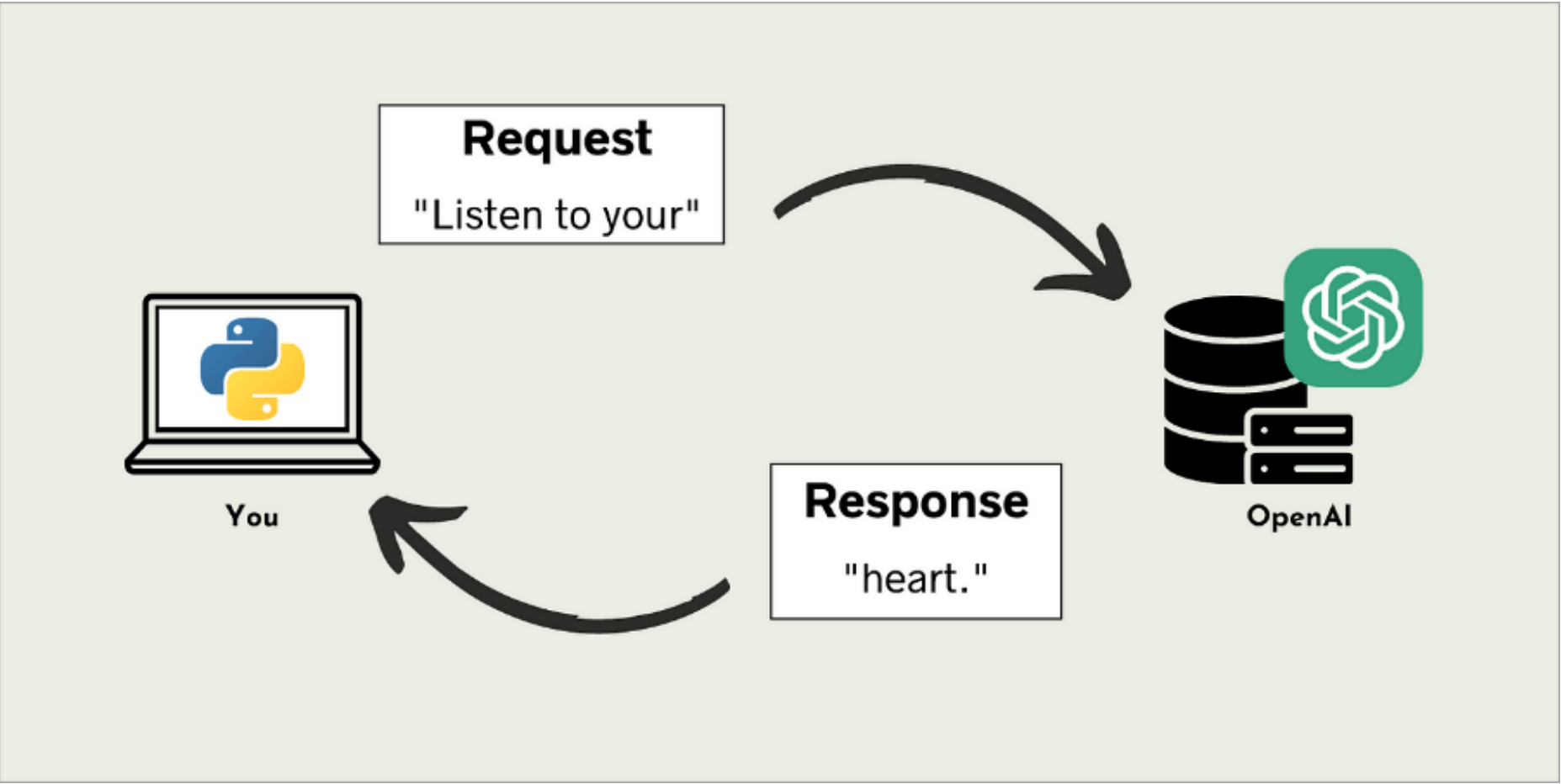
芯片

服务器

算力集群

OpenAI定义了应用调用AI能力的高层API接口，事实标准

- Chat Completions API (gpt-4o, gpt-4-turbo, gpt-3.5, ...)
- Embeddings API (text-embedding-ada-002, ...)
- Image generation API (dall-e-3, ...)
- ...



商业模式变化，MAAS商业模式

Model	Pricing
gpt-4o	US\$5.00 / 1M input tokens US\$15.00 / 1M output tokens
dall-e-3	\$0.040 / image
ada v2	US\$0.10 / 1M tokens



从上层模型至底层工具链，大部分厂商均支持OpenAI接口；按照以Token量计费的MAAS模式进行商业运营。

观察2：大模型应用门槛急剧降低，应用快速涌现

模型广场

模型类型: 大语言模型 视觉大模型 Embedding模型 MoE模型 多模态模型

支持芯片: AMD NVIDIA DCU CA100

搜索您想要查找的模型...

DeepSeek-V2-Chat

大语言模型 MoE模型 DCU NVIDIA AMD

DeepSeek-V2, 这是一种强大的专家混合 (MoE) 语言模型, 其特点是经济的训练和高效的推理。DeepSeek-V2实现了...

2024/07/17 查看模型详情

Meta-Llama-3.1-405B

大语言模型 NVIDIA DCU AMD CA100

Meta AI开发并发布了Meta Llama 3.1系列的大型语言模型 (LLM), 这是一系列预训练和指令调整的生成性文本模型, ...

2024/07/23 查看模型详情

Qwen2-72B-Instruct-GPTQ-

大语言模型 NVIDIA DCU AMD

Qwen2 是 Qwen 大型语言模型的新系列。对于 Qwen2, 发布了许多基础语言模型和指令调优语言模型, 范围从 0.5 ...

2024/06/18 查看模型详情

Bash调用示例:

1.将令牌保存到环境变量

```
export PERF_X_API_KEY="sk-xxxxxxx"
```

2.curl方式发送请求

```
curl https://cloud.perfxlab.cn/v1/chat/completions \
-H "Content-Type: application/json" \
-H "Authorization: Bearer $PERF_X_API_KEY" \
-d '{
  "model": "llama3.1:405b",
  "messages": [
    {"role": "system", "content": "want you to be a chatterbox expert and answer the following in a friendly"},
    {"role": "user", "content": "Say this is a test! "}
  ],
  "temperature": 1,
  "max_tokens": 16,
  "n": 1,
  "stop": ["<|eot_id|>", "<|start_header_id|>", "<|end_header_id|>"],
  "presence_penalty": 0,
  "frequency_penalty": 0,
}
```

大模型服务平台直接调用API服务接口，用户只需编写少量代码，即可在短时间内构建并发布一个大模型应用。
















大模型能力的“涌现”，带来了大模型应用的涌现，国外大量行业应用正在改变人们原有的工作、生活方式。

观察3：开源与闭源大模型相互竞争促进，持续迭代

PerfXLab

开源模型+模型微调之后，能力提升显著

Open-source Model Fine-Tuning Leaderboard

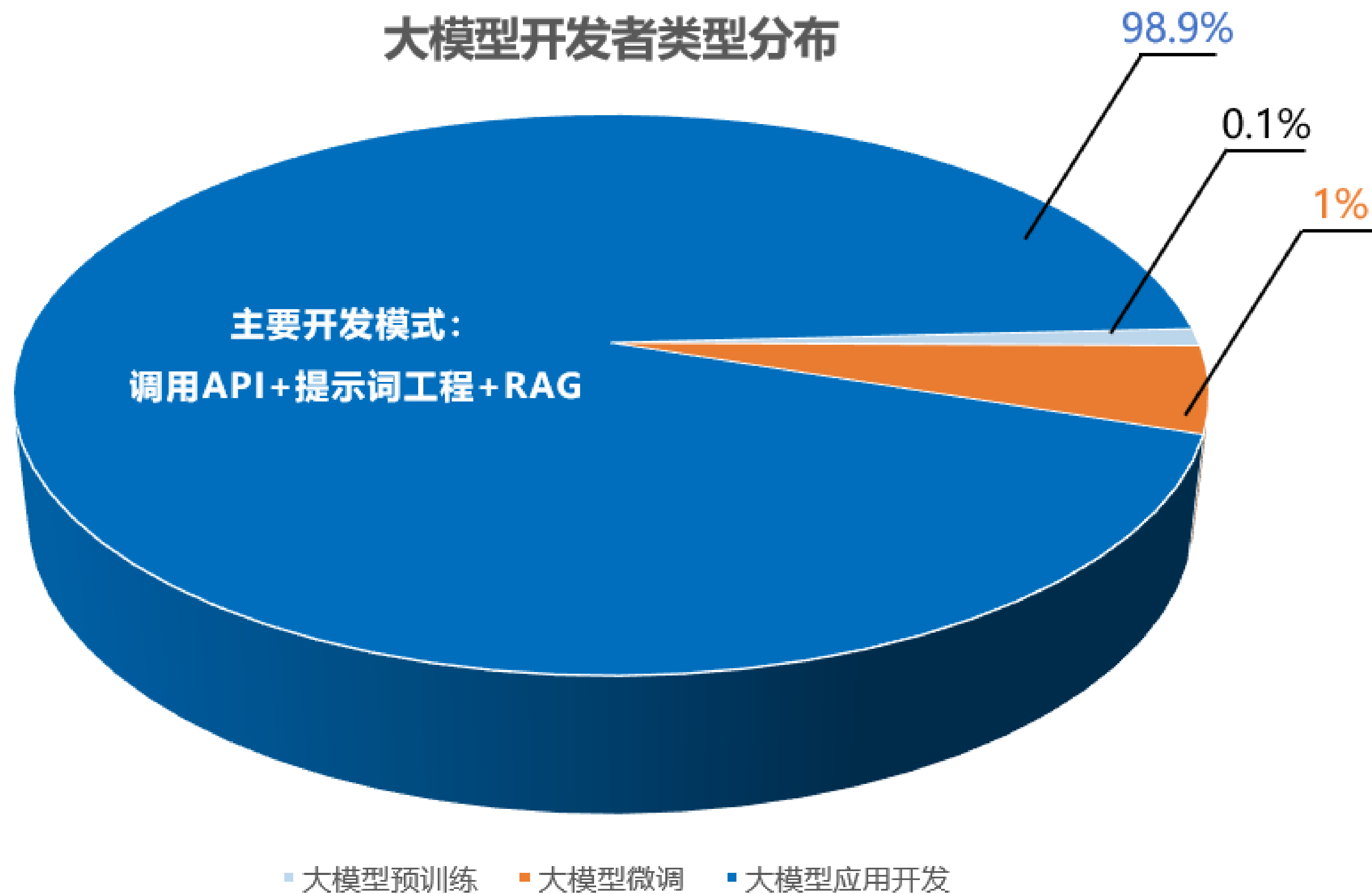
Developer	Model	Performance
	llama-3-8b	0.75
	phi-3-4k	0.74
	zephyr-7b-beta	0.74
	llama-3-8b-instruct	0.74
	mistral-7b	0.73
	mistral-7b-instruct	0.72
	llama-2-7b-chat	0.71
	llama-2-7b	0.70
	phi-2	0.68
	gpt-4	0.66
	gemma-2b	0.66
	gemma-7b-instruct	0.66
	gemma-7b	0.65
	gemma-2b-instruct	0.64
	gpt-3.5-turbo	0.60

开源大模型能力快速提升：中文开放式生成评估结果

模型	开源/闭源	总分
gpt-4-1106-preview	闭源	8.01
DeepSeek-V2 Chat (RL)	开源	7.91
erniebot-4.0-202404 (文心一言)	闭源	7.89
DeepSeek-V2 Chat (SFT)	开源	7.74
gpt-4-0613	闭源	7.53
erniebot-4.0-202312 (文心一言)	闭源	7.36
moonshot-v1-32k-202404 (月之暗面)	闭源	7.22
Qwen1.5-72B-Chat (通义千问)	开源	7.19
Yi-34B-Chat (零一万物)	开源	6.12
gpt-3.5-turbo-0613	闭源	6.08
DeepSeek-V2-Lite 16B Chat	开源	6.01

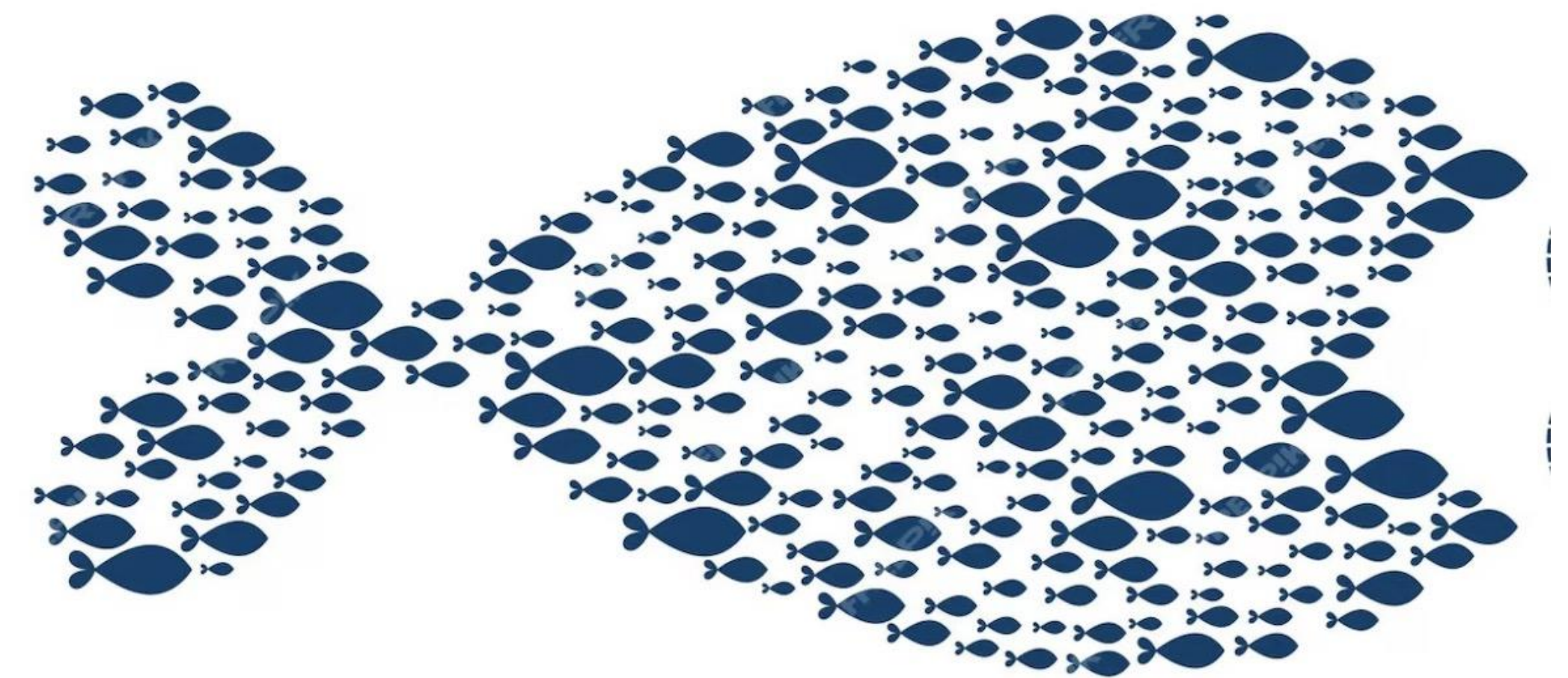
公共智算中心 和 私有算力集群，都将同时运行数十款大模型，并持续迭代。

推论：算力工厂/超级AI Foundry必然出现



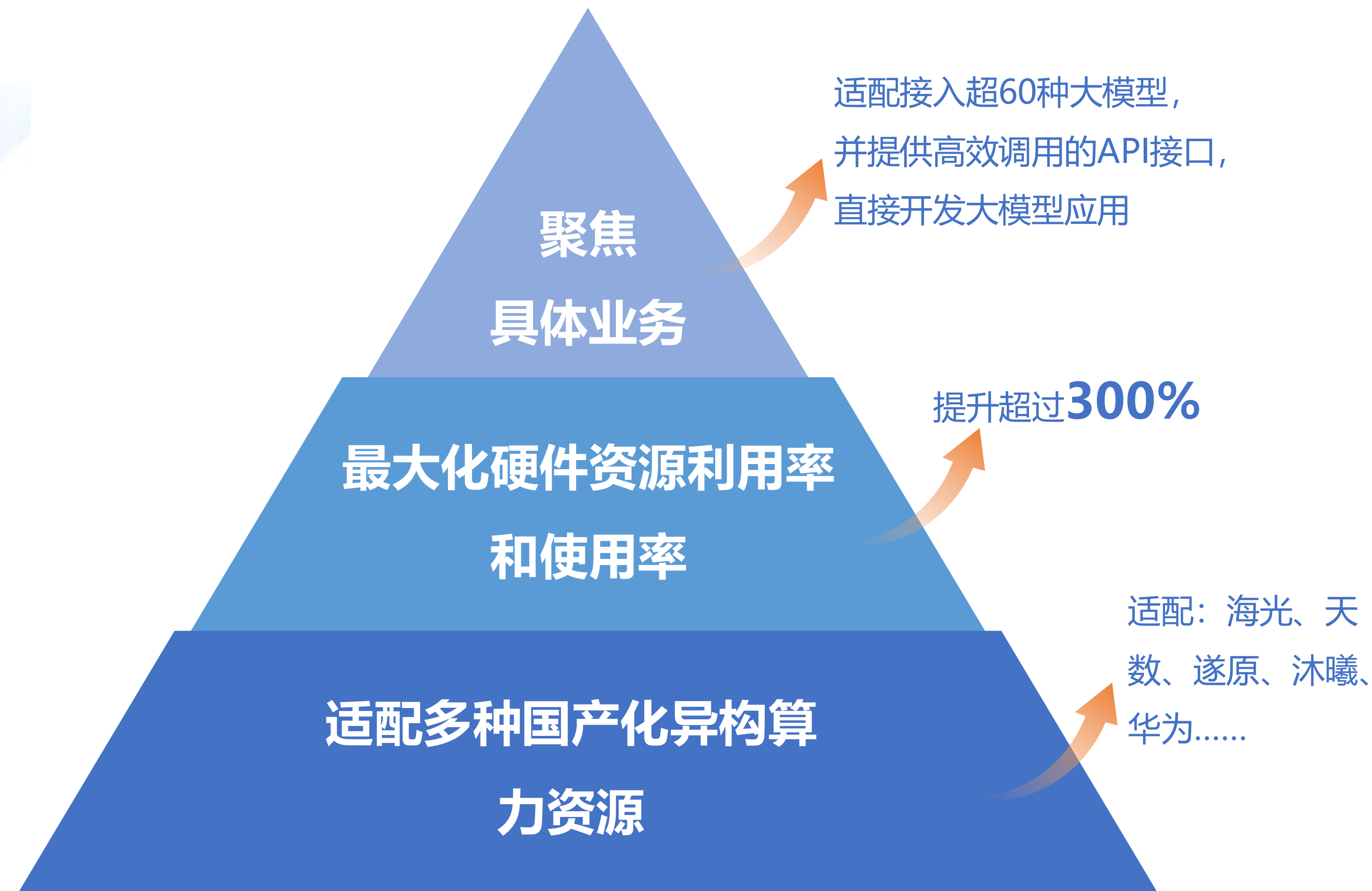
根据机构测算在众多的大模型开发类型中：

- 大模型预训练占比仅0.1%
- 大模型微调开发占比1%
- 大模型应用开发占比**98.9%**



基于大模型应用开发的场景将爆发式
增长，共同形成AI Foundry

PerfXCloud为开发者和企业量身打造的AI开发和部署平台。它专注于大模型的微调 and 推理需求，为用户提供极致便捷的一键部署体验。为算力中心提供大模型AI、科学与工程计算的整体运营解决方案，助力算力中心升级成为“AI超级工厂”。





云端和边缘端一体

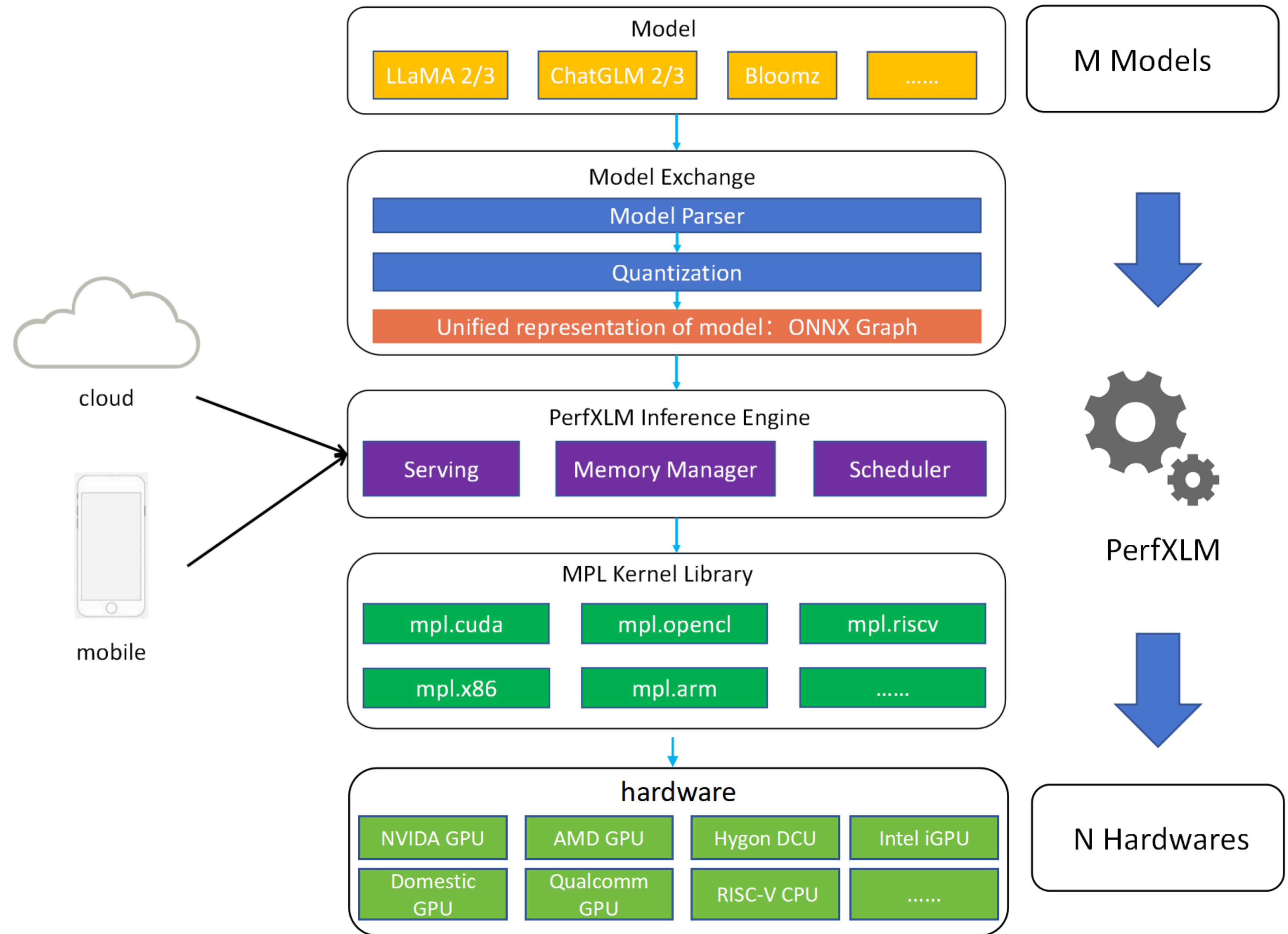
- 支持云端GPU/NPU
- 支持嵌入式GPU/NPU

支持多种硬件

- 国际主流硬件：NVIDIA GPU, Intel iGPU, 高通SoC等
- 国产GPU/NPU硬件：RISC-V CPU, Hygon DCU, 燧原等。

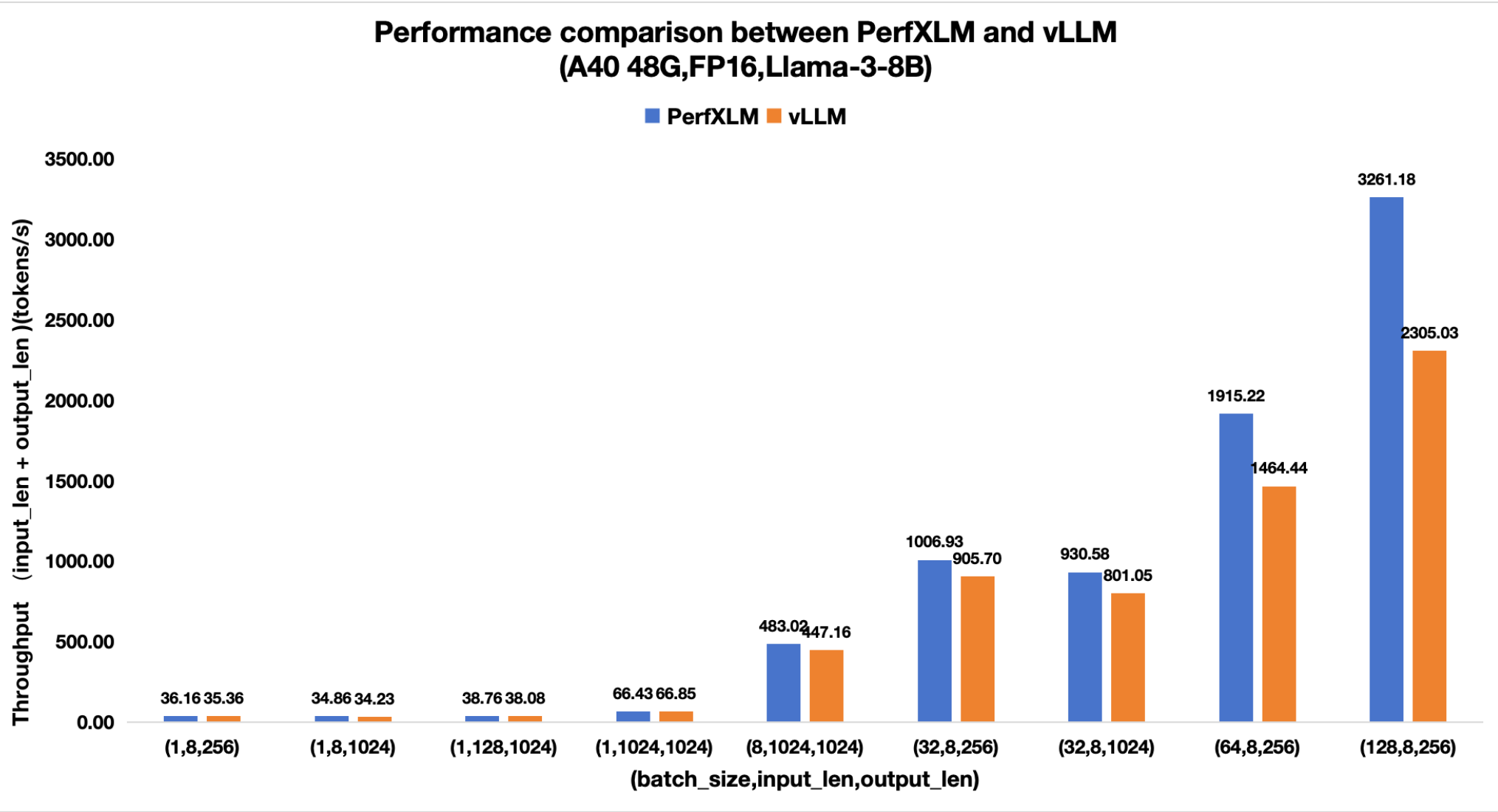
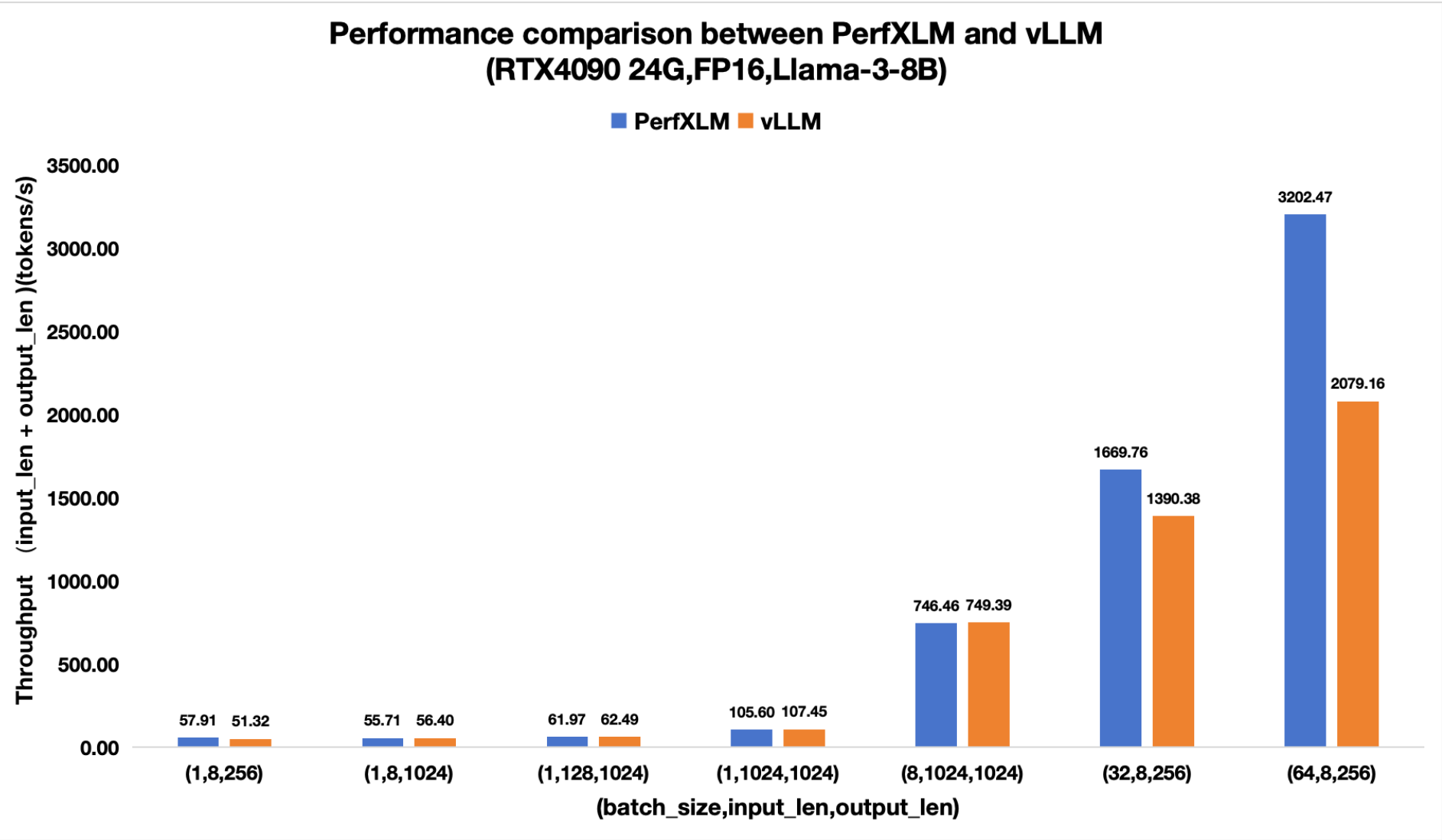
深入性能优化

- 多种算子融合与优化技术
- 核心计算库优化等
- 量化技术
- 多卡并行
- 内存优化



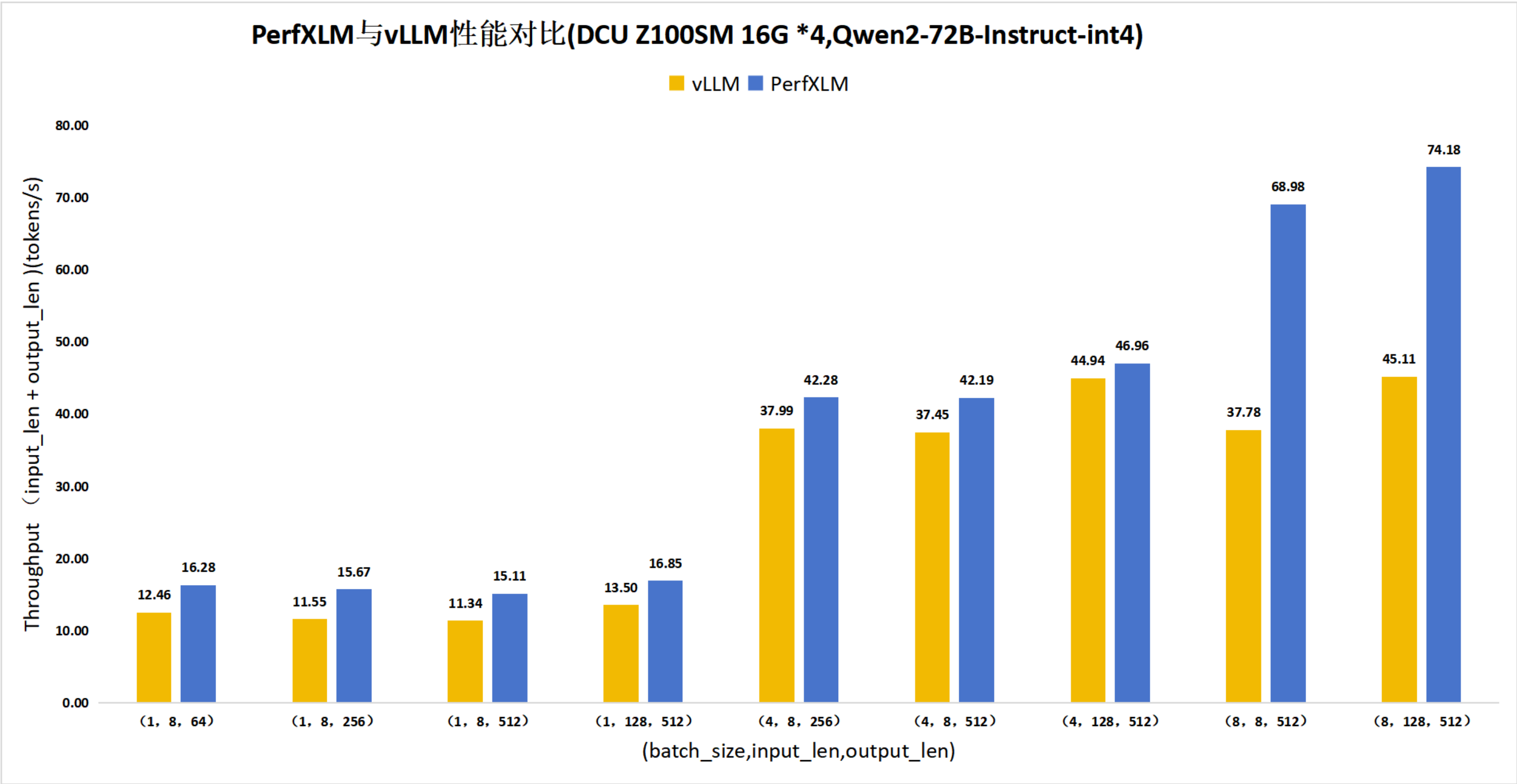
NVIDIA平台初步结果

- Llama3 8B
- RTX4090和A40



DCU平台初步结果

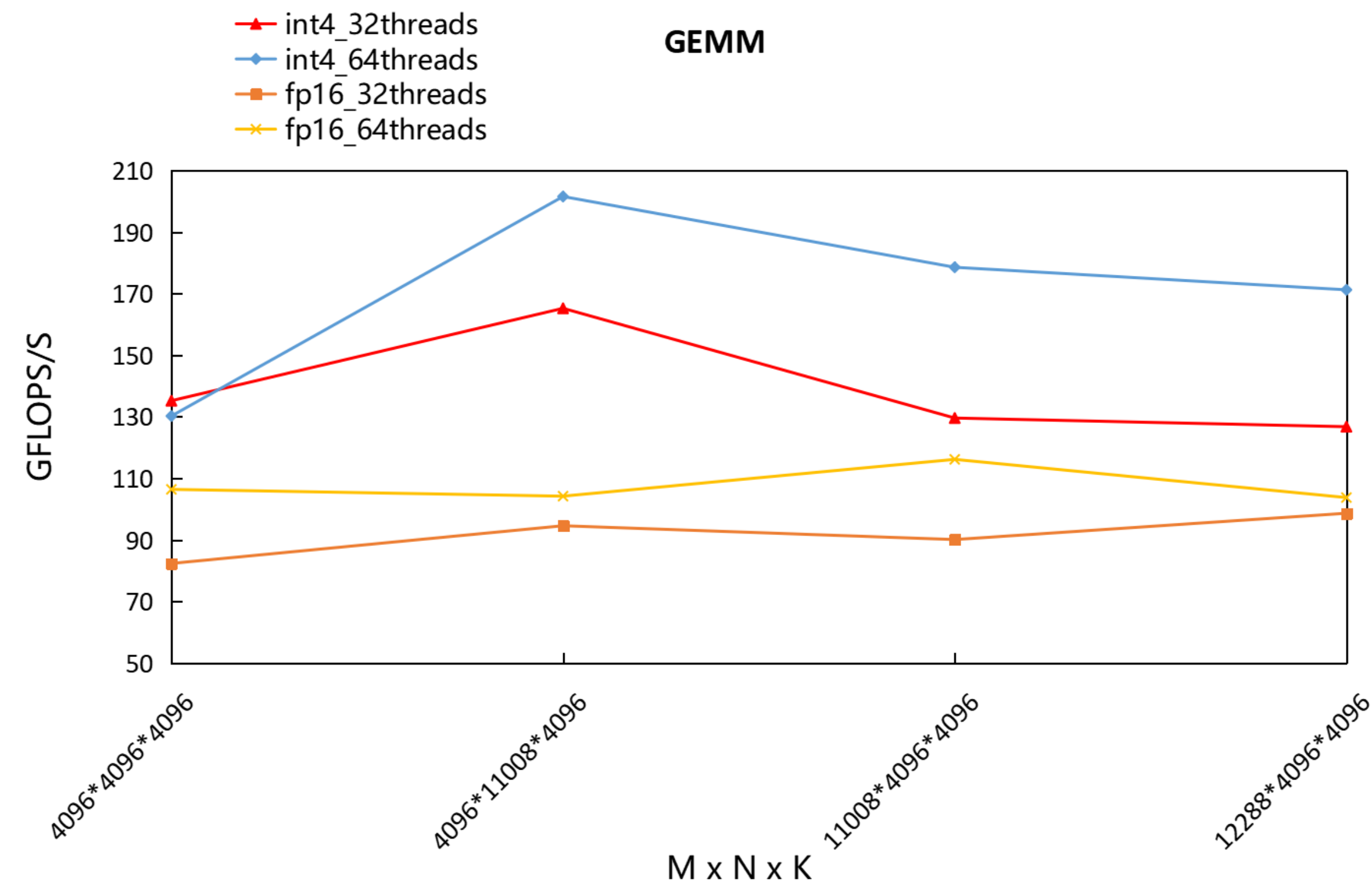
- Qwen2 72B int4



PerfXLM on RISC-V CPU

GEMM optimization for prefill stage:

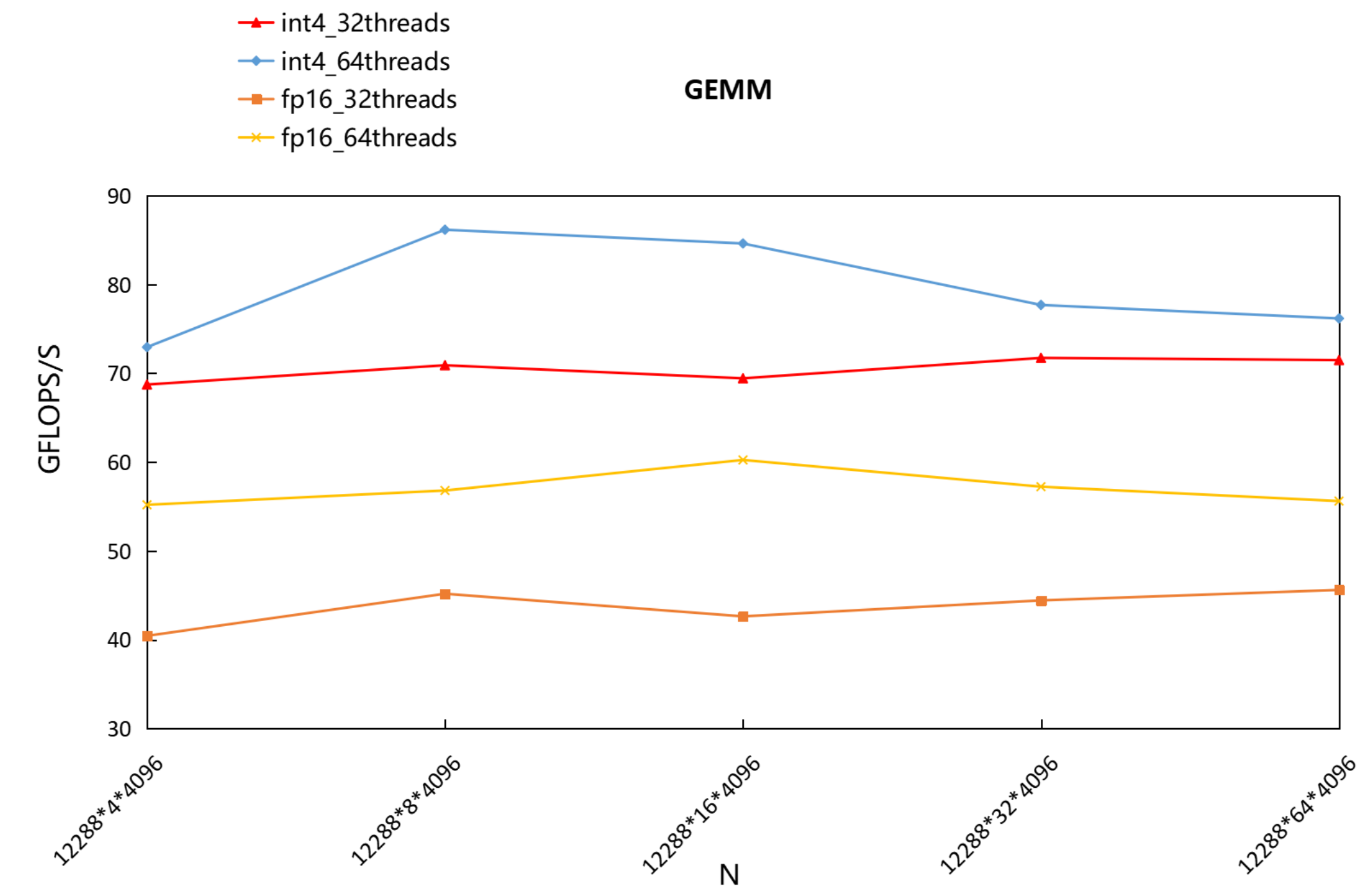
- Single batch: tall-and-skinny matrix
- Vector instruction optimization
- GEMM blocking
- Multi-threading



GEMM Performance (Big and square matrices)

RISC-V CPUs Testbed

- SG2042 CPU
 - Xuantie C920 2.0GHz
 - RVV 0.7.1, 128-bit
 - L1-D 64KB
 - 4 cores/cluster
 - Shared L2 1MB
 - 64 cores (16 cluster)
 - Shared L3 64MB
- 4x DDR4 memory controllers

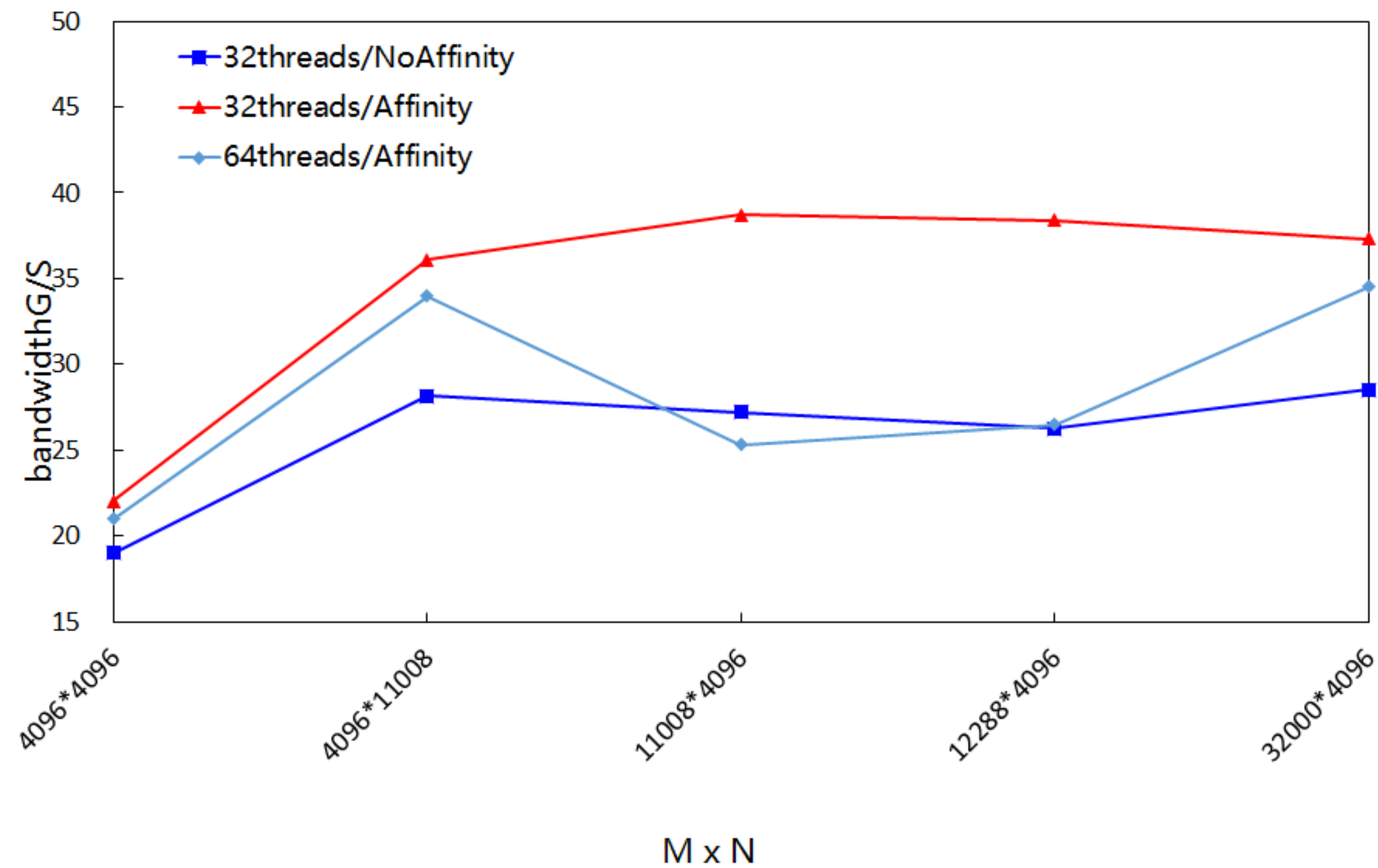


GEMM Performance (tall-and-skinny matrices)

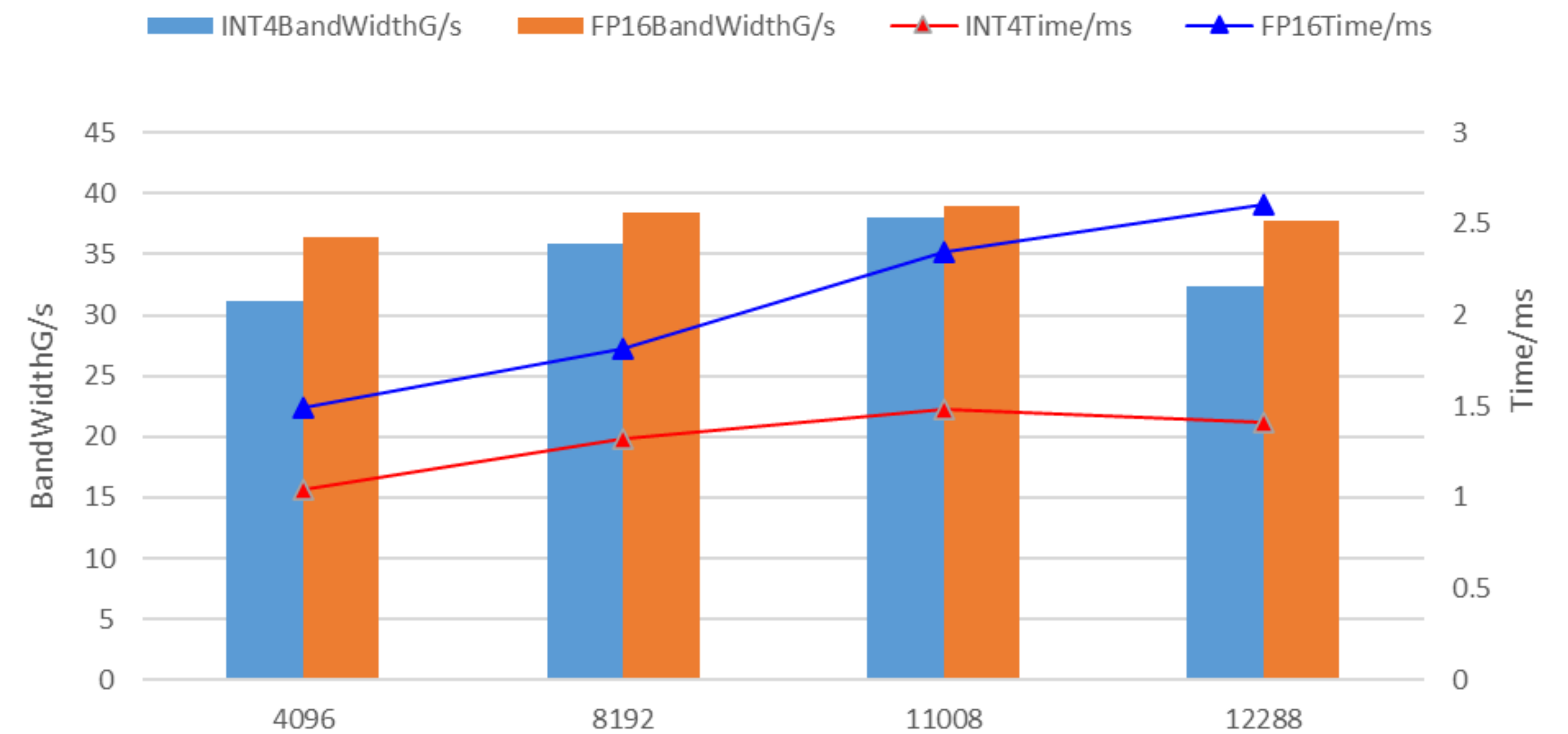
PerfXLM on RISC-V CPU

GEMV optimization for decoding stage:

- Vector instruction optimization



GEMV performance under different number of threads and affinity.



GEMV Performance (32 of threads)

Performance of the GEMV operator after optimization. The input of the GEMV performance test is: n =4096, m changes.


```
Welcome to Ubuntu 23.04 (GNU/Linux 6.1.42 riscv64)

* Documentation: https://help.ubuntu.com
* Management:   https://landscape.canonical.com
* Support:       https://ubuntu.com/pro

System information as of Wed Aug 21 13:16:21 CST 2024

System load: 17.35      Temperature:   46.0 C
Usage of /:   98.3% of 902.73GB Processes:    743
Memory usage: 1%        Users logged in: 0
Swap usage:  0%         IPv4 address for eth0: 192.168.5.11

=> / is using 98.3% of 902.73GB

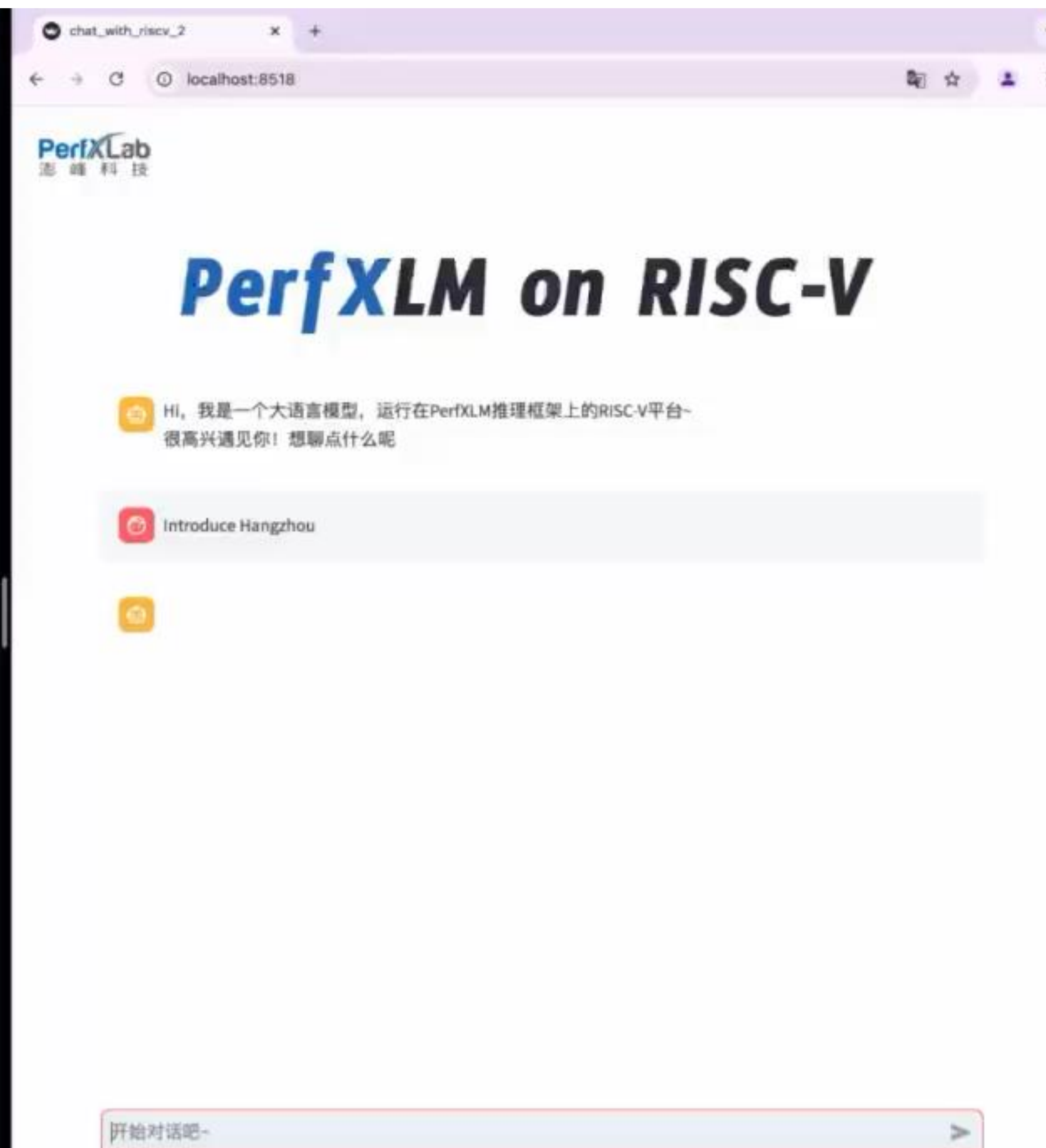
* Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
  just raised the bar for easy, resilient and secure K8s cluster deployment.

https://ubuntu.com/engage/secure-kubernetes-at-the-edge

@ updates can be applied immediately.

Last login: Wed Aug 21 13:10:47 2024 from 192.168.5.10
ubuntu@riscv02:~$ cd /home/ubuntu/yuxinan/perfxlm/build
ubuntu@riscv02:~/yuxinan/perfxlm/build$ export THREAD_POOL_SIZE=32
ubuntu@riscv02:~/yuxinan/perfxlm/build$ bin/llama_example
Total ranks: 1.
P0 is running with CPU.

[PT]: listenOnServer start.
connected to eventl...
{"messages": [{"role": "user", "content": "Introduce Hangzhou"}], "model": "deepseek-v2-chat",
"stop": ["<|eot_id|>", "<|start_header_id|>", "<|end_header_id|>"], "stream": true}
*****perfxlm_request_begin*****
Model: deepseek-v2-chat
Role: user
Content: Introduce Hangzhou
Temperature: 0
Top_p: 0
Max_length: 512
Stream: true
*****perfxlm_request_end*****
line: <s>[INST] Introduce Hangzhou [/INST]
[PT]: llama.forward start
[]
```



算力经济中的超级AI Foundry条件具备

- API接口标准化, 商业模式
- 应用需求爆发
- 开源大模型能力提升

PerfXLM+PerfXCloud+RISC-V的机会

- 模型推理和微调平台, API兼容接口, MAAS模式
- 用户替换0成本

大家都在问: AI/大模型超级应用在哪里?

- 需要一定等待超级应用出现吗
- 各行各业受益, 将来数以百万级AI应用遍地开花

