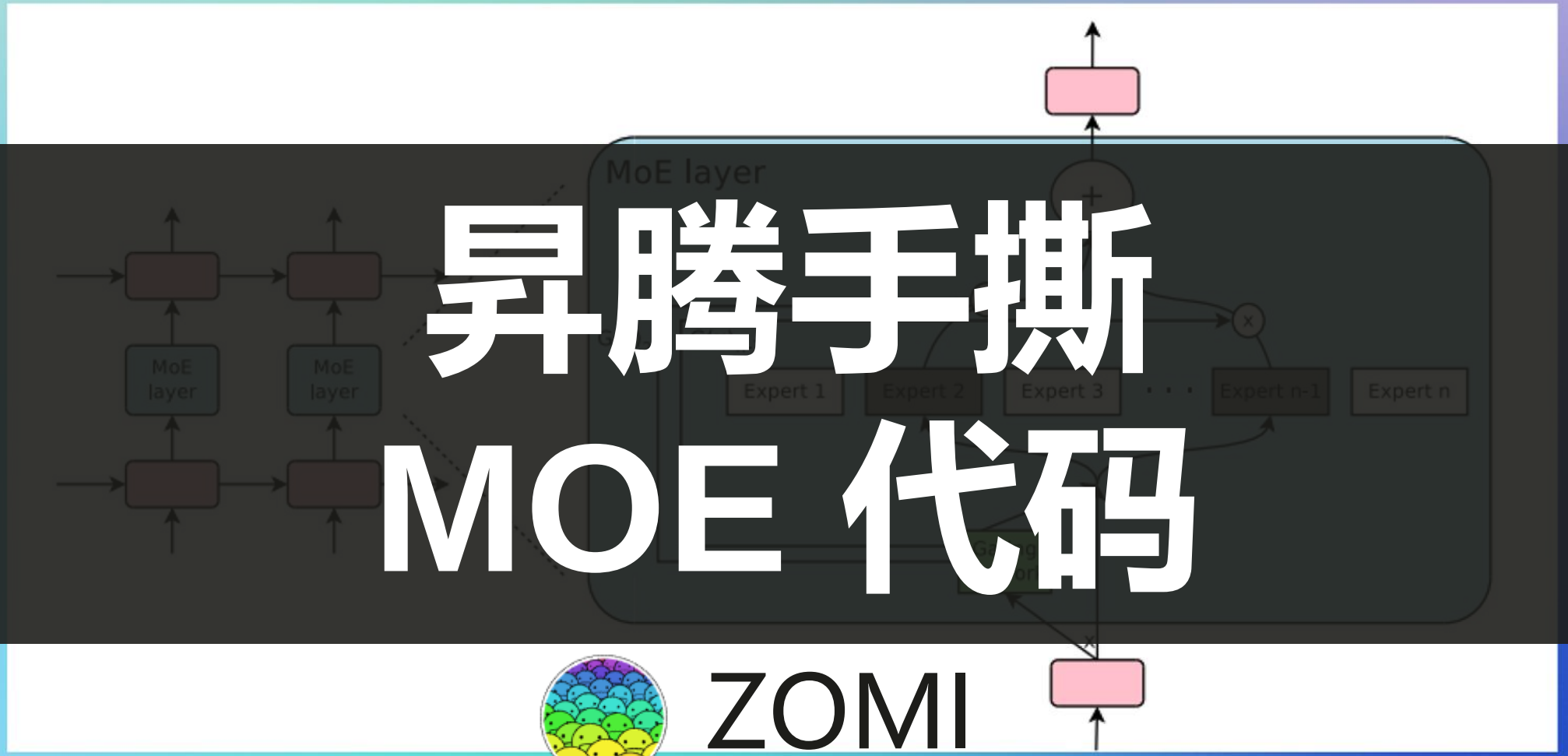
Mixture of Experts (MoE)

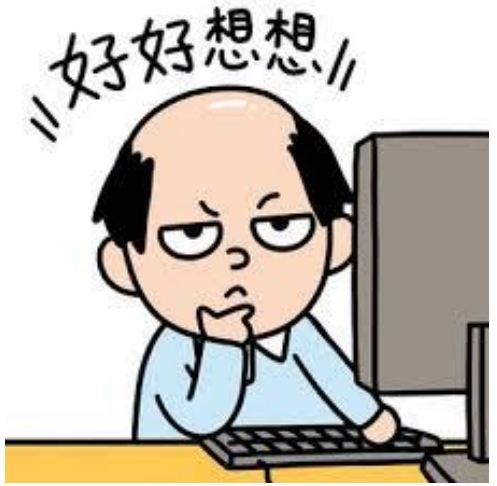


昇腾手撕 MOE 代码

ZOMI

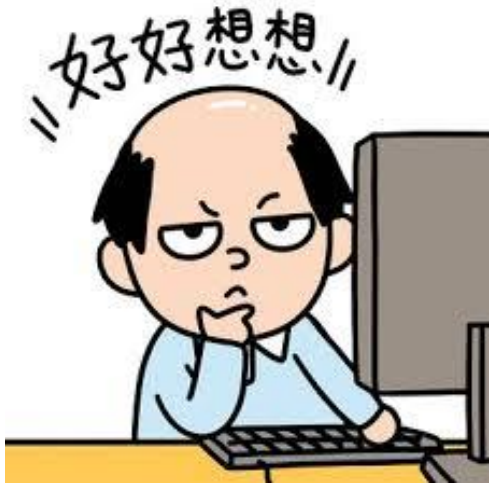
视频目录大纲

1. 单卡版 MoE
2. 单机八卡版 MoE
3. Profiling MoE 看计算 & 通信耗时



视频目录大纲

- MoE 混合专家
  - MoE 基本介绍 4
  - MoE 前世今生 5
  - MoE 核心工作 6
  - MoE 架构原理 3
  - 大模型遇上 MoE 4
  - 手撕 MoE 代码
    - 简单版从零实现一个 MOE
    - DeepSpeed 中 MoE 模型（Gshard 架构）
    - MoE 并行（专家并行）
    - 专家并行通信加速（All2All）
    - Megatron MoE 代码解读
  - MoE 的未来 2



01

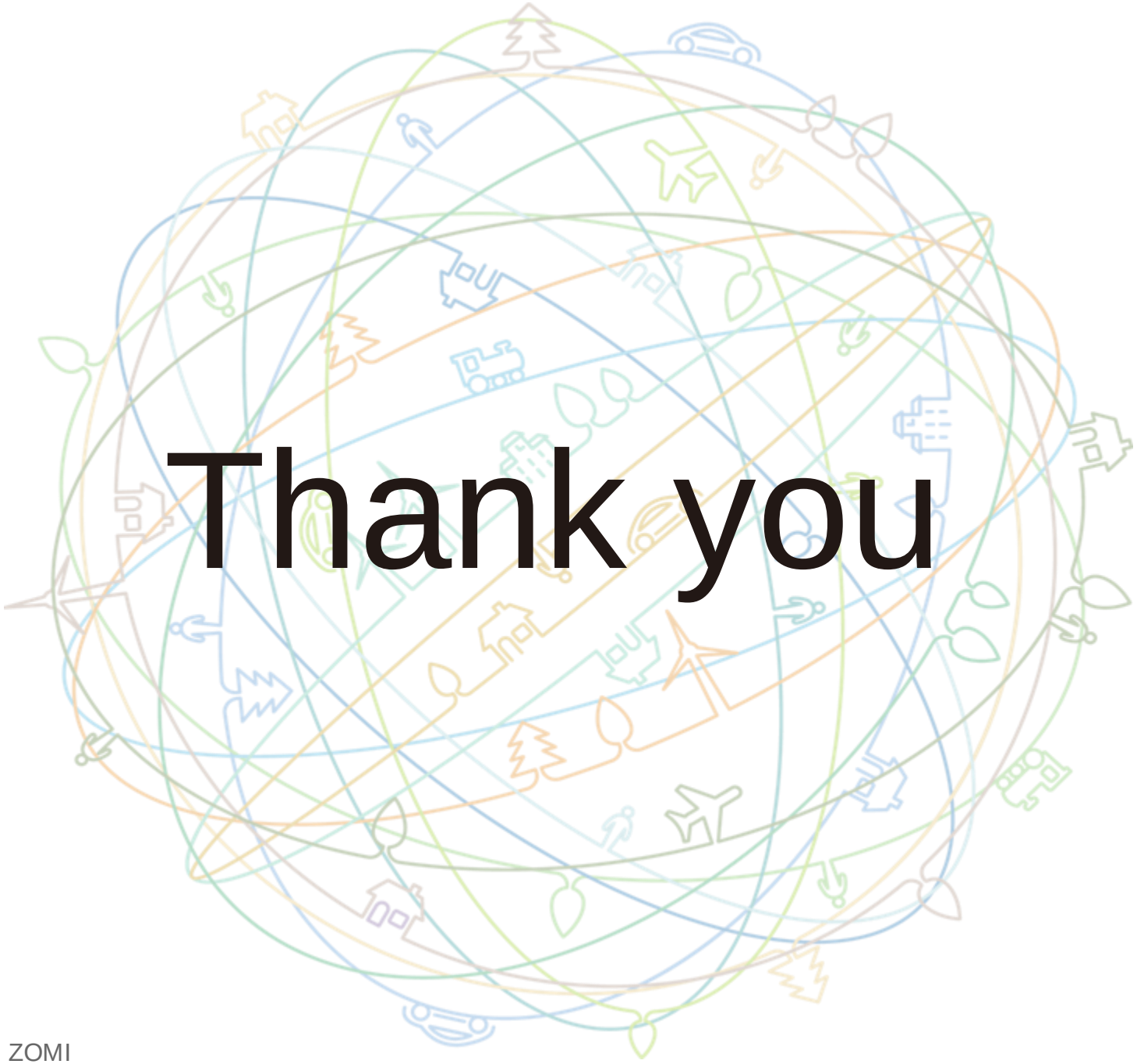
单卡版 MoE 实现

02

单机八卡版 MoE

03

Profiling MoE



Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

GitHub https://github.com/chenzomi12/AIFoundation

引用与参考

- https://mp.weixin.qq.com/s/6kzCMsJuavkZPG0YCKgeig
- https://www.zhihu.com/tardis/zm/art/677638939?source_id=1003
- https://huggingface.co/blog/zh/moe
- https://mp.weixin.qq.com/s/mOrAYo3qEACjSwcRPG7fWw
- https://mp.weixin.qq.com/s/x39hqf8xn1cUlnxEIM0_ww
- https://mp.weixin.qq.com/s/ZXjwnO103e-wXJGmmKi-Pw
- https://mp.weixin.qq.com/s/8Y281VYaLu5jHoAvQVvVJg
- https://blog.csdn.net/weixin_43013480/article/details/139301000
- https://developer.nvidia.com/zh-cn/blog/applying-mixture-of-experts-in-llm-architectures/
- https://www.zair.top/post/mixture-of-experts/
- https://my.oschina.net/IDP/blog/16513157

- PPT 开源在：
- https://github.com/chenzomi12/AIInfra