

Mixture of Experts (MoE)



昇腾手撕
MOE 代码



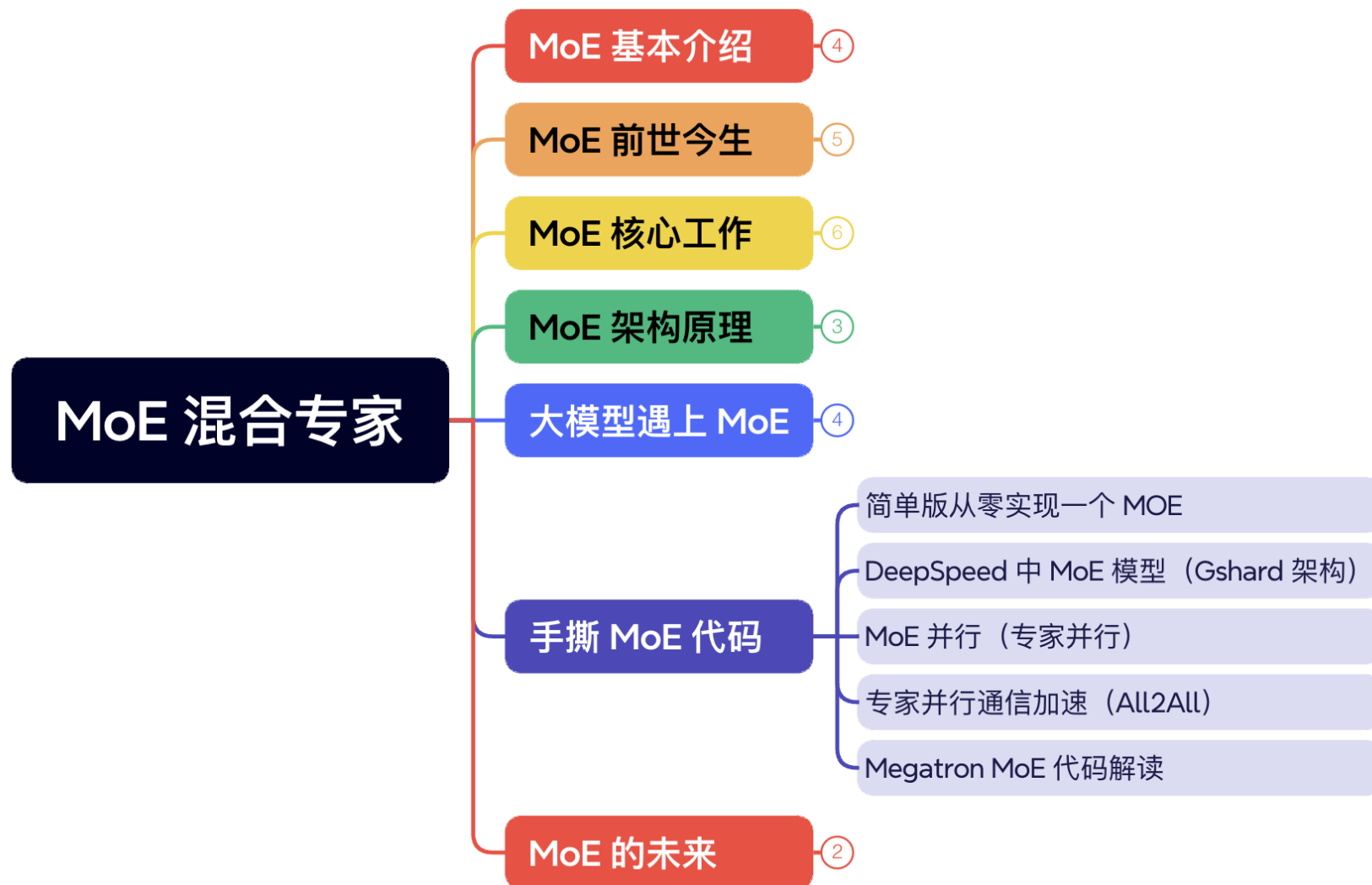
ZOMI

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



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.

Copyright © 2024 XXX Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.



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/AllInfra>

