

十万卡AI集群思考



ZOMI



思考

- **构建十万卡集群是一项复杂的系统工程：**
 - 不仅意味着算力指数级增长，还涉及复杂的技术和运营挑战
 - 需要解决高效能计算、高能耗管理、高密度机房设计、高稳定性训练等一系列问题
 - 最终能否将算力有效释放，还取决于算法、软件架构的优化与调度能力



目录

- 十万集群都在训什么?
- 功耗挑战?
- 互联挑战?
- 可靠挑战?



01

十万集群都 训练什么



十万集群目标

- **谁在建?**
 - 多个大型 AI 实验室 (OpenAI/微软、xAI、Meta) 竞相建立 >10 万卡 AI 集群。
- **有什么用?**
 - AI 下一步利用海量多模态数据, 训练超大规模参数大模型。
- 虽然目前还没有团队完成, 但该领域竞争已经十分激烈。



算力对比

- 通过对标 GPT-4 训练来了解 10 万卡 AI 集群能提供多少算力：
 - OpenAI ~2 万块 A100 上对 GPT-4 持续训练 90-100 天 (FP16 2.15×10^{25} FLOPs) , 集群峰值吞吐为 6.28 BF16 ExaFLOP/s;
 - 如果使用 10 万卡 H100 集群, 峰值吞吐可以飙升至 198/FP8 or 99/FP16 ExaFLOP/s。与 2 万卡 A100 集群相比, AI 集群训练算力可提升 31.5 倍;



GPT4 遥遥领先

- GPT4 在 2023.6 发布，OpenAI GPT-4 完成一次训练约三个月时间，使用 25k NVIDIA A100 GPU。如今已经十万卡 H100 集群，为什么到现在为止还没有出现超越 GPT4 的模型？

- 大模型训练不是堆算力就能解决
- 但是不堆算力绝对解决不了



单芯片能力

- 研发单 GPU 更多并行处理核心,努力提高运行频率。
- 通过优化高速缓存, 减少 GPU 访存延迟, 于是越来越多近存计算的新架构出现
- 优化浮点数表示格式, 探索从 FP16 到 FP8/HiF8 浮点数表示格式, 芯片引入新计算精度



超节点计算能力

- 提高万卡集群 NPU 卡间互联的网络利用率和模型利用率 (MFU) , 实现通信时延减少, 带宽能力跃升, 支持更高频次、更大带宽和更低延迟通信特性:
 1. NPU 节点内集成类 Scale up 能力 Switch 芯片, 优化卡间南向互联效率和规模, 增强 TP/PP/SP 等数据传输能力
 2. 引入节点内 Switch 芯片, 增强卡间 P2P 带宽, 有效提升节点内网络传输效率, 满足大模型互联和带宽的增长需求
 3. 对 NPU 卡间互联协议进行系统性优化和重构, 以提升 All2All 下通信效率, 如华为的灵渠总线架构协议、AMD、Google 组成的 UALink
 4. 重新设计 NPU 卡间数据报文格式、引入 CPO/NPO、提高和优化 SerDes 传输速率、优化网络拥塞控制/重传机制



网络计算能力

- DPU 上实现存储后端接口，基于 RDMA 网络功能连接块存储集群、对象存储集群、文件存储集群及文件存储集群
- 降低多机多卡间端到端通信时延，提升多机间端到端通信带宽，构建节点间数据交换的高速通道
- AI 计算场景流量特征是流数少、单流带宽大，端口级负载均衡技术或算网协同负载均衡技术能提供更好的吞吐



智能管控

- 统一对计算、存储、网络、光模块等进行资源、性能、告警、日志、拓扑等信息采集
- 提供资源管理、服务编排、监控、作业运维等功能，实现 AI 集群智能运维
- 环境健康检查、作业故障诊断、集群环境/资源管理、服务器管理以及监控分析等云化能力



02

功耗挑战



耗电

- 十万卡 AI 集群资本支出 >40 亿刀， AI 集群通过集中部署来利用芯片与芯片间高速网络能力，因此受到 IDC 功耗和电力不足严重限制。
- 一个 10 万卡 AI 集群需要 >150MW 功耗容量，并且每年耗费 1.59 TWh 电量，按 0.078 \$/kwh 标准费率来计算，每年电费高达 1.239 亿刀。

100k H100 GPU Cluster - Annual Electricity Costs		
Region	Tariff (USD/kWh)	Annual Cost \$M
USA - Average	\$0.083	\$131.9
USA - North Dakota	\$0.074	\$117.6
USA - Utah	\$0.068	\$108.0
USA - Arizona	\$0.078	\$123.9
USA - ND Wind PPA	\$0.033	\$52.4
USA - Solar PPA (CAISO)	\$0.033	\$52.4



还是耗电

- 单卡 GPU 能耗为 700W，每个 H100 节点，CPU、网络接口卡（NIC）和电源装置（PSU）额外增加每 GPU 400~500W
- 节点外，AI 集群还需存储服务器、网络交换机、CPU 节点、光模块和其他设备，总能耗约占 IT 能耗 10%
- 建设 10 万卡 AI 集群时，通常指一片园区内部署，而不是单栋建筑内部署。xAI 将田纳西州孟菲斯旧工厂改造成智算中心



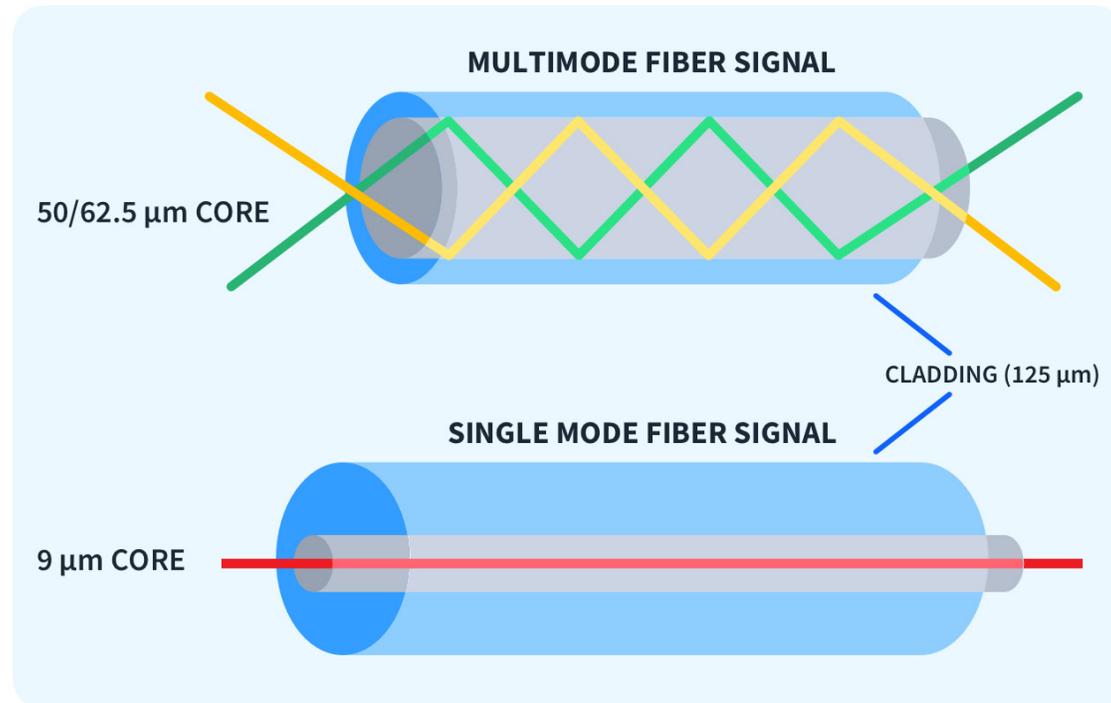
03

网络互联挑战



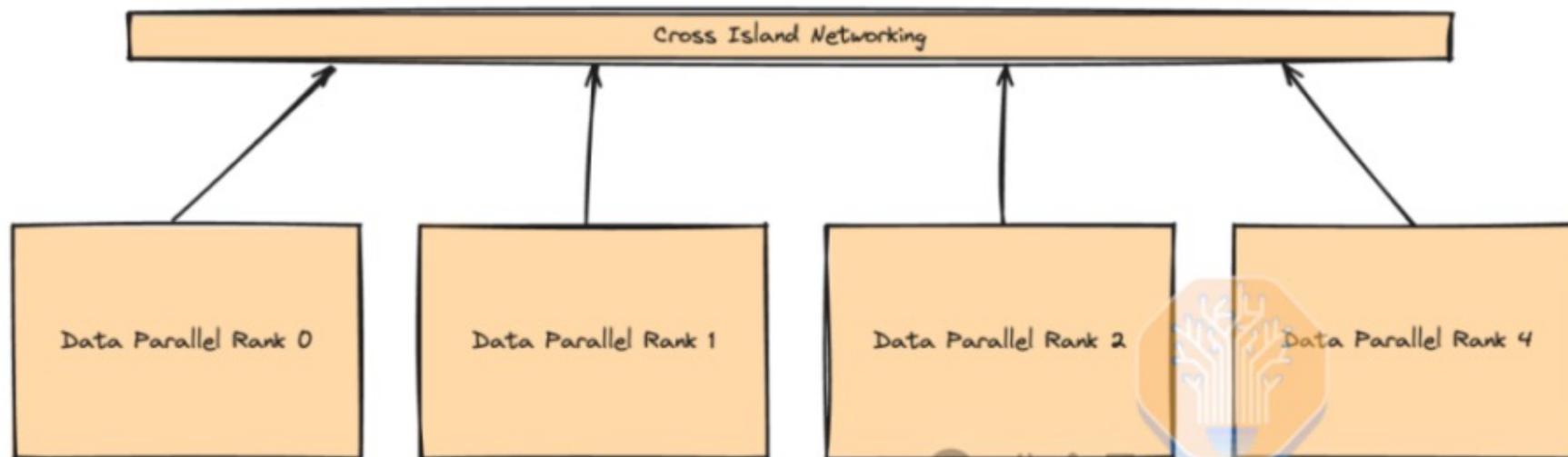
光模块传输距离

- AI 集群通过光模块互联，而光模块成本与覆盖范围成正比：
 - 多模 SR 和 AOC 光模块支持 ~50m 传输距离，长距离单模 DR 和 FR 光模块支持 500m ~ 2km 信号
 - DR 是 SR 多模光模块 ~2.5 倍成本，园区级相干光模块传输距离能 >2 千米，但价格 >10 倍以上



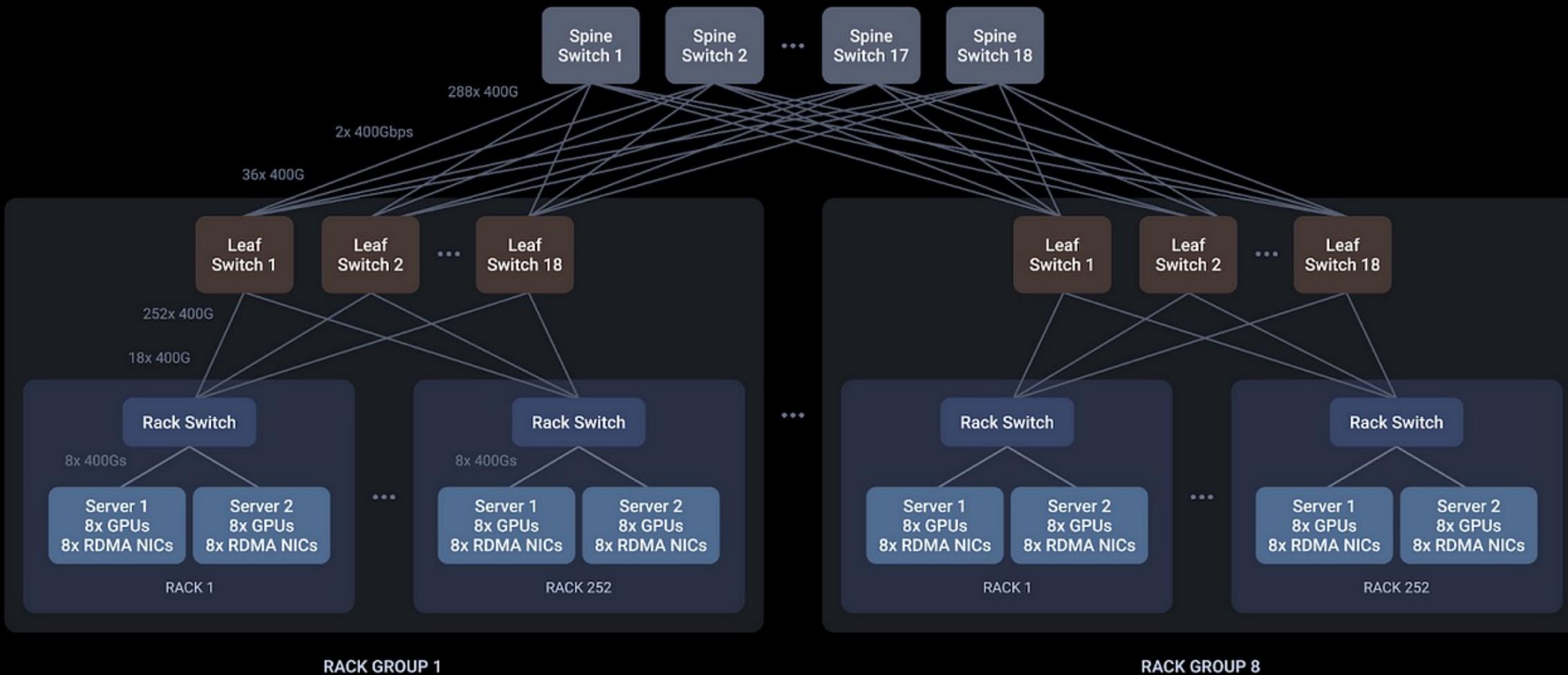
岛内和岛外互联的选择

- H100 小集群通常只使用多模光模块，通过一到两层交换机网络实现 400G 带宽卡间互联
- 一般每栋楼包含一个或多个计算节点，它们之间用铜缆和多模光模块进行连接
- 较长距离光模块来实现岛（island）际互连，计算岛内具备高速带宽，计算岛外带宽较低



公众号: DETACHED UNCLE
semicondysis





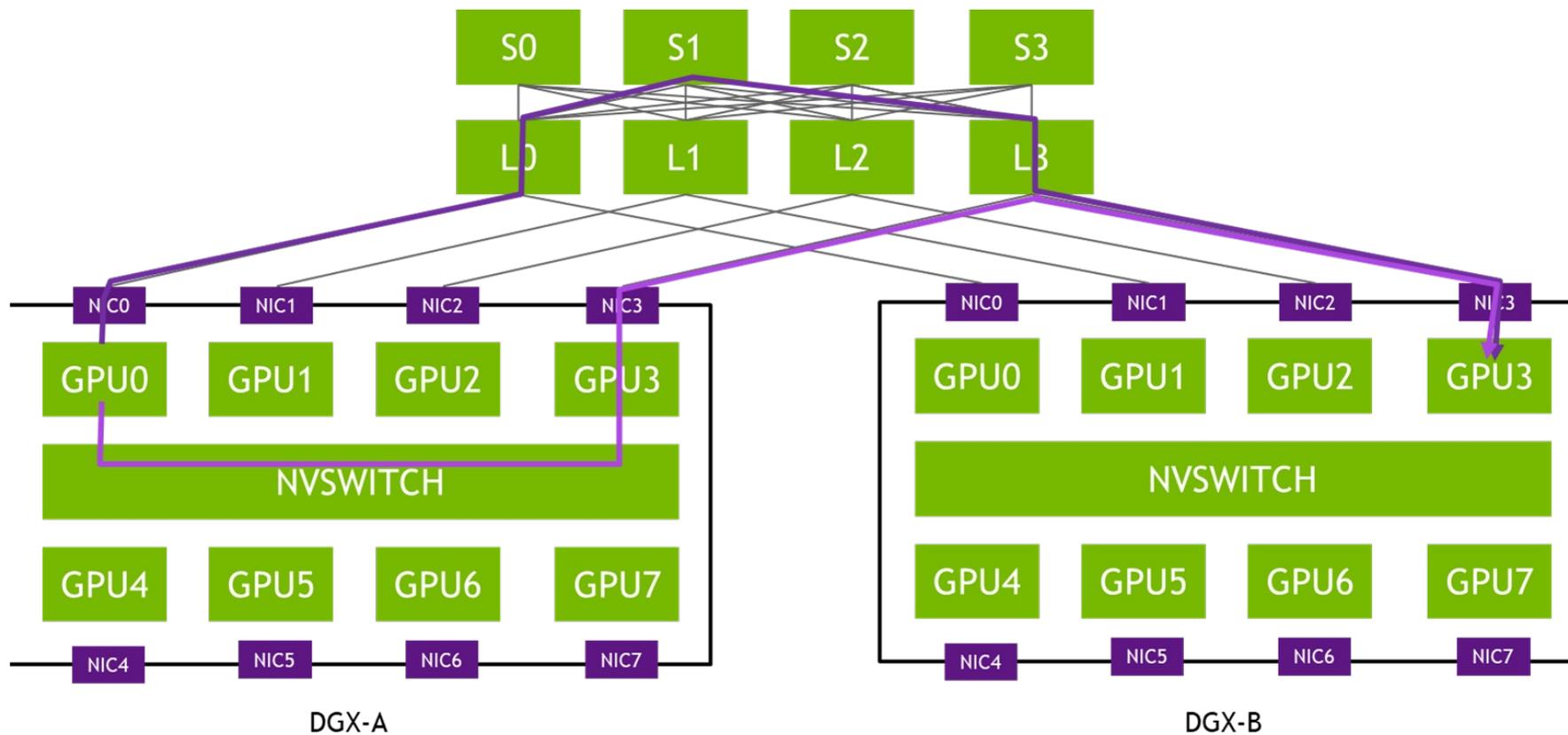
16*252 = 4032 GPUs/RACK GROUP * 8 RACK GROUPs
= 32256 GPU Clusters

- Meta 3.2 万卡 GPU 集群中，设计 8 个全速带宽计算岛，计算岛之上一层交换机网络带宽过载比为 7:1，岛际互连网络速度比岛内网速慢 7 倍。



轨道优化 (Rail Optimized) & 中间机架 (Middle of Rack)

- 为提高可维护性并增加铜缆网络 (连接距离 <3m) 和多模网络 (连接距离 <50m), 部分不采用轨道优化设计, 转而采用中间机架设计



Question

- 中间机架 (Middle of Rack) 是什么意思?
- 轨道优化 (Rail Optimized) 又是什么意思?



Question

- 轨道优化 (Rail Optimized) 又是什么意思?

<https://space.bilibili.com/517221395/channel/collectiondetail?sid=3666664>

我的合集和视频列表 > 合集·【大模型】集合通信库

▶ 播放全部

合集 | 9个视频 | 9-10更新

MPI 是集合通信库的鼻祖，英伟达 NVIDIA 大量的参考和借鉴 MPI 通信库相关的内容从而提出了业界集合通信库的标杆 NCCL。本将会从 MPI 开始，介绍业界的各种主流集合通信库的变种 XCCL。然后深入地剖析 NCCL 相关的实现算法、对外 API 等，最后还...

默认排序

升序排序

编辑

+
去创作中心添加视频

1
大模型系列-NCCL/HGCL_01
XCCL的基础 MPI通信
13:29

NCCL/HGCL 的基础 MPI 通信介绍! #大模型 #集合通信 #MPI
▶ 4215 8-16

2
大模型系列-NCCL/HGCL_02
业界XCCL集合通信库介绍 1
11:11

业界集合通信库XCCL大串烧(基本介绍)上篇 #大模型 #集合通信
▶ 2819 8-17

3
大模型系列-NCCL/HGCL_03
业界XCCL集合通信库介绍 2
10:44

业界集合通信库XCCL大串烧(基本介绍)下篇 #大模型 #集合通信
▶ 2409 8-19

4
大模型系列-NCCL/HGCL_04
英伟达 NCCL 通信库剖析
19:01

了解英伟达NCCL通信库一个视频就够 #大模型 #集合通信 #NCCL
▶ 5704 8-20

5
大模型系列-NCCL/HGCL_05
英伟达 NCCL API 解读
14:37

英伟达NCCL通信库到底怎么用! #大模型 #集合通信 #NCCL
▶ 2895 9-5

6
大模型系列-NCCL/HGCL_06
通信算法&网络拓扑协同优化
11:58

终于搞清楚通信算法与网络拓扑啥关系了! #大模型 #集合通信
▶ 2521 9-6

7
大模型系列-NCCL/HGCL_07
NCCL核心算法双二叉树
16:13

万卡集群通信优化算法双二叉树! #大模型 #集合通信 #NCCL
▶ 2515 9-9

8
大模型系列-NCCL/HGCL_08
华为 HCCL 集合通信库解读!
21:17

华为HCCL 集合通信库开源(放)啦! 难得呀! #大模型 #集合通信
▶ 2564 9-10

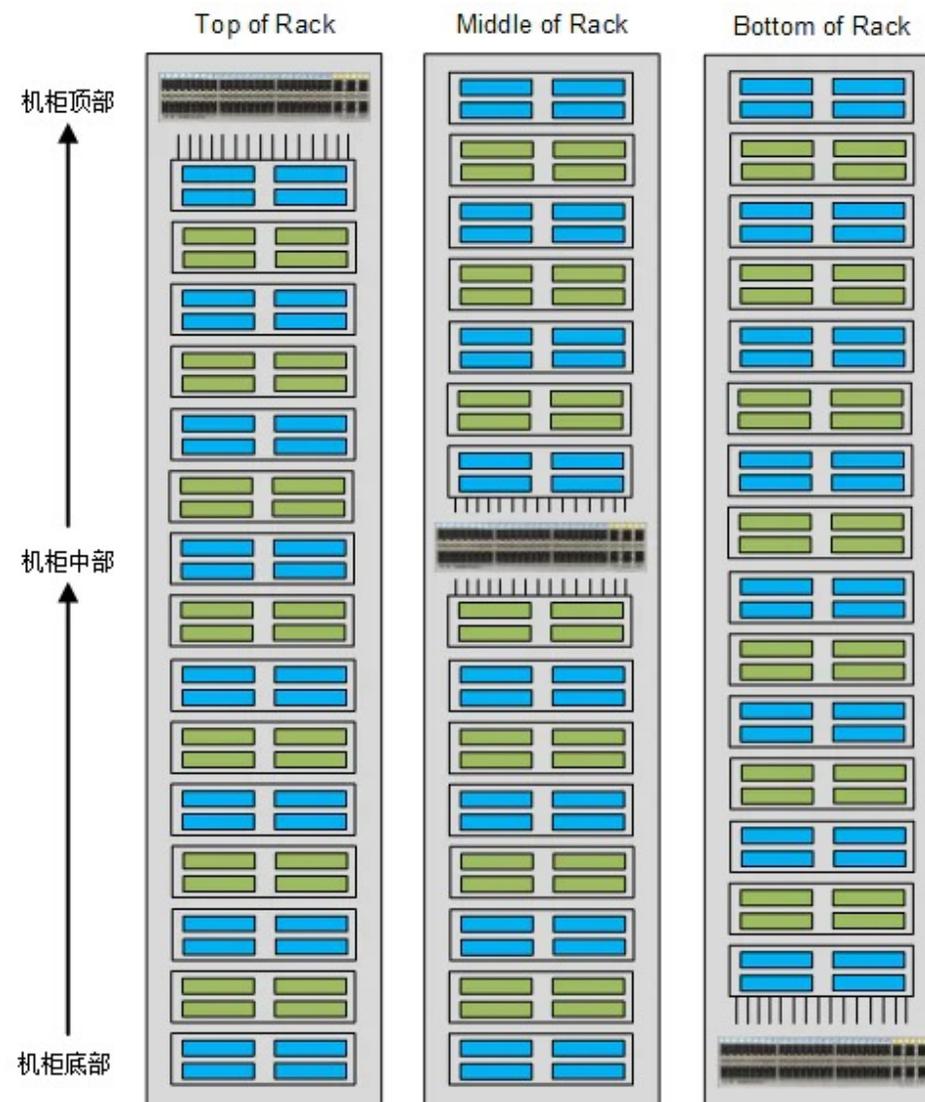
9
大模型系列-NCCL/HGCL_10
对集合通信影响因素进行建模
11:50

XCCL网络建模了解对大模型通信的影响 #大模型 #集合通信 #NCCL
▶ 1931 9-7



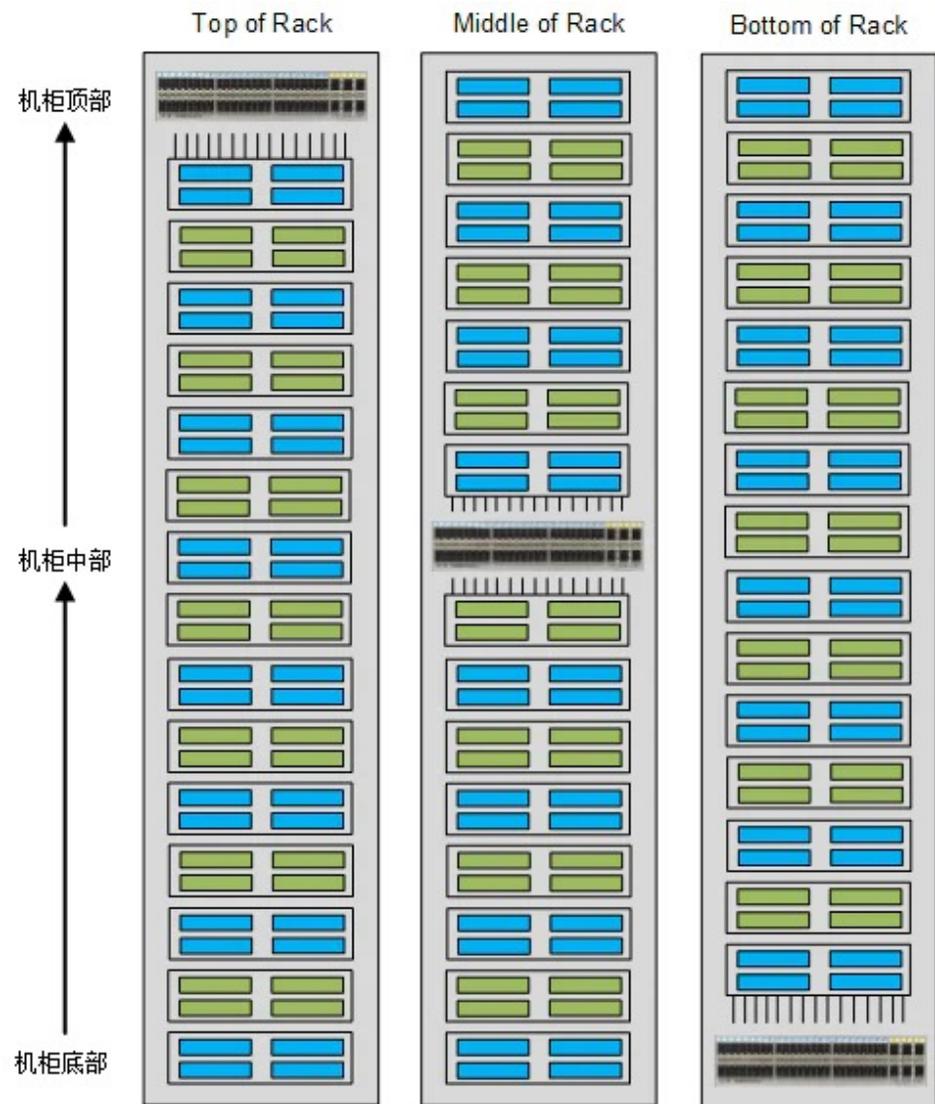
Middle of Rack

- 交换机在 IDC 机柜上部署位置，分为机柜顶部（Top of Rack）、机柜中部（Middle of Rack）& 机柜底部（Bottom of Rack）。
- 多轨优化让每台 H100 节点连接到 8 个不同叶交换机，每 GPU 只需 1 个交换跃点（switch hop）就能与另外 GPU 通信。



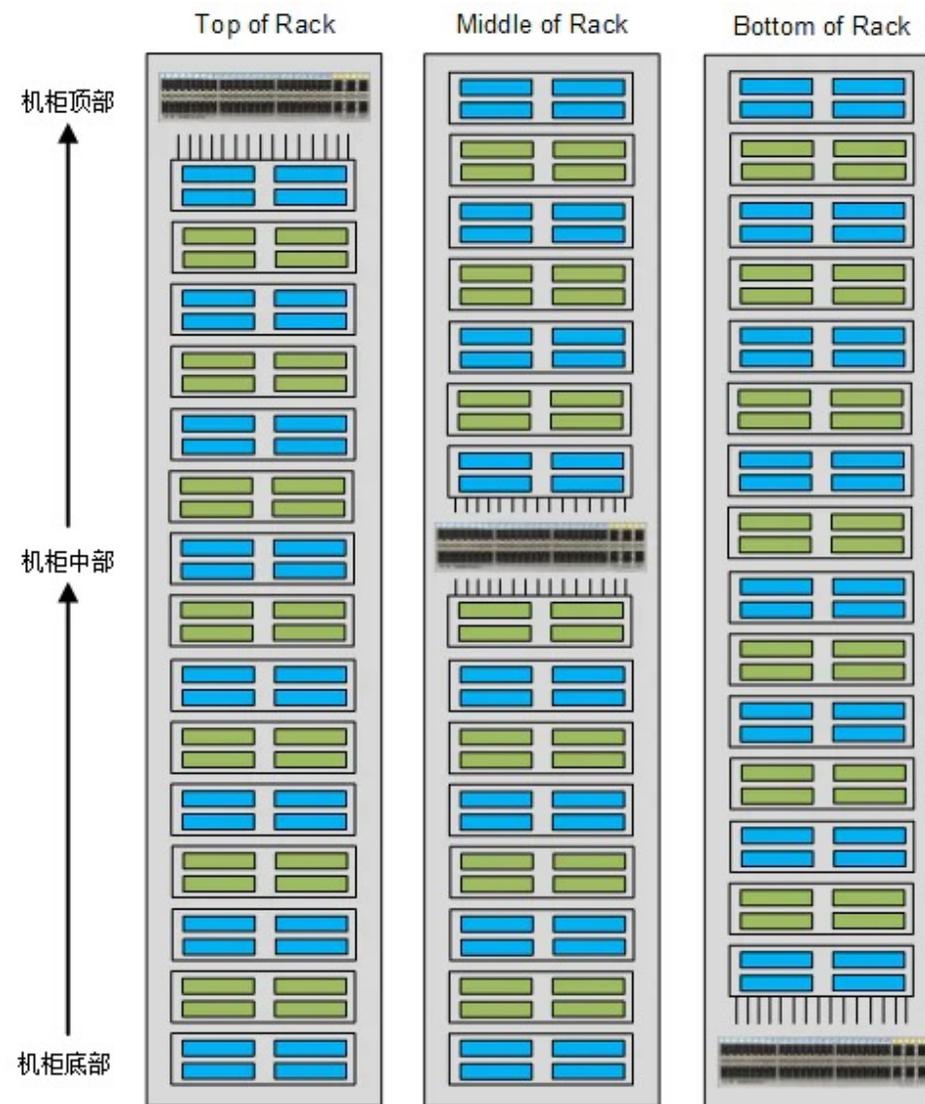
Middle of Rack

- 缺点是节点中 8 块 GPU 必须连接到不同距离叶交换机，而不是更靠近服务中间机架上。
- **单轨方案**：当交换机位于同一机架时，可以使用无源直连电缆（DAC）和有源电缆（AEC）
- **多轨方案**：交换机不一定位于同一机架，必须使用光模块。此外，叶交换机到脊交换机的距离可能 > 50m，必须使用单模光模块。



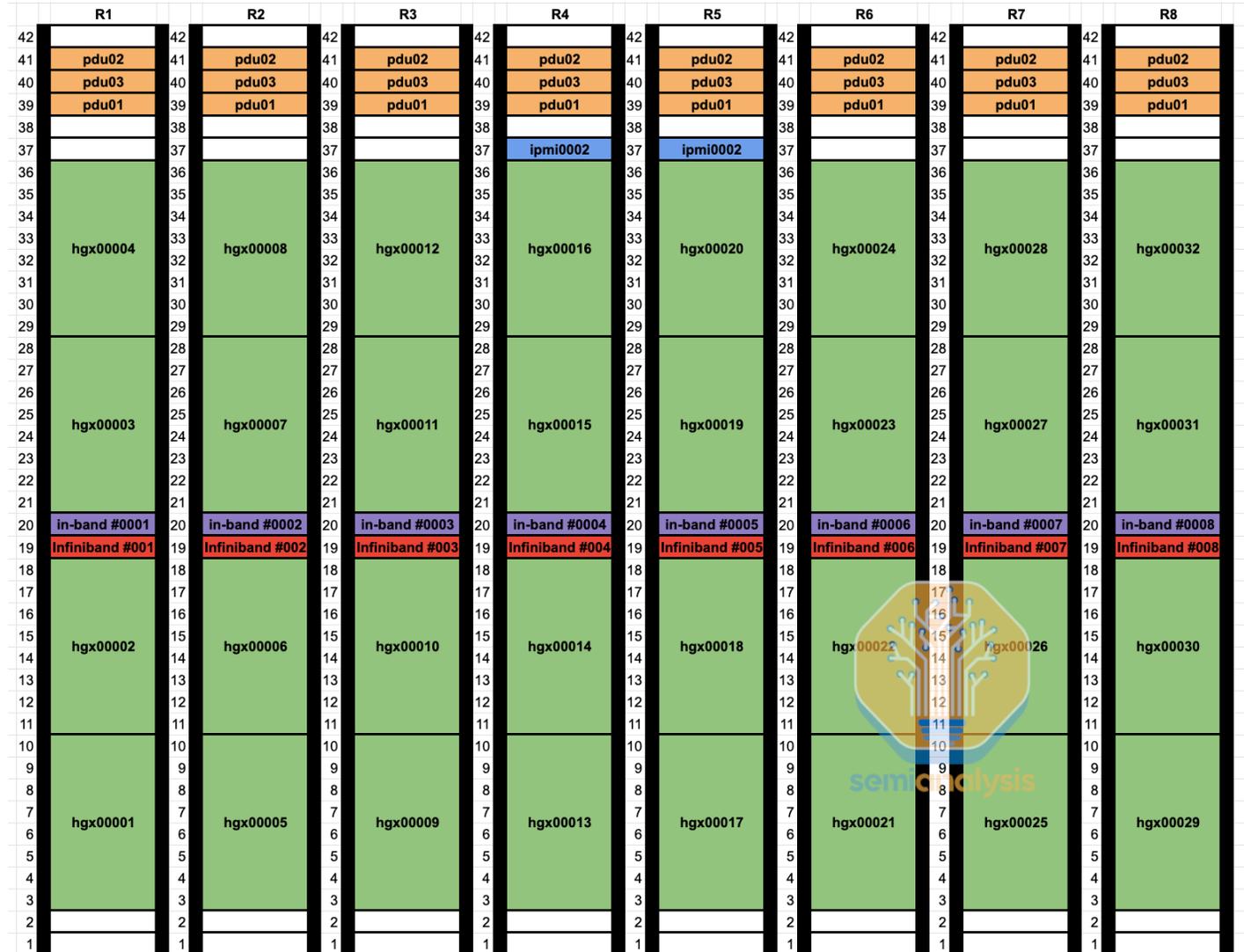
Middle of Rack

- 使用无轨优化的中间机架设计，可以用更廉价的铜缆取代连接 GPU 和叶交换机的光模块，在 GPU 网络中实现 25-33% 铜缆覆盖。



Middle of Rack

- 每个 GPU 与叶交换机连接不再是先连接到电缆托架、再从侧面穿过 9 个机架连接到专用的导轨优化叶交换机上；
- 将叶交换机放在机架中间，从而让每个 GPU 都能使用 DAC 铜缆；
- DAC 电缆运行温度更低、能耗更低、可靠性更高，实现更少 flapping（网络链路间歇性瘫痪）和故障；



对比

- 一个Quantum-2 IB Spine 交换机，使用 DAC 铜缆耗电 ~747W，使用多模光模块时，能耗会增加到 1,500W

	R1	R2	R3	R4	R5	R6	R7	R8	R9
42									
41	pdu02								
40	pdu03								
39	pdu01								
38									
37				ipmi0002	ipmi0002				
36									
35									
34									
33	hgx00004	hgx00008	hgx00012	hgx00016	hgx00020	hgx00024	hgx00028	hgx00032	
32									
31									Infiniband #008
30									
29									
28									Infiniband #007
27									
26									
25	hgx00003	hgx00007	hgx00011	hgx00015	hgx00019	hgx00023	hgx00027	hgx00031	Infiniband #006
24									
23									
22									Infiniband #005
21									
20	in-band #0001	in-band #0002	in-band #0003	in-band #0004	in-band #0005	in-band #0006	in-band #0007	in-band #0008	
19									Infiniband #004
18									
17									Infiniband #003
16									
15	hgx00002	hgx00006	hgx00010	hgx00014	hgx00018	hgx00022	hgx00026	hgx00030	
14									Infiniband #002
13									
12									
11									Infiniband #001
10									
9									
8									
7	hgx00001	hgx00005	hgx00009	hgx00013	hgx00017	hgx00021	hgx00025	hgx00029	
6									
5									
4									
3									
2									
1									

	R1	R2	R3	R4	R5	R6	R7	R8
42								
41	pdu02							
40	pdu03							
39	pdu01							
38								
37				ipmi0002	ipmi0002			
36								
35								
34								
33	hgx00004	hgx00008	hgx00012	hgx00016	hgx00020	hgx00024	hgx00028	hgx00032
32								
31								
30								
29								
28								
27								
26								
25	hgx00003	hgx00007	hgx00011	hgx00015	hgx00019	hgx00023	hgx00027	hgx00031
24								
23								
22								
21								
20	in-band #0001	in-band #0002	in-band #0003	in-band #0004	in-band #0005	in-band #0006	in-band #0007	in-band #0008
19	Infiniband #001	Infiniband #002	Infiniband #003	Infiniband #004	Infiniband #005	Infiniband #006	Infiniband #007	Infiniband #008
18								
17								
16								
15	hgx00002	hgx00006	hgx00010	hgx00014	hgx00018	hgx00022	hgx00026	hgx00030
14								
13								
12								
11								
10								
9								
8								
7	hgx00001	hgx00005	hgx00009	hgx00013	hgx00017	hgx00021	hgx00025	hgx00029
6								
5								
4								
3								
2								
1								



03

故障恢复与运维



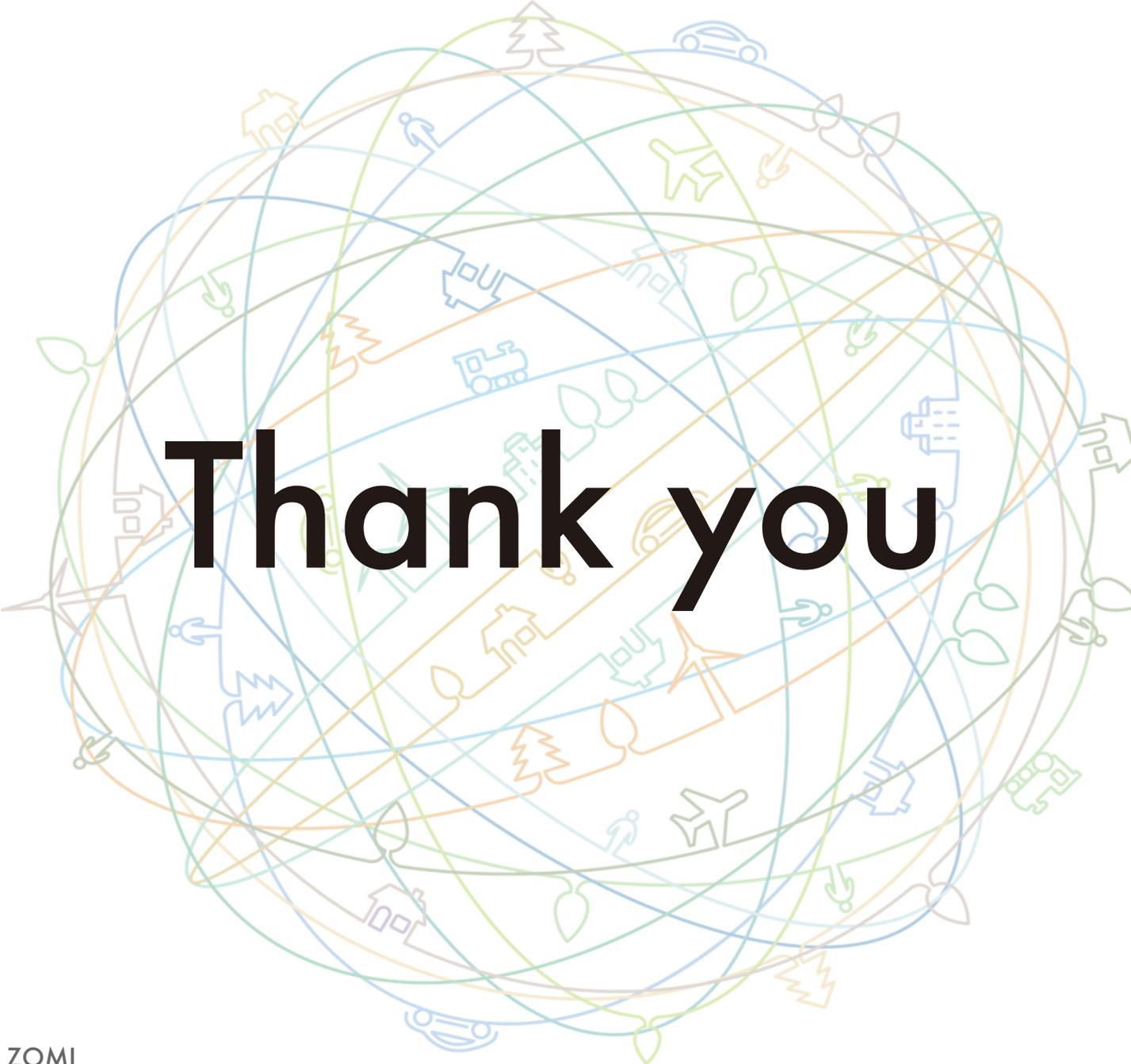
可靠性问题

- GPU HBM ECC 错误、GPU 驱动器卡死、光模块故障、网卡过热以及计算节点宕机或出错等
- 训练时为防止 HBM ECC 等错误，需要对 CPU 内存或 NAND SSD 进行 Checkpointing 检查

Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency	Dependency	2	0.5%
CPU	Host	2	0.5%
System Memory	Host	2	0.5%

Table 5 Root-cause categorization of unexpected interruptions during a 54-day period of Llama 3 405B pre-training. About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.





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>

Reference 引用

1. <https://www.youtube.com/embed/Jf8EPSBZU7Y>
2. <https://www.youtube.com/watch?v=Jf8EPSBZU7Y&t=1s>

