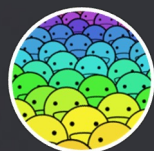


大模型系列 – 集合通信库

通信模型&影响因素



ZOMI

时事热点

智能体

RAG

具身智能

智能驾驶

工业大模型

...

大模型训推

6. 大模型数据&算法

数据&模型评估

Prompt 工程, 模型评估算法和测评体系

大模型算法

Scaling Law、Transform 结构, LLM/MLM 模型

7. 大模型训练

分布式训练

TP/DP/PP/SP/EP 并行, Megatron、DeepSpeed 分布式并行库介绍

微调

全参微调、底参微调(LoRA/QLoRA 等)、指令微调

8. 大模型推理

推理框架

VLLM、推理框架的架构, 推理框架线程池等构架

推理优化

大模型推理加速(XXXAttention)、长序列推理优化算法

编译计算架构

4. 计算架构

传统编译器

传统编译器 GCC与LLVM

AI 编译器

AI编译器发展与架构定义, 未来挑战与思考

前端优化

前端优化(算子融合、内存优化等)

后端优化

后端优化(Kernel优化、Auto Tuning)

多面体

复杂的循环依赖关系映射到高维几何空间

5. 通信架构

集合通信

通信原语、通信原理、集合通信算法

NCCL/HCCl

集合通信库、网络拓扑、通信方式、通信算法, NCCL 架构

硬件体系结构

3. AI 集群

集群管理运维

K8s集群运维、K8s容器、集群监控等工具

集群性能指标

稳定性、吞吐、线性度等

集群训推一体化

训练、推理大模型执行, 训练推理显存分析

机房建设

风火水电、夜冷、柜板等知识

1. AI 芯片

AI 计算体系

AI 计算模式与计算体系架构

AI 芯片基础

CPU、GPU、NPU等芯片体基础原理

英伟达GPU

英伟达GPU TensorCore、NVLink剖析

国外AI芯片

谷歌、特斯拉等专用AI处理器核心原理

国内AI芯片

寒武纪、燧原科技等专用AI处理器原理

2. 通信与存储

通信

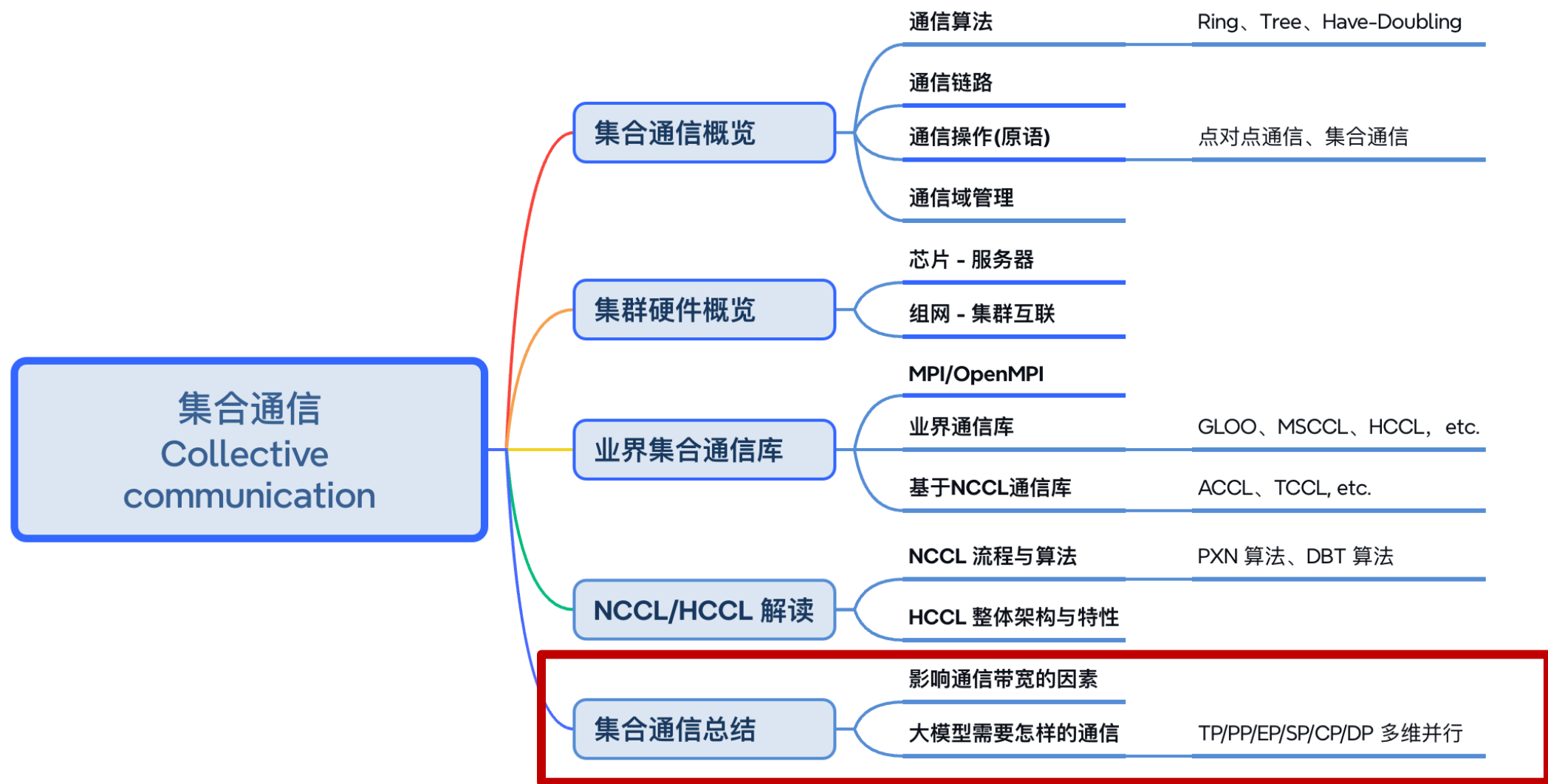
路由器、交换机基本原理和网络拓扑

存储

DRAM、SRAM、存储 POD 到大模型存储 CKPT 算法



思维导图 XMind



Question

1. 对集群通信成本进行建模是一个重要又具有挑战性的工作，了解通信性能模型的发展、作用，以及现有的各类集合通信模型分别做什么、为何这样设计、有哪些进一步的贡献、优缺点等等

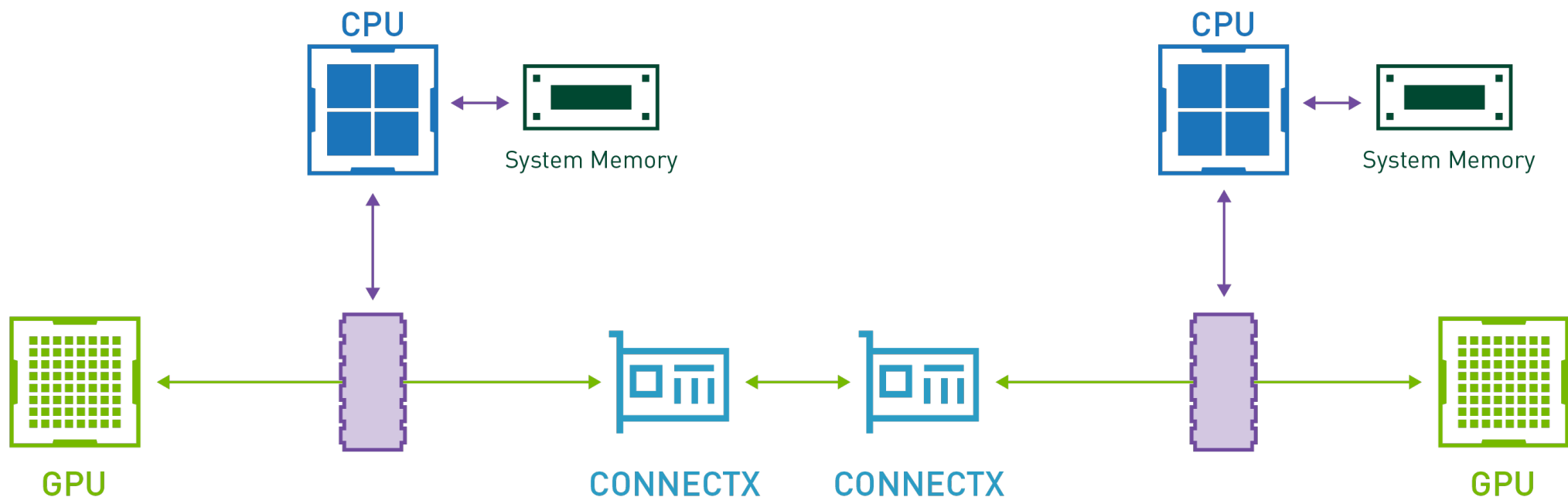


01. 基础通信模型 的作用



基础通信模型

- 集群的计算节点越来越复杂，以同一节点上的进程间通信为例，它们可以使用共享内存作为缓冲区通信（GPU Direct），也可以通过高性能网络（如 InfiniBand/RoCE v2）使用操作系统模块、操作系统旁路或远程内存访问（RDMA）进行消息的直接传输。



概念与形式

- 通信的性能模型使用一系列参数对一次点对点通信的时间消耗进行建模，因为点对点通信（P2P）的成本可以推广到集合通信当中；
- 集合通信在大模型训练和推理过程中，占用了相当高的耗时，因此许多研究都致力于提高 XCC L 的集合通信性能
- **通信性能模型主要能从两点提供帮助：**
 1. 为集合通信算法设计提供参考
 2. 具体分析某一通信模式的性能



通信性能模型 A

- 为集合通信算法设计提供参考：
 - 主要是理论层面分析，分析集合通信算法底层的点对点通信顺序与步数，表示总体通信时间成本，从而评价通信算法的优劣，反过来指导通信优化算法的设计。



通信性能模型 B

- 具体分析某一通信模式的性能：
 - 主要是实践层面，通过建立通信性能模型来估计集合通信优化算法的性能，能够基于当前网络拓扑和优化算法计算出一个具体通信耗时，更进一步让 XCCL 库可以自适应地选择具体的通信优化算法，从而优化集合通信操作效率。



02. 基础通信模型



Postal 模型

- Postal 模型是出现最早通信分析模型，由 1992 年 Bar-Noy 和 Kipnis 提出，希望通过引入网络延迟对全连接网络的通信进行建模分析。
- Postal 模型中，点对点通信的时间 $T_{p2p}(m) = \lambda$ ，Postal 模型假设同一时间一个设备（CPU）仅能发送或接收一个消息。

- Amotz Bar-Noy and Shlomo Kipnis. 1992. Designing broadcasting algorithms in the postal model for message-passing systems. In Proceedings of the 4th Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA'92). ACM, New York, NY, 13–22.



Hockney 模型

- Hockney 模型也被称为 $\alpha - \beta$ 模型, Postal 模型基础上引入了带宽, 并限制每次传输数据量;
- Hockney 模型一次点对点通信成本 $T_{p2p}(n) = \alpha + n * \beta$ 。Hockney 模型假设一个设备节点 (CPU) 可以同时执行一个发送操作和一个接收操作。
- 其中,
 - α : 节点间的固定时延
 - β : 每 byte 数据传输耗时
- Roger W. Hockney. 1994. The communication challenge for MPP: Intel paragon and Meiko CS-2. Parallel Computing 20, 3 (March 1994), 389–398.



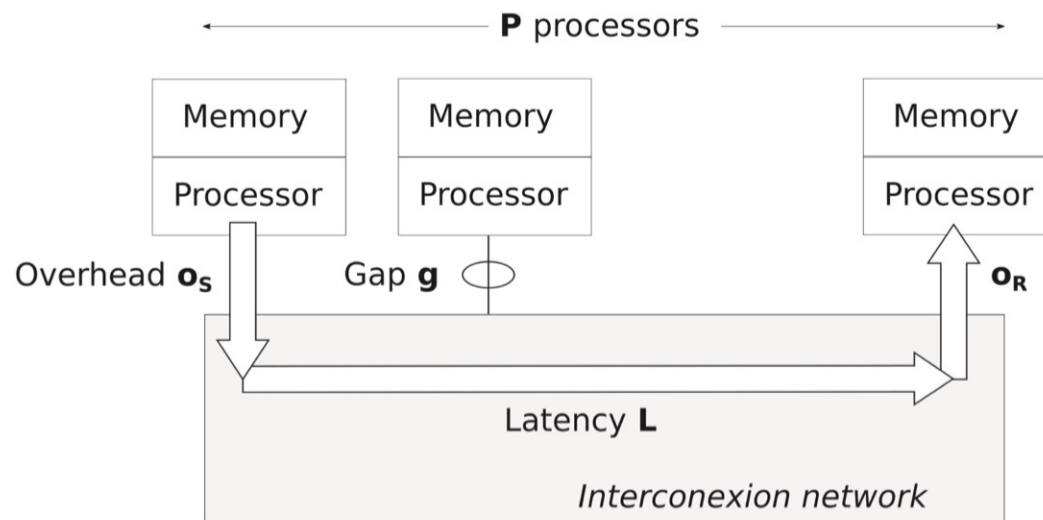
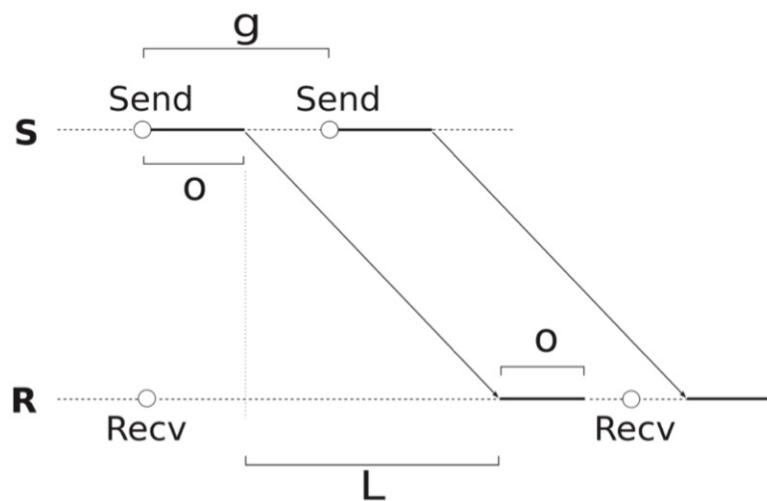
LogP 模型

- LogP 模型名称中的四个字母代表了模型中四项参数：
 - **L (Latency)** : 网络延迟, 发送方发送完最后一个字节到接收方接收完最后一个字节的时间。
 - **o (Overhead)** : 准备开销, 处理器为发送/接收消息花费时间, 包括准备消息、发送队列排队、向 NIC 发送信号等。
 - **g (Gap per message)** : 间隔。NIC 向链路注入相邻两个数据包之间所需要的最小时间, 其倒数反应了网络带宽, 处理器在网络中被阻塞之前能够发送 $\lfloor L/g \rfloor$ 个消息。
 - **P (Process)** : 进程总数。



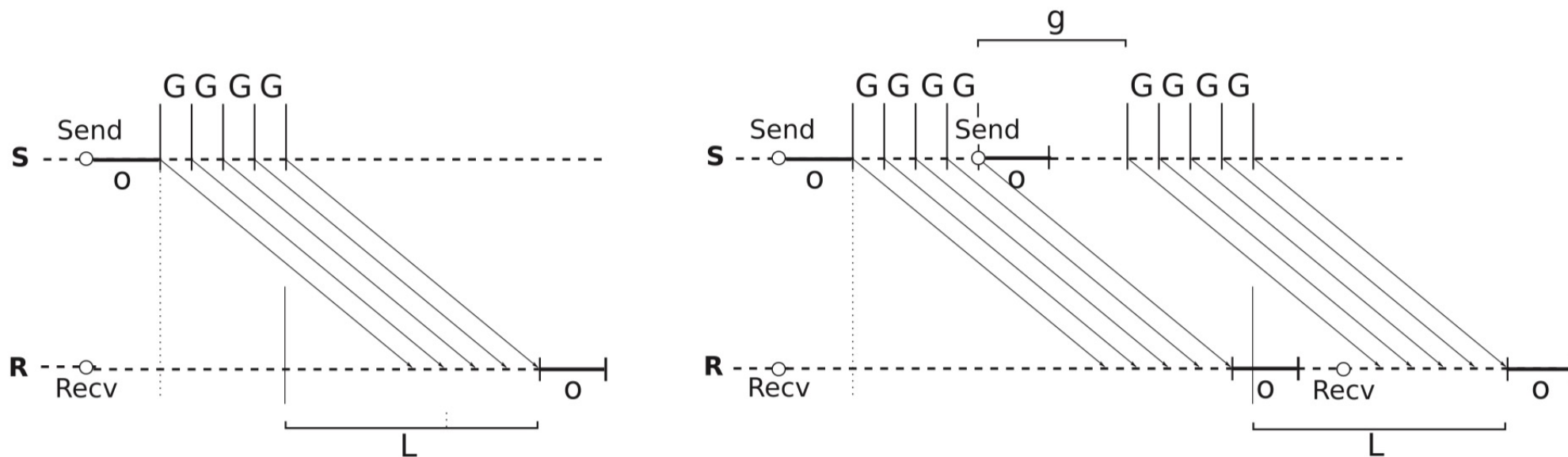
LogP 模型

- LogP 模型发送一个短消息时间成本为 $T_{p2p}(m) = o + L + o$, Culler 等人提出进一步区分发送方和接收方的开销, 即 o_s 与 o_r 。
- LogP 模型主要进展是考虑到处理器对通信延迟影响, g 存在意味着处理器能够做一些非通信的开销工作, 因此可以存在计算与通信重叠。



LogGP 模型

- LogP 模型有着重要的地位，同时也是后续许多模型的基础，其限制是仅考虑短消息。
- LogGP 模型多引入了一个参数 G (Gap per byte)，用来向网络中注入两个字节最小间隔，对于长消息而言 $1/G$ 代表网络带宽，对应于Hockney中 β (每 byte 数据传输耗时)。



- LogGP 模型对网络带宽建模无疑是更准确，发送单个消息的时间成本可表示为：

$$T_{p2p} = 2o + L + (m - 1)G$$

- 发送 n 个消息为：

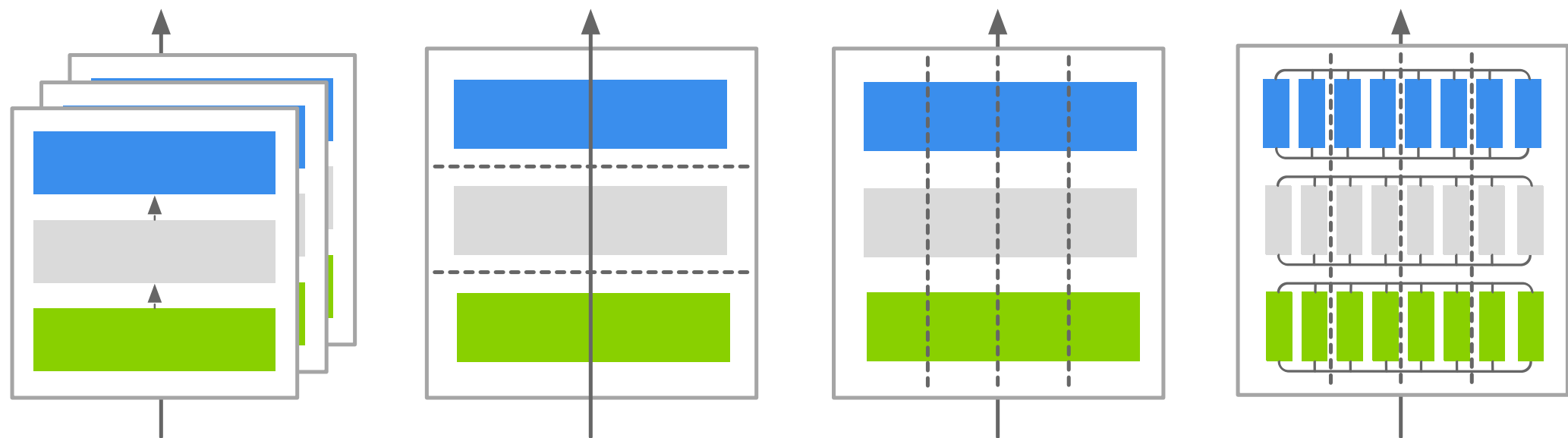
$$T_{p2p} = 2o + n * (L + (m - 1)G) + (n - 1) * g$$



03.集合通信算法带 宽影响因素



大模型典型通信模型



类型	操作	节点规模	数据量	备注
TP	allreduce、allgather、reducescatter	64chip	MB-GB	Matmul流水，隐藏部分通信
PP	Send/Recv	2chip	KB-MB	可隐藏流水
DP	allreduce	1000+ chip	GB	计算通信重叠，可隐藏流水
EP	all2all	256chip	KB-MB	计算通信串行，不可隐藏



集合通信算法带宽

- 集合通信算法带宽 = $\frac{\text{Data}}{\sum_{k=1}^n \text{argmax}(X_{ij,k})}$
- $X_{ijk} = \frac{\text{Data}}{\text{物理带宽} \times \text{带宽利用率} \times \text{收敛比} \times \text{拥塞比}} + T_{\text{通信抖动}} + T_{\text{通信启动}} + T_{\text{静态时延}}$
- 其中:
 - $X_{ij,k}$ 第 K 个数据切片在 rank i 跟 rank j 之间数据传输完成时间
 - argmax: 集合通信性能受限于短板效应



集合通信算法带宽

- 集合通信算法带宽 = $\frac{\text{Data}}{\sum_{k=1}^n \text{argmax}(X_{ij,k})}$
- $X_{ijk} = \frac{\text{Data}}{\text{物理带宽} \times \text{带宽利用率} \times \text{收敛比} \times \text{拥塞比}} + T_{\text{通信抖动}} + T_{\text{通信启动}} + T_{\text{静态时延}}$

互联拓扑

- 其中:
 - $X_{ij,k}$ 第 K 个数据切片在 rank i 跟 rank j 之间数据传输完成时间
 - argmax: 集合通信性能受限于短板效应



集合通信算法带宽

- 集合通信算法带宽 = $\frac{\text{Data}}{\sum_{k=1}^n \text{argmax}(X_{ij,k})}$
- $X_{ijk} = \frac{\text{Data}}{\text{物理带宽} \times \text{带宽利用率} \times \text{收敛比} \times \text{拥塞比}} + T_{\text{通信抖动}} + T_{\text{通信启动}} + T_{\text{静态时延}}$

硬件能力
IO Die & 交换芯片

- 其中:
 - $X_{ij,k}$ 第 K 个数据切片在 rank i 跟rank j 之间数据传输完成时间
 - argmax: 集合通信性能受限于短板效应



集合通信算法带宽

- 集合通信算法带宽 = $\frac{\text{Data}}{\sum_{k=1}^n \text{argmax}(X_{ij,k})}$
- $X_{ijk} = \frac{\text{Data}}{\text{物理带宽} \times \text{带宽利用率} \times \text{收敛比} \times \text{拥塞比}} + T_{\text{通信抖动}} + T_{\text{通信启动}} + T_{\text{静态时延}}$

物理调度 & 网络拓扑
& 通信算法 HCCL

- 其中:
 - $X_{ij,k}$ 第 K 个数据切片在 rank i 跟 rank j 之间数据传输完成时间
 - argmax: 集合通信性能受限于短板效应



集合通信算法带宽

- 集合通信算法带宽 = $\frac{\text{Data}}{\sum_{k=1}^n \text{argmax}(X_{ij,k})}$
- $X_{ijk} = \frac{\text{Data}}{\text{物理带宽} \times \text{带宽利用率} \times \text{收敛比} \times \text{拥塞比}} + T_{\text{通信抖动}} + T_{\text{通信启动}} + T_{\text{静态时延}}$

通信协议
短距离 UBC & 长距离 RoCE

- 其中:
 - $X_{ij,k}$ 第 K 个数据切片在 rank i 跟 rank j 之间数据传输完成时间
 - argmax: 集合通信性能受限于短板效应





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 © 2023 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.



ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/AIFoundation