

分布式集群系列

AI集群架构



BUILDING A BETTER CONNECTED WORLD

Ascend & MindSpore

www.hiascend.com
www.mindspore.cn

关于本内容

1. 内容背景

- AI集群+大模型+分布式训练系统

2. 具体内容

- **AI集群服务器架构**：参数服务器模式 – 同步与异步并行 - 环同步算法
- **AI集群软硬件通信**：通信软硬件实现 - 通信实现方式
- **分布式通信原语**：通信原语
- **框架分布式功能**：并行处理硬件架构 – AI框架中的分布式训练

◦ **大模型算法**：挑战 – 算法结构 – SOTA大模型

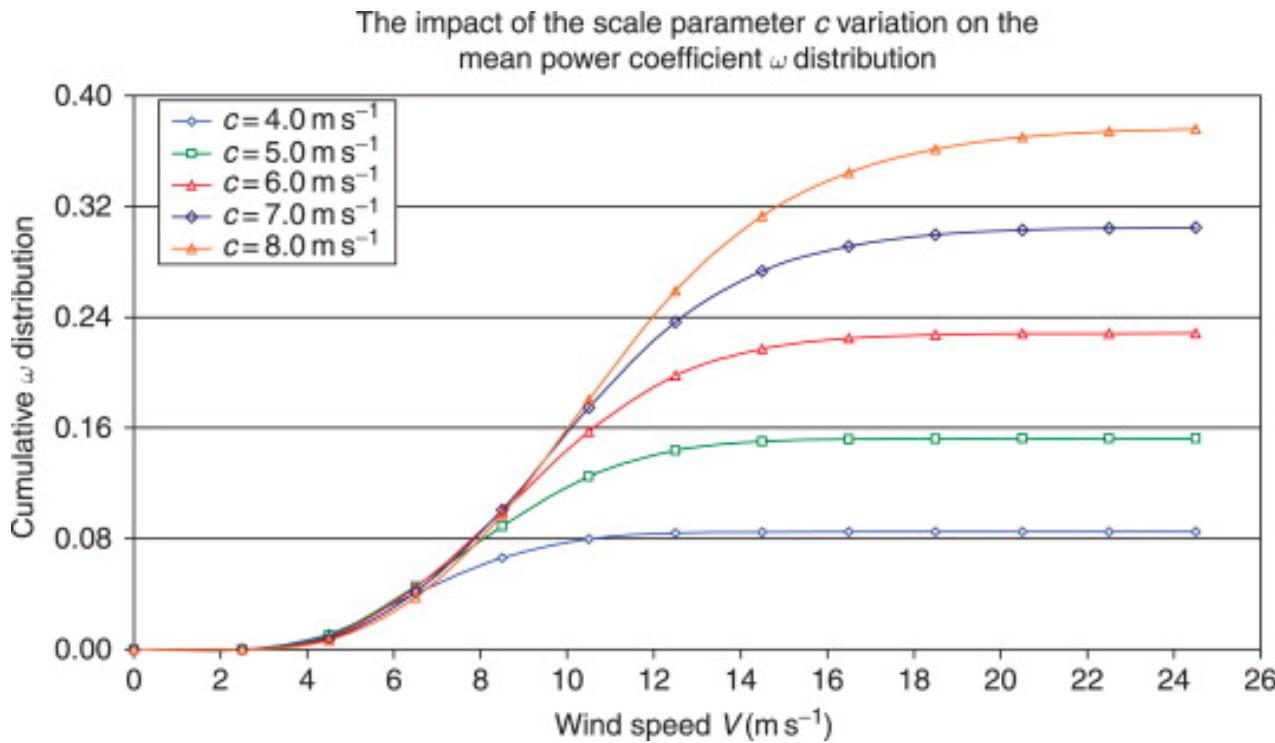
◦ **分布式并行**：数据并行 – 张量并行 – 自动并行 – 多维混合并行

分布式系统：分布式并行架构



加速比

$$\text{scale factor} = \frac{T_n}{nT}$$



分布式系统：分布式并行架构

- 深度学习训练耗时：

$$\text{训练耗时} = \text{训练数据规模} \times \text{单步计算量} / \text{计算速率}$$


模型相关，相对固定 可变因素

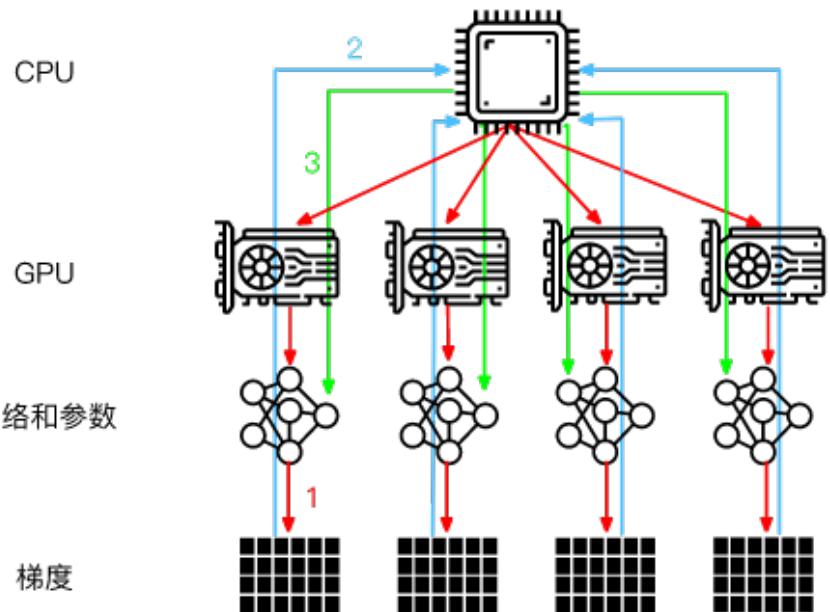
- 计算速率：

$$\text{计算速率} = \text{单设备计算速率} \times \text{设备数} \times \text{多设备并行效率（加速比）}$$

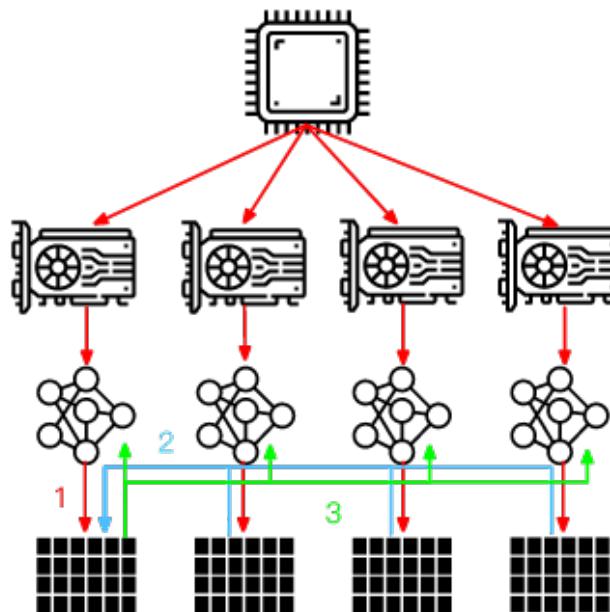

混合精度 服务器架构 数据并行
算子融合 通信拓扑优化 模型并行
梯度累加 流水并行

分布式架构：参数服务器

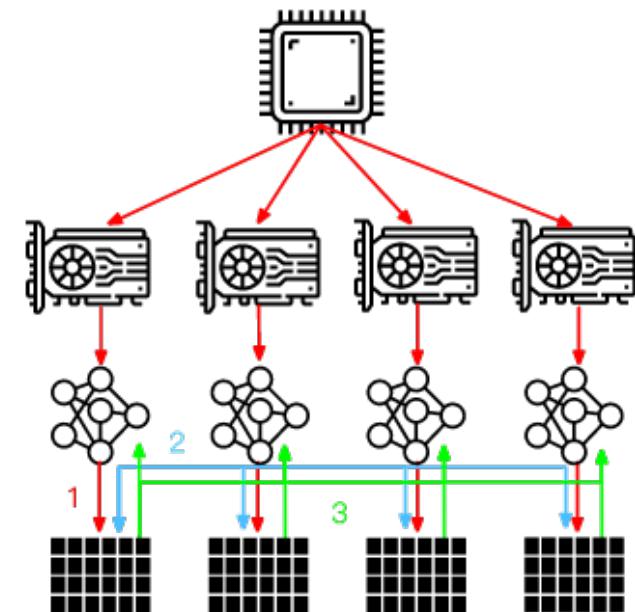
(1) 计算损失和梯度 (2) 梯度聚合 (3) 参数更新并参数重新广播



CPU做参数服务器



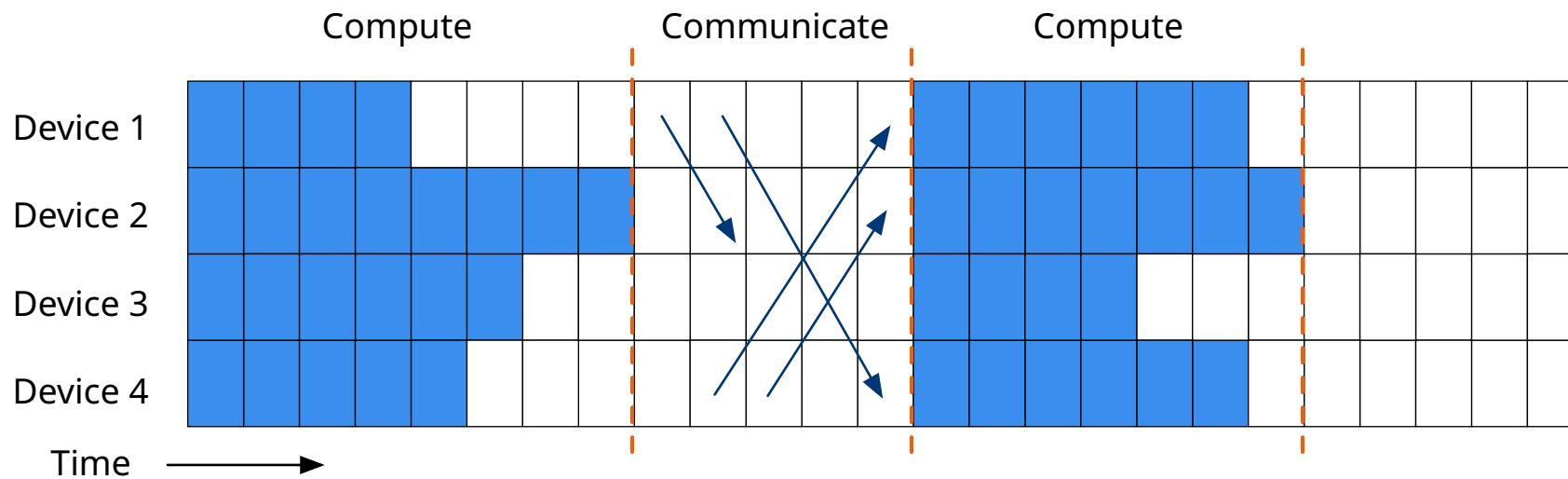
GPU0做参数服务器



参数服务器分布在所有GPU上

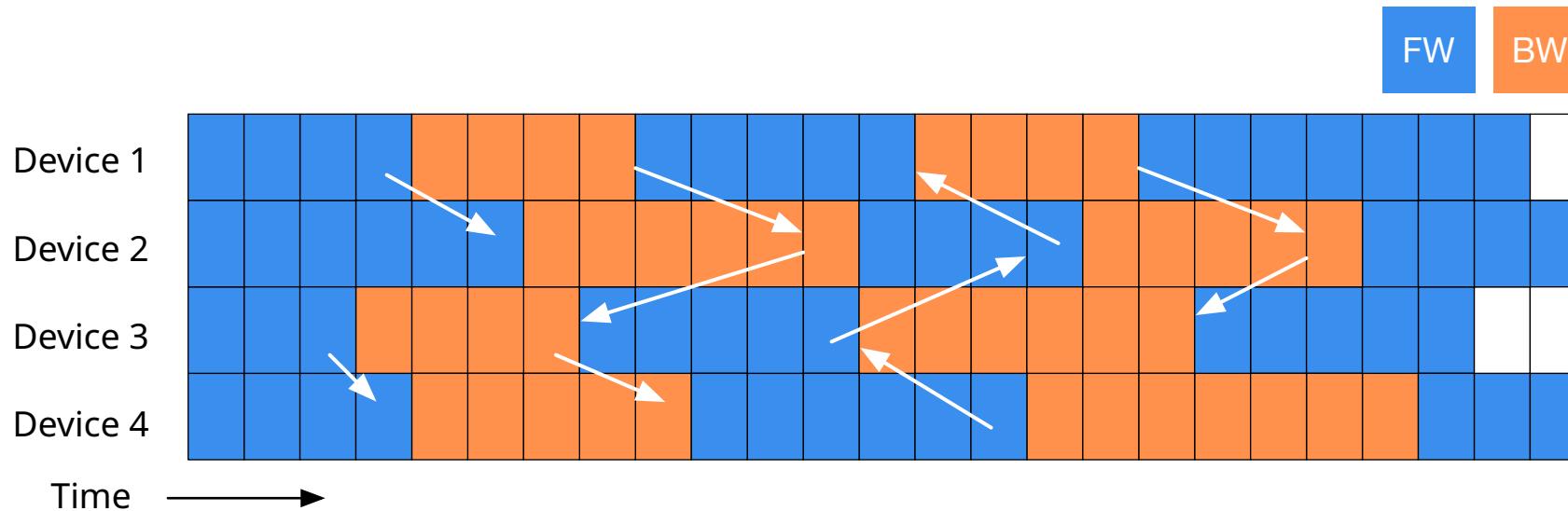
分布式 同步并行

- 必须等全部工作节点完成了本次通信之后才能继续下一轮本地计算
- 优点**：本地计算和通信同步严格顺序化，能够容易地保证并行的执行逻辑于串行相同
- 缺点**：本地计算更早的工作节点需要等待其它工作节点处理，造成了计算硬件的浪费。



分布式 异步并行

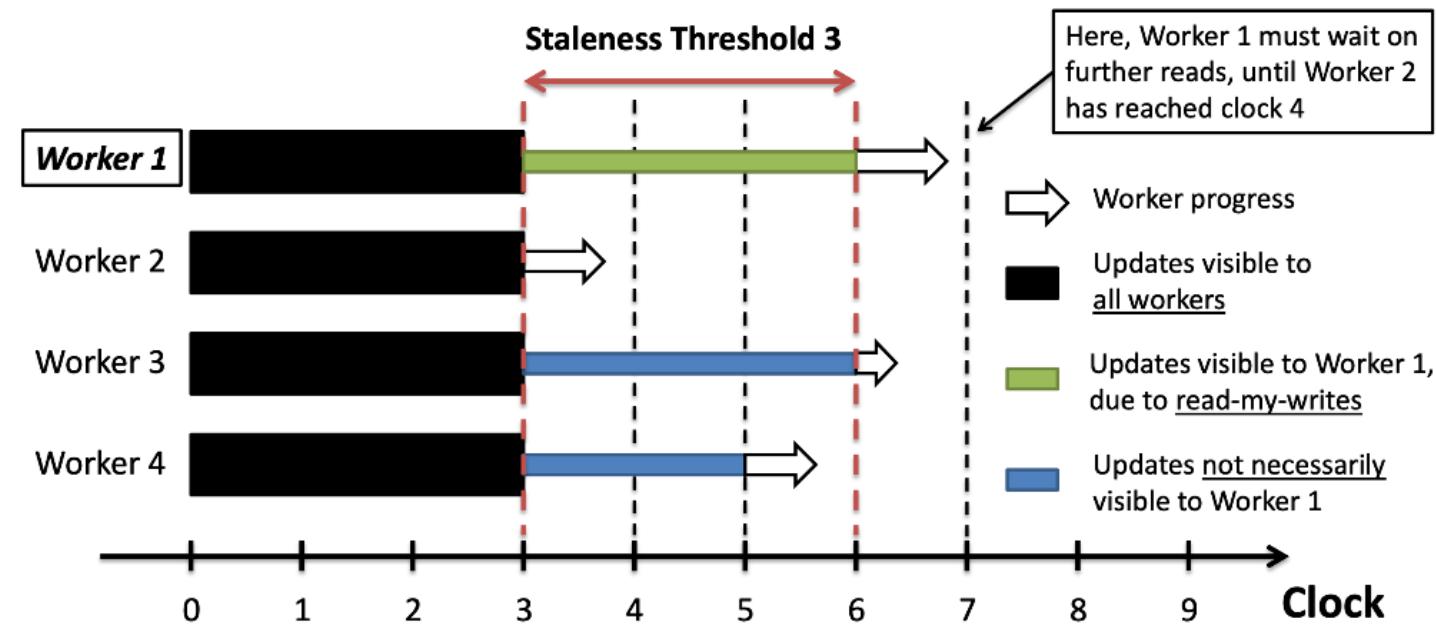
- 当前batch迭代完后与其他服务器进行通信传输网络模型参数
- **优点**：执行效率高，中间除了单机通信时间以外没有任何通信和执行之间的阻塞等待
- **缺点**：网络模型训练不收敛，训练时间长，模型参数反复使用导致无法工业化



分布式 半同步并行

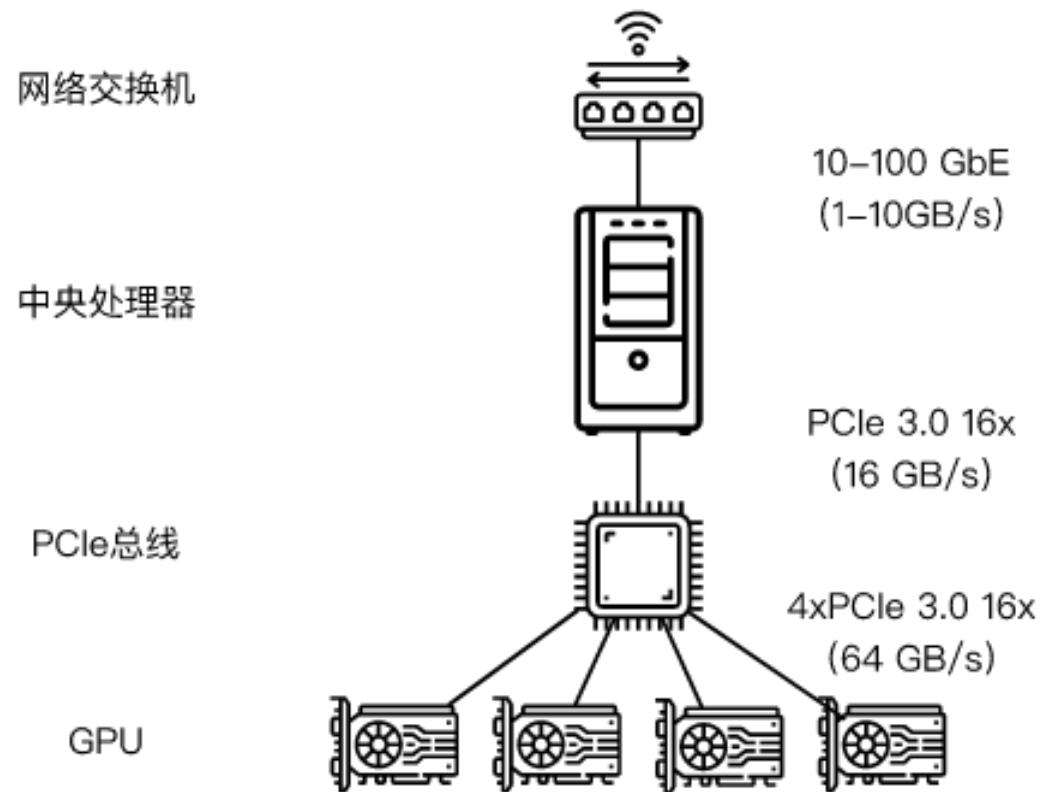
- 通过动态限制进度推进范围，有限定的宽松同步障的通信协调并行
- 跟踪各节点进度并维护最慢节点，保证计算最快和最慢节点差距在一个预定的范围内

SSP: Bounded Staleness and Clocks

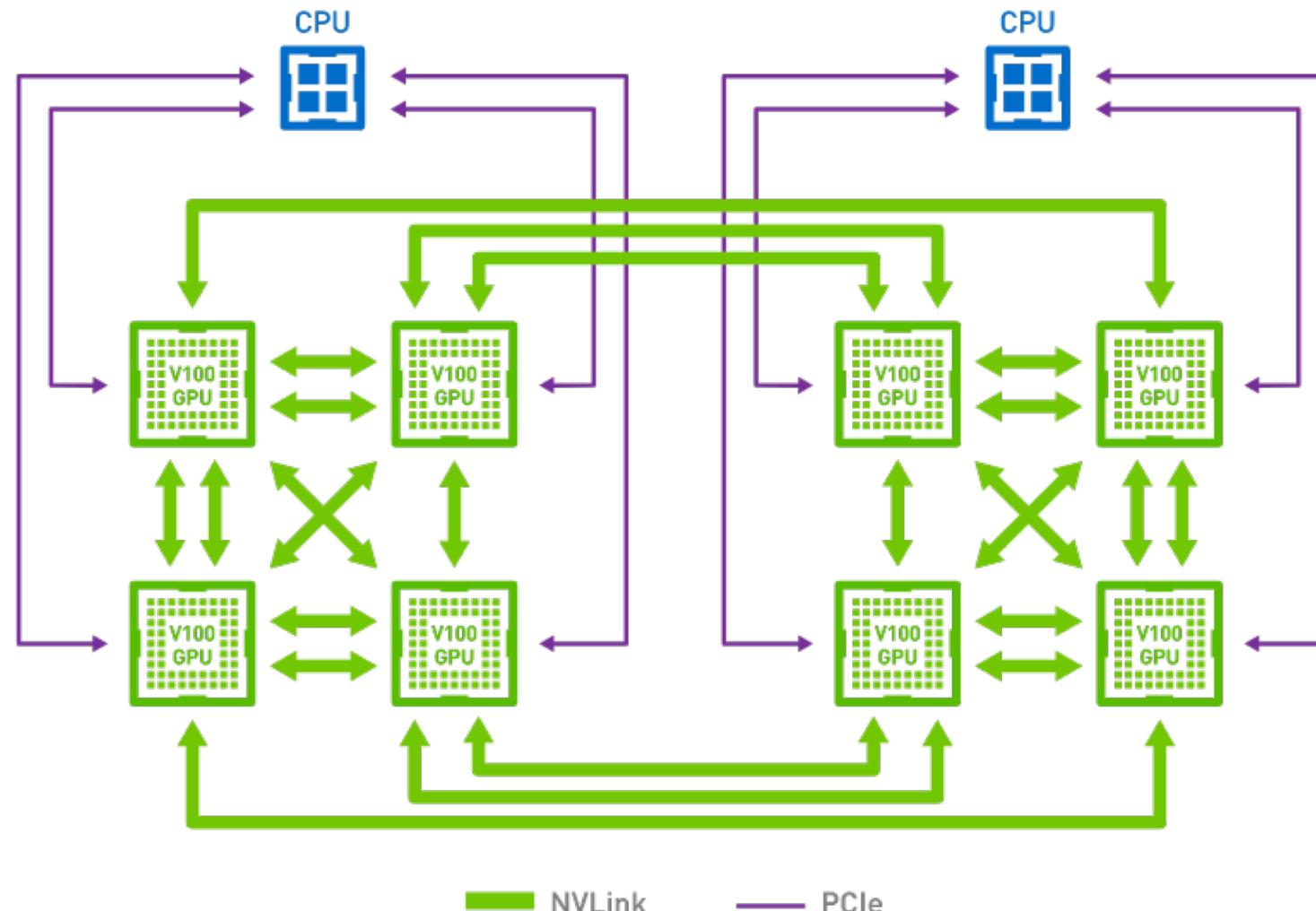


环同步 (Ring Synchronization)

聚合的NVLink带宽明显高于PCIe带宽，问题是如何有效地使用环

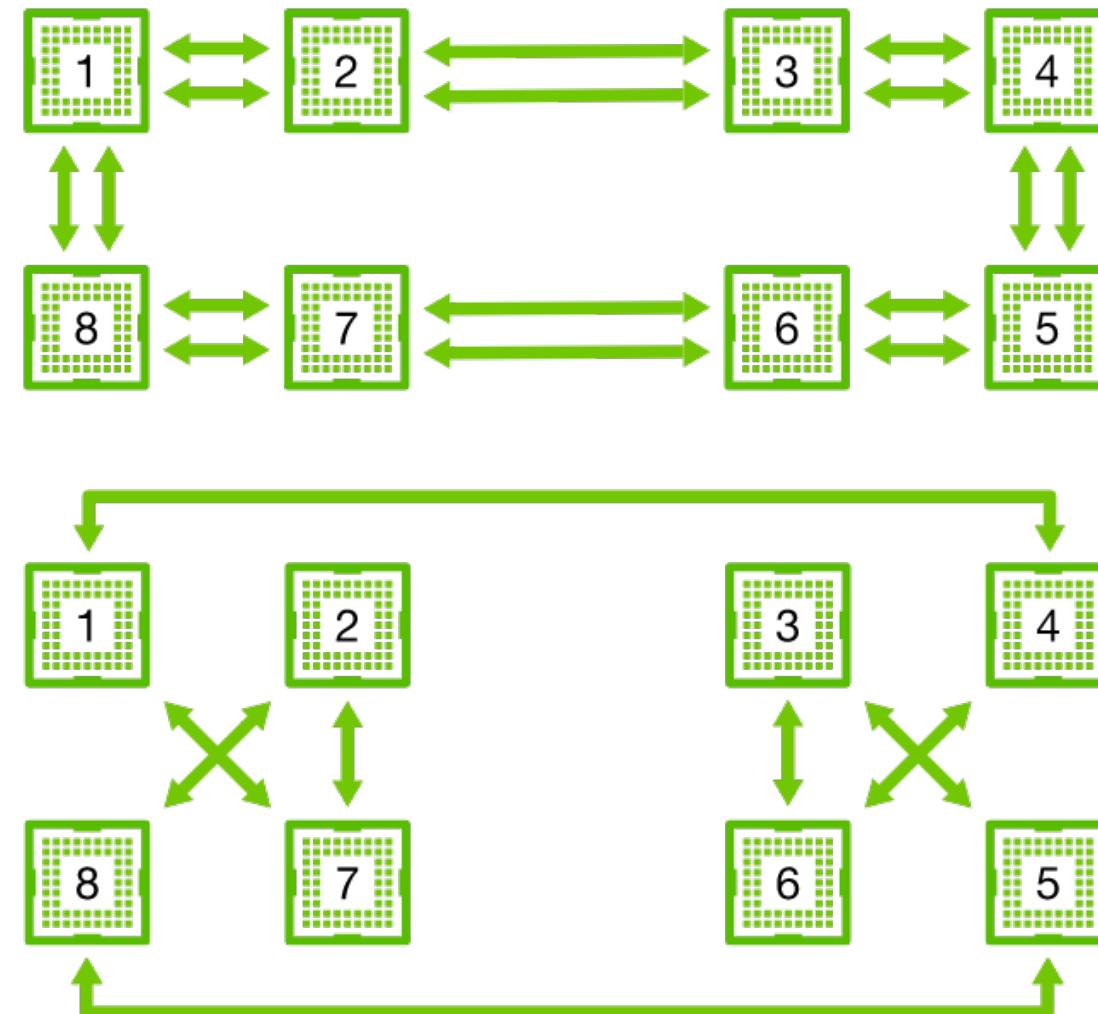


环同步 (Ring Synchronization)



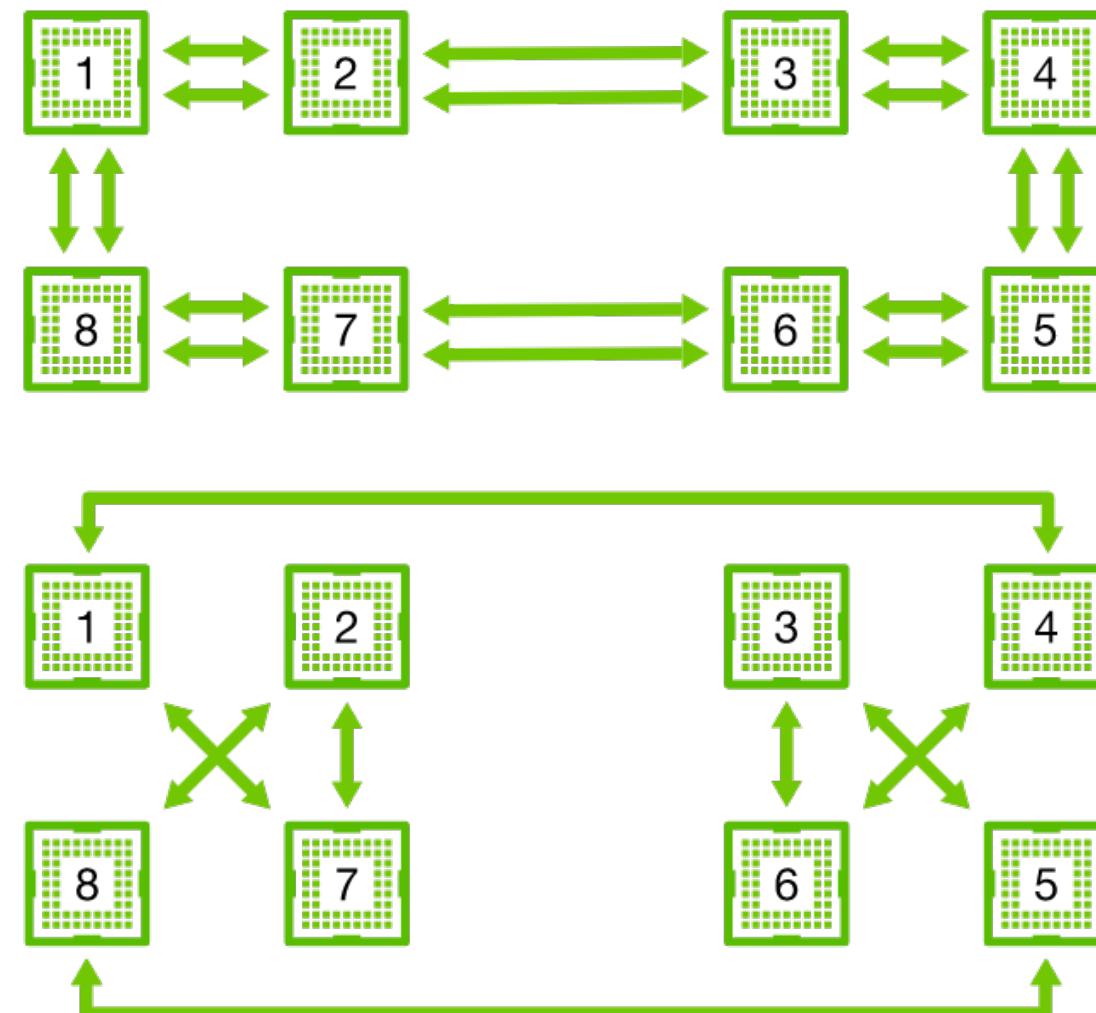
在8台V100 GPU服务器上连接NVLink

环同步 (Ring Synchronization)



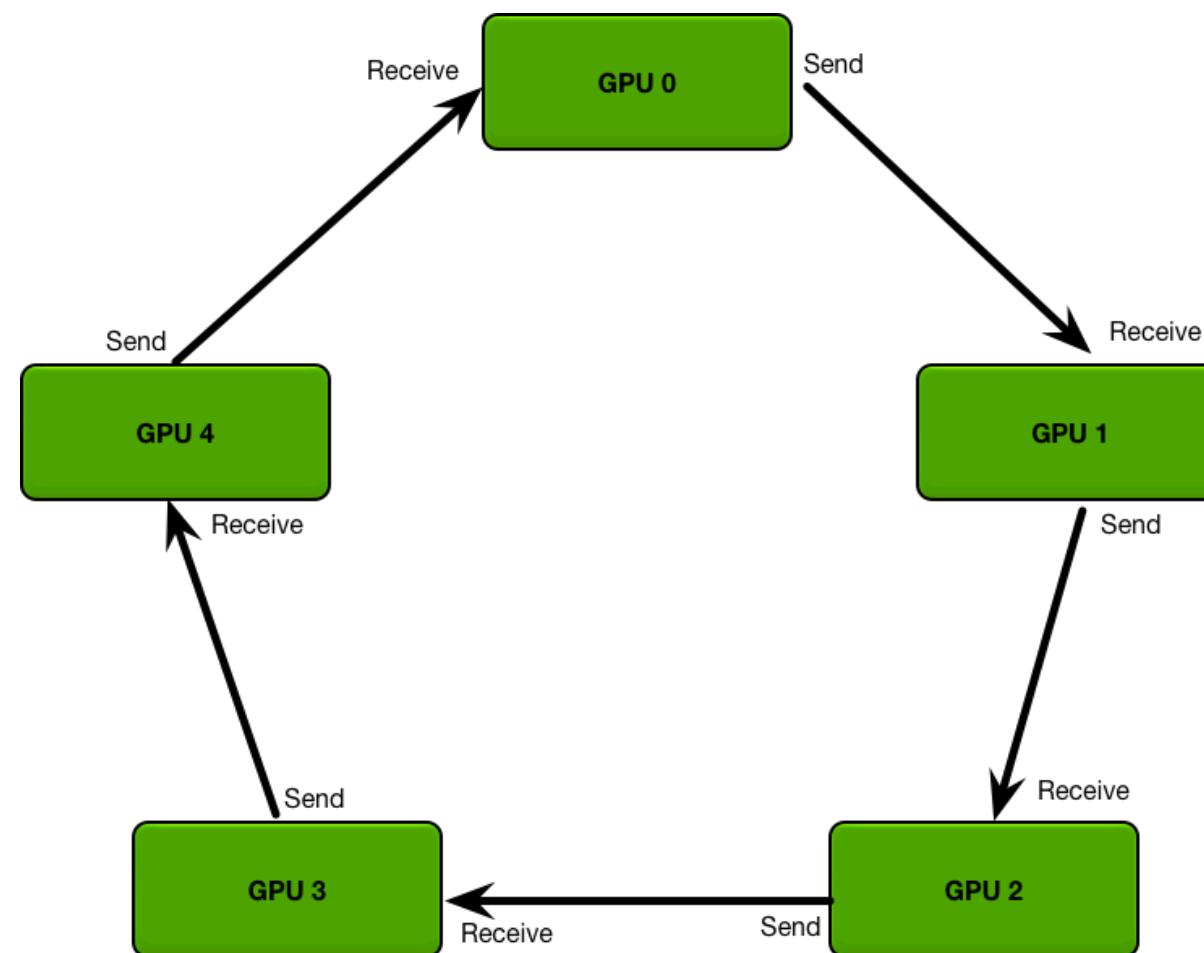
环 (1-2-3-4-5-6-7-8-1)

环同步 (Ring Synchronization)

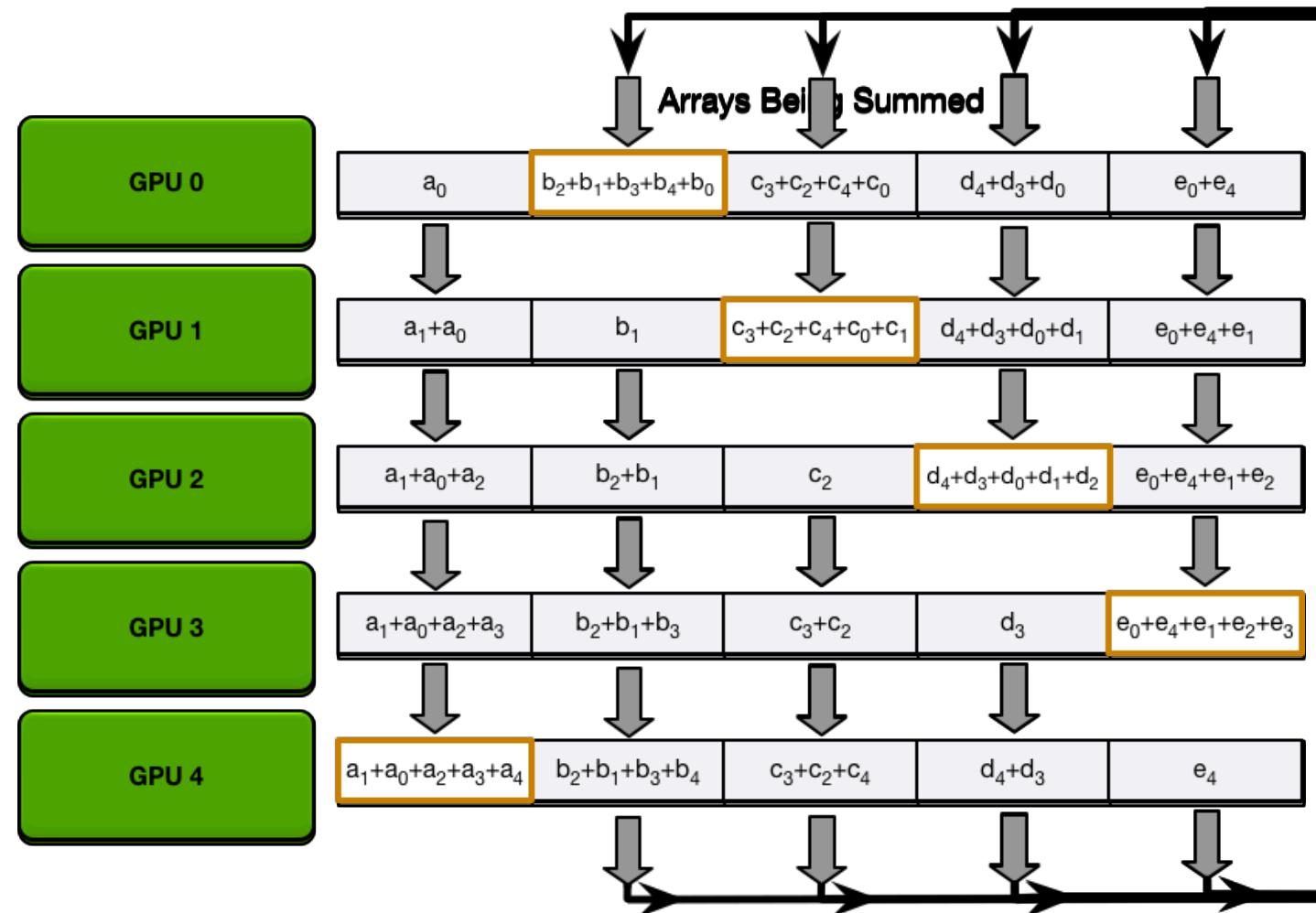


环 (1-4-6-3-5-8-2-7-1)

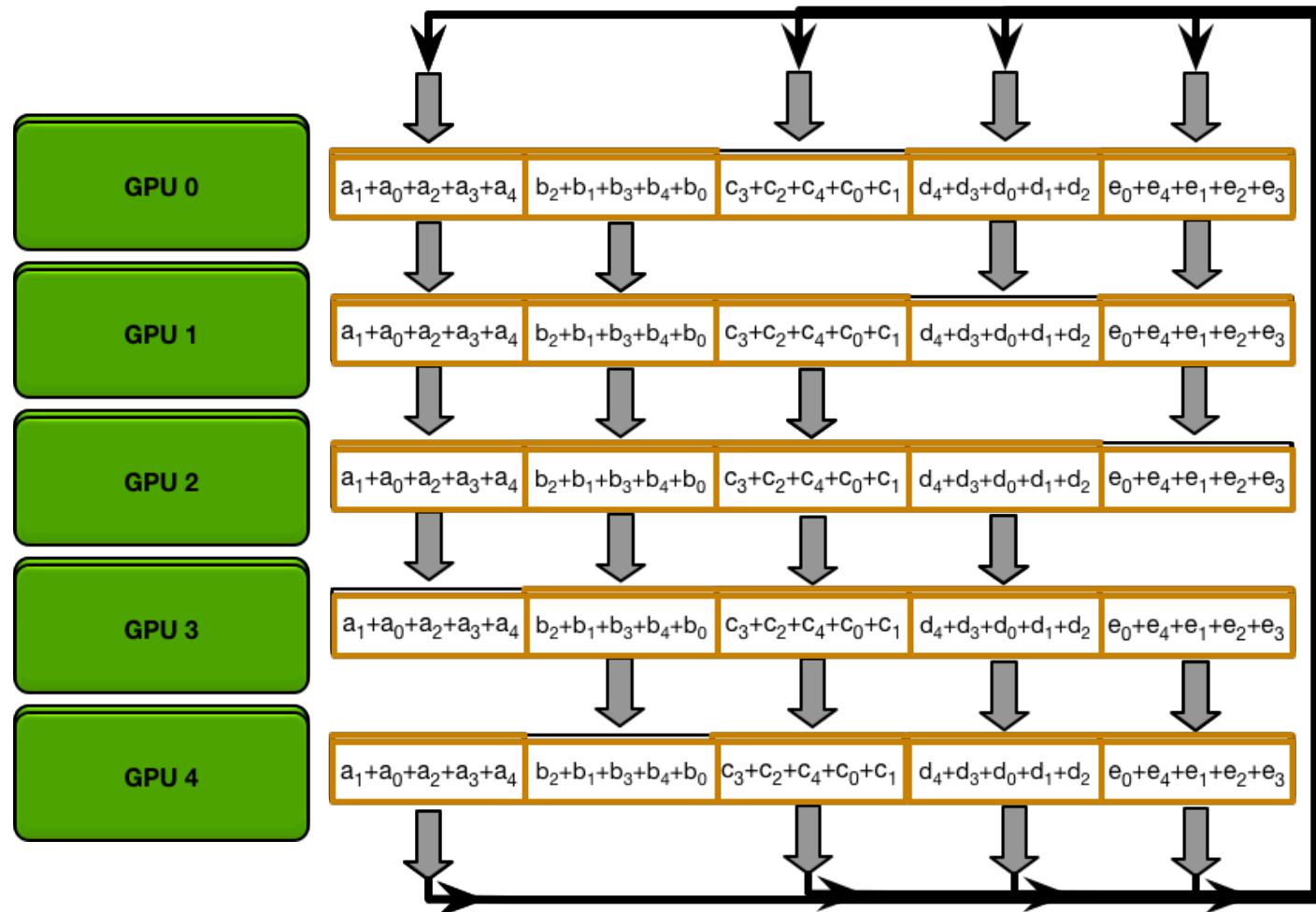
环同步 Ring All Reduce(I)



环同步 Ring All Reduce(II) Scatter-reduce

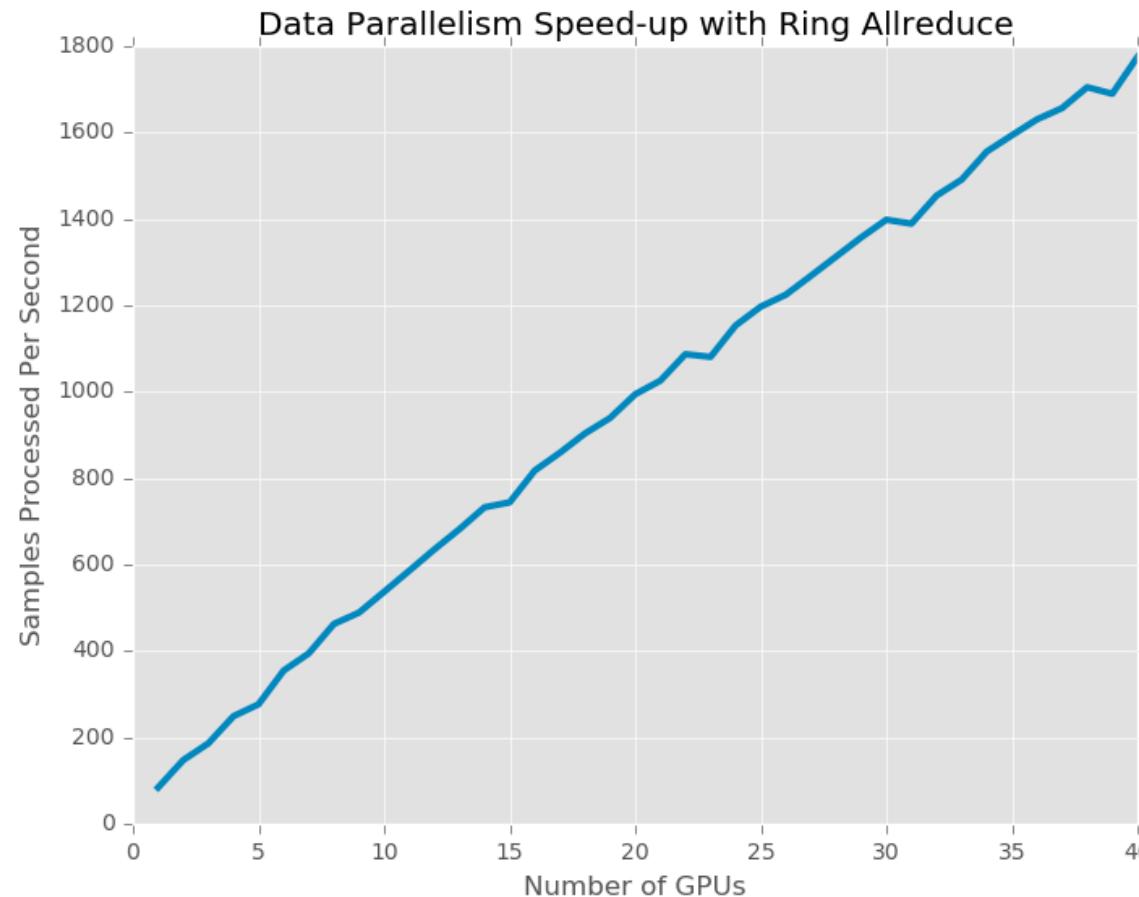


环同步 Ring All Reduce(III) All Gather



环同步 : Ring All Reduce

- 3亿参数语言模型每秒处理的样本数量与同时进行同步训练的GPU数量呈线性关系



Summary

1. 大规模分布式训练中主要使用参数服务器架构模式（ PS ），参数服务器分布在多个 GPU 是 PS 模式的一种特殊形态；
2. PS 架构下通过集合通信来实现环同步，从而同步分布在多个 GPU 中的参数， Ring All Reduce 是环同步的经典同步方式；



BUILDING A BETTER CONNECTED WORLD

THANK YOU

Copyright©2014 Huawei 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. Huawei may change the information at any time without notice.