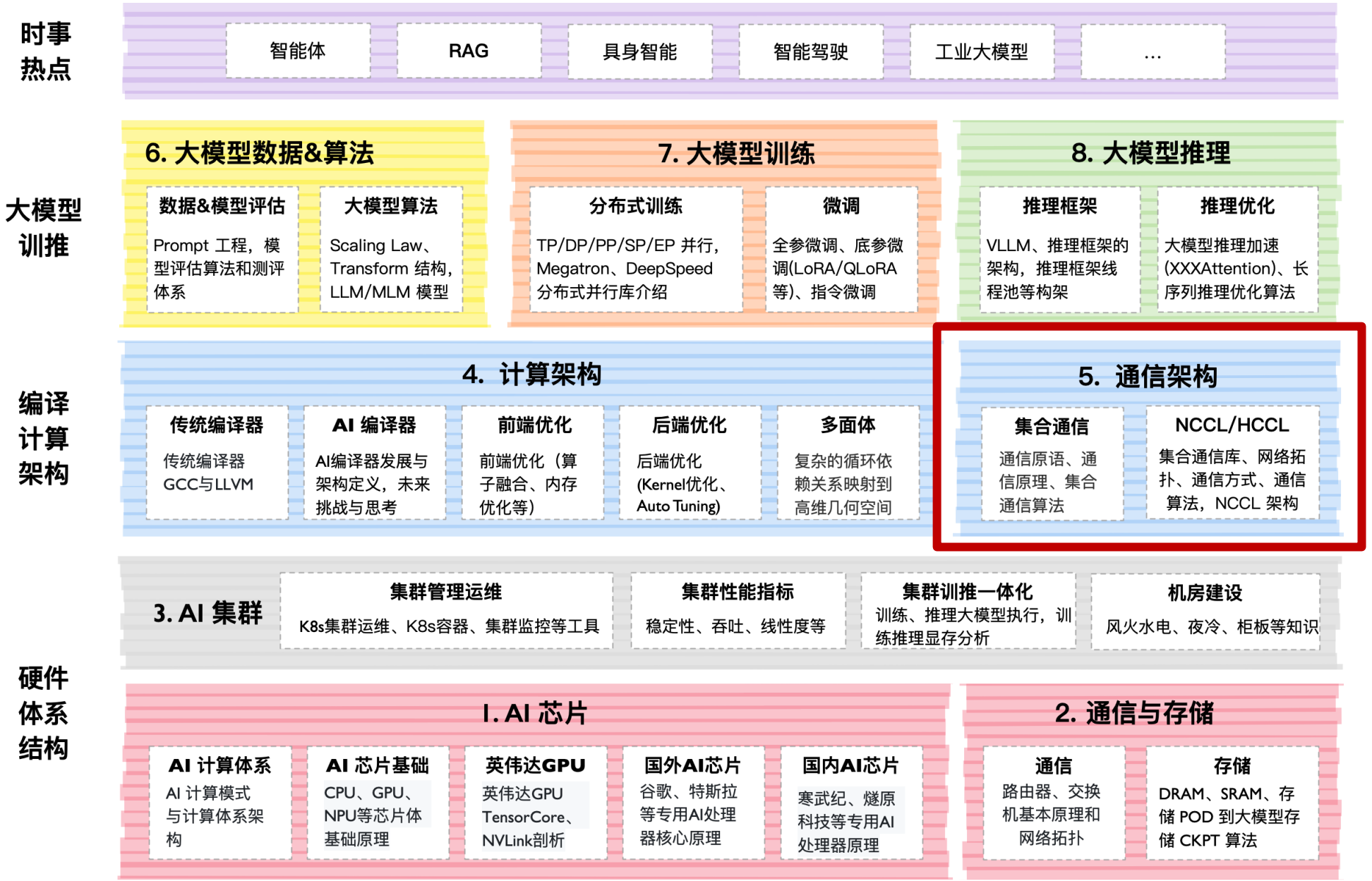
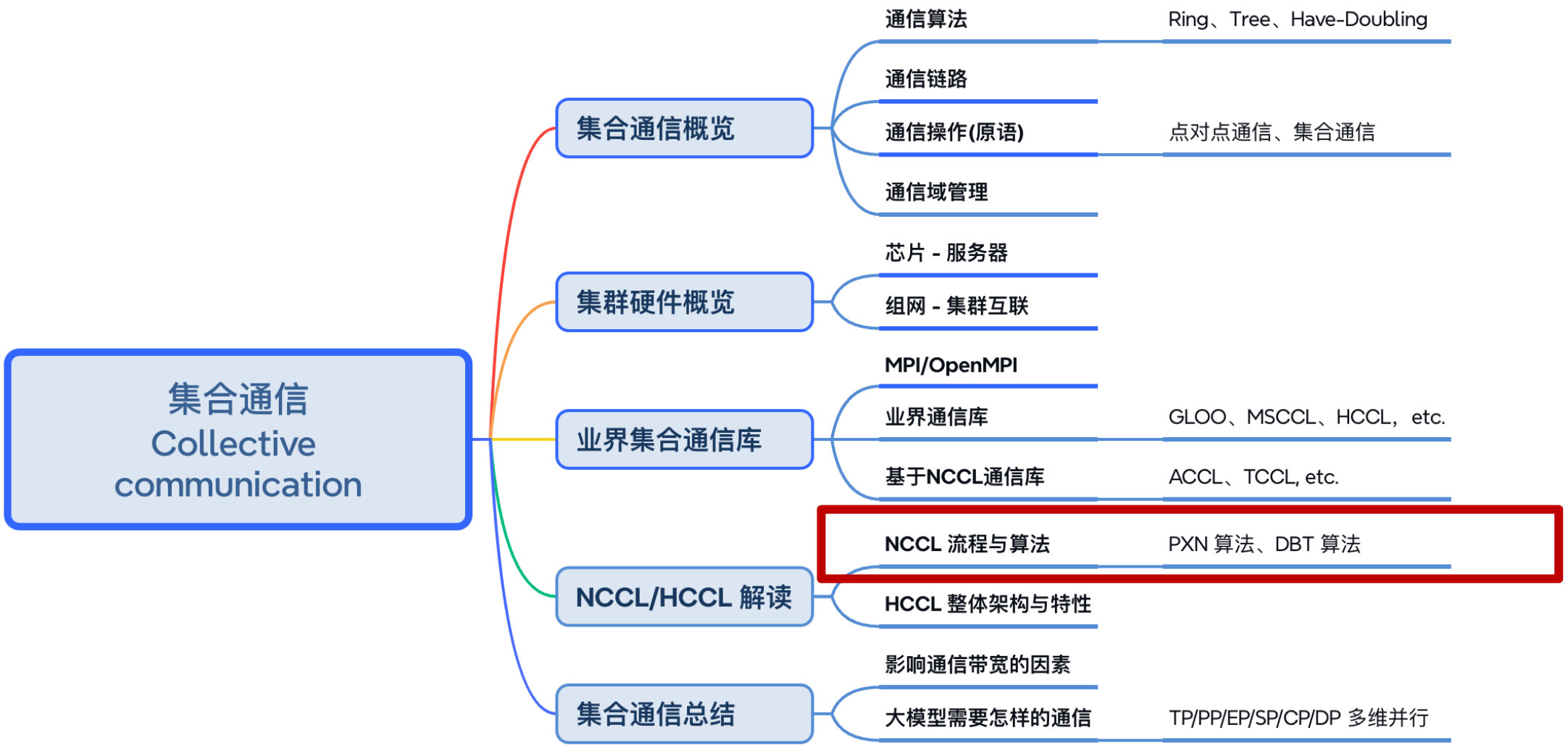
大模型系列 – 集合通信库

拓扑与算法结合

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


集合通信
Collective
communication
集合通信概览
通信算法
Ring、Tree、Have-Doubling
通信链路
通信操作(原语)
点对点通信、集合通信
通信域管理
集群硬件概览
芯片 - 服务器
组网 - 集群互联
业界集合通信库
MPI/OpenMPI
业界通信库
GLOO、MSCCL、HCCL，etc.
基于NCCL通信库
ACCL、TCCL, etc.
NCCL/HCCL 解读
NCCL 流程与算法
PXN 算法、DBT 算法
HCCL 整体架构与特性
集合通信总结
影响通信带宽的因素
大模型需要怎样的通信
TP/PP/EP/SP/CP/DP 多维并行

Question

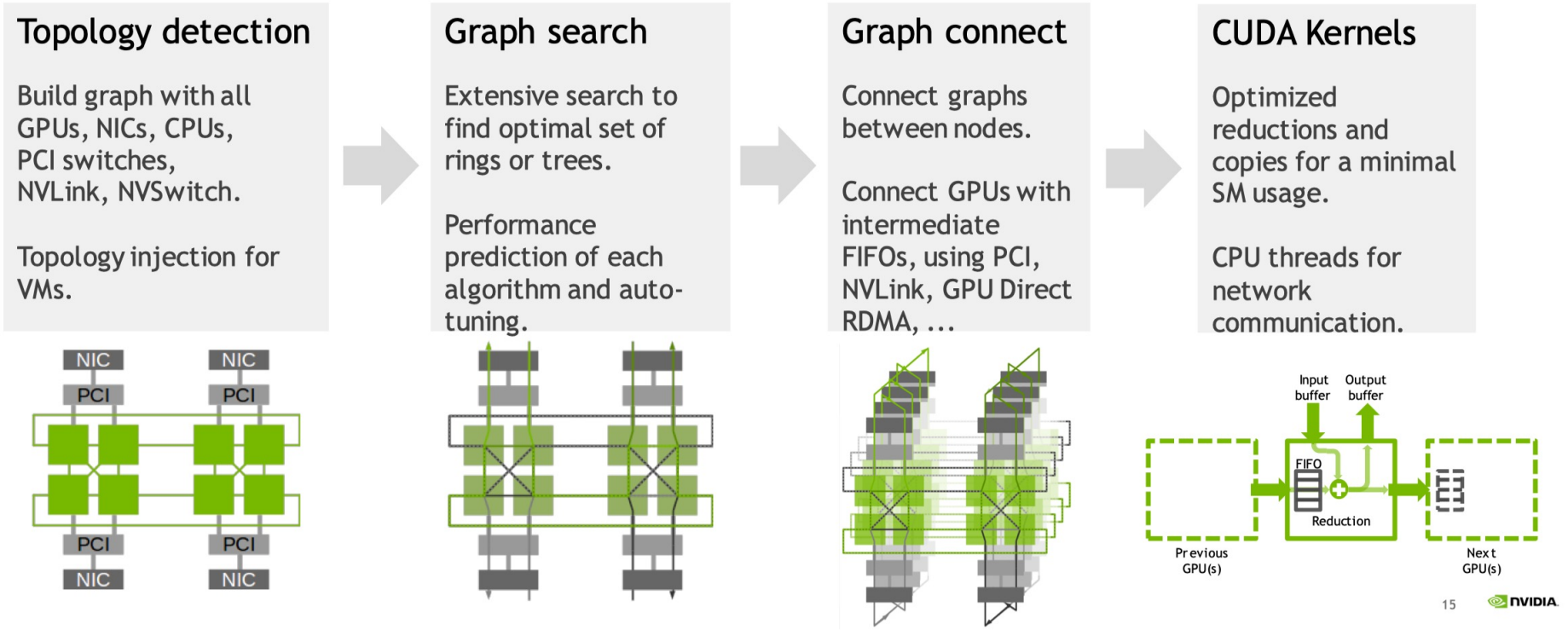
1. 如何在 GPU 集群上，网络拓扑和通信算法协同优化，提升集合通信的性能?

01. NCCL 简介

Basic Concept

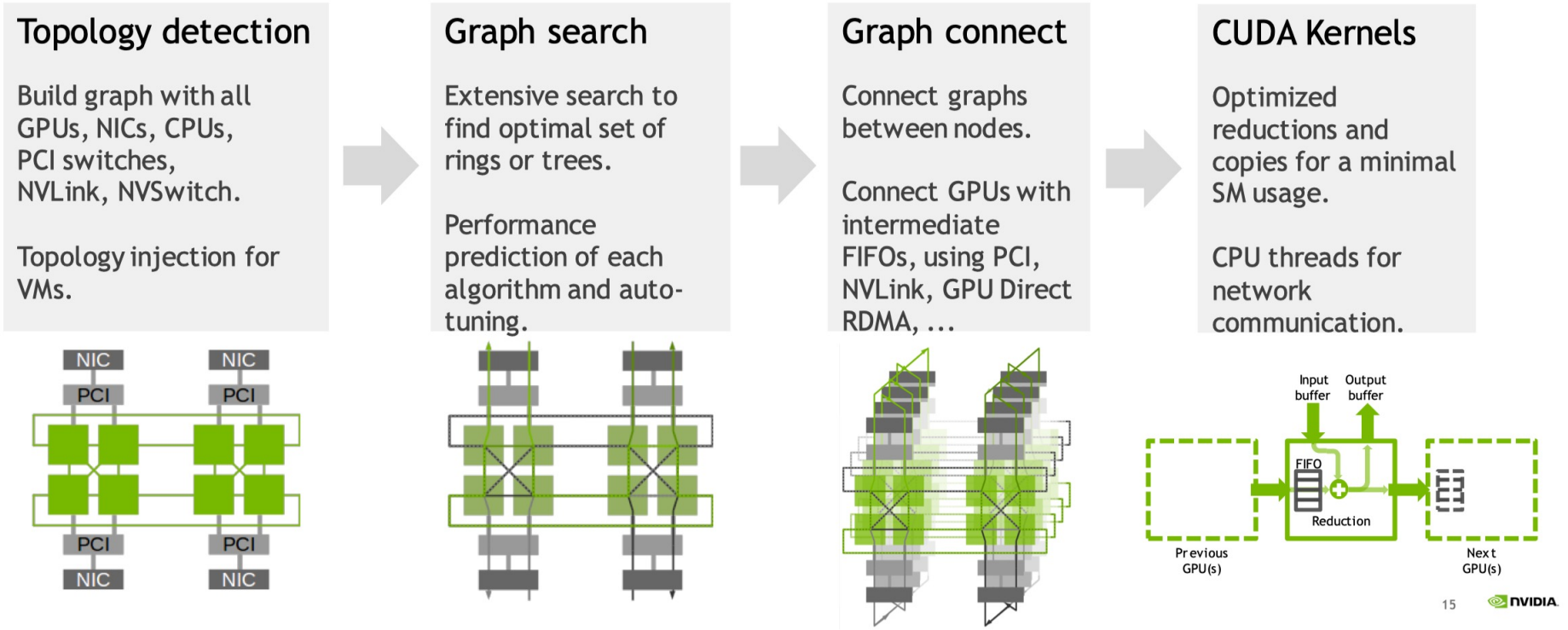
NCCL 基本介绍

- NVIDIA 集合通信库 (NCCL) 是一个类 MPI 通信库，可实现 GPU 的集合通信算法相关操作：
- all-gather / all-reduce / broadcast / reduce / reduce-scatter / point-to-point send and receive



NCCL 基本介绍

- NCCL 具有拓扑感知能力，经过优化可通过 PCIe、NVLink、以太网和 InfiniBand 互连实现高带宽和低延迟。NCCL GCP 插件支持自定义网络连接的云环境中实现高性能 NCCL 操作。



02. 通信结合

网络拓扑

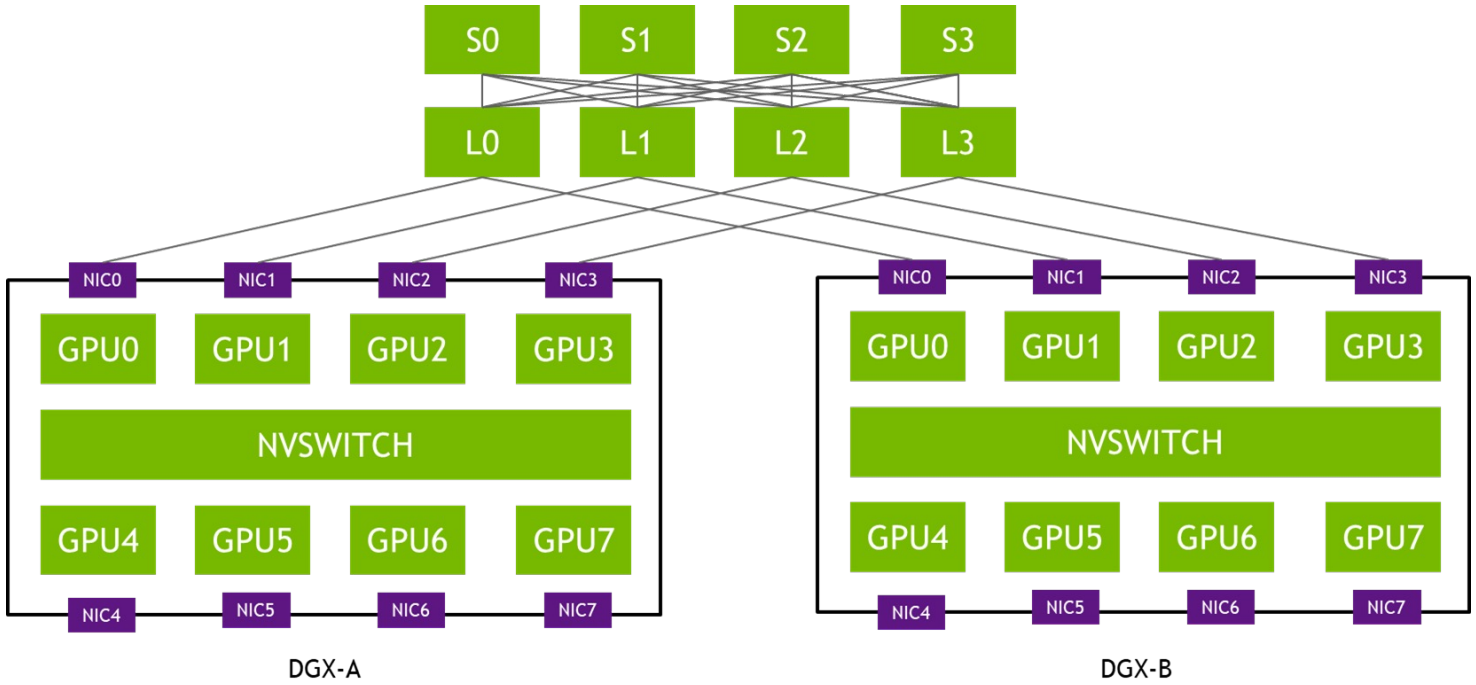
PXN： PCI × NVLink

- 2022 年 NCCL 2.12 引入新功能 PXN，即 PCI × NVLink，使 GPU 能够通过 NVLink 与节点上的 NIC 进行通信。
- 不需要使用 QPI 或其他 CPU 协议。使得每个 GPU 仍然尝试尽可能多地使用本地 NIC 提升集合通信算法，当然也可以访问其他NIC。

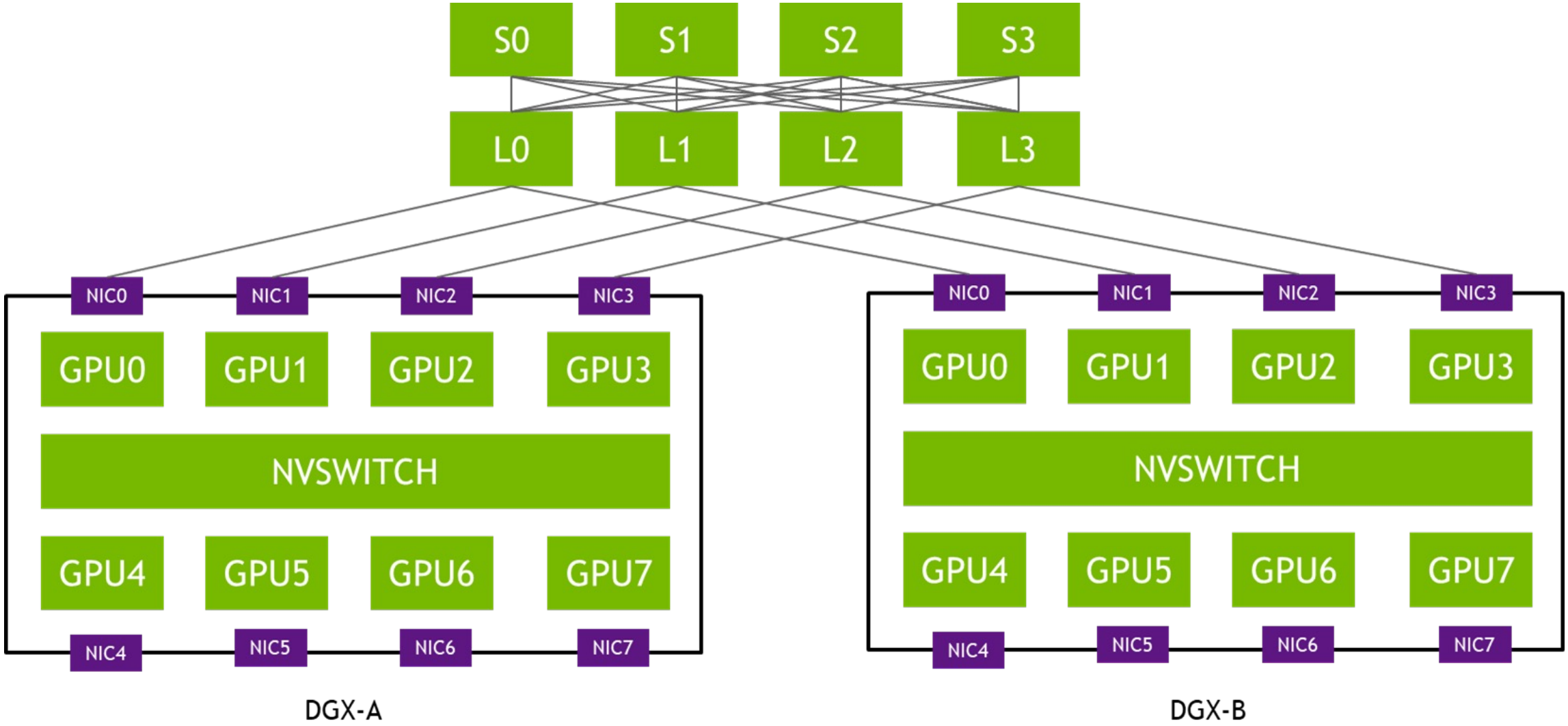
- 快速通道互联（Intel QuickPath Interconnect，QPI），Intel 开发并使用的点对点处理器互联架构，用来实现CPU之间的互联。

PXN： PCI × NVLink

- GPU 不是在本地内存上准备缓冲区供本地 NIC 发送，而在中间 GPU 上准备缓冲区，通过 NVLink 写入该缓冲区。接着通知管理该 NIC 的 CPU 代理数据已经准备好，而不是通知其自己 CPU 代理。

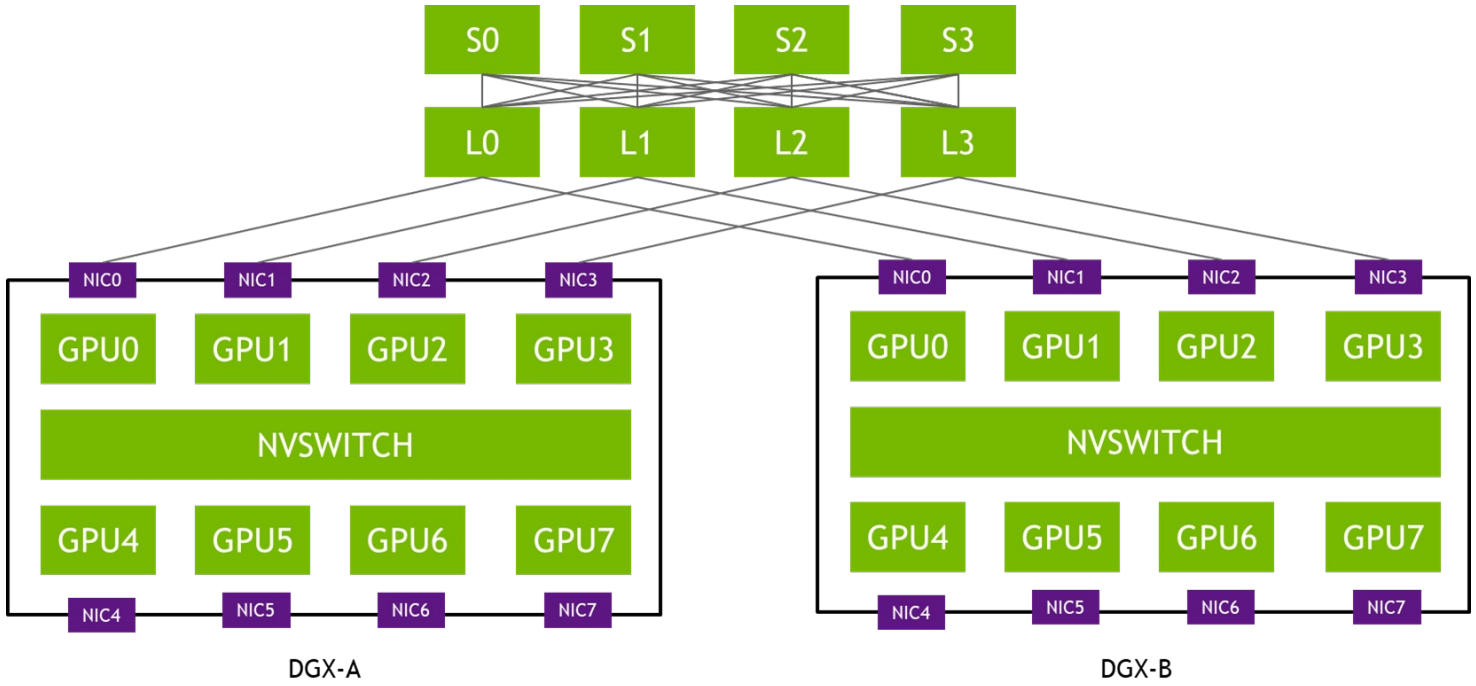


多轨通信

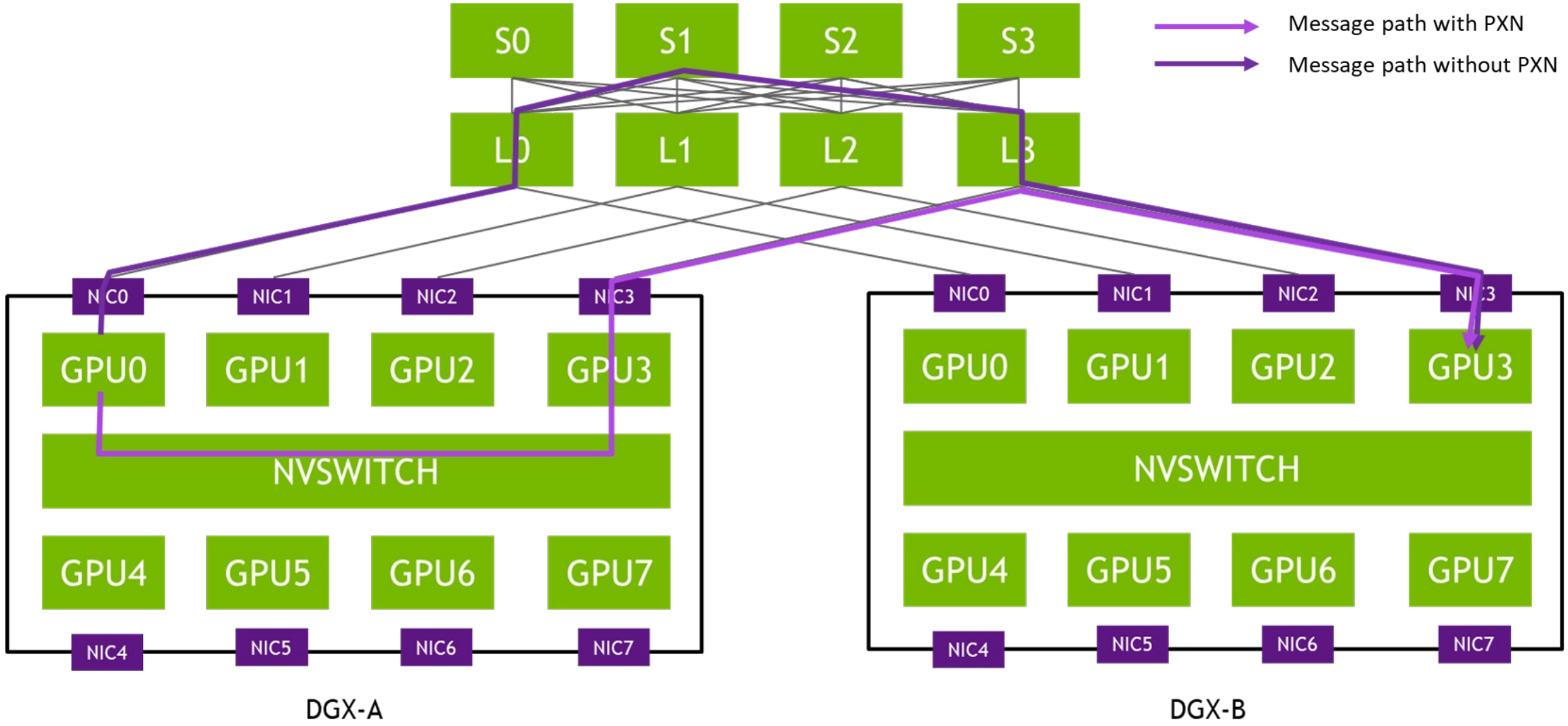

S0
S1
S2
S3
L0
L1
L2
L3
NIC0
NIC1
NIC2
NIC3
GPU0
GPU1
GPU2
GPU3
NVSWITCH
GPU4
GPU5
GPU6
GPU7
NIC4
NIC5
NIC6
NIC7
DGX-A
NIC0
NIC1
NIC2
NIC3
GPU0
GPU1
GPU2
GPU3
NVSWITCH
GPU4
GPU5
GPU6
GPU7
NIC4
NIC5
NIC6
NIC7
DGX-B

多轨通信

- PXN 利用节点内 GPU 间 NVSwitch 连接，首先将数据移动到与目的地位于同一轨道上 GPU，然后将其发送到目的地而无需跨轨道。这可以实现消息聚合和网络流量优化。



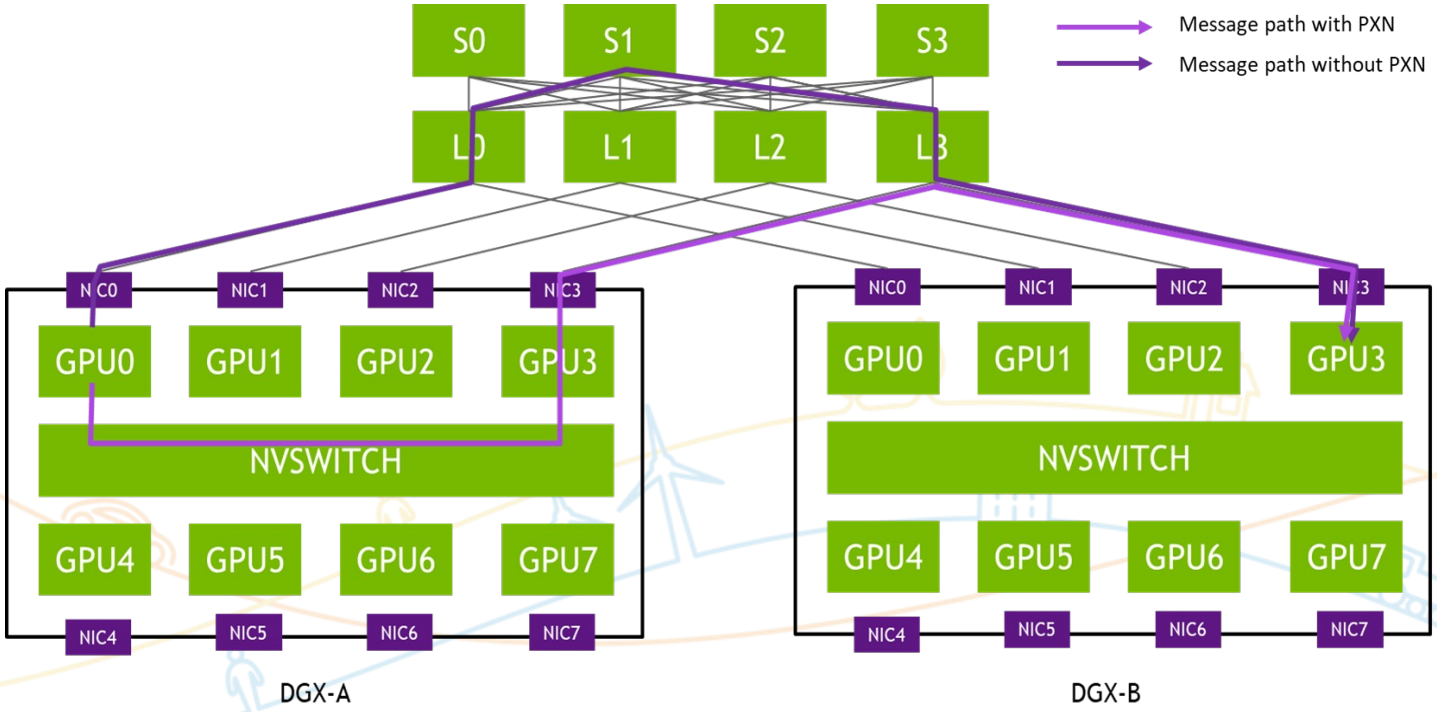
多轨通信


S0
S1
S2
S3
L0
L1
L2
L3
Message path with PXN
Message path without PXN
NIC0
NIC1
NIC2
NIC3
GPU0
GPU1
GPU2
GPU3
NVSWITCH
GPU4
GPU5
GPU6
GPU7
NIC4
NIC5
NIC6
NIC7
DGX-A
DGX-B

Question

I. 多轨通信的好处在哪里？

- 在同一对 NIC 之间传递的数据被聚合，最大限度地提高有效数据传输速率和网络带宽。



03 数据聚合

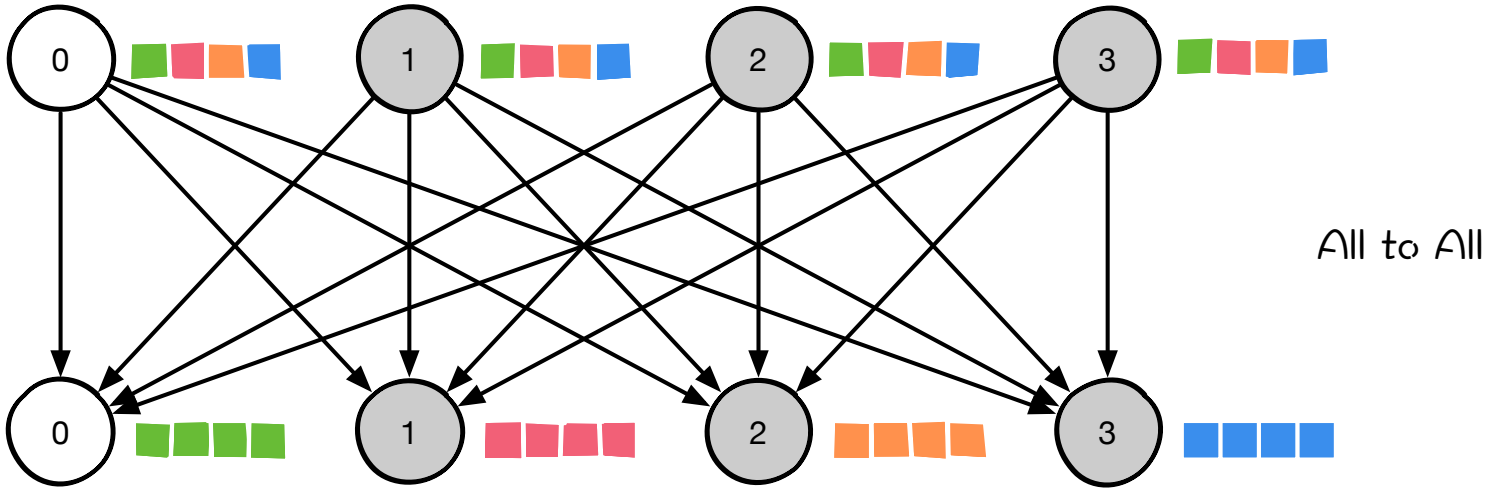
Message Aggregation

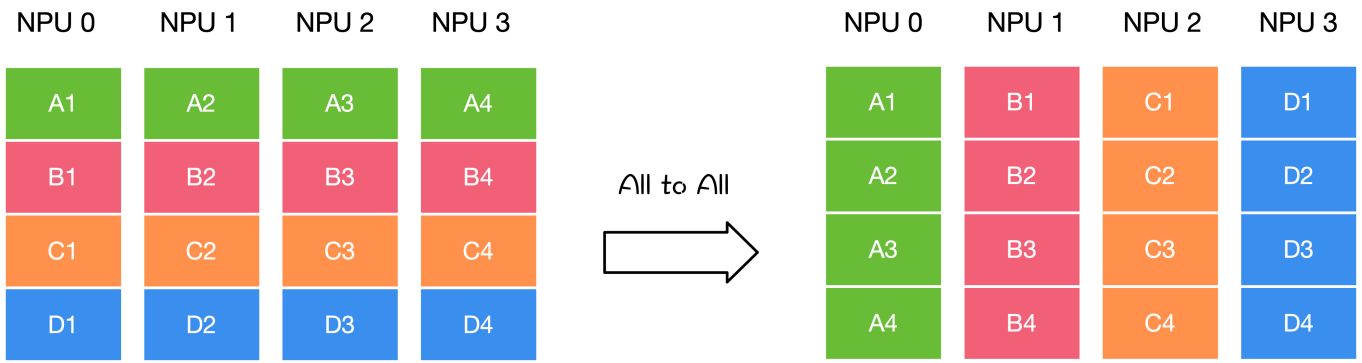
聚合操作

- Aggregation：给定节点上所有 GPU 数据，移动到目的单个 GPU。
- PXN 作用：使网络层能进行消息聚合，建立的网络连接更少，提升路由效率，减少 GPU 负载。

All2All

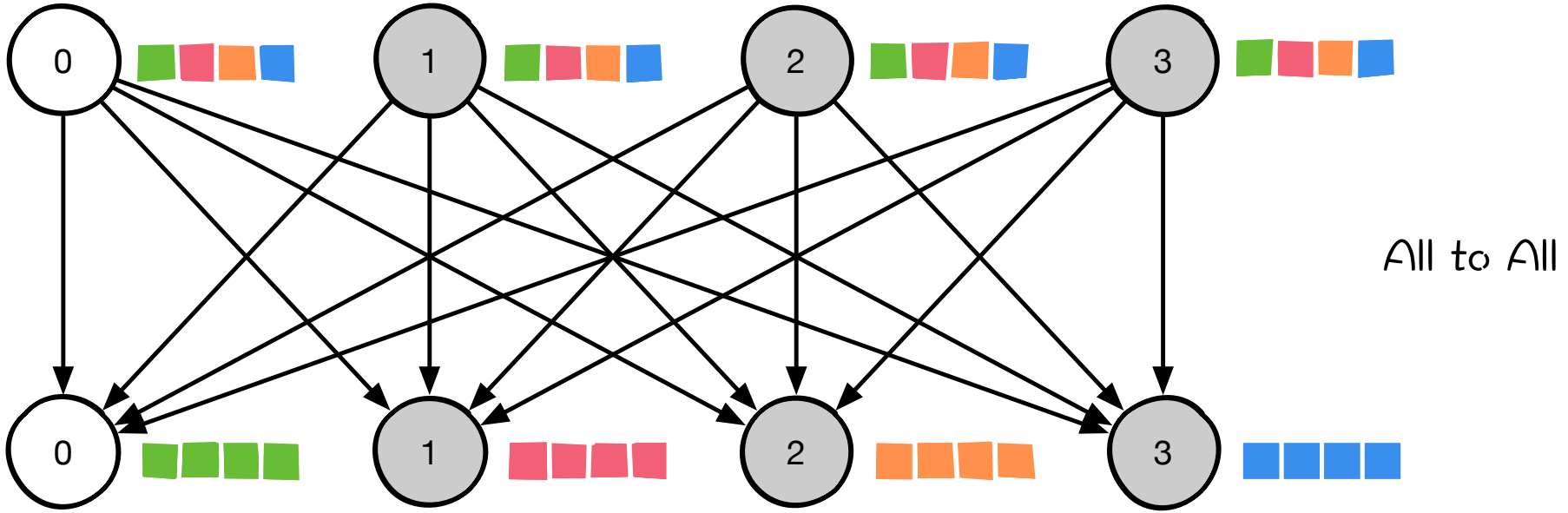
- **算法放方式：**
  - All-Gather 扩展，不同节点向某一节点收集到数据是不同
  - 每个进程与其他所有进程进行通信
- **主要应用场景：**
  - 应用于模型并行(TP/SP/EP)
  - 模型并行里的矩阵转置



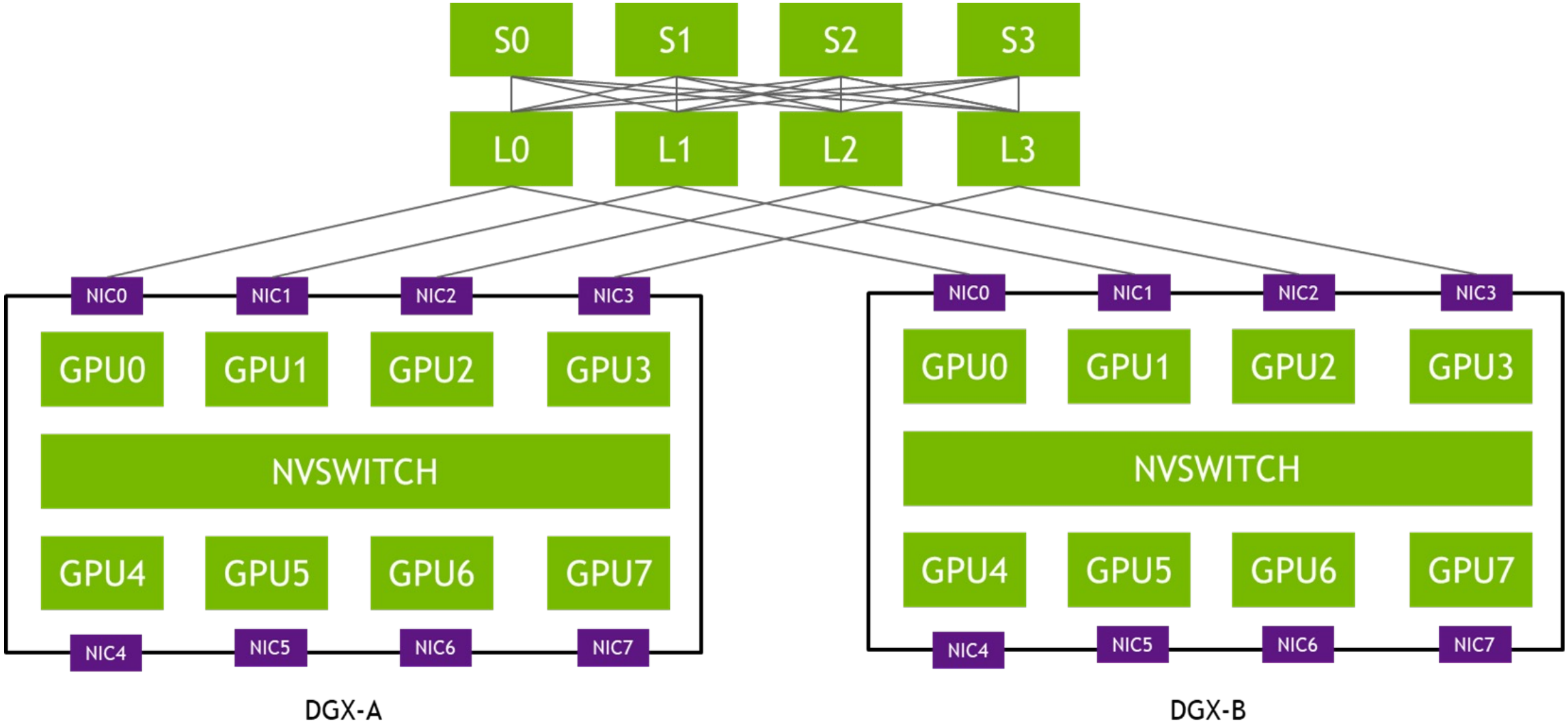


All2All 操作

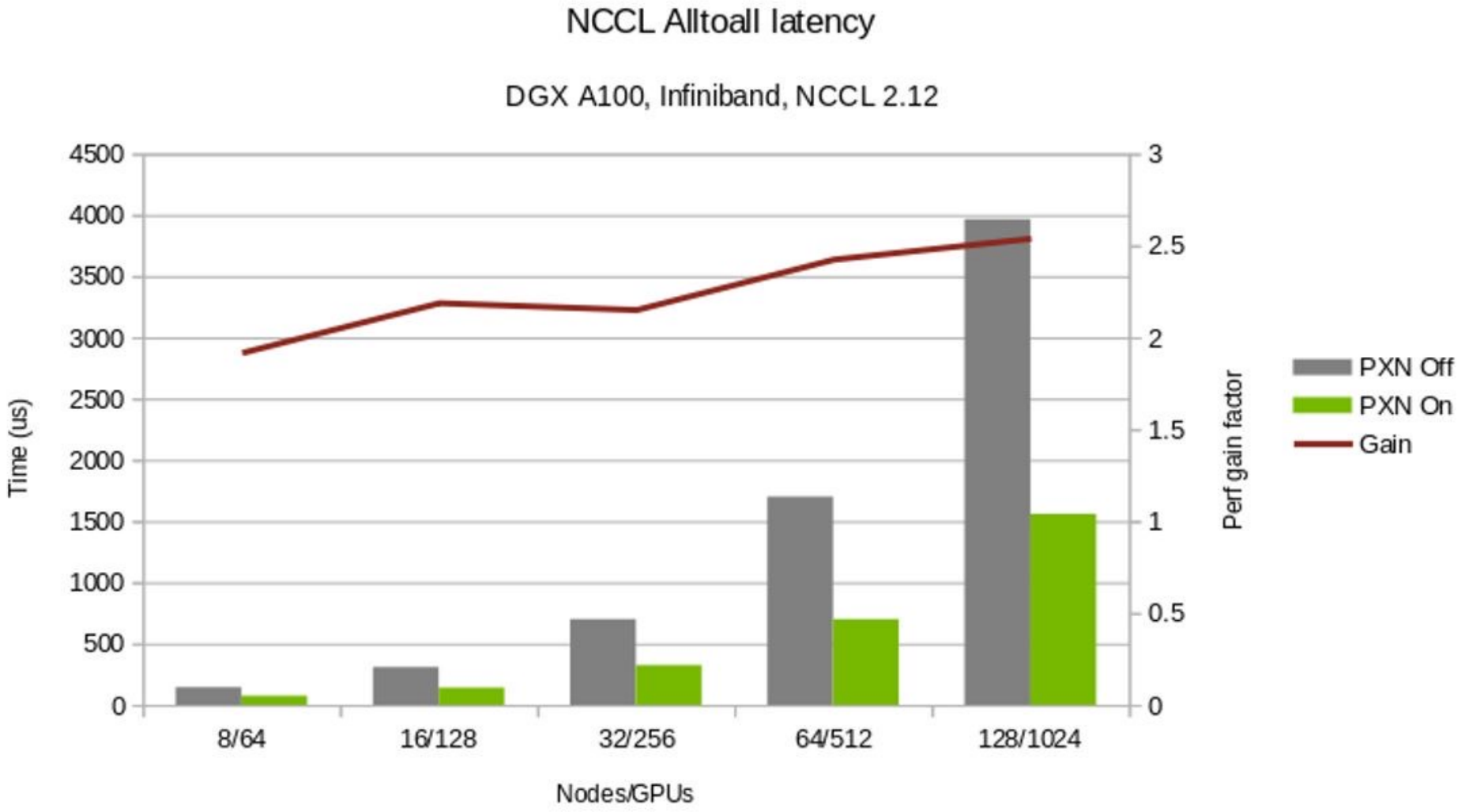
- 如果节点上 GPU 执行 All2All 操作，要从远程节点的八个 GPU 接收数据：
  - 接收方：NCCL 会通过多个接收器，调用八个缓冲区的数据；
  - 发送方：网络层可以等到所有八个发送都准备就绪，然后一次发送 8 条消息。



All2All 操作


S0
S1
S2
S3
L0
L1
L2
L3
NIC0
NIC1
NIC2
NIC3
GPU0
GPU1
GPU2
GPU3
NVSWITCH
GPU4
GPU5
GPU6
GPU7
NIC4
NIC5
NIC6
NIC7
DGX-A
NIC0
NIC1
NIC2
NIC3
GPU0
GPU1
GPU2
GPU3
NVSWITCH
GPU4
GPU5
GPU6
GPU7
NIC4
NIC5
NIC6
NIC7
DGX-B

PXN All2All


NCCL Alltoall latency
DGX A100, Infiniband, NCCL 2.12
Time (us)
4500
4000
3500
3000
2500
2000
1500
1000
500
0
3
2.5
2
1.5
1
0.5
0
Perf gain factor
8/64
16/128
32/256
64/512
128/1024
Nodes/GPUs
PXN Off
PXN On
Gain

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

Course chenzomi12.github.io

GitHub github.com/chenzomi12/AIFoundation

参考资料

- http://www.njwdr.com/newsdetail_3252489.html