白皮书

VIAVI Solutions

AI/ML 数据中心网络验证

评估 AI/ML 网络基础设施的概念、挑战和解决方案

AI/ML 数据中心网络验证

简介

AI 正在重塑数据中心

人工智能(AI)和机器学习(ML)工作负载的规模和复杂性已经重新定义了数据中心设计。为了高效训练万亿参数模型，数据中心正在高速互连网络上部署数千个 GPU 和 xPU。这些分布式集群必须作为一个紧密同步的计算平台运行。

这种转变给网络带来了巨大的压力。AI 工作负载要求低延迟、高吞吐量和无损通信。任何数据包丢失或延迟都可能导致整个训练过程停滞。因此，网络已经成为 AI 基础设施的关键性能组件。[参考文献 1]

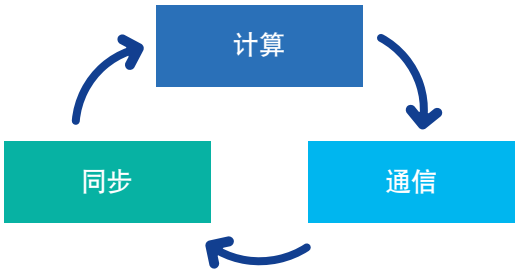
VIAVI 测试解决方案可帮助组织确保其 AI 数据中心网络经过优化、运行可靠，并具备扩展能力。

AI 工作负载的关键网络概念

AI 流量模式与集合通信库(CCL)

AI 工作负载依赖于一种独特的流量模式，专为分布式系统中的大容量并行数据处理而设计。训练现代 AI 模型涉及将大规模数据集分成块，并将它们分布在多个 xPU（例如，GPU 或自定义加速器）上进行同时处理。这些操作的结果必须在训练期间反复同步，需要精确的协调和超可靠的网络通信。

- 包含大量大象流的流量模式
- 数据和计算密集型工作负载
- 需要大量短小的远程内存访问操作
- 节点同时开始传输
- 任何一个流发生延迟，都会拖慢所有节点的处理进度
- AI 应用对网络弹性高度敏感



集合通信库是一个软件库，旨在促进并行和分布式计算环境中多个进程之间的高效数据交换和同步。它提供了集合通信操作的优化实现，使多组进程能够在通信任务上协同工作。

NVIDIA 的 NCCL 实现了针对 NVIDIA GPU 和网络优化的多 GPU 与多节点通信原语。NCCL 提供全收集、全归约、广播、归约、归约-分散以及点对点发送和接收等例程，这些例程经过优化，可通过节点内的 PCIe 和 NVLink 高速互连以及跨节点的 NVIDIA Mellanox 网络实现高带宽和低延迟。[参考文献 2]

NCCL 在深度学习框架中具有重要应用，其中 AllReduce（全归约）集合通信被广泛用于神经网络训练。通过 NCCL 提供的多 GPU 与多节点通信，可实现神经网络训练的高效扩展。

RingAllReduce

RingAllReduce（环状全归约）是一种分布式算法，主要用于深度学习，在分布式训练设置中跨多个设备（如 GPU 或节点）有效地平均梯度。设备排列在逻辑环中（可视为一个通信器）。每台设备只与其直接相邻的设备（环中的下一台和上一台设备）通信。

它分为两个阶段：ReduceScatter（聚合阶段）与 AllGather（分发阶段）。[参考文献 3]

ReduceScatter（归约分散）阶段：

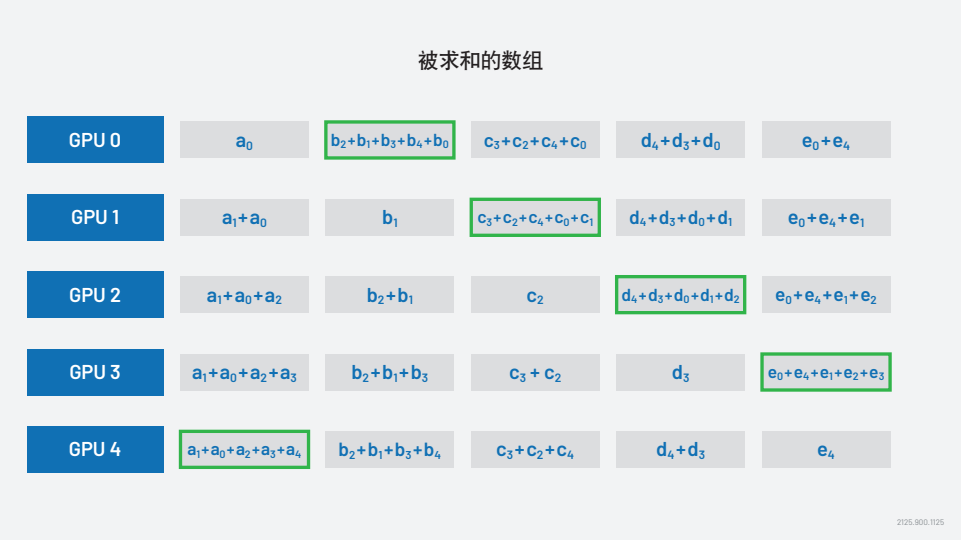
1. 总数据被分成 N 个块，其中 N 是设备（或进程）的数量。
2. 每台设备向下一台设备发送一个数据块，同时接收来自前一台设备的数据块。
3. 在接收到块时，每台设备执行将接收到的数据与其本地块相结合的归约操作。本文档将以求和操作为例进行说明。
4. 这个过程持续 N-1 个步骤，之后每个设备将持有归约结果的一个块。

ReduceScatter（归约分散）的第一步

在第一次发送和接收完成后，每个 GPU 将拥有一个块，该块由两个不同 GPU 上相应块的总和组成。例如，第二 GPU 上的第一个块将是来自第二 GPU 和第一 GPU 的该块中的值的总和。

ReduceScatter（归约分散）的第一次迭代完成后的中间和

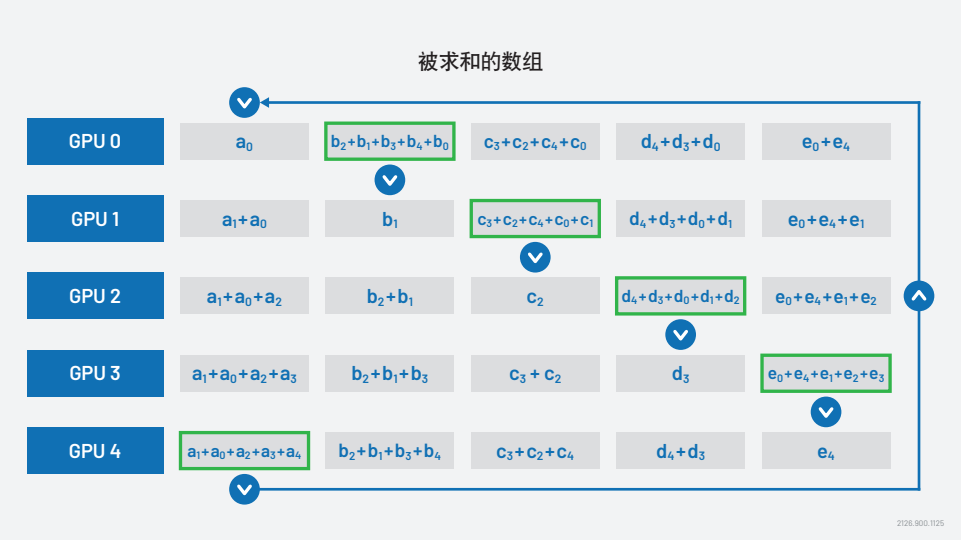
在 NCCL 运算参数 -o,—op <sum/prod/min/max/avg/all> 中，指定要执行的归约运算。仅与归约运算相关，如 AllReduce（全归约）、Reduce（归约）或 ReduceScatter（归约分散）。默认值：Sum [参考文献 4]



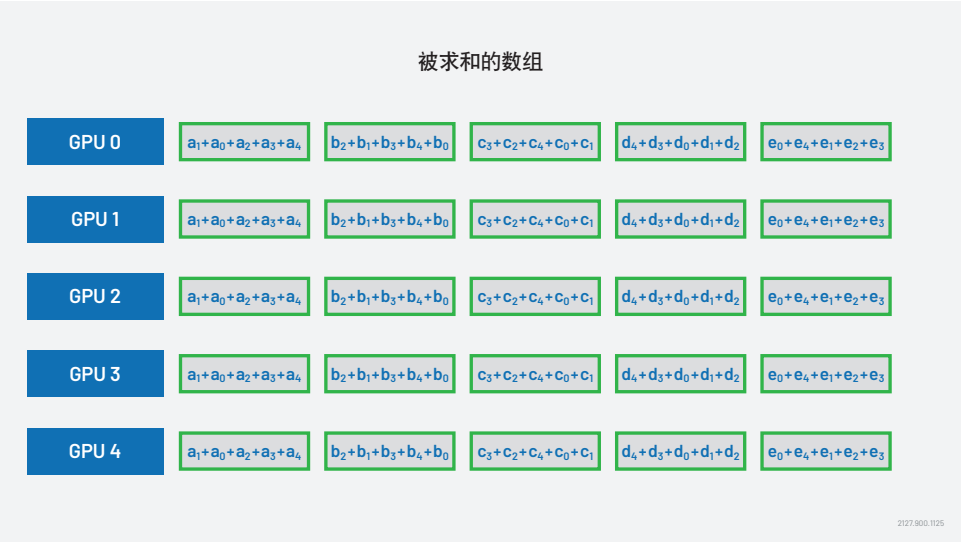
所有 ReduceScatter（归约分散）传输后的最终状态

AllGather（全收集）阶段：

1. 每台设备将其归约的块传递给下一台设备，同时从前一台设备接收另一个缺失的块。
2. 在 N - 1 个步骤之后，所有设备都将接收到完整的聚合数据。



AllGather（全收集）的第一步中的数据传输

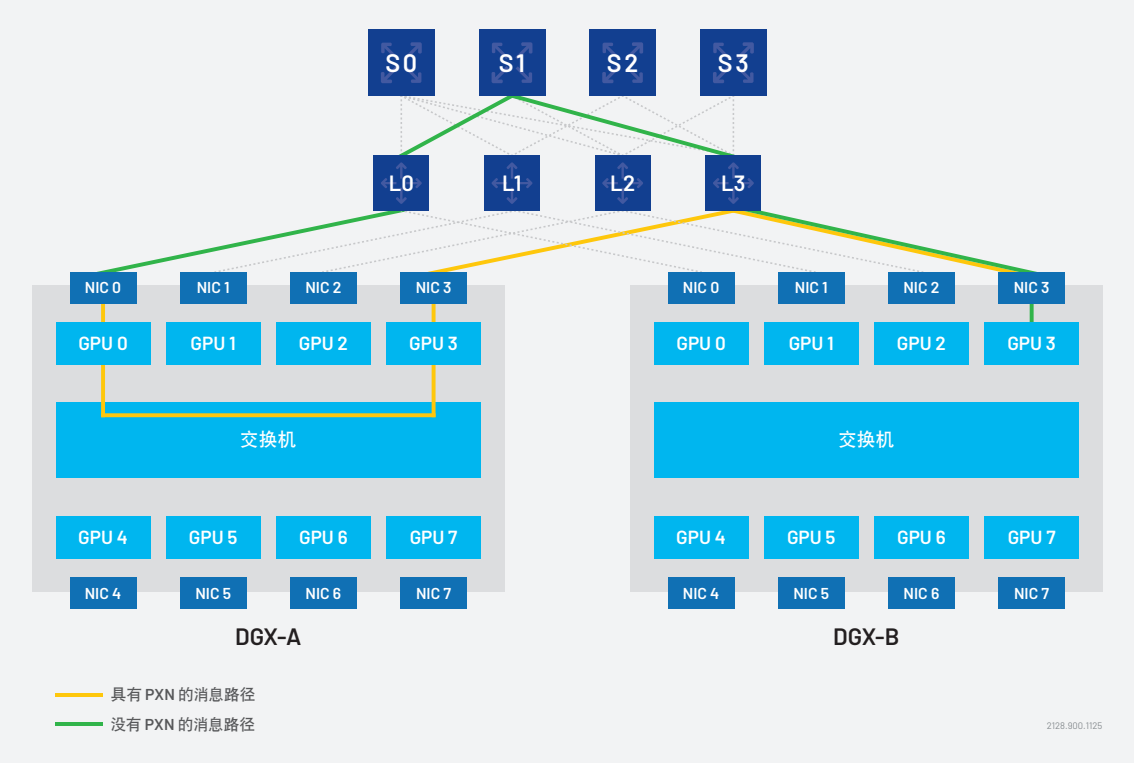


所有 AllGather（全收集）传输后的最终状态

在 AllGather（全收集）阶段结束时，每个 GPU 都将拥有整个数组的完整累积值。所有节点中的所有梯度都已同步。

AlltoAll

最具挑战性的通信模式之一是 **AlltoAll**（全对全），在这种模式下，集群中的每个处理器都要与其他所有处理器交换数据。这导致了极其密集的通信流，并对交换结构提出了很高的要求。全对全通信是一种通信模式，其中每个 xPU 与其他 xPU 交换数据。在 NCCL 2.12 中，NVIDIA 引入了一个名为 PXN 的新功能，该功能可以优化消息路径以提高效率。[参考文献 5]

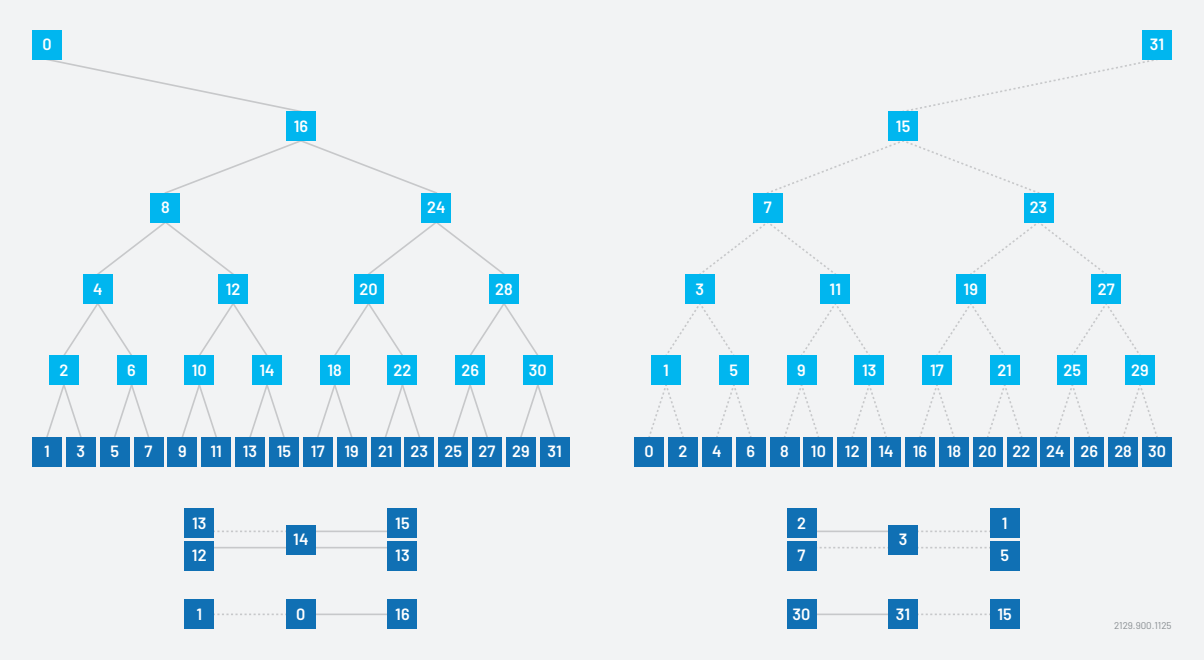


从 DGX-A 中的 GPU0 到 DGX-B 中的 GPU3 的 AlltoAll（全对全）示例消息路径

在 Rail 优化网络拓扑中，每个 DGX 系统的 NIC-0 连接到同一个叶交换机 (L0)，NIC-1 连接到同一个叶交换机 (L1)，依此类推。如果没有 PXN，上图中的消息将经过三跳网络交换机（L0、S1 和 L3），这可能会导致争用，并因其他流量的干扰而变慢。在同一对 NIC 之间传递的消息被聚合，以最大化有效消息速率和网络带宽。

双二叉树

NCCL 2.4 引入了双二叉树，它提供了满带宽和对数级延迟，甚至低于 2D 环的延迟。在双二叉树中，第一二叉树中的一半秩是节点，另一半秩是叶。第二个二叉树将其反转，使用叶作为节点，反之亦然，每个二叉树皆如此。该图说明了如何使用这种模式通过翻转树来反转节点和叶，从而构建双二叉树。[参考文献 6]



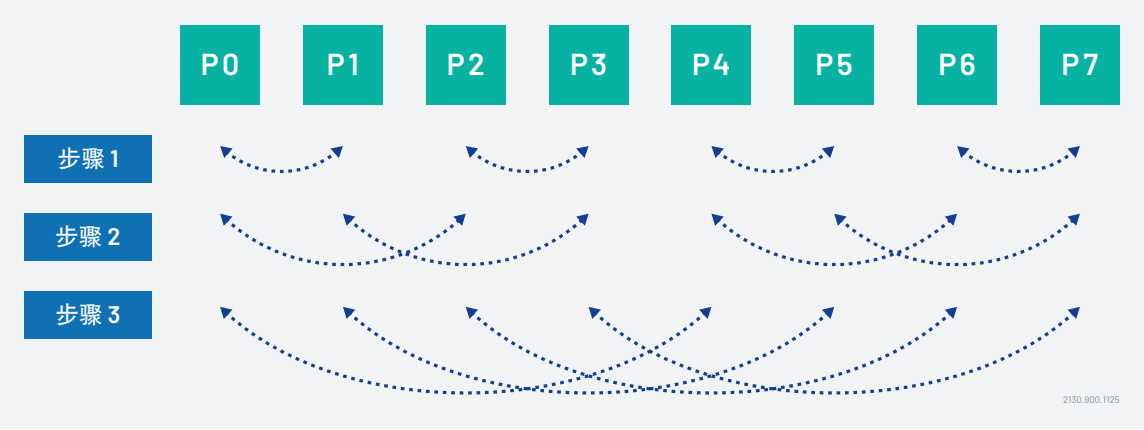
两个互补的二叉树

如果将这两棵树叠加起来，除了根秩具有一个父级和一个子级之外，所有秩都有两个父级和两个子级。如果我们用两棵树中的每一棵分别处理一半的数据，那么每个秩最多会接收和发送两次半量数据；在发送和接收的数据量方面，这与环算法一样最优。

Halving Doubling（折半倍增）

这种算法结合了 ReduceScatter（归约分散），通过递归向量折半和距离倍增实现；随后是 AllGather（全收集），通过递归向量倍增和递归距离折半来实现（用于 AllReduce（全归约））。[参考文献 7]

在第一步中，进程的数量 p 减少到 2 的幂值。前 2r 个进程成对执行，从每个偶数秩向奇数秩（秩 + 1）传递输入向量的后半部分，从每个奇数秩向偶数秩（秩 – 1）传递输入向量的前半部分。所有 2r 进程各自计算其对应一半部分的归约。第一步结束时，将每个奇数进程(1 ..... 2r - 1) 的结果发送至秩 - 1，放入缓冲区的第二部分。



ReduceScatter（归约分散）的递归折半与倍增

在第二步中，偶数秩/奇数秩的进程将其缓冲区的后半部分/前半部分发送到秩 +1/秩 -1。在接下来的步骤中，缓冲区递归折半，距离倍增。这个过程一直持续到 ReduceScatter（归约分散）阶段完成。

AllGather（全收集）是 ReduceScatter（归约分散）的逆过程，不涉及归约操作。

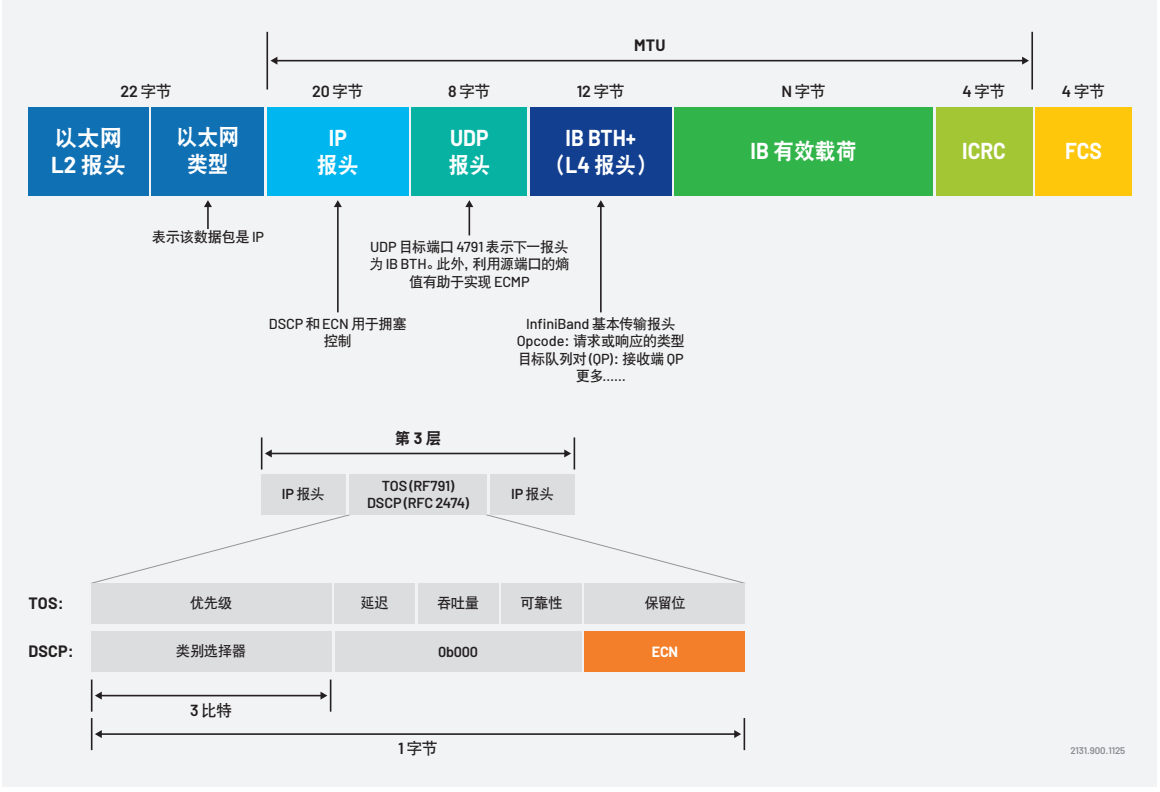
根据 xPU 的数量、向量长度和归约例程，在训练期间将调用一个或多个流量模式。TestCenter 可以跨多个端口模拟这些流量模式的组合。此外，非 AI 训练流量可以作为背景流量添加。

基于融合以太网的 RDMA 第 2 版 (RoCEv2)

由于 CPU 处理和内核交互，传统的网络堆栈会引入延迟。远程直接内存访问 (RDMA) 通过在没有 CPU 参与的情况下在不同节点的内存之间实现直接数据传输，消除了这些瓶颈。这对于 AI 工作负载尤其有利，因为同步事件必须频繁、快速且延迟最小。

RoCEv2（基于融合以太网的 RDMA 第 2 版）是数据中心最广泛采用的 RDMA 协议。它通过标准以太网运行，并支持使用 UDP 封装跨第 3 层网络进行路由。这种灵活性允许 AI 工作负载跨大型、多机架甚至多站点部署进行扩展。

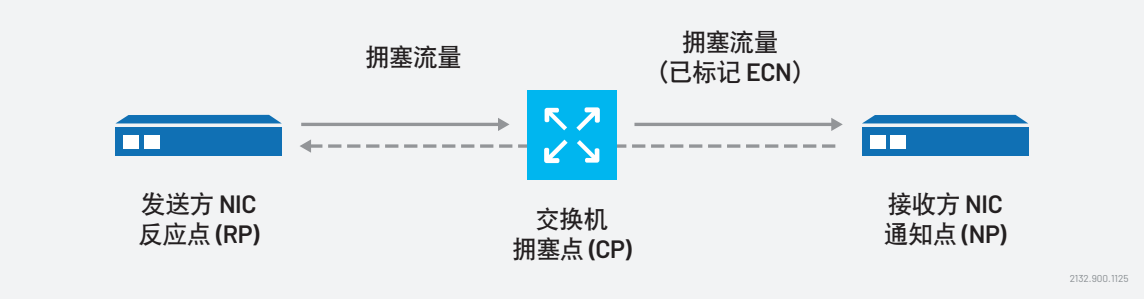
自 2014 年推出以来，RoCEv2 因其低延迟、高带宽的能力而被广泛应用于数据中心。它的拥塞控制机制利用 IP ECN（显式拥塞通知）位进行拥塞标记，利用拥塞通知数据包 (CNP) 进行确认和流量调节。[参考文献 8]



RoCEv2 数据帧格式

为了使 RoCEv2 有效运行，底层网络必须支持无损传输。这需要先进的拥塞管理协议，包括：

- **优先级流量控制 (PFC)：**当缓冲区已满时，通过暂停每个优先级类别的流量来防止第 2 层的数据包丢失。
- **显式拥塞通知 (ECN)：**标记经历拥塞的数据包，以便端点可以降低其传输速率。
- **数据中心量化拥塞通知 (DCQCN)：**一种拥塞控制算法，在拥塞时成倍地降低传输速率，在拥塞解除时逐渐提高速率。这有助于防止网络拥塞和数据包丢失，同时在基于 RDMA 的网络中保持高吞吐量和低延迟。
- **拥塞通知数据包 (CNP)：**DCQCN（数据中心量化拥塞通知）的一部分，用于通知发送方自适应降低速率。



DCQCN 拥塞控制流程

数据中心量化拥塞通知(DCQCN)

由交换机和 NIC 促成的关键拥塞控制机制是 DCQCN 和优先级流量控制(PFC)。在这些大规模环境中，不正确或未经优化的网络设置会导致应用程序性能低下。因此，在拥塞情况下，验证交换结构性能、优化配置并确保网络稳定性至关重要。

DCQCN 是专门为 RoCEv2 设计的拥塞控制算法。它有助于防止网络拥塞和数据包丢失，同时在基于 RDMA 的网络中保持高吞吐量和低延迟。当发送方收到 CNP 时，它会成倍地降低传输速率。如果一段时间内没有检测到拥塞，发送方会逐渐增加其速率。这确保了网络利用率保持较高，同时避免了拥塞。

作为接收方：

- 如果某个流的标记数据包到达，并且在最近 N 微秒内没有为该流发送 CNP，则立即发送 CNP。
- 如果在该时间窗口内到达的任一数据包被设置了 ECN 标志位，则接收方最多每 N 微秒（CNP 生成间隔）为该流生成一个 CNP 数据包 [参考文献 9]

作为发送方：

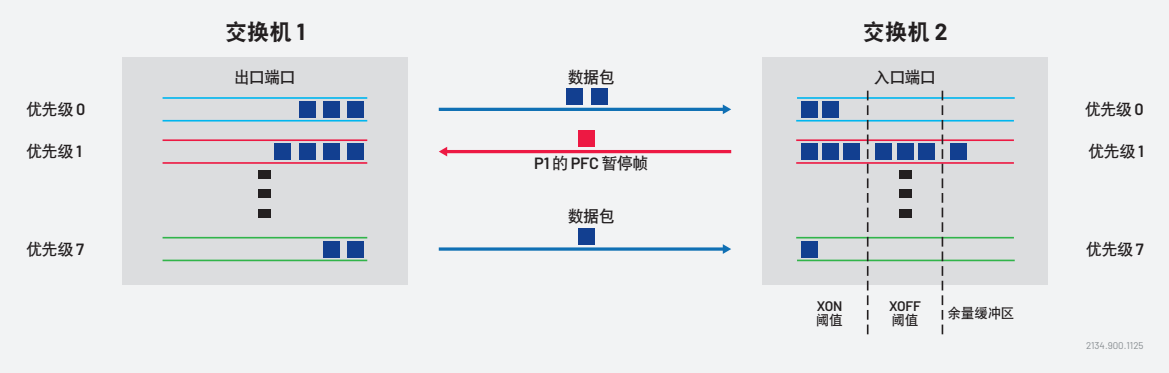
自该 QP 最近一次收到 RoCEv2 CNP 以来，当经过了可配置的时间和/或在该 QP 上已传输了可配置的字节数时，发送方应提高该 QP 上的注入速率。此外，它还维护一个名为 Alpha 的参数，该参数用于估计网络中的拥塞程度，并用于降低速率。

在速率降低阶段，如果收到该 QP 的 CNP，则 QP 速率降低。Alpha 用于控制下降速率，直到达到最小速率，公式如下：RC（当前速率）= RC(1 - Alpha/2)(0 < Alpha < 2)。如果连续收到 CNP，速率将降低到配置的最小速率。

当最后一次减少后达到时间复位或字节复位阈值时，系统进入增加阶段。速率将增加到目标速率。

优先级流量控制(PFC)

PFC 在第 2 层（数据链路层）运行，基于 IEEE 802.1Qbb。网络流量分为 8 个优先级(0-7)。每个优先级都可以独立控制。



PFC 操作

- 当交换机缓冲区接近溢出时（由 Xoff 阈值指示，表示特定优先级队列中的高缓冲区利用率），交换机发送 PFC 暂停帧，提醒上游端口暂停数据传输。
- 一旦缓冲区使用率低于 Xon 阈值，交换机就会提示上游端口恢复流量。[参考文献 10]

每个 PFC 暂停帧包括一个针对每个优先级的 2 字节定时器值，指示流量应该暂停的持续时间。定时器以暂停量为单位进行计量，其中一个暂停量表示在该端口速率下传输 512 位所需的时间，取值范围为 0 到 65535。Xon 帧对已启用优先级的 Time 值为 0，用于恢复该优先级的流量传输。

当第 2 层或第 3 层接口发生拥塞时，PFC 可用于防止流量丢失。在第 2 层，VLAN 报头中的优先级代码点(PCP)可以标识流量的优先级。流量的第 3 层 IP 报头中的区分服务代码点(DSCP)值被映射到第 2 层的 PCP 值。

PFC 和 DCQCN 的结合使用优化了 RDMA 性能。PFC 通过减慢发送速率来有效管理拥塞。DCQCN 通过向端点发送数据路径上任何地方的拥塞信号，有效地缓解了每个流的拥塞。DCQCN 可作为主要的拥塞管理机制，而 PFC 则充当故障安全解决方案。

常见 AI 测试挑战

AI 训练流量与传统网络流量有很大不同，带来了独特的挑战，需要专门的测试方法。一个突出的问题是 AI 工作负载产生的大量同步数据爆发。这些密集的数据流可能会使网络缓冲区和队列不堪重负。传统的测试方法通常基于企业或 Web 流量模式，无法充分模拟这些苛刻的条件。

另一个关键因素是 AI 集群中东西向流量的主导地位。与优先考虑南北向流量（客户端-服务器交互）的典型数据中心运营不同，AI 训练涉及 GPU 或 xPU 之间大量的横向数据移动。这种全网状通信对交换结构提出了很高的要求，要求它们支持跨多个节点的同步高速通信，而不会产生瓶颈。

通过优化控制协议（如优先级流量控制(PFC)和数据中心量化拥塞通知(DCQCN)）来有效管理拥塞也至关重要。拥塞管理配置不当可能导致数据包丢失、训练作业延迟或网络链路利用率不足。

服务质量(QoS)配置错误，如不正确的 VLAN 标记、不正确的队列映射或缓冲区分配不足，会悄无声息地降低性能，使故障排除变得特别困难。此外，AI 工作负载的规模和相互依赖性使根因定位变得更加复杂，因为问题通常会同时涉及多个组件和网络层。

通过精确的流量模拟和高级分析进行测试，帮助组织准确识别性能下降发生的位置和原因。

揭示 AI 网络健康状况的统计数据

监控 AI 网络结构的健康和性能需要捕获和分析具体的、可操作的指标。例如，数据包丢失表示训练期间所需的关键更新失败。作业完成时间(JCT)反映了训练工作负载在网络中传输的整体速度与效率。

尾部延迟代表最坏情况下的延迟，会显著降低训练作业的速度，而丢弃的数据包计数和重新排序的数据包等指标则突出了拥塞问题或等价多路径(ECMP)路由行为的问题。跟踪发送(Tx)和接收(Rx)速率的偏差，将识别未充分利用的链路和流量不平衡。

此外，对 ECN、CNP 和 PFC 活动的可见性提供了对流量控制操作的基本洞察，使网络团队能够在拥塞处理问题变成性能问题之前主动对其进行诊断。

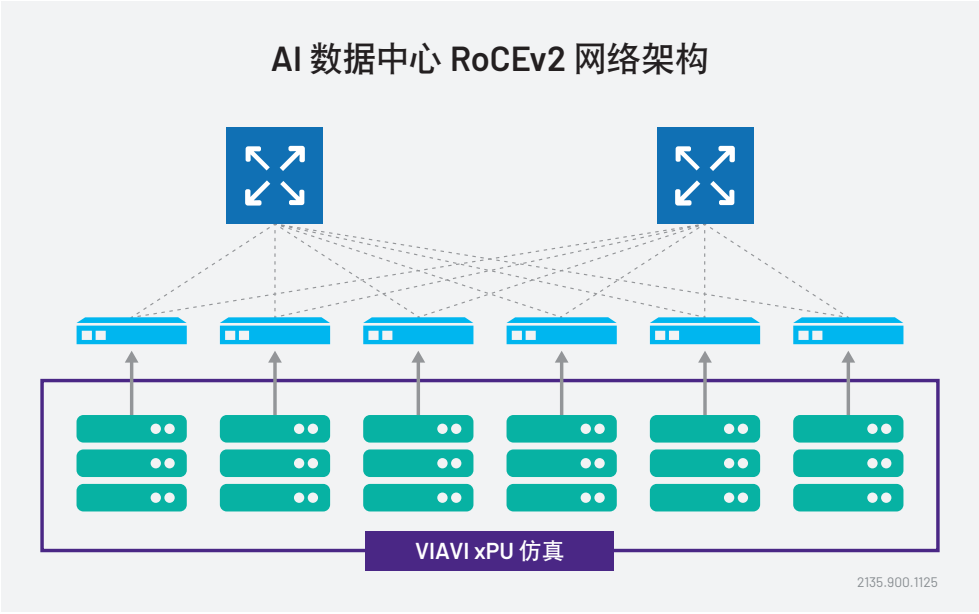
精确定位哪个队列对(QP)或端口在峰值负载期间遭受数据包丢失，隔离造成尾部延迟的有问题的链路，并验证拥塞管理机制是否按预期运行。通过尽早发现这些问题，VIAVI 确保网络问题不会影响实时 AI 训练操作。

问题指标示例

观测到的问题	可能的原因	测试洞察
高峰负载期间的数据包丢失	同步训练事件期间 PFC 阈值配置错误或交换机缓冲器溢出	查明受影响的队列对(QP)或交换机端口，并确定丢失开始时的负载级别
训练中的长尾部延迟	特定交换阶段的流路径不平衡或资源争用	揭示哪些链路或流被延迟，并将延迟与拓扑和配置相关联
高 JCT 方差	不一致的 ECN/CNP 响应或队列堆积	比较负载下的算法性能，跟踪 JCT 变化以及性能下降的迭代轮次
无速率下降的拥塞	ECMP 算法或网络拓扑需要优化	证在发生拥塞但未出现数据包丢失时，STC 是否会在拥塞期间降低流量速率

TestCenter AI 测试解决方案概述

TestCenter 为 AI 数据中心环境提供高密度、多速率的性能测试。它支持模拟真实的 AI 工作负载，包括基于 RDMA(RoCEv2)协议和集合通信库(CCL)模式（如 AlltoAll、RingAllReduce 等）的流量。这使得它非常适合验证必须支持同步、高带宽、低延迟 AI 训练通信的网络架构。



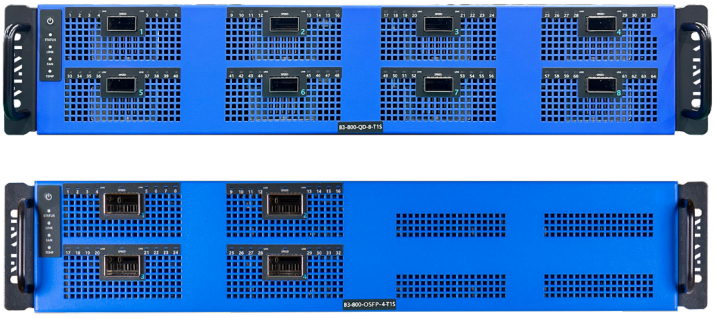
AI 数据中心 RoCEv2 网络架构

VIAVI 测试设备包括 A1 和 B3 系列高速以太网设备。

A1-400-QD-16 平台提供多达 16 个 400G 以太网端口，支持 100G、200G 和 400G 上的 RoCEv2，使其高度灵活，适合需要多用户、多速率环境的实验室。

A1 400G 16 端口设备

B3 平台支持 QSFP-DD 和 OSFP 800G 接口，提供高达 6.4 Tbps 的流量生成，具备业界领先的端口密度优势。



B3 800G 8 端口和 4 端口设备

通过内置对 RoCEv2 特性的支持，如 DCQCN、ECN 和 CNP，以及用于性能基准测试的自动化框架，VIAVI 使用户能够模拟跨叶主干拓扑的东西向 AI 流量模式，在动态拥塞和流量不平衡条件下对网络进行压力测试，并在高性能环境中验证 DCQCN 和 PFC 等流量控制机制。

TestCenter 的自动化工具允许对不同的帧大小、数据大小、端口速率和流量模式进行测试，并支持持续集成（CI/CD）工作流。该解决方案通过高级报告和交互式仪表板提供可操作的结果，使网络架构师能够微调设置、排查问题并验证 AI 工作负载的就绪性。

实现可靠、可扩展的 AI 数据中心运营

随着 AI 工作负载的不断扩大，底层网络架构的性能在确保及时、高效的模型训练和推理方面发挥着至关重要的作用。在分布式环境中，提供一致的吞吐量、低延迟和可靠的同步的压力尤其大，在这种环境中，即使很小的延迟也会导致显著的性能下降或资源利用不足。

为了应对这些挑战，组织需要反映 AI 流量独特需求的测试策略和工具，特别是集合通信模式、RoCEv2 传输行为以及拥塞和流量控制的影响。及早了解压力下的网络行为是优化系统设计、配置和部署的关键。

模拟特定于 AI 的流量模式并测量作业完成时间、数据包丢失和尾部延迟的能力，有助于工程师识别瓶颈、验证配置并优化性能调优。借助这些功能，团队可以在交换结构、拥塞管理机制和 QoS 策略影响实时训练操作之前，对其进行评估。

通过将流量仿真、性能基准测试和流级分析集成到开发和测试流程中，网络和基础设施团队能够更好地做出明智的决策，降低部署风险，并提供满足 AI 规模计算需求的基础设施。

要探索支持可扩展、高性能 AI 网络的测试策略，请访问 viavisolutions.com/zh-cn/node/123953。

参考文献

[参考文献 1] https://www.thefastmode.com/expert-opinion/39865-how-ai-changes-the-game-for-high-speed-ethernet

[参考文献 2] https://developer.nvidia.com/nccl

[参考文献 3] https://andrew.gibiansky.com/blog/machine-learning/baidu-allreduce/

[参考文献 4] https://github.com/nvidia/nccl-tests

[参考文献 5] https://developer.nvidia.com/blog/doubling-all2all-performance-with-nvidia-collective-communication-library-2-12/

[参考文献 6] https://developer.nvidia.com/blog/massively-scale-deep-learning-training-nccl-2-4/

[参考文献 7] Rolf Rabenseifner, Optimization of Collective Reduction Operations, International Conference on Computational Science, June 7–9, Krakow, Poland, LNCS, Springer-Verlag, 2004.

[参考文献 8] https://en.wikipedia.org/wiki/RDMA_over_Converged_Ethernet

[参考文献 9] https://medium.com/@ravikishorechitakani/optimizing-ai-ml-and-hpc-workloads-exploring-rdma-rocev2-for-high-performance-data-center-8d130cda74ae

[参考文献 10] https://medium.com/@ravikishorechitakani/optimizing-ai-ml-and-hpc-workloads-exploring-rdma-rocev2-for-high-performance-data-center-8d130cda74ae

北京 电话：+8610 8233 0055

上海 电话：+8621 6859 5260

上海 电话：+8621 2028 3588
（仅限 TeraVM 及 TM-500 产品查询）

深圳 电话：+86 755 8869 6800

网站：www.viavisolutions.cn

aiml-dc-wp-hse-nse-zh-cn
30194901 900 1225

viavisolutions.cn