



开放数据中心委员会
Open Data Center Committee

[编号 ODCC-2021-05001]

数据中心智能无损网络 白皮书

开放数据中心委员会
2021-09-15 发布

目 录

前 言	iii
版权说明	iv
数据中心智能无损网络白皮书	1
1. 介绍	1
1.1. 范围	1
1.2. 目的	1
2. 让数据中心焕发生机	1
2.1. 一个到处都是数据的新世界	1
3. 数据中心需求和技术不断提升	3
3.1. 原有数据中心桥接标准	3
3.2. 需求演化	4
3.3. AI 计算的特点	5
3.3.1. 模型并行计算	6
3.3.2. 数据并行计算	6
3.4. 技术演进	8
3.4.1. SSDs 和 NVMeoF: 高吞吐量低时延网络	8
3.4.2. GPU: 用于并行计算的超低时延网络	11
3.4.3. SmartNICs	12
3.4.4. 远程直接内存访问 (RDMA)	14

3.4.5. GPU DirectRDMA	16
4. 当今数据中心网络面临的挑战	19
4.1. 平衡高吞吐量和低时延	19
4.2. 无死锁无损网络	21
4.3. 大规模数据中心网络的拥塞控制问题	23
4.4. 拥塞控制算法的配置复杂性	26
4.4.1. 自适应 PFC Headroom 计算	26
4.4.2. 动态 ECN 阈值设置	27
5. 解决新数据中心问题的新技术	28
5.1. 低时延和高吞吐量的混合传输	28
5.2. 基于拓扑识别的 PFC 死锁预防	30
5.3. 改善拥塞的通知	32
5.3.1. 反应点 (RP)	33
5.3.2. 阻塞点 (CP)	33
5.3.3. 通知点 (NP)	33
5.4. 解决拥塞控制算法的配置复杂性	35
5.4.1. 优化缓存区以降低 PFC headroom 配置的复杂性 ..	35
5.4.2. 智能 ECN 阈值优化	35
6. 结论	37

前 言

由中国信通院云大所、百度、腾讯、美团、京东、移动、电信、华为、思科、博通、英伟达等 ODCC（开放数据中心委员会）成员单位联合编制的《智能无损数据中心网络白皮书》正式发布。该白皮书内容翔实，分别从数据中心的重要性、应用发展需求、网络面临的挑战、相应的解决方案和标准化工作进展等方面开展了介绍。

2017 年起，ODCC 牵头制定**无损网络技术标准以及测试规范**等，相继发布行业标准、技术报告等 10 多项成果，得到了产业界的广泛参与，搭建起一个 **DCN 技术** 热点讨论平台，相关技术的标准化推动工作也在紧锣密鼓进行当中。在国家高度重视新基建的环境下，白皮书发布为我国数据中心技术、产品和服务走出去打下了良好的基础。

起草单位：中国信息通信研究院（云计算与大数据研究所）、百度在线网络技术（北京）有限公司、中国移动通信集团有限公司、中国电信集团有限公司、深圳市腾讯计算机系统有限公司、华为技术有限公司、NVIDIA(英伟达)中国有限公司、思科(中国)有限公司、博通公司、北京三快在线科技有限公司、北京京东世纪贸易有限公司

起草者：郭亮、李洁、高峰、顾戎、赵继壮、程传胜、殷悦、宋庆春、刘军、何宗应、孙黎阳、唐广明、权皓、陶春雷、王少鹏、赵精华

版权说明

ODCC（开放数据中心委员会）发布的各项成果，受《著作权法》保护，编制单位共同享有著作权。

转载、摘编或利用其它方式使用 ODCC 成果中的文字或者观点的，应注明来源：“开放数据中心委员会”。

对于未经著作权人书面同意而实施的剽窃、复制、修改、销售、改编、汇编和翻译出版等侵权行为，ODCC 及有关单位将追究其法律责任，感谢各单位的配合与支持。



数据中心智能无损网络白皮书

1. 介绍

1.1. 范围

白皮书研究了支持现代数据中心网络需求的网络技术，包括高性能计算和人工智能应用，提出了需求演变和新时代技术挑战的解决方案。

1.2. 目的

白皮书旨在为现代数据中心网络存在的问题和面临的挑战，提供高层次解决方案。白皮书梳理了数据中心的建设现状和技术演进，介绍了数据中心发展过程中面临的问题，并基于分析研究，提出增强数据中心网络能力和运营效率的技术解决方案，契合持续变化的应用需求。

2. 让数据中心焕发生机

2.1. 一个到处都是数据的新世界

数字化转型正在改变着我们的个人生活和职业生活。工作流程和人际交往正转向基于云、移动设备和物联网的数字化流程和自动化工具。支撑数字化转型的技术是人工智能（AI）。数据中心在运行拥有海量数据的人工智能应用程序时，要将这些数据重新转换为相关性信息、自动化人工交互和细致化决策制定（如图1）。在增强现实、语音识别和上下文搜索需求强劲的当今世界，满足数据中心实时交互需求比以往任何时候都更加重要。为满足实时需求，数据中心网络必须具备更强大的性能、规模和可靠性。

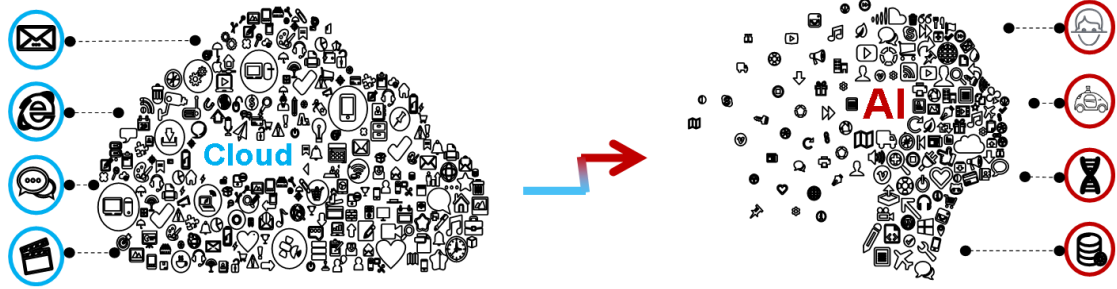


图 1 AI 时代的数字化转型

云时代的数据中心专注于应用转型和服务的快速部署。在 AI 时代，数据中心提供了实现数字化生活所需的信息和算法。高速存储和人工智能分布式计算的结合，将大数据转化为快速数据，供人、机、物访问。高性能、大规模、无丢包的数据中心网络对数字转换的顺利进行至关重要。

人工智能、网络性能等高性能应用的关键指标包括吞吐量、时延和拥塞。吞吐量是指快速传输大量数据的网络总容量。时延是指跨数据中心网络事务的总延迟。当流量超过网络容量时，会发生拥塞。丢包是严重影响吞吐量和时延的因素。

当前，各行业正在加速数字化转型。据估计，有 64% 的企业已经成为数字转型的探索者和实践者¹。在 2000 家跨国公司中，67% 的 CEO 将数字化作为企业战略的核心²。现实世界中的数字化转型趋势正在引领数据中心网络支持“以数据为中心”的计算模式。

数字化过程中产生的大量数据成为核心资产，人工智能应用也随之出现。根据华为全球产业展望的预测，到 2025 年，新增数据量将达到 180ZB³。然而，数据并不是“自我终结”。从数据中获取的知识和智慧拥有永恒价值。非结构化数据（如原始语音、视频、图像数据）的比例不断增加，未来将占到所有数据的 95%（如图 2）。现有的大数据分析无法适应数据的快速增长，需要进行性能优化，从原始数据中挖掘更多价值。基于深度学习的人工智能方法可以过滤掉大量

¹ Orange, “Finding the competitive edge with digital transformation,” 03 June 2015. [Online]. Available: <https://www.orange-business.com/en/magazine/finding-the-competitive-edge-with-digital-transformation>.

² Wiles, J., “Mobilize Every Function in the Organization for Digitalization,” Gartner, 03 December 2018. [Online]. Available: <https://www.gartner.com/smarterwithgartner/mobilize-every-function-in-the-organization-for-digitalization/>. [Accessed 10 June 2020].

³ Huawei, “Touching an Intelligent World,” Huawei, 2019. [Online]. Available: https://www.huawei.com/minisite/giv/Files/whitepaper_en_2019.pdf. [Accessed 15 March 2021].

无效数据，并自动提取有用信息，提供更有效的决策建议和行为指导。

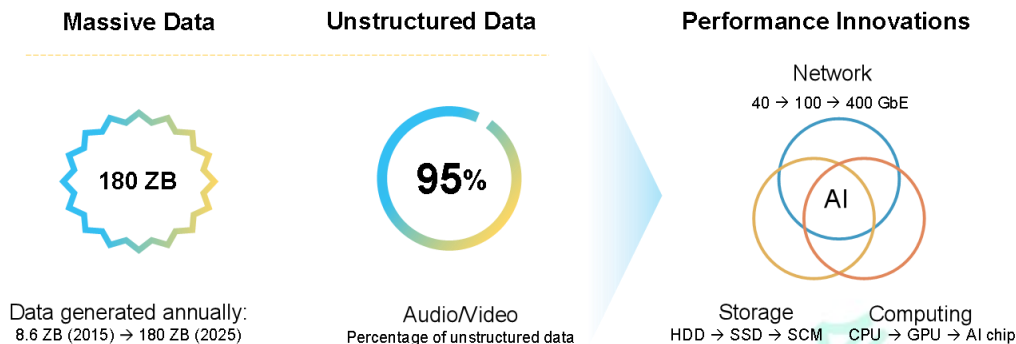


图 2 新兴的 AI 应用

总体来说，云数据中心架构提高了应用性能，扩大了应用规模。云平台允许 IT 资源快速分配，创建以应用程序为中心的服务模型。在 AI 时代，应用程序需要消耗前所未有的数据量，必要的性能创新增强了云数据中心架构的负载处理能力。在现有的云数据中，很难实现性能创新和新人工智能应用程序的无缝衔接。知道如何根据人工智能应用的需求实现数据有效处理，至关重要。实现成功的关键因素是有机结合应用程序的存储资源和计算资源之间的数据流。

3. 数据中心需求和技术不断提升

3.1. 原有数据中心桥接标准

在 10Gbps 以太网的早期，ODCC 工作组就开始关注数据中心桥接（DCB）。DCB 任务组针对数据中心环境中所用的以太网、网桥和相关协议定义了一组增强功能。使用案例和重点应用是集群和存储区域网络，使用了传统的专用技术，如 InfiniBand™和光纤通道⁴。以太网的重要目标是消除拥塞造成的损失，并在链路上为特定流量分配带宽。数据中心桥接的关键参数包括：

- **优先级流量控制（PFC）**：一种链路层流量控制机制，消除了数据包丢失风险，可以独立应用于各种流量。
- **增强型传输选择（ETS）**：一种队列调度算法，允许流量带宽分配。

⁴ InfiniBand 是 InfiniBand®贸易协会的商标和服务标志。

- **拥塞通知：**一种检测拥塞的二层端到端拥塞管理协议，通过跨二层网络的信号来限制发送端的传输速率，避免丢包。
- **数据中心桥接能力交换协议（DCBX）：**一个识别和性能交换协议，与链路层发现协议（LLDP）共同作用，用于传输上述参数的功能和配置。

这些参数对于将以太网扩展到集群计算和存储区域网络的专业市场非常重要。然而，随着环境和技术的变化，还需要不断优化。目前，使用三层协议和高度协调管理系统的数据中心已经实现规模部署。以太网链路已经从 10Gbps 提高到 400 Gbps，并计划将速度提高到 Tbps 范围。人工智能等新应用程序的出现，对基础设施提出了新的要求，推动了体系结构变化。为进一步扩大以太网在现代数据中心中的应用范围，还需要继续创新。

3.2. 需求演化

人工智能应用给数据中心网络带来了压力。自动驾驶汽车的人工智能训练就是一个例子。深度学习算法严重依赖海量数据和高性能计算技术。每天收集的训练数据接近 PB 级（1PB=1024TB），如果使用传统硬盘存储和普通 CPU 来处理数据，可能至少需要一年才能完成训练。这显然是不切实际的。为了提高人工智能的数据处理效率，需要在存储和计算领域进行革命性的变革。例如，存储性能需要提高一个数量级才能实现每秒 100 万次以上的输入/输出操作（IOPS）⁵。

为了满足实时数据的访问要求，存储介质已经从硬盘驱动器（HDD）发展到固态驱动器（SSD），再到存储类内存（SCMs），存储介质延迟缩短了 1000 倍以上。如果在网络延迟方面没有类似的改进，这些存储优化就无法实现，只能简单地将瓶颈从介质转移到网络上。对于网络固态硬盘（SSD），通信时延占端到端存储总时延的 60%以上。如果转向存储类内存（SCMs），除非网络性能得到改善，否则这一比例可能会增加到 85%。这就造成了存储介质有一半以上的时间处于闲置状态。同时优化存储媒介和 AI 计算处理器，会使得通信时延占总时延的 50%

⁵ Handy, J. and T. Coughlin, "Survey: Users Share Their Storage," 12 2014. [Online]. Available: <https://www.snia.org/sites/default/files/SNIA%20IOPS%20Survey%20White%20Paper.pdf>. [Accessed 14 May 2020].

以上，限制技术进步，造成资源浪费⁶。

人工智能应用程序和应用场景的范围和复杂性持续增加。例如 2015 年微软的 Resnet 实现 7 百亿亿次计算，有 6000 万个参数。2016 年百度在训练深度语音系统时，实现 20 百亿亿次计算和 3 亿个参数。2017 年谷歌 NMT 实现 105 百亿亿次计算和 87 亿个参数⁷。AI 计算的新特性对数据中心网络的发展提出更高要求。

传统协议已经不能满足日常生活中新应用程序的服务需求。举个简单的例子，美团线上外卖业务增长在过去 4 年里大约增长了 5 倍⁸。仅在用餐高峰期的几个小时里，交易量就从 21.49 亿增加到 123.6 亿。美团智能调度系统为用户、商家和超过 60 万名外卖员设计了一个复杂的多人多点实时决策过程。该系统每天更新 50 亿次定位数据，这些数据为外卖员计算可选路径并在 0.55 毫秒内选择最佳路线。当后端服务器使用 TCP/IP 协议时，内核缓存区、应用缓存区和网卡缓存区之间的数据量副本使得 CPU 和内存总线资源紧张，导致延迟增加，无法满足应用程序的需求。新远程直接内存访问（Remote Direct Memory Access, RDMA）协议消除了数据副本，释放了 CPU 资源，能够完成路径选择和取出顺序计算。RDMA 效率的提高给网络带来了更大的压力，将瓶颈转移到数据中心网络基础设施上，低时延和无损行为成为了新的必要需求。

3.3. AI 计算的特点

传统的数据中心服务（web、数据库和文件存储）是以事件为基础，计算结果通常是确定的。对于这样的任务，单个事件和相应网络通信之间几乎没有相关性或依赖性。传统事件的发生和持续时间是随机的。然而，AI 计算并非如此。这是一个迭代收敛的优化问题。它导致数据集和计算算法之间存在高度的空间相关

⁶ Huawei, “AI, This Is the Intelligent and Lossless Data Center Network You Want!” 13 March 2019. [Online]. Available: <https://www.cio.com/article/3347337/ai-this-is-the-intelligent-and-lossless-data-center-network-you-want.html>. [Accessed 14 May 2020].

⁷ Karuppiah, E. K., “Real World Problem Simplification Using Deep Learning / AI,” 2 November 2017. [Online]. Available: https://www.fujitsu.com/sg/Images/8.3.2%20FAC2017Track3_EttikanKaruppiah_RealWorldProblemSimplificationUsingDeepLearningAI%20.pdf. [Accessed 14 May 2020].

⁸ Yanqin, D., “The “Ultra Brain” weapon behind Meituan's delivery of 30 million orders in a single day,” 19 September 2019. [Online]. Available: <https://www.infoq.cn/article/2Kt8Ru9oD75idBNHKBrp>. [Accessed 15 March 2021].

性，在通信流之间形成时间相关性。

AI 计算用于大数据，要求快数据。为了满足这一点，它必须与“分而治之”的问题并行运作。计算模型和输入数据集较大（例如 100MB 节点条件下，10K 规则的 AI 模型需要超过 4TB 的内存）。单个服务器无法提供足够的存储容量和处理资源，使得问题无法按顺序解决。需要 AI 计算和存储节点并行，缩短处理时间。这种分布式 AI 计算和存储要求需要快速、高效和无损的数据中心网络，该网络构建起两种不同的并行计算模式——模型并行计算和数据并行计算。

3.3.1. 模型并行计算

模型并行计算中，每个节点承担了整个算法的一部分计算。每个节点处理相同的数据集，不同的算法部分，完成了对不同参数集的估计。通过节点交换算法估计，得到收敛于所有数据参数的最佳估计。模型并行计算最初是将公共数据集分布到分布式节点，然后把来自每个分布式节点的单个参数进行集合。图 3 显示了在并行操作模式下，整个模型的参数如何分布在计算节点上。

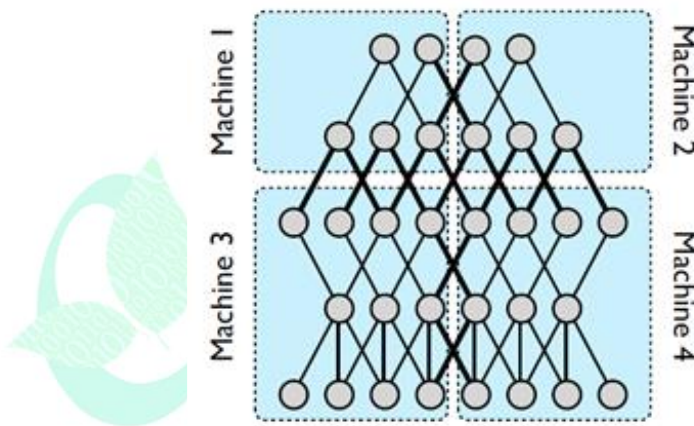


图 3 并行计算模型⁹

3.3.2. 数据并行计算

在数据并行计算中，每个节点都承载了整个 AI 算法模型，但只处理部分输

⁹ Dean, Jeffrey, Greg S. Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc' Aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Andrew Y. Ng, Large Scale Distributed Deep Networks, Google Inc., Mountain View, CA. Available: <https://storage.googleapis.com/pub-tools-public-publication-data/pdf/40565.pdf>. [Accessed 19 May 2021].

入数据。每个节点都试图使用不同的数据视图来估计相同的参数集。当一个节点完成一轮计算时，由公共参数服务器加权并聚合参数，如图 4 所示。更新加权参数要求所有节点同步更新信息。

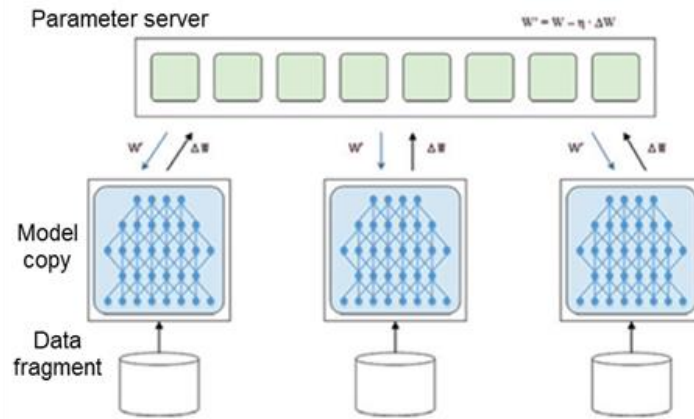


图 4 数据并行计算⁹

无论采用哪种并行计算方法，数据中心网络都要承受更大的通信压力。当网络成为瓶颈时，计算资源的等待时间会超过工作完成时间的 50%¹⁰。

对于所有的 AI 应用程序，计算模型都在不断迭代，且存在一个会造成网络 incast 阻塞的同步步骤。图 5 显示了 AI 训练中发生 incast 堵塞的方式。训练过程在不断迭代，在每次迭代都会产生很多同步参数。应用程序在下载模型时会同步将下一次计算得到的结果 (ΔM) 上传到参数服务器。上传到参数服务器中的过程会造成 incast。应用新兴计算技术能够缩短计算时间，但网络压力和由此产生的 incast 也会随之增加。

¹⁰ Cardona, O., "Towards Hyperscale High Performance Computing with RDMA," 12 June 2019. [Online]. Available: https://pc.nanog.org/static/published/meetings/NANOG76/1999/20190612_Cardona_Towards_Hyperscale_High_v1.pdf. [Accessed 14 May 2020].

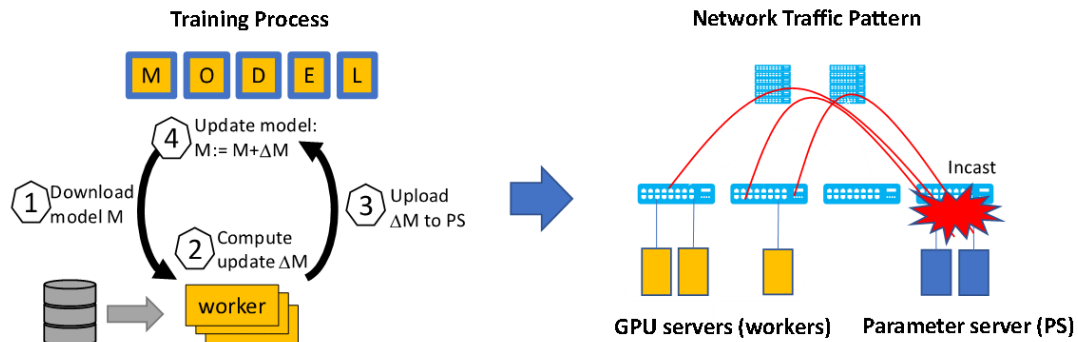


图 5 训练期间的周期性 incast 拥塞

工作节点和参数服务器间的通信构成了相互依赖的网络流集。分布式 AI 计算的迭代过程中，大量突发流量会在几毫秒内将数据分配到工作节点，当传递和更新中间参数时，发送到参数服务器的小规模流量会发生 incast。在这些流交换的过程中，网络可能会出现丢包、拥塞、负载失衡等问题。因此，一些流的流完成时间（FCT）被延长。如果有一些流发生延迟，可能会导致存储和计算资源无法得到充分利用。进而延迟了整个应用程序的完成时间。

分布式 AI 计算具有同步性，在理想情况下，可以预测到计算完成时间。当没有拥塞时，低网络动态时延使得平均 FCT 是可预测的，因此，整个应用程序的性能也可以被预测。当拥塞导致的动态延迟增加到丢包临界点时，就无法预测 FCT 了。完成时间远远大于平均完成时间的流，会发生所谓的尾部时延。系统对输入/输出（I/O）请求的全部响应中，尾部时延仅占系统响应时间的一小部分，与大部分响应时间相比，它花费的时间最长。尽可能缩短尾部延迟对于并行算法和整个分布式计算系统的成功至关重要。为了最大限度地利用数据中心中的计算资源，尾部时延需要被解决。

3.4. 技术演进

当不断变化的需求和技术相协调时，即意味着进步。新需求通常会驱动新技术研发，而新技术能支撑新用例，这些用例又会促成新需求。网络存储、分布式计算、系统架构和网络协议的突破推动下一代数据中心的发展。

3.4.1. SSDs 和 NVMeoF：高吞吐量低时延网络

在网络存储中，一个文件被分发到多个存储服务器，实现输入/输出的加速和冗余。当数据中心应用程序读取文件时，它会同时从不同的服务器访问数据的不同部分。数据几乎同时通过数据中心交换机进行聚合。数据中心应用程序写入文件时，数据写入会在分布式存储节点和冗余存储节点之间触发一系列存储事务。图 6 显示了由网络存储服务模型触发的数据中心通信示例。

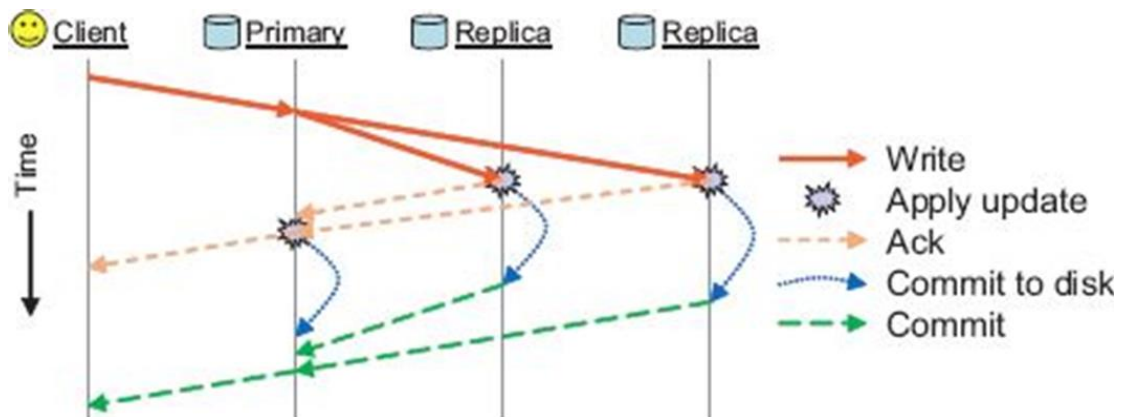


图 6 网络存储服务模型

该示例强调了网络同时支持高吞吐量和低时延的重要性。写入主存储服务器的大量数据会分多次传输到副本。小规模的确认真和提交消息必须进行排序，并在事务完成之前传递给发起客户端，说明了超低时延的必要性。

随着使用非易失性存储器高速(NVMe)接口规范的技术从 HDD 发展到 SSD，再发展到 SCM，存储性能得到了巨大提升。通过 NVMe 访问存储介质所花时间相比以前的硬盘技术减少了 1000 倍。不同技术之间的样本搜索时间分别为：HDD= 2-5 毫秒，SATA SSD = 0.2 毫秒，NVMe SSD = 0.02 毫秒。SCM 通常比 NVMe 闪存 SSD 快三到五倍。

NVMe-over-fabrics（简称 NVMeoF）是指用于网络存储的 NVMe 配置。介质的访问速度越快，网络瓶颈越大，网络时延的影响也越显著。图 7 展示了网络时延如何成为更快 NVMe 存储的主要瓶颈。网络时延是端到端网络硬盘存储时延中可以忽略的一部分，但随着网络化 SCM 存储的发展，网络时延将成为一个重要的组成部分。为了最大化新介质的 IOPS 性能，首先必须解决网络时延问题。

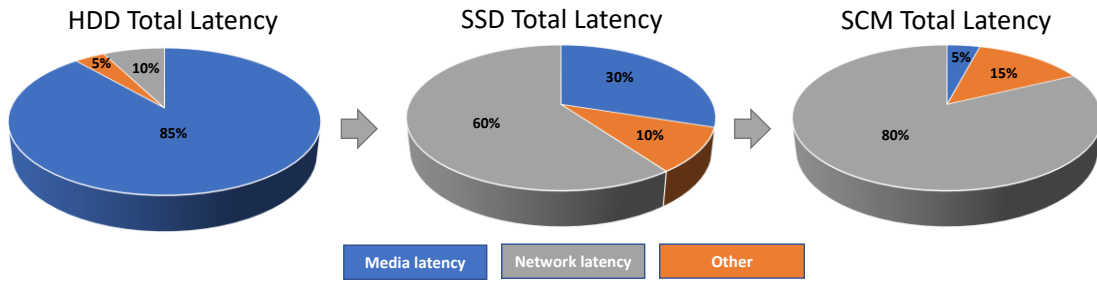


图 7 HDD 和 SSD 的端到端时延故障

时延由静态时延和动态时延两种类型组成。静态时延包括串行数据时延、设备转发时延和光/电传输时延。这种时延类型取决于交换硬件的性能和数据传输的距离。它通常是固定的，而且很容易预测。图 8 显示，当前静态时延的行业测量值通常为纳秒（ 10^{-9} 秒）或亚微秒（ 10^{-6} 秒）级别，在端到端网络总时延中占比不到 1%。

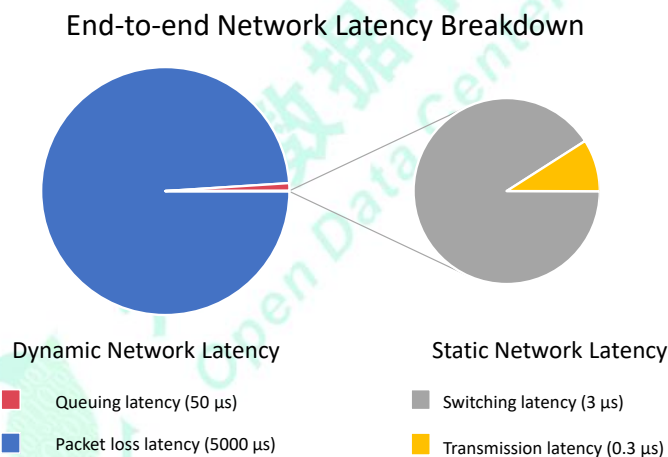


图 8 网络时延故障

动态时延对端到端网络总时延的作用更强，并且受通信环境条件的影响更多。动态时延是由于内部排队和数据包重传引起，这些时延的原因是网络拥塞和数据包丢失。并行 AI 计算模型会形成独特的流量模式，造成严重的网络拥塞。降低端到端网络时延的关键是解决动态时延，而解决动态时延的关键是缓解拥塞。

动态时延的主要来源是丢包导致的数据包重传。丢包时延比排队时延大一个数量级，并对应用程序有严重影响。当交换机缓存区因拥塞而溢出时，就会发生丢包（需要注意的是，此处忽略传输过程中由于低概率比特错误而造成的丢包）。

导致丢包的两种主要拥塞分别是网络内拥塞和 incast 拥塞。当网络结构中的链路过载时，交换机之间的链路就会发生网络内拥塞，这可能是由于负载不平衡。当许多源同时向相同目的地发送数据时，网络边缘就会出现 incast 拥塞。AI 计算模型本身所具有的阶段，即在处理迭代之后对数据进行聚合，很容易发生 incast 拥塞（多打一）。

3.4.2. GPU：用于并行计算的超低时延网络

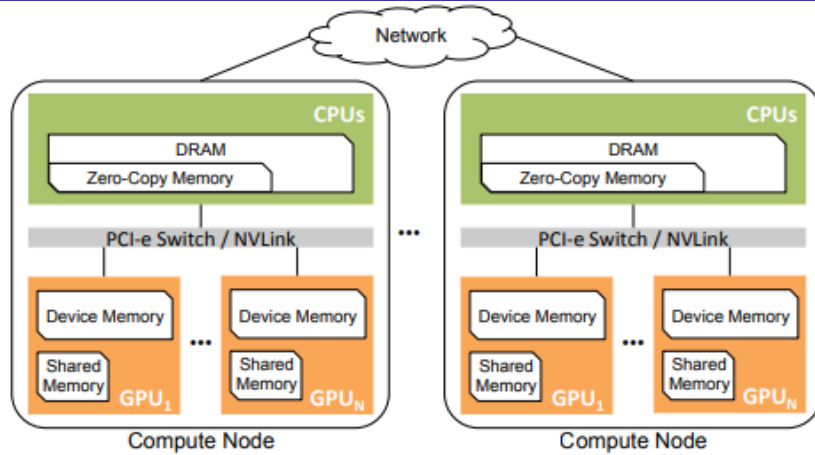
今天的 AI 计算架构包括中央处理器（CPU）和图形处理器（GPU）。GPU 最初是为了高速渲染电子游戏而发明的，现在在数据中心有了新的用途。GPU 是一个拥有数千内核的处理器，能够同时执行数百万次数学运算。所有的人工智能学习算法都能进行复杂的统计计算，并且可以处理大量的矩阵乘法运算——这非常适用于 GPU。然而，要扩展 AI 计算架构以满足当前数据中心对 AI 应用程序的需求，GPU 必须是分布式和网络化的。这就对通信量和性能提出了更高要求。

Facebook 最近测试了分布式机器学习平台 Caffe2，这个平台为实现并行加速，使用了最新的多 GPU 服务器。测试时发现，8 台服务器的计算任务导致 100Gbit/s InfiniBand 网络的资源未得到充分利用。网络和网络争用的出现使解决方案的性能降低到线性范围以下¹¹。因此，网络性能极大地限制了人工智能系统的横向扩展。

GPU 提供了比现在的 CPU 架构高得多的内存带宽。多节点 GPU 由于其高能效和硬件并行性，被用于高性能计算。图 9 展示了多 GPU 节点架构，每个节点由一台主机（CPU）和多个 GPU 设备组成，这些设备通过 PCI-e 交换机或 NVLink 连接。每个 GPU 都能够直接访问其本地相对较大的设备内存、更小更快的共享内存，以及主机节点 DRAM 的一小块固定区域，即零拷贝内存¹²。

¹¹ Morgan, T. P., “Machine Learning Gets an Infiniband Boost with Caffe2,” 19 April 2017. [Online]. Available: <https://www.nextplatform.com/2017/04/19/machine-learning-gets-infiniband-boost-caffe2/>. [Accessed 14 May 2020].

¹² Jai, Z., Y. Kwon, G. Shipman, P. McCormick, M. Erez and A. Aiken, “A distributed multi-GPU system for fast graph processing,” in VLDB Endowment, 2017.

图 9 分布式 AI 计算架构¹²

3.4.3. SmartNICs

在过去的几年里，CPU 的速度提高和以太网链路的性能优化已经相互抵消。图 10 展示了以太网链路的历史速度提升情况和 CPU 性能的基准提升¹³。在过去的某些时期，传统 CPU 的处理能力足以承载以太网链路的负载，而简化的 NIC 可以节约成本，还可以在软件中灵活处理整个网络堆栈，优势明显。而在其他时期，处理器无法适应链路速度提升，因此在使用以太网链路时，需要使用更昂贵、更复杂的 SmartNIC 和专业可卸载硬件。随着时间的推移，SmartNIC 卸载逐渐成熟，其中一些特性已经成为标准，并涵盖在现在通用的 NIC 基本特性中。这种现象随着 TCP 卸载引擎（TOE）的出现而出现，TOE 支持 TCP 校验和卸载、大数据段发送和接收端扩展。

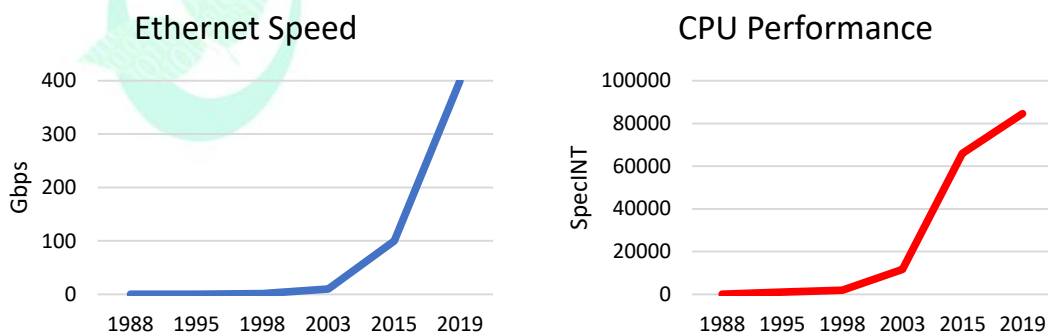


图 10 历史性能比较

¹³ Rupp, K., “42 Years of Microprocessor Trend Data,” February 2018. [Online]. Available: <https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/>. [Accessed 22 July 2020].

当今世界,有迹象表明摩尔定律正在消失,而以太网链路速度却在持续飙升,可达到 400Gbps。这种变化差异还包括现代数据中心中软件定义网络、虚拟化技术、存储、消息传递和安全协议方面的复杂性,有一种观点认为,SmartNIC 体系结构会继续存在。那么,当今的数据中心 SmartNIC 到底是什么呢?

图 11 展示了一个包含 SmartNIC 的数据中心服务器架构。SmartNIC 涵盖了全部典型的 NIC 功能,还包括卸载功能,加快了应用程序在服务器 CPU 和 GPU 的运行速度。SmartNIC 不是 CPU 或 GPU 的替代,而是通过网络卸载对 CPU 或 GPU 进行补充。一些关键卸载包括虚拟机接口支持、数据包灵活匹配、覆盖隧道的终止和发起、加密、流量计量、塑形和每流统计。此外,SmartNICs 通常包括整个协议卸载和直接数据放置,支持 RDMA 和 NVMe-oF 存储接口。

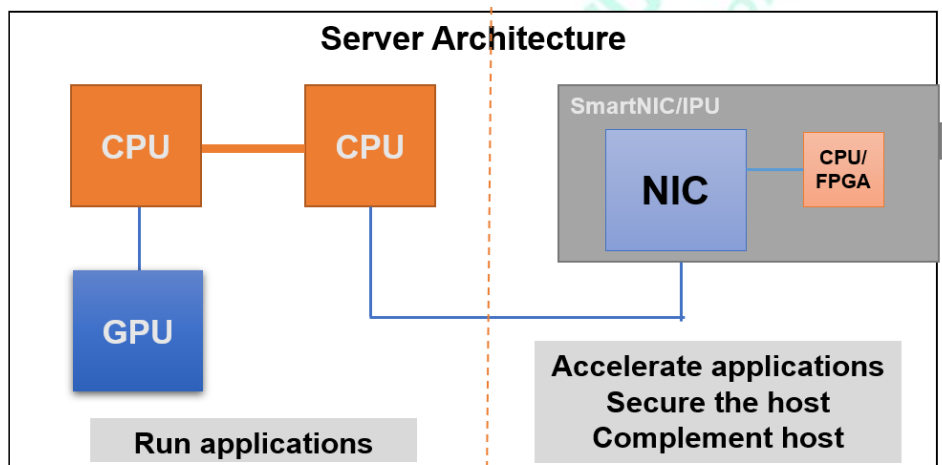


图 11 使用 SmartNIC 的服务器架构

现在, SmartNIC 一个新的关键特征是可编程性。过去对 SmartNIC 的质疑是它们无法满足快速变化的网络环境。早期的云数据中心倾向于将 CPU 用于大多数网络功能,因为 NIC 所需特性集的发展速度超过了硬件开发周期。然而,今天的 SmartNIC 拥有开放灵活的编程环境。它们实质上是开源环境中计算机前面的一台计算机,开源环境基于 Linux 和其他软件定义网络工具,如 Open vSwitch¹⁴。将智能技术无缝集成到开源生态系统中,能够快速开发特性并进行利用。

¹⁴ The Linux Foundation, “Open vSwitch,” 2016. [Online]. Available: <https://www.openvswitch.org/>. [Accessed 23 July 2020].

数据中心 SmartNIC 提高了网络的整体利用率和负载。它们使网络链路充分、迅速饱和，加剧了拥塞影响。同时，它们可以快速响应来自网络的拥塞信号，减轻间歇性冲击，避免丢包。SmartNIC 的可编程性使它能够适应新的协议，这些协议可以与网络相协调，以避免 incast 等情况。

3.4.4. 远程直接内存访问（RDMA）

RDMA 是一种新技术，旨在解决网络应用中服务器端数据处理的高延迟问题。RDMA 可以让数据直接从一台计算机的存储器传输到另一台计算机，而不需要任何操作系统的干预。它允许高带宽、低时延的网络通信，特别适用于大规模并行计算环境。图 12 展示了 RDMA 协议的原理。

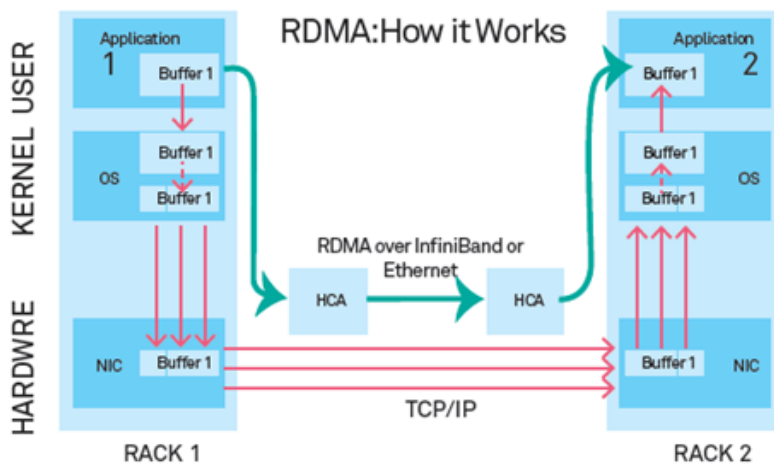


图 12 RDMA 协议的工作原理

RDMA 协议有三种不同的传输方式：Infiniband、iWarp 和 RoCEv1/RoCEv2。

- **Infiniband**

2000 年，InfiniBand 贸易协会（IBTA）首次发布了支撑 RDMA 的 InfiniBand 准则。InfiniBand 适用于高效硬件设计，可以确保数据传输的可靠性和访问远程节点内存的直接性。Infiniband 作为一种特定的网络解决方案，需要专门的 Infiniband 交换机和 Infiniband 接口卡。

- **iWarp**

iWarp 是一种 RDMA 协议，2014 年，IETF 规定 iWarp 需在 TCP 上运行。使用 TCP 作为传输工具，可以使 iWarp 覆盖互联网和广域网，以及标准以太网和数据中心。虽然 iWarp 可以在软件中实现，但要获得所需的性能，还需要数据中心使用专门的 iWarp 网卡。

● 融合以太网上的 RDMA (RoCE)

2010 年 4 月，IBTA 发布了 RoCEv1 规范，该规范增强了 Infiniband 体系结构规范性，支持以太网 Infiniband (IBoE)。RoCEv1 标准直接在以太网链路的顶层指定了一个 Infiniband 网络层。因此，RoCEv1 规范不支持 IP 路由。由于 Infiniband 依赖于无损物理传输，所以 RoCEv1 规范依赖于无损以太网环境。

现代数据中心倾向于使用三层技术来支持大规模和更大流量控制。RoCEv1 规范需要端到端的二层以太网传输，而在三层网络中不能有效运行。2014 年，IBTA 发布了 RoCEv2，它扩展了 RoCEv1，用 IP 和 UDP 报头替换了 Infiniband 全局路由报头 (GRH)。现在的 RoCE 是可路由的，它很容易集成到首选的数据中心环境中。然而，为了获得所需的 RDMA 性能，RoCE 协议被卸载，由特定网络接口卡取代。这些网卡实现了全部的 RoCEv2 协议，包括 UDP 堆栈、拥塞控制和任何重传机制。虽然 UDP 的重量比 TCP 轻，但附加支持在提高 RoCEv2 可靠性的同时，增加了网卡实现的复杂性。RoCEv2 仍然依赖 Infiniband 传输协议，该协议要在无损 Infiniband 环境中运行，因此 RoCEv2 仍然受益于无损以太网环境。

表 1 RDMA 网络技术的比较

技术	数据速率 (Gbit/s)	时延	关键技术	优势	劣势
以太网 TCP/IP	10, 25, 40, 50, 56, 100 或 200	500-1000ns	TCP/IP 套接 字编程接口	应用范围 广，价格低 廉，兼容性 好	网络利用率 低，平均性 能差，链路 传输速率不 稳定
Infiniband	40, 56, 100 或 200	300-500ns	InfiniBand 网 络协议和架 构动词编程	性能良好	不支持大规 模网络，需 要特定的

			接口		NIC 和交换机
RoCE/RoC Ev2	40, 56, 100 或 200	300-500ns	InfiniBand 网络层或传输层和以太网链路层动词编程接口	与传统以太网技术兼容, 性价比高, 性能好	特定的 NIC 仍然有许多挑战
Omni-Path	100	100ns	OPA 网络架构动词编程接口	性能良好	单个厂商和特定的 NIC 和交换机

图 13 显示了最常见的 RDMA 协议栈及相关标准。表 1 比较了不同的实施细节。大型云数据中心中, RDMA 成为选择高速存储、人工智能和机器学习的协议。现实世界中, 有成千上万的服务器在生产中使用 RDMA。使用 RDMA 极大的提高了应用程序性能¹⁵。例如, 机器学习分布式训练加快了 100 多倍, 用 RDMA 代替 TCP/IP 进行通信, 使得网络化 SSD 存储的 I/O 速度提高了约 50 倍。这些优化来自于 RDMA 的硬件卸载特性。

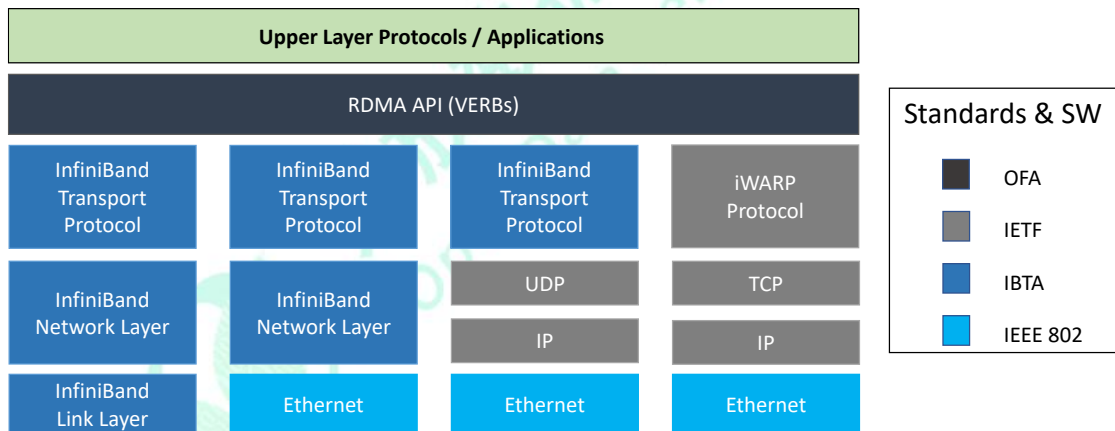


图 13 RDMA 协议栈和标准

3.4.5. GPU DirectRDMA

把两个好想法结合起来往往能创造一个突破性的想法。GPU DirectRDMA 包括 PCIe PeerDirect 技术和网络 RDMA 技术, 可以将数据直接发送到 GPU 内存。任何 PCIe 对等机都能支持此项技术, 如 NVIDIA GPU、XEON PHI、AMD

¹⁵ Li, Y., R. Miao, H. H. Liu, Y. Zhuang, F. Feng, L. Tang, Z. Cao, M. Zhang, F. Kelly, M. Alizadeh and M. Yu, "HPCC: High Precision Congestion Control," in Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM '19), New York, NY, USA, 2019.

GPU、FPGA 等。

GPU 通信使用“固定”缓存区来进行数据移动。SmartNIC 还可以使用“固定”内存与网络中的远程“固定”内存通信。这两种类型的“固定”内存是专用于 GPU 和 SmartNIC 主机内存的独立部分。

在 GPU DirectRDMA 之前，当一个 GPU 向远程服务器中的另一个 GPU 传输数据时，源 GPU 需要将数据从 GPU 内存复制到其固定的 CPU 内存中。然后主机 CPU 将数据从 GPU 固定内存复制到 SmartNIC 固定内存中。接下来，SmartNIC 使用 RDMA 将数据通过网络传输到远程服务器。在远程服务器端进行相反的过程。数据到达 SmartNIC 固定内存，CPU 再将数据复制到 GPU 固定内存中，最终数据会传输到远端 GPU 内存。图 14 显示了在使用 GPU DirectRDMA 之前，从 GPU 到 GPU 的数据拷贝过程。

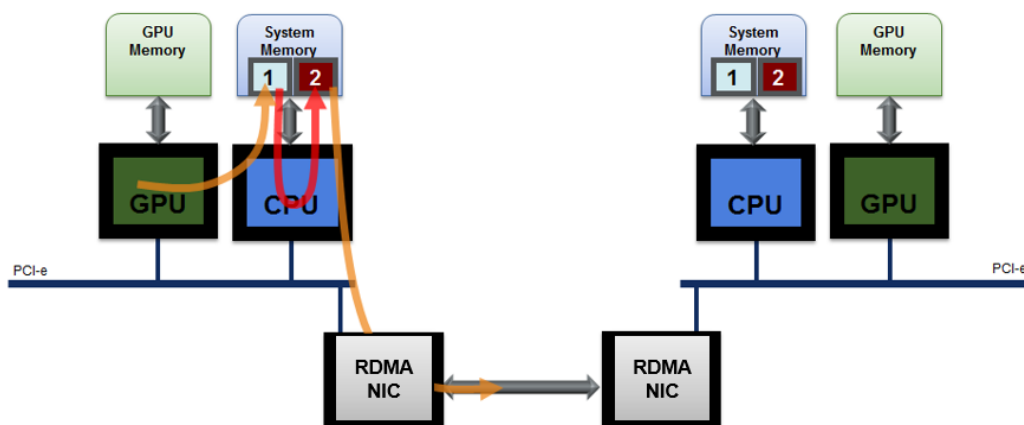


图 14 GPU DirectRDMA 之前的数据传输

虽然在 GPU 和 CPU 之间复制数据的成本比使用 TCP 在 GPU 之间传递数据的成本低得多，但它仍然存在以下问题：

1. GPU 资源消耗。CPU 可能成为数据传输过程中的瓶颈副本。
2. 时延增加，带宽降低。多余的内存副本会耗费时间，减少 I/O 带宽。
3. 主机内存消耗。多重固定缓存区会减少主机可用内存，影响应用程序性能，增加系统的 TCO。

写合并和 GPU 计算与数据传输重叠等优化，使网络 and GPU 共享“固定”缓存区。免除了在主机内存中复制数据的过程，数据可以直接通过 RDMA 传输。在接收端，数据经由 RDMA 到达后，直接写入 GPU 固定的主机缓存区。这种技术消除了 CPU 和 GPU 之间的缓存区复制过程，称为直接 GPU 技术（如图 15 所示）。

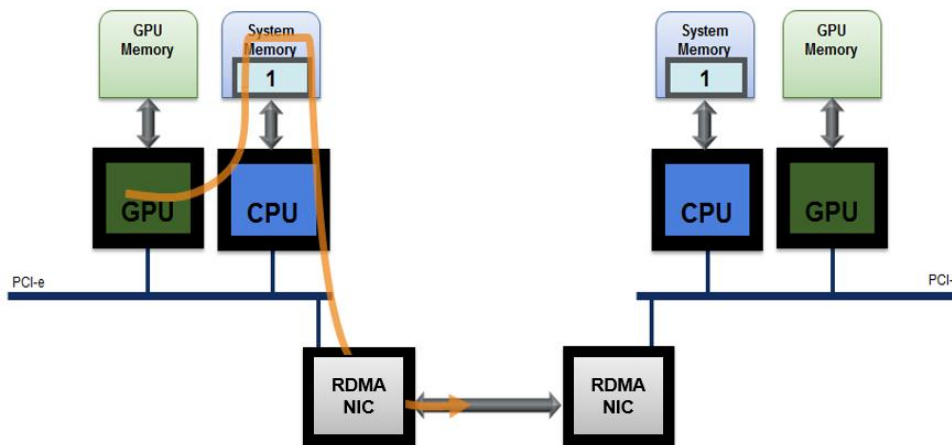


图 15 使用 GPU Direct 的数据传输

在本地 GPU 内存和远程 GPU 内存之间创建 RDMA 通道，消除 CPU 带宽和延迟瓶颈，实现进一步优化，极大的提高了 GPU 远程节点之间的通信效率。为了实现这一优化，CPU 协调 GPU 和 SmartNIC 的 RDMA 通信。SmartNIC 可直接访问 GPU 内存，向远端 GPU 内存发送和接收数据。这种技术被称为 GPU DirectRDMA 技术（如图 16 所示）。

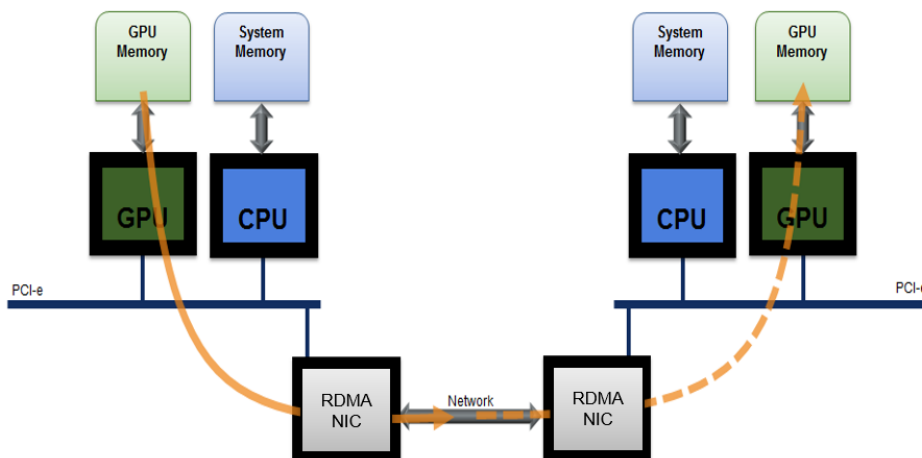


图 16 基于 GPU DirectRDMA 的数据传输

图 17 显示了 GPU DirectRDMA 技术如何将 GPU 通信性能提高到传统方法的 10 倍。这些改进方案使 GPU DirectRDMA 技术成为 HPC 和 AI 应用程序的必备组成部分，提高了应用性能，增强了可扩展性。

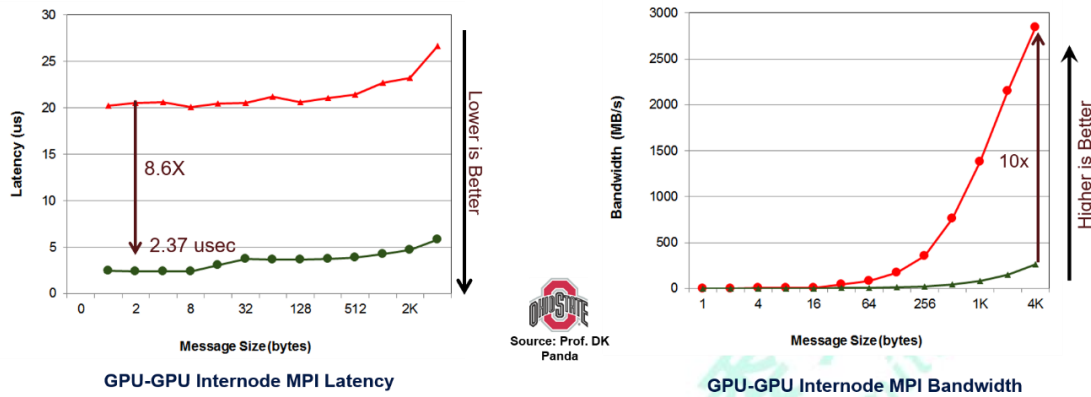


图 17 GPU DirectRDMA 性能（来自 OSU）

4. 当今数据中心网络面临的挑战

4.1. 平衡高吞吐量和低时延

在大规模数据中心中很难同时实现低时延和高吞吐量。为了实现低时延，必须以线速率开始传输，同时维持几乎空白的交换机队列流动。以线速率开始流动会使它们立即消耗所有可用的网络带宽，并可能导致汇聚点的极端拥塞。大缓存交换机可以吸收临时拥塞以避免丢包，但延长了敏感数据包传递时延。虽然大缓存交换机提供了充足的资源来平衡低时延和高吞吐量，但构建大缓存交换机越来越困难。交换容量随着链路速度和端口密度的提高而不断增加，但普通交换芯片的缓存区大小却无法相适应。图 18 显示了 Broadcom 制造的顶级数据中心交换机芯片的硬件趋势¹⁶。

¹⁶ Goyal, P., P. Shah, N. Sharma, M. Alizadeh and T. Anderson, “Backpressure Flow Control,” in Proceedings of the 2019 Workshop on Buffer Sizing (BS '19), New York, NY, USA, 2019.

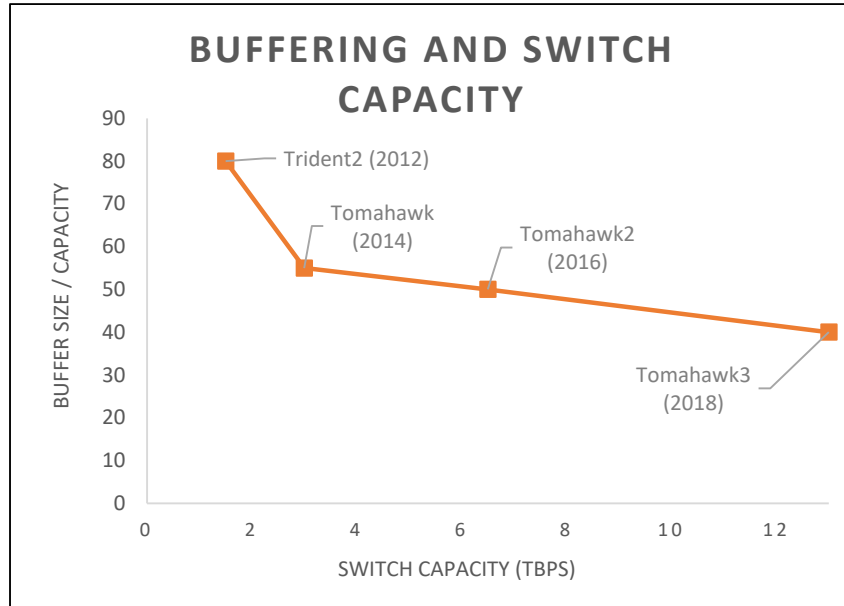
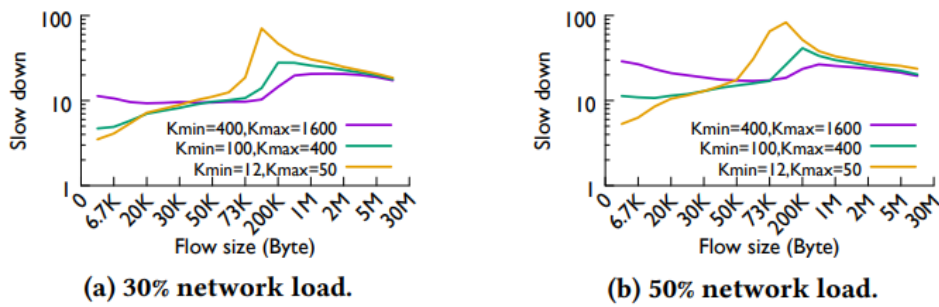


图 18 交换机芯片缓存区趋势

较低的 ECN 标记阈值可以减缓恶意流量并保持交换队列级别为空，但这会降低吞吐量。高吞吐量数据流有助于形成更大的交换机队列和更高的 ECN 标记阈值，以防止对临时拥塞的过度反应和非必要减速。

实验表明，在改变算法、参数、流量模式和链路负载后，可以权衡高吞吐量和低时延¹⁷。图 19 展示了在输入公共 RDMA WebSearch 流量工作负载的对照实验中，如何通过不同的 ECN 标记阈值 (K_{min} , K_{max})，使流量完成时间 (FCT) 延迟超过其最小理论值。较低的 K_{min} 和 K_{max} 值将导致 ECN 标记更快出现，并迫使流量大幅减速。如图所示，当使用低 ECN 阈值时，对于时延敏感的小流量，其 FCT 减速较慢，而接入带宽的大流量，FCT 减速更剧烈。网络负载越高，这种趋势越明显（图 19(b) 的平均链路负载为 50%）。

¹⁷ Li, Y., R. Miao, H. H. Liu, Y. Zhuang, F. Feng, L. Tang, Z. Cao, M. Zhang, F. Kelly, M. Alizadeh and M. Yu, “HPCC: High Precision Congestion Control,” in Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM '19), New York, NY, USA, 2019.

图 19 WebSearch 不同 ECN 阈值的 FCT 放缓分布¹⁷

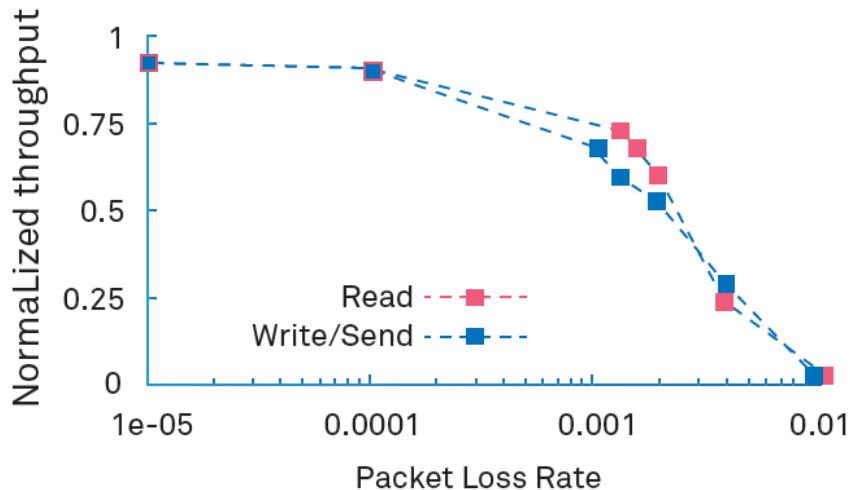
4.2. 无死锁无损网络

RDMA 相较于 TCP 的优势包括低时延、高吞吐量和低 CPU 使用率。然而，与 TCP 不同的是，RDMA 需要一个无损网络；即交换机不会因为缓存区溢出而导致数据包丢失¹⁸。RoCE 协议在 UDP 顶层运行，使用后退 N 帧重传策略，当重传被调用时，会严重影响性能。因此，RoCE 要求基于优先级的流量控制，确保数据中心网络中不出现丢包。图 20 显示了 RoCE 业务吞吐量随着丢包率增加而快速下降。哪怕只有千分之一的数据包丢失，也会使 RoCE 服务性能降低约 30%¹⁹。

基于优先级的流量控制（PFC）功能是当接收设备输入缓存区的占用超过设定阈值时，暂停上游发送设备，防止因缓存区溢出造成丢包。虽然提供了 RoCE 所必需的无损环境，但 PFC 的大规模使用也存在一些问题，包括 PFC 死锁的可能性。

¹⁸ Guo, C., H. Wu, Z. Deng, G. Soni, J. Ye, J. Padhye and M. Lipshteyn, “RDMA over Commodity Ethernet at Scale,” in In Proceedings of the 2016 ACM SIGCOMM Conference (SIGCOMM ’16), 2016.

¹⁹ Zhu, Y., H. Eran, D. Firestone, C. L. M. Guo, Y. Liron, J. Padhye, S. Raindel, M. H. Yahia and M. Zhang, “Congestion Control for Large-Scale RDMA Deployments,” in Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication (SIGCOMM ’15), London, United Kingdom, 2015.

图 20 丢包对 RDMA 吞吐量的影响¹⁹

PFC 式反向压力的死锁无损网络中的研究已经持续很多年了²⁰。当数据中心网络的交换机之间发生循环依赖缓存（CBD）时，就会发生 PFC 死锁。当一组交换机中的一个从属交换机等待组中其他交换机在传输数据包之前的可用缓存区时，CBD 就被创建了。如果 CBD 涉及的交换机正在使用 PFC，并且交换机序列通过物理连接形成回路，就会发生 PFC 死锁。Clos 数据中心网络中的 RDMA 流分布在多个等价路径上，以尽可能实现最大吞吐量和最低时延。虽然逻辑拓扑中没有循环，但这些路径原本就包含物理拓扑中的循环。网络中的 PFC 死锁可以完全停止网络流量。

考虑图 21 中的示例。该图展示了形成 PFC 死锁的四个阶段。第一阶段，四个流量在 Clos 架构上负载均衡，网络平稳运行。第二阶段，红色十字表示拓扑中的临时故障或永久故障，如链路故障、端口故障或路由故障。在本例中，故障使 H1 和 H7（绿线和黄线）之间的流量被重新路由。如第二阶段所示，重新路由推动更多流量通过分支 2 和分支 3，导致主干 1 和主干 2 中出现潜在溢出。在这个例子中，我们假设首先对主干 1 施加压力。为了避免损失，主干 1 的交换机使 PFC 向分支 3 方向偏移，如第三阶段所示。现在，分支 3 中的流量发生拥堵，进一步导致拓扑周围拥堵，PFC 信息沿着环路向初始拥塞点反向倾泻。第四阶段出

²⁰ Hu, S., Y. Zhu, P. Cheng, C. Guo, K. Tan, J. Padhye and K. Chen, "Tagger: Practical PFC Deadlock Prevention in Data Center Networks," in Proceedings of the 13th International Conference on emerging Networking Experiments and Technologies (CoNEXT '17), 2017.

现 PFC 死锁。

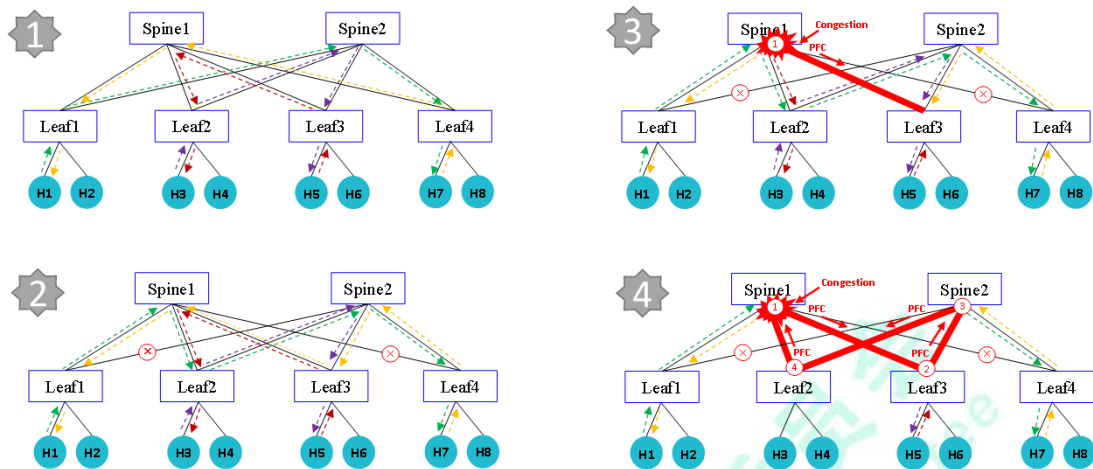


图 21 PFC 死锁示例

当网络规模较小时，PFC 死锁的概率较低。然而，随着 RoCE 协议规模的扩大和高性能要求的提高，PFC 死锁的概率显著增加。实现更大规模和最佳性能是未来智能无损数据中心网络的关键目标。第 5 节讨论了一种预防 PFC 死锁的新技术。

4.3. 大规模数据中心网络的拥塞控制问题

RDMA 技术最初是客户在受限、保守、小规模的环境中使用的，例如高性能集群计算或目标存储网络。至少在某种程度上，网络运营商可以调配专用环境所需的资源。然而，事实证明，RDMA 的性能优势可以应用于许多应用程序环境，人们强烈希望能够大规模使用 RDMA。图 22 显示了一个大型 RoCE 网络的示例。在该示例中，整个数据中心网络都基于以太网。计算集群和存储集群采用 RDMA 协议，X86 服务器集群采用传统的 TCP/IP 协议。

在大规模数据中心网络场景中，由于不同的原因，TCP 和 RoCE 流量可以在网络公共部分中使用。具有高速存储后端的传统网络应用程序为了读写数据，将终端用户的 TCP 请求和 RDMA 存储请求混合在一起。当使用 RoCE 进行数据通信时，RDMA 设备的管理和软件定义的控制平面通常基于 TCP。AI/ML 应用程序使用 RoCE 实现 GPU 和 CPU 互连，但也可能使用 TCP 存储方案。这使得 TCP 和 RoCE 在计算与计算、存储与存储以及计算与存储系统之间出现多种组合。

理论上，在网络中分离 TCP 和 RoCE 流量应该很容易。根据可提供的 8 类服务，基于不同队列调度算法，映射到 8 个队列。

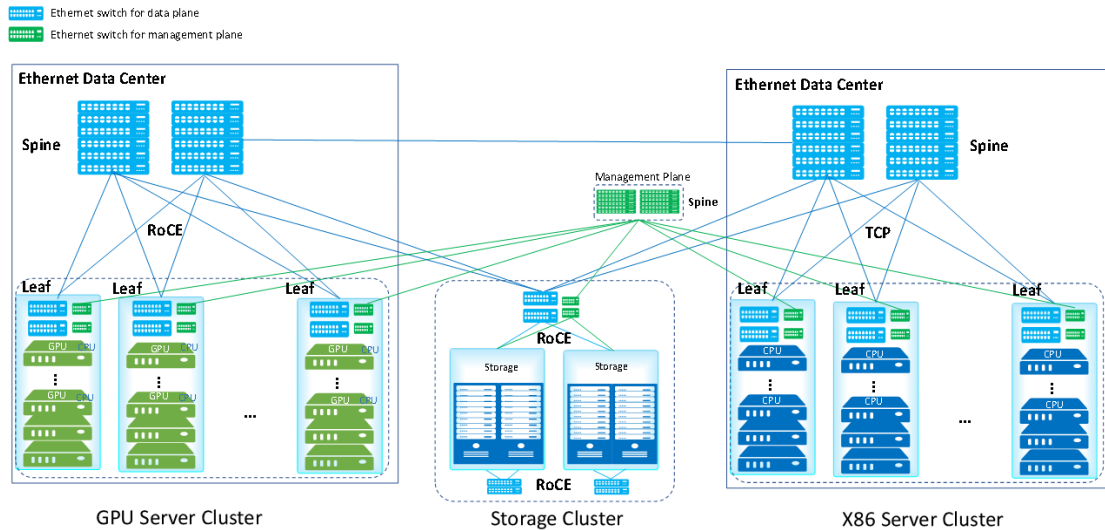


图 22 大型数据中心网络的 RoCE 应用

不同的交换机队列可用于区分不同的流量类型。随着每个交换机芯片的端口数增加，每个端口上的每个队列分配到充足的专用内存，可以吸收微爆发流量，避免数据包丢失，但这种方式非常昂贵且技术上具有挑战性。为了解决这个问题，交换机芯片供应商采用了一种智能缓存机制，它混合了固定缓存和共享缓存。

智能缓存的核心思想是创建动态共享缓存区。目标是通过减少专用缓存区的数量，优化缓存区利用率和突发吸收，同时所有端口间提供一个动态的自调整共享池来处理临时的突发情况²¹。

智能缓存区架构如图 23 所示。每个端口针对其各个队列设立了固定数量的专用缓存池和一个集中式通用动态缓存池。该架构认为，典型数据中心环境中的拥塞在任意给定时间点都会发生在出口端子集中，很少同时发生在所有端口上。这种假设要求集中式芯片缓存区的内存占用率在总成本和功耗方面达到“适当大小”，同时在需要时通过自调整阈值为拥塞端口提供资源。

²¹ Das, S. and R. Sankar, “Broadcom Smart-Buffer Technology in Data Center Switches for Cost-Effective Performance Scaling of Cloud Applications,” April 2012. [Online]. Available: <https://docs.broadcom.com/docs-and-downloads/collateral/etp/SBT-ETP100.pdf>. [Accessed 24 June 2020].

相比其他交换机架构中的每个端口的静态缓存方案,智能缓存区方法显著提高了缓存区利用率,优化了数据中心应用程序的性能。然而,在拥塞时,共享动态池会影响流量类别分离。TCP 和 RoCE 流量在经过公共链路时,虽然流量类别不同,但仍可能相互影响。它们采用不同的拥塞控制机制、不同的重传策略和不同的流量类别配置方案。算法和配置可能导致公共资源的分配不当。图 23 显示了交换机负载过重时存在的问题。网络运营商根据网络的业务需求,将网络带宽分配给不同的流量类别,但是随着时间的推移,在拥塞时,无法满足带宽分配。不同的拥塞控制方法会产生不同的流量行为,影响智能缓存机制合理分配动态共享缓存池的能力。在这种情况下,即使 TCP 和 RoCE 被分配给不同的流量类别,TCP 也会抢占 RoCE 带宽。RoCE 流量完成时延增加了 100 倍。ODCC 进行了几项测试来验证流量共存的问题²²。

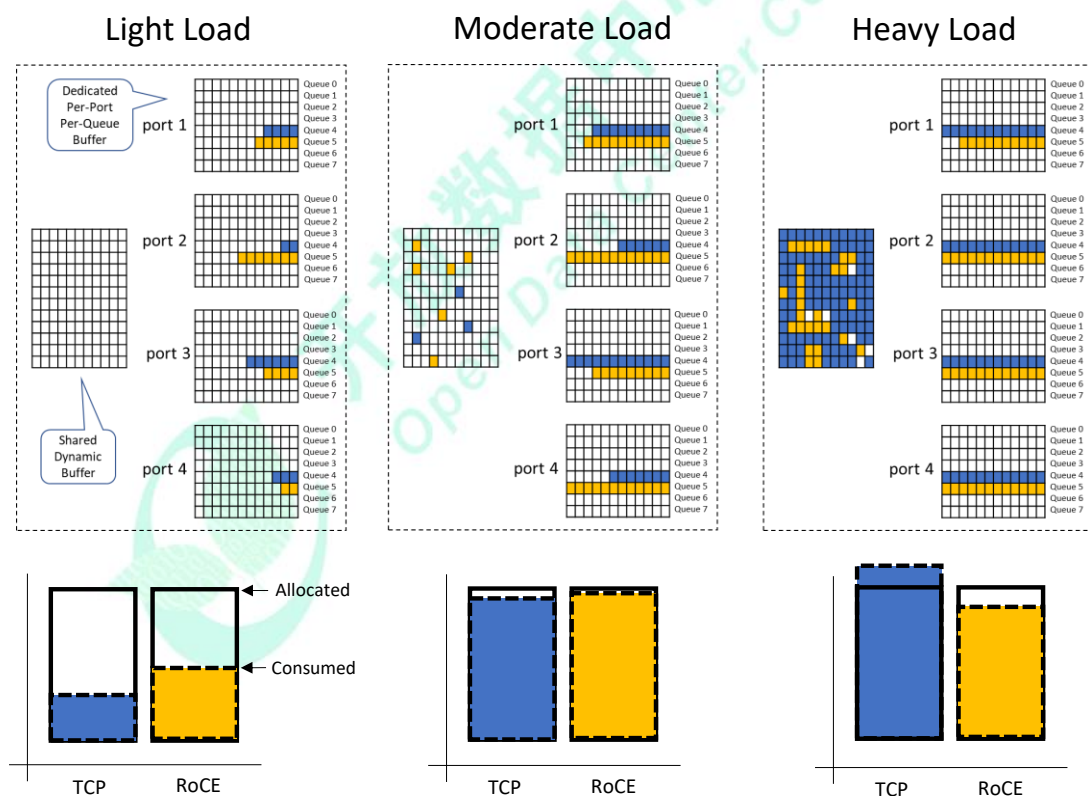


图 23 TCP 和 RoCE 的智能缓存共存

²² 开放数据中心标准推进委员会. ODCC 无损网络测试报告 (最终版) [R]. 北京: 开放数据中心标准推进委员会, 2019[2019-09-02]. <http://www.odcc.org.cn/download/p-1169553273830920194.html>.

4.4. 拥塞控制算法的配置复杂性

过去，HPC 数据中心网络规模很小，可以通过手动配置进行优化。然而，智能无损数据中心网络的目标是使 HPC 和 AI 数据中心能够发展到云规模，并实现自动配置。手动配置和手动调参在云规模下是不可能的，但是 HPC 数据中心的正常运行需要针对多个属性，在网络范围内配置一致。关键属性包括：

- 网络优先级与交换流量类别（即交换队列）的一致映射。

- 将应用流量一致地分配给网络优先级。
- 在无损流量类别上一致启用 PFC。
- 使用增强传输调度（ETS）为流量类别分配带宽。
- 为 PFC 设置缓存阈值，确保有足够的 headroom 来避免损失。
- 设置 ECN 标记的缓存阈值。

数据中心桥接交换协议（DCBX）可以使多个数据中心的配置属性在发现、配置和错配检测方面实现自动化。DCBX 利用链路层发现协议（LLDP）交换网络对等体配置属性的子集，如果对等体“愿意”接受推荐设置，两个对等体就可以实现配置一致。如果所有设备都运行 DCBX，这种一致的配置可以在整个数据中心网络中传播。然而，该协议并不交换数据中心网络的全部关键属性。特别是，它不能自动设置缓存区阈值，而缓存区阈值的确定非常复杂，且对于网络的正常运行至关重要。

4.4.1. 自适应 PFC Headroom 计算

PFC 缓存阈值决定了何时发送暂停帧，如图 24 所示。如果接收端的缓存区内存超过了 XOFF 阈值，接收器将发送一个暂停帧。当缓存区耗尽到 XON 阈值以下或者清零时，接收端可能会发送非暂停帧以取消先前暂停，或是仅让原始暂停超时。XOFF 阈值必须设置为允许接收动态帧。超出 XOFF 阈值的可用缓存区内存通常称为 headroom，它必须可用以确保无损操作。找到最佳 XON/XOFF 阈值很难。过高的估计阈值是不现实的，因为会浪费了宝贵的交换机内存，并减少可以支持的无损流量类别的数量。过低的阈值会导致丢包和协议（如 RoCE）性

能下降。找到最佳设置很困难，因为它需要对许多模糊的参数进行复杂的计算²³。

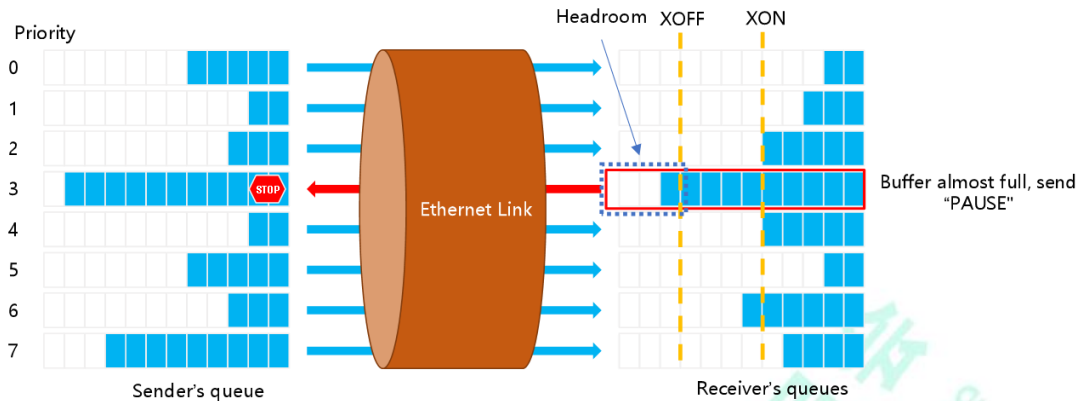


图 24 基于优先级的流量控制（PFC）

这些模糊的参数包括：

- 网络最大帧尺寸
- 链接速度
- 电缆长度
- 内部交换机和收发器时延
- 发送端响应时间
- 接收端缓存结构的内部存储单元大小

显然，这些参数不是网络运营商可以轻易获得的。许多都是在交换机内部实现的，并且因供应商而异。此外，包括链路速度和电缆长度的乘积在内的传播时延，可能因网络的每个端口而异。由于需要配置数千个端口，网络运营商将从配置 PFC headroom 的自动化方案中受益。

4.4.2. 动态 ECN 阈值设置

在拥塞的数据包中标记显式拥塞通知（ECN）的阈值，是保证网络平稳运行的另一个重要配置。如图 18 所示，设置低 ECN 阈值有助于实现低时延，但代价

²³ Cisco Systems, Inc., “Priority Flow Control: Build Reliable Layer 2 Infrastructure,” 2009. [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/switches/nexus-7000-series-switches/white_paper_c11-542809.pdf. [Accessed 15 Dec 2020].

是较大流量出现高吞吐量问题。面向吞吐量的流量，设置较高的 ECN 阈值会有更好的性能，但对延迟敏感的小流量，会缩短流量完成时间。随着数据中心网络内工作负载的变化，理想的解决方案是动态调整 ECN 阈值，以权衡高吞吐量和低时延。

ECN 的拥塞控制算法涉及到网络适配器和网络交换机之间的协作。随着工作负载的变化，需要协调交换机上的 ECN 阈值、NIC 运价减成和响应参数、端站协议栈。这种协调可能导致一组需要实时更新的配置参数无法维持。许多网络运营商只使用基于工程师长期经验推荐的静态配置。但是，静态配置不适应由应用程序的输入/输出和通信子协议的可测波动所驱动的网络流量实时变化。对于相同的应用程序，不同的静态配置可能导致不同的服务性能，而对于不同的应用程序，使用相同的设置可能导致数据中心网络中应用程序集合的性能不佳。测量一组应用程序的输入/输出和通信子协议的网络流量特征，可以产生一种预测算法，该算法可以动态调整交换机中的 ECN 阈值以及终端运价减成和响应参数。

5. 解决新数据中心问题的新技术

5.1. 低时延和高吞吐量的混合传输

传统的数据中心传输协议，如 DCTCP[3]和带有 DCQCN[44]的 RoCEv2 是发送端驱动的。他们试图测量和匹配沿途可用的瞬时带宽，将数据推入信道，等待接收端的反馈或测量。它们不断将更多数据推入信道，直到出现拥塞，此时它们会降低发送速率以避免丢包。有许多方法可以确定拥塞发生时间并相应调整发送速率，但发送端驱动传输的基本前提是相同的——根据可用信道的带宽估计不断调整发送速率。这是一种众所周知的、成熟的拥塞传输控制方法，已经在互联网等高度多样化的网络中取得了成功。可用带宽的准确估计不仅要检测拥塞，还要创造拥塞。拥塞信号的延迟和发送速率调整的不及时都可能导致队列剧烈波动，从而使吞吐量和延迟出现差异。路由器和交换机中的大型缓存区可以吸收这些波动，避免数据包丢失。

接收端驱动的传输（如快速通道²⁴）能够避免剧烈的队列波动，最大程度减小从发送端到接收端路径上的缓存。在接收端驱动的传输中，发送端的传输由接收端的时间表决定。在充分利用网络带宽的同时，使用请求授权协议或基于信用的协议可以调整发送端的速度，避免拥塞。这种方法尤其适用于多个并发发送端在接收端超限运行的 incast 拥塞情况。接收端驱动传输的问题是，接收端必须立刻估计路径上的可用带宽。可以使用相似的拥塞检测技术，而接收端驱动方法的优势在于可以第一个接收到拥塞信号。对于接收端驱动的传输，一个更大的挑战是发送端在初始缓存区请求中可能的固有延时。在大多数情况下，最初的请求授权交换不利于对延迟敏感的小流量，而这些小流量组成了数据中心网络中的大部分流量。

混合驱动传输，如 NDP²⁵或 Homa²⁶，发挥发送端驱动传输和接收端驱动传输的优势，通过避免拥塞减少时延，增加吞吐量。混合方法使发送端将一定量的未调度流量传输到网络时，不需要等待接收端的缓存区授权，但是在发送未调度流量后，它必须转换成接收端调度驱动。未调度流量没有额外的延时损失，对小流量有好处，但也会造成缓存区占用率的轻微波动，从而导致一定程度的丢包。由于未调度流量的数量很少，因此缓存区总体占用率仍然很低，这将导致更有限的时延和更低频的丢包。基于试探法调整未调度流量有助于在保持低缓存区利用率的同时，优化网络以实现高吞吐量和低时延。图 25 显示了每一种传输类型的高级传输方法以及缓存区利用率随时间变化的概念图。

²⁴ Cho, I., K. Jang and D. Han, “Credit-Scheduled Delay-Bounded Congestion Control for Datacenters,” in Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17), New York, 2017.

²⁵ Handley, M., C. Raiciu, A. Agache, A. Voinescu, A. W. A. G. Moore and M. Wojcik, “Re-architecting data center networks and stacks for low latency and high performance,” in SIGCOMM '17, Los Angeles, 2017.

²⁶ Montazeri, B., Y. Li, M. Alizadeh and J. Ousterhout, “Homa: A Receiver-Driven Low-Latency Transport Protocol Using Network Priorities,” 26 03 2018. [Online]. Available: <https://arxiv.org/abs/1803.09615v1>. [Accessed 22 May 2018].

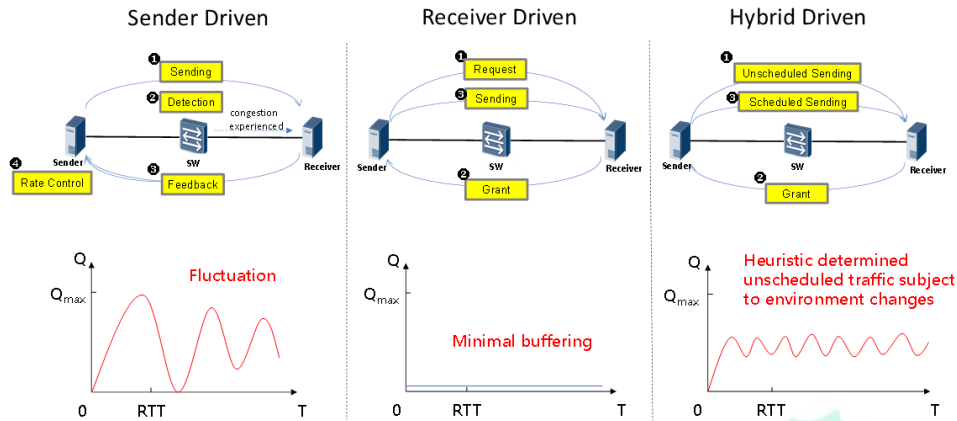


图 25 传输类型和概念上的网络缓存区含义

5.2. 基于拓扑识别的 PFC 死锁预防

均衡的 **Clos 网络**上，流量是无环路的，通常在入口端从上行链路流向下行链路，在出口端从下行链路流向上行链路。然而，当检测到瞬态链路故障时，会重新路由，流量重新路由的概率为 10^{-5} 左右²⁷。虽然 10^{-5} 的概率很小，但对于大流量和大规模数据中心网络，死锁仍有可能发生，即使是最小的死锁概率也会产生严重后果。PFC 死锁是真实存在的！规模越大，PFC 死锁的概率就越高，这个关键资源的服务可用性就越低。

有一种机制可以通过发现和规避 CBD 循环来防止 PFC 死锁。无死锁算法的核心思想是通过识别产生死锁的流量，打破循环依赖。**实现这一目标的第一步是发现拓扑结构，了解网络中每个交换机端口的端口方向**。创新的分布式拓扑和角色自动发现协议（role auto-discovery protocol）用于识别数据中心网络的网络位置和角色。

拓扑和角色发现协议能自动判断拓扑设备层级和每个设备端口方向。拓扑结构中的层级是指从网络边缘开始的跳数。例如，一个服务器或存储端点位于 0 级，那么与该服务器或存储端点相连的机架顶部交换机位于 1 级。一个端口的端口方向可以是上行链路、下行链路或交叉链路。例如，上行链路方向是根据连接到另

²⁷ Hu, S., Y. Zhu, P. Cheng, C. Guo, K. Tan, J. Padhye and K. Chen, "Tagger: Practical PFC Deadlock Prevention in Data Center Networks," in Proceedings of the 13th International Conference on emerging Networking Experiments and Technologies (CoNEXT '17), 2017.

一个更上层设备的端口确定的。

该协议始于识别已知条件。服务器和存储端点始终位于 0 级，它们的端口方向总是上行链路。交换机初始化时不需要知道它们的层级或端口方向，但当信息通过发现协议传播时，算法会收敛到一个准确的视图。图 26 显示了一个简单 Clos 网络中的拓扑结构和角色发现。

Discovery protocol exchange automatically determines:

1. Topology level of devices in network
 - 0 = End-station or server edge
 - 1 = Leaf
 - n+1 = Spine
2. Port orientation for each link
 - Uplink
 - Downlink
 - Crosslink

HINT: Servers are always at level 0 with uplinks.

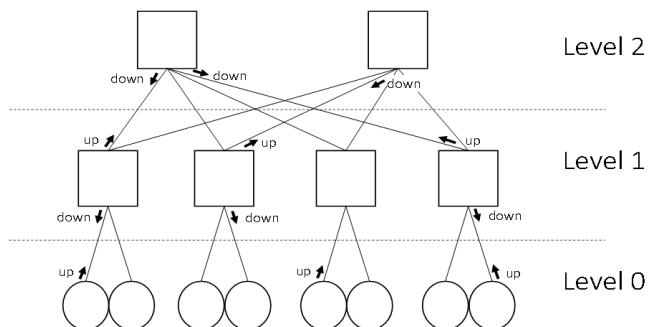


图 26 拓扑和角色发现

一旦协议识别出拓扑和端口角色，无死锁机制就可以确定网络中潜在的 CBD 点，调整转发平面以打破缓存区依赖。图 27 显示了如何识别拓扑中的潜在 CBD 点。在一个正常运行的 Clos 网络中，不存在 CBD，流量通常会穿过一个交换机入口和出口端口对，该端口对具有四种可能的端口方向组合中的三种。该流量可以从下行链路的定向端口通过到上行链路的定向端口。在脊柱网络中，流可以从下行链路的定向端口通过到下行链路的另一个定向端口。最后，当流量到达其目的地时，它可以从上行链路定向端口传递到下行链路定向端口。如果流量被重新路由，并且已经从上行链路的定向端口传输到上行链路的另一个定向端口时，可能存在一个 CBD。

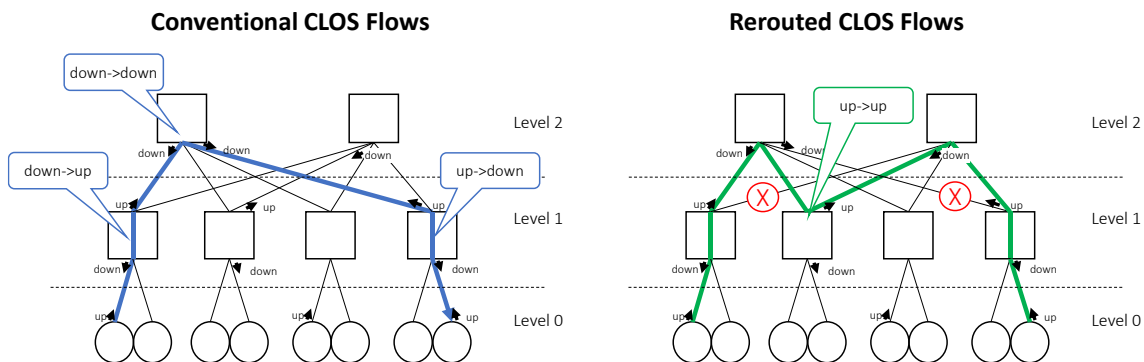


图 27 识别重路由流中的 CBD 点

识别出 CBD 点之后，转发平面负责破坏 CBD。CBD 的存在是因为一组具有相同分类的流量，通过一系列由于流量重新路由而形成环路的交换机。缓存区依赖是公共流量分类（例如交换队列）的共享缓存区内存。为打破 CBD，需要将重新路由的流量的数据包转发到一个单独的队列。这些包可以被识别，因为它们从一个上行链路的定向端口流动到上行链路的另一个定向端口。图 28 说明了交换机内队列重新映射的过程。该示例中，绿色流量重新映射到一个隔离队列将消除 PFC 死锁。不同的流量可以安全地通过潜在 CBD 点上的不同队列。

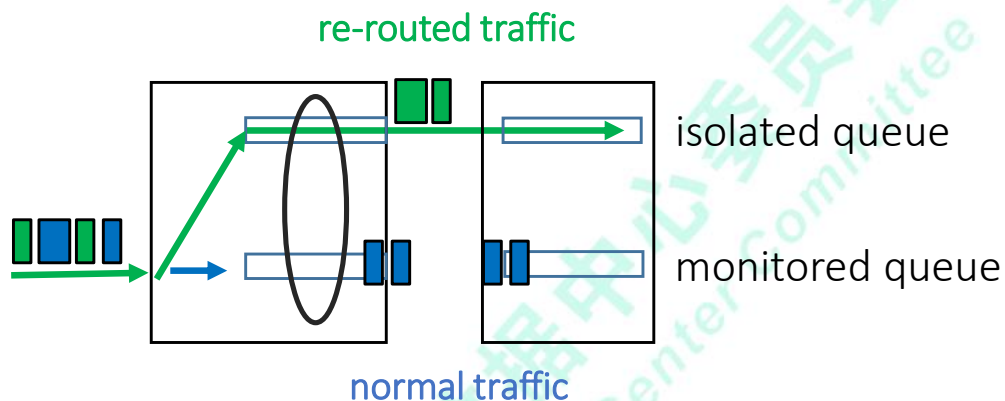


图 28 根据 CBD 重路由流识别交换机队列

ODCC 与许多网络供应商合作进行测试，验证了无死锁算法²²。

5.3. 改善拥塞的通知

在今天的數據中心中，用于 RoCEv2 协议的最先进的拥塞控制机制是数据中心量化拥塞通知（**DCQCN**）²⁸。DCQCN 结合 ECN 和 PFC，形成了大规模无损数据中心网络。图 29 展示了 DCQCN 的三个关键组件：反应点（RP）、阻塞点（CP）和通知点（NP）。

²⁸ Huawei, “Configuration Guide Low Latency Network,” [Online]. Available: <https://support.huawei.com/enterprise/en/doc/EDOC1100040243/c28a82e4/buffer-optimization-of-lossless-queues>. [Accessed 14 July 2020].

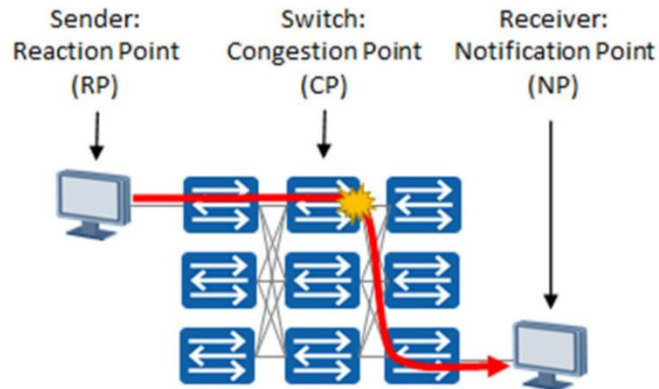


图 29 基于 DCQCN 的 RoCE 拥塞控制的三个组件

5.3.1. 反应点（RP）

反应点（RP）负责调节数据包注入网络的速率。它通常在发送网卡上实现，并在网络中检测到拥塞时，响应 NP 发出的拥塞通知包（CNP）。当一个 CNP 被接收，RP 会降低当前的注入速率。如果 RP 在指定的时间内没有收到 CNP，它会使用 DCQCN 指定的量化算法提高发送速率。

5.3.2. 阻塞点（CP）

阻塞点（CP）包含在发射器（RP）和接收器（NP）之间的路径开关中。当一个出口队列出现拥塞时，CP 负责用 ECN 标记数据包。拥塞是根据出口队列长度和可配置阈值（ K_{min} 和 K_{max} ）评估确定的。当队列长度小于 K_{min} 时，不标记流量。当队列长度大于 K_{max} 时，所有通过队列的数据包都会被标记。当队列长度在 K_{min} 和 K_{max} 之间时，根据 DCQCN 规定标记概率，随队列长度的程度而增加。

5.3.3. 通知点（NP）

通知点（NP）负责在一个流量数据包在网络中遇到拥塞时通知 RP。当带有 ECN 标志的数据包到达接收端时，如果在过去的 N 微秒内没有发送过 CNP 数据包，则 NP 会将 CNP 数据包发送回 RP。可以将 N 设为 0，这样 NP 就会为每个设置了 ECN 标志的数据包发送一个 CNP。

随着数据中心网络规模的扩大和支持的流量数增多，分配给每个流量的平均

带宽可能会变小。在这种环境中发生拥塞的流量可能会导致其数据包延迟，从而使到达 NP 的 ECN 标记也会延迟。如果 ECN 标记数据包的到达速率大于 RP 用来增加注入速率的间隔，可能会出现问题。即 RP 在应该降低注入速率的时候，却开始增加注入速率，出现这个现象是因为流量拥塞，且丢失的 CNP 信息只是被延迟了。在这种情况下，端到端拥塞控制环路不能正常工作。

无损网络中，端到端拥塞控制环路失效会导致拥塞扩散。这种非必要拥塞会增加 PFC 信息和链接暂停时间。这些 PFC 信息进一步延迟了 ECN 标记数据包的传输，加剧问题的严重性。在这种场景下，PFC 和 ECN 的结合变得无效。

想要解决这个问题，可以由 NP 发送 CNP 数据包进行网络智能化补充。智能化包括考虑出口端口的拥塞水平、接收到 ECN 标记数据包的时间间隔、RP 增加 DCQCN 速率的时间间隔。收到 ECN 标记数据包后，CP 会跟踪收到 ECN 标记数据包频率和拥塞包序列号。当 CP 出口队列发生拥塞时，CP 可能会根据接收到的 ECN 标记包速率和 RP 处 DCQCN 速率增加的间隔时间，主动补充 CNP。由于 CP 知道被标记的 ECN 数据包被延迟，且来自该 NP 的后续 CNP 数据包也会被进一步延迟，因此补充的 CNP 数据包可以防止端到端拥塞控制环路失效。只有当 CP 出口队列严重拥塞时，才会执行补充 CNP 操作，因此当 DCQCN 在正常非拥塞状态下运行时，时延和吞吐量不会受到影响。解决方案如图 30 所示。

ODCC 测试了增强的拥塞控制机制，效果良好²⁹。测试结果显示，在 TCP:RoCE = 9:1 时，性能可以提高 30% 以上。

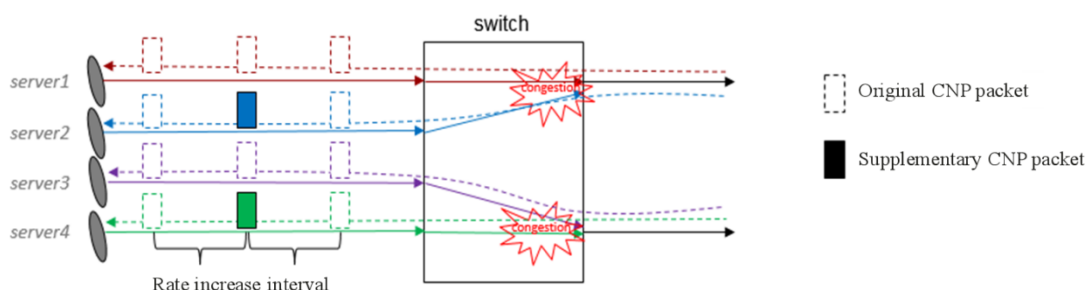


图 30 智能补充 CNP

²⁹ 开放数据中心标准推进委员会. ODCC 无损网络测试基准[S/OL]. 北京: 开放数据中心标准推进委员会, 2020[2020-09-03]. <http://www.odcc.org.cn/auth/v-1300974311558307841.html>.

5.4. 解决拥塞控制算法的配置复杂性

由于需要配置数千台交换机和数万个端口，网络运营商需要自动解决方案来正确配置数据中心网络中负责管理拥塞控制的参数。数据中心交换协议（DCBX）在简化一些配置和错误检测方面取得了长足进步，但是还需要不断改进。需要设置和调整交换机缓存区阈值的自动化解决方案。

5.4.1. 优化缓存区以降低 PFC headroom 配置的复杂性

成功设置 PFC XOFF 阈值的关键是确保在发出暂停帧后，有足够的 headroom 来吸收动态数据。在发送暂停帧和发送方停止数据传输之间有一个自然时延。在此时延期间，headroom 必须提供足够的缓存区用来接收数据，但对所需内存量的计算可能相当复杂。组成延迟的大部分是在交换机内部实现的，并且保持相对静态。例如，接口延迟和更高层次的延迟没有改变特定的配置和执行。这些静态延迟组件可以在网络上的对等体之间进行通信，但目前没有能够这样做的标准协议。介质的传播延迟取决于传输速度和电缆长度。为了准确地获得这一延迟组件，需要进行测量。

ODCC 工作组定义了一种可选时间戳，可用于测量点对点链路上两个对等点之间的电缆延时。然而，它针对的是受限环境中对时间敏感的应用，例如音频/视觉、工业和汽车网络。它的主要关注点是启用精确时间协议（PTP），用于在整个计算机网络中同步时钟。虽然细粒度的同步时钟在数据中心的很有价值，但在数据中心交换设备中支持全套功能的负担可能很繁重。另一方面，延迟测量设备在数据中心中可以实现 PFC 阈值的自动配置。要实现完全自动化配置，必须具有发现并在对等点之间交流的能力，以及其他 DCBX 特质。

5.4.2. 智能 ECN 阈值优化

ECN 阈值决定了交换机指示信息数据包发生拥塞的紧迫度，以及后续发送端调整传输速率所需的频率。最佳阈值设置取决于网络的当前状态和争夺公共资源的通信流类型。如前所述，低阈值设置有益于对时延敏感的小流量，而高阈值设置可以提高对吞吐量敏感的大流量性能。这些流量的混合方法和它们的通信模

式是不断变化的，但已经证明，基于模拟应用程序流量行为的机器学习技术是可预测的³⁰。预测数据中心网络流量模式的机器学习模型可用于动态调整 ECN 阈值，以优化低时延和高吞吐量之间的权衡。智能缓存方案中动态内存池的不公平共享问题也可以通过动态调整 TCP 和 RoCE 流量的 ECN 阈值来解决。

为了训练数据中心的网络流量模式模型，AI/ML 系统需要大量来自网络的实时数据。数据采集系统需要捕获数据中心和大规模网络设备之间的时间关系。基于 SNMP 和/或 NetConf 的传统网络监控系统使用轮询从设备中“拉出”数据。这种方法存在扩展问题，扩大了网络流量，并且增加了关联收集数据的困难度。需要获得直接从网络设备流出的基本参数遥测流。遥测技术是一种网络监测技术，旨在从物理或虚拟设备上快速收集性能数据。遥测技术不同于传统的网络监测技术，因为它使网络设备能够实时、高速地将高精度性能数据“推送”到数据存储库中。它提高了数据采集过程中设备和网络资源的利用率。

利用来自网络设备的遥测数据流，AI/ML 系统可以建立一个模型来监控整个网络上所有队列的拥塞状态。参数流可以用来训练和再训练网络模型，使得网络设备上的推理引擎预测数据中心环境发生变化，并自调整其 ECN 阈值。该模型的输入可以远超传统网络监测系统所获得的现有计数器。基本输入参数包括以下内容：

- 出口端口的 incast 比率 (N:1) 快照
- 入口端口的老鼠流和大象流混合体
- 交换机缓存区占用率的变化

其他更传统的网络指标可能包括以下内容：

- 端口级信息
 - ✧ 发送和接收字节
 - ✧ 发送和接收数据包

³⁰ Mozo, A., B. Ordozgoiti and S. Gomez-Canaval, “Forecasting short-term data center network traffic load with convolutional neural networks,” PLoS ONE, vol. 13(2), no. e0191939. <https://doi.org/10.1371/journal.pone.0191939>, 2018.

- ◇ 发送和接收方向丢失的数据包
- ◇ 接收的单播数据包、多播数据包和广播数据包
- ◇ 发送的单播数据包、多播数据包和广播数据包
- ◇ 发送和接收的错误数据包
- ◇ 输入端带宽利用率和输出端带宽利用率
- ◇ ECN 数据包
- 队列级信息
 - ◇ 输出队列缓存区利用率
 - ◇ headroom 缓存区利用率
 - ◇ 接收的 PFC 帧
 - ◇ 发送的 PFC 帧

另一种遥测类型称为带内遥测，在单个数据包穿网络时，提供有关其经历的实时信息。信息由数据平面收集并嵌入到数据包中，而不涉及控制平面。收集的信息量比传统遥测数据流更有限，因为它必须包含有限规模的原始数据包内容。然而，数据包内的信息与数据包存在于网络时观察到的网络状态直接相关。沿着路径的每一跳都可以被要求插入表示交换跳状态的本地数据。基本资料可能包括以下内容：

- 输入端和输出端数量
- 输入和输出的本地时间戳
- 输出链路利用率
- 输出队列缓存区利用率

从本地设备获取实时遥测输入的 AI 模型可以预测需要对 ECN 阈值进行的调整，实现低时延和高吞吐量之间的最佳平衡。检查带内遥测信号的目的是快速向发送端传送适当的拥塞信号，从而避免包丢失和流量完成时间的长尾延迟。

6. 结论

数据中心网络必须继续扩展，不断开发新技术，以满足人工智能和机器学习

应用在高速计算和存储方面的需求增长。白皮书探讨了现如今云级高性能计算数据中心的技术挑战和新解决方案，致力于研制出新的混合传输协议，可以更好地平衡人工智能和机器学习的高吞吐量和低时延通信需求。白皮书阐述了一种使用拓扑识别算法来防止 PFC 死锁的解决方案，该算法基于现有广泛部署的链路层发现协议（LLDP），研究了如何通过交换机补充拥塞信号，缩短拥塞通告信息反馈周期，描述了基于高级遥测系统开发的自动协议和人工智能模型，降低交换缓存区阈值配置复杂性的方法。这些创新方案，加之对开放和标准化工作的推进，可以将以太网的使用提前，将其作为现如今云级高性能数据中心的初始网络结构。



开放数据中心委员会
Open Data Center Committee



ODCC服务号



ODCC订阅号

www.ODCC.org.cn

开放数据中心委员会（秘书处）

地址：北京市海淀区花园北路 52 号

电话：010-62300095

邮箱：ODCC@odcc.org.cn