开放数据中心委员会

[编号 ODCC-2023-03004]

# 数据中心高性能网络拥塞检测技术白皮书（2023 年）

中移（苏州）软件技术有限公司

中国信息通信研究院云计算与大数据研究所

2023-09 发布

# 版权声明

# 编写组

**项目经理：**

| | |
|---|---|
| 赵兴华 | 中国移动云能力中心 |

**工作组长：**

| | |
|---|---|
| 王超 | 阿里云计算有限公司 |

**贡献专家：**

| | |
|---|---|
| 徐军 | 中国移动云能力中心 |
| 刘军卫 | 中国移动云能力中心 |
| 姚军 | 中国移动云能力中心 |
| 孟令坤 | 中国移动云能力中心 |
| 王东旭 | 中国移动云能力中心 |
| 张胜举 | 中国移动云能力中心 |
| 孙伟 | 云脉芯连科技有限公司 |
| 张久仙 | 中国移动云能力中心 |
| 季忠铭 | 中国移动云能力中心 |
| 许治国 | 中国移动云能力中心 |
| 潘训营 | 中国移动云能力中心 |
| 史成龙 | 中国移动云能力中心 |
| 陈继磊 | 中国移动云能力中心 |
| 杨亚军 | 中国移动云能力中心 |
| 王晓辉 | 中国移动云能力中心 |
| 郝泉澄 | 中国移动云能力中心 |
| 薛迁 | 中国移动云能力中心 |
| 徐军 | 中国移动云能力中心 |

# 目 录

# 术语与缩略语

| Term | Meaning |
|---|---|
| RDMA | Remote Direct Memory Access |
| RoCE | RDMA over Converged Ethernet |
| iWarp | internet Wide Area RDMA Protocol |
| GPU | Graphics Processing Unit |
| IOPS | Input/Output Operations Per Second |
| SRD | Scalable Reliable Datagram |
| AWS | Amazon Web Services |
| DPU | Data Processing Unit |
| RNIC | RDMA Network Interface Card |
| ECN | Explicit Congestion Notification |
| DCQCN | Data Center Quantized Congestion Notification |
| HPCC | High Precision Congestion Control |
| PFC | Priority Flow Control |
| RED | Random Early Detection |
| AQM | Active Queue Management |
| RTT | Round Trip Time |
| INT | In-Net Telemetry |
| ECMP | Equal-Cost Multi-Path |
| TCD | Ternary Congestion Detection |
| CBFC | Credit-Based Flow Control |
| PFC | Priority-based Flow Control |
| QCN | Quantized Congestion Notification |
| RPS | Random Packet Spraying |
| CONGA | Distributed Congestion-Aware Load Balancing |
| FCT | Flow Complete Time |
| RED | Random Early Detection |
| BCN | Backward Congestion Notification |
| FECN | Forward Explicit Congestion Notification |
| PCN | Pre-Congestion Notification |
| HPQ | High Priority Queue |
| LPQ | Low Priority Queue |

# 前 言

《“十四五”数字经济发展规划》中指出数字经济是继农业经济、工业经济之后的主要经济形态，是以数据资源为关键要素，以现代信息网络为主要载体，以信息通信技术融合应用、全要素数字化转型为重要推动力，促进公平与效率更加统一的新经济形态。

随着数字经济的持续发展，算力需求呈爆发性增长，逐步成为新时代的核心生产力。算力的发展带动了网络的变革，构建了高效、灵活、敏捷的数据中心网络新型基础设施，成为算力网络驱动和演进的关键。

远程直接内存访问（Remote Direct Memory Access，RDMA）网络是一种高性能网络传输技术。通过绕过操作系统内核，RDMA 可以直接在网络适配器和内存之间传送数据，从而减少了数据传输过程带来的延迟和 CPU 开销，提高了数据传输的效率和吞吐量。近年来，高性能网络广泛应用于高性能计算、云计算、大数据处理等领域，成为当下网络领域的研究热点之一。

高性能网络的重要性在于，为各种应用提供了快速、可靠、安全的数据传输能力，并将数据中心、云计算和大数据处理等领域的计算资源、存储资源和网络资源紧密结合，提高了整个系统的效率和性能。同时，高性能网络还可以支持更多的应用和服务，促进了科学研究、产业发展和社会进步。因此，高性能网络的发展和研究是当前网络领域的重要方向。

本白皮书通过阐明和分析高性能网络技术发展的过程与现状，以网络拥塞这一关键问题展开详述当前业界拥塞管理控制技术的架构体系，并聚焦拥塞管理控制过程中面临不同需求所产生的拥塞检测机制。本白皮书旨在通过对拥塞检测技术的研究，推动高性能网络技术的深入发展、生态链建设和产业落地。

## 一、 高性能网络的机遇与挑战

在需求端强力驱使下，过去的 10 年中，数据中心网络链路传输带宽经历了从 1 Gbps 到 100Gbps 的快速增长，并且这一增长趋势仍在持续。因此，作为未来数据中心服务的提供者，云计算厂商面临着越来越严苛的数据中心网络建设需求。

目前，传统数据中心应用的 TCP/IP 网络已经难以高效地满足新的需求。一方面，快速膨胀的链路速率导致了极高的 CPU 占用率，每增加一个用于 TCP 网络传输的 CPU 资源意味着云计算厂商能够出售的虚拟机减少了一个，这将降低整体的经济效益。另一方面，机器学习、搜索等业务所要求的超低的网络延迟（低于 10 us/跳），传统的 TCP/IP 协议的性能是很难达到的。

为解决这一问题，远程直接内存获取（Remote Direct Memory Access，RDMA）技术开始逐渐广泛地应用于数据中心网络中（本文提及的 RDMA 无损网络针对更广泛应用的以太网络，如无特殊声明，适用协议为 RoCEv2）。

相较于传统的 TCP/IP，RDMA 有着如下的优势：

1）降低了 CPU 占用率。数据传输过程不再需要 CPU 的持续介入，而是通过硬件卸载的形式完成数据传输。

2）降低了传输时延，避免了数据拷贝过程中频繁的用户态和内核态切换。因此，通过硬件卸载、内核旁路，RDMA 完成了数据传输和计算的解耦，从而实现高效的并行计算处理。

正因为以上的技术优势，高性能网络已经成为云计算领域应用广泛核心基础设施之一。据公开文献[1]显示，在微软 Azure 存储集群中，RDMA 流量已经占据了超过一半的比例。在可以预见的未来，高性能网络技术都将作为云计算领域的核心基础设施之一，深刻地影响数据中心技术格局。

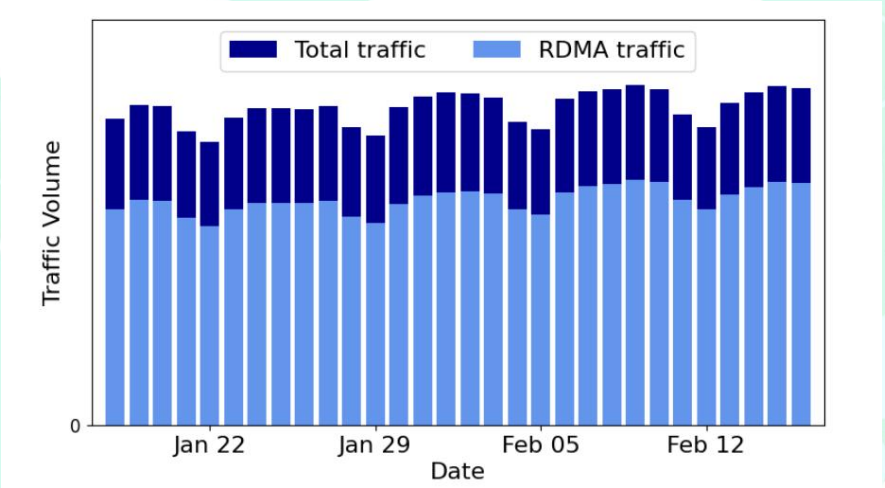



图 1 微软 Azure 存储集群流量占比[1]

## （一）应用背景与现状

随着云计算技术的发展，高性能网络的应用场景日益增多。本节主要从分布式云存储、内存池化、键值存储、智算中心四个方向的应用，对高性能网络的应用场景和应用现状进行概述。

### 1. 分布式储存场景

分布式存储是云计算中的一个核心应用。各家云厂商都会提供高达百万输入/输出操作每秒（IOPS）的高性能存储实例，旨在满足对性能要求极高的应用场景。

由于百万 IOPS 云硬盘需要同时处理大量的读取和写入请求，这就要求了网络要提供极高的吞吐量和极低的响应时间。因此，主流云厂商普遍选择 RDMA 作为高性能分布式存储的网络解决方案，如公开文献中阿里云、微软云等关于分布式云存储的工作[1]，[2]。

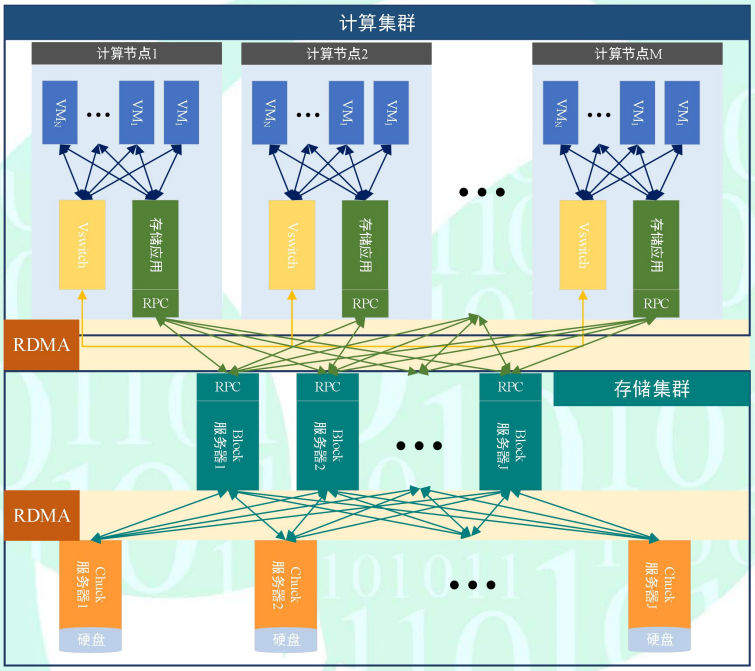



图 2 云存储基本架构图

阿里云 EBS 云存储中应用的阿里自研网络协议栈 Solar[3]，对云存储 IO 延迟进行了全面优化。论文中给出了 EBS 产品详细的网络延迟性能测评。图 3 中的数据为阿里云超过 10 万个计算节点一周时间的测试结果。在图中，Kernal 是传统的 TCP/IP 协议，Luna 是用户态加速协议栈，Solar 是阿里自研的 RDMA 网络，FN 是计算是存储的前端网络，BN 是存储集群后端网络，SSD 是落盘网络，SA 是阿里自研的 SPDK 软件。该实验很好的对比了内核态、用户态、RDMA 对于存储业务的影响。可以看到，整体 IO 延迟性能上，Solar RDMA

协议有明显的优势。同时，RDMA 网络协议栈还在很大程度上改善了整个网络的长尾时延问题，性能实现了数量级的提升。

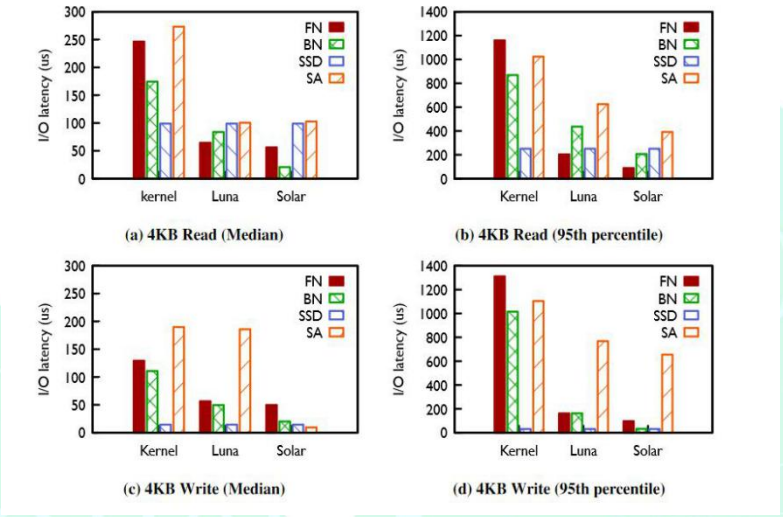



图 3 阿里云 EBS 网络性能对比测试

## 2. 内存池化场景

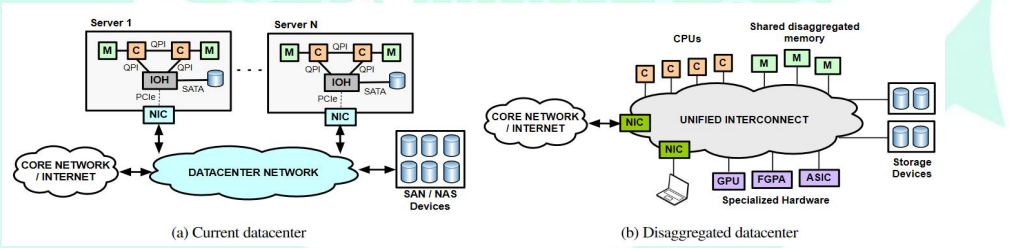



图 4 内存池化的分布式数据中心

现有的数据中心是通过服务器构建的，每个服务器紧密集成了计算任务所需的各种资源（CPU、内存、存储）。虽然这种以服务器为中心的架构已经持续使用了几十年，但最近的研究表明，未来即将出现一种向分解式数据中心（Disaggregated Datacenter，DDC）转变的范式。其中，每种资源类型都作为独立的资源池进行构建，而网络结构则用于连接这些资源池[4]。

资源池化的一个关键的促进（或阻碍）因素将是网络。因为将 CPU 与内存、磁盘分解开来，原本需要在服务器内部进行的资源间通信，而现在必须通过网络进行。因此，为了支持良好的应用级性能，网络结构必须提供低延迟的通信以应对这种负载更大的情况。

因此，RDMA 高性能网络作为一个解决方案在内存池化的场景已经有广泛的研究[5], [6]。RDMA 有效地提升了内存池化数据中心的效率。尽管没有完全解决资源池化场景的网络互连问题，但其仍然是未来分布式数据中心的一个有力的网络技术方案。

### 3．键值存储场景

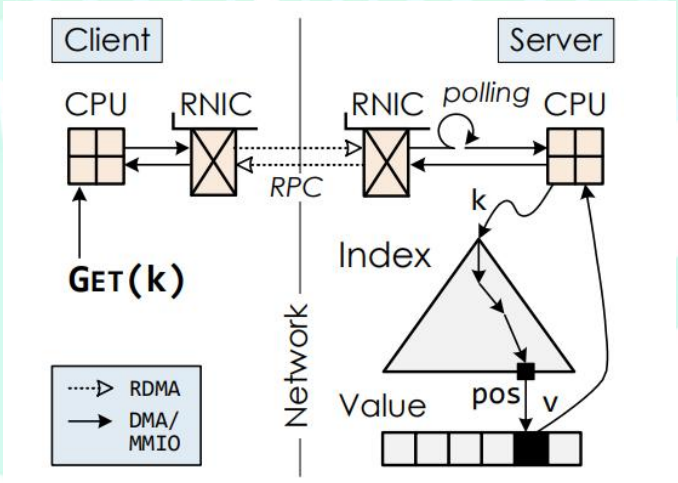



图 5 基于 RDMA 的键值存储系统[7]

键值存储（Key-Value Store）是一种数据存储方法，它以键值对（Key-Value Pair）的形式存储和访问数据。与传统的关系型数据库相比，键值存储通常更加简单、灵活、高效，并且可以处理更大规模的数据。键值存储不要求数据具有固定的结构和模式，因此

可以轻松地存储和检索各种类型的数据。键值存储还支持高度可扩展性和分布式部署，可以轻松地在多个节点上进行水平扩展和数据复制以提高性能和可靠性。

在常见应用中，Redis 就是一种流行的键值存储系统。它支持多种数据类型，包括字符串、哈希、列表、集合和有序集合等。与关系型数据库不同，Redis 不支持复杂的 SQL 查询语句，而是提供了一组简单的操作命令，如 GET、SET、INCR、DECR、LPUSH、RPUSH、SADD、SMEMBERS 等，以实现键值对的读写和操作。

然而，在键值存储中，CPU 是一个显而易见的性能瓶颈。而 RDMA 技术通过绕过内核的方式直接访问内存，这能够保证 CPU 资源的高效利用。因此，RDMA 技术在键值存储系统中的应用也逐渐被更多的讨论[7], [8]。同时，阿里云也公开声明了其 eRDMA 技术在 Redis 产品中的应用[9]。从测试结果可以看出，无论是 GET 测试还是 SET 测试，eRDMA 相对于 TCP 带来了至少 40%以上的性能测试数据提升。

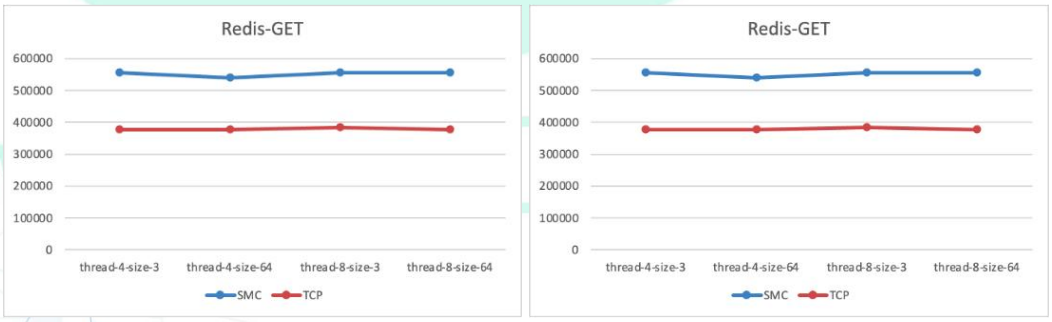



图 6 RDMA 技术加速 Redis 服务

## 4. 智能算力场景

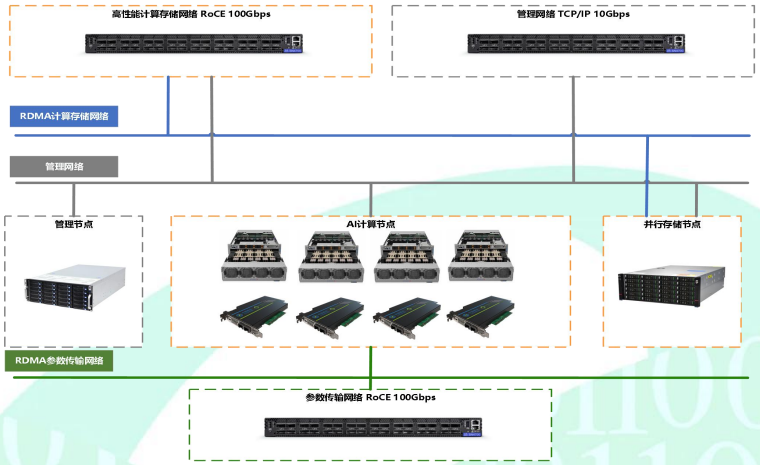



图 7 智算中心高性能网络组网方案

近年来，大型语言模型如 GPT 等在自然语言处理任务上的强大能力引起了广泛关注。这些模型通过预训练在海量文本数据上获取语言知识，然后进行微调应用于下游任务。大模型以极大的模型尺寸、大量数据和计算资源进行训练。其一系列成果显示了大模型具备了通过无监督学习获取语言理解能力的潜力。

但是训练大模型也带来了巨大的计算和环境成本，需要大规模高速互联的智算中心，其原因如下：

a) 模型参数量巨大，单机单卡无法加载整个模型。而使用多机多卡可以将训练的参数梯度分布在不同设备上。

b) 训练时间长。如果只使用单机单卡，训练大模型往往需要非常长的时间。多机多卡情况下，并行计算可以大幅减少训练时间。

c) 训练数据量大。多机多卡并行读取数据后汇总梯度，产生了大量的网络带宽需求。

因此，在智算中心场景下，高性能 RDMA 网络实现多个服务器、多个 GPU 的互联，打造多通道、无收敛、多路径的参数网络（如图 5 所示），是当前的主流技术方案之一。AWS 在其超算、智算服务中广泛的提供 SRD 高性能网络服务[10]，进一步的引起了行业内对高性能网络技术的大规模投入。

## （二）高性能网络拥堵问题与挑战

高性能网络已经成为云计算领域应用广泛核心基础设施之一。然而，RDMA 网络中出现拥塞问题将会大幅降低网络的吞吐和延迟性能，这也成为了限制 RDMA 网络应用规模的重要因素。当网络中的数据流量超过了网络链路的处理能力或带宽限制或者当多个节点同时进行 RDMA 通信时，网络链路无法及时处理或传输所有的数据包，就会发生拥塞。

拥塞一方面会导致交换机的缓存队列增大，数据包传输的延迟等比例的延长，使网络服务质量下降；另一方面，交换机中数据包堆积，会触发 PFC 机制，以保证 RoCE 网络的无损特性，这导致网络中会出现一系列相应的风暴、死锁等问题[11]。这也一定程度上限制了 RDMA 网络在以太网环境的部署规模和网络性能。因此，近年来在

RDMA 高性能网络方向聚焦拥塞问题，产生了大量的前沿研究和工程实践工作。

总之，随着未来数据中心网络带宽需求的不断增长，RDMA 高性能网络在云计算、人工智能等领域具有巨大的机遇。同时，拥塞问题作为 RDMA 网络中限制规模、性能的主要瓶颈，形成标准化、规范化的拥塞管控系统，将已有技术进行归纳延伸，是当前数据中心网络中迫切要完成的一项工作。拥塞检测技术中，有如下几点挑战亟需解决:

a) 精度、频率和开销的矛盾。对于网络拥塞信息的检测，当前存在多种主流方案，其获取的拥塞信息都不相同，但都遵循“没有免费的午餐”这一规则。更高的测量精度、更快的测量频率，都会带来额外的网络带宽开销（例如 INT 对比 ECN）。这需要对不同的场景需求进行深入的研究，以实现最佳的拥塞检测效果。

b) 标准和兼容性：RDMA 技术存在多种标准和实现，如 InfiniBand、RoCE（RDMA over Converged Ethernet）和 iWARP（Internet Wide Area RDMA Protocol）。其中，RoCE 网络的发展近年来尤为迅猛。原有的以太网拥塞检测机制和协议在 RDMA 网络中该如何规范化，这也是未来不同 RoCE 网络设备厂商和用户潜在的问题。

c) 跨层级应用：不同的拥塞检测机制可以在更多的拥塞管控技术层级进行应用。比如，RTT、ECN 的拥塞信息可以作为流量调度、负载均衡的参考权重。这些研究工作虽然已经较多，但哪些拥塞检测机制适合哪种层级的拥塞管控协议仍是需要进一步探讨的问题。

本白皮书通过阐明和分析高性能网络拥塞管控的技术发展的过程与现状，整理、探讨不同方案中关键的拥塞检测机制，归纳其技术路线与演进，从而推动高性能网络技术的深入发展，助力完整的生态链建设和产业落地。

## 二、 拥塞管理与控制技术体系

为了缓解高性能网络中的拥塞问题，RoCE 高性能网络协议已经构建了多层的拥塞管理和控制技术体系。这一体系中，细分主要包含拥塞控制、负载均衡、链路控制、流量调度等。形成了从用户层到链路层的多层次拥塞管理和控制体系。

其中，拥塞控制协议、链路控制的响应快、周期短，通过调整流的发送速率实现拥塞的避免，且主流方案通过闭环控制技术实现，因此归类为拥塞控制技术；负载均衡、流量调度，往往通过管理的方式，对数据进行调度分流，通过更高效的利用网络拓扑资源实现拥塞的避免，因此归类为拥塞管理技术。

本章中重点对现有拥塞管理与控制技术进行了归纳整理。以便系统的给出后续第 4 部分拥塞检测技术的技术发展方向。

### （一）拥塞控制技术

拥塞控制，顾名思义，可知其在网络拥塞问题处理中的核心位置。拥塞控制是为了防止网络过载而采取的一种流量调节机制。当网络拥塞时，路由器队列堆积，丢包和延迟增加。拥塞控制算法(如 TCP 的滑动窗口)可以通过监控网络状况来动态调整发送方的发送速率。比如，在拥塞开始时降低发送速率，拥塞消除后逐渐增加发送速率。这种闭环反馈机制可以使网络稳定运行在最优状态，最大化网络的吞吐量，是保证网络顺畅运行的重要机制。

## 1. 基于 ECN 的拥塞控制

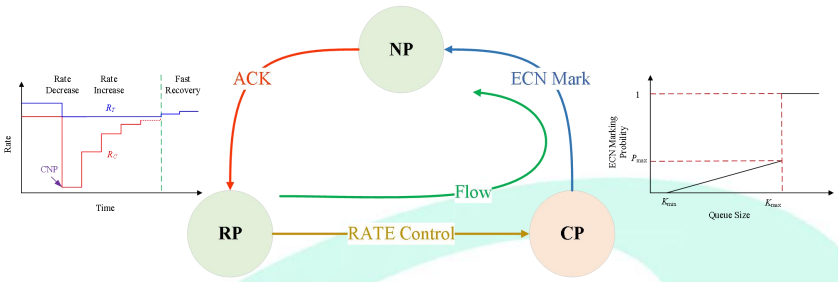



图 8 DCQCN 原理

为了缓解网络拥塞，文献[12]中微软提出了一种端到端的 RoCE 拥塞控制协议 DCQCN，这也是近几年来 RoCE 高性能网络拥塞控制技术的开端。DCQCN 相比于 PFC 是一种更精细的控制算法。它以 ECN（Explicit Congestion Notification，ECN）作为交换机拥塞程度的量化标记信息，根据生成的 CNP（Congestion Notification Packet）报文来触发式的调节网卡传输速率。DCQCN 的设计理念结合了既有的 QCN[13]和 DCTCP[14]的算法思想。一方面避免了 QCN 方法局限于 L2 网络的缺点，另一方面降低了 DCTCP 中 ACK 报文的通信开销。DCQCN 的使用大幅度缓解了 PFC 的触发，目前仍是最广泛应用的 RDMA 网络拥塞控制技术。

## 2. 基于时延的拥塞控制

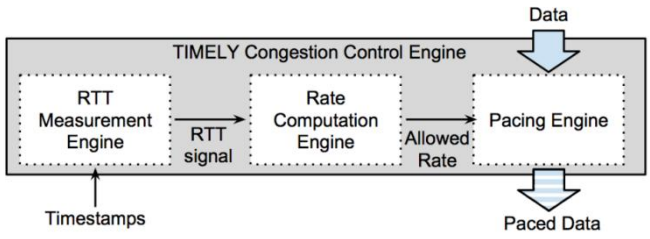



图 9 TIMELY 原理[15]

同期，谷歌在文献[15]中提出了一种基于时延的拥塞控制方案 TIMELY。TIMELY 使用数据流的往返传递时间（Round Trip Time，RTT）作为量化链路拥塞的信息，并设计了一套相应的梯度调速算法。相较于传统的软件测量的 RTT，谷歌方案在他们的智能网卡中集成了专有的 RTT 硬件测量电路，这使得 RTT 测量拥塞的方案得以实用化。同时 RTT 相比于 ECN 是一个快速、多位的数据，能够提供更丰富的网络拥塞信息。

### 3. 基于 INT 的拥塞控制

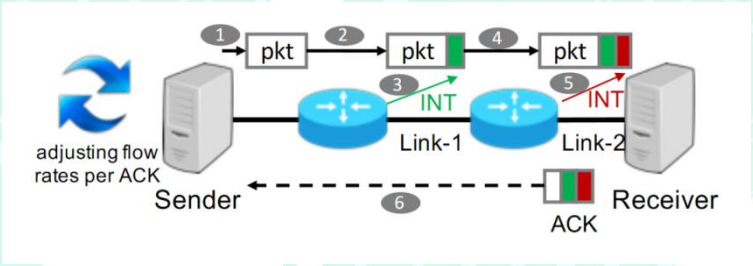



图 10 HPCC 原理[16]

微软的 DCQCN 和谷歌的 TIMELY 在 RDMA 网络拥塞控制方面虽然各有所长，但仍存在各自难以突破的局限性。2019 年，阿里云提出了一种基于带内遥测（In-Net Telemetry，INT）的拥塞控制协议 HPCC[16]。相比于 DCQCN 和 TIMELY，HPCC 方法牺牲了一定的带宽引入了 INT 能力，同时也获得了超高精度的拥塞控制性能。HPCC 可以实现快速的算法收敛以更优的利用闲置带宽，同时保持交换机始终处于近零队列，从而实现超低的数据传输延迟。

### 4. 其他技术方案

尽管 HPCC 在处理拥塞方面的性能得到了普遍的认可，但它过高的网络带宽占用仍然为后续的技术改进留下了空间。此后，更多的技术方案进入了学术界的讨论。

其中，一类是基于已有方案的变种。

(1)ECN 方案变种：文献[17]中提出了 ACC 来自动调节 DCQCN 中 ECN 参数，大幅度降低了运维大规模 RDMA 集群过程中调试算法参数的工作难度；文献[18]旨在改进 DCQCN 在 Incast 场景的性能表现，通过自适应的选取 DCQCN 的参数，实现大规模多打一场景小步长，小规模多打一场景大步长的控制效果；文献[19]中提出了 IRN，通过改变 RoCE 的重传机制和基于 BDP 的算法，从原理上改进 DCQCN 和 TIMELY 方案下拥塞场景的 RDMA 网络性能；文献[20]改进了交换机的 ECN 标记机制，将传统的两态标记优化为三态标记 TCD，提升了 ECN 标记携带的信息量，从而改进了 DCQCN 的控制效果。

(2)RTT 方案变种：文献[21]中建立了 DCQCN 和 TIMELY 的流体模型，分别就二者的拥塞控制效果进行了对比研究。基于二者性能上的差别，其提出了使用 PI 控制器的改进 TIMELY 算法，一定程度上提升了基于 RTT 方法的控制性能；文献[22]为解决 TIMELY 的收敛点不固定这一问题，重新设计了 TIMELY 的调速算法，使用了 AIMD（Additive-Increase Multiplicative-Decrease，AIMD）调速算法，

从而实现了更好的算法稳定性；文献[23]通过将 ECN 信号与 RTT 信号结合考虑，提出了 EAR 拥塞控制协议，在多个场景实现了更好的完成时间性能。

(3) INT 方案变种：文献[24]为了解决 HPCC 中 INT 包头带来的明显的网络带宽占用问题，提出了概率性带内遥测（Probabilistic In-band Network Telemetry，PINT）方案。PINT 使用了概率性编码的方式，大幅度的降低了 HPCC 中网络带宽占用的问题。文中的结果显示，在包头从 46 B 降低到 1 B 的基础上，HPCC-PINT 实现了与 HPCC-INT 接近的流完成时间分布，HPCC-PINT 在长流上略优但在短流上略差。该方案难以大规模部署的局限在于 P4 可编程交换机的资源瓶颈。

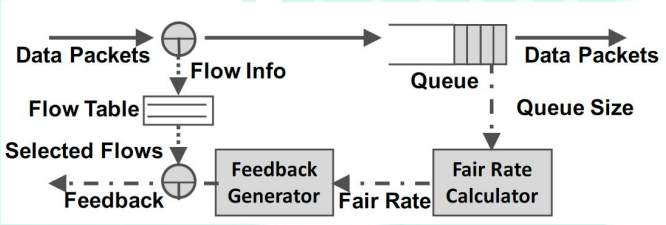



图 11 RoCC 原理[25]

另一类兴起的是基于接收方的方案。思科首先提出了 RoCC[25]。在该方案中，将传统的由发送方驱动的端到端拥塞控制协议，改进为接收方驱动。在文献[25]中认为，多个不同发送方做出的调速决策时常是矛盾的，这导致了最终控制效果的反复波动。而由接收方驱动的拥塞控制协议中，交换机作为拥塞的感知方，可以进行更准确的拥塞反馈。但 RoCC 本身受限于交换机的计算能力和可编程能力，

并未实现大规模商用。其后，文献[26],[27]实现了网卡作为接收方的拥塞控制协议，在 Incast、公平性的场景有明显优势。

总体上看，拥塞问题是 RDMA 网络目前应用的核心难题，而拥塞控制协议作为处理拥塞问题的最短回路在学术研究和工业应用角度都极具研究价值。目前，拥塞控制协议主流应用的方案仍然是 DCQCN、TIMELY 和 HPCC，但显然这三种方案都存在不同的缺陷。后续提出的方案有些偏向解决某一特定流量场景的问题，有些需要更新的硬件支持，难以大规模商用并完全解决 RDMA 无损网络拥塞控制的问题。

## 5. 拥塞控制总结

表 1 拥塞控制协议功能实现对比

| | DCQCN | TIMELY | HPCC | RoCC |
|---|---|---|---|---|
| 公平性 | 公平 | 公平 | 公平 | 公平 |
| 快速收敛 | 慢 | 快 | 快 | 快 |
| 带宽利用率 | 低 | 高 | 较高 | 高 |
| 稳定性 | 稳定 | 局部收敛 | 稳定 | 稳定 |
| 鲁棒性 | 中 | 差 | 中 | 鲁棒 |
| 近零队列 | 不满足 | 不满足 | 满足 | 满足 |
| 兼容性 | 强 | 强 | 差 | 差 |
| 易用性 | 差 | 差 | 易用 | 易用 |

根据上一节中总结的当前拥塞控制协议的核心功能点，本文就当前具有代表性的四种拥塞控制协议：DCQCN、TIMELY、HPCC、RoCC，进行了定性的对比，如表 1 所示。

DCQCN 协议强项在于兼容性，这也直接决定了 DCQCN 成为了现在最为主流的应用方案。具体来讲，DCQCN 协议对于交换机要求较低，只需满足 ECN 标记功能，当前商用 RoCE 交换机已普遍支持。但是，DCQCN 在收敛速度、带宽利用率方面与稳定性存在一个博弈，即往往需要牺牲稳定性来实现更高的带宽利用率，二者不可兼得。其次，DCQCN 在易用性方面有明显缺陷，端侧和网侧有十余个参数需要调试且参数相互耦合，维护成本很高。同时，DCQCN 使用了 AIMD（Additive Increase Multiplicative Decrease，AIMD）调速机制是不基于模型的调速算法。对于不同场景下的扰动，有较强的鲁棒性。但由于输入信息位数受限，进一步增大了维护成本。

TIMELY 协议的优势主要是检测精确，RTT 测量的拥塞信息本身就是多位的，这解决了 DCQCN 算法中最大的一个局限点。因此，TIMELY 在收敛速度和带宽利用率上更高。但是，TIMELY 在公平性和稳定性方面也是存在一组矛盾[21]：收敛到特定稳定点则不能保证公平，保证公平则不能保证达到目标收敛点。其次，TIMELY 的易用性偏差，主要难点在于目标时延的选择困难。同时，TIMELY 采用的是基于模型的控制算法，这保证了精确计算的同时，带来了对于扰动抑制能力差的问题。

HPCC 协议在公平性、收敛速度、稳定性方面都达到了相比于 DCQCN 和 TIMELY 更好的整体效果。同时，HPCC 实现了近零队列控制功能，这有效的保证了网络的平均流完成时间(Flow complete time,

FCT)性能，降低了排队时延。HPCC 主要问题在于，INT 消耗了 10%左右的网络带宽，虽然在 PINT 中有所缓解，但资源的开销被转移到了交换机，这方面的开销导致 HPCC 仍有改进空间。此外，HPCC 对交换机设备的要求较高，需要可编程交换机的支撑，这限制了它的推广使用。最后，HPCC 的算法是以 BDP 精确计算为基础的，这本质上还是基于模型的控制方法。对于更复杂的场景，模型和参数的鲁棒性存在潜在的问题，有待进一步研究验证。

RoCC 协议是针对其他方案的缺陷设计的，其实现上解决了绝大多数问题，性能指标趋于理想。相比于 HPCC，RoCC 在其基础上采用了 PI 控制器，该控制方法是不依赖于模型的，在实际应用中相比 HPCC 具有更好的鲁棒性。但是，RoCC 与 HPCC 存在一个类似的缺点，需要可编程交换机作为运算的主要载体，这对于交换机的性能提出了更高的要求。目前，可编程交换机技术已经开始快速发展，但其芯片架构决定了不适合执行高精度的运算任务。随着未来芯片技术和可编程交换机技术的发展，RoCC 可能是更优的拥塞控制协议。

综上所述，目前的拥塞控制协议很难在兼容、易用的基础上，还具备完善的功能、性能。

## （二）链路控制技术

链路控制技术主要通过速率限制和流量控制来实现，可以动态调整 RDMA 网络中的数据发送速率，避免拥塞的产生,从而提高网络效率。链路控制对于构建可靠、低延迟的 RDMA 网络至关重要。

相比于 L4 传输层的拥塞控制协议，链路控制协议工作在 L2 链路层。相应的，链路控制技术响应的时间相比拥塞控制更加短，速率的调节也更加及时。对应的拥塞检测机制也明显有别于 L4 层面，要求响应的速度更快。链路层控制中，往往并不获取网络内的拥塞程度信息，而是用事件触发产生的报文进行进行控制。

因此，链路控制中的拥塞检测与控制结合的更紧密，往往就是一组报文的交互过程，而非细化到速率的计算与调整。

本节受限于篇幅，主要介绍 Infiniband 中常用的信用机制和 RoCE 网络中常用的 PFC 和 QCN 机制。

### 1．信用

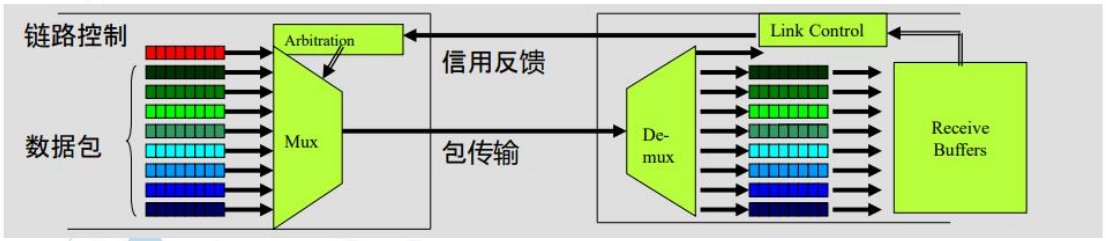



图 12 信用机制原理图

在 InfiniBand 高性能网络协议中，提供了一种基于信用的链路控制（Credit-Based Flow Control）机制，用于优化数据传输。在 InfiniBand 中，每个端口都有一个缓冲区，用于存储接收到的数据包。当发送端发送数据包时，它会向接收端发送一定数量的信用（Credit），表示接收端有多少可用的缓冲区来存储数据包。在接收端，当缓冲区被占满时，它会向发送端发送信号，表示不能再接收更多的数据包，从而避免了网络拥塞的发生。

基于信用的链路控制可以显著提高网络的吞吐量和性能。它可以避免数据包的丢失和重传，并减少网络拥塞和延迟。此外，由于每个端口都有自己的缓冲区，它也可以实现流量隔离和保障，从而提高网络的可靠性和安全性。

通常，信用机制这种传输控制与 TCP 中的滑动窗口有相似之处，两者都是收发两端协调控制发送数据量的链路层流量控制机制。相比滑动窗口使用 ACK 报文确认的方式，信用机制为了更高的性能，设计了特殊信用帧来反馈未使用的信用量。

信用机制目前是 Infiniband 定制的链路控制技术。这一技术目前未在开放标准的以太网中广泛应用，这也导致了 RoCE 网络中普遍需要使用 PFC 来保障网络的无损特性，同时也导致了 RoCE 网络没能继承 Infiniband 中细粒度拥塞控制的底层基础。

## 2. PFC

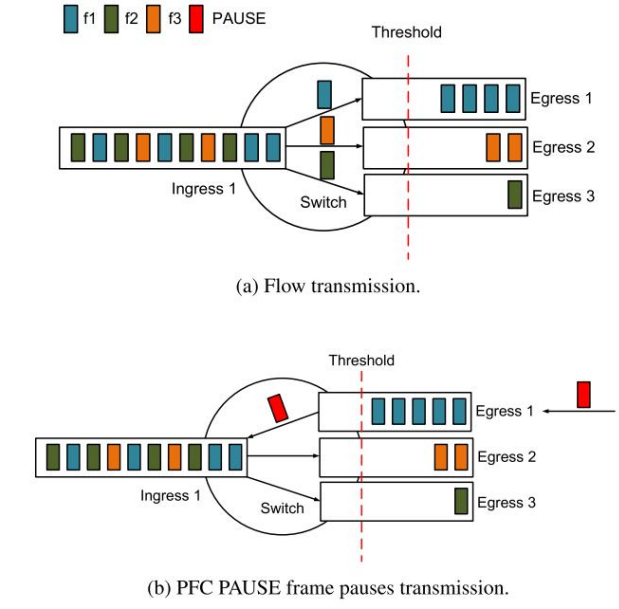



图 13 PFC 控制原理图

优先级流控制（Priority-based Flow Control，PFC）[28]是 IEEE 802.1Qbb 定义的一项用于数据中心的无丢包网络流量控制协议，主要用于确保网络的无损特性。无损网络意味着网络不会因为拥塞而导致数据包丢失。上图展示了在交换机层级之间实现 PFC 的示例。

PFC 通过 Pause 帧触发反压的方式实现无丢包。当交换机队列接近满时（达到 ON/OFF 阈值），交换机将向上游交换机发送一个 Pause 帧，告知上游不要继续发送数据包。待拥塞缓解后，再通知上游继续发送数据包。同时，PFC 通过虚拟队列将数据包分成不同的优先级。即使某个优先级受到拥塞阻塞，仍然可以通过更高优先级发送数据包，以确保重要分组的及时传递。

PFC 解决了无丢包的问题，但会带来一系列问题:

a)导致出现受害者流，无法正常传输数据，如上图中(b)所示，Egress 2 和 3 的数据包尽管未发生拥塞，也会被停止发送，即 HoL（Head of Line）阻塞。

b)PFC 风暴，Pause 帧会反向逐级传递，形成网络内大量的设备停止发送.

c)PFC 死锁，当系统中出现 Pause 帧的 CBD（Circular Buffer Dependency）现象时，PFC 发生死锁导致网络传输长时间中止，如下图。

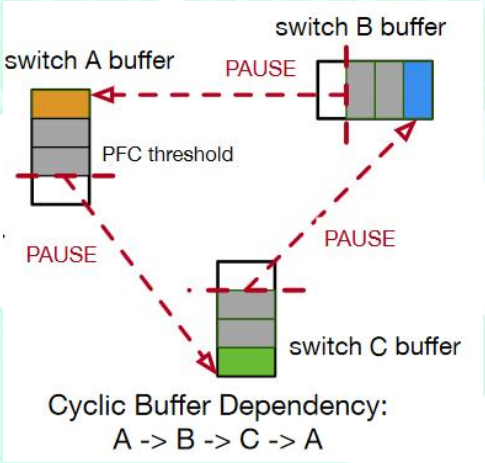



图 14 PFC 死锁[11]

总而言之，PFC 技术是 RoCE 高性能网络发展的重要过渡技术。它保障了 RoCE 网络可以进入生产部署阶段，但它的一系列技术上的缺陷又局限了 RoCE 网络的大规模部署。

近年来，就如何摆脱 PFC 限制，学术界和工业界有大量的研究，包括 PFC 自恢复机制、更高效的拥塞控制协议、选择重传[19]等一系列的解决方案。相关技术的后续研究进程将决定 RoCE 高性能网络与传统 Infiniband 高性能网络在技术上的实质差距。

### 3. QCN

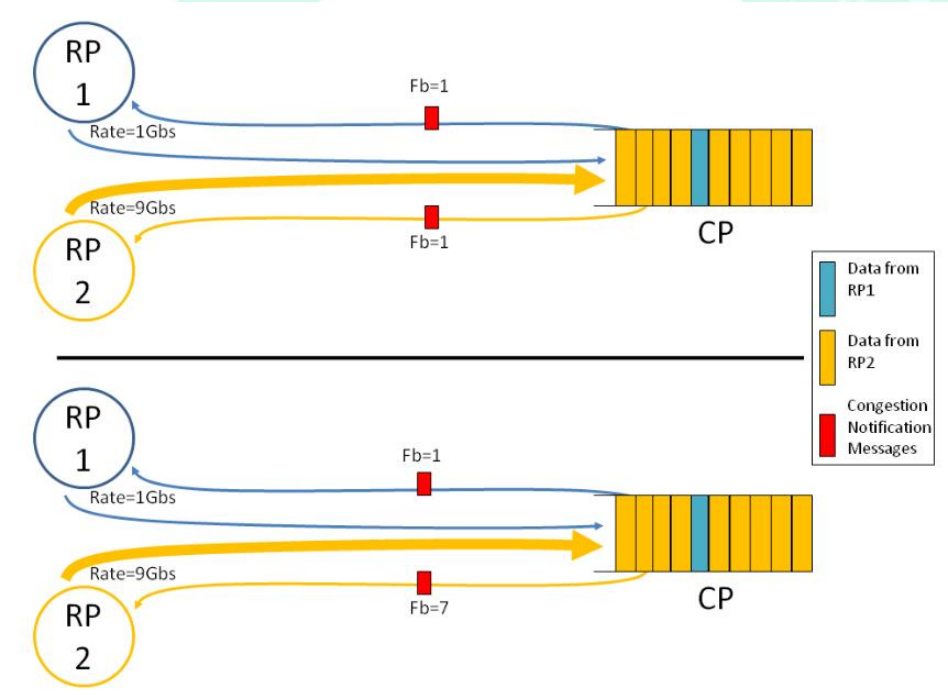



图 15 QCN 技术原理

QCN[13]是一种 L2 链路层的网络拥塞控制技术，旨在提高网络的性能和可靠性，避免网络拥塞引起的数据丢失、延迟和抖动等问题。QCN 通过在交换机中实时监测网络拥塞情况向终端设备发送拥塞通知，从而调整数据传输速率，以适应网络的拥塞状况。

具体流程上，一旦中间交换机或者路由器发生拥塞，拥塞点会：

从交换机读出队列长度；

基于队列长度信息计算反馈的值;

格式化一个特殊反馈值的 QCN 帧，使用 源 MAC 地址将该帧返回 QCN 源端;

根据 QCN 算法指定的动态信息更新队列的采样速率，通过拥塞通告可以一定程度上解决拥塞热点的问题。然而在真实环境中很少实现，因为它会高度依赖拥塞点反应时间，通过网络发送 QCN 帧的时间和反应点反应时间；并且它只能运行在二层网络上，很难适应数据中心大量的三层隧道功能。

同时，如能解决设备支持的因素，将 QCN 与 DCQCN 结合实现一个类似于 IB 的链路层+传输层拥塞控制方案，将是对 RoCE 网络缺少 credit 链路控制技术的一个有力补充，这都有待于进一步的研究和工程实践。

**4. 链路控制总结**

在链路控制层面，由于现有的 RoCE 协议中缺少了 Infiniband 协议中的信用机制，这一定程度上破坏了原有的双环控制系统的完整性。用 PFC 替代细粒度的信用机制，在整个控制系统的角度看，可以认为是将一个内环控制系统置换成了一个非线性死区，这一定程度上降低了整个系统的阶数。

从实际的网络角度意味着传输速率的调节，RoCE 网络相比 Infiniband 会更加的迟钝。这个技术上的差距是不能通过调节 DCQCN 参数弥补的。因此，RDMA 网络在以太网上的应用更为重要的

技术突破要在链路控制的层面进行。但同时，链路层的改动需要对协议内容进行更深入的探讨和精巧的设计，同时需要在硬件层面进行大量的研发工作，其研发周期往往更长。

### （三）负载均衡技术

负载均衡技术通过在网络中多个服务器或链路间分配流量，实现网络流量的均衡分配，防止热点的产生。负载均衡机制的核心目标在于利用好网络中冗余的传输路径，使得流量均匀分布。

负载均衡技术对于网络拥塞起到了有效的缓解作用，从拥塞的源头上，降低了拥塞发生的概率。本节按业内通用的分类，按流级别、包级别、flowlet 级别对常见的典型负载均衡机制进行了分类总结。

#### 1．流级别

ECMP 全称等价多路径（Equal-cost multi-path），它是一种基于流的负载均衡路由策略。当路由器发现同一目的地址存在多条等价路径时，路由器会依据相应算法将不同流量分布到不同的链路上，以增进网络带宽利用率。



ECMP 的路径选择策略有多种算法:

哈希，例如利用流的五元组哈希为流选择路径;

轮询，各个流在多条路径之间轮询选择;

基于路径权重，根据权重系数，系数大的分配流量多。

ECMP 是一种简单的负载均衡策略，但在实际应用中存在许多问题。

a) 可能加重网络链路的拥塞问题。由于 ECMP 仅使用哈希或轮询方法进行负载均衡，它无法感知到链路的拥塞情况。因此，在已经存在拥塞的链路上使用 ECMP 可能会进一步加剧拥塞情况。

b) ECMP 无法解决非对称网络的性能损失。当数据中心网络发生故障时，网络结构可能会出现非对称情况导致无法实现网络物理链路的均衡分布，进而造成流量不平衡的问题。

c) 在流量大小分布均匀的情况下，ECMP 效果较好。然而，在同时存在大流量和小流量的情况下，ECMP 的效果并不理想。假设有一条大流量和一条小流量同时到达路由器，ECMP 会将这两条流量均匀分配到等价路径上，但显然这种情况下等价路径并没有被高效利用。

因此，尽管 ECMP 是一种简单的负载均衡策略，但它存在上述问题，限制了其在某些场景下的有效性。在解决这些问题的同时，可以考虑使用更复杂的负载均衡策略或结合其他技术来改善网络性能和流量分配的均衡性。

因此，将 ECMP 直接部署在数据中心这种突发流量多，大象流与老鼠流并存的环境中，需要仔细考虑环境的问题。尽管后续的研究（如 Hedera[29]，BurstBalancer[30]）很多考虑了不同的数据流

量特征（大象流、burst 等）。但是，ECMP 由于其工程复杂度低、性能可接受仍然广泛应用在数据中心网络中。

### 2. 包级别

随机包喷洒（Random Packet Spraying，RPS）是一种基于包级别的负载均衡策略。当路由器发现有多条等价路径指向同一目的地址时，RPS 会将数据包以单个包为单位分散到这些路径上。与 ECMP 不同，RPS 以数据包为单位进行操作，而 ECMP 则是以流为单位进行操作。RPS 将同一流中的不同数据包转发到不同的等价路径上。

RPS 的优点在于简单易实施，并且能够充分利用网络链路。在没有突发流或流大小差异的情况下，RPS 能够避免网络出现不均衡的情况，能够实现更好的负载均衡并提高网络性能。同时，RPS 也有一些限制。由于数据包的随机分布，可能会导致同一流中的数据包到达目的地的顺序不同，这可能对某些应用程序造成影响。此外，RPS 对网络中的路由器和交换机的支持程度也可能存在差异。

RPS 技术往往需要 RDMA 网卡在传输层支持乱序传输，这对于当前市场上已有的 RNIC，是一个相对苛刻的硬件要求，这也导致了当前 RPS 方式负载均衡的使用范围。



### 3. Flowlet 级别

流级别的太过粗糙，包级别的粒度太细。Flowlet 作为一个折中方案，就成为了一个研究点。M. Alizadeh 等人于 2015 年提出 CONGA[31]，它是一种基于网络的分布式拥塞感知负载平衡系统。其设

计目标是在不增加传输层复杂度的前提下，通过分布式方式实现全局负载平衡。

CONGA 基于数据中心网络的特点将流进一 步细分为间隔粒度在微秒级别的小流(Flowlets)，负载均衡也针对每一个 Flowlet 的第一个包，之后每个 Flowlet 使用相同的链路。上行链路交换机搜集链路拥塞状况并交给收端交换机，保存一个来自各叶节点的拥塞状况，并反馈给源端交换机。CONGA 通过负载均衡提升了数据中心网络传输性能进而提高吞吐量，但 CONGA 仍然需要网络负载与实际容量相匹配。当实际容量无法满足时，CONGA 的性能无法得到保证。

此外，这一领域的研究内容也逐渐细化，但整体上讲，应用范围相比 ECMP 更加局限。

### 4. 负载均衡总结

负载均衡作为网络领域的一个传统问题，旨在合理分配网络资源以实现流量的均衡和优化性能。无论是在传统的 TCP 网络还是在 RDMA 网络中，负载均衡都是一个重要的考虑因素。

在传统的 TCP 网络中，负载均衡通常通过在网络层或应用层进行实现。常见的负载均衡方法包括基于轮询、哈希算法、最短队列优先等。这些方法旨在将传入的请求或数据流量分发到多个服务器或网络路径上，以实现负载均衡和避免单一节点或路径的过载。

在 RDMA 网络中，负载均衡的目标和原则与传统 TCP 网络并无区别。然而，由于 RDMA 网络具有独特的特性和协议，需要特别考虑一些因素。RDMA 网络的部署和配置可能涉及特定的硬件和软件要求。负载均衡解决方案需要考虑 RDMA 适配器、交换机和路由器的支持程度，以及与 RDMA 协议栈的集成。对于无损的 RDMA 网络和有损的 RDMA 网络，其负载均衡机制上仍应有不同。结合未来更多的高性能网络应用场景，负载均衡机制上仍有优化的空间，与拥塞的实时检测信息结合可能是未来的潜在研究方向。

### （四）流量调度技术

流量调度技术不同于前述的几项技术，它更多的是在给特定的数据流量进行优先级调配，解决全局的网络服务质量问题。

数据中心中的流量调度技术主要通过软件定义网络(SDN)来实现。SDN 控制器整合拓扑发现模块和流量监控模块获取全网视图，再根据业务优先级、网络状态、服务器负载状态等，用开放流协议(OpenFlow)下发流表规则到数据平面，协调网络设备实现动态的流量调度。这可以按需分配网络资源，绕过拥塞链路，根据业务需求分割带宽，还可以按照负载将流量导向闲置服务器。流量调度提高了网络利用率，保证了关键业务的质量。

目前来看，流量调度技术分为两类，一类是开环的管理与分配，这种往往由设置特定的规则来进行调度。本节以调度的特点来分类，对流量调度的典型技术进行了归纳。

## 1．基于规则的调度技术

基于规则的流量调度技术在学术界讨论的非常广泛，例如，pFabric[32]，PDQ[33]，PIAS[34]，FastPass[35]，Homa[36]，AuTo[37]等。它们的研究方法通常是设定特优的规则来给不同的数据流进行优先级分类。例如，对较长的流，减少包丢弃；对时延敏感的流进行优先级排序，平衡网络负载，使其快速通过。

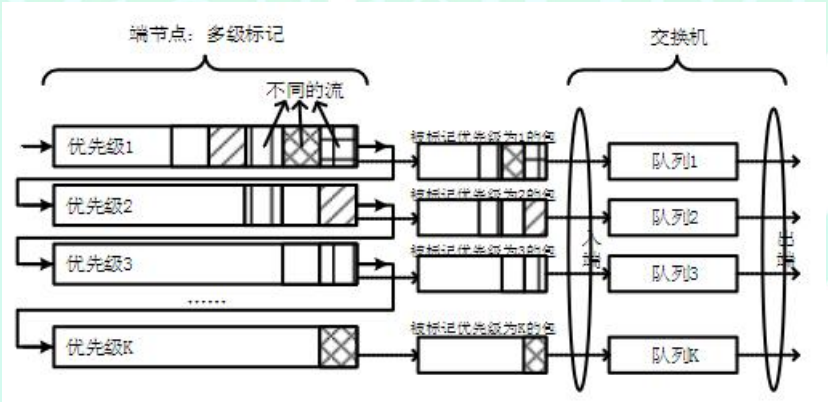



图 16 PIAS 调度原理[38]

由于流量调度的研究较多且领域更加细化，与应用结合较多，但总体的研究思路是类似的。限于篇幅，此处以 PIAS 举例说明。PIAS 是一个流调度算法，目的是达成短流高优先级，长流低优先级。PIAS 假定流分布是不可知的，从而寻求将 FCT 最小化。

PIAS 借鉴了模拟最短作业（SJF）工作原理来最小化 FCT。利用现有交换机中可用的多个优先级队列来实现多级反馈队列（MLFQ）。在这种队列中，PIAS 流会根据其发送字节数逐渐从高优先级队列降级为低优先级队列。因此，短流能在前几个高优先级队列中完成，这使得 PIAS 能够在无法预先知道流量大小情况下模拟 SJF。

如图 15 所示，PIAS 只部署在端节点，而不用在交换机上部署。问题的难点在于如何准确划分包长阈值 K，以及端与端之间的配合。阈值不准确和优先级之间失匹配，都会导致性能损失。文中，作者虽然通过建模给出了如何准确计算阈值和解决失匹配问题，但是仍然需要很长的时间使模型收敛。同时，它的优先级匹配的过程，存在慢启动的问题。即长流可能需要很长时间才能达到一个准确的优先级，导致性能损失。

PIAS 论文的研究工作中，我们可以看出基于规则的流量调度技术更多的是需要针对特定的网络流量环境，通过精细的平衡各方面的 tradeoff，实现某种边界条件下的网络性能最优化。这类工作在科研学术的角度，有研究价值，也可以提升封闭环境下的网络性能。但是在更广泛应用的云数据中心中，失去了流量特征的种种理想假设，这种敏感的规则平衡几乎不可能达成。这也一定程度上限制了流量调度技术的落地和广泛应用。

### 2．基于反馈的实时调度

流量调度本质上与交通、物流等领域的调度问题没有区别，而实时调度在这些领域中有广泛的应用。而网络流量的调度技术中，结合反馈信息的实时调度研究相对较少。其根本原因在于，网络环境相比传统的交通、物流，信息更难测量获取。

一些研究确实也引入了一些实时调度的思路，例如 D2TCP[39]、D3[40]等基于完成时间的流调度技术。这类技术在整体上将原有的从流量的特征制定规则的角度转到实时的完成的截止时间上。

但是，不管是 D3 还是 D2TCP，使用的时间都是截止时间和已发送时间。这种本质上虽然也形成了回路，但系统的传感器是本地时钟，这种实时的反馈调度仍然是规则化的。如何利用更多的网内信息，实时的调整调度规则，是未来研究的一个潜在方向。

### 3．流量调度总结

流量调度相比于拥塞控制、链路控制、负载均衡，更接近用户层，与业务耦合更紧密，能有效的优化特定业务场景下的业务服务质量。但是，由于其软件调度的局限性，它很难完成快速的拥塞避免，这一特点也决定了它需要的检测手段要具有周期长且准确特点。

同时，基于简单规则的调度技术难以在复杂的流量环境下广泛的应用，流量调度技术更明显的在向实时调度的方向演进。随着拥塞检测技术的进步，更丰富的网络实时信息将给流量调度技术带来

更大的操作空间，也给流量调度技术在未来云数据中心的更广泛应用提供了更多的契机。

### （五）本章小结

本章归纳了高性能网络中，用于处理和缓解拥塞的技术体系，主要包括拥塞控制和链路控制组成的拥塞控制技术，负载均衡和流量调度组成的拥塞管理技术。拥塞管理和控制的技术体系，目前仍然是高性能网络的核心，将更为合适拥塞检测技术更为广泛的集成到特定的管控层面，是未来高性能网络的一个重要课题。

同时，在本章中，讨论了各项技术与拥塞检测技术已有以及潜在的结合点。总体上看，拥塞检测在硬件实现更多、响应速度更快的拥塞控制协议、链路控制协议中应用更加广泛。但随着网络观测技术的进步，在负载均衡、流量调度技术方向上，也有较大的潜在应用。

# 三、 高性能网络拥塞检测技术

拥塞检测技术本身的出现早于高性能网络，ECN、RTT、INT 等拥塞测量的方案在传统的 TCP 网络中就已经被广泛探讨[14], [41], [42]，以优化 TCP 网络的传输效率。但在高性能网络，特别是 RoCE 协议下，拥塞问题的重要性进一步上升，也对拥塞检测技术提出了更高要求。

在高性能网络的拥塞管理与控制技术体系中，存在一个直观规律。以拥塞控制协议为例，不同的拥塞控制协议往往对应不同拥塞检测技术，如 DCQCN 对应 ECN、TIMELY 对应 RTT、HPCC 对应 INT 等。

由此可见，拥塞检测在拥塞控制方案中是决定性的。各种不同协议中控制器的算法之所以存在区别，归根结底是拥塞检测的实现方案区别。例如，DCQCN 中使用 CNP 报文的事件驱动型控制，其算法设计上采用 AIMD 来进行逐拍的控制；TIMELY 使用 RTT 作为拥塞信息，其算法则可以使用 PID 进行线性控制。拥塞检测方案是整个协议的起点，也是不同的拥塞控制方案的本质区别。

同时，在之前的研究工作中，拥塞控制的设计缺乏系统性的思考，检测环节、处理环节、控制环节通常没有细分的设计。这也导致了控制算法设计上很多与检测环节强耦合，工程实现上缺乏通用性，且参数繁多。

因此，本章对当前的拥塞检测技术进行系统的归纳，主要以拥塞检测的主体为分类依据，以交换机、网卡、端网协同三个类别，分别对高性能网络的拥塞检测技术的演进思路开展探讨。

## （一）网侧拥塞检测

在高性能网络中，交换机是拥塞发生最为频繁的设备节点。因此，交换机设备的拥塞检测结果，往往代表了整条传输链路中最大拥塞程度。本节中，以几个典型的 RDMA 网络交换机拥塞检测技术为切入点进行了归纳总结。

### 1. ECN 检测

显式拥塞通知[43], [44]（Explicit Congestion Notification，ECN）是对 Internet 协议和传输控制协议（TCP）的扩展，定义在 RFC 3168（2001）中。ECN 允许在不丢弃数据包的情况下，通知网络拥塞的发生。ECN 在以太网中是一个可选的功能，在底层网络基础设施也支持的情况下，可以在两个支持 ECN 的终端之间使用。

ECN 在 Internet 层和传输层都需要特定的支持，在 TCP/IP 中，路由器在 Internet 层运作，而传输速率由传输层的端点处理。拥塞可能仅由发送器处理，但由于只有在发送了一个数据包之后才知道发生了拥塞，因此接收器必须将拥塞指示回传给发送器。在没有 ECN 的情况下，拥塞指示回传是通过检测丢失的数据包间接实现的。而有了 ECN，拥塞通过将 IP 数据包中的 ECN 字段设置为 CE（Congestion Experienced）来指示，并通过接收器在传输协议的头部设置适当的位来回传给发送器。

ECN 广泛的应用于以太网络，因此在 RoCE 高性能网络协议中，ECN 作为拥塞检测方案存在广泛硬件基础。事实上，应用广泛的

DCQCN 和 DCTCP 协议都使用了 ECN 作为其拥塞检测方案，其参数设置上并不完全相同，如下图所示。

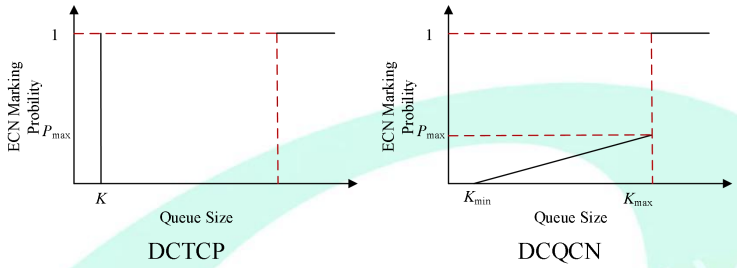



图 17 ECN 标记参数含义

ECN 在使用的过程中存在不同的标记算法区别，以下给出了典型的 RED 和 Blue 标记算法。

1）RED 标记

ECN 在使用的过程中，通常与 RED 功能结合使用。RED 是由 Sally Floyd 和 Van Jacobson 在 1990 年代初发明的交换机队列管理机制[45]。RED 会监控平均队列大小，并根据统计概率丢弃（或在与 ECN 结合使用时标记）数据包。

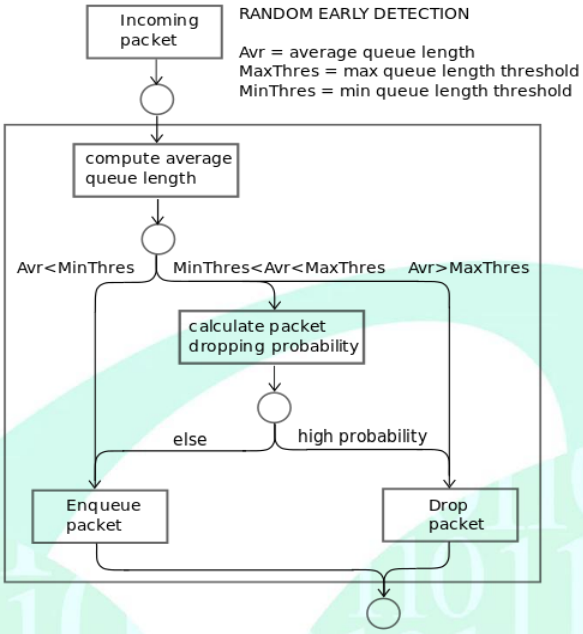



图 18 RED 标记原理

随机早期检测（Random Early Detection，RED）是一种适用于拥塞避免的网络调度器队列规则。

在传统的尾部丢弃算法中，路由器或其他网络组件缓存尽可能多的数据包，并简单地丢弃无法缓存的数据包。如果缓冲区不断满载，表示网络拥塞。尾部丢弃不公平地分配缓冲区空间给各个流量流。尾部丢弃还可能导致 TCP 全局同步，因所有 TCP 连接会同时"退缩"并同时"前进"。网络会交替地被低效利用和淹没，形成波动现象。

RED 通过在缓冲区完全满载之前预先丢弃数据包来解决这些问题。它使用预测模型来决定要丢弃的数据包。如果缓冲区几乎为空，则接受所有传入的数据包。随着队列的增长，丢弃传入数据包的概率也会增加。当缓冲区已满时，概率达到 1，所有传入的数据包都会被丢弃。

RED 后续结合 QoS 等信息，进一步衍生出了 WRED（Weighted RED）和 RIO(RED with In and Out)。

2) Blue 标记

RED 是一种依赖队列长度的标记算法。根据排队论中的著名结果，只有当数据包的到达时间服从泊松分布时，队列长度才直接与活动源的数量和真正的拥塞水平相关。不幸的是，在网络链路上，数据包到达时间往往不服从泊松分布。

Blue[46]是一种网络调度器的调度策略，由密歇根大学的研究生冯武昌（Wu-chang Feng）为 Kang G. Shin 教授以及 IBM Thomas J. Watson 研究中心的其他人在 1999 年开发而成。

Blue 使用了数据包丢失和链路利用率历史来管理拥塞。通过维护一个单一的概率，用于在数据包排队时标记（或丢弃）数据包。如果由于缓冲区溢出而导致队列持续丢包，Blue 会增加标记概率，从而增加发送拥塞通知的速率。相反地，如果队列变为空或链路处于空闲状态，BLUE 会减小其标记概率。BLUE 相对于 RED 在减少数据包丢失方面有一定的优越性，即使在使用较小的缓冲区时也是如此。基于 Blue 的机制，还提出并评估了一种新的机制，用于有效地和可扩展地实现大量流之间的公平性。

当前的交换机中普遍使用 RED 与 ECN 结合，通过 RED 标记机制生成 ECN 标记。但随之 Lossy 逐渐在主流 RoCE 网卡中普及，对于未

来的高性能网络，Blue 以及近年来的一些其他主动队列管理的方法可能存在更多的应用场景。

## 2．TCD 检测

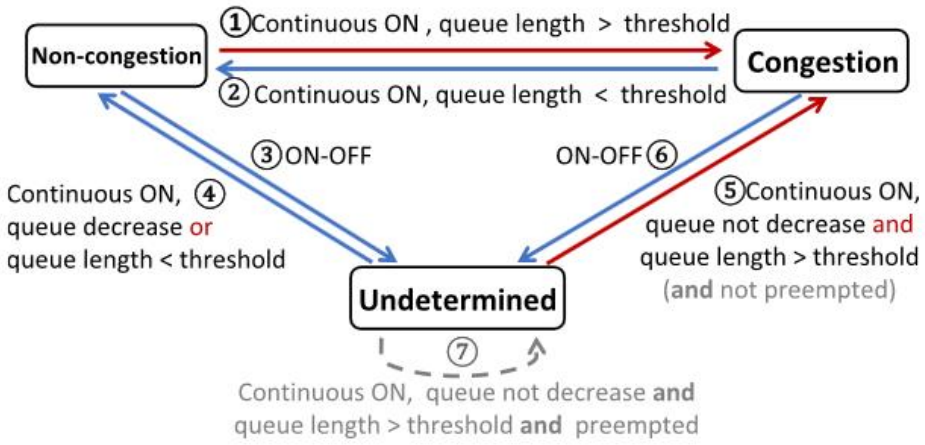



图 19 TCD 的状态转移原理图[20]

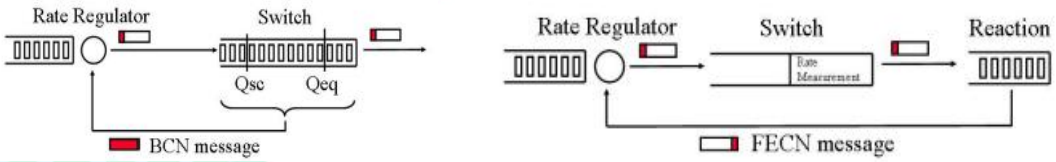
在文献[20]中，意识到了拥塞检测对于整个系统的重要作用。该论文的观点中，ECN 检测给出的结果是具有 ON-OFF 特性的，而这种 ON-OFF 发送模式可能对交换机中的拥塞检测行为产生意外影响，包括引起队列积压并影响暂停端口的实际输入速率。

以此为启发，该论文中提出了一种三元拥塞检测技术来实现 RDMA 网络的拥塞检测。它将网络设备的端口状态不再用 0-1 标记，而是区分成三种状态，拥塞、非拥塞和不确定。这三个状态用上图所示的方式进行转移。

尽管 TCD 意识到了拥塞检测对于高性能网络拥塞的关键作用，但其工作仍然是受限于状态的转移。本质上讲，TCD 是通过扩展 ECN

标记的数据位宽实现更准确的拥塞检测。相比于其三元状态的转移逻辑的设计，该研究工作中体现出的高位宽带来搞检测精度最终带来拥塞控制系统性能的提升结果更加令人振奋。

### 3．其他检测技术




（a）BCN 系统模型　　（b）FECN 系统模型

图 20 BCN、FECN 拥塞检测

相比于 TCD 和 ECN 在 RDMA 网络中的应用，一些在以太网中也定义的其他的拥塞通知协议如 Backward Congestion Notification（BCN）[47]，Forward Explicit Congestion Notification[48]，Pre-Congestion Notification[49]等，在 21 世纪初的 IEEE 802.1Qau 工作组中都有广泛的讨论。这些拥塞检测技术同样有应用到 RoCE 网络中的潜在可能。

## （二）端侧拥塞检测

交换机和断网协同完成拥塞检测，往往意味着需要特定的交换机支持。由于大规模的云厂商对于设备的供应链、兼容性、规范性都有苛刻的要求，专属定制的交换机对于大多数云服务提供商来说是难以接受的。因此，不依赖交换机，在端侧通过云厂商自研的

DPU、网卡就能独立完成拥塞检测的方案，对于云数据中心也具有很大的潜在应用价值。

### 1．RTT 检测

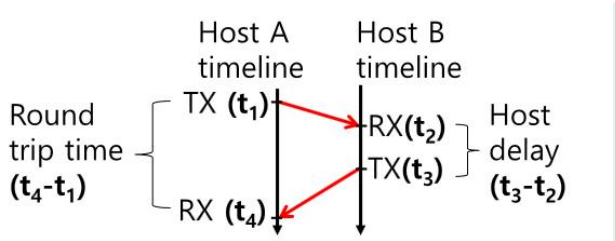



图 21 RTT 测量原理

RTT（Round Trip TIme）是数据包传输的往返时间，在应用于 RDMA 高性能网络之前，就被广泛的应用于 TCP 网络协议的传输控制中，如 BBR、Reno、Vegas 等，其表达式如下:

$$RTT = \mathrm{t_{completion}} - \mathrm{t_{send}} - \frac{\text{seg. size}}{\text{NIC line rate}}$$

使用 RTT 进行拥塞检测遵循一个朴素的哲学，即发生拥塞的链路中数据包的传输时延增大。当网卡实测的 RTT 与端到端正常的 RTT 发生变动时，RTT 的变化量即代表了数据包排队引起端到端延迟。

在 ECN 方案中，具有不同优先级的多个队列共享同一个输出链路，但 ECN 标记仅提供了超过阈值的队列的信息。低优先级的流量可能会经历较大的排队延迟，而不一定会积累大量的队列。此外，ECN 标记描述了单个交换机上的行为。在高度利用的网络中，拥塞发生在多个交换机上，而 ECN 信号无法区分它们之间的差异。

相比于 ECN，RTT 积累了关于端到端路径的信息，包括可能出现拥塞的网络接口卡（NIC）。RTT 提供的信息是一个更为精炼、聚合了端到端拥塞的最终量化指标，从这个角度讲，RTT 是更为直接的测量指标。

但是，文献[21]也对 RTT 和 ECN 方法进行了客观的对比，RTT 测量还是会存在对时钟抖动敏感等问题。同时，使用 RTT 测量的 TIMELY 的算法也存在设计上的问题，导致同期提出的 RTT 方案相比 ECN 方案不具备优势。

但是，RTT 相比与 ECN，在测量的角度仍有其优势。在解决了控制器设计的问题后，其方案简单、测量精度高、端到端、设备依赖弱等特性，使其在未来云数据中心的应用前景更加广阔。

## 2．优先级队列检测

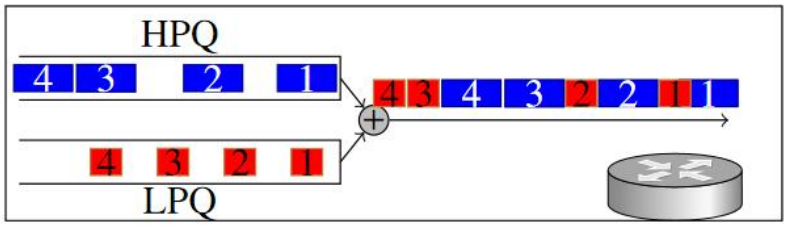



图 22 优先级队列测量原理[52]

文献[52]中提出了一种使用不同优先级队列消息，实现拥塞检测的技术方案。该方案中，使用商用交换机中可用的基本特性（优先级队列），无需对交换机进行修改或在主机上实施任何复杂的算法。它使用了 Scout 服务技术，Scout 服务基于一个简单而有效的想法，即使用交换机中的优先级队列来获取拥塞状态。

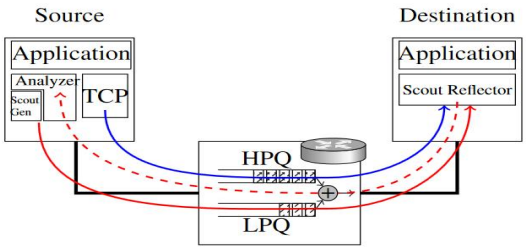



图 23 DWTCP 方案[52]

当高优先级队列（HPQ）变得更加繁忙（链路利用率更高）时，低优先级队列（LPQ）获得的服务机会更少（如图 1 所示）。因此，测量 LPQ 队列中的消息的 RTT 时延，观察到链路的状态，并且，这一检测可以在观察到 HPQ 建立之前几个 RTT 检测到拥塞，相比于传统的拥塞检测技术，能有提前预测拥塞的功能。

### （三）端侧协同拥塞检测

交换机虽然通常是链路中拥塞的瓶颈点，但单独使用交换机完成的拥塞检测，不能对网卡的拥塞程度有直接的测量。因此，一些研究工作中提出，要使用端网协同的拥塞检测方案来实现全链路的拥塞检测。

#### 1. INT 检测

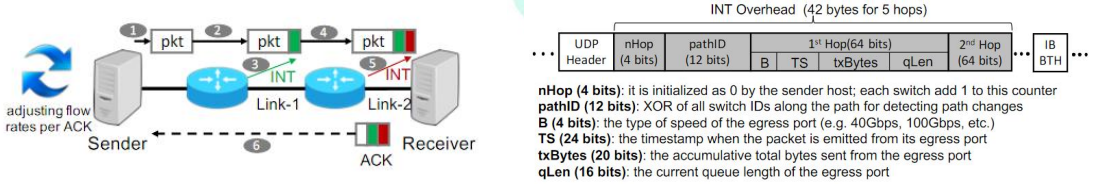



图 24 INT 检测及在 HPCC 中的应用[16]

当前，交换机在数据平面上变得更加开放和灵活。其中，网络内部遥测（In-Network Telemetry，INT）正在迅速普及。我们所了解的几乎所有交换机供应商都已经在其新产品中启用了 INT 功能（例如，Barefoot Tofino 、Broadcom Tomahawk3 、Broadcom Trident3 等），越来越多的商用交换机也在逐步支持 INT。

通过 INT，发送者可以通过 ACK 数据包准确了解流经路径上链路的负载情况，从而便于发送者进行准确的流量调整。例如，在如图所示的 HPCC 拥塞控制协议中，阿里云的研究人员通过自定义了拥塞检测的 INT 报文，准确的获取了链路中网络设备的队列长度、时间、发送数据总量等信息，实现了全链路细粒度拥塞检测。

但 INT 带来检测精度提升是通过增加包问头的形式完成的，这就造成了一定的带宽浪费。如何平衡 INT 带来的检测精度提升和造成的 overhead，成为了使用 HPCC 不得不考虑的一个问题。后续的研究工作中，文献[24]提出了概率添加的 PINT 来减少 INT 的 overhead。但是，这项工作并未完全解决 INT 带来 overhead 的问题，而是给精度和 overhead 这一组 tradeoff 提供了一个归一化的调节参数，如何平衡这一矛盾仍是一个问题。



### 2．ECN#检测

ECN 已经在生产数据中心广泛使用，以提供高吞吐量和低延迟的通信。尽管取得了成功，但之前基于 ECN 的传输机制存在一个重

要的缺点：在计算瞬时 ECN 标记阈值时采用了固定的往返时间（RTT）值，忽视了实际中的 RTT 变化。

然而，在数据中心中往返时间（RTT）的变化很常见，因为不同的流量通过不同的处理组件，例如网络堆栈、虚拟化管理程序和中间件。与服务内部的流量相比，服务之间的流量经历了来自第四层负载均衡器的额外处理延迟。此外，给定组件的处理延迟也会根据工作负载的不同而变化。据研究显示，这一波动往往会达到 3 倍以上。

文献[50]中提出了 ECN#，它基于瞬时和持续的拥塞状态对数据包进行标记。当满足以下条件之一时，数据包将被标记：

当存在大的瞬时队列时，ECN#会主动标记数据包以避免缓冲区溢出。

当存在持续队列时，ECN#会保守地标记数据包以减少排队延迟。

ECN#是对 ECN 标记的一个补充，它结合了网卡侧的 RTT 信息，使 ECN 标记的阈值设置成动态的。根据文献[50]中的评估，ECN#对于短流的平均流完成时间（FCT）可以降低高达 23.4%（99th 百分位数为 31.2%），同时为大流提供类似的 FCT。

## 3．ConEx 检测

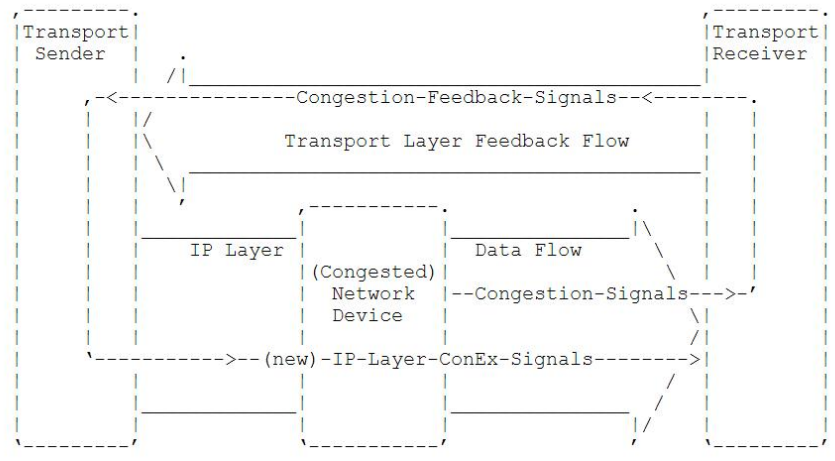



图 25 ConEx 架构[51]

ConEx（Congestion Exposure）[51]是 IETF 标准组织中提出的一种旨在提高网络的性能和可靠性，避免网络拥塞引起的数据丢失、延迟和抖动等问题的拥塞检测技术。ConEx 通过在数据包头部添加拥塞信息，向网络设备和终端设备传递拥塞信息，从而调整数据传输速率，以适应网络的拥塞状况。

在网络中的特定测量点，“剩余路径拥塞”（也称为“下行拥塞”）是一个流预计在测量点和其最终目标之间经历的拥塞水平。“上行拥塞”是在测量点之前经历的拥塞。

如果网络中的流量支持 ECN（显式拥塞通知），则路由器可以在中间节点监测 ECN 信号，并根据该信号量测上行拥塞情况。与之不同的是，ConEx 信号将插入 IP 头中，从源端到目的端包含了整个网络路径中的拥塞情况。因此，如果监测点检测到这两个信号，它

可以通过将 ConEx 的路径拥塞情况减去 ECN 的上行拥塞情况，计算出数据包在监测点和目标之间可能遇到的拥塞情况，也就是剩余路径拥塞。这个计算可以对所有流量进行汇总。

ConEx 目前还没有公开的文献讨论其在 RoCE 网络中的应用，但剩余路径拥塞检测无疑是当前 ECN 检测方案的一个有力补充，它更好的利用了端侧能提供的信息。

## 4. 本章小结

本章归纳了高性能网络中，拥塞检测相关的技术，以网侧、端侧、端网协同为依据，将现有的拥塞检测技术及其典型应用进行了简单的归纳。不同的拥塞检测机制存在明显的优缺点，这决定了其在高性能网络中的应用也需要各有侧重。

同时，在本章中讨论了各项拥塞检测技术设计的本质。总体上看，当前的拥塞检测机制设计上，检测、处理、控制多个环节紧耦合的现状下，拥塞检测机制难以标准化、模块化。需要在后续的数据中心网络领域开展更为系统的研究。

## 四、 总结与展望

随着未来数字经济发展，算力网络宏观战略日益落实，东数西算、大模型等新兴应用场景与算力需求形成了交替驱动的螺旋上升趋势。同时，算力需求的膨胀带动了网络互连带宽需求的直线上升，未来数据中心网络随着云计算的发展将占据更多的市场份额，这都给高性能网络技术的发展提供了新的机遇。

而在这一新的机遇期，RoCE 网络由于其开放兼容的优势，毫无疑问将成为这次技术浪潮中的主角。而相比 Infiniband 传统高性能网络，RoCE 网络在拥塞管理控制技术方面，存在一定的差距。

首先，本白皮书中就高性能网络的背景和现状进行了研究，总结了当前数据中心中分布式存储、内存池化、键值存储、智能算力等场景下高性能网络的应用情况，并分析了高性能网络中的拥塞问题。然后，本白皮书进一步总结归纳了高性能网络的拥塞管理控制技术体系。从网络层、传输层、链路层逐级分解，对已有的拥塞管理控制技术体系进行了深度的剖析。本白皮书中，以网侧驱动、端侧驱动、端网协同为划分依据，对现有的拥塞检测技术进行了细致的分类，同时深入讨论了不同拥塞检测技术方案设计的优缺点，探讨了不同方案的本质特点，对工业部署广泛、学术影响深远的技术方案和方法进行了系统解读。

本白皮书为探索更多拥塞检测技术应用，实现标准化、规范化、模块化的高性能网络拥塞检测技术落地，提供一些理论和实践参考。

目前，拥塞检测技术仍存在如下的挑战：

1、检测效果与资源占用不可兼得的矛盾。提高检测的精度和频率通常需要增加系统的复杂度和提高资源的占用，如何权衡方案的损失收益，目前没有明确的评价规范。例如，INT 带来的拥塞控制效果提升，与其 overhead 占用的资源，怎样评估利弊。这一规范需要进一步系统的研究和审慎的论证。

2、检测环节与处理、控制环节的耦合。当前的拥塞管理与控制方案中，检测、处理、控制三个环节往往通过一个整体来设计，这也导致离开整体系统的特定硬件，检测环节难以独立工作。这需要当前 RoCE 网络的标准化工作更加系统的对拥塞检测技术进行分解，实现拥塞检测技术的标准化、规范化、模块化。

3、检测技术多层次应用。当前的拥塞检测技术普遍的应用于拥塞控制协议中，构建闭环的控制系统。闭环控制系统能够有效的抵御外部扰动，这对于高性能网络在复杂的云场景应用有重要意义。而更上层的负载均衡、流量调度技术中，目前较少应用反馈的机制，这也导致很多新技术在复杂场景不具应用前景。

总的来说，高性能网络拥塞管理与控制的技术体系目前越发清晰。作为其中的基石技术，拥塞检测技术在各种技术方案中具有核心影响。而未来，拥塞检测技术是整个拥塞管理控制技术体系的基石，其关键的系统化、标准化、模块化工作，将成为高性能网络进一步规模应用的重要技术支撑。

# 参考文献

[1] W. Bai *et al.*, “Empowering Azure Storage with RDMA”.

[2] Y. Gao *et al.*, “When Cloud Storage Meets RDMA,” in *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*, 2021, pp. 519–533.

[3] R. Miao *et al.*, “From luna to solar: the evolutions of the compute-to-storage networks in Alibaba cloud,” in *Proceedings of the ACM SIGCOMM 2022 Conference*, Amsterdam Netherlands: ACM, Aug. 2022, pp. 753–766. doi: 10.1145/3544216.3544238.

[4] P. X. Gao *et al.*, “Network Requirements for Resource Disaggregation,” in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, Savannah, GA: USENIX Association, Nov. 2016, pp. 249–264. [Online]. Available: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/gao

[5] P. Zuo, J. Sun, L. Yang, S. Zhang, and Y. Hua, “One-sided RDMA-Conscious Extendible Hashing for Disaggregated Memory”.

[6] A. Dragojević, D. Narayanan, O. Hodson, and M. Castro, “FaRM: Fast Remote Memory,” in *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, in NSDI’14. USA: USENIX Association, 2014, pp. 401–414.

[7] X. Wei, R. Chen, and H. Chen, “Fast RDMA-based Ordered Key-Value Store using Remote Learned Cache”.

[8] X. Jin *et al.*, “NetCache: Balancing Key-Value Stores with Fast In-Network Caching,” in *Proceedings of the 26th Symposium on Operating Systems Principles*, Shanghai China: ACM, Oct. 2017, pp. 121–136. doi: 10.1145/3132747.3132764.

[9] “5_最佳实践-使用 SMC 和 ERI 透明加速 Redis 应用 - OpenAnolis 代码库.” https://openanolis.cn/sig/high-perf-network/doc/735934915657042794 (accessed Aug. 24, 2023).

[10] L. Shalev, H. Ayoub, N. Bshara, and E. Sabbag, “A Cloud-Optimized Transport Protocol for Elastic and Scalable HPC,” *IEEE Micro*, vol. 40, no. 6, Art. no. 6, Nov. 2020, doi: 10.1109/MM.2020.3016891.

[11] S. Hu *et al.*, “Tagger: Practical PFC Deadlock Prevention in Data Center Networks,” in *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, Incheon Republic of Korea: ACM, Nov. 2017, pp. 451–463. doi: 10.1145/3143361.3143382.

[12] Y. Zhu *et al.*, “Congestion Control for Large-Scale RDMA Deployments,” in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, London United Kingdom: ACM, Aug. 2015, pp. 523–536. doi: 10.1145/2785956.2787484.

[13] “802.1Qau – Congestion Notification |.” https://1.ieee802.org/dcb/802-1qau/ (accessed Apr. 10, 2023).

[14] M. Alizadeh *et al.*, “Data Center TCP (DCTCP),” in *Proceedings of the ACM SIGCOMM 2010 Conference*, in SIGCOMM ’10. New York, NY, USA: Association for Computing Machinery, 2010, pp. 63–74. doi: 10.1145/1851182.1851192.

[15] R. Mittal *et al.*, “TIMELY: RTT-based Congestion Control for the Datacenter,” in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, London United Kingdom: ACM, Aug. 2015, pp. 537–550. doi: 10.1145/2785956.2787510.

[16] Y. Li *et al.*, “HPCC: high precision congestion control,” in *Proceedings of the ACM Special Interest Group on Data Communication*, Beijing China: ACM, Aug. 2019, pp. 44–58. doi: 10.1145/3341302.3342085.

[17] S. Yan, X. Wang, X. Zheng, Y. Xia, D. Liu, and W. Deng, “ACC: automatic ECN tuning for high-speed datacenter networks,” in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, Virtual Event USA: ACM, Aug. 2021, pp. 384–397. doi: 10.1145/3452296.3472927.

[18] Y. Gao, Y. Yang, T. Chen, J. Zheng, B. Mao, and G. Chen, “DCQCN+: Taming Large-Scale Incast Congestion in RDMA over Ethernet Networks,” in *2018 IEEE 26th International Conference on Network Protocols (ICNP)*, Cambridge:

IEEE, Sep. 2018, pp. 110–120. doi: 10.1109/ICNP.2018.00021.

[19] R. Mittal *et al.*, “Revisiting network support for RDMA,” in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, Budapest Hungary: ACM, Aug. 2018, pp. 313–326. doi: 10.1145/3230543.3230557.

[20] Y. Zhang, Y. Liu, Q. Meng, and F. Ren, “Congestion detection in lossless networks,” in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, Virtual Event USA: ACM, Aug. 2021, pp. 370–383. doi: 10.1145/3452296.3472899.

[21] Y. Zhu, M. Ghobadi, V. Misra, and J. Padhye, “ECN or Delay: Lessons Learnt from Analysis of DCQCN and TIMELY,” in *Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies*, Irvine California USA: ACM, Dec. 2016, pp. 313–327. doi: 10.1145/2999572.2999593.

[22] G. Kumar *et al.*, “Swift: Delay is Simple and Effective for Congestion Control in the Datacenter,” in *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, Virtual Event USA: ACM, Jul. 2020, pp. 514–528. doi: 10.1145/3387514.3406591.

[23] G. Zeng, W. Bai, G. Chen, K. Chen, D. Han, and Y. Zhu, “Combining ECN and RTT for Datacenter Transport,” in *Proceedings of the First Asia-Pacific Workshop on Networking*, Hong Kong China: ACM, Aug. 2017, pp. 36–42. doi: 10.1145/3106989.3107002.

[24] R. Ben Basat, S. Ramanathan, Y. Li, G. Antichi, M. Yu, and M. Mitzenmacher, “PINT: Probabilistic In-band Network Telemetry,” in *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, Virtual Event USA: ACM, Jul. 2020, pp. 662–680. doi: 10.1145/3387514.3405894.

[25] P. Taheri, D. Menikkumbura, E. Vanini, S. Fahmy, P. Eugster, and T. Edsall, “RoCC: robust congestion control for RDMA,” in *Proceedings of the 16th*

*International Conference on emerging Networking EXperiments and Technologies*, Barcelona Spain: ACM, Nov. 2020, pp. 17–30. doi: 10.1145/3386367.3431316.

[26] X. Zhong, J. Zhang, Y. Zhang, Z. Guan, and Z. Wan, "PACC: Proactive and Accurate Congestion Feedback for RDMA Congestion Control," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, London, United Kingdom: IEEE, May 2022, pp. 2228–2237. doi: 10.1109/INFOCOM48880.2022.9796803.

[27] J. Zhang *et al.*, "Receiver-Driven RDMA Congestion Control by Differentiating Congestion Types in Datacenter Networks," in *2021 IEEE 29th International Conference on Network Protocols (ICNP)*, Dallas, TX, USA: IEEE, Nov. 2021, pp. 1–12. doi: 10.1109/ICNP52444.2021.9651938.

[28] "802.1Qbb – Priority-based Flow Control |." https://1.ieee802.org/dcb/802-1qbb/ (accessed Sep. 03, 2022).

[29] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: Dynamic Flow Scheduling for Data Center Networks".

[30] Z. Liu *et al.*, "BurstBalancer: Do Less, Better Balance for Large-scale Data Center Traffic".

[31] M. Alizadeh *et al.*, "CONGA: distributed congestion-aware load balancing for datacenters," in *Proceedings of the 2014 ACM conference on SIGCOMM*, Chicago Illinois USA: ACM, Aug. 2014, pp. 503–514. doi: 10.1145/2619239.2626316.

[32] M. Alizadeh *et al.*, "pFabric: minimal near-optimal datacenter transport," in *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, Hong Kong China: ACM, Aug. 2013, pp. 435–446. doi: 10.1145/2486001.2486031.

[33] C.-Y. Hong, M. Caesar, and P. B. Godfrey, "Finishing flows quickly with preemptive scheduling," in *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*, Helsinki Finland: ACM, Aug. 2012, pp. 127–138. doi: 10.1145/2342356.2342389.

[34] W. Bai, L. Chen, K. Chen, D. Han, C. Tian, and H. Wang, “Information-Agnostic Flow Scheduling for Commodity Data Centers”.

[35] J. Perry, A. Ousterhout, H. Balakrishnan, D. Shah, and H. Fugal, “Fastpass: a centralized ‘zero-queue’ datacenter network,” in *Proceedings of the 2014 ACM conference on SIGCOMM*, Chicago Illinois USA: ACM, Aug. 2014, pp. 307–318. doi: 10.1145/2619239.2626309.

[36] B. Montazeri, Y. Li, M. Alizadeh, and J. Ousterhout, “Homa: a receiver-driven low-latency transport protocol using network priorities,” in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, Budapest Hungary: ACM, Aug. 2018, pp. 221–235. doi: 10.1145/3230543.3230564.

[37] L. Chen, J. Lingys, K. Chen, and F. Liu, “AuTO: scaling deep reinforcement learning for datacenter-scale automatic traffic optimization,” in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, Budapest Hungary: ACM, Aug. 2018, pp. 191–205. doi: 10.1145/3230543.3230551.

[38] 杜鑫乐, “数据中心网络的流量控制：发展现状与趋势,” 计算机学报, 2020.

[39] B. Vamanan, J. Hasan, and T. N. Vijaykumar, “Deadline-aware datacenter tcp (D2TCP),” in *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*, Helsinki Finland: ACM, Aug. 2012, pp. 115–126. doi: 10.1145/2342356.2342388.

[40] C. Wilson, H. Ballani, T. Karagiannis, and A. Rowtron, “Better never than late: meeting deadlines in datacenter networks,” in *Proceedings of the ACM SIGCOMM 2011 conference*, Toronto Ontario Canada: ACM, Aug. 2011, pp. 50–61. doi: 10.1145/2018436.2018443.

[41] L. S. Brakmo, S. W. O’Malley, and L. L. Peterson, “TCP Vegas: new techniques for congestion detection and avoidance,” in *Proceedings of the conference on Communications architectures, protocols and applications - SIGCOMM ’94*, London, United Kingdom: ACM Press, 1994, pp. 24–35. doi:

10.1145/190314.190317.

[42] D. Scholz, B. Jaeger, L. Schwaighofer, D. Raumer, F. Geyer, and G. Carle, “Towards a Deeper Understanding of TCP BBR Congestion Control,” in *2018 IFIP Networking Conference (IFIP Networking) and Workshops*, Zurich, Switzerland: IEEE, May 2018, pp. 1–9. doi: 10.23919/IFIPNetworking.2018.8696830.

[43] J. H. Salim and U. Ahmed, “Performance Evaluation of Explicit Congestion Notification (ECN) in IP Networks,” Internet Engineering Task Force, Request for Comments RFC 2884, Jul. 2000. doi: 10.17487/RFC2884.

[44] K. K. Ramakrishnan and S. Floyd, “A Proposal to add Explicit Congestion Notification (ECN) to IP,” Internet Engineering Task Force, Request for Comments RFC 2481, Jan. 1999. doi: 10.17487/RFC2481.

[45] S. Floyd and V. Jacobson, “Random early detection gateways for congestion avoidance,” *IEEE/ACM Trans. Networking*, vol. 1, no. 4, pp. 397–413, Aug. 1993, doi: 10.1109/90.251892.

[46] W. Fengy, D. D. Kandlurz, D. Sahaz, and K. G. Shiny, “BLUE: A New Class of Active Queue Management Algorithms”.

[47] B. Nandy, N. Seddigh, and J. H. Salim, “A proposal for Backward ECN for the Internet Protocol (IPv4/IPv6),” Internet Engineering Task Force, Internet Draft draft-salim-jhsbnns-ecn-00, Jun. 1998. Accessed: Aug. 25, 2023. [Online]. Available: https://datatracker.ietf.org/doc/draft-salim-jhsbnns-ecn-00

[48] R. Jain, “Enhanced Forward Explicit Congestion Notification (E-FECN) Scheme for Datacenter Ethernet Networks”.

[49] P. Eardley, “Pre-Congestion Notification (PCN) Architecture,” Art. no. rfc5559, Jun. 2009, doi: 10.17487/RFC5559.

[50] J. Zhang, W. Bai, and K. Chen, “Enabling ECN for datacenter networks with RTT variations,” in *Proceedings of the 15th International Conference on Emerging Networking Experiments And Technologies*, Orlando Florida: ACM, Dec. 2019, pp. 233–245. doi: 10.1145/3359989.3365426.

[51] B. Briscoe, R. Woundy, and A. Cooper, “Congestion Exposure (ConEx)

Concepts and Use Cases," Art. no. rfc6789, Dec. 2012, doi: 10.17487/RFC6789.

[52] S. Abbasi, S. Ketabi, A. Munir, M. Bahnasy, and Y. Ganjali, "DWTCP: Ultra Low Latency Congestion Control Protocol for Data Centers," no. arXiv:2207.05624. arXiv, Jul. 12, 2022. Accessed: Jul. 14, 2022. [Online]. Available: http://arxiv.org/abs/2207.05624

ODCC公众号

ODCC订阅号