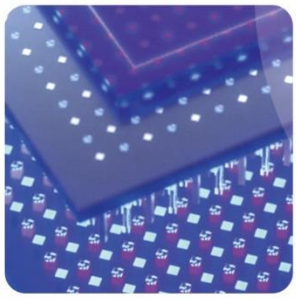
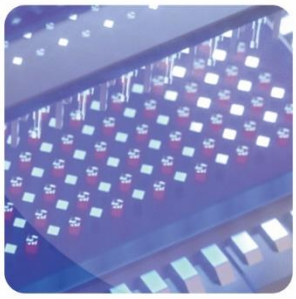
HUAWEI

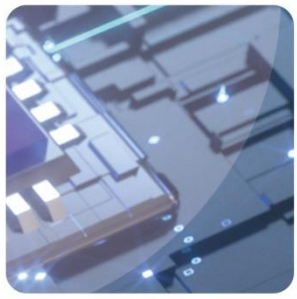
智能无损网络技术白皮书

中国科学院
高能物理
研究所

江苏省
未来网络
创新研究院







紫金山
实验室

北京
邮电大学

构建万物互联的智能世界

Building a Fully Connected, Intelligent World

目 录

1 概述

计算、存储、网络是数据中心的三要素，三个要素互相促进、共同发展。GPU/AI 芯片异构计算蓬勃发展，近五年性能提升 600 倍。固态硬盘相比机械硬盘访问性能提升了 100 倍，NVMe 又相比固态硬盘提升了 100 倍；计算和存储的发展，都对网络提出了新的要求：更大的带宽、更低的时延，智能无损网络就是解决方案。

2 背景

2.1 AI 异构计算，处理性能爆炸增长

AI 时代已经来临，随着人工智能应用数据的急速增加，AI 算法规模不断增加，AI 模型的训练任务提出海量规模的计算量需求。根据 OpenAI[1]的分析数据，自 2012 年起，最新的模型训练所需的计算量每 3.4 个月翻一倍。模型的训练时间也随之延长，严重影响 AI 模型部署的迭代周期。为了满足 AI 业务市场需求，缩短模型训练时长，对 AI 算力提出越来越高的要求。

为了满足 AI 时代模型训练的算力需求，AI 加速硬件 GPU 的算力持续着近乎摩尔定律的极限增长速度。然而，GPU 算力的增长速度仍然远低于 AI 应用对计算量的需求的增速。因此，分布式 AI 集群系统成为了满足 AI 应用算力需求的首选。

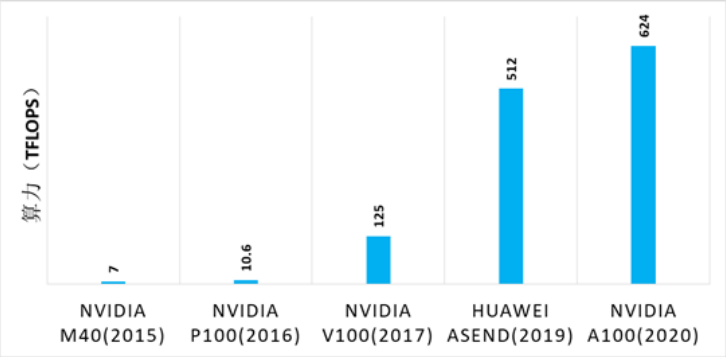


图 1. GPU 算力演进

随着 AI 训练集群规模的增大，以及单节点算力的增长，分布式 AI 集群系统已经逐渐从计算约束转换为网络通信约束。一方面，在过去 5 年，GPU 算力增长了近 90 倍，而网络带宽仅增长了 10 倍。另一方面，当前的 AI 集群系统中，当 GPU 集群达到一定规模以后，随着计算节点数的增加，由于分布式 AI 集群节点之间的通信代价的增加，可能导致集群每秒训练的图片数量不增反减。

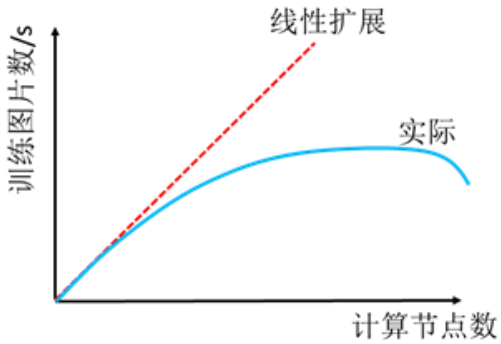


图 2. GPU 集群图片训练速度曲线

为了充分发挥 AI 集群的 GPU 算力，加速 AI 模型的训练，分布式 AI 集群系统对网络提出了更高的需求：

- **分布式 AI 集群需要高性能的 RDMA 网络。**

RDMA 网络的内核旁路以及内存零拷贝特性，大大缩短了协议栈的通信处理时延以及数据搬移时延。同时，RDMA 网络相比 TCP 能提供更高的单流通信带宽。具体而言，100GE 网络环境下，RDMA 单流能跑满带宽，而 TCP 单流仅能跑到 17Gbps 的带宽。因此，RDMA 网络能为 GPU 集群系统提供更好的通信性能。

Uber 在其发布的 HOROVOD 平台测试了 RDMA 与 TCP 两种 25GE 网络环境下，VGG16 模型的训练性能。结果显示，在 512 块 GPU 规模时，RDMA 性能比 TCP 性能高 50%。

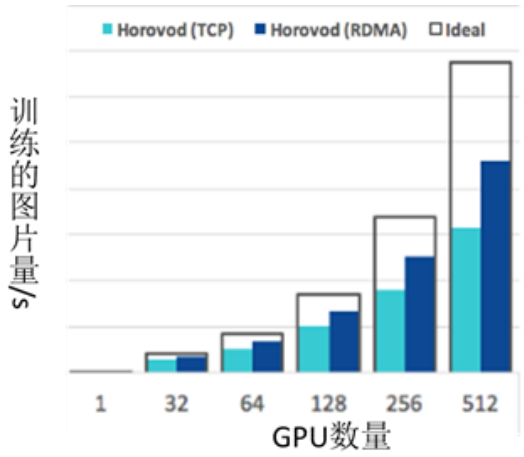


图 3：25GE 网络，VGG16 模型训练，RDMA 比 TCP 高约 50%的性能

--数据来源：https://eng.uber.com/horovod/

为了进一步测试更高带宽下，GPU 集群在 RDMA 与 TCP 网络的性能差异，我们在 100GE 环境下，基于 Tensorflow Horovod 测试了不同节点规模的集群性能。结果显示，在 8 节点的 VGG16 模型训练，RDMA 性能是 TCP 的 8+倍。

综合两个测试结果可知，当网络带宽更高时，RDMA 较 TCP 的性能优势将更明显。因此，在 25GE 以上的网络，应优选 RDMA 网络。

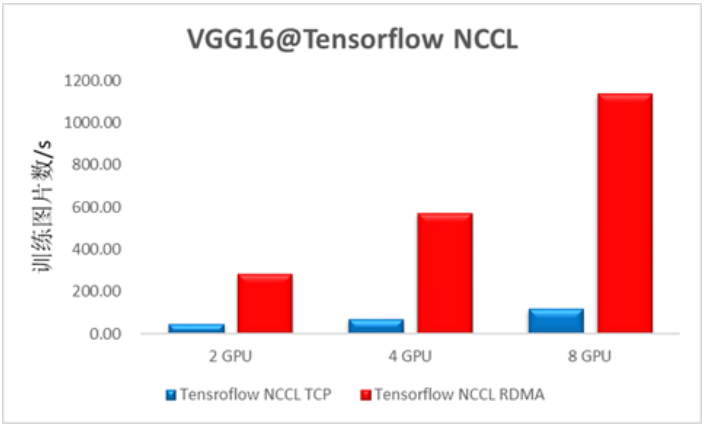


图 4：100GE，Tensorflow NCCL 架构，DNN 模型训练，RDMA 提供 TCP 的 8+倍性能

--数据来源：实验室自测数据

- **分布式 AI 集群需要 100GE 以上的高带宽。**

分布式 AI 集群承载的 AI 模型训练业务，是一个计算-通信的迭代过程。模型训练过程中，各节点基于训练样本计算出新梯度后，需要将各节点的梯度进行同步更新，下一轮迭代计算开始，依赖新的模型参数。因此，计算节点之间的同步时间，直接影响 GPU 集群的效率。高带宽能有效降低参数同步时间，提升集群训练效率。

研究机构[2]测试了 VGG16 模型在 64GPU 集群不同网络带宽下的扩展效率，结果表明，100GE 相比 25GE 能提高接近 2 倍的性能。因此，为了充分发挥 GPU 的算力，GPU 集群建议配置 100GE 以上的带宽。

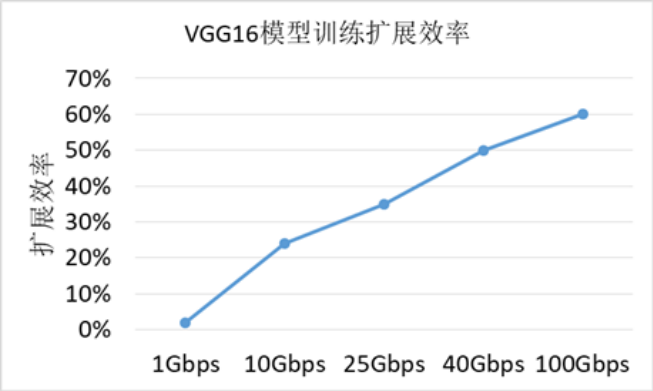


图 5 不同网络带宽下，GPU 集群的扩展效率

--数据来源：Is Network the Botleneck of Distributed Training, Johns Hopkins University & AWS

2.2 存储介质的进步，需要更高性能的网络

当前存储介质 SSD（solid-state drive，固态硬盘）的访问性能相比传统分布式存储 HDD（hard-disk drive，机械硬盘）已提升了 100 倍，对于采用 NVMe（Non-Volatile Memory express）接口协议的 SSD（简称 NVM 介质）时，访问性能相比 HDD 甚至可以提升 10000 倍。在存储介质的时延已大幅降低的情况下，网络的时延占比已从原来的小于 5%变化到 65%左右，也就是说，宝贵的存储介质有一半以上的时间是空闲通信等待。如何降低网络时延成为提升 IOPS（Input/OutputOperations Per Second，每秒读写次数）的核心。

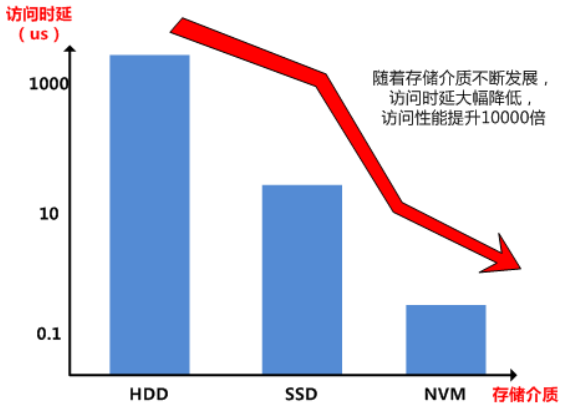


图 6 存储介质访问时延趋势图

3 智能无损网络架构

智能无损网络从 2 个方面来提升应用性能，一方面是网络自身的优化，另一方面是网络与应用系统的融合优化。

网络自身优化的目标是使整网吞吐最高、时延最低，其包括三个层次：

1.流的控制（Flow control）：用于解决发送端与接收端的速率匹配问题，做到无丢包。

2.拥塞控制（Congestion control）：用于解决网络拥塞时对流量的速率控制问题，做到满吞吐与低时延。

3.流量调度（Traffic scheduling）：用于解决业务流量与网络链路的负载均衡性问题，做到不同业务流量的服务质量保障。

网络与应用系统的融合优化，其关键点即在于发挥网络设备负责连通性的天然物理位置优势，与计算、存储系统进行一定层次的配合，以提高应用系统的性能。目前智能无损网络正在开展的融合优化包括如下两方面：

1.针对 HPC 场景 MPI 通信的网算一体特性（INC：Intergrated Network and Forwarding），网络设备参与计算过程，减少任务完成时间。

2.针对 NOF 存储系统的智能无损存储网络特性（NOF+），网络与存储系统，共同完成存储系统的访问控制、故障倒换、长距双活等功能。

3.1 软件架构

智能无损存储网络从软件架构层面，包括网络设备软件、控制器软件、分析器软件三方面。

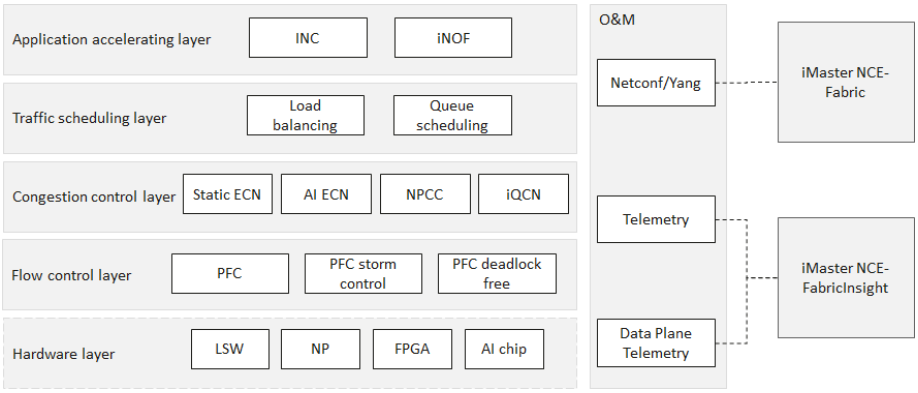


图 7：智能无损网络软件架构

系统分类	特性	描述
流控技术	PFC	优先级流控：由 IEEE 802.1Qbb 定义的一个协议，主要解决拥塞导致的丢帧问题。
	PFC storm control	PFC 风暴控制：主要解决由 PFC 风暴引起网络断流的问题，也称为 PFC 死锁检测。
	PFC deadlock free	PFC 死锁预防：通过识别环形缓存依赖并破除其产生的必要条件，从而解决 PFC 死锁的问题，提高网络可靠性。
拥塞控制	Static ECN	全称为 Explicit Congestion Notification，即显式拥塞通知，是一种端到端的网络拥塞通知机制，它允许网络在发生拥塞时不丢弃报文，在 RFC 3168 (2001)中定义。
	AI ECN	通过 iLossless 智能无损算法动态调节 ECN 门限，以获得最大带宽与最小时延。
	NPCC	全称为 Network Proactive Congestion Control，在网络设备上基于本设备的拥塞状态主动控制每条流的发送速率，主要用于长距无损网络。
	iQCN	全称为 Intelligent Quantized Congestion Notification，在 TCP 与 RoCE 混跑场景用于控制 RoCE 流量的时延。
流量调度	负载分担	负载分担是一种在存在多条路径时选择转发路径的技术，其目的是达到更平衡的网络负载，典型的应用场景是链路捆绑与 ECMP。
	队列调度	队列调度用于控制不同队列之间流量的发送策略，以为不同队

系统分类	特性	描述
		列的流量提供差异化的质量保障
应用加速	INC	网算一体功能，全称为 Intergrated Network and Computing，在网络设备上通过对 MPI 通信数据的聚合以提高 HPC 应用性能。
	NOF+	智能无损存储网络特性，主要用于基于 RoCE 的存储网络，以提升存储系统的易用性、可靠性，以及支持存储系统的同城双活场景。
O&M	Netconf/Yang	由 IETF 定义的一组网络管理协议，主要用于控制器或网管系统对网络设备进行管理，是当前主流的设备北向接口。
	Telemetry	网络设备的测量系统，广义上的 Telemetry 为所有支撑网络设备数据采集的技术，狭义上的 Telemetry 通常特指基于 gRPC 框架按照 protobuf 编码的网络数据采集技术，此技术广泛应用于网络分析器中。
	Syslog	由 IETF 定义的一个协议，用于传输网络设备的日志信息。
	Data Plane Telemetry	数据平面 Telemetry 功能，是在网络设备的数据转发平面进行测量的一种技术，通常用于测量与数据流相关的指标，例如某条流的速率、时延等信息。
控制器	iMaster NCE-Fabric	iMaster NCE 自动驾驶网络管理与控制系统实现了物理网络与商业意图的有效连接，向下实现全局网络的集中管理、控制和分析，面向商业和业务意图使能资源云化、全生命周期自动化，以及数据分析驱动的智能闭环；向上提供开放网络 API 与 IT 快速集成
分析器	iMaster NCE-FabricInsight	FabricInsight 系统利用 CE 系列交换机设备的 Telemetry 特性采集设备、接口、队列等性能 Metrics 数据进行分析，主动监控、预测网络异常。

3.2 硬件架构

设备类型	设备款型	设备主要规格描述
框式交换机	CloudEngine 16800	包括 CloudEngine 16804、CloudEngine 16808、CloudEngine 16816 三种款型，支持 36 x 100GE; 18 x 100GE; 36 x 40GE; 24 x 40GE; 48 x 10GE 类型线卡。
	CloudEngine 12800	包括 CE12804/CE12808/CE12812/CE12816，支持高密度的 36*100GE/36*40GE/144*25GE/144*10GE 线速线卡;
盒式交换机	CloudEngine 6866-48S8CQ-P	CloudEngine 6866-48S8CQ-P 系列交换机提供 48 个 25GE SFP28 接口或 48 个 50GE SFP56 接口，8 个 100GE QSFP28 接口或 8 个 200GE QSFP56 接口。
	CloudEngine 8851-32CQ8DQ-P	CloudEngine 8851-32CQ8DQ-P 交换机提供 32 个 100GEQSFP28 接口或 32 个 200GEQSFP56 接口， 8 个 400GE QSFPDD 接口。
	CloudEngine 6865-48S8CQ-EI	CloudEngine 6865-48S8CQ-EI 系列交换机提供 48 个 25GE SFP28 接口，8 个 100GEQSFP28 接口 。
	CloudEngine 8850-64CQ-EI	CloudEngine 8850-64CQ-EI 系列交换机提供 64 个 100GE QSFP28 接口。
	CloudEngine 8861-4C-EI	CloudEngine 8861-4C-EI 系列交换机提供 2U 高度，支持四个半宽灵活插卡，支持以下系列子卡: 1) CE88-D24S2CQ 系列插卡提供 24 个 10GE/25GE SFP28 和 2 个 100GE QSFP28 接口。 2) CE88-D24T2CQ 系列插卡提供 24 个 10GE (BASE-T) 和 2 个 100GE (QSFP28) 接口。 3) CE88-D8CQ 系列插卡提供 8 个 100GE (QSFP28) 接口。 4) CE88-D16Q 系列插卡提供 16 个 40GE (QSFP+) 接口。 5) CE88-D24S2CQ-U 系列插卡提供 24 端口 25GE 以太网/16G FC 光接口（SFP28）和 2 端口 40GE/100GE 以太网光接口 （QSFP28）接口。
	CE9860-4C-EI	CE9860-4C-EI 为 4U 高度，支持四个全宽灵活插卡，插卡形态为 CE98-D32CQ ，提供 32 端口 100GE/40GE QSFP28 接

设备类型	设备款型	设备主要规格描述
		口卡。

4 网络流量控制

4.1 流控技术

流控技术是保障网络零丢包的基础技术。在数据通信中，流量控制提供了一种机制，此机制作用于接收方，由接收方来控制数据传输的速率，以防止快速的发送方压倒慢速的接收方。本章节介绍流控相关的技术，包括如何对流量进行优先级映射，PFC 优先级流控的原理，如何解决 PFC 的死锁问题。

4.1.1 流量映射

优先级映射用来实现报文携带的外部优先级与设备的内部优先级之间的转换，并利用 DiffServ 域来管理和记录外部优先级和内部优先级之间的映射关系，从而设备可以根据内部优先级提供有差别的 QoS 服务质量。

对于不同的优先级概念介绍如下：

- **外部优先级**

又称为报文优先级、QoS 优先级，即使用报文中某些特定字段比如 VLAN 报文的 802.1p 值、IP 报文的 DSCP 值等记录 QoS 信息。需要注意的是，设备只可以根据设备内部优先级处理收到的报文，为不同的业务提供不同的 QoS 服务，所以上述外部优先级在进入设备后会映射为设备内部的优先级。

- **内部优先级**

又称为服务等级（Class of Service）、PHB 行为（Per Hop Behavior）、本地优先级，支持 8 种取值，优先级从高到低依次为 CS7、CS6、EF、AF4、AF3、AF2、AF1、BE，与出端口的 8 个队列从高到低依次对应。因此内部优先级决定了报文在设备内部所属的队列。当针对某一个队列配置 QoS 业务时，即对所有通过该队列转发的报文都设置了相同的 QoS 服务。

用户可以根据网络规划在不同网络中使用不同的外部优先级字段，例如在 VLAN 网络中使用 802.1p，IP 网络中使用 DSCP。当报文在网络中传输时：

- 所有进入设备的报文，其外部优先级字段（包括 802.1p、DSCP）都被映射为内部优先级，然后根据内部优先级与队列之间的映射关系确定报文进入的队列，从而针对队列进行流量整形、拥塞避免、队列调度等处理。
- 设备发出报文时，则将内部优先级映射为某种外部优先级字段，以便其他设备根据报文的外部优先级提供相应的 QoS 服务。
- 在配置 PFC 功能之前先结合 QoS 功能进行优先级规划，首先需要确定采用 802.1p 值还是 DSCP 值映射的优先级来承载 RoCEv2 流量。其次需要对队列进行规划，一般情况下，推荐的队列规划如下图所示。（对于本文推荐的款型，优先级与队列为一一对应关系）
- 队列 7 和 6 一般用于承载各种协议报文，为确保 CNP 报文的及时传递，CNP 报文可以单独用队列 6 进行承载。

- 需要至少一个队列用来单独承载 RoCEv2 业务，映射到该队列的优先级需要使能 PFC 功能。推荐使用队列 5、4 和 3 来承载 RoCEv2 业务流量。
- 若为 TCP/RoCE 流量混跑场景，可以使用队列 2 和 1 来承载 TCP 流量。
- 推荐队列 7 和 6 采用 SP 调度（命令行为 qos pq），其他队列采用 WDRR 调度（命令行为 qos drr），当采用 WDRR 调度时，推荐权重从大到小为：承载 RoCEv2 的队列>承载 TCP 的队列>0 队列。

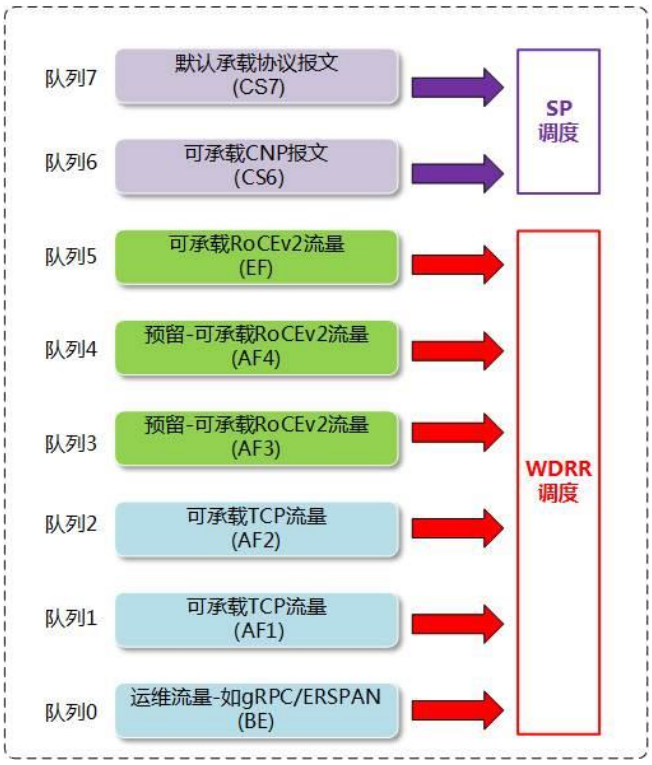


图 8 推荐的队列规划

4.1.2 Pause 帧与 PFC

通过以太 PAUSE 帧实现的流控（IEEE 802.3 Annex 31B）是以太网的一项基本功能。当下游设备发现接收能力小于上游设备的发送能力时，会主动发 PAUSE 帧给上游设备，要求暂停流量的发送，等待一定时间后再继续发送数据。

如下图所示，端口 A 和 B 接收报文，端口 C 向外转发报文。如果端口 A 和 B 的收包速率之和大于端口 C 的带宽，那么部分报文就会缓存在设备内部的报文 buffer 中。当 buffer 的占用率达到一定程度时，端口 A 和 B 就会向外发送 PAUSE 帧，通知对端暂停发送一段时间。PAUSE 帧只能阻止对端发送普通的数据帧，不能阻止发送 MAC 控制帧。

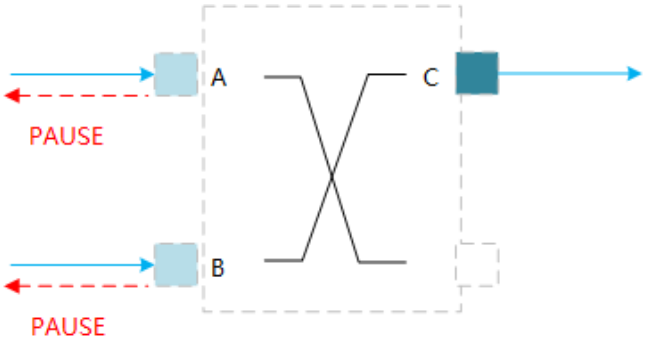


图 9：以太 PAUSE 帧应用示意图

以上的描述有个先决条件，那就是端口 A 和 B 工作在全双工模式下，并且使能了流控功能，同时对端的端口也要开启流控功能。需要注意的是，有的以太网设备只能对 PAUSE 帧做出响应，但是并不能发送 PAUSE 帧。

以太PAUSE机制的基本原理不难理解，比较容易忽视的一点是——端口收到PAUSE帧之后，停止发送报文多长时间？其实，如下图所示，PAUSE 帧中携带了时间参数。

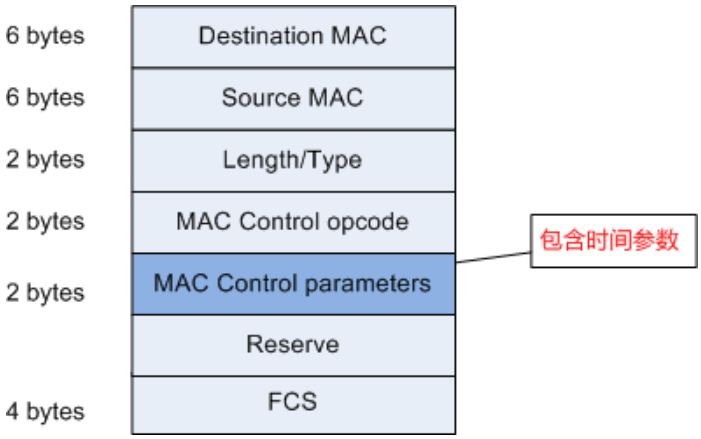


图 10：PAUSE 帧格式

- PAUSE 帧的目的 MAC 地址是保留的 MAC 地址 0180-C200-0001，源 MAC 则是发送 PAUSE 帧的设备的 MAC 地址。
- MAC Control Opcode 域的值是 0x0001。其实，PAUSE 帧是 MAC 控制帧的一种，其他类型的 MAC 控制帧使用不同的 opcode 值。因此，通过 opcode，交换机可以识别收到的 MAC 控制帧是否是 PAUSE 帧。
- MAC Control Parameters 域需要根据 MAC Control Opcode 的类型来解析。对于 PAUSE 帧而言，该域是个 2 字节的无符号数，取值范围是 0~65535。该域的时间单位是 pause_quanta，每个 pause_quanta 相当于 512 比特时间。收到 PAUSE 帧的设备通过简单的解析，就可以确定停止发送的时长。对端设备出现拥塞的情况下，本端端口通常会连续收到多个 PAUSE 帧。只要对端设备的拥塞状态没有解除，相关的端口就会一直发送 PAUSE。

基于以太 PAUSE 机制的流控虽然可以预防丢包，但是有一个不容忽视的问题。PAUSE 帧会导致一条链路上的所有报文停止发送，即在出现拥塞时会将链路上所有的流量都暂停，在服务质量要求较高的网络中，这显然是不能接受的。为了解

决这个问题，IEEE 在 802.1qbb 中引入了优先级流控功能 PFC（Priority-based Flow Control，基于优先级的流量控制）也称为 Per Priority Pause 或 CBFC（Class Based Flow Control），是对 PAUSE 机制的一种增强。

PFC 允许在一条以太网链路上创建 8 个虚拟通道，并为每条虚拟通道指定一个优先等级，允许单独暂停和重启其中任意一条虚拟通道，同时允许其它虚拟通道的流量无中断通过。这一方法使网络能够为单个虚拟链路创建零丢包类别的服务，使其能够与同一接口上的其它流量类型共存。

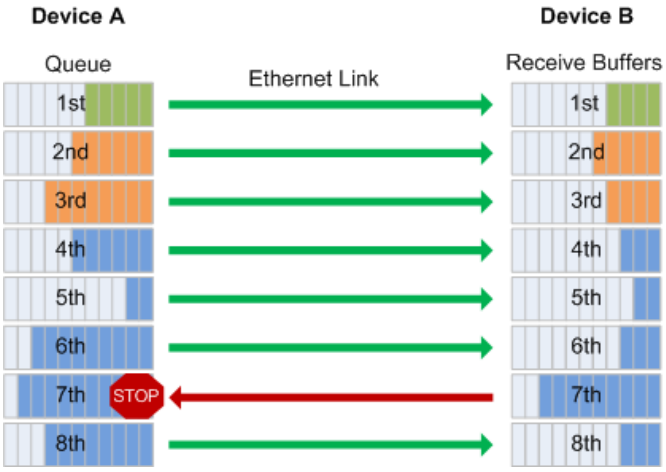


图 11：PFC 的工作机制

如上图所示，DeviceA 发送接口分成了 8 个优先级队列，DeviceB 接收接口有 8 个接收缓存（buffer），两者一一对应（报文优先级和接口队列存在着一一对应的映射关系），形成了网络中 8 个虚拟化通道，缓存大小不同使得各队列有不同的数据缓存能力。

当 DeviceB 的接口上某个接收缓存产生拥塞时，即某个设备的队列缓存消耗较快，超过一定阈值（可设定为端口队列缓存的 1/2、3/4 等比例），DeviceB 即向数据进入的方向（上游设备 DeviceA）发送反压信号“STOP”。

DeviceA 接收到反压信号，会根据反压信号指示停止发送对应优先级队列的报文，并将数据存储在本地接口缓存。如果 DeviceA 本地接口缓存消耗超过阈值，则继续向上游反压，如此一级级反压，直到网络终端设备，从而消除网络节点因拥塞造成的丢包。

“反压信号”实际上是一个以太帧，其具体报文格式如图 12 所示。

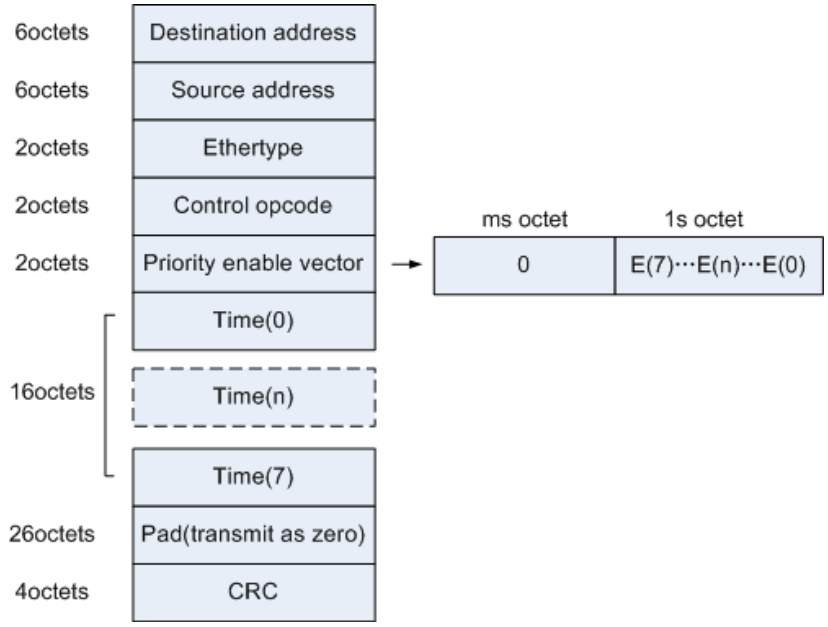


图 12：PFC 帧格式

表4-1 PFC 帧的定义

项目	描述
Destination address	目的 MAC 地址，取值固定为 01-80-c2-00-00-01。
Source address	源 MAC 地址。
Ethertype	以太网帧类型，取值为 88-08。
Control opcode	控制码，取值为 01-01。
Priority enable vector	反压使能向量。 其中 E(n)和优先级队列 n 对应，表示优先级队列 n 是否需要反压。当 E(n)=1 时，表示优先级队列 n 需要反压，反压时间为 Time(n)；当 E(n)=0 时，则表示该优先级队列不需要反压。
Time(0)～Time(7)	反压定时器。 当 Time(n)=0 时表示取消反压。
Pad	预留。 传输时为 0。
CRC	循环冗余校验。

总而言之，设备会为端口上的 8 个队列设置各自的 PFC 门限值，当队列已使用的缓存超过 PFC 门限值时，则向上游发送 PFC 反压通知报文，通知上游设备停止发包；当队列已使用的缓存降低到 PFC 门限值以下时，则向上游发送 PFC 反压停止报文，通知上游设备重新发包，从而最终实现报文的零丢包传输。

由此可见，PFC 中流量暂停只针对某一个或几个优先级队列，不针对整个接口进行中断，每个队列都能单独进行暂停或重启，而不影响其他队列上的流量，真正实现多种流量共享链路。而对非 PFC 控制的优先级队列，系统则不进行反压处理，即在发生拥塞时将直接丢弃报文。

4.1.3 PFC 死锁检测

服务器网卡故障引起其不断发送 PFC 反压帧，网络内 PFC 反压帧进一步扩散，导致出现 PFC 死锁，最终将导致整网受 PFC 控制的业务的瘫痪。

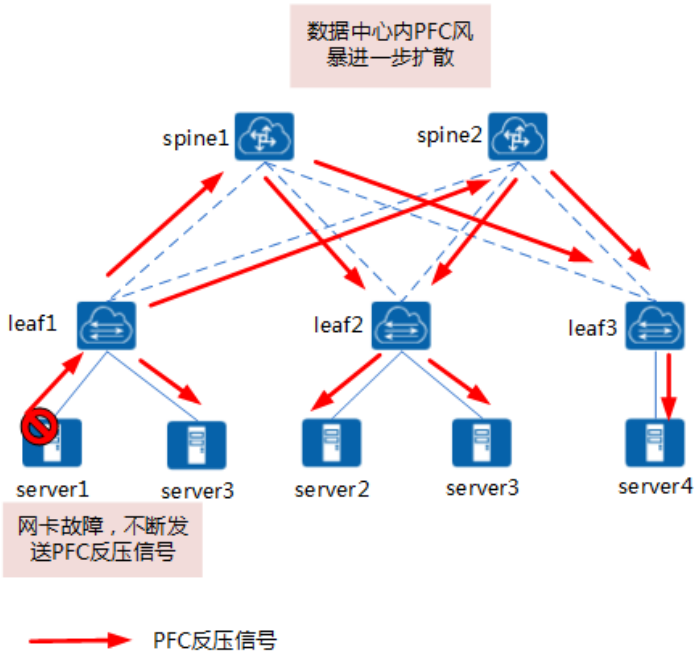


图 13：服务器网卡故障引起 PFC 风暴形成 PFC 死锁示意图

由 PFC 死锁的各个场景可知，一旦出现 PFC 死锁，若不及时解除，将威胁整网的无损业务，智能无损网络为每个设备提供了 PFC 死锁检测功能，通过以下几个过程对 PFC 死锁进行全程监控，当设备在死锁检测周期内持续收到 PFC 反压帧时，将不会响应。

1.死锁检测

Device2 的端口收到 Device1 发送的 PFC 反压帧后，内部调度器将停止发送对应优先级的队列流量，并开启定时器，根据设定的死锁检测和精度开始检测队列收到的 PFC 反压帧。

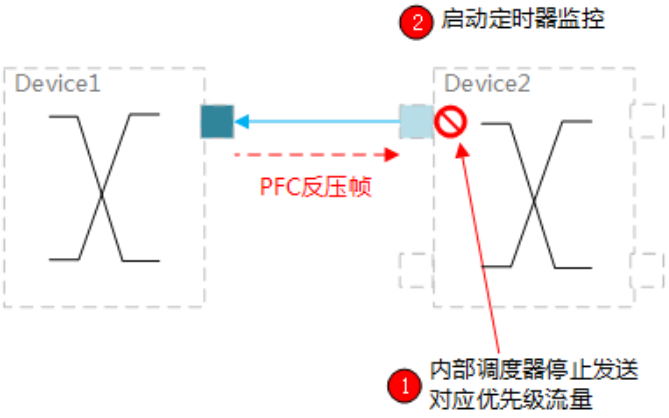


图 14：开启死锁检测

2. 死锁判定

若在设定的 PFC 死锁检测时间内该队列一直处于 PFC-XOFF（即被流控）状态，则认为出现了 PFC 死锁，需要进行 PFC 死锁恢复处理流程。

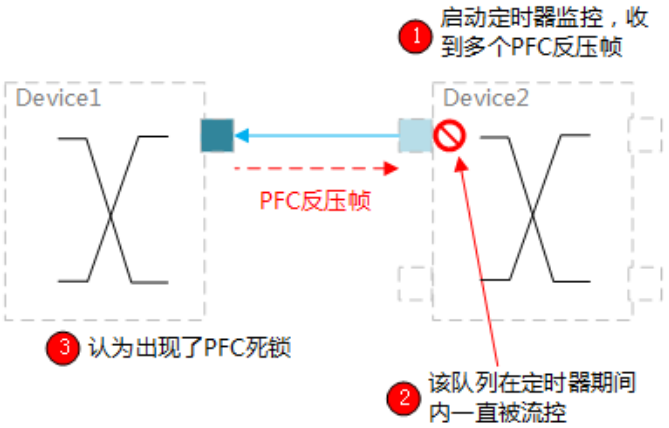


图 15：判断出现了死锁

3. 死锁恢复

在 PFC 死锁恢复过程中，会忽略端口接收到的 PFC 反压帧，内部调度器会恢复发送对应优先级的队列流量，也可以选择丢弃对应优先级的队列流量，在恢复周期后恢复 PFC 的正常流控机制。若下一次死锁检测周期内仍然判断出现了死锁，那么将进行新一轮周期的死锁恢复流程。

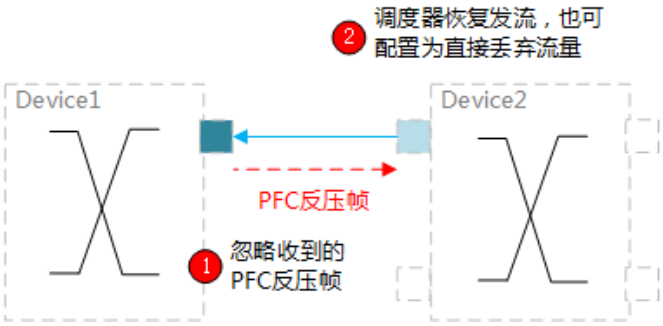


图 16：进入死锁恢复流程

4. 死锁控制

若上述死锁恢复流程没有起到作用，仍然不断出现 PFC 死锁现象，那么用户可以配置在一段时间内出现多少次死锁后，强制进入死锁控制流程。比如设定一段时间内，PFC 死锁触发了一定的次数之后，认为网络中频繁出现死锁现象，存在极大风险，此时进入死锁控制流程，设备将自动关闭 PFC 功能，需要用户手动恢复。

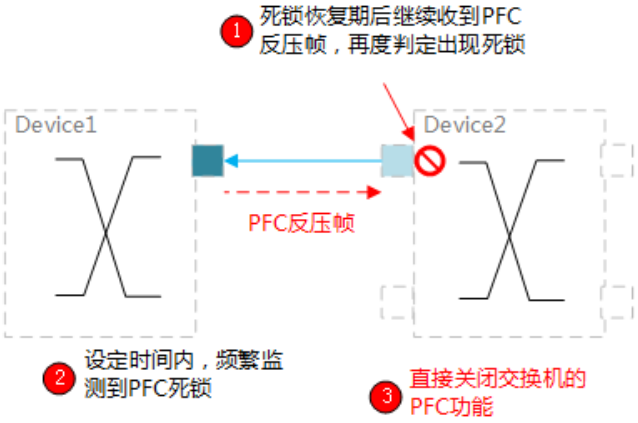


图 17：频繁出现死锁可关闭 PFC 功能

4.1.4 PFC 死锁预防

特殊情况下，例如发生链路故障或设备故障时，路由重新收敛期间可能会出现流量绕行，即对于一台交换机，从上行接口收到的流量又从另一个上行口转发出去，从而出现了环形缓存依赖。如下图所示，当 4 台交换机都达到 PFC 门限，都同时向对端发送 PFC 反压帧，这个时候该拓扑中所有交换机都处于停流状态，由于 PFC 的反压效应，整个网络或部分网络的吞吐量将变为零。即使在无环网络中形成短暂环路时，也可能发生死锁。虽然经过修复短暂环路会很快消失，但它们造成的死锁不是暂时的，即便重启服务器中断流量，死锁也不能自动恢复。

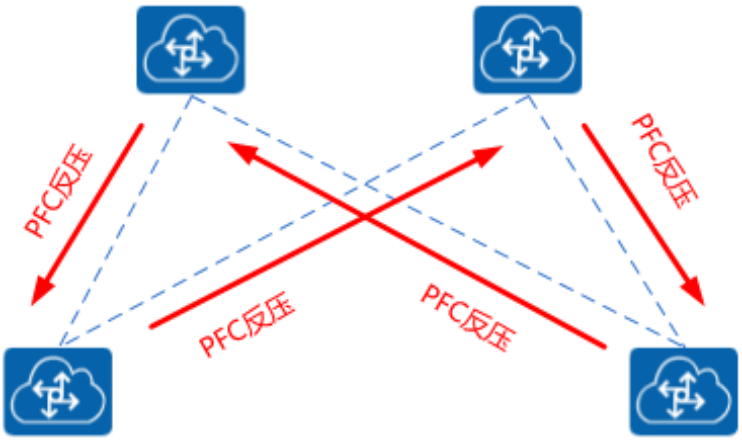


图 18：循环缓冲区依赖形成 PFC 死锁示意图

PFC 死锁预防正是针对数据中心网络典型的 CLOS 组网的一种事前预防的方案，通过识别易造成 PFC 死锁的业务流，修改队列优先级，改变 PFC 反压的路径，让 PFC 反压帧不会形成环路。

如图 19 所示，一条业务流从 Server1-Leaf1-Spine1-Leaf2-Server4，这种正常的业务转发过程不会引起 PFC 死锁。然而若 Leaf2 与 Server4 间出现链路故障，或者 Leaf2 因为某些故障原因没有学习到 Server4 的地址，都将导致流量不从 Leaf2 的下游端口转发，而是从 Leaf2 的上游端口转发。这样 Leaf2-Spine2-Leaf1-Spine1 就形成了一个循环依赖缓冲区，当 4 台交换机的缓存占用都达到 PFC 反压帧触发门限，都同时向对端发送 PFC 反压帧停止发送某个优先级的流量，将形成 PFC 死锁状态，最终导致该优先级的流量在组网中被停止转发。

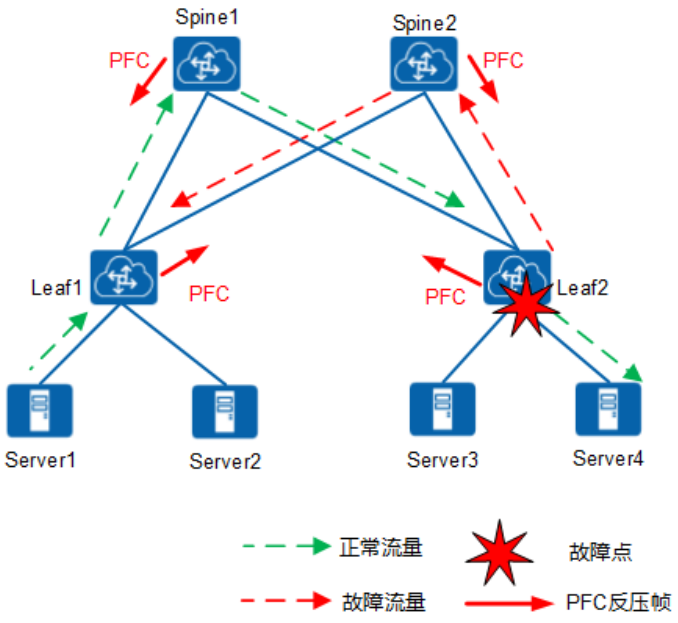


图 19：CLOS 架构下的 PFC 死锁

PFC 死锁预防功能中定义了 PFC 上联端口组，用户可以将一个 Leaf 设备上与 Spine 相连的接口，例如图 20 中的 interface1 与 interface2，都加入 PFC 上联端口组，一旦 Leaf2 设备检测到同一条业务流从属于该端口组的接口内进出，即说明该业务流是一条高风险的钩子流，易引起 PFC 死锁的现象。

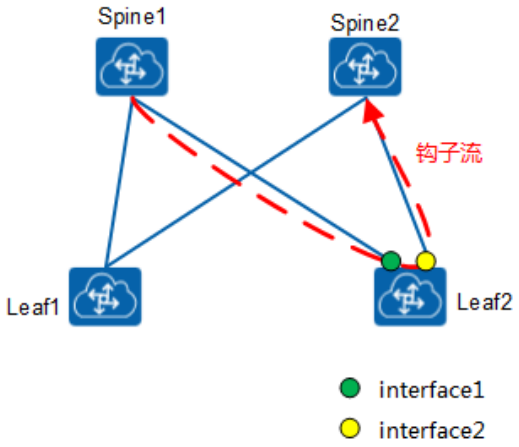


图 20：PFC 钩子流

如图 21 所示，Device2 识别到一条从 Device1 发过来的走队列 a 的流量为钩子流。此时 Device2 会修改该流的优先级并修改其 DSCP 值，使其从队列 b 转发。这样若该流在下游设备 Device3 引起了拥塞，触发了 PFC 门限，则会对向 Device2 的队列 b 进行反压，让 Device2 停止发送队列 b 对应优先级的流量，不会影响队列 a，避免了形成循环依赖缓冲区的可能，从而预防了 PFC 死锁的发生。

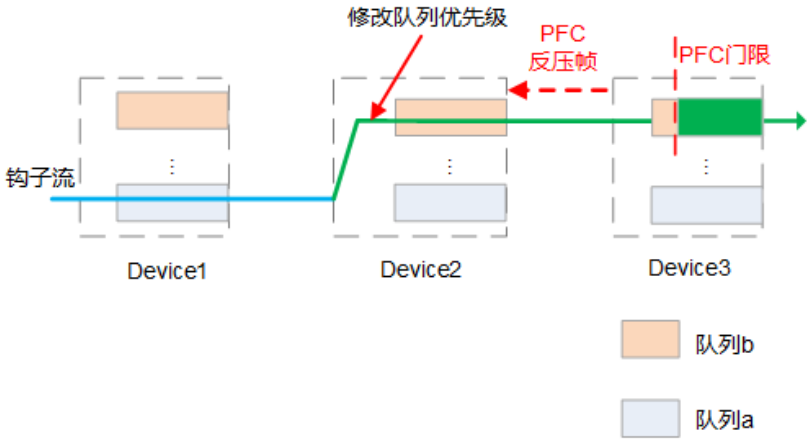


图 21：PFC 死锁预防原理

4.2 拥塞控制技术

拥塞控制是指对进入网络的数据总量进行控制，使网络流量保持在可接受水平的一种控制方法。拥塞控制与流量控制的区别在于，流量控制作用于接收者，而拥塞控制作用于网络。在当前的 RoCEv2 网络中，DCQCN 是应用最广泛的一种拥塞控制方法，华为智能无损网络对此拥塞控制算法进行了专门的功能增强与性能优化，以满足更广泛的应用场景。本章节介绍 ECN、DCQCN 的基本原理、AI ECN 的 iLossless 智能无损算法、ECN overlay 特性等。

4.2.1 ECN

ECN（Explicit Congestion Notification）是指流量接收端感知到网络上发生拥塞后，通过协议报文通知流量发送端，使得流量发送端降低报文的发送速率，从而从早期避免拥塞而导致的丢包，实现网络性能的最大利用，有如下优势：

- 所有流量发送端能够早期感知中间路径拥塞，并主动放缓发送速率，预防拥塞发生。
- 在中间交换机上转发的队列上，对于超过平均队列长度的报文进行 ECN 标记，并继续进行转发，不再丢弃报文。避免了报文的丢弃和报文重传。
- 由于减少了丢包，发送端不需要经过几秒或几十秒的重传定时器进行报文重传，提高了时延敏感应用的用户感受。
- 与没有部署 ECN 功能的网络相比，网络的利用率更好，不再在过载和轻载之前来回震荡。

那么，流量接收端是如何感知到网络上发生拥塞的呢？这里，需要先介绍一下 IP 报文中的 ECN 字段。

根据 RFC791 定义，IP 报文头 ToS（Type of Service）域由 8 个比特组成，其中 3 个比特的 Precedence 字段标识了 IP 报文的优先级，Precedence 在报文中的位置如下图所示。

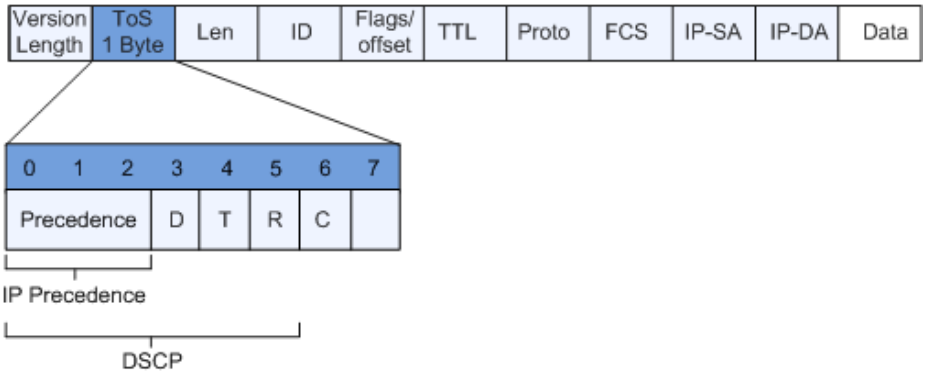


图 22：IP Precedence/DSCP 字段

比特 0～2 表示 Precedence 字段，代表报文传输的 8 个优先级，按照优先级从高到低顺序取值为 7、6、5、4、3、2、1 和 0。优先级是 7 或 6，经常是为路由选择或更新网络控制通信保留的，用户级应用仅能使用 0～5。

而比特 0～5 为 IP 报文的 DSCP，比特 6～7 为 ECN 字段。协议对 ECN 字段进行了如下规定：

- ECN 字段为 00，表示该报文不支持 ECN。
- ECN 字段为 01 或者 10，表示该报文支持 ECN。
- ECN 字段为 11，表示该报文的转发路径上发生了拥塞。

因此，中间交换机通过对将 ECN 字段置为 11，就可以通知流量接收端本交换机是否发生了拥塞。

在 CE 系列交换机中，ECN 功能是与 WRED 策略相结合应用的，应用这两个功能后，交换机会对收到的报文队列进行识别：

- **当实际队列长度小于报文丢包的低门限值时：**
  - 不对报文进行任何处理，直接进行转发。
- **当实际队列长度处于报文丢包的低门限值与高门限值之间时：**
  - 若设备接收到 ECN 字段为 00 的报文时，对报文按照概率进行丢包处理。
  - 若设备接收到 ECN 字段为 01 或者 10 的报文时，对报文按照概率将 ECN 字段修改为 11 后进行转发。
  - 若设备接收到 ECN 字段为 11 的报文时，对报文不做处理，直接进行转发。
- **当实际队列长度大于报文丢包的高门限值时：**
  - 若其 ECN 字段为 00 的报文时，对报文进行丢包处理。
  - 若其 ECN 字段为 01 或者 10 的报文时，对报文将 ECN 字段修改为 11 后进行转发。
  - 若其 ECN 字段为 11 的报文时，对报文不做处理，直接进行转发。

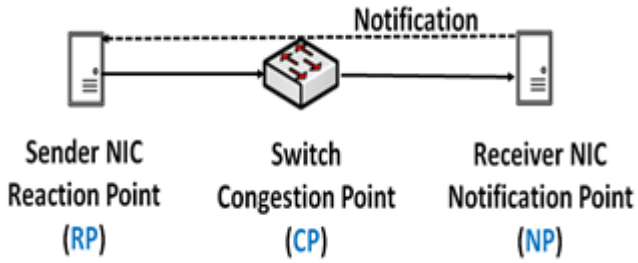
这样，当流量接收端收到 ECN 字段为 11 的报文时，就知道网络上出现了拥塞。这时，它向流量发送端发送协议通告报文，告知流量发送端存在拥塞。流量发送端收到该协议通告报文后，就会降低报文的发送速率，避免网络中拥塞的加剧。

当网络中拥塞解除时，流量接收端不会收到 ECN 字段为 11 的报文，也就不会往流量发送端发送用于告知其网络中存在拥塞的协议通告报文。此时，流量发送端收不到协议通告报文，则认为网络中没有拥塞，从而会恢复报文的发送速率。

4.2.2 DCQCN

DCQCN 全称为 Data Center Quantized Congestion Notification，是目前在 RoCEv2 网络中使用最广泛的拥塞控制算法，它融合了 QCN 算法和 DCTCP 算法，DCQCN 只需要可以支持 WRED 和 ECN 的数据中心交换机（市面上大多数交换机都支持），其他的协议功能在端节点主机的 NICs 上实现。DCQCN 可以提供较好的公平性，实现高带宽利用率，保证低的队列缓存占用率和较少的队列缓存抖动情况。

DCQCN 算法由三个部分组成，如下图所示：



- **交换机（CP，congestion point）**

CP 算法与 DCTCP 相同，如果交换机发现出端口队列超出阈值，在转发报文时就会按照一定概率给报文携带 ECN 拥塞标记（ECN 字段置为 11），以标示网络中存在拥塞。标记的过程由 WRED（Weighted Random Early Detection）功能完成。

WRED 是指按照一定的丢弃策略随机丢弃队列中的报文。它可以区分报文的服务等级，为不同的业务报文设置不同的丢弃策略。WRED 在丢弃策略中设置了报文丢包的高/低门限以及最大丢弃概率，（该丢弃概率就是交换机对到达报文标记 ECN 的概率）。并规定：

- 当实际队列长度低于报文丢包的低门限值时，不丢弃报文，丢弃概率为 0%。
- 当实际队列长度高于报文丢包的高门限值时，丢弃所有新入队列的报文，丢弃概率为 100%。
- 当实际队列长度处于报文丢包的低门限值与高门限值之间时，随机丢弃新到来的报文。随着队列中报文长度的增加，丢弃概率线性增长，但不超过设置的最大丢弃概率。

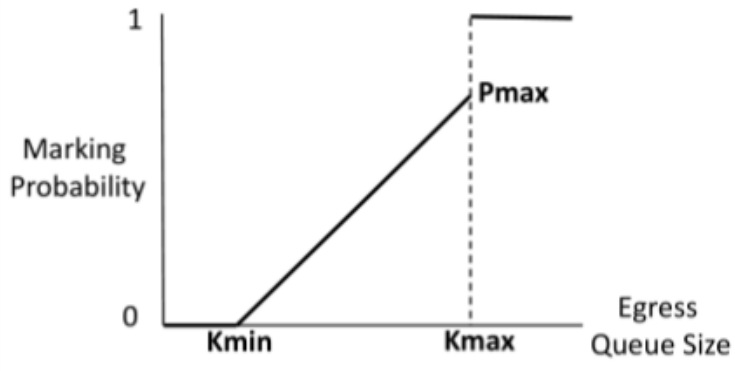


图 23：报文被标记的概率与队列长度关系

- **接收端（NP，notification point）**

接收端 NP 收到报文后，发现报文中携带 ECN 拥塞标记（ECN 字段为 11），则知道网络中存在拥塞，因此向源端服务器发送 CNP 拥塞通知报文（Congestion Notification Packets），以通知源端服务器进行流量降速。

NP 算法说明了 CNPs 应该什么时间以及如何产生：如果某个流的被标记数据包到达，并且在过去的 N 微秒的时间内没有相应 CNP 被发送，此时 NP 立刻发送一个 CNP。NIC 每 N 微秒最多处理一个被标记的数据包并为该流产生一个 CNP 报文。

- **发送端（RP，reaction point）**

当发送端 RP 收到一个 CNP 时，RP 将减小当前速率 Rc，并更新速率降低因子 α，和 DCTCP 类似，并将目标速率设为当前速率，更新速率过程如下：

RT = Rc

Rc = Rc * (1 - α/2)

α = (1-g)*α+g

如果 RP 在 K 微秒内没有收到 CNP 拥塞通知，那么将再次更新 α，此时 α = (1 - g) * α。注意 K 必须大于 N 即 K 必须大于 CNP 产生的时间周期。

进一步，RP 增加它的发送速率，该过程与 QCN 中的 RP 相同。

4.2.3 AI ECN

无损队列的动态 ECN 门限功能可以根据网络流量 N 对 1 的 Incast 值、大小流占比来动态调整无损队列的 ECN 门限，在尽量避免触发网络 PFC 流控的同时，尽可能的兼顾时延敏感小流和吞吐敏感大流。然而现网中的流量场景复杂多变，动态 ECN 门限功能并不能一一覆盖所有流量场景，无法帮助无损业务达到最优性能。而结合了 iLossless 智能无损算法的无损队列的 AI ECN 门限功能可以根据现网流量模型进行 AI 训练，对网络流量的变化进行预测，并且可以根据队列长度等流量特征调整 ECN 门限，进行队列的精确调度，保障整网的最优性能。

如下图所示，设备会对现网的流量特征进行采集并上送至 AI 业务组件，AI 业务组件将根据预加载的流量模型文件智能的为无损队列设置最佳的 ECN 门限，保障无损队列的低时延和高吞吐，从而让不同流量场景下的无损业务性能都能达到最佳。

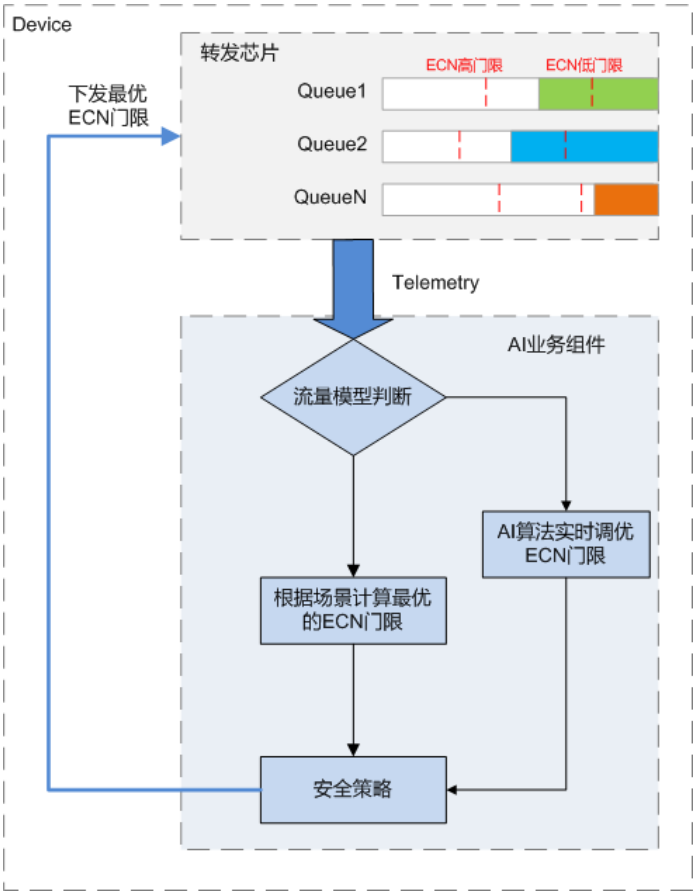


图 24：无损队列的 AI ECN 功能实现原理

1. Device 设备内的转发芯片会对当前流量的特征进行采集，比如队列缓存占用率、带宽吞吐、当前的 ECN 门限配置等，然后通过 Telemetry 技术将网络流量实时状态信息推送给 AI 业务组件。

2. AI 业务组件收到推送的流量状态信息后，将根据预加载的流量模型文件对当前的流量进行场景识别，判断当前的网络流量状态是否是已知场景。如果是已知场景，AI 业务组件将从积累了大量的 ECN 门限配置记忆样本的流量模型文件中，推理出与当前网络状态匹配的 ECN 门限配置。

3. 如果是未知的流量场景，AI 业务组件将结合 iLossless 智能无损算法，在保障高带宽、低时延的前提下，对当前的 ECN 门限不断进行实时修正，最终计算出最优的 ECN 门限配置。

4. 最后，AI 业务组件将符合安全策略的最优 ECN 门限下发到设备中，调整无损队列的 ECN 门限。

5. 对于获得的新的流量状态，设备将重复进行上述操作，从而保障无损业务的最佳性能。

无损队列的 AI ECN 门限功能可以根据现网流量模型进行 AI 训练，对网络流量的变化进行预测，并且可以根据队列长度等流量特征调整 ECN 门限，进行队列的精确调度，保障无损业务的最优性能。

同时，与拥塞管理技术（队列调度技术）配合使用时，无损队列的 AI ECN 门限功能可以实现网络中 TCP 流量与 RoCEv2 流量的混合调度，保障 RoCEv2 流量的无损传输的同时实现低时延和高吞吐。

4.2.4 ECN overlay

ECN Overlay 功能即将 ECN（Explicit Congestion Notification）功能应用到 VXLAN 网络中，使 Overlay 网络中的拥塞状态可以及时被流量接收端感知，让流量接收端知会流量发送端进行降速，缓解网络拥塞。

根据 VXLAN 网络中拥塞发生点不同，ECN Overlay 功能的工作原理如下所示：

- **在入 VXLAN 隧道的设备上发生拥塞（Ingress NVE 场景）**

如果报文在进入 VXLAN 隧道前已经发生拥塞，即 Ingress NVE 对报文进行 VXLAN 封装前，其原始报文中的 ECN 字段已置为 CE 标记，那么 ECN 字段在 VXLAN 网络中转发的流程如图 25 所示。

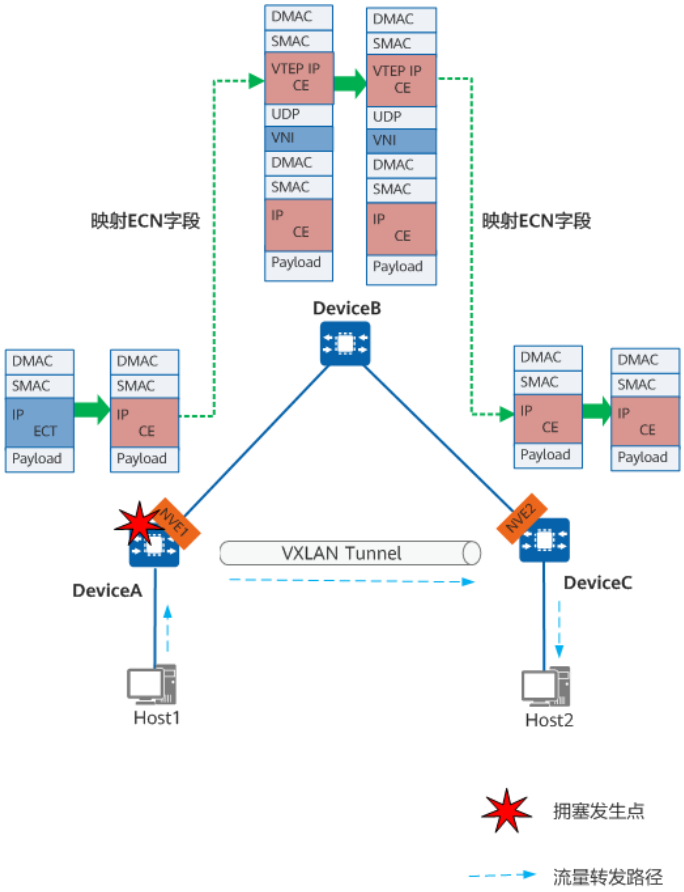


图 25：Ingress NVE 场景下的 ECN 拥塞标记

a. 报文从 Host1 进入 DeviceA 设备，此时报文所在队列在 DeviceA 出现拥塞，DeviceA 将报文中的 ECN 字段置为 CE。

b. 在 NVE1 上对原始报文进行 VXLAN 封装，报文进入 VXLAN 隧道进行转发。此时，原始报文 IP 头中的 ECN 字段将映射到 VXLAN 外层 IP 头中。

c. Overlay 中，通过 Underlay 设备 DeviceB，根据外层 IP 头将报文的 ECN 字段携带至 NVE2。

d. 在 NVE2 上对 VXLAN 报文进行解封装，根据 VXLAN 外层 IP 头中的 ECN 字段映射到原始报文 IP 头中，最终通过原始报文将 ECN 拥塞标记传递到 Host2。

- **在 VXLAN 隧道的转发设备中发生拥塞（Overlay 网络场景）**

如果报文在 Overlay 网络中进行转发时发生拥塞，即报文在中间转发设备上才将 ECN 字段置为 CE 标记，那么 ECN 字段在 VXLAN 网络中转发的流程如图 26 所示：

a. 报文从 Host1 进入 DeviceA 设备，此时报文所在队列在 DeviceA 没有出现拥塞，报文中的 ECN 字段一直保持为 ECT 状态。

b. 在 NVE1 上对原始报文进行 VXLAN 封装，报文进入 VXLAN 隧道进行转发。此时，原始报文 IP 头中的 ECN 字段将映射到 VXLAN 外层 IP 头中。

c. 报文在 Overlay 网络中进行转发时发生拥塞，即报文所在队列在 Underlay 设备 DeviceB 上出现拥塞，DeviceB 将 VXLAN 外层 IP 头的 ECN 字段置为 CE。

d. 在 NVE2 上对 VXLAN 报文进行解封装，根据 VXLAN 外层 IP 头中的 ECN 字段映射到原始报文 IP 头中，最终通过原始报文将 ECN 拥塞标记传递到 Host2。

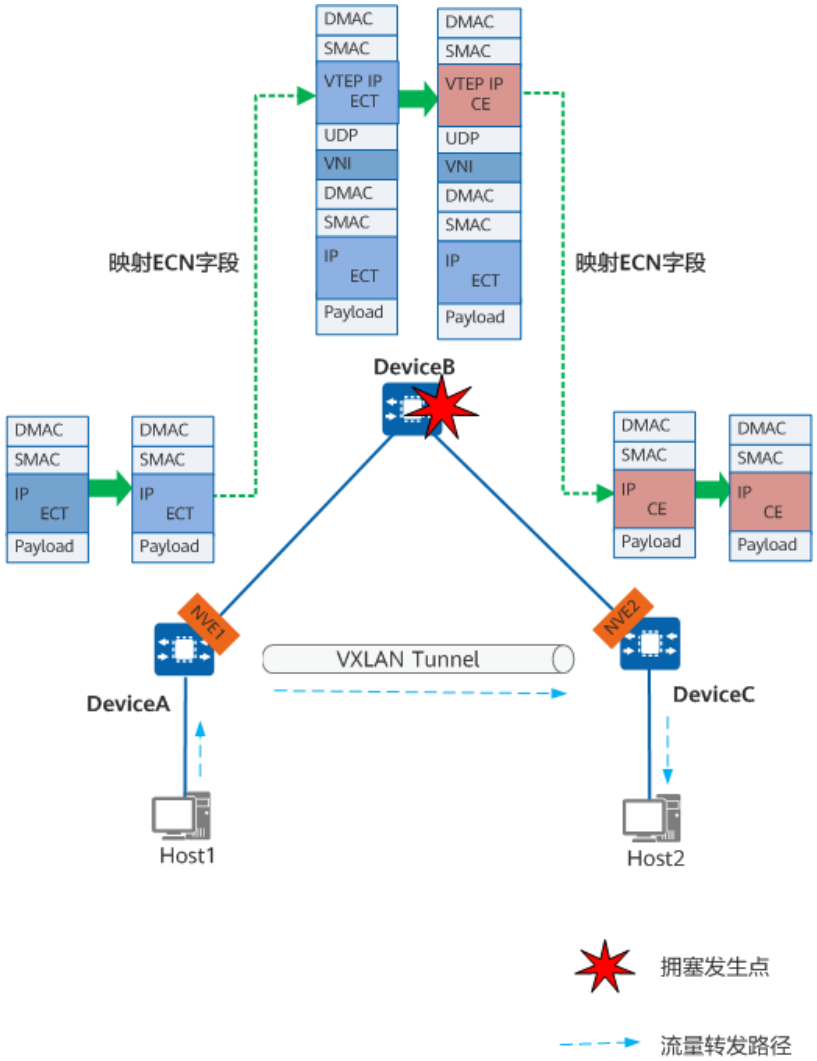


图 26：Overlay 网络场景下的 ECN 拥塞标记

- **在出 VXLAN 隧道的设备上发生拥塞（Egress NVE 场景）**

如果报文出 VXLAN 隧道的设备上发生拥塞，即 Egress NVE 对报文进行 VXLAN 接封装后，报文中的 ECN 字段才置为 CE 标记，那么 ECN 字段在 VXLAN 网络中转发的流程如图 27 所示：

a. 报文从 Host1 进入 DeviceA 设备，此时报文所在队列在 DeviceA 没有出现拥塞，报文中的 ECN 字段一直保持为 ECT 状态。

b. 在 NVE1 上对原始报文进行 VXLAN 封装，报文进入 VXLAN 隧道进行转发。此时，原始报文 IP 头中的 ECN 字段将映射到 VXLAN 外层 IP 头中。

c. Overlay 中，通过 Underlay 设备 DeviceB，根据外层 IP 头将报文的 ECN 字段携带至 NVE2。

d. 在 NVE2 上对 VXLAN 报文进行解封装，根据 VXLAN 外层 IP 头中的 ECN 字段映射到原始报文 IP 头中。

e. 报文所在队列在 DeviceC 上出现拥塞，DeviceC 将原始报文 IP 头的 ECN 字段置为 CE，最终通过原始报文将 ECN 拥塞标记传递到 Host2。

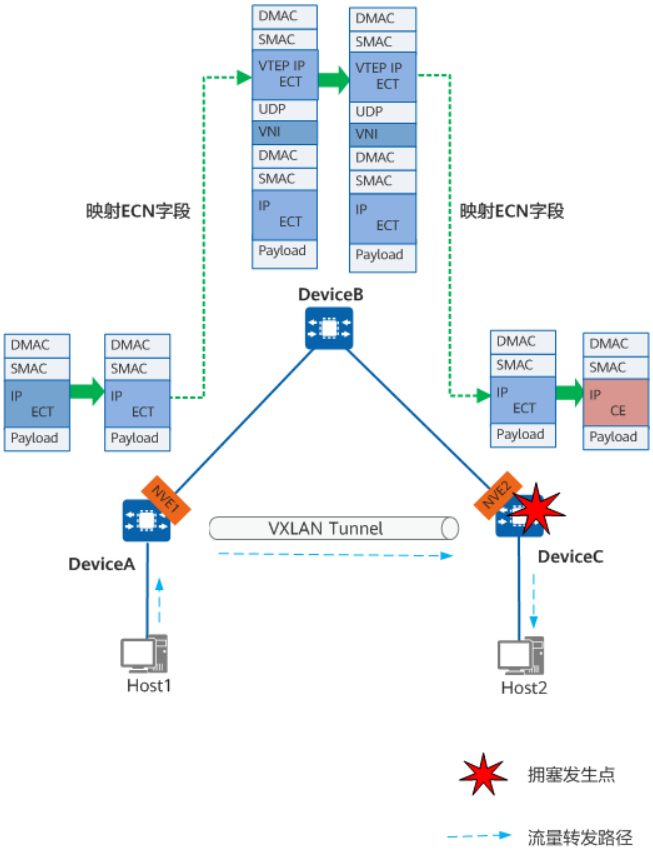


图 27：Egress NVE 场景下的 ECN 拥塞标记

4.2.5 iQCN

iQCN（intelligent Quantized Congestion Notification）通过让转发设备智能的补偿发送 CNP（Congestion Notification Packets）拥塞通知报文，解决流量发送端网卡未及时收到 CNP 报文而迅速升速带来的网络拥塞加剧的问题。iQCN 的工作原理如下图所示。

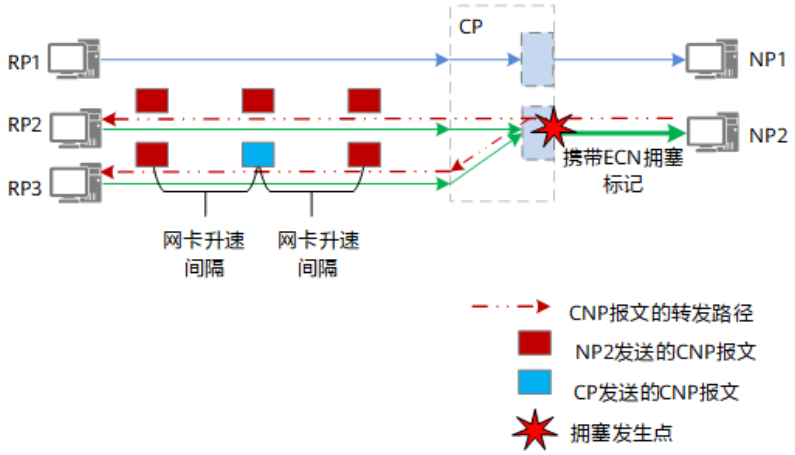


图 28：iQCN 工作原理图

1. 报文从 RP1 发往 NP1，若 CP 没有发生拥塞，流量正常转发。

2. 报文从 RP2 和 RP3 发往 NP2，在 CP 的端口出方向发生了拥塞，CP 对报文中进行 ECN 拥塞标记后将报文转发给 NP2。

3. NP2 收到携带了 ECN 拥塞标记的报文后，获知网络中出现了拥塞，NP2 的网卡向 RP2 和 RP3 发送 CNP 拥塞通知报文，通知 RP2 和 RP3 的网卡降低发送报文的速率。若 CP 的出端口持续拥塞，则 NP2 将持续发送 CNP 拥塞通知报文。

4. 使能了 iQCN 功能的 CP 会对收到的 CNP 报文进行记录，维护包含 CNP 报文信息和时间戳的流表。同时 CP 会对本设备的端口拥塞程度进行持续监测，端口拥塞较为严重时，将收到 CNP 报文的时间间隔与网卡升速时间进行比较：

- 若发现从 NP 收到 CNP 报文的时间间隔小于 RP 的网卡升速时间，判断网卡可以正常降速，CP 正常转发 CNP 报文。
- 若发现从 NP 收到 CNP 报文的时间间隔大于 RP 的网卡升速时间，判断网卡不能及时降速且存在升速风险，CP 将会主动补偿发送 CNP 报文。

4.3 流量调度技术

4.3.1 负载分担

负载分担（Load Balance），指的是网络节点在转发流量时，将负载（流量）分摊到多条链路上进行转发，要在网络中存在多条路径的情况下，比如 all-to-all 流量模型下，实现无损网络，达成无丢包损失、无时延损失、无吞吐损失，需要引入该机制。数据中心中常用的负载分担机制为等价多路径路由 ECMP 和链路聚合 LAG。

ECMP 负载分担

等价多路径路由 ECMP（Equal-Cost Multi-Path routing）实现了等价多路径负载均衡和链路备份的目的。

ECMP 应用于多条不同链路到达同一目的地址的网络环境中。当多条路由的路由优先级和路由度量都相同时，这几条路由就称为等价路由，多条等价路由可以实现负载分担。当这几条路由为非等价路由时，就可以实现路由备份。如果不使用 ECMP 转发，发往该目的地址的数据包只能利用其中的一条链路，其它链路处于备份状态或无效状态，并且在动态路由环境下相互的切换需要一定时间。而 ECMP 可以在该网络环境下同时使用多条链路，不仅增加了传输带宽，并且可以无时延无丢包地备份失效链路的数据传输。

当实现 ECMP 负载分担时，路由器将数据包的五元组（源地址、目的地址、源端口、目的端口、协议）作为哈希因子，通过 HASH 算法生成 HASH-KEY 值，然后根据 HASH-KEY 值在负载分担链路中选取一条成员链路对数据包进行转发。当五元组相同时，路由器总是选择与上一次相同的下一跳地址发送报文。当五元组不同时，路由器会选取相对空闲的路径进行转发。

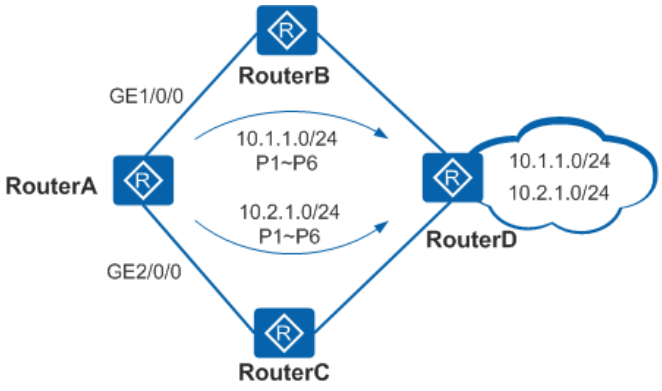


图 29：负载分担组网图

如图 29 所示，RouterA 已经通过接口 GE1/0/0 转发到目的地址 10.1.1.0/24 的第 1 个报文 P1，随后又需要分别转发报文到目的地址 10.1.1.0/24 和 10.2.1.0/24。其转发过程如下：

- 当转发到达 10.1.1.0/24 的第 2 个报文 P2 时，发现此报文与到达 10.1.1.0/24 的第 1 个报文 P1 的五元组一致。所以之后到达该目的地的报文都从 GE1/0/0 转发。
- 当转发到达 10.2.1.0/24 的第 2 个报文 P2 时，发现此报文与到达 10.1.1.0/24 的第 1 个报文 P1 的五元组不一致。所以选取从 GE2/0/0 转发，并且之后到达该目的地的报文都从 GE2/0/0 转发。

正常情况下，路由器采用主路由转发数据。当主链路出现故障时，主路由变为非激活状态，路由器选择备份路由中优先级最高的路由转发数据。这样，也就实现了主路由到备份路由的切换。当主链路恢复正常时，由于主路由的优先级最高，路由器重新选择主路由来发送数据。这样，就实现了从备份路由回切到主路由。

LAG 负载分担

随着网络规模不断扩大，用户对骨干链路的带宽和可靠性提出越来越高的要求。在传统技术中，常用更换高速率的设备的方式来增加带宽，但这种方案需要付出高额的费用，而且不够灵活。

采用链路聚合（Link Aggregation）技术可以在不进行硬件升级的条件下，通过将多个物理接口捆绑为一个逻辑接口，达到增加链路带宽的目的。在实现增大带宽目的的同时，链路聚合采用备份链路的机制，可以有效的提高设备之间链路的可靠性。

链路聚合组 LAG（Link Aggregation Group）是指将若干条以太链路捆绑在一起所形成的逻辑链路。每个聚合组唯一对应着一个逻辑接口，这个逻辑接口称之为聚合接口或 Eth-Trunk 接口。

如图 30 所示，DeviceA 与 DeviceB 之间通过三条以太网物理链路相连，将这三条链路捆绑在一起，就成为了一条逻辑链路，这条逻辑链路的最大带宽等于原先三条以太网物理链路的带宽总和，从而达到了增加链路带宽的目的；同时，这三条以太网物理链路相互备份，有效地提高了链路的可靠性。

图 30：Eth-Trunk 示意图

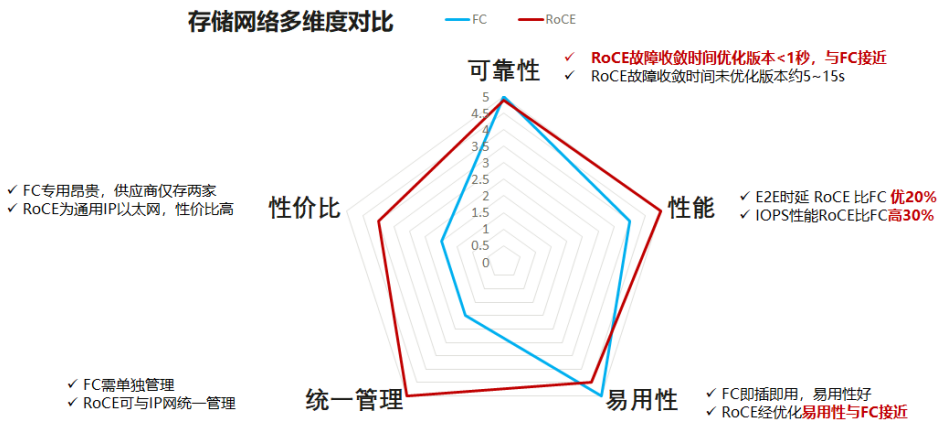
Eth-Trunk 接口可以作为普通的以太网接口来使用，实现各种路由协议以及其它业务。与普通以太网接口的差别在于：转发的时候 LAG 需要从 Eth-Trunk 成员接口中选择一个或多个接口来进行数据转发。

因此，在使用 Eth-Trunk 转发数据时，由于聚合组两端设备之间有多条物理链路，就会产生同一数据流的第一个数据帧在一条物理链路上传输，而第二个数据帧在另外一条物理链路上传输的情况。这样一来同一数据流的第二个数据帧就有可能比第一个数据帧先到达对端设备，从而产生接收数据包乱序的情况。

为了避免这种情况的发生，Eth-Trunk 采用逐流负载分担的机制，这种机制把数据帧中的地址通过 HASH 算法生成 HASH-KEY 值，然后根据这个数值在 Eth-Trunk 转发表中寻找对应的出接口，不同的 MAC 或 IP 地址 HASH 得出的 HASH-KEY 值不同，从而出接口也就不同，这样既保证了同一数据流的帧在同一条物理链路转发，又实现了流量在聚合组内各物理链路上的负载分担。

5 网络与存储协同

传统的存储区域网络（SAN）有两个主要的技术，一个是 FC，一个是以太。长期以来，因为 FC 在性能、可靠性等方面的优势，同时由于对 SAN 有使用需求的应用场景对性能均极度敏感，FC 占据了存储网络的绝大部分份额。但是，随着全闪存储的普及，以及随之而来的 NVMe over Fabric 技术，情况正在发生变化。RoCEv2 在性能、性价比、统一管理等角度比 FC 有显著优势，劣势部分在易用性、存储业务多路径倒换可靠性方面。为了弥补 RoCEv2 在存储网络的劣势，开发了智能无损存储网络（NOF+），主要包括两方面技术：存储网络区域划分、网络故障与存储多路径联动。



5.1 存储网络区域划分

SAN 网络中一个常用的概念叫做 Zone，Zone 具备两个层面的功能，第一是让处于相同 Zone 内的主机与存储彼此可见，第二是让处于不同 Zone 之间成员转发隔离，主要从安全角度考虑。

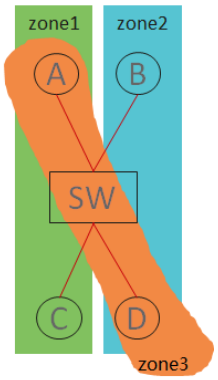
以下图为例说明 Zone 功能，A、B、C、D 四个终端通过单个交换机互联，划分三个区域：

zone1：包括 A、D 两个接入终端

zone2：包括 B、C 两个接入终端

zone3：包括 B、C 两个接入终端

网络需要做到相同 zone 内的终端可以互访，但不同 zone 之间的终端不可以互访。



对于 MAC 转发的网络，直接的办法是采用 VLAN 作为隔离手段；对于 IP 转发的网络，直接的办法是采用 VRF 作为隔离手段，但是问题在于：一个较大规模的网络会划分上万个 Zone，不论采用 VLAN 亦或 VRF 作为隔离手段，都会面临资源不足的问题。为了解决这个问题，我们设计了全新的 NOF+协议。如下图所示：整网设置其中一台或两台设备可以作为反射器（Reflector），其余设备作为客户端（Client）。每个 NOF+反射器和 NOF+客户端之间都需要建立 NOF+连接从而可以传输 NOF+报文。NOF+客户端之间不需要建立 NOF+连接，只需要与主机（Host）直连。NOF+反射器之间不需要建立 NOF+连接，两者互为备份，与主机直连的接入设备也可以作为 NOF+反射器。

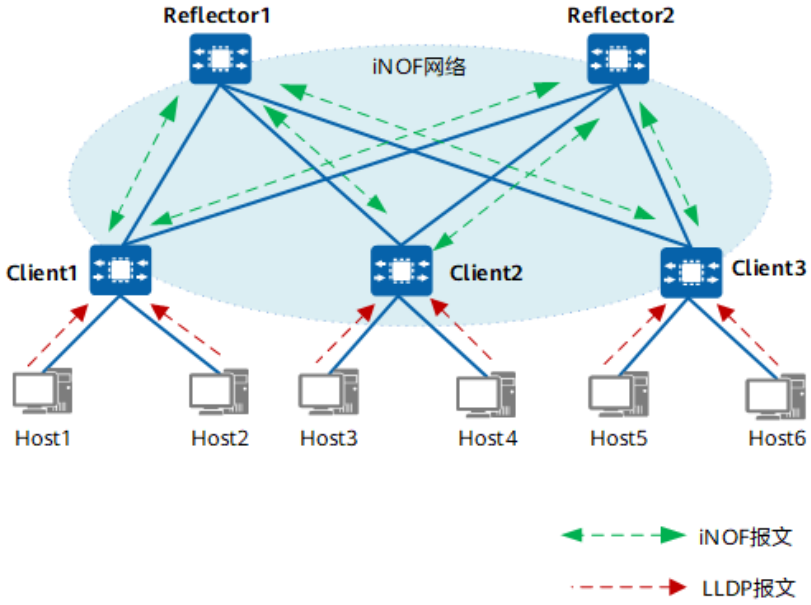


图 31：NOF+原理图

NOF+报文是 TCP 封装的报文，TCP 端口号范围为 10000 到 57999，缺省值为 19516，包含 NOF+关键信息的内容承载在 TCP 报文的 Data 字段内。客户端可以通过 NOF+报文将 NOF+关键信息发送给反射器，反射器汇总后再发往其他客户端。通过 NOF+报文可以传输以下几类信息：

- 建连信息：NOF+反射器和客户端之间需要通过互相交换 NOF+报文来建立 NOF+连接，具体的建立过程类似 TCP 建连。
- 域配置信息：NOF+系统中，设备可以通过域（Zone）对接入的主机进行管理，NOF+反射器上完成 NOF+域的相关配置后，会通过 NOF+报文把域配置信息发往各个客户端。若一个 NOF+系统中同时存在两个反射器，只有在两个反射器上相同的域配置才可以生效。
- 主机动态信息：主机的网卡和 NOF+设备均需要启用 LLDP 功能，当有新的主机接入客户端或者离开客户端时，主机会主动向客户端发送 LLDP 报文，报文内记录了 LLDP 邻居信息的变化，让 NOF+系统内的其他设备感知到主机动态信息。

5.2 网络故障与存储多路径联动

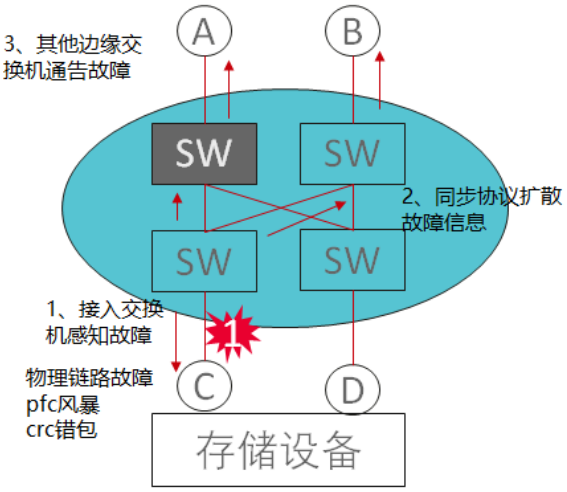
多路径软件功能介绍：在主机与磁阵的连接中，通常会使用两个 HBA 卡连接磁阵的 A、B 两个控制器。在没有多路径的情况下，对于磁阵上的一个 LUN，在主机会看到两个大小一样的硬盘。在安装了多路径软件后，会形成一个虚拟的设备，操作系统只需对虚拟的设备进行读写，底层具体通过哪个通道访问磁阵由多路径软件完成，达到 loadbalance 和 failover 的效果。当主机与存储之间通过网络相连时，如果网络连接中断，多路径切换依赖与 keepalive 机制感知超时，通常切换时间约 10 秒钟，网络故障与存储多路径联动的目的是提升此种场景下的切换性能。

场景 1：接入端口的故障

步骤 1：接入交换机感知故障，故障类型包括物理链路故障、PFC 风暴、CRC 错包等

步骤 2：采用 NOF+同步协议，将接入端口关联不可达 IP 信息同步至整网

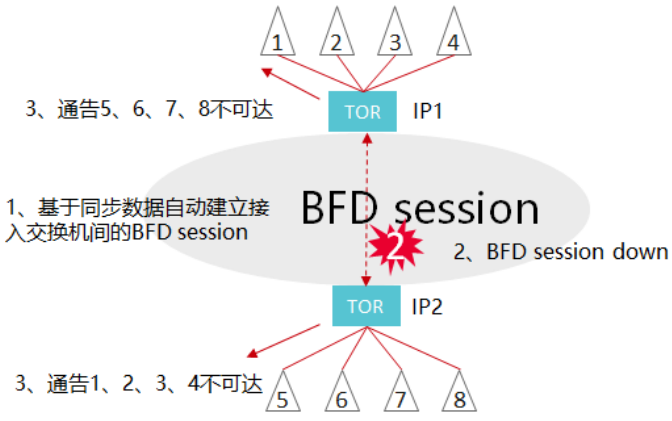
步骤 3：其他边缘交换机检查与故障 IP 处于相同 Zone 内的 IP，向关联的 IP 进行通告



场景 2：网络级联端口故障

步骤 1：基于 NOF+的同步数据，建立 TOR 之间的 BFD 测量会话

步骤 2：BFD 测量会话两端的 TOR，在感知到 BFD session down 时，分别通告对端接入的 IP 不可达。如下图所示：



6 网络与计算协同

通过对超算网络的流量模型进行分析，可以得到超算网络的典型特性：80%以上均为小于 16 字节 payload 的报文，极端考验设备的静态时延。典型的以太网芯片静态时延约 500ns，IB 芯片时延约 90ns。从静态时延角度看，以太相比 IB 存在劣势，这种劣势通过网络与计算的协同可以改善甚至消除。

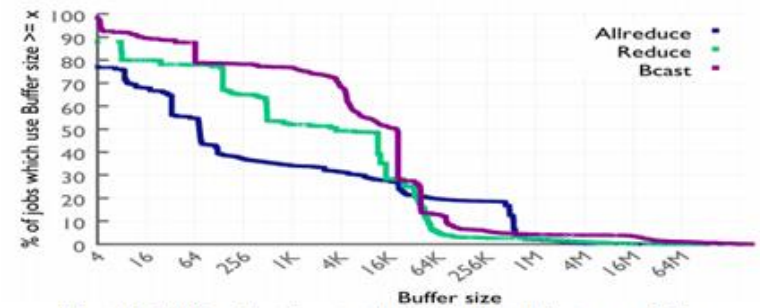


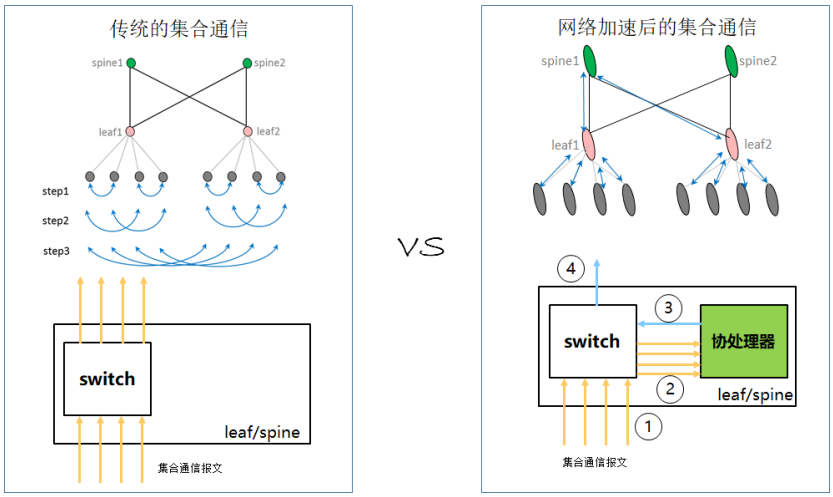
Fig. 16: MPI collectives total accumulated bytes on Mira

参考《Characterization of MPI Usage on a Production Supercomputer》

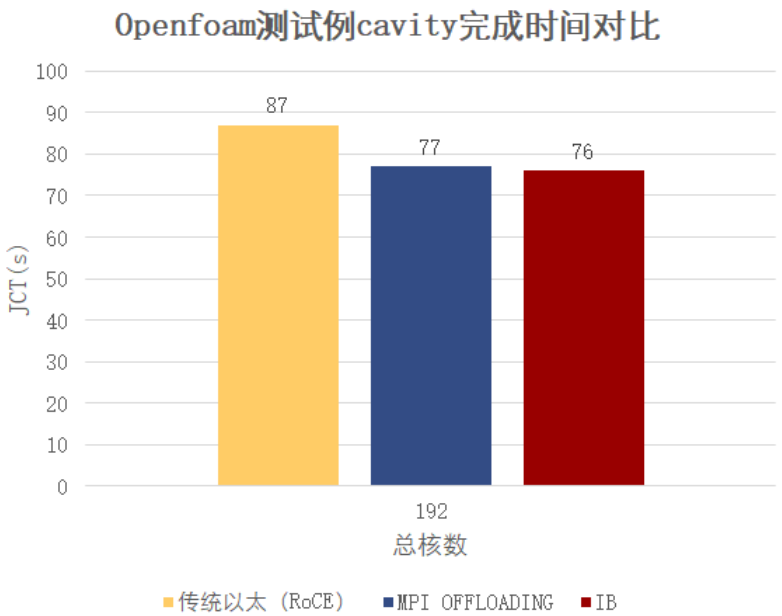
6.1 集合通信加速

传统集合通信过程中的计算在服务器侧完成，网络只负责转发；采用交换设备对集合通信加速，卸载一部分集合通信的计算过程，可以有效地提高集合通信的效率，从而降低总的任务完成时间。

如下图所示，左侧为传统集合通信方式，8 个节点进行 allreduce 计算总共需要 3 个批次的通信，复杂度为 logN；右侧为网络加速后的集合通信方式，8 个节点进行 allreduce 计算，由接入 leaf 进行第一次汇聚，由 spine 交换机进行第二次汇聚，总的通信批次只与网络的层次数量有关，复杂度变为常量。



基于 Openfoam 测试例 cavity 测试任务完成时间，使用网络加速的集合通信方案，相较于采用普通 RoCEv2 的传统集合通信方法，完成时间缩短了 12%；相较采用 IB 网络的传统集合通信方法偏差在 1.3%。



7 智能无损网络运维

为了采集智能无损网络的流量状态，CloudEngine 系列交换机提供了可以向 iMaster NCE-FabricInsight 推送网络设备的各项高精度性能 Metrics 数据的 Telemetry 技术。

7.1 Temetry 原理介绍

Telemetry 是一项从物理设备或虚拟设备上远程高速采集性能数据的网络监控技术。相比于传统的网络监控技术，Telemetry 通过推模式（Push Mode）高速且实时的向 FabricInsight 推送网络设备的各项高精度性能数据指标，提高了采集过程中设备和网络的利用率。

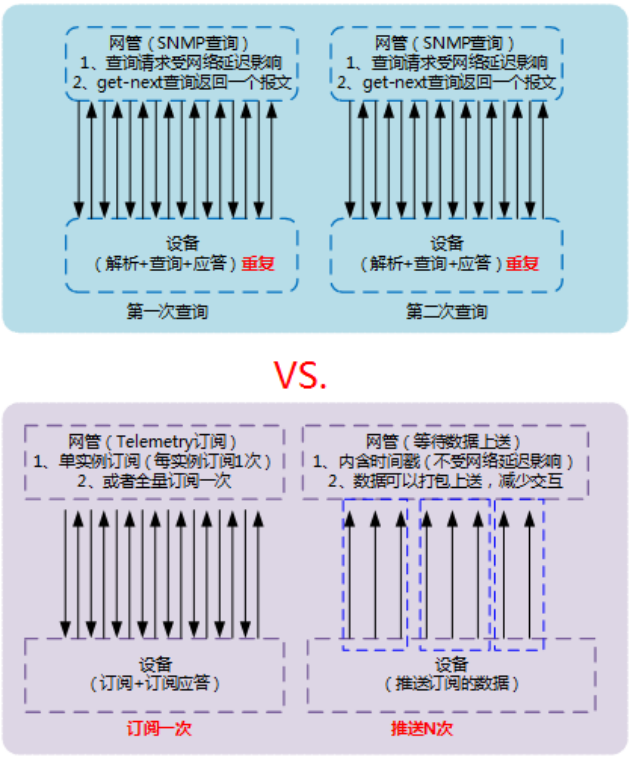


图 32：SNMP 查询过程与 Telemetry 采样过程对比

如上图所示，与传统网络监控技术（SNMP-get）相比，Telemetry 具有如下优势：

- **通过推模式主动上送采样数据，扩大了被监控节点的规模**

在传统网络监控技术中，网管与设备之间是一问一答式交互的拉模式。假设 1 分钟内需要交互 1000 次数据才能完成查询过程，则意味着设备解析了 1000 次的查询请求报文。第 2 分钟设备将再次解析 1000 次的查询请求报文，如此持续下去。实际上，第 1 分钟和第 2 分钟解析的 1000 次查询请求报文是一样的，后续设备每分钟都需要重复解析 1000 次的查询请求报文。查询请求报文的解析需要消耗设备的 CPU 资源，因此为了不影响设备的正常运行，则必须限制设备被监控节点的数量。

在 Telemetry 技术中，网管与设备之间采用的是推模式。在第 1 分钟内，网管向设备下发 1000 次的订阅报文，设备解析 1000 次的订阅报文，在解析订阅报文的过程中，设备记录下网管的订阅信息。后续每分钟内，网管不再向设备下发订阅报文，设备根据记录的订阅信息自动且持续的向网管推送数据。这样每分钟都节省了 1000 次订阅报文的解析，也就节省了设备的 CPU 资源，使得设备能够被监控更多的节点。

- **通过打包方式上送采样数据，提高了数据采集的时间精度**

在传统网络监控技术中，设备每分钟内都要解析大量的查询请求报文，且对于一个查询请求报文只上送一个采样数据。而查询请求报文的解析也需要消耗设备的 CPU 资源。因此为了不影响设备的正常运行，必须限制网管下发查询请求报文的频度，也就降低了设备数据采集的时间精度。通常来说，传统网络监控技术的采样精度为秒级。

在 Telemetry 技术中，只有第 1 分钟设备需要解析订阅报文，其他时间内设备都不需要解析订阅报文，且对于一个订阅报文可以通过打包方式上送多个采样数据，进一步减少了网管与设备之间交互报文的次数。因此，Telemetry 技术的采样精度可以达到毫秒级乃至亚秒级。

- **通过携带时间戳信息，提升了采样数据的准确性**

在传统网络监控技术中，采样数据中没有时间戳信息，由于网络传输时延的存在，网管监控到的网络节点数据并不准确。

在 Telemetry 技术中，采样数据中携带时间戳信息，网管进行数据解析时能确认采样数据的发生时间，从而避免了网络传输延迟对采样数据的影响。

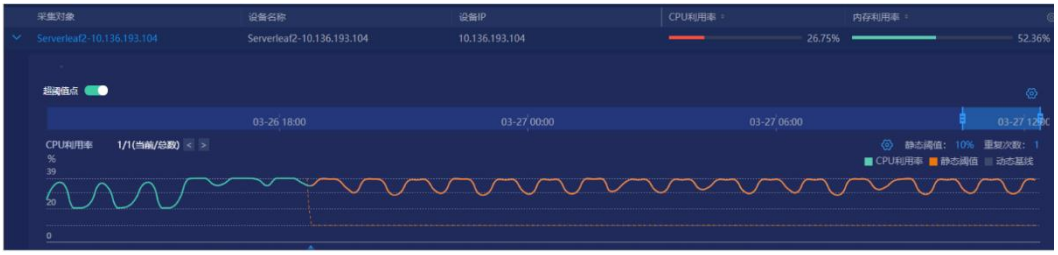
通过 Telemetry 技术，可以对整个无损网络中各个无损队列的拥塞状态，比如 PFC 报文、ECN 报文、拥塞丢弃报文等等，进行实时监控，协助拥塞控制算法进行参数优化，实现 0 丢包、低时延、高吞吐的智能无损网络，最终帮助客户构建与传统以太网兼容的 RDMA，引领数据中心网络进入极速无损的高性能时代。

7.2 智能无损网络可视化

iMaster NCE-FabricInsight 系统利用 CloudEngine 系列交换机设备的 Telemetry 特性采集设备、接口、队列等性能 Metrics 数据进行分析，主动监控、预测网络异常。通过大数据平台，对数据进行统一采集、存储和分析，具有高效的大数据处理能力。

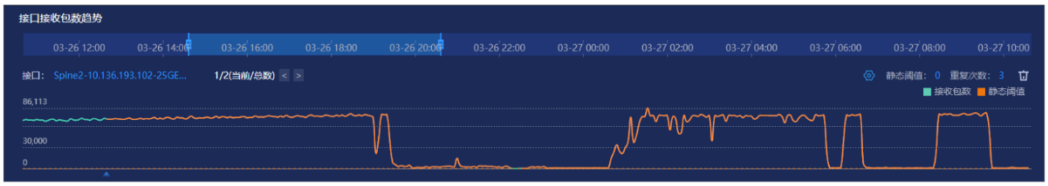
- **设备级信息**

支持的指标包括：CPU 使用率、内存使用率



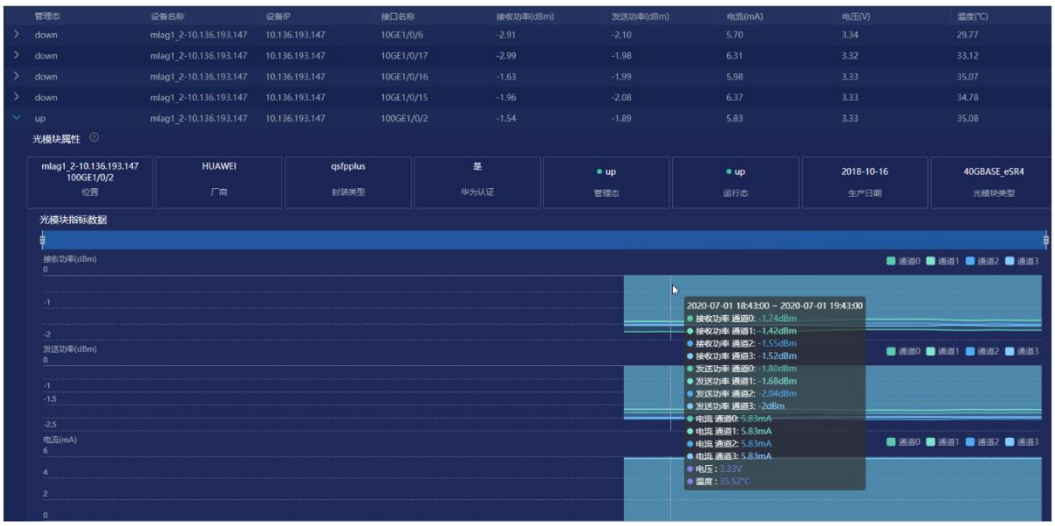
- **端口级信息**

支持查看的指标包括：接收字节数、发送字节数、接收包数、发送包数、接收丢包数、发送丢包数、单播/组播/广播接收包数、单播/组播/广播发送包数、接收错误包数、发送错误包数、接收带宽占用率、发送带宽占用率、ECN 报文数等。



- **光模块信息**

支持查看光模块的基本属性和各种性能指标。

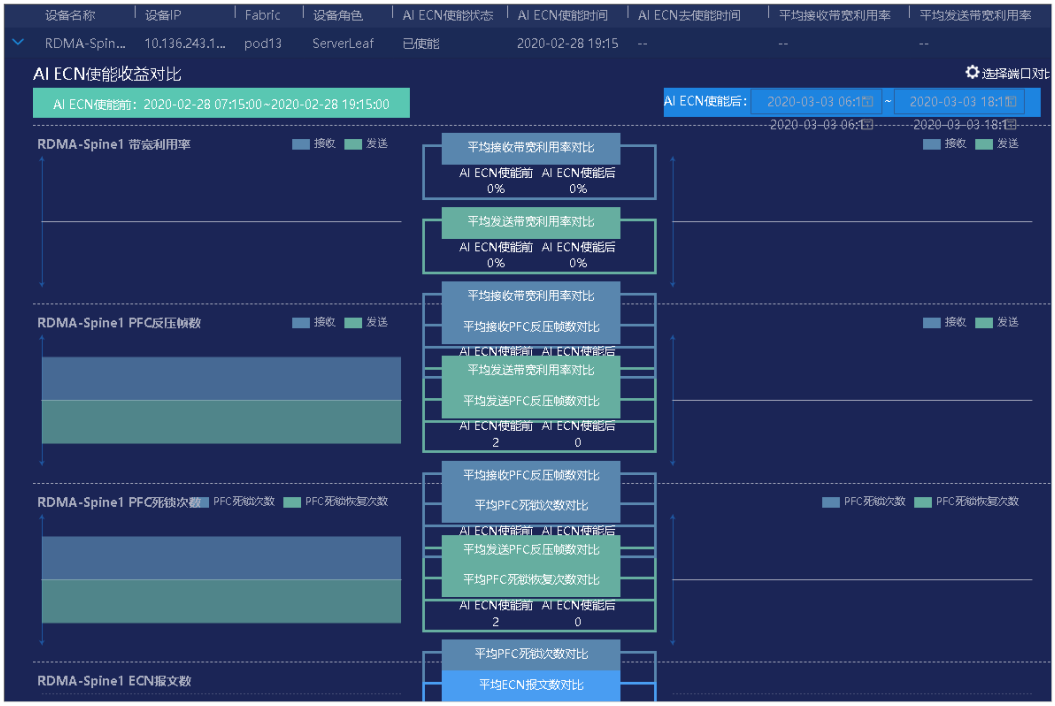


- **队列级信息**

支持查看的指标包括：队列缓存、已使用 Headroom 缓存、接收 PFC 反压帧数、发送 PFC 反压帧数、PFC 死锁监控次数、PFC 死锁恢复次数、已使用 Guaranteed 缓存。

- **AI ECN 特性信息**

支持查看 AI ECN 使能后的性能指标详情，对比图从带宽利用率、PFC 反压帧数、PFC 死锁次数、ECN 报文数等性能指标进行对比，展示使能 AI ECN 给 RoCE 网络设备带来的收益。



8 智能无损网络性能测试

美国 Tolly Group 实验室是国际权威评测机构，华为智能无损网络与业界其他主流厂商基于 RDMA 的数据中心交换机组网方案进行了性能对比测试，Tolly 选取 AI 时代数据中心的高性能计算（HPC）、人工智能/机器学习（AI/ML）和分布式存储三大典型应用场景，在全 100GE 和 100GE/25GE 相同 Spine-leaf 组网环境下，进行了 IMB（Intel MPI Benchmark）模型、TensorFlow 模型和吞吐率、丢包率及时延等性能指标的全面对比测试评估。经过严格测试证明，华为智能无损网络全面领先，性能平均高出 30%左右，并且 Tolly 工程师确认了了华为智能无损网络方案组网内无丢包。

8.1 与思科 Nexus 交换机组网性能对比

华为和思科的方案均基于 RDMA over ConvergedEthernet（RoCEv2）。在所有三大场景中，华为智能无损网络解决方案的性能均优于思科。

- **高性能计算（HPC）**

组网拓扑图如下图所示，全网一共有 9 台计算节点，每根蓝色连接代表 3 条 100G 链路；每根灰色连接代表 1 条 100G 链路。设备选型如表 8-1 所示。

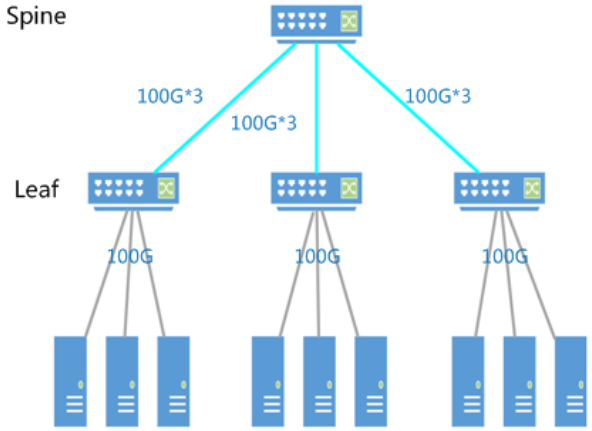


图 33：高性能计算（HPC）以及分布式 AI 训练场景测试拓扑

表8-1 HPC 场景设备选型

厂商	设备	数量	角色	组网
Huawei-智能无损网络	CE8850-64CQ-EI	1	Spine	RoCEv2
	CE8861-4C-EI	2	Leaf	
Cisco	Nexus 92300YX	1	Spine	RoCEv2
	Nexus 93180YC-EX	2	Leaf	
	Nexus 93180YC-FX	1	Leaf	

在 CE 交换机上配置智能无损网络中的 PFC、无损队列的缓存空间优化和动态 ECN 门限功能，在 Cisco 上完成 ROCE 配置。设置网卡模式为 RoCEv2，运行 Intel MPI Benchmarks（IMB）测试工具，分别测试 8 字节和 16 字节下的任务完成时间。

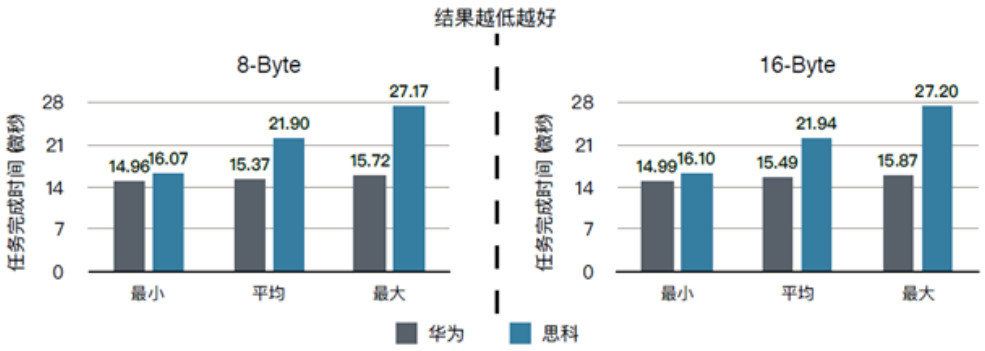


图 34：高性能计算（HPC）场景 IMB 测试结果（MPI_Allreduce）

显然，智能无损网络方案中的任务完成时间更低，说明整网时延更低，并且 Tolly 工程师确认了了华为智能无损网络方案组网内无丢包。

- **分布式 AI 训练场景**

AI 场景中的组网拓扑和设备选型均与 HPC 中一致，在 CE 交换机上配置智能无损网络中的 PFC、无损队列的缓存空间优化、动态 ECN 门限和动态负载分担功能，在 Cisco 上完成 ROCE 配置并使能 Cisco 的 PFC 功能。服务器安装 GPU，一共 9 台计算节点，每节点装有一块 NVIDIA Tesla P100（16G）GPU。设置网卡模式为 RoCEv2，编译安装 Tensorflow 1.10，测试 AI-Tensorflow 图片训练性能。

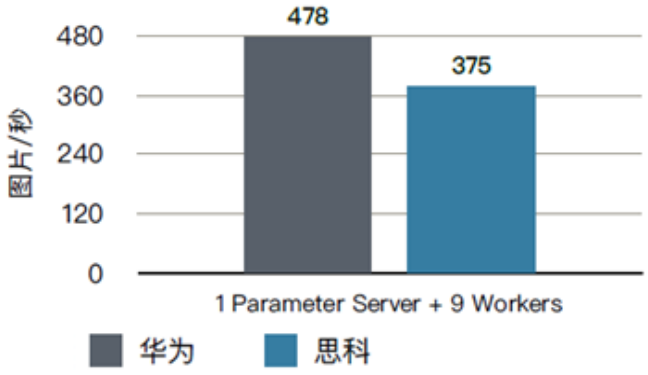


图 35：TensorFlow-GPU 分布式 AI 训练下加速比

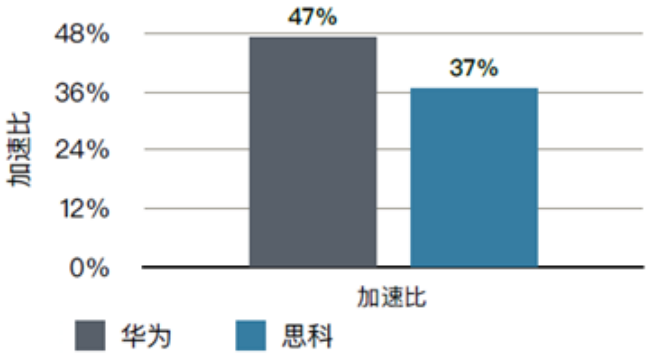


图 35：TensorFlow-GPU 分布式 AI 训练下每秒钟处理训练图片的数量:

加速比指分布式场景整体训练性能/（单节点训练性能 * 节点个数）的百分比，是验证 AI 性能的主要判定指标。可以看出，无论是每秒钟处理图片的数量还是加速比，都说明在智能无损网络组网下，能更好的支撑 GPU 的计算，提升 AI 训练的效果。

- **分布式存储场景**

组网拓扑图如下图所示，全网一共有 4 台存储节点，12 台计算节点。每根蓝色连接代表 4 条 100G 链路，每根灰色连接代表 1 条 25G 链路。设备选型如表 8-2 所示。

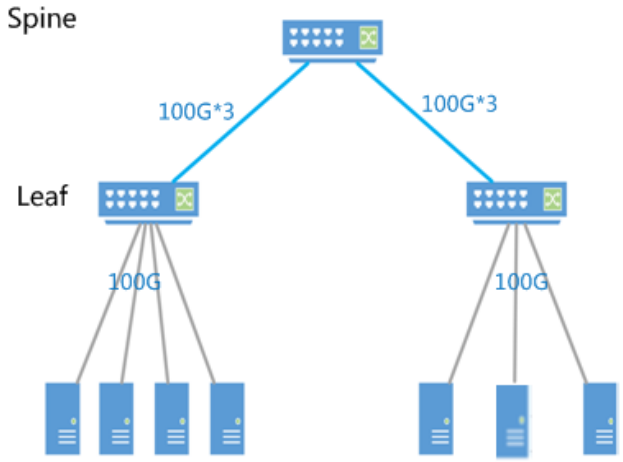


图 36：分布式存储场景测试拓扑

表8-2 分布式存储场景设备选型

厂商	设备	数量	角色	组网
Huawei-智能无损网络	CE8850-64CQ-EI	1	Spine	RoCEv2
	CE6865-48S8CQ-EI	2	Leaf	
Cisco	Nexus 92300YX	1	Spine	RoCEv2
	Nexus 93180YC-EX	2	Leaf	
	Nexus 93180YC-FX	1	Leaf	

在 CE 交换机上配置智能无损网络的 PFC、无损队列的缓存空间优化和动态 ECN 门限功能，在 Cisco 上完成 ROCE 配置。服务器的 4 个存储节点安装 NVME SSD 硬盘，部署分布式存储系统，将所有磁盘池化，计算节点安装 FIO 工具。存储资源池创建 36 个卷，3 份冗余，每个卷 20GB；每个计算节点均挂载 3 个卷。设置网卡模式为 RoCEv2，6 台计算节点运行 FIO 对存储进行读写作为背景流。另外 6 台计算节点运行 FIO 对存储进行读写作为测试流，测试参数：

- 背景流：-rwmixread=70, -rw=randwrite, -bs=128K, -iodepth=256, -runtime=3000, -numjobs=1；
- 测试流：-rwmixread=30, -bs=4K, -iodepth=1, -runtime=30, -numjobs=1, -rw=(randwrite, randread,write,read,rw, randrw)。

观察是否出现丢包，并记录 IOPS（吞吐）和时延。

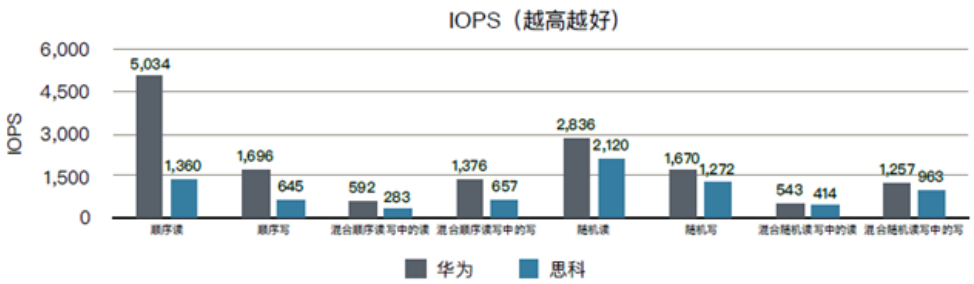


图 37：分布式存储场景 FIO 工具测试结果-IOPS

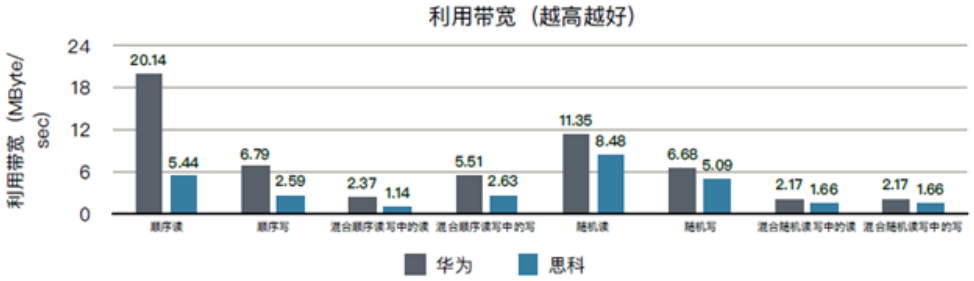


图 38：分布式存储场景 FIO 工具测试结果-利用带宽

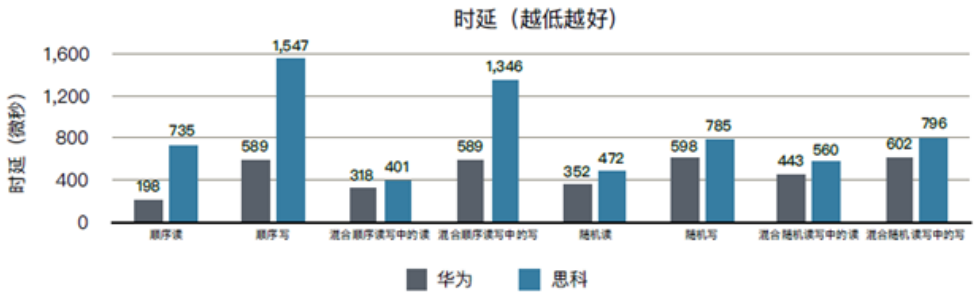


图 39：分布式存储场景 FIO 工具测试结果-时延

华为智能无损网络方案中的随机读写 IOPS 及利用带宽的性能是思科的 1.3 倍，时延是思科的 0.7；顺序读写 IOPS 及利用带宽的是思科的 2-3 倍，时延则是思科的 0.2-0.3。显然华为智能无损网络在无丢包的前提下具有更高的吞吐、带宽和更低的时延。

9 最佳实践

9.1 Atlas AI 集群

Atlas 900 是华为在全联接大会 2019 上发布的 AI 训练集群，总算力达到 256P~1024P FLOPS @FP16，相当于 50 万台 PC 的计算能力，只需 59.8 秒就可完成 ResNet-50@ImageNet 训练，排名全球第一。

Atlas 900 系统

采用 CloudEngine 系列交换机组成智能无损的全网 0 丢包以太网络。Atlas900 就是这样的 0 丢包以太网连接而成， 0 丢包的以太网为 Atlas 集群内的每一个 AI 服务器提供 8*100GE 的接入能力，从而实现百 TB 全互联无阻塞 0 丢包专属参数同步网络。

Atlas 900 系统网络联接架构示意图

10 参考资料

[1] Is Network the Botleneck of Distributed Training, Johns Hopkins University & AWS

[2] Congestion Control for Large-Scale RDMADeployments, Yibo Zhu1,3 Haggai Eran2 Daniel Firestone1 Chuanxiong Guo1 Marina Lipshteyn1Yehonatan Liron2 Jitendra Padhye1 Shachar Raindel2 Mohamad Haj Yahia2 Ming Zhang11Microsoft 2Mellanox 3U. C. Santa Barbara

[3] Characterization of MPI Usage on a Production Supercomputer, Sudheer Chunduri, Scott Parker, Pavan Balaji, Kevin Harms and Kalyan Kumaran, Argonne National Laboratory, {sudheer, sparker, balaji, kharms, kumaran}@anl.gov

[4] Huawei AI Fabric Intelligent Lossless Data Center Network Solution Performance Evaluation vs. Mellanox Switches, http://reports.tolly.com/DocDetail.aspx?DocNumber=219119

[5] Huawei AI Fabric Intelligent Lossless Data Center Network Solution Performance Evaluation vs. Cisco Nexus Switches, https://reports.tolly.com/DocDetail.aspx?DocNumber=219118