RDMA over Commodity Ethernet at Scale(Background)

Our data center network is an Ethernet-based multilayer Clos network [1, 3, 19, 31] as shown in Figure 1. Twenty to forty servers connect to a top-of-rack (ToR) switch. Tens of ToRs connect to a layer of Leaf switches. The Leaf switches in turn connect to a layer of tens to hundreds of Spine switches. Most links are 40Gb/s, and we plan to upgrade to 50GbE and 100GbE in near future [11, 25]. All switches use IP routing. The servers typically use copper cables of around 2 meters to connect to the ToR switches. The ToR switches and Leaf switches are within the distance of 10 Figure 2: How PFC works. - 20 meters, and the Leaf and Spine switches are within the distance of 200 - 300 meters. With three layers of switches, tens to hundreds of thousands of servers can be connected in a single data center. In this paper, we focus on supporting RDMA among servers under the same Spine switch layer. RoCEv2: We deployed RDMA over Converged Ethernet v2 (RoCEv2) [5] for both technical and economical reasons. RoCEv2 encapsulates an RDMA transport [5] packet within an Ethernet/IPv4/UDP packet. This makes RoCEv2 compatible with our existing networking infrastructure. The UDP header is needed for ECMP-based [34] multi-path routing. The destination UDP port is always set to 4791, while the source UDP port is randomly chosen for each queue pair (QP) [5]. The intermediate switches use standard five-tuple hashing. Thus, traffic belonging to the same QP follows the same path, while traffic on different QPs (even between the same pair of communicating end points) can follow different paths. PFC and buffer reservation: RoCEv2 uses PFC [14] to prevent buffer overflow. The PFC standard specifies 8 priority classes to reduce the head-of-line blocking problem. However, in our network, we are able to use only two of these eight priorities for RDMA. The reason is as follows. PFC is a hop-by-hop protocol between two Ethernet nodes. As show in Figure 2, the sender's egress port sends data packets to the receiver's ingress port. At the sending egress port, packets are queued in up to eight queues. Each queue maps to a priority. At the receiving ingress port, packets are buffered in corresponding ingress queues. In the shared-buffer switches used in our network, an ingress queue is implemented simply as a counter – all packets share a common buffer pool. Once the ingress queue length reaches a certain threshold (XOFF), the switch sends out a PFC pause frame to the corresponding upstream egress queue. After the egress queue receives the pause frame, it stops sending packets. A pause frame carries which priorities need to be paused and the pause duration. Once the ingress queue length falls below another threshold (XON), the switch sends a pause with zero duration to resume transmission. XOFF must be set conservatively to ensure that there is no buffer overflow, while XON needs be set to ensure that there is no buffer underflow. It takes some time for the pause frame to arrive at the upstream egress port, and for the switch to react to it. During this time, the upstream port will continue to transmit packets. Thus, the ingress port must reserve buffer space for each priority to absorb packets that arrive during this "gray period". This reserved buffer is called headroom. The size of the headroom is decided by the MTU size, the PFC reaction time of the egress port, and most importantly, the propagation delay between the sender and the receiver. The propagation delay is determined by the distance between the sender and the receiver. In our network, this can be as large as 300 meters. Given that our ToR and Leaf switches have shallow buffers (9MB or 12MB), we can only reserve enough headroom for two lossless traffic classes even though the switches support eight traffic classes. We use one lossless class for real-time traffic and the other for bulk data transfer. Need for congestion control: PFC works hop by hop. There may be several hops from the source server to the destination server. PFC pause frames propagate from the congestion point back to the source if there is persistent network congestion. This can cause problems like unfairness and victim flow [42]. In order to reduce this collateral damage, flow based congestion control mechanisms including QCN [13], DCQCN [42] and TIMELY [27] have been introduced. We use DCQCN, which uses ECN for congestion notification, in our network. We chose DCQCN because it directly reacts to the queue lengths at the intermediate switches and ECN is well supported by all the switches we use. Small queue lengths reduce the PFC generation and propagation probability. Though DCQCN helps reduce the number of PFC pause frames, it is PFC that protects packets from being dropped as the last defense. PFC poses several safety issues which are the primary focus of this paper and which we will discuss in Section 4. We believe the lessons we have learned in this paper apply to the networks using TIMELY as well. Coexistence of RDMA and TCP: In this paper, RDMA is designed for intra-DC communications. TCP is still needed for inter-DC communications and legacy applications. We use a different traffic class (which is not lossless), with reserved bandwidth, for TCP. Different traffic classes isolate TCP and RDMA traffic from each other.

我们的数据中心网络是基于以太网的多层 Clos 网络 [1,3,19,31]，如图 1 所示。二十到四十台服务器连接到架顶式 (ToR) 交换机。数十个 ToR 连接到一层 Leaf 交换机。叶子交换机又连接到数十到数百个主干交换机的层。大多数链路为 40Gb/s，我们计划在不久的将来升级到 50GbE 和 100GbE [11, 25]。所有交换机都使用 IP 路由。服务器通常使用约 2 米的铜缆连接到 ToR 交换机。 ToR 交换机和 Leaf 交换机的距离在 10 以内。 图 2：PFC 的工作原理。 - 20米，Leaf和Spine交换机的距离在200-300米以内。通过三层交换机，单个数据中心可以连接数万到数十万台服务器。在本文中，我们重点关注同一 Spine 交换机层下服务器之间的 RDMA 支持。 RoCEv2：出于技术和经济原因，我们部署了基于融合以太网 v2 的 RDMA (RoCEv2) [5]。 RoCEv2 将 RDMA 传输 [5] 数据包封装在以太网/IPv4/UDP 数据包中。这使得 RoCEv2 与我们现有的网络基础设施兼容。基于 ECMP 的 [34] 多路径路由需要 UDP 标头。目标 UDP 端口始终设置为 4791，而源 UDP 端口是为每个队列对 (QP) [5] 随机选择的。中间交换机使用标准五元组哈希。因此，属于同一 QP 的流量遵循相同的路径，而不同 QP 上的流量（甚至在同一对通信端点之间）可以遵循不同的路径。 PFC 和缓冲区预留：RoCEv2 使用 PFC [14] 来防止缓冲区溢出。 PFC 标准指定了 8 个优先级来减少队头阻塞问题。然而，在我们的网络中，我们只能将这八个优先级中的两个用于 RDMA。理由如下。 PFC 是两个以太网节点之间的逐跳协议。如图2所示，发送方的出端口将数据包发送到接收方的入端口。在发送出口端口，数据包在最多八个队列中排队。每个队列都映射到一个优先级。在接收入口端口，数据包被缓存在相应的入口队列中。在我们网络中使用的共享缓冲区交换机中，入口队列只是作为计数器实现的 - 所有数据包共享一个公共缓冲池。一旦入口队列长度达到某个阈值（XOFF），交换机就会向相应的上游出口队列发送 PFC 暂停帧。出口队列收到暂停帧后，停止发送报文。暂停帧携带需要暂停哪些优先级以及暂停时长。一旦入口队列长度低于另一个阈值 (XON)，交换机就会发送持续时间为零的暂停来恢复传输。 XOFF 必须保守设置，以确保不发生缓冲区溢出，而 XON 则需要设置以确保不发生缓冲区下溢。暂停帧到达上游出口端口以及交换机对其做出反应需要一些时间。在此期间，上游端口将继续传输数据包。因此，入口端口必须为每个优先级预留缓冲空间，以吸收在此"灰色时期"到达的数据包。这个保留的缓冲区称为净空。净空大小由 MTU 大小、出口端口的 PFC 反应时间以及最重要的发送器和接收器之间的传播延迟决定。传播延迟由发送器和接收器之间的距离决定。在我们的网络中，长度可达 300 米。鉴于我们的 ToR 和 Leaf 交换机具有浅缓冲区（9MB 或 12MB），即使交换机支持八个流量类别，我们也只能为两个无损流量类别保留足够的空间。我们将一种无损类别用于实时流量，另一种用于批量数据传输。拥塞控制的需要：PFC逐跳工作。从源服务器到目标服务器可能有多个跃点。如果存在持续的网络拥塞，PFC 暂停帧会从拥塞点传播回源。这可能会导致不公平和受害者流等问题[42]。为了减少这种附带损害，引入了基于流的拥塞控制机制，包括 QCN [13]、DCQCN [42] 和 TIMELY [27]。我们在网络中使用 DCQCN，它使用 ECN 进行拥塞通知。我们选择 DCQCN 是因为它直接对中

> 间交换机的队列长度做出反应，并且我们使用的所有交换机都很好地支持 ECN。较小的队列长度会降低 PFC 的生成和传播概率。虽然 DCQCN 有助于减少 PFC 暂停帧的数量，但 PFC 是保护数据包不被丢弃的最后一道防线。 PFC 提出了几个安全问题，这些问题是本文的主要焦点，我们将在第 4 节中讨论这些问题。我们相信，我们在本文中学到的经验教训也适用于使用 TIMELY 的网络。 RDMA和TCP的共存：在本文中，RDMA是为DC内通信而设计的。 DC 间通信和遗留应用程序仍然需要 TCP。我们对 TCP 使用不同的流量类别（非无损）并保留带宽。不同的流量类别将 TCP 和 RDMA 流量相互隔离。

Ethernet packet

计算机网络在传输数据时，为了保证所有共享网络资源的计算机都能公平、迅速地使用网络，通常把数据分割成若干小块作为传输单位进行发送，这样的传输单位我们通常称之为包，也叫“数据包”。以太网数据包有四种分类

IPv4 packet

IPv4（Internet Protocol version 4）是互联网协议中的一种，它是互联网上数据通信的基础协议之一。IPv4 协议的工作原理主要包括以下几个方面：

1. IP 地址的分配和分类：IPv4 使用 32 位二进制数表示 IP 地址，根据网络的规模和需求，IP 地址被分为 A 类、B 类、C 类等不同的分类，用于标识不同的网络和主机。
2. IP 数据包的封装与传输：IPv4 将数据分割成一系列的数据包（也称为 IP 数据报），每个数据包包含一个 IP 报头和一个数据部分。IP 报头包含源 IP 地址、目标 IP 地址、协议类型、数据包长度等重要信息，用于在网络中传输和路由数据包。
3. IP 地址的解析和路由选择：当一个 IP 数据包从源主机发出时，它首先被传输到网络中的第一个路由器，路由器根据目标 IP 地址和路由表的信息，选择下一个路由器或直接将数据包传输到目标主机。这个过程被称为 IP 地址解析和路由选择。
4. IP 地址的转换和映射：为了解决 IP 地址不足和网络隔离等问题，IPv4 引入了一些特殊的地址，如子网掩码、网络地址转换（NAT）、端口转发等技术，用于实现 IP 地址的转换和映射。

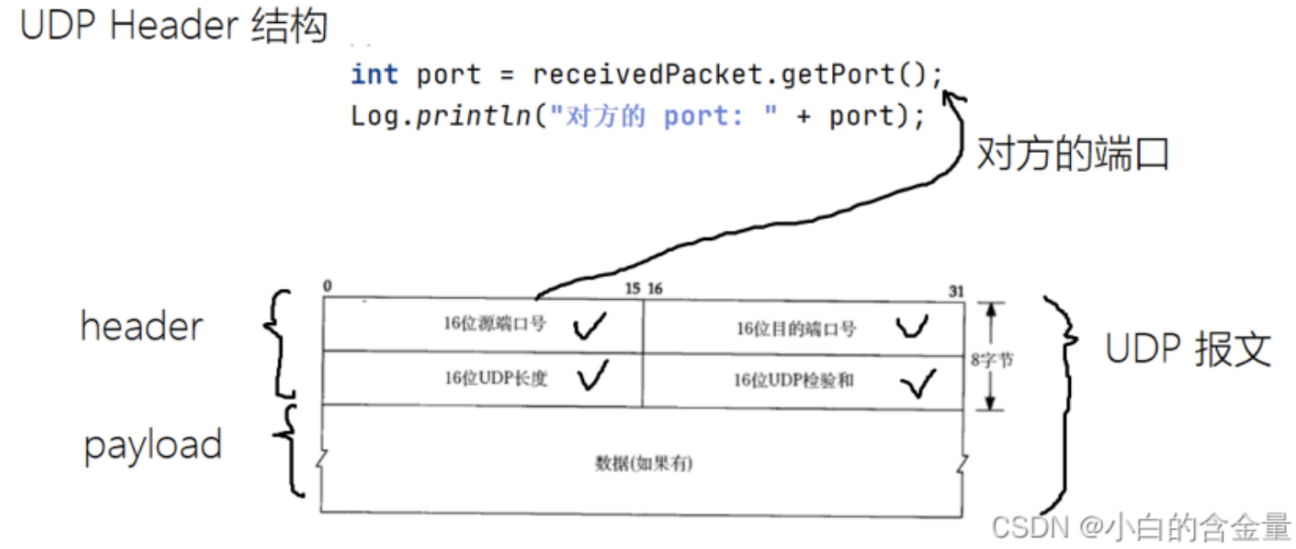
总的来说，IPv4 协议的工作原理是通过分配和解析 IP 地址来标识不同的网络和主机，并将数据分割成数据包进行传输和路由选择。在传输过程中，IPv4 还支持地址转换和映射等技术，以解决一些实际问题。

UDP packet

显而易见

UDP header

UDP的header放了双方的IP和端口，长度8个字节
图中为整个报文结构，UDP长度就是报文长度
payload就是除去header剩下的报文长度

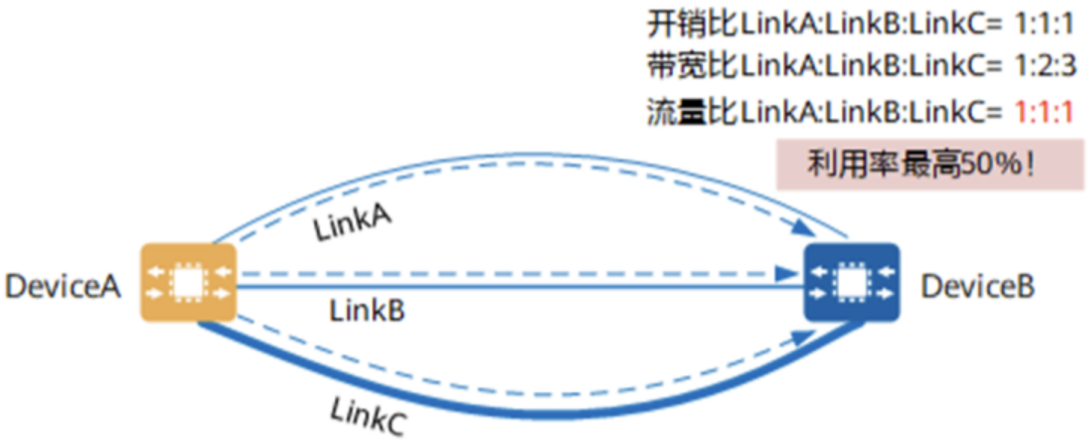


ECMP-based multi-path routing（ 多路径路由技术 ）

https://forum.huawei.com/enterprise/zh/thread/659558832164192256

ECMP是一种路由技术，实现将数据流量包在不同的路径上传输，流量在这些链路上是平均分配的，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。ECMP最大的特点是实现了等值情况下，多路径负载均衡和链路备份的目的，在静态路由和OSPF中基本上都支持ECMP功能。

ECMP的缺点是在路径间带宽差异大时，带宽利用率低。例如，流量在三条路径上负载分担，其中路径的带宽分别是10Mbps、20Mbps和30Mbps，如果部署ECMP，则总带宽只能达到30Mbps，利用率最多只能到50%。



the MTU size

最大传输单元（Maximum Transmission Unit，MTU）用来通知对方所能接受数据服务单元的最大尺寸，说明发送方能够接受的有效载荷大小。[1]

是包或帧的最大长度，一般以字节记。如果MTU过大，在碰到路由器时会被拒绝转发，因为它不能处理过大的包。如果太小，因为协议一定要在包(或帧)上加上包头，那实际传送的数据量就会过小，这样也划不来。大部分操作系统会提供给用户一个默认值，该值一般对用户是比较合适的。

QCN

https://zhuanlan.zhihu.com/p/643007675?utm_id=0

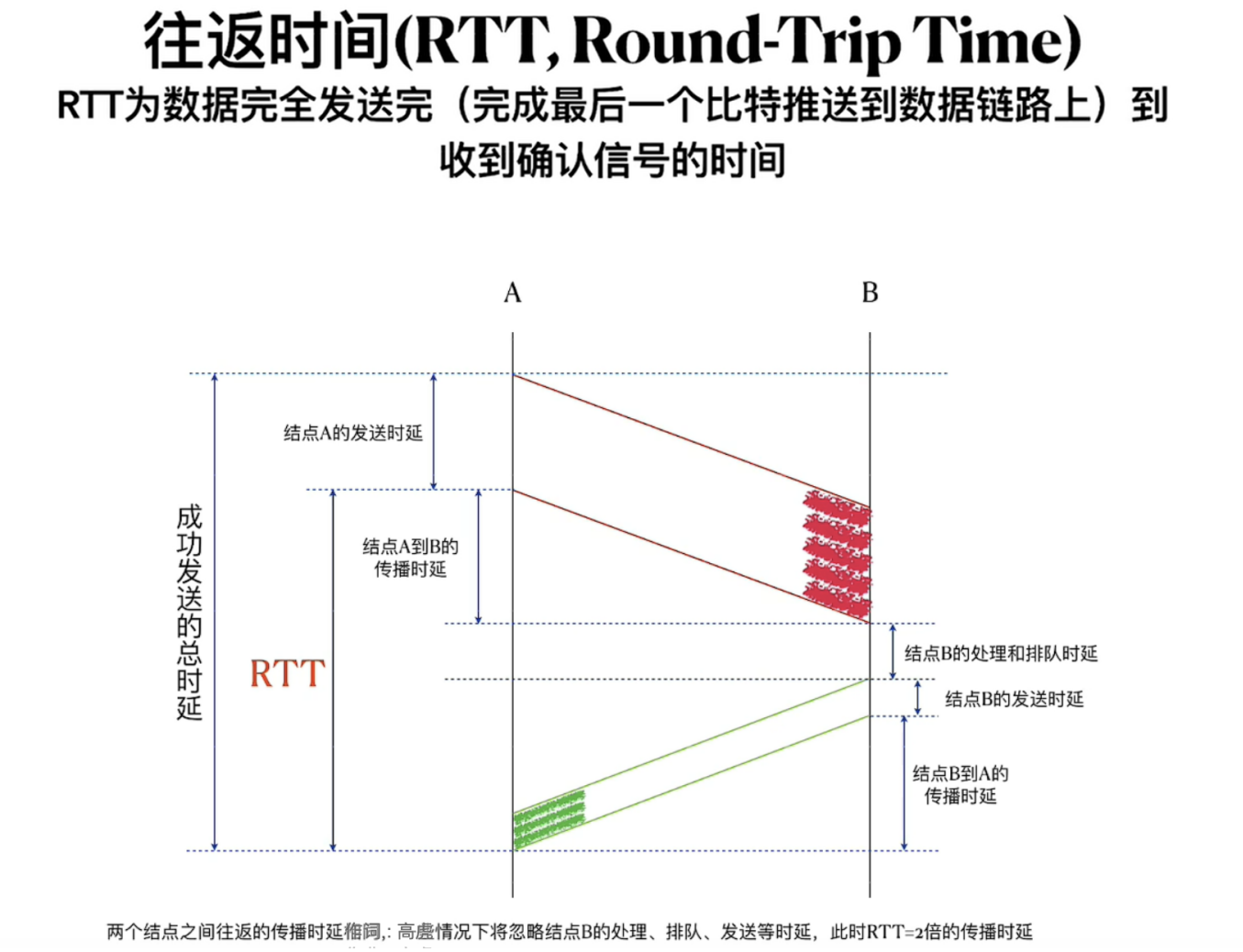
DCQCN

https://zhuanlan.zhihu.com/p/643007675?utm_id=0

TIMELY

https://zhuanlan.zhihu.com/p/570169667

RTT（Round-Trip Time）



ECN

https://zhuanlan.zhihu.com/p/643007675?utm_id=0

intra-DC communications

inter-DC communications