7. RELATED WORK

> This paper focuses on how to safely deploy RoCEv2 on a large-scale. Besides RoCEv2, there are two other RDMA technologies: Infiniband [6] and iWarp [30]. Infiniband is a complete networking protocol suite, which has its own layer-1 to layer-7 protocols. An Infiniband network is composed of multiple Infiniband subnets which are connected via Infiniband routers. Within a subnet, servers are connected via Infiniband switches. However, to the best of our knowledge, there are still no Infiniband routers in production. Hence Infiniband does not meet our scalability requirement. Furthermore, Infiniband is not compatible with Ethernet, which is the dominant networking technology for data centers. iWarp runs RDMA over TCP/IP. The TCP/IP protocol is offloaded to the NIC. Since TCP guarantees reliable delivery even if some of the packets are dropped, iWarp does not necessarily need a lossless network. iWarp has one advantage over RoCE in that it can be use for inter-DC communications. But since iWarp uses TCP for packet transmission, it faces the same issue of TCP: long latency caused by packet drops and retransmission timeout. As we have discussed in 6.3, we expect new transport protocols different from the Infiniband transport and TCP to be introduced in the future driven by new hardware innovations. Deadlock is well studied in the literature and it is well known cyclic buffer dependency is necessary for deadlock [12, 18, 22, 33, 36]. Due to the specific Clos network topology, we once thought our network should be free from deadlock since it should be free of cyclic buffer dependency. But the 'conspiracy' of PFC and Ethernet packet flooding has shown that deadlock can happen in Clos networks. TCP performance issues such as TCP incast [35, 38, 39] and long latency tail [41] have been studied extensively. These solutions are still within the existing TCP framework. They either tune the retransmission timer (as in [35]), or control the TCP receiving window ([39]), or tune the ECN parameter ([38]). RDMA provides a different approach.Compared to [41] which still uses TCP, RDMA bypasses the OS kernel, so that the latency introduced by the kernel is eliminated. Our work shows that RDMA can be safely deployed at scale for intra-DC communications. As we have shown in Figure 5.4, RDMA greatly reduces the high percentile latency compared with TCP. RDMA has been used to build systems including storage, key-value stores, and distributed transaction systems [17, 26, 28, 37]. Most of these systems use Infiniband or RoCE with tens of servers. In this paper, we have shown that we can scale RDMA to much larger networks using RoCEv2. Hence much larger in-memory systems can be built in the future.

本文重点讨论如何大规模安全部署RoCEv2。除了 RoCEv2 之外，还有另外两种 RDMA 技术：Infiniband [6] 和 iWarp [30]。

1. Infiniband是一个完整的网络协议套件，拥有自己的第1层到第7层协议。 Infiniband 网络由多个 Infiniband 子网组成，这些子网通过 Infiniband 路由器连接。在子网内，服务器通过 Infiniband 交换机连接。
2. 然而，据我们所知，目前还没有 Infiniband 路由器投入生产。因此Infiniband不能满足我们的可扩展性要求。此外，Infiniband 与以太网不兼容，而以太网是数据中心的主导网络技术。
3. iWarp 通过 TCP/IP 运行 RDMA。 TCP/IP 协议被卸载到 NIC。**由于即使某些数据包丢失，TCP 也能保证可靠传输，因此 iWarp 不一定需要无损网络**。
4. iWarp 相对于 RoCE 的一个优势在于它可以用于 DC 间通信。但由于iWarp使用TCP进行数据包传输，因此它面临着与TCP相同的问题：丢包和重传超时导致的长时间延迟。
5. 正如我们在 6.3 中讨论的，我们预计未来将在新硬件创新的推动下引入不同于 Infiniband 传输和 TCP 的新传输协议。死锁在文献中得到了充分研究，众所周知，循环缓冲区依赖性对于死锁是必要的[12,18,22,33,36]。由于特定的 Clos 网络拓扑，我们曾经认为我们的网络应该没有死锁，因为它应该没有循环缓冲区依赖。但 PFC 和以太网数据包泛洪的"阴谋现象"表明，Clos 网络中可能会发生死锁。
6. TCP 性能问题，例如 TCP incast [35,38,39] 和长延迟尾部 [41] 已被广泛研究。这些解决方案仍然在现有的 TCP 框架内。它们要么调整重传定时器（如[35]），要么控制TCP接收窗口（[39]），要么调整ECN参数（[38]）。
7. RDMA提供了一种不同的方法。与仍然使用TCP的[41]相比，RDMA绕过了操作系统内核，从而消除了内核引入的延迟。我们的工作表明，RDMA 可以安全地大规模部署用于 DC 内通信。如图 5.4 所示，与 TCP 相比，RDMA 大大降低了高百分比延迟。
   RDMA 已用于构建包括存储、键值存储和分布式事务系统在内的系统 [17,26,28,37]。这些系统大多数都使用 Infiniband 或 RoCE 以及数十台服务器。在本文中，我们展示了可以使用 RoCEv2 将 RDMA 扩展到更大的网络。因此，未来可以构建更大的内存系统。

-什么是Infiniband?

InfiniBand（英特尔 InfiniBand）是一种高性能、低延迟的计算和存储网络技术，最初由一家叫做 InfiniBand Trade Association (IBTA) 的行业联盟制定的标准。InfiniBand 技术旨在支持高性能计算和数据中心应用，提供*低延迟*、*高带宽和可扩展性*，特别适用于高性能计算集群、超级计算机、存储系统和云计算环境。

以下是 InfiniBand 技术的一些关键特点和组成部分：

1. 高带宽和低延迟：InfiniBand 提供了高带宽和低延迟的数据传输，适用于需要快速数据交换的应用，如科学计算、金融交易、大数据分析等。
2. 双向通信：InfiniBand 支持双向通信，使设备能够同时发送和接收数据，从而提高了通信效率。
3. 硬件卸载：InfiniBand 硬件卸载了一些网络任务，如数据包处理、路由和流量管理，减轻了主机 CPU 的负担，提高了性能。
4. 灵活的拓扑：InfiniBand 支持多种拓扑，包括点对点、星型、环形和多层拓扑，可根据应用需求进行选择。
5. RDMA 支持：InfiniBand 支持远程直接内存访问（RDMA），允许主机之间直接访问彼此的内存，而无需中央处理单元（CPU）的干预，从而提高数据传输效率。
6. 可靠性和故障恢复：InfiniBand 提供多种机制来确保数据的可靠传输和网络的高可用性，包括====错误检测和校正（EDC）以及路径故障恢复。
7. 技术标准：InfiniBand 技术是基于一系列标准的，由 InfiniBand Trade Association 制定和管理，这有助于确保不同供应商的设备和系统之间的互操作性。

InfiniBand 技术广泛应用于高性能计算领域，例如超级计算机、科学研究、天气模拟和仿真等。此外，它还在数据中心中用于构建高性能存储和网络基础设施，以满足大规模数据传输和处理的需求。虽然以太网也在数据中心中起到重要作用，但在某些要求极高性能的应用中，InfiniBand 技术提供了一个强大的选择。

-为什么目前还没有 Infiniband 路由器投入生产?

InfiniBand 路由器的市场份额相对有限，主要原因包括以下几点：

1. 市场需求：InfiniBand 技术主要用于高性能计算、数据中心和科学应用，这些领域对低延迟、高带宽和高性能的网络要求较高。然而，这些市场是相对狭窄的，与广泛部署的以太网相比，市场规模较小。
2. 成本：InfiniBand 基础设施的成本通常较高，包括网卡、交换机和相关设备。这对于一些组织来说可能不划算，尤其是在较小规模的网络中。
3. 以太网的主导地位：以太网是目前数据中心和企业网络中的主导网络技术，因为它是一种广泛部署的标准，并且在不同规模的网络中表现良好。许多组织更倾向于使用以太网，因为它的成本相对较低，而且有更大的供应商生态系统和更多的支持。
4. 软件支持：InfiniBand 技术需要相应的软件支持，包括操作系统和应用程序的支持。虽然一些高性能计算和科研应用程序支持 InfiniBand，但并非所有应用程序都能够充分利用这种技术。

-为什么Infiniband 与以太网不兼容?

InfiniBand 和以太网是两种不同的网络技术，它们在*物理层、数据链路层和协议层面*都有显著的差异，因此它们不兼容的主要原因如下：

1. 物理层差异：InfiniBand 使用了不同的物理层标准和电信号传输机制，通常基于双绞线、铜缆或光纤连接，而以太网则使用以太网电缆和光纤标准。这两者的物理层特性和接口不同，因此不兼容。
2. 数据链路层差异：InfiniBand 和以太网使用不同的数据链路层协议，分别称为InfiniBand数据链路层协议和以太网数据链路层协议（通常是Ethernet的各种形式）。这些协议定义了数据包的格式、帧结构和处理规则，它们之间也不兼容。
3. 路由和交换：InfiniBand 和以太网采用不同的路由和交换方式，它们的网络拓扑和路由算法不同，因此不同类型的路由器和交换机用于支持它们的网络。
4. 协议栈差异：InfiniBand和以太网在协议栈的各个层面都有不同的协议，包括网络层、传输层和应用层。这些协议用于管理连接、数据传输和通信，它们也不互通。
   虽然 InfiniBand 和以太网是两种不兼容的网络技术，但有些设备和技术可以用于在它们之间建立桥接或互操作。例如，InfiniBand 网络可以使用网关设备将数据传输到以太网网络，从而实现两者之间的通信。但这种桥接通常需要专门的硬件和协议转换，因此并不是完全透明的互操作。
   总的来说，InfiniBand 和以太网是两种在不同用例和场景中使用的网络技术，它们各自具有自身的特点和优势。用户在选择网络技术时需要考虑应用需求、性能、可扩展性和成本等因素，以确定哪种技术最适合其特定的用途。

-iwarp是什么?

iWARP（Internet Wide Area RDMA Protocol）是一种网络协议，它建立在以太网上，并允许远程直接内存访问（RDMA）在广域网（WAN）和互联网上进行高性能数据传输。
iWARP 的设计目标是在不需要特殊硬件支持的情况下，在标准以太网基础设施上提供低延迟和高带宽的 RDMA。

以下是 iWARP 协议的一些关键特点和特性：

1. RDMA 支持：iWARP 允许应用程序在网络中直接访问远程主机的内存，而无需中央处理单元（CPU）的干预。这提供了低延迟和高吞吐量的数据传输。
2. 基于标准以太网：iWARP 协议建立在标准以太网网络基础设施上，而不需要特殊的硬件或网络拓扑。这降低了部署成本和复杂性。
3. 安全性：iWARP 协议支持数据加密和认证，以确保数据在传输过程中的机密性和完整性。
4. 应用领域：iWARP 技术通常用于需要高性能数据传输的应用领域，如存储区域网络（SAN）、云计算、高性能计算（HPC）和远程备份等。
5. 协议堆栈：iWARP 协议通常建立在传输控制协议（TCP）之上，从而利用 TCP 的可靠性和流控制功能。它还可以使用一些协议来提供 RDMA 功能，如 Internet Engineering Task Force（IETF）定义的 RDMA Protocol Extensions（RPE）。
   iWARP 技术是一种在标准以太网上实现高性能 RDMA 的方式，它允许应用程序利用 RDMA 的优势来加速数据传输，而无需特殊硬件或网络设备。这对于需要在广域网或互联网上进行数据传输的应用非常有用，因为它提供了低延迟和高带宽的选项。

-iwarp为什么可以在DC之间通信

iWARP（Internet Wide Area RDMA Protocol）是一种可用于在数据中心（DC）之间进行高性能通信的协议。它能够在不同数据中心的计算机之间实现高效的远程直接内存访问（RDMA），这意味着可以通过网络直接访问远程计算机的内存而不涉及主机CPU的数据传输。

iWARP 之所以可以在不同数据中心之间通信，是因为它**建立在标准的以太网技术**上，并**使用标准的TCP/IP协议栈**。它具有以下特点和优势：

1. RDMA功能：iWARP利用RDMA技术，使数据传输更加高效。它允许数据中心内的计算机可以直接访问远程计算机的内存，无需主机CPU的干预。这降低了通信的延迟并提高了吞吐量。
2. 基于以太网：iWARP使用标准以太网硬件，因此可以轻松集成到数据中心网络中，而无需昂贵的专用硬件。

3. TCP/IP支持：iWARP在TCP/IP协议栈上运行，这使得它与现有的网络基础设施兼容，包括路由器、交换机和防火墙。这意味着可以在不同数据中心之间建立连接，并使用TCP/IP协议进行通信。
4. 安全性：iWARP可以利用现有的网络安全机制，如IPsec，以确保通信的机密性和完整性。
   总之，iWARP是一种在不同数据中心之间通信的有效方式，因为它融合了高性能的RDMA技术和标准的以太网和TCP/IP技术，同时保持了网络的互操作性和安全性。这使得它成为处理大规模数据中心之间通信需求的一种有吸引力的解决方案。

-DC之间与DC内部常用的通信协议

在数据中心（DC）内部和数据中心之间，通信协议的选择取决于具体的应用和需求。以下是一些常见的通信协议，它们在数据中心内部和数据中心之间都有广泛的应用：

（1）在数据中心内部常用的通信协议：

1. Ethernet（以太网）：以太网是数据中心内部通信的主要协议，用于连接服务器、交换机、路由器和其他网络设备。不同速度的以太网如1 GbE、10 GbE、25 GbE、40 GbE、100 GbE等都用于数据中心内部的连接。
2. TCP/IP：TCP/IP协议套件是互联网通信的基础，也在数据中心内部广泛使用。它包括TCP（传输控制协议）和IP（互联网协议），用于可靠的数据传输和路由。
3. InfiniBand：InfiniBand是一种高性能网络技术，特别适用于数据中心内部的高性能计算和存储集群。它支持低延迟和高吞吐量通信，并广泛用于高性能计算 (HPC) 和大规模数据中心。
4. RDMA（Remote Direct Memory Access）：RDMA协议允许远程计算机直接访问另一台计算机的内存，从而提供了低延迟和高性能的通信。它可以用于不同的传输协议，如InfiniBand、iWARP等。

（2）在数据中心之间常用的通信协议：

1. BGP（Border Gateway Protocol）：BGP是用于在不同数据中心之间进行路由和互联的标准协议。它在全球互联网中广泛使用，用于确定最佳的数据包传输路径。
2. MPLS（Multiprotocol Label Switching）：MPLS是一种用于创建虚拟专用网络（VPN）和流量工程的协议。它常用于跨多个数据中心之间建立虚拟专用网络连接。
3. VXLAN（Virtual Extensible LAN）：VXLAN是一种用于创建虚拟局域网（VLAN）的技术，允许数据中心之间的虚拟机通信。它通常用于云计算环境中。
4. OTN（Optical Transport Network）：OTN是一种用于数据中心之间长距离高速光通信的协议，适用于数据中心互联和数据中心与远程站点之间的通信。
   这些是一些常见的数据中心内部和数据中心之间通信协议。具体的选择取决于数据中心的架构、性能需求、可用技术和成本等因素。不同数据中心可能会采用不同的组合来满足其特定需求。

（3）TCP/IP协议可以用于数据中心之间通信，也可以用于DC内部的通信：
实际上，TCP/IP协议是互联网和大多数企业网络中通信的基础协议。它也在数据中心之间广泛使用，尤其是用于跨不同数据中心的通信。
数据中心之间的通信通常需要以下几个要求：

1. 可靠性：通信必须是可靠的，以确保数据的完整性和可用性。
2. 安全性：通信应该是安全的，以保护敏感数据不受未经授权的访问。
3. 可扩展性：通信需要支持数据中心的扩展和增加带宽容量。
4. 路由：通信必须能够在不同数据中心之间进行路由，以确保数据包能够正确到达目的地。
   TCP/IP协议套件，包括TCP（传输控制协议）和IP（互联网协议），可以满足这些要求。它提供可靠的数据传输（通过TCP），同时通过IP支持数据包的路由。此外，数据中心通常会使用附加的安全性层，如VPN（虚拟专用网络）或TLS（传输层安全性），以提供加密和认证，从而确保通信的安全性。
   虽然有一些专用的数据中心通信协议，如InfiniBand、RoCE（RDMA over Converged Ethernet）和iWARP，它们可以提供更高性能和低延迟的通信，但TCP/IP仍然是广泛使用的通用协议，尤其是在需要跨不同数据中心进行通信时。因此，TCP/IP在数据中心之间通信中有广泛的应用。

-RoCE为什么不可以在DC之间通信?

（1）RoCE是什么：

RoCE (RDMA over Converged Ethernet) 是一种网络协议，它允许在以太网上实现远程直接内存访问 (RDMA)。RDMA 是一种高性能的数据传输技术，允许计算机系统在没有主机 CPU 的干预下直接从内存中读取或写入数据。RoCE 将这种高性能的数据传输能力扩展到以太网环境，提供了低延迟和高吞吐量的数据传输选项。

以下是 RoCE 的主要特点和组成部分：

1. RDMA 支持：*RoCE 协议支持 RDMA，这意味着网络中的两个节点可以直接相互访问其内存，而无需将数据包传递到主机 CPU 进行处理*。这降低了数据传输的延迟和 CPU 开销，从而提高了性能。

2. 以太网基础：RoCE 利用以太网网络基础，这意味着它*可以在现有的以太网基础设施上运行，而不需要特殊的硬件或网络拓扑*。这降低了部署成本和复杂性。
3. 高性能：RoCE 提供低延迟和高吞吐量的数据传输，适用于需要大数据传输或实时数据处理的应用，如数据中心、高性能计算 (HPC)、存储系统等。
4. 安全性：RoCE 协议支持数据加密和认证，以确保数据在传输过程中的机密性和完整性。
   RoCE 可以被视为一种在以太网上实现 RDMA 的标准，它在高性能计算和大规模数据中心环境中得到广泛应用。在使用 RoCE 的网络中，通常需要支持 RoCE 的适配器卡（网卡）和交换机，以确保协议的正常运行。这使得 RoCE 成为支持高性能计算和数据传输的重要技术之一。

（2）RoCE为什么不可以在DC之间通信?

RoCE（RDMA over Converged Ethernet）协议通常用于数据中心内部通信，而不太适用于数据中心之间的通信，有以下原因：

1. 范围问题：RoCE 是一种基于以太网的协议，通常用于局域网或数据中心内部网络。在数据中心之间的通信通常涉及更广泛的范围，涵盖跨足够大的距离，可能涉及多个数据中心或甚至多个地理位置。以太网通常不适用于长距离通信，因为它在物理层面上有距离限制。
2. 延迟问题：RoCE 通常设计用于提供低延迟和高性能的数据中心内部通信。在跨足够远的距离内，延迟会变得更高，因为数据需要在不同的路由器、交换机和链路之间传输。这不符合 RoCE 的设计目标。
3. 需要专用硬件：RoCE 需要特殊的网卡和网络设备来支持。在数据中心之间的通信中，往往需要考虑多种不同类型的网络和设备，而不是仅仅使用以太网。RoCE 需要特定的硬件支持，因此可能不适用于所有类型的网络。
   为了实现数据中心之间的通信，通常会使用其他协议和技术，例如 IP（Internet Protocol）以及基于 TCP/IP 的协议，以适应更广泛的范围和网络类型。这些协议可以在较长的距离上提供通信，并且通常适用于不同类型的硬件和网络设备。
   因此，RoCE 主要用于数据中心内部通信，而不适合跨足够远的距离或不同类型的网络的数据中心之间的通信。

（3）为什么以太网通常不适用于长距离通信？为什么它在物理层面上有距离限制?

以太网通常不适用于长距离通信，因为它在物理层面上存在距离限制，这是由于以太网的电信号衰减和时序问题引起的。以下是导致以太网有距离限制的主要原因：

1. 信号衰减：以太网使用电信号来传输数据，而电信号在传输过程中会受到信号衰减的影响，即信号的强度会随着传输距离的增加而减弱。*随着距离的增加，电信号变得越来越弱，到达接收端时可能无法可靠解码*。这限制了以太网的传输距离。
2. 时序问题：以太网在物理层上使用时分多路复用（Time Division Multiplexing，TDM）来管理数据传输。每个以太网标准都规定了特定的时钟频率和时间间隔，以确定如何传输数据帧。在较长的距离上，时序问题可能会导致数据帧到达接收端时出现时钟漂移或数据损坏。
3. 延迟和抖动：**以太网通常设计用于局域网或数据中心内部通信，其中延迟较低且稳定性很高**。在长距离通信中，信号必须通过更多的中继设备和传输链路，从而增加了延迟和时钟抖动。这可能不适合某些实时应用或需要高性能的应用。
   为了克服以太网的距离限制，通常会采用其他传输媒体和协议，如光纤、长距离以太网（Long Range Ethernet，LRE）、光纤通道（Fibre Channel）等。这些技术可以支持更长的通信距离，并提供更好的性能和稳定性，适用于广域网（WAN）和跨足够远的距离的通信需求。不同的以太网标准和物理媒体也具有不同的传输距离限制，因此在设计网络时需要考虑具体的要求和技术。

（4）以太网的时钟漂移现象是什么?

以太网的时钟漂移是指*在网络中使用多个设备和链路时，由于每个设备的时钟或时序参考存在微小的不匹配，导致数据传输的时钟不同步*。这种时钟漂移可能会导致数据帧的到达时间和频率出现微小的不一致，从而影响网络性能和数据可靠性。

时钟漂移的主要原因包括以下几点：

1. 晶振不稳定性：网络设备通常使用晶振来产生时钟信号，但不同的晶振可能存在微小的频率不匹配和波动。这些不匹配可能会导致设备的时钟信号存在微小的漂移，这就是所谓的时钟漂移。
2. 数据链路长度：在以太网中，数据通过电缆或光纤等链路传输，不同链路的长度可能会略有不同，这也会导致数据的传输速率略有不同，从而引发时钟漂移。
3. 信号传输延迟：数据在不同的链路上传输需要一定的时间，这导致信号的延迟存在微小差异，从而影响到数据到达的时序。

时钟漂移通常在高性能网络中需要特别注意，因为在需要高精度和同步的应用中，如音视频传输、金融交易和科学实验等领域，时钟不同步可能会导致数据帧的丢失、重复或乱序，从而影响系统性能和稳定性。

为了解决时钟漂移问题，网络设备通常采用时钟同步技术，如网络时间协议（Network Time Protocol，NTP）或精确时间协议（Precision Time Protocol，PTP）。这些协议允许设备校准其时钟以匹配网络中的统一时钟，从而减小时钟漂移。同时，高性能网络设备通常具有更准确的时钟源和更高精度的时钟同步机制，以减小时钟漂移对系统的影响。

-TCP incast [35,38,39] 和长延迟尾部 [41]是什么?

TCP incast 现象和长延迟尾部现象是与 TCP 协议相关的两种不同网络现象，它们都可以影响网络性能和吞吐量。

1. TCP Incast 现象（TCP Incast Congestion）：
   - TCP Incast 现象**通常发生在数据中心网络中，特别是在集群计算或云计算环境**中。

- 当*多个客户端同时请求从相同的服务器或存储设备获取数据*时，数据中心*交换机或路由器可能会面临来自多个客户端的并发数据请求*。
- 如果这些请求同时到达服务器，服务器需要为每个请求生成响应，并将数据发送回客户端。在短时间内，服务器可能会面临大量的并发请求。
- 当这些响应返回到网络时，它们可能在网络交换机或路由器的缓冲区中排队等待传输。如果缓冲区容量有限，数据包可能会被阻塞，从而导致*延迟增加，吞吐量下降，甚至丢失数据包*。
- TCP Incast 现象的解决方法通常包括优化网络配置、增加缓冲区容量、改进调度算法等。

2. 长延迟尾部现象（Long Tail Latency）：
   - 长延迟尾部现象指的是网络中存在一小部分数据包的传输延迟明显超过大多数数据包的现象。
   - 这种现象通常发生在拥塞或高负载网络中，其中大多数数据包能够以低延迟传输，但一些数据包可能因为竞争资源或拥塞而经历明显更长的传输延迟。
   - 长延迟尾部现象可能对实时应用或需要低延迟的应用产生负面影响，因为它会导致某些数据包的到达时间不稳定，从而降低了系统的性能和可预测性。
   - 解决长延迟尾部现象的方法可能包括使用更好的拥塞控制算法、改进网络配置、使用服务质量（Quality of Service，QoS）策略等。

这两种现象都是在特定网络条件下出现的，而且通常需要特定的网络设计和调优来减轻它们的影响。理解 TCP Incast 现象和长延迟尾部现象对于高性能计算、云计算和数据中心网络的管理和性能优化至关重要。

- 长延迟尾部现象中，小部分数据包的延迟明显高于大多数数据包，这是因为网络中可能存在拥塞、竞争资源或其他因素导致部分数据包在传输过程中遇到了特殊情况，使它们的延迟显著增加。这些特殊情况可能包括：

1. 竞争资源：在拥塞的网络中，多个数据包可能竞争相同的网络资源，如带宽或路由器缓冲区。一些数据包可能在竞争中获胜，能够更快地传输，而其他数据包可能需要等待，从而导致它们的延迟增加。
2. 拥塞：在网络中的某些部分可能会出现拥塞，即网络资源不足以满足所有传输请求。这会导致数据包排队等待传输，导致延迟增加。在长延迟尾部现象中，某些数据包可能更容易受到拥塞的影响，因此它们的延迟更高。
3. 传输路由：不同的数据包可能采用不同的路由路径，这可能导致一些数据包经历更多的中继设备或网络拓扑中的拓扑变化，从而增加它们的延迟。
4. 拥挤窗口：TCP 协议中的拥塞窗口大小可能会变化，取决于网络状况和拥塞控制算法。*一些数据包的拥塞窗口可能较小，导致它们在传输时采用较慢的速率，从而增加延迟*。
   虽然大多数数据包在正常情况下具有较低的延迟，但长延迟尾部现象是由于一些数据包在特殊情况下经历了高延迟，而这些情况通常是短暂的、随机的或临时的。这些特殊情况可以导致延迟分布的尾部拉长，使少数数据包的延迟显著增加，而大多数数据包的延迟仍保持在较低水平。这种现象对于需要低延迟和可预测性的应用可能具有负面影响，因此网络管理和优化通常需要采取措施来减轻长延迟尾部现象的影响。

-iwarp还需要tcp辅助，那这项技术不是很鸡肋吗?

> 不鸡肋！把上面查阅的详解全部阅读一遍就知道了