

issue_thinking

【1】PCP

ask: 是什么?

Ans:

PCP 是 Priority Code Point 的缩写, 它是用于以太网帧 (Ethernet frames) 标头中的一个字段, 用于定义数据包的优先级。具体来说, PCP 是 IEEE 802.1Q 标准中定义的一个字段, 用于支持虚拟局域网 (VLAN) 和服务质量 (Quality of Service, QoS) 功能。

1. PCP 是一个 3 位二进制字段, 可以表示 0 到 7 的值, 其中 0 表示最低优先级, 而 7 表示最高优先级。这个字段用于帮助网络设备 (例如交换机和路由器) 确定如何处理经过网络的数据包, 以确保在拥挤的网络环境中能够满足不同类型数据流的服务质量需求。
2. PCP 值的具体含义和如何影响数据包处理取决于网络中所使用的 QoS 策略和设备的配置。
3. 通常, 较高的 PCP 值表示更高的优先级, 网络设备可能会优先处理这些数据包, 以确保它们获得更好的服务质量。这对于实现流量控制和不同应用之间的优先级区分非常重要, 以确保网络资源得到有效分配。

【2】“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” 在我们网络中使用的共享缓冲区交换机中, 入口队列仅作为计数器实现 - 所有数据包共享一个公共缓冲池

Ask: 为什么?

Ans:

这句话意思是, 在我们的网络中使用的共享缓冲交换机中, 入口队列被实现为一个简单的计数器, 所有数据包共享一个公共缓冲池。

具体来说, 这句话描述了网络交换机中数据包的处理方式。在这种情况下, 所有到达的数据包都被放置在同一个缓冲池中, 而不是为每个数据包创建单独的缓冲区。每个数据包都与一个计数器相关联, 用于跟踪缓冲池中的可用缓冲区数量。当数据包到达时, 它们会占用或释放缓冲池中的一些缓冲区。

这种方式的优点是简化了入口队列的管理, 因为所有数据包共享相同的资源, 但缺点是如果流量拥塞或过多, 可能会导致数据包排队等待处理, 从而影响网络性能。这种共享缓冲池的方法通常用于一些低成本的网络设备, 但在高性能要求的网络环境中, 通常会使用更复杂的队列管理和缓冲策略来确保更好的性能和服务质量。

【3】共享缓冲池的缺点是, 在高负载或拥塞情况下, 它可能导致数据包排队和竞争资源

Ask: 此时有没有更优化的方法?

Ans: 在高负载或拥塞情况下, 共享缓冲池的确可能导致数据包排队和资源竞争, 从而影响服务质量。为了应对这些问题, 网络设计和管理可以采用一些优化方法, 以改善性能和避免拥塞。

以下是一些可能的优化方法:

1. **流控制和拥塞管理**: 使用流量控制和拥塞管理机制, 如流控制协议 (例如TCP)、拥塞避免算法、队列管理算法 (如RED、WRED), 以帮助平滑和控制拥塞情况。
2. **增加缓冲区大小**: 增加缓冲池的大小, 以容纳更多数据包。这可以减少数据包排队等待时间, 但要确保不消耗过多内存资源。
3. **分级缓冲区**: 将缓冲池分为不同的优先级或服务类别。不同的通信流可以被分配到不同的缓冲区, 从而确保高优先级流量不受低优先级流量的干扰。
4. **速率限制**: 在入口点对数据包进行速率限制, 以避免过多的流量进入网络, 从而减轻拥塞风险。
5. **智能队列管理**: 使用智能队列管理策略, 如最小带宽保证 (Minimum Bandwidth Guarantee) 和最大带宽限制 (Maximum Bandwidth Limitation), 以确保不同类型的流量得到适当的带宽分配。
6. **深度缓冲区**: 采用深度缓冲区技术, 以支持瞬时的数据包爆发, 确保不会丢失关键数据包。
7. **网络设计优化**: 优化网络拓扑和路径选择, 以降低拥塞风险, 例如使用多路径路由。

这些方法可以根据具体的网络需求和情况来选择和组合。在高负载或拥塞情况下, 使用这些策略可以改善网络性能, 减少拥塞, 并提供更好的服务质量。

我的思考: 有没有一种类似于单独队列处理的方法? 不同的数据包走不同的队列, 使得信息处理更有序, 也可以避免上述问题?

Gpt—Ans: 您的思考是正确的。一种常见的方法是使用不同的队列来处理不同的数据包, 这被称为队列管理或队列调度。这种方法可以帮助避免共享缓冲池时可能出现的负面影响, 如拥塞、延迟和服务质量下降。以下是一些相关的队列管理方法:

1. **Weighted Fair Queuing (WFQ)**: WFQ 是一种队列管理算法, 它为不同的数据包分配不同的优先级和带宽份额。这允许高优先级流量在拥塞时获得更多的带宽份额, 确保服务质量。
2. **Class-Based Queueing (CBQ)**: CBQ 允许将流量分类到不同的队列, 并为每个队列分配带宽限制和优先级。这可以用来实现服务质量策略, 确保关键应用的性能。
3. **Weighted Round Robin (WRR)**: WRR 将队列分配为不同的权重, 每个队列按照其权重依次发送数据包。这有助于平衡不同队列的服务质量。
4. **Differentiated Services (DiffServ)**: DiffServ 使用 Differentiated Services Code Points (DSCP) 字段来标记数据包, 路由器和交换机根据这些标记来确定如何处理数据包。这允许不同的数据包被路由到不同的队列, 以获得不同的服务质量。

5. **Quality of Service (QoS) Policies**: QoS 策略可以配置为为不同的流量类型分配不同的处理队列和服务水平。这允许网络管理员根据应用需求和服务级别要求来管理流量。

这些队列管理方法可以帮助更有序地处理不同类型的数据包，提供更好的服务质量，降低拥塞的风险，并确保关键应用的性能。选择适当的方法取决于网络的需求和硬件能力。

【4】“we can only reserve enough headroom for two lossless traffic classes even though the switches support eight traffic classes.”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余量

Ask: 为什么只支持两个？

Ans:

通常情况下，能够为无损流量类别保留足够的头程（headroom）或缓冲区空间的能力受到多种因素的限制。这些因素包括网络设计、硬件资源、性能需求和成本等。以下是一些可能的原因：

1. **硬件资源限制**: 交换机的硬件资源，包括缓冲区大小、内存、处理能力等，可能有限。在一个网络设备中支持多个无损流量类别需要更多的硬件资源，这可能导致成本增加或性能下降。
 2. **流量类别的需求**: 并非所有网络都需要多个无损流量类别。在某些网络环境中，只有少数一两个无损流量类别是必需的，因此为更多的无损流量类别保留头程可能是不必要的。
 3. **服务质量要求**: 一些应用和服务对无损传输的要求较高，而其他则可以容忍有损传输。因此，网络管理员可能会根据实际需求来确定要为哪些流量类别保留头程，以满足关键应用的需求。
 4. **成本和复杂性**: 增加无损流量类别的支持可能会增加交换机的成本和复杂性。在一些情况下，维护多个无损流量类别可能会变得不划算，特别是如果网络中的无损流量较少或不需要额外的保障。
- 因此，虽然交换机理论上支持多个无损流量类别，但实际上，网络管理员可能会选择只为其中的一部分类别分配足够的头程，以平衡性能需求和资源约束。这个选择通常基于具体网络的需求和可用资源。

我的思考：如果能够支持的超过两个，会不会效果更好？具体会好在哪里？

一些潜在的好处：

- **更好的流量隔离**: 通过将流量分类到多个无损流量类别中，可以更好地隔离不同类型的流量，防止它们相互干扰，从而提高网络的可靠性和服务质量。
- **更好的服务质量控制**: 每个流量类别可以具有不同的服务质量参数，如最小带宽、延迟和丢包率要求。这允许网络管理员更精确地控制不同类型的流量，以满足其性能需求。
- **更好的适应性**: 随着网络负载和流量类型的变化，可以根据需要重新分配头程和资源，以适应不同的情况。
- **支持多样性的应用**: 不同的应用程序和服务可能需要不同的服务质量保障。支持多个无损流量类别可以满足各种应用的需求，包括实时音频/视频、云应用、数据备份等。

【5】“use one lossless class for real-time traffic and the other for bulk data transfer”使用一种无损类别用于实时流量，另一种用于批量数据传输

Ask: 实时流量need无损类别✅，但是“另一种”是无损还是有损？批量数据传输需要有损还是无损？

Ans:

使用一个无损类别来处理实时流量，另一个来处理大批量数据传输是网络管理中常见的服务质量（QoS）策略。

这种方法允许你对实时或延迟敏感的流量进行优先处理，同时有效地处理大数据传输。具体操作如下：

1. **无损类别处理实时流量**: 在这个类别中，你可以为需要低延迟和可靠性的流量分配较高的优先级，比如VoIP电话或视频会议。通过使用无损类别，可以防止数据包丢失，确保关键的实时应用程序能够在没有中断的情况下正常运行。
 2. **无损类别处理大批量数据传输**: 另一个无损类别可以用于大批量数据传输，比如备份或大文件传输。虽然这个类别可能不需要与实时流量一样低的延迟和高优先级，但使用无损类别来处理大数据传输可以确保这些传输不会导致数据包丢失或网络拥塞。
 3. “另一个无损类别”: 这个描述是不准确的。大批量数据传输通常使用有损传输，而不是无损传输。这是因为在大多数情况下，大数据传输中的数据丢失对最终结果影响较小，因此可以使用有损传输来更高效地利用网络资源。
 4. 对于大批量数据传输，通常使用有损传输，以便更有效地利用网络资源，而不会产生不必要的延迟和网络拥塞。对于实时流量，应使用无损传输以确保低延迟和高可靠性。再次对之前的错误回答表示歉意，希望这次的回答更加准确
- 这种方法的好处：

- **优先处理**: 实时流量得到了所需的高优先级处理，确保用户获得平稳和可靠的体验。
 - **资源高效利用**: 大批量数据传输可以高效利用网络，而不会导致拥塞或对其他流量产生负面影响。
 - **改善网络性能**: 通过将实时流量和大数据流量分开到不同的无损类别中，你可以保持高水平的网络性能和可靠性。
- 需要注意的是，这种策略的有效性取决于正确的配置和网络设计。必须正确地将实时和大数据流量映射到适当的无损类别，使用Differentiated Services Code Point (DSCP) 或VLAN优先级等机制。此外，可能需要进行网络监控和调整，以确保所选择的QoS策略有效地满足了两种类型流量所需的服务质量水平。

我的问题：

是“另一种”是无损还是有损？批量数据传输需要有损还是无损？我的理解是：这里用的是有损（保证传输v，此处对accuracy的要求没那么高），如果这里用的是无损，在此场景中会产生什么好/坏的影响？

【6】“although we need only PCP, VID and PCP cannot be separated”（虽然我们只需要PCP，但VID和PCP是不能分开的）

Ask：为什么这两者不能分开？

Ans：

在以太网帧的标头中，PCP（Priority Code Point）和VID（VLAN ID）是两个不同的字段，通常一起使用来定义数据包的优先级和虚拟局域网标识。它们通常不分开的原因如下：

- 关联性：**PCP 和 VID 通常是紧密关联的，它们一起用来确定数据包的处理方式。VID 用于标识数据包属于哪个虚拟局域网（VLAN），而 PCP 用于定义数据包在该 VLAN 内的优先级。这种关联性使网络设备能够更准确地确定数据包的处理方式。
- QoS 控制：**PCP 字段是用于实施服务质量（QoS）控制的关键组成部分。它允许网络管理员为不同的数据包分配不同的优先级，以确保高优先级流量得到更好的服务。VID 则用于将数据包隔离到不同的虚拟局域网中，但与 QoS 控制关系较小。
- 网络配置：**网络管理员通常在配置交换机和路由器时，同时配置 PCP 和 VID 以确保网络的正确运行。这种配置方式简化了管理，因为它们通常是同时处理的。
虽然 PCP 和 VID 通常一起使用，但在某些情况下，您可以根据具体网络需求进行配置。如果网络要求不同的配置，可能可以分别处理 PCP 和 VID，但这需要特定设备和配置支持，并需要更复杂的网络管理。在大多数情况下，将它们一起使用是更常见和方便的做法。

我的思考：如果分开会怎么样？对于原问题会产生更好还是更负面的影响？

Gpt—ans：

将PCP（Priority Code Point）和VID（VLAN ID）分开处理会根据具体的网络需求和设计而有不同的影响，既可以带来好处，也可能带来负面影响。以下是一些可能的影响：

好处：

- 更灵活的配置：**分开处理PCP和VID允许更灵活的配置，可以根据网络需求和服务质量要求，单独调整PCP或VID，而不影响另一个字段。
- 定制化的QoS控制：**分开处理PCP可以提供更精细的服务质量（QoS）控制。您可以更准确地定义不同优先级的数据包处理方式，以满足特定应用的需求。
- 不同层次的虚拟化：**VID用于虚拟局域网（VLAN）标识，而PCP用于QoS。分开处理它们可以允许不同层次的虚拟化，以更好地满足不同网络部署的需求。

负面影响：

- 复杂性增加：**将PCP和VID分开处理可能会增加网络配置和管理的复杂性。网络管理员需要更多的精细控制，这可能会增加操作风险。
- 错误配置风险：**分开处理可能增加配置错误的风险。如果PCP和VID之间的配置不匹配，可能导致数据包被错误地处理。
- 性能影响：**过于复杂的配置可能导致性能下降。网络设备需要更多的处理能力来处理不同的数据包类型和配置。
总的来说，将PCP和VID分开处理可以根据特定网络需求提供更多的灵活性，但需要更谨慎的配置和管理，以确保正确运行并满足性能和服务质量要求。对于不同的网络环境，需要仔细权衡和评估是否分开处理更有利。在大多数情况下，将它们一起使用是更常见和方便的做法。

【7】“The mapping between DSCP values and PFC priorities can be flexible and can even be many-to-one.”

DSCP值和PFC优先级之间的映射可以是灵活的，甚至可以是多对一的。

Ask：为什么可以多对一？

Ans：

1.背景：

DSCP（Differentiated Services Code Point）值和PFC（Priority-based Flow Control）优先级之间的映射可以是灵活的，有时甚至可以是多对一的关系。这意味着网络管理员可以根据特定的网络需求配置它们，以满足不同应用和服务的要求。

具体来说：

- 灵活的映射：**网络管理员可以自定义DSCP值和PFC优先级之间的映射，以适应特定的网络要求。这允许他们根据不同应用和服务的需要来调整网络的行为。
- 多对一的映射：**在某些情况下，多个不同的DSCP值可以映射到相同的PFC优先级。这意味着即使数据包使用不同的DSCP值标记，它们都可以分配到相同的PFC优先级。这通常用于在PFC方面赋予某些类型的流量相同的重要性，而不考虑其DSCP标记。
- 优先处理关键流量：**关键或高优先级的流量，如实时语音或视频，通常会映射到特定的PFC优先级，以确保其获得优先处理和低延迟传输。
- 根据应用需求调整：**不同的应用或服务可能根据其服务质量需求使用不同的DSCP标记。DSCP到PFC优先级的映射可以进行调整，以确保网络提供所期望的服务质量。

这种映射的灵活性允许网络管理员对网络行为进行微调，根据组织的具体需求来设置流量的优先级。这种灵活性对于确保关键应用获得所需的网络资源和低延迟传输非常有价值，同时在网络基础设施内高效地使用PFC。

我的问题：多对一的映射会不会导致多个流量包同时进入同一队列，导致拥塞问题？

Gpt—ans：是的，多对一的映射，即多个不同的DSCP值映射到相同的PFC优先级，可能会导致多个流量流同时进入同一队列，从而增加队列拥塞的潜在风险。这可以在高流量情况下导致以下问题：

1. **拥塞风险**：如果多个不同类型的流量都映射到相同的PFC优先级，并且这些流量同时进入队列，可能会导致队列拥塞，特别是在网络容量有限或拥挤的情况下。
 2. **延迟增加**：拥塞可能导致数据包排队等待传输，从而增加了延迟。对于对延迟敏感的应用，这可能会导致性能下降。
 3. **丢包可能性增加**：在拥塞的情况下，可能需要丢弃某些数据包，以减轻队列的压力。这可能会影响数据完整性和服务质量。
- 因此，在配置DSCP到PFC优先级的映射时，需要谨慎考虑，以确保关键流量获得适当的优先级，并避免引入潜在的拥塞问题。网络管理员需要根据网络的负载情况、带宽资源和服务质量需求来平衡这些映射，以确保高优先级的流量不会因多对一的映射而受到不必要的竞争。此外，适当的流量控制和拥塞管理策略也可以帮助减轻拥塞问题的影响。

【8】为什么S5没有出问题？

Ask：不知道我的理解对不对：

- 1) 右上角的三个path就是数据发送的路径
- 2) dead server说明此服务器坏了（要发送PFC暂停帧）
- 3) 具体解释：

1. 紫色数据从s1->t0->la->t1->s3，但是此时s3坏了，因此s3发出pfc暂停帧（帧a），帧a从s3一路回流到s1；
2. S1收到消息后，不往外发出消息了（冻结时间ing） => 导致原本传输黑色数据的路径被迫“冻结” => s5本身虽然没问题，但是现在起也收不到数据了
3. S2坏了，发出pfc暂停帧，沿着path_blue逆流而上，导致s4进入冻结时间
- 4) 这幅图片想要突出的是：s5本身没毛病，但是受s3的干扰（s3的帧a使得s1进入冻结时间），它此时没法接收数据，这就是所谓的congested port

Ans：完全错误✗

正确解答：

1. 服务器 S1 通过路径 {T0, La, T1} 向 S3 和 S5 发送数据包。紫色数据包发送至 S3，黑色数据包发送至 S5。S3 已失效，因此在端口 T1.p3 收到的紫色数据包将被洪泛到 T1 的其余端口（包括 p4）。一旦紫色数据包位于队列头部，T1.p4 的出口队列就会丢弃它们，因为目标 MAC 不匹配。但在此之前，这些紫色数据包就在那里排队。此外，由于来自 S1 和其他来源的组播流量，T1.p2 也发生拥塞。因此，黑色数据包在 T1 中排队。结果，T1.p3 的入口端口开始暂停La.p1的出口端口。
 2. 因此，当黑色和紫色数据包在 La 中建立队列时，La.p0 的入口端口开始暂停 T0.p2 的出口端口。出于同样的原因，T0.p0 的入口端口开始暂停 S1。
 3. 服务器S4开始通过路径{T1, Lb, T0}向S2发送蓝色数据包。不幸的是，S2也死了。然后，端口 T0.p3 将蓝色数据包泛洪到包括 T0.p2 在内的其余端口。由于T0.p2出端口的所有数据包（包括蓝色数据包）无法被排出，因此T0.p3的入端口开始暂停Lb.p0。
 4. 结果，Lb.p1的入端口开始暂停T1.p4，T1.p1开始暂停S4。请注意，即使在 T1.p2 的拥塞消失后黑色数据包离开 T1 到 S5，T1.p3 也会继续暂停 La.p1。这是因为紫色数据包无法被排出，因为 T1.p4 被 Lb 暂停。然后在四个开关之间形成PFC暂停帧环路。因此出现死锁。一旦发生死锁，即使我们重新启动所有服务器，死锁也不会消失。
- 这种死锁是众所周知的循环缓冲区依赖性的具体示例（参见[12,18,22,36]及其中的参考文献）。然而，循环依赖的原因是“新的”。这是由交换机的洪泛行为引起的。在以太网交换机中，一旦数据包的目的MAC地址未知，数据包就会被洪泛到除接收端口之外的所有端口。正如我们在上面的示例中所示，这种“合法”行为会导致依赖圈的形成。

答而后惑：

P2中的灰色队列是什么？为什么不像其他端口一样画上的是紫色数据包

【9】“Compared with the number of pause frames, pause intervals can reveal the severity of the congestion in the network more accurately”与暂停帧数相比，暂停间隔可以更准确地揭示网络拥塞的严重程度✔

Ask：为什么？

Ans：这句话的含义是，当评估网络中的拥塞情况时，观察和分析暂停间隔（pause intervals）可能会提供比单纯计算暂停帧数量更准确的信息。以下是为什么这种说法成立的解释：

1. **暂停帧数量可能不足以提供完整的信息**：暂停帧是一种特殊的以太网帧，通常用于通知网络设备减缓或停止数据传输，以避免网络拥塞。然而，仅仅统计暂停帧的数量可能无法提供足够的信息来了解拥塞的程度。暂停帧的数量可以告诉你拥塞事件发生的频率，但它们不能提供拥塞期间持续的时间或拥塞的严重性。
 2. **暂停间隔提供更多上下文信息**：暂停间隔指的是两个连续暂停帧之间的时间间隔。通过分析暂停间隔，您可以了解拥塞事件的持续时间和间隔。如果拥塞事件频繁且持续时间长，那么网络可能面临更严重的拥塞问题。相比之下，暂停帧数量可能无法区分短暂的拥塞事件和持续的拥塞。
 3. **综合考虑拥塞严重性**：通过分析暂停间隔，您可以更全面地了解拥塞的严重性。较长或更频繁的暂停间隔可能表示网络中存在更大的问题，而较短的暂停间隔可能表明问题相对较小。
- 总之，暂停帧数量是一个重要的指标，但暂停间隔提供了更多的上下文信息，能够更准确地揭示网络拥塞的严重程度。网络管理员可以通过同时考虑这两个因素来更好地评估和应对拥塞问题，从而改善网络性能和可靠性。

【10】“Pause intervals, unfortunately, are not available for the switches we currently use.” 不幸的是，暂停间隔不适用于我们当前使用的开关。✔

Ask：为什么

Ans：“Pause intervals”（暂停间隔）通常是与“暂停帧”（pause frames）相关的一个概念，它用于管理网络设备之间的流量控制。在以太网网络中，暂停帧用于通知接收设备降低或停止数据传输，以防止拥塞。然而，不是所有以太网交换机和网络设备都支持或提供关于暂停间隔的信息，因此有一些可能的原因解释为什么暂停间隔不适用于某些网络设备：

1. **设备不支持暂停间隔报告**：某些较旧或低成本的以太网交换机和网络设备可能没有支持报告或记录暂停间隔的功能。这意味着它们不能提供有关暂停间隔的信息。
2. **设备配置或监视的限制**：即使某些设备理论上支持暂停间隔，但它们可能被配置为不生成或记录这些信息，或者管理员可能没有启用相关的监视和报告功能。
3. **协议或固件版本**：有些设备可能需要特定的协议或固件版本才能支持暂停间隔的报告。使用较旧的设备或未经升级的设备可能无法提供这些信息。
4. **限制性因素**：特定的网络拓扑、配置或限制条件可能导致某些设备无法提供暂停间隔的信息。例如，某些网络设备在虚拟化环境中可能受到限制。因此，如果某个网络中的交换机或网络设备不提供有关暂停间隔的信息，那么网络管理员可能需要依赖其他指标和工具来监视和评估网络性能，以确保网络正常运行。这可能包括查看暂停帧数量、带宽利用率、丢包率等其他可用的性能指标。

【11】为什么RDMA不能改变incast流量模式？✔

【12】什么是Infiniband？✔

【13】为什么目前还没有 Infiniband 路由器投入生产？✔

【14】为什么Infiniband 与以太网不兼容？✔

【15】iwarp是什么？✔

【16】iwarp为什么可以在DC之间通信 & RoCE为什么不可以在DC之间通信？✔

【17】TCP incast [35,38,39] 和长延迟尾部 [41]是什么？✔

【18】iwarp还需要tcp辅助，这项技术不是很鸡肋吗？✔