白皮书

Cisco Nexus 3100 平台和 9000 系列交换机上的 IEEE 1588 PTP

目录

概述

本文档介绍了如何启用一个高度精确的计时解决方案，该解决方案可以为当今的数据中心网络和金融交易应用提供亚微秒级的精确度，方法是在 Cisco Nexus® 3100 平台交换机和 Cisco Nexus 9000 系列交换机上使用 IEEE 1588-2008 精确时间协议(PTP) 版本 2 (PTPv2)。PTP 是适用于数据包网络的具有纳秒级精确度的分布式时间同步协议。

本文档解释了当今的网络和应用所面临的挑战，说明了为什么需要使用 PTP 提供亚微秒级的精确度，并介绍了该协议的工作原理。本文档还解释了 PTP 概念并比较了硬件和软件时间戳。它还介绍了 Cisco Nexus 3100 平台和 9000 系列支持的 PTP 功能，包括封装远程 SPAN (ERSPAN) 中的 PTP 时间戳。它提供了使用 Cisco Nexus 3100 平台和 9000 系列的 PTP 最佳实践。本文档还介绍了 Cisco Nexus 3100 平台和 9000 系列可以实现的 PTP 精确度，并解释了如何执行测量。

简介

准确且精确的计时信息对于当今的数据中心网络和金融交易应用至关重要。网络和系统管理员需要能够确切了解网络上正在发生的情况以及每个事件发生的时间。应用开发人员和管理员需要在一个庞大而复杂的计算环境内将各种事件日志与流程和应用相关联。合规性和数字调查分析还要求为每个数据事务精确设置时间戳。当今的数据网络的一项基本要求是一个可靠、精确且可部署的时间同步协议，以便可以向数据通信网络的所有相关元素（包括路由器、交换机、服务器和应用）提供精确的计时信息。

挑战

传统上，网络时间协议 (NTP) 一直用于在基于数据包的网络中提供毫秒级的计时。但是，由于上一部分提到的原因，毫秒级精确度已经无法满足要求。

现在，如果要确切了解每个流程中以及服务器和交换机中发生的情况，组织需要一个可以在整个网络中提供微秒级详细信息的计时同步协议。

全球定位系统 (GPS) 可以提供 +/-100 纳秒 (ns) 的精确度，但是它需要一个专用的介质将信号分发给最终用户，这意味着网络中的每台设备使用一个采用 IRIG-B 的 BNC 或某个其他串行接口，从单独的网络接收 GPS 计时信息。该要求使 GPS 无法部署在数据中心网络中，因为在数据中心网络中，即使最小的服务器群也具有数百或数千台服务器以及路由器、交换机及其他网络元素，它们不具有专用的特殊时间协议接口。实际上，组织需要具有精确且易于实施和管理的基于数据包的解决方案。

解决方案

IEEE 1588-2008 PTPv2 是适合当今的数据中心和金融应用的计时解决方案。它提供以下优势：

- 空间上局部化的系统，具有适合较大系统的选项
- 基于数据包的计时分发和同步
- 纳秒到亚微秒级的精确度
- 管理和维护很轻松，管理操作非常少
- 管理冗余和容错系统的能力
- 非常适用于高端和低端设备的低成本、低资源使用率的解决方案

PTP 的精确度源自交换机和服务器网络接口卡 (NIC) 对 PTP 的硬件支持。它使该协议可以精确补偿网络中的消息延迟和变化。对 PTP 提供的这一硬件支持使该协议能够实现纳秒级精确度。

PTP 概念

本部分介绍了 PTP 的一些主要概念。

PTP 时钟类型

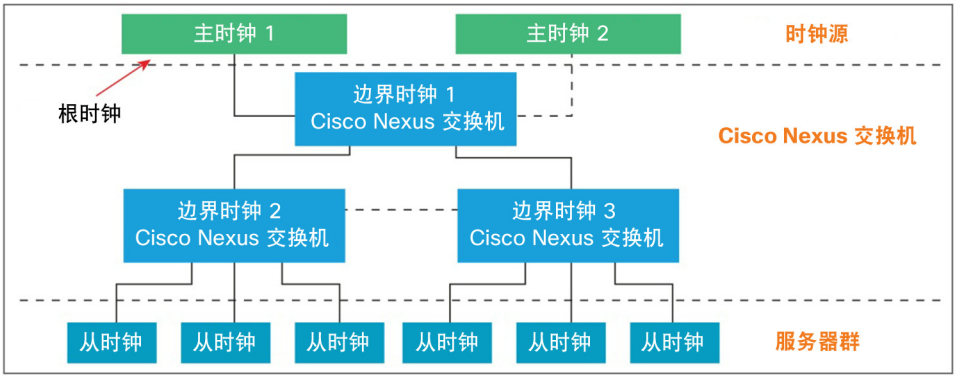
下面总结了主要的 PTP 时钟类型。

- 根时钟 (Gm)：根时钟是其 PTP 域内的最高级时钟，并且是所有其他 PTP 元素的主要参考源。
- 从时钟：从时钟通过将自身与主时钟同步，从主时钟接收时间信息。它不会将时间重新分发给其他时钟。在数据中心内，服务器通常是 PTP 从时钟。
- 普通时钟：普通时钟是具有单个 PTP 端口的 PTP 时钟。它可以是主时钟（根时钟）或从时钟。
- 边界时钟 (BC)：边界时钟是在 PTP 根时钟与其 PTP 从客户端之间的中间设备。它在某个域内具有多个 PTP 端口，并维护该域内使用的时标。边界时钟上的不同端口可以是主端口或从端口。边界时钟终止 PTP 流，恢复时钟和时间戳，然后重新生成 PTP 流。从端口恢复时钟和主端口以重新生成 PTP 数据包。
  透明时钟 (TC)：透明时钟测量 PTP 事件消息经过该设备所需的时间，然后补偿数据包延迟。

PTP 拓扑

最佳主时钟算法 (BMCA) 用于选择每条链路上的主时钟，并最终选择整个 PTP 域的根时钟。它在普通和边界时钟的每个端口本地运行，将本地数据集与从 Announce 消息接收的数据进行比较，以选择链路上的最佳时钟。图 1 显示了数据中心内的 PTP 拓扑示例。

图 1. 数据中心内的 PTP 拓扑



在 IEEE 1588-2008 PTPv2 网络上，需要先在 PTP 域中建立主-从分层时钟拓扑，然后再进行时钟同步。这种树形拓扑类似于生成树拓扑，根时钟是此分层系统中最精确的时钟，因此每个 PTP 从时钟与其同步。在 PTP 网络中，普通和边界时钟的 PTP 端口检查在端口上接收的所有 PTP Announce 消息的内容，然后每个端口运行独立的 PTP 状态机以确定端口状态。使用 BMCA、Announce 消息以及与普通或边界时钟相关联的数据集，可以确定 PTP 端口处于以下三种状态之一：

- 主：该端口是该端口所服务的路径上的时间来源。具有主端口的时钟变成其下游 PTP 节点的主时钟（不是根时钟）。
- 从：该端口与处于主状态的端口的路径上的设备同步。
- 被动：该端口不是路径上的主端口，也不与主设备同步。此状态可防止 PTP 级别的计时环路。

图 1 显示，当网络具有多个主时钟时（例如因为新的根时钟添加到系统中），最终只有一个主时钟被选择为根时钟，并且它变成主-从拓扑的根。连接到主时钟 2 的边界时钟 1 上的端口将过渡到被动状态，并且将不会在这两个时

钟之间建立主-从关系。图中的虚线表示两个边界时钟之间的主-从关系没有形成，并且其中一个端口处于由端口状态机确定的被动状态

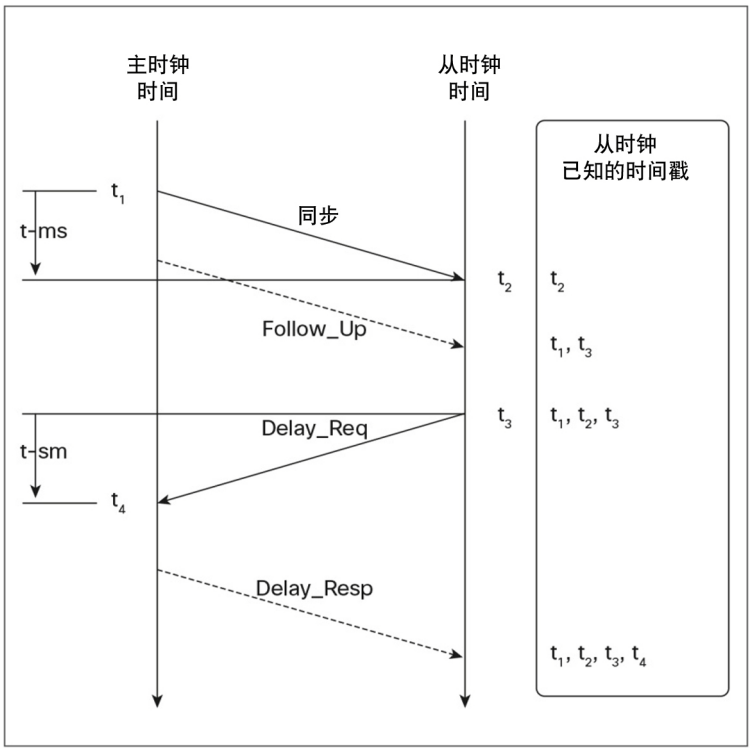
PTP 算法

总的来说，PTP 交换含有某个时间戳的数据包，该时间戳表示接收设备的时钟需要调整到的当前时间。例如，假定 PTP 来源发送一条通告时间为下午 1:00:00 的 PTP 消息。但是，这条消息到达其目的地需要时间。例如，如果 PTP 数据包到达其来源需要 1 秒，则当来源接收一个指明时间是下午 1:00:00 的 PTP 数据包时，时间将是下午 1:00:01。PTP 如何补偿网络延迟？

此同步是通过主时钟与从时钟之间交换的一系列消息实现的，如图 2 所示。

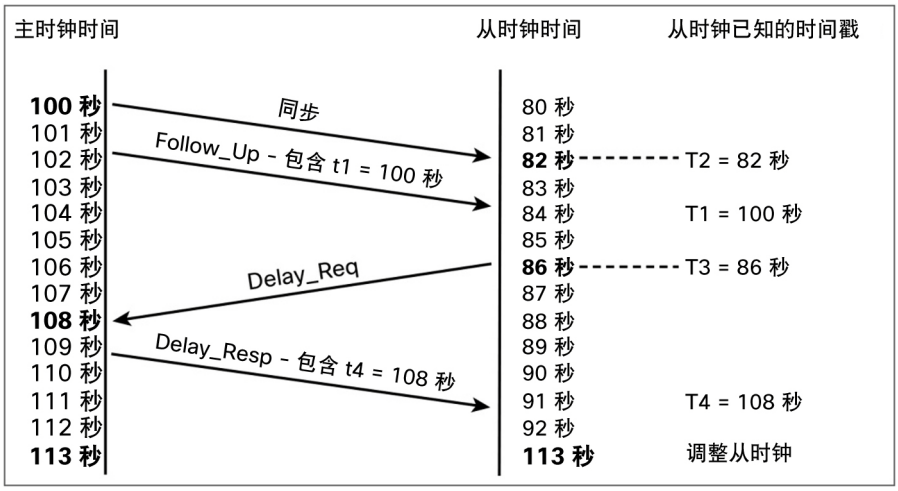
1. 主时钟发送 Sync 消息。Sync 消息离开主时钟的时间带有时间戳 t1，该时间戳可以嵌入到 Sync 消息本身（单步操作）或在 Follow_Up 消息中发送（两步操作）。
2. 从时钟接收 Sync 消息；t2 是从时钟接收 Sync 消息的时间。
3. 从时钟发送 Delay_Req 消息，该消息在离开从时钟时带有时间戳 t3，而在主时钟接收该消息时，该消息带有时间戳 t4。
4. 主时钟以包含时间戳 t4 的 Delay_Resp 消息做出响应。

图 2. 时钟同步流程



在图 3 的示例中，主时钟时间值最初是 100 秒，从时钟时间值是 80 秒。需要调整从时钟时间以便与主时钟时间相匹配。

图 3. PTP消息交换示例



时钟偏差（主时钟与子时钟之间的差异）的计算方法如下所示：

```
Offset = t2 - t1 - meanPathDelay
```

由主时钟到从时钟的链路延迟等于 t2 - t1。

由从时钟到主时钟的链路延迟等于 t4 - t3。

IEEE 1588-2008 假定主时钟与子时钟之间的路径延迟是对称的，因此平均路径延迟的计算方法如下所示：

```
meanPathDelay = ((t2 - t1) + (t4 - t3)) / 2
```

您可以使用此公式计算上一个示例中的主时钟与子时钟之间的偏差：

```
Mean Path Delay = ((t2 - t1) + (t4 - t3)) / 2
                = (-18 + 22) / 2
                = 2


Offset = t2 - t1 - Mean Path Delay
       = 82 - 100 - 2
       = -20
```

在本例中，子时钟比主时钟慢 20 秒，因此会将其时钟调整为 +20 秒。

实施此算法的 PTP 设备被称为两步时钟，因为时间信息由 Sync 和 Follow_Up 消息提供。

该算法的另一个版本（单步操作）在 Sync 消息中发送时间戳，而不是使用单独的 Follow_Up 消息。单步 PTP 硬件可能比较精确，但是在硬件中实施起来可能比较复杂。这种复杂性主要源于需要实时更新时间戳和修改校验和。

单步 PTP 的优势是它可以防止链路和 CPU 的负载增加。

硬件和软件时间戳设置

交换机上的 PTP 实施的目标是向已连接的服务器提供 PTP 计时信号，以便可以准确同步系统时钟。

在软件 PTP 实施中，CPU 会为数据包设置时间戳。与硬件时间戳设置相比，此解决方案更易于设计和实施，并且还具有成本效益。但是它的精确度较低：数据包需要进入 CPU 才可设置时间戳，并且 CPU 运行整个操作系统，而操作系统除了包括 PTP 进程之外，还包括许多其他进程。即使 PTP 进程被给予最高优先级，软件 PTP 也无法与硬件 PTP 一样精确。

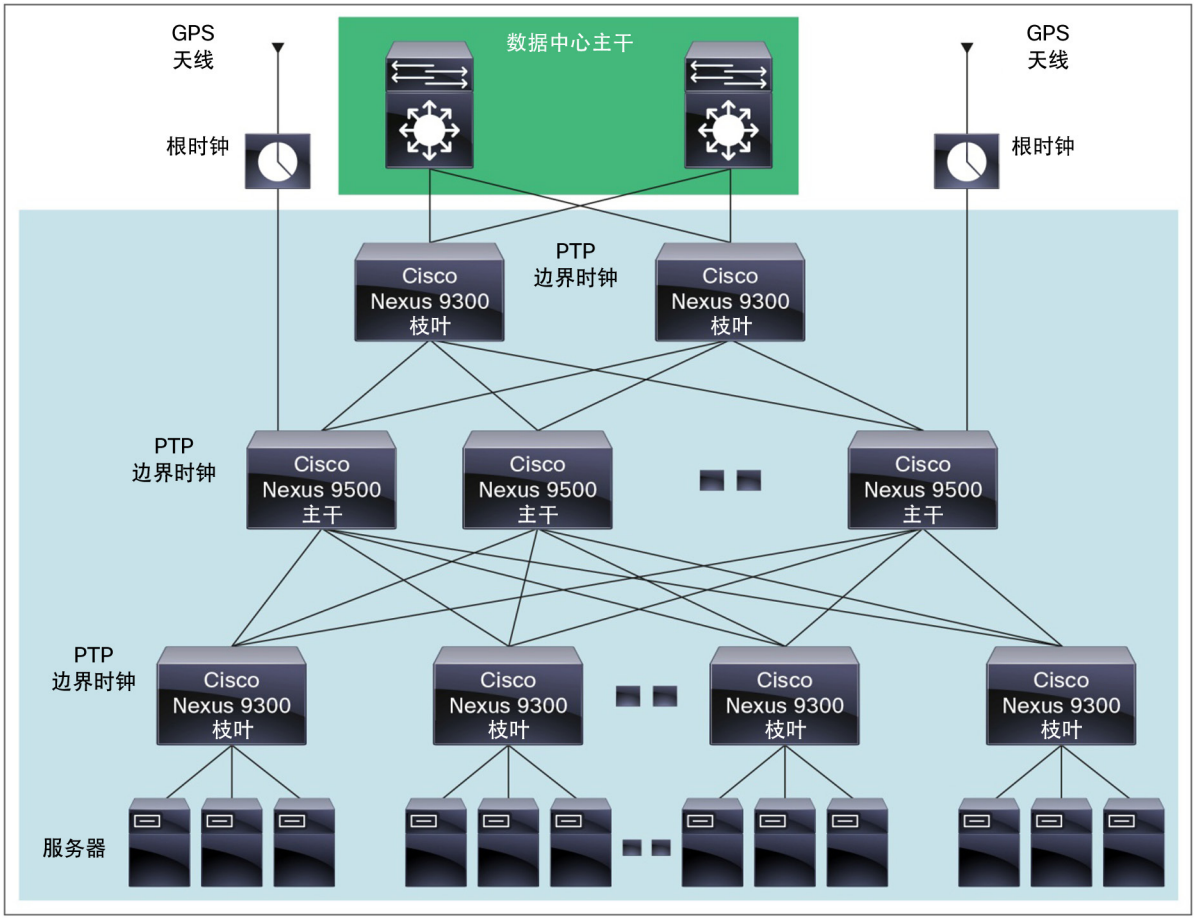
在硬件 PTP 实施中，PTF 数据包中的时间戳的捕获和插入由交换机的特定应用集成电路 (ASIC) 或服务器的 NIC 执行。此过程最好在物理 (PHY) 级别执行。对于入口，执行此过程的最佳时间是刚在线路上接收到数据包时。对于出口，执行此过程的最佳时间就在线路上序列化数据包之前且在数据包离开缓冲区之后，这样时间戳精确度就不会因拥塞或速度不匹配而受到潜在缓冲区使用率影响。

Cisco Nexus 3100 平台和 9000 系列的端到端数据中心 PTP 部署

拓扑和组件

由于 PTP 基于以太网，因此它可以在现有的数据中心架构上运行，并且通常不要求重新设计数据中心。图 4 显示了采用 PTP 组件的典型枝叶-主干式数据中心设计。

图 4. 数据中心 PTP 部署



这种端到端的数据中心 PTP 部署的组件包括根时钟和 Cisco Nexus 交换机。

主时钟

主时钟通常是专用的特定设备。此设备通常连接到 GPS，GPS 提供准确的时间源输入。然后，主时钟通过网络生成 PTP 数据包。

Symmetricom TimeProvider 5000 是 PTP 根时钟（图 5）的一个示例。

图 5. Symmetricom TimeProvider 5000 PTP 根时钟

已启用 PTP 的 Cisco Nexus 3100 平台和 9000 系列数据中心交换机

Cisco Nexus 3100 平台交换机是适用于架顶式 (ToR) 数据中心部署或行尾式 (EoR) 部署的紧凑型单机架单元和双机架单元（1RU 和 2Ru）交换机。它们使用数据中心级 Cisco® NX-OS 软件操作系统提供线速的第 2 层和第 3 层交换。

Cisco Nexus 9000 系列交换机包括 Cisco Nexus 9500 平台模块化交换机和 Cisco Nexus 9300 平台固定配置交换机。Cisco Nexus 9000 系列交换机既可以在思科以应用为中心的基础设施 (ACI) 模式下运行，又可以在思科 NX-OS 模式下运行。

在 ACI 模式下，Cisco Nexus 9000 系列交换机在与思科应用策略基础设施控制器 (APIC) 结合使用时提供以应用为中心的基础设施。

在 NX-OS 模式下，Cisco Nexus 9000 系列交换机起到经典交换机的的作用。Cisco Nexus 9000 系列交换机配备增强版的思科 NX-OS 作为操作系统，通过传统方式提供网络连接，但是具有优异的性能、增强的网络恢复能力和可编程的自动化功能。

本文档重点介绍 NX-OS 模式下的 Cisco Nexus 9000 系列交换机上的 PTP。

从思科 NX-OS 版本 7.0(3)I1(1) 开始，以下 Cisco Nexus 3100 平台交换机和 9000 系列交换机支持 PTP：Cisco Nexus 3164PQ（图 6）、9396Px（图 7）和 9332Pq（图 8）交换机以及带有 Cisco X9636PQ 线卡的 9500 平台机箱（图 9）。采用 PTP 的 X9636PQ 线卡可用于 Cisco Nexus 9504 或 9508 机箱。

图 6. Cisco Nexus 3164Q

图 7. Cisco Nexus 9396PX

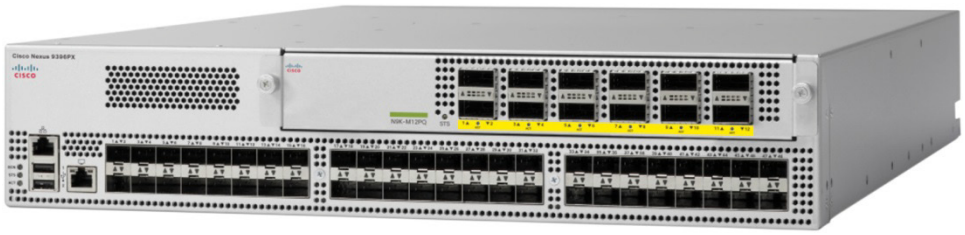


图 8. Cisco Nexus 9332PQ

图 9. 带有 X936PQ 线卡的 Cisco Nexus 9500 机箱

从思科 NX-OS 版本 7.0(3)I1(1) 开始，Cisco Nexus 3100 平台和 9000 系列以太网交换机上支持 PTP。它包括以下功能：

- IEEE 1588-2008 PTPv2 标准
- 两步的边界时钟
- 基于 Ipv4 组播的用户数据报协议 (UDP)，使用 IEEE 1588 标准中定义的组播地址 224.0.1.129
- 硬件辅助的 PTP 实施
- 默认情况下通过处理和转发具有较高优先级的 PTP 消息来实现高效处理网络拥塞；无需执行其他步骤以配置额外的服务质量 (QoS)
- 支持 ERSPAN 报头版本为 3 的 ERSPAN，该版本包括时间戳
- 在第 2 层和第 3 层接口上支持 PTP

请注意：

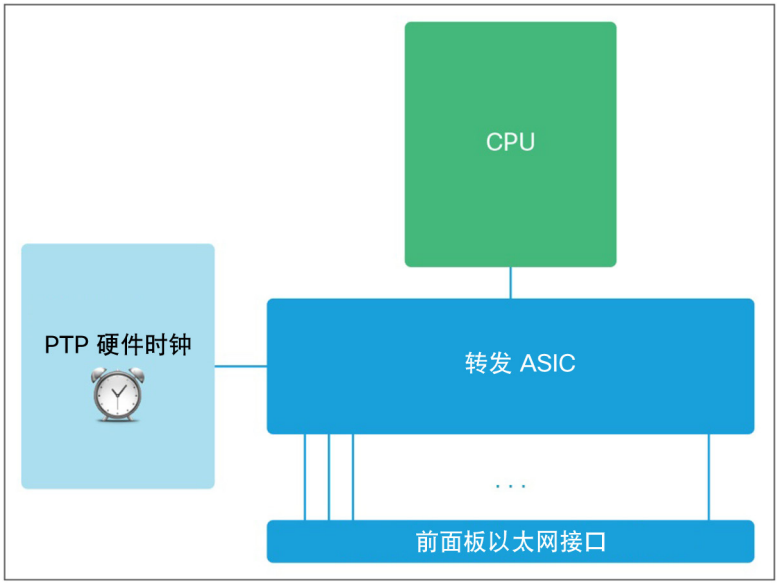
- 在 Cisco Nexus 9332PQ 交换机上，最后六个物理端口不支持 PTP。
- 在 Cisco Nexus 9396PX 交换机上，所有 40 Gbps 物理端口不支持 PTP。
- 在带有 X9636PQ 线卡的 Cisco Nexus 9504 和 9508 交换机上，所有物理端口均支持 PTP。
- 在 Cisco Nexus 3164Q 交换机上，所有物理端口均支持 PTP。

Nexus 3100 平台和 9000 系列 PTP 架构

Cisco Nexus 3100 平台和 9000 系列支持硬件辅助的 PTP 操作。转发 ASIC 在硬件中为入口和出口方向的 PTP 数据包设置时间戳。

图 10 显示了 Cisco Nexus 3100 平台和 9000 系列 PTP 架构。

图 10. Cisco Nexus 3100 平台和 9000 系列 PTP 架构



在 Cisco Nexus 9300 平台交换机上，PTP 硬件时钟在应用枝叶引擎 (ALE) 和 ALE-2 ASIC 中。在 Cisco Nexus 3100 平台交换机，它在一个单独的现场可编程门阵列 (FPGA) 中。在 Cisco Nexus 9500 平台模块化交换机上，它在所有线卡使用的管理引擎卡上存在的 FGPA 中。

Cisco Nexus 3100 平台和 9000 系列 PTP 数据包流

两种可能的情况：在一种情况下，Cisco Nexus 3100 平台或 9000 系列交换机的某个前面板端口是 PTP 从时钟，在另一种情况下，该端口是 PTP 主时钟。

当 Cisco Nexus 3100 平台或 9000 系列交换机的某个前面板端口是 PTP 从时钟时，交换机从 PTP 主时钟接收时间信息并纠正其自己的时钟。数据包流如下所述：

1. 交换机在从端口上的 PTP 主时钟接收 PTP Sync 消息。转发 ASIC 具有一个同步到 PTP 硬件时钟的时钟。转发 ASIC 记录到达时间 Sync 数据包的到达时间。它会使用内部报头将此时间戳添加到该数据包并将其转发到 CPU。在 CPU 上运行的 PTP 软件进程接收该数据包并存储该时间戳。时间戳是图 2 中的 t2。
2. PTP 主时钟将一条 PTP Follow_Up 消息发送到交换机。该消息包含由主时钟在发送 PTP Sync 数据包时记录的 t1 时间戳。Cisco Nexus 3100 平台或 9000 系列交换机 PTP 软件进程接收此数据包并存储其时间戳。PTP 软件进程现在具有 t1 和 t2。
3. Cisco Nexus 3100 平台或 9000 系列交换机将 Delay_Req 消息发送到主时钟。当 Delay_Req 消息离开转发 ASIC 时，会记录离开的时间戳。PTP 软件进程存储此时间戳，即 t3。PTP 软件进程现在具有 t1、t2 和 t3。
4. PTP 主时钟记录其收到 Delay_Req 消息的时间。此时间戳是 t4。它会将包含 t4 的 Delay_Resp 信息发送到 Cisco Nexus 3100 平台或 9000 系列交换机从时钟。PTP 软件进程现在具有 t1、t2、t3 和 t4。
5. Cisco Nexus 3100 平台或 9000 系列交换机上的 PTP 软件根据本文档前面部分“PTF 算法”中介绍的公式计算机与 t1、t2、t3 和 t4 的偏移量。Cisco Nexus 3100 平台或 9000 系列交换机的 PTP 硬件时钟相应地调整。该调整正是作为 PTF 边界时钟运行的设备执行所谓的时钟恢复操作的原因所在。

Cisco Nexus 3100 平台或 9000 系列交换机的 PTP 从端口上的 PTP 消息交换到此结束。此消息交换在 Cisco Nexus 3100 平台或 9000 系列交换机从端口与其连接到的主端口之间持续执行，因此，交换机的时钟与主时钟始终同步。

当 Cisco Nexus 3100 平台或 9000 系列交换机的某个前面板端口是 PTP 主时钟时，交换机将向与其同步的从设备提供该时间。数据包流如下所述：

1. Cisco Nexus 3100 平台或 9000 系列交换机将 PTP Sync 消息发送到从时钟。它使用其 PTP 硬件时钟记录时间戳 t1，该时间戳指明数据包离开转发 ASIC 的时间。
2. Cisco Nexus 3100 平台或 9000 系列交换机将该 t1 时间戳通过 PTP Follow_Up 数据包发送到从时钟。
3. Cisco Nexus 3100 平台或 9000 系列交换机接收来自从时钟的 PTP Delay_Req 消息。它记录时间戳 t4，该时间戳指明收到数据包的时间。
4. Cisco Nexus 3100 平台或 9000 系列交换机将 t4 通过 PTP Delay_Resp 数据包发送到从时钟。
5. 从设备现在可以根据它从主时钟接收并记录的时间戳调整自己的时钟。Cisco Nexus 3100 平台或 9000 系列交换机在任何时候都不会调整自己的时钟，因为此 PTP 消息交换在主端口上进行。

Cisco Nexus 3100 平台或 9000 系列交换机 PTP 实施不会受拥塞或缓冲影响。当数据包为序列化做好准备时以及在缓冲数据包后，交换机会在 PHY 级别添加时间戳。此方法可产生非常高的 PTP 精确度，如本文档后面的“Cisco Nexus 3100 平台和 9000 系列的数据中心 PTP 性能”部分所讨论。

Cisco Nexus 3100 平台和 9000 系列 PTP 配置

本部分介绍 Cisco Nexus 3100 平台和 9000 系列上最重要的 PTP 配置命令。有关配置的全面指南，请参阅位于以下网址的系统管理配置指南：http://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/system_management/configuration/guide/b_Cisco_Nexus_9000_Series_NX-OS_System_Management_Configuration_Guide_7x/b_Cisco_Nexus_9000_Series_NX-OS_System_Management_Configuration_Guide_7x_chapter_0100.html。

必需的配置

此处显示的命令必须存在，PTP 才可在 Cisco Nexus 3100 平台和 9000 系列上运行。

全局配置命令

```
switch(config)# feature ptp
```

此命令在交换机上启用 PTP。

```
switch(config)# ptp source <ip>
```

此命令为交换机生成的 PTP 数据包指定源 IP 地址。

接口配置命令

```
switch(config)# interface Ethernet slot/port
switch(config-if)# ptp
```

这些命令在端口上启用 PTP。Cisco Nexus 3100 平台和 9000 系列交换机是边界时钟，因此它同时具有主端口和从端口。主端口与从端口之间不存在配置差异。两者均使用 **ptp** 选项进行配置，BMCA 确定端口是 PTP 从端口还是主端口。

可选的配置

此处列出的命令是可选的。

```
switch(config)# clock protocol ptp
```

此命令配置交换机，以便它使用 PTP 更新系统日历。此配置使交换机的时钟与 PTP 保持同步。如果您不启用此命令，交换机仍将在其主端口上传播 PTP 时钟。但是，时间源将是 Cisco Nexus 本地时钟。

```
switch(config)# interface ethernet slot/port
switch(config-if)# sync interval value
```

这些命令在接口上配置 PTP Sync 消息数据包速率。如果未指定值，默认值为 0。

PTP logSyncInterval (**sync interval**) 值表示每秒在接口上发送的 PTP Sync 消息数据包数量（表 1）。

表 1. PTP logSyncInterval 值

logSyncInterval 值	每秒的 PTP Sync 消息数量
-3	8
-2	4
-1	2
0	1
1	每 2 秒 1 条

您应该使默认的 **sync interval** 值设置保留为 0，除非您确实需要更高的精确度。

Cisco Nexus 3100 平台和 9000 系列 PTP 配置验证

此处显示的命令可用于在 Cisco Nexus 3100 平台和 9000 系列交换机上验证 PTP。

```
switch# show clock
```

此命令可用于验证交换机时钟是否与根时钟同步。您不能使用此命令行界面 (CLI) 命令验证准确的精确度，但是如果已配置 **clock protocol ptp**，您至少可以验证时间是否与根时钟相匹配。

```
switch# show ptp clock
```

此命令显示本地时钟的属性，包括时钟身份。

以下是 **show ptp clock** 输出的示例：

```
switch# show ptp clock
PTP Device Type: Boundary clock
Clock Identity :  88:f0:31:ff:fe:2a:fa:e1
Clock Domain: 0
Number of PTP ports: 3
Priority1 : 196
Priority2 : 255
Clock Quality:
        Class : 248
        Accuracy : 254
        Offset (log variance) : 65535
Offset From Master : 0
Mean Path Delay : 0
Steps removed : 0
Local clock time:Wed Jan 28 17:56:19 2015
9396px-nd1#


switch# show ptp parent
```

此命令显示 PTP 父时钟的属性。它可用于验证父时钟的身份。

以下是 **show ptp parent** 输出的示例：

```
switch# show ptp parent

PTP PARENT PROPERTIES

Parent Clock:
Parent Clock Identity:  00:14:01:ff:fe:00:00:01
Parent Port Number: 1
Observed Parent Offset (log variance): N/A
Observed Parent Clock Phase Change Rate: N/A
```

```
Grandmaster Clock:
Grandmaster Clock Identity:  00:14:01:ff:fe:00:00:01
Grandmaster Clock Quality:
        Class: 6
        Accuracy: 35
        Offset (log variance): 0
        Priority1: 1
        Priority2: 1

switch# show ptp brief
```

此命令显示所有接口的 PTP 状态。PTP 端口可以处于以下三种状态之一：

- 主：该端口是该端口所服务的路径上的时间来源。
- 从：该端口与处于主状态的端口的路径上的设备同步。
- 已禁用：未在此端口上启用 PTP。
- 被动：该端口不是路径上的主端口，也不与主设备同步。

由于 Cisco Nexus 3100 平台或 9000 系列交换机是 PTP 边界时钟并且仅支持一个 PTP 域，因此交换机只能具有一个从端口。如果交换机已有一个从端口，则连接到第二个根时钟的第二个端口将处于被动状态。当第一个根时钟或第一个从端口发生故障时，BMCA 会将之前被动的端口移至从状态。通过此过程可以实现根时钟冗余。

以下是 **show ptp brief** 输出的示例：

```
switch# sh ptp brief

PTP port status
-----------------------
Port         State
-------  --------------
Eth1/1      Slave
Eth1/2      Master
Eth1/3      Master
Eth1/4      Master
Eth1/5      Master
Eth1/6      Master
...
Switch#

switch# show ptp corrections
```

此 CLI 命令显示最后几个 PTP 纠正。

以下是 **show ptp corrections** 输出的示例：

```
PTP past corrections
-------------------------------------------------------
Slave Port SUP Time Correction(ns) MeanPath Delay(ns)
---------- ---------------------------------------------
Eth1/46 Mon Dec 23 09:52:11 2013 48581 -1 293
Eth1/46 Mon Dec 23 9:52:12 2013 49318 3 297
Eth1/46 Mon Dec 23 9:52:13 2013 49193 -8 297
Eth1/46 Mon Dec 23 9:52:14 2013 49208 12 298
Eth1/46 Mon Dec 23 9:52:15 2013 48625 -3 298
Eth1/46 Mon Dec 23 9:52:16 2013 47607 -13 295
Eth1/46 Mon Dec 23 9:52:17 2013 49091 0 295
Eth1/46 Mon Dec 23 9:52:18 2013 47961 2 295
Eth1/46 Mon Dec 23 9:52:19 2013 48005 -1 295
Eth1/46 Mon Dec 23 9:52:20 2013 48350 0 296
Eth1/46 Mon Dec 23 9:52:21 2013 48507 -5 292
Eth1/46 Mon Dec 23 9:52:22 2013 48105 2 292
Eth1/46 Mon Dec 23 9:52:23 2013 48188 12 301
Eth1/46 Mon Dec 23 9:52:24 2013 48021 6 301
Eth1/46 Mon Dec 23 9:52:25 2013 48239 -12 296
```

Cisco Nexus 3100 平台和 9000 系列的数据中心 PTP 性能

PTP 性能衡量定义和概念

本部分介绍用于描述 PTP 性能的基本定义和概念。

偏差是两个时钟上的时间之间的差异。在以下测试结果中，它用于衡量从时钟与主时钟的同步有多准确。它指明了作为边界时钟的 Cisco Nexus 3100 平台或 9000 系列交换机所提供的测量值的不准确程度。此值越小越好。

平均路径延迟是 PTP 帧在主从时钟之间流动花费的平均时间。此测量值不指示交换机或服务器的性能或准确度。较小的平均路径延迟有助于获取基准结果。如果平均路径延迟较大且具有大量抖动，则意味着是一个复杂的数据中心，具有缓冲和延迟尖峰、控制协议正在运行、流量速率高等。在处理密集型环境的实际 PTP 性能测试中，此测量值可能非常值得关注。

标准偏差指明与平均值的差异或离差。低标准偏差表示数据点往往十分接近该平均值（也称为预期值）；高标准差价表示数据点分布在较大的值范围。

在本文档的其余部分，以下定义用于描述 PTP 性能：

- 平均偏移量
- 偏移量标准偏差
- 最小和最大偏移量峰值
- 平均路径延迟

外部工具使用相同的稳定时钟参考测量出这些值。

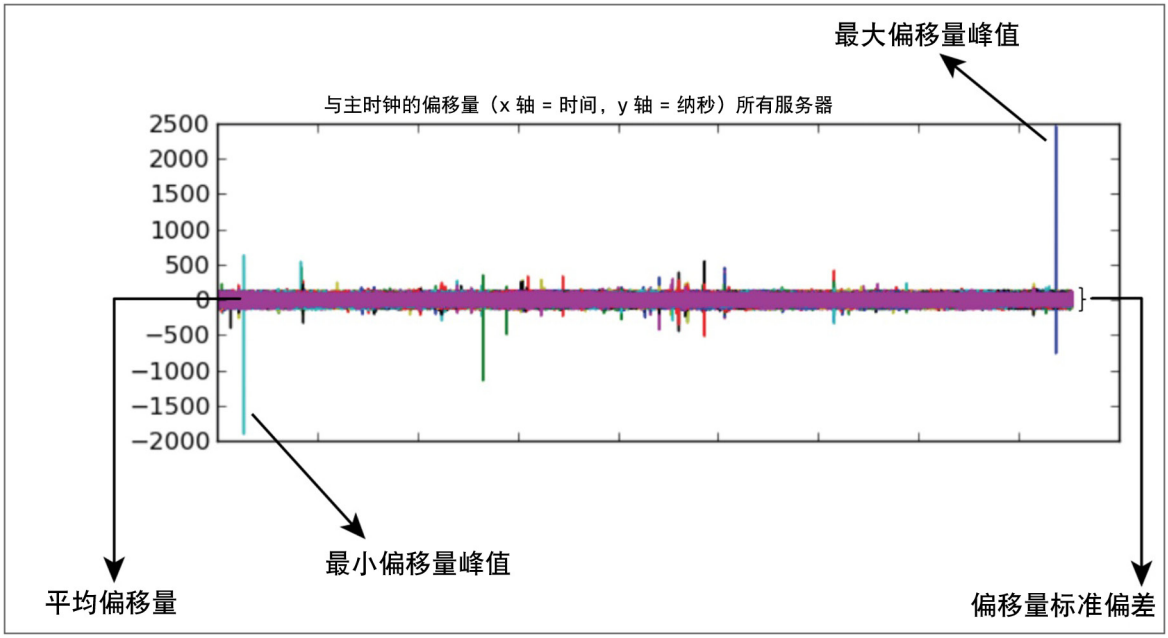
平均偏移量通常接近 0，因为在计算平均值时，正和负时钟偏移量会相互调平。

最小和最大偏移量峰值非常有用；但是，单单这些值并未指明达到这些峰值的频率。因此，除了平均偏移量和偏移量峰值外，偏移量标准偏差也很重要。了解偏移量标准偏差的另一种方法是通过抖动。

这四个数据点和平均路径延迟让您可以非常好地了解设备的 PTP 性能。

图 11 是一个偏移量图表示例，显示了平均偏移量、偏移量标准偏差以及最小和最大偏移量峰值。纵轴是偏移量值，水平轴是 PTP 客户端集。

图 11. 偏移量值的图表示例



Cisco Nexus 3100 平台和 9000 系列性能测量方法和设备

此处描述的测试将来自 Spirent 的流量生成器用于 PTP 消息生成、主从模拟和后台流量。它使用一个带有 HyperMetrics CV 8 个万兆以太网增强型小型封装热插拔 (SFP+) 卡的 Spirent 9RU 机箱 (SPT-9000)。使用 Spirent 版本 4.33.0086。

Spirent 9RU 机箱连接到 Symmetricom TimeProvider 5000 根时钟。Spirent 从计时接口接收时钟。Symmetricom TimeProvider 5000 根时钟从嵌入式 GPS 接收器获取其时间源。

Spirent 9RU 机箱有一个主端口连接到接受测试的 Cisco Nexus 交换机。Spirent 9RU 机箱上的其余端口是从 Cisco Nexus PTP 主端口获取时钟的 PTP 从端口，这些主端口从 Spirent 9RU 机箱重新分发时钟。因此，Spirent 可以计算从其主端口发送的时钟与在其从端口上收到的时钟之间的偏移量。此结果指明接受测试的 Cisco Nexus 交换机的精确度。

图 12 显示了性能测量方法。

图 12. 性能测量方法

Cisco Nexus 3100 平台和 9000 系列使用思科 NX-OS 版本 7.0(3)I1(1)。

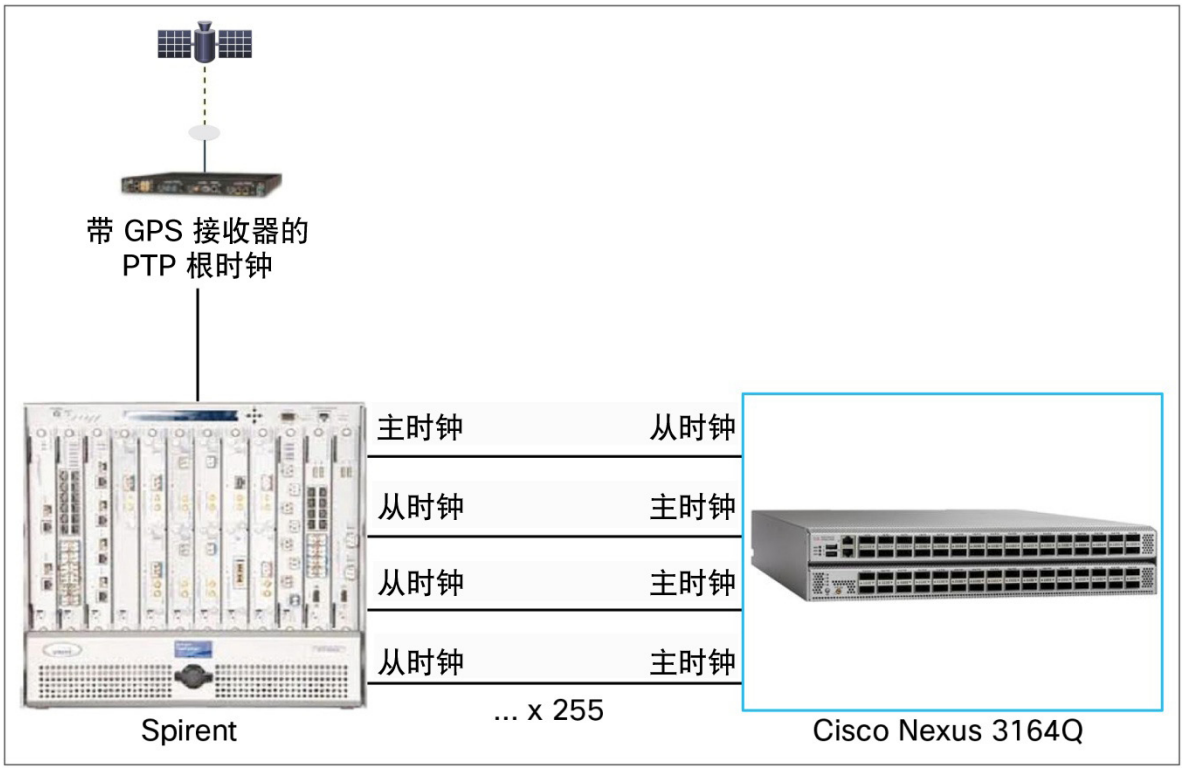
对于所有测试，PTP 配置是 Cisco Nexus 交换机上的默认设置。特别是，PTP **sync interval** 参数保留默认值 0。

Cisco Nexus 3164Q PTP 性能

Cisco Nexus 3164Q 以 4 个 10-Gbps 分支模式配置，具有 44 个连接到 Spirent 的物理 10-Gbps 端口。此配置会产生 43 个 PTP 从端口。此外，Spirent 中配置了 212 个模拟 PTP 从端口。此配置共计有 255 个 PTP 从端口，Cisco Nexus 3164Q 需要向其提供时间。

图 13 显示了采用 Spirent 的 Cisco Nexus 3164Q 性能拓扑。

图 13. 采用 Spirent 的 Cisco Nexus 3164Q 性能拓扑



在 6 小时后，作为所有端口平均值的结果如下：

- 平均偏移量：4 纳秒
- 偏移量标准偏差：50 纳秒
- 最小偏移量峰值：-300 纳秒
- 最大偏移量峰值：302 纳秒
- 平均路径延迟：110 纳秒

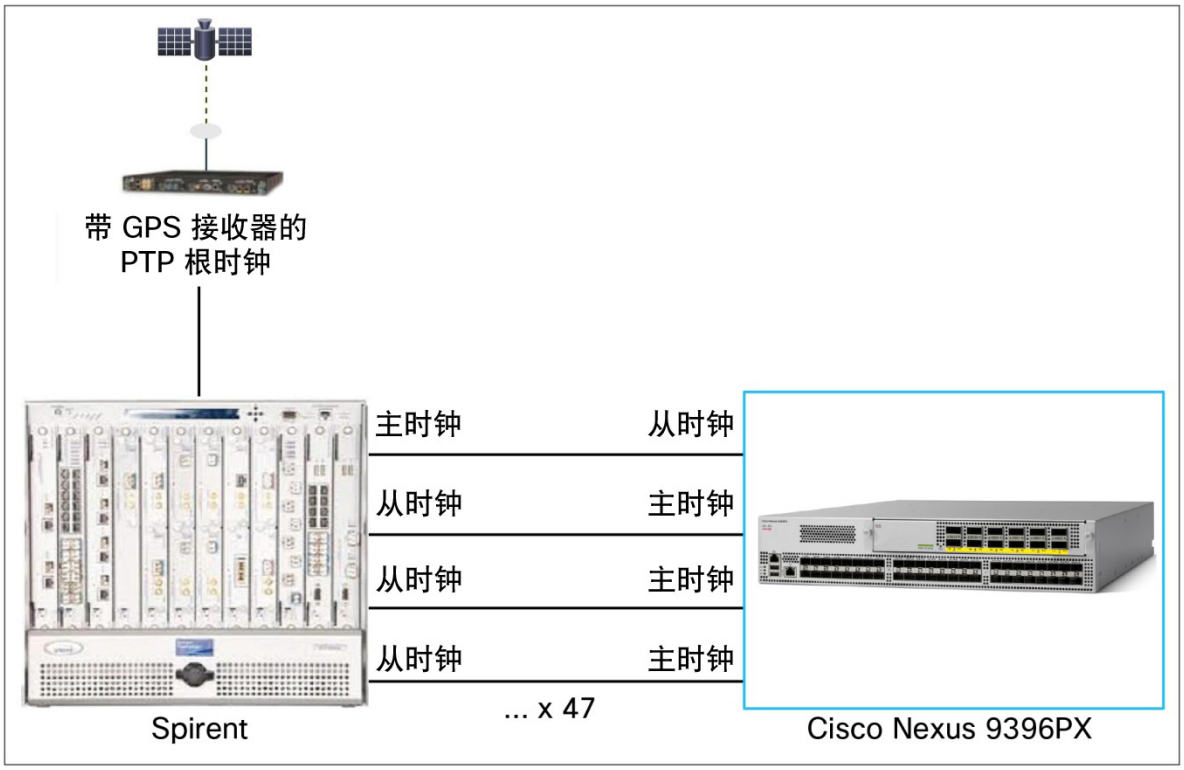
平均偏移量非常接近 0。偏移量标准偏差为 50 纳秒表明，在测试持续期间，平均偏移量值聚集在平均值附近。因此，Cisco Nexus 3164Q 在 PTP 客户端时钟同步流程中引入非常小的偏移量，即使具有大量的 PTP 从端口。

Cisco Nexus 9396PX PTP 性能

Cisco Nexus 3164Q 配置有 44 个连接到 Spirent 的物理 10-Gbps 端口。此配置会产生 43 个 PTP 从端口。此外，Spirent 中配置了 4 个模拟 PTP 从端口。此配置共计有 47 个 PTP 从端口，Cisco Nexus 9396PX 需要向其提供时间。

图 14 显示了采用 Spirent 的 Cisco Nexus 9396PX 性能拓扑。

图 14. 采用 Spirent 的 Cisco Nexus 9396PX 性能拓扑



在 6 小时后，作为所有端口平均值的结果如下：

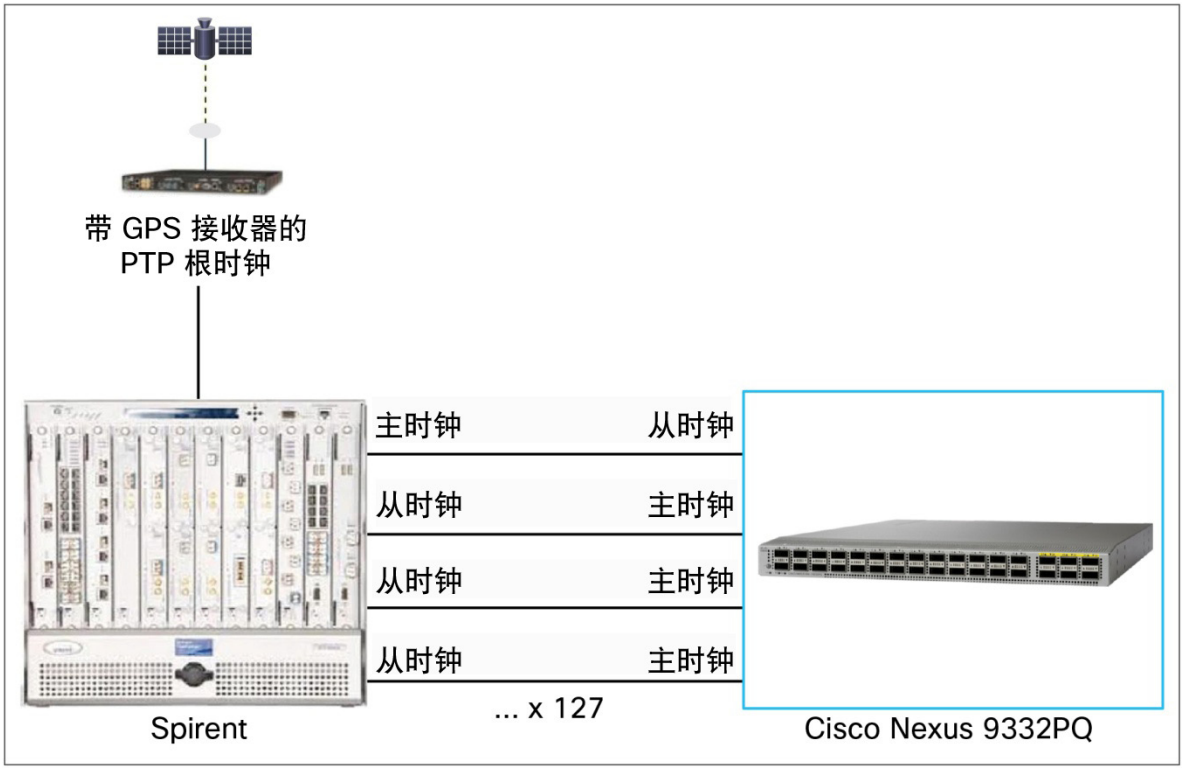
- 平均偏移量：2 纳秒
- 偏移量标准偏差：50 纳秒
- 最小偏移量峰值：-200 纳秒
- 最大偏移量峰值：217 纳秒
- 平均路径延迟：115 纳秒

Cisco Nexus 9332PQ PTP 性能

Cisco Nexus 9332PQ 以 4 个 10-Gbps 分支模式配置，具有 44 个连接到 Spirent 的物理 10-Gbps 端口。此配置会产生 43 个 PTP 从端口。此外，Spirent 中配置了 84 个模拟 PTP 从端口。此配置共计有 127 个 PTP 从端口，Cisco Nexus 9332PQ 需要向其提供时间。

图 15 显示了采用 Spirent 的 Cisco Nexus PQ 性能拓扑。

图 15. 采用 Spirent 的 Cisco Nexus 9332PQ 性能拓扑



在 6 小时后，作为所有端口平均值的结果如下：

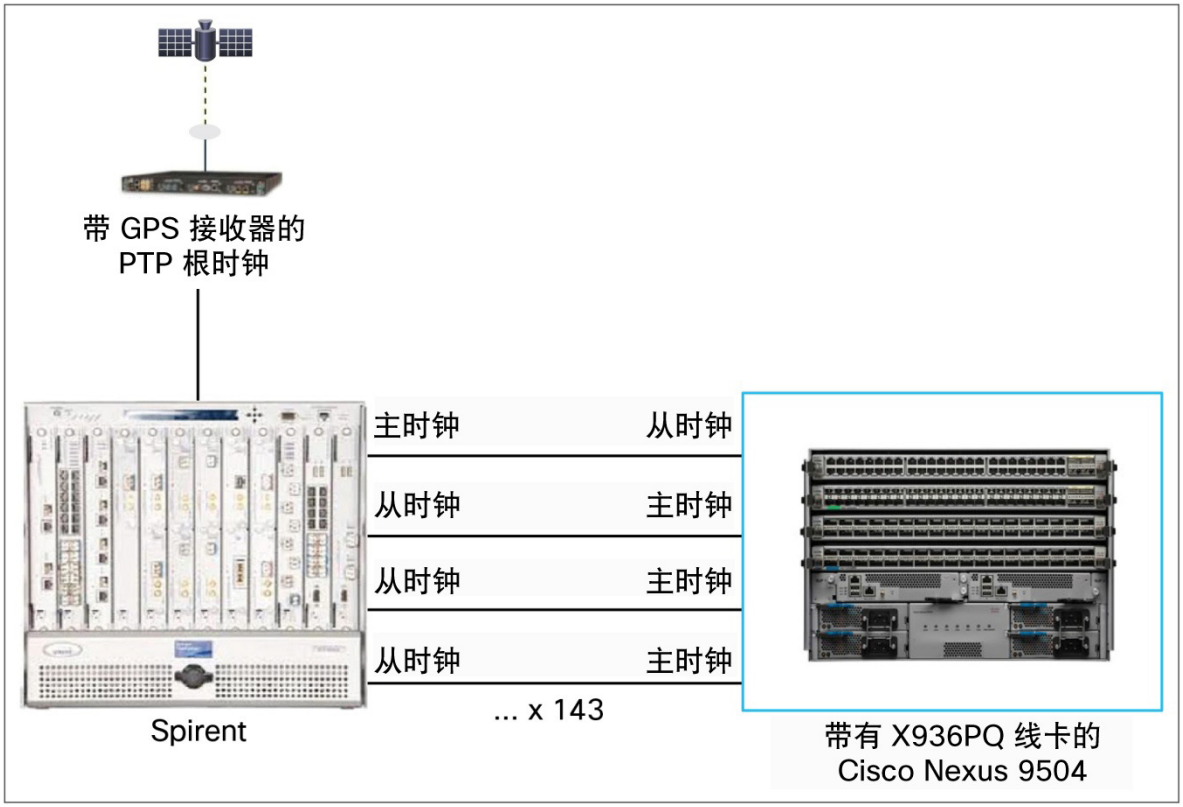
- 平均偏移量：-5 纳秒
- 偏移量标准偏差：11 纳秒
- 最小偏移量峰值：-101 纳秒
- 最大偏移量峰值：98 纳秒
- 平均路径延迟：120 纳秒

带有 X9636PQ 线卡的 Cisco Nexus 9504 PTP 性能

Cisco Nexus 9500 平台交换机上的 Cisco Nexus X9636PQ 线卡以 4 个 10-Gbps 分支模式配置，具有 44 个连接到 Spirent 的物理 10-Gbps 端口。此配置会产生 43 个 PTP 从端口。此外，Spirent 中配置了 99 个模拟 PTP 从端口。此配置共计有 143 个 PTP 从端口，Cisco Nexus X9636PQ 线卡需要向其提供时间。

图 16 显示了采用 Spirent 的 Cisco Nexus X9636PQ 性能拓扑。在图中，线卡连接到 Cisco Nexus 9504 交换机。

图 16. 采用 Spirent 的 Cisco Nexus X9636PQ 性能拓扑



在 6 小时后，作为所有端口平均值的结果如下：

- 平均偏移量：3 纳秒
- 偏移量标准偏差：9 纳秒
- 最小偏移量峰值：-95 纳秒
- 最大偏移量峰值：81 纳秒
- 平均路径延迟：108 纳秒

平均偏移量接近 0 纳秒。它始终接近此平均值（标准偏差为 9 纳秒）。即使具有大量的 PTP 从端口，带有 Cisco Nexus X9636PQ 线卡的 Cisco Nexus 9500 机箱也保持极高的 PTP 精确度。

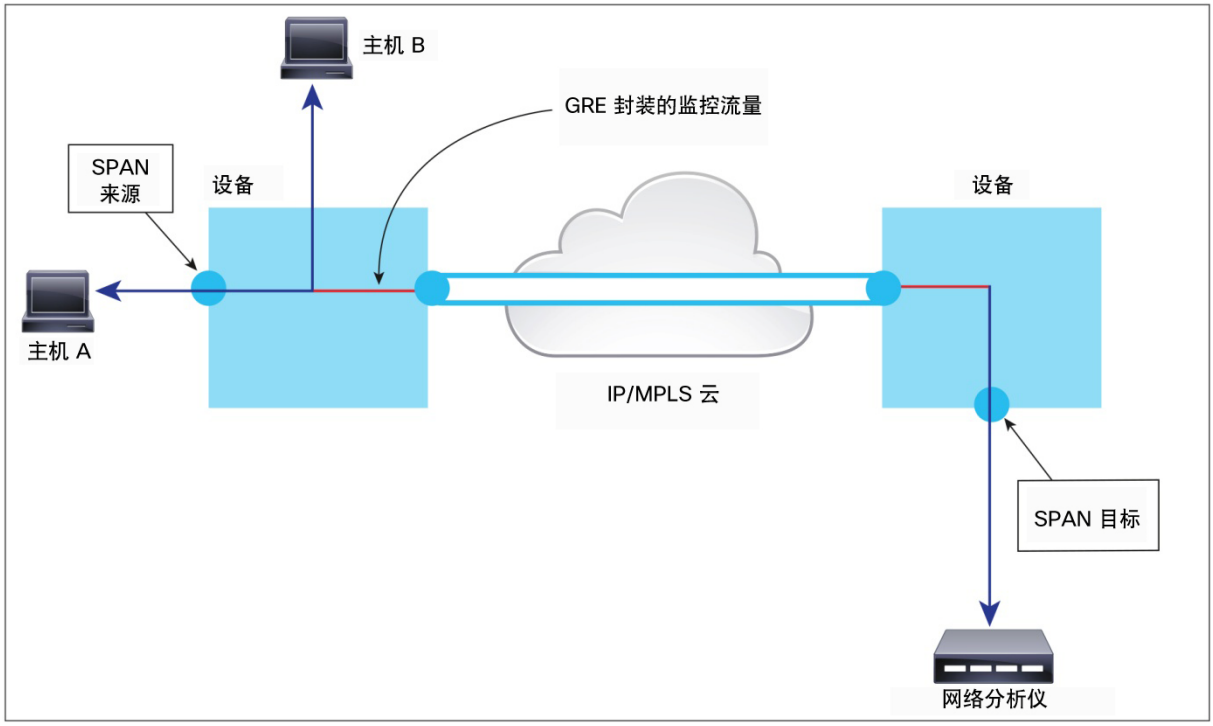
Cisco Nexus 9000 系列 ERSPAN 和 PTP 时间戳设置

ERSPAN 的概念

ERSPAN 通过 IP 网络传输镜像流量。流量在源路由器封装，然后在网络上传输。数据包在目标路由器解封，然后发送到目标接口。

ERSPAN 包括 ERSPAN 源会话、可路由 ERSPAN 通用路由封装 (GRE) 封装的流量和 ERSPAN 目标会话。您可分别配置不同交换机上的 ERSPAN 源会话和目标会话。图 17 显示了 ERSPAN 拓扑。

图 17. ERSPAN 拓扑



Cisco Nexus 9000 系列支持包含时间戳信息的 ERSPAN 类型 III。它可用于计算网络的不同部分之间的数据包延迟。

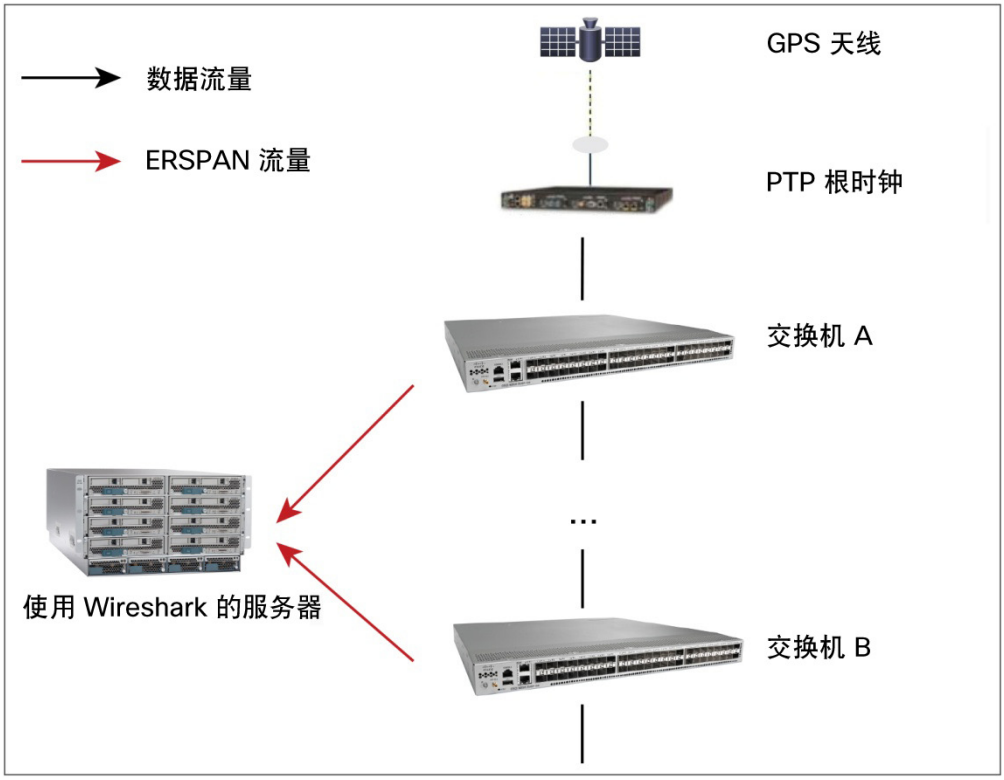
图 18 显示了数据中心内的延迟监控典型示例。目标是确定交换机 A 与交换机 B 之间的延迟。

为监控延迟，在每台交换机上配置了 ERSPAN 源会话。此会话的源端口是延迟受到监控的流量的入口端口。会话的目的 IP 地址是运行数据包捕获分析器（如 Wireshark）的服务器的 IP 地址。两台交换机的时钟使用 PTP 进行同步，因此在 ERSPAN 报头中记录的时间戳具有相同的时间参考。在分析器服务器上，接收来自两台交换机的数据包（因为配置了相同的 ERSPAN 目的 IP 地址）。

利用此信息，您可以轻松计算接收的数据包的 ERSPAN 时间戳之间的增量以获取网络延迟。

图 18 显示了如何配置 ERSPAN 和 PTP 以在典型网络中执行延迟监控。

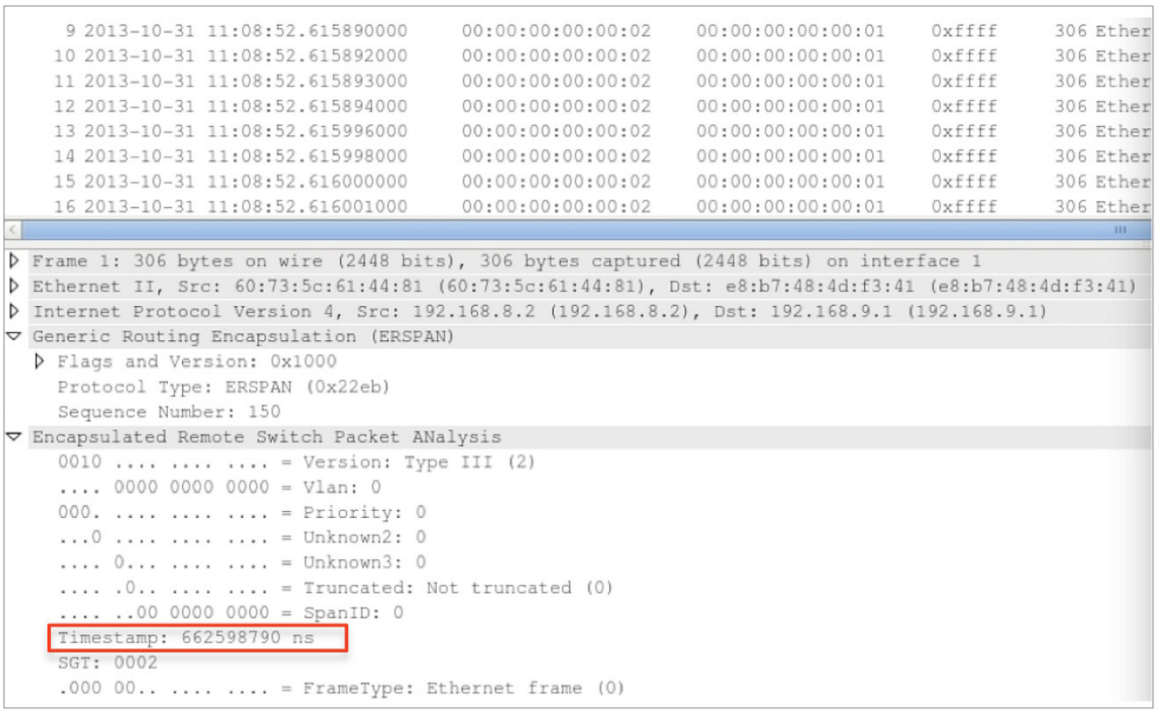
图 18. 使用 ERSPAN 和 PTP 的延迟监控



需要注意的是，未在此场景中的任何交换机上配置 ERSPAN 目标会话。ERSPAN 目标从数据包解封 GRE 和 ERSPAN 报头。因此，时间戳将会丢失。相反，会使用 Wireshark，它可以读取 ERSPAN 报头并显示时间戳。

图 19 显示了 ERSPAN 数据包的 Wireshark 捕获的示例。时间戳以红色突出显示。

图 19. Wireshark ERSPAN 捕获



Cisco Nexus 9000 系列上的 ERSPAN 支持

配备 ALE 或 ALE-2 ASIC 的 Cisco Nexus 9300 平台交换机支持 ERSPAN 类型 III 报头。从思科 NX-OS 版本 7.0(3)I1(1) 开始，这些交换机包括 Cisco Nexus 93128TX、9396PX、9396TX、9372PX、9372TX 和 9332PQ。

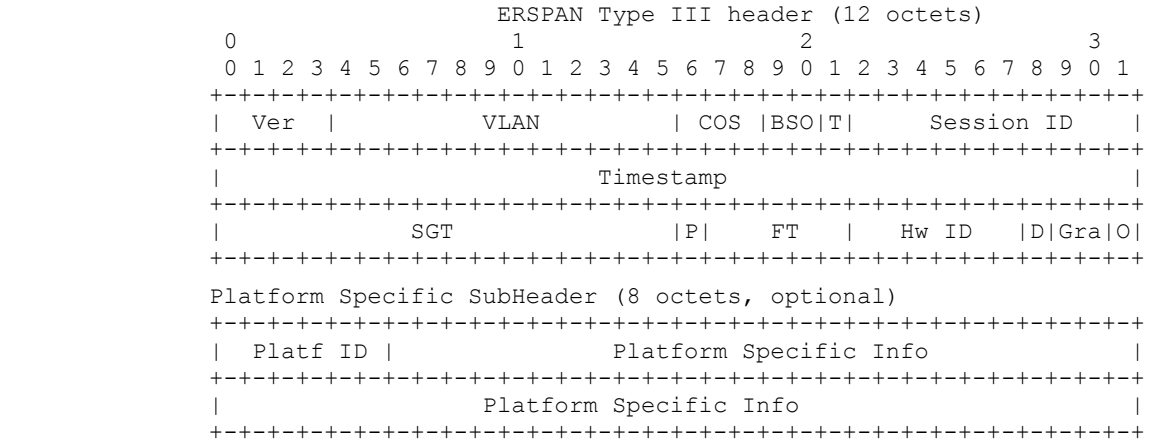
Cisco Nexus 3100 和 9500 平台不支持数据包扩展副本中的 ERSPANv3 报头，因此无法在受监控的数据包中包含时间戳。在这些交换机上，数据包封装为 GRE 隧道数据包。ERSPAN 报头不会添加到数据包。

Cisco Nexus 9000 系列上的 ERSPAN 数据包格式

ERSPAN 类型 III 数据包格式在 GRE 报头上包括额外的 12 或 20 字节报头。因此，它将总共最多 62 个字节添加到原始帧长度：14 (MAC) + 20（IPv4，IPv6 则为 40）+ 8 (GRE) + 20 (ERSPAN)。

20 字节的 ERSPAN 报头包含 32 位的时间戳字段。ERSPAN 报头在 ERSPAN 数据包的第 42 个字节开始。时间戳字段是 ERSPAN 数据包中的第 46 至 49 字节。

以下是 ERSPAN 类型 III 数据包格式的摘要：



前一格式摘要中的复合报头后面紧跟原始镜像帧，其后是标准的 4 个 8 位二进制数以太网循环冗余校验和 (CRC)。

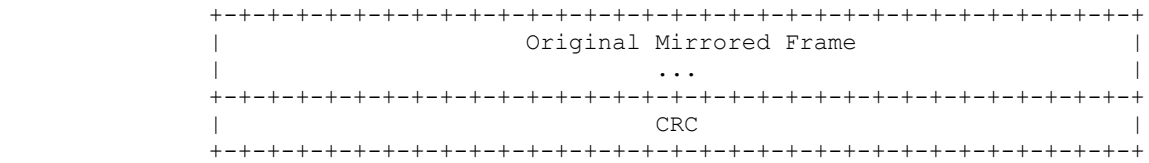


表 2 描述了 ERSPAN 类型 III 数据包报头中的每个字段。

表 2. ERSPAN 类型 III 数据包报头字段

字段	长度	定义
与 ERSPAN 类型 II 相同的报头字段		
Ver	4	ERSPAN 封装格式版本
		类型 II：0x1。类型III：0x2。版本 0x0 已过时。
VLAN	12	ERSPAN 源会话监控的帧的 VLAN。对于入口监控，这将是原始源 VLAN，对于出口监控，这将是目标 VLAN。
Cos	3	受监控帧的服务类别 (CoS)。入口或出口 CoS 值取决于监控类型。
BSO	2	该 2 位值指明 ERSPAN 传输的负载的完整性。
		00：负载是不具有错误或未知完整性的好帧。
		11：负载是具有 CRC 或对齐错误的坏帧。
		01：负载是一个短帧。
		10：负载是一个过大的帧。
T	1	表示封装的帧被截断。封装的数据包超过 ERSPAN 源会话的最大传输单位 (MTU) 设置。
Session#	10	与每个 ERSPAN 源会话相关联的标识号。标识要在目标解封的数据流。

字段	长度	定义
特定于 ERSPAN 类型 III 报头的字段		
Timestamp	32	32 位的时间戳。根据指定的时间戳精细程度，记录大约 40 亿个时间单位。
SGT	16	受监控帧的安全组标记。
P	1	此位指明负载是协议帧还是桥接协议数据单元 (BPDU) 帧。
FrameType	5	应为 00000。
Hardware ID	6	ERSPAN 引擎的唯一标识符。
Direction	1	入口或出口中受 SPAN 约束的原始帧。入口 = 0，出口 = 1。
Timestamp Granularity	2	00：精细度 = 100 毫秒（强制默认值）
		01：精细度 = 100 纳秒
		10：1588 PTP 11 = 1 纳秒
Optional SubHeader	1	**O** 标志指明可选的子报头是否存在。
		当 **O** 等于 0b ERSPAN 时，负载在 O 标志后开始。
		当 **O** 等于 1b ERSPAN 时，负载在 O 标志的 8 字节后开始。
Supplemental Info	64	可选信息。

Cisco Nexus 9000 系列上的 ERSPAN 及时间戳配置

以下网址提供的配置指南介绍了 ERSPAN 配置：
http://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/system_management/configuration/guide/b_Cisco_Nexus_9000_Series_NX-OS_System_Management_Configuration_Guide_7x/b_Cisco_Nexus_9000_Series_NX-OS_System_Management_Configuration_Guide_7x_chapter_010001.html。

ERSPAN 类型 III 报头包含 32 位时间戳。此时间戳表示自 1970 年 1 月 1 日以来的时间。启用 ERSPAN 时间戳所必需的 CLI 选项是 ERSPAN 会话配置中的 **header-type 3** 。

下面显示了具有 **header-type 3** 的 ERSPAN 配置示例：

```
switch(config)# monitor session 1 type erspan-source
switch(config-erspan-src)# vrf default
switch(config-erspan-src)# source interface Ethernet1/30 rx
switch(config-erspan-src)# destination ip 192.168.8.1
switch(config-erspan-src)# header-type 3
switch(config-erspan-src)# no shut
```

仅当通过上行链路接口（即连接到 ALE 或 ALE-2 ASIC 的接口）解析 ERSPAN 目的 IP 地址路由时，ERSPAN 类型 III 报头才添加到数据包。在 Cisco Nexus 93128TX、9396PX 和 9396TX 上，上行链路是插槽 2 上的所有接口。在 Cisco Nexus 9372PX 和 9372TX 上，上行链路是 e1/49 至 e1/54。在 Cisco Nexus 9332PQ 上，上行链路是 e1/27 至 e1/32。

ERSPAN 精细度和标记数据包

ERSPAN 时间戳的分辨率为 100 皮秒 (ps)。因此，ERSPAN 报头中的 32 位时间戳字段将每大约 0.43 秒记录一次。这不足以表示全天，它要求 64 位。因此，思科 NX-OS 提供定期发送一个标记数据包的能力，该数据包中包含完整的协调世界时 (UTC) 时间戳。ERSPAN 时间戳和合并的标记数据包值提供恢复 ERSPAN 时间戳全部值的能力。思科 NX-OS 每秒发送 10 个标记数据包。

标记数据包包含在目的端口为 8880 的 UDP 数据包中。它使用与 ERSPAN 会话相同的源和目的 IP 地址。

下面显示了具有标记数据包功能的 ERSPAN 配置示例：

```
switch(config)# monitor session 1 type erspan-source
switch(config-erspan-src)# marker-packet
```

表 3 显示了标记数据包格式。

表 3. 标记数据包格式

字段	位置（字节：位）	长度（位数）	定义
Align		16	插入 2 字节的对齐位以将数据包的其余部分与4 字节的边界对齐。该值是 0xFF，表示标记数据包的开头。
Version		4	版本号默认值为版本 1。
Type		4	标记数据包的类型。默认值为 0。
ssid		8	ERSPAN 源会话的会话 ID。
granularity		8	3 位的硬件时间戳的精细度。 该值是 0x100，表示精细度为 100 ps。
Utc_offset		8	ASIC 时钟与 CPU UTC 时钟之间的 UTC 偏移量。默认值为 0： 当前设置为 0。
timestamp		32	32 位的 ASIC 硬件时间戳。
UTC sec		32	来自 Cisco Nexus 9000 系列交换机的 CPU 时钟的 UTC 时间戳的高 32 位。此值是 ERSPAN 报头时间戳字段的时间恢复的基准。
UTC usec		32	来自 Cisco Nexus 9000 系列交换机的 CPU 时钟的 UTC 时间戳的低 32 位。
序列		32	标记数据包的序列号。
Reserved		32	已保留供将来使用。
Signature		32	值为 0xA5A5A5A5。

在硬件中执行 ERSPAN 时间戳设置。当 ERSPAN 监控的数据包到达 ALE 或 ALE-2 ASIC 时，会记录时间戳。

请注意，PTP 时间戳和 ERSPAN 时间戳的含义不同：

- PTP 时间戳是 PTP 控制数据包中使用的时间戳，这些数据包使 PTP 可以操作时钟并在网络上同步时钟。
- ERSPAN 时间戳是 ERSPAN 报头中的时间戳。它独立于 PTP 时间戳。ERSPAN 时间戳设置实际上可以在交换机未启用 PTP 的情况下运行。在这种情况下，ERSPAN 时间戳将来源于交换机的本地振荡器，并将不会与任何来源同步。这是仅使用一台交换机的基本延迟监控的可接受解决方案，因为它仍显示交换机中的数据包到达时间之间的增量。

结论

IEEE 1588-2008 PTPv2 为需要纳秒级或亚微秒级精确度的数据中心提供可靠、高度精确的分布式时间同步解决方案。PTP 易于实施，只需很少的管理工作，并且可以通过内置的容错机制忍受网络和时钟故障。

Cisco Nexus 3100 平台和 9000 系列交换机有一个非常精确的 PTP 实施。在具有大量同步 PTP 子时钟的情况下，Cisco Nexus 3100 平台和 9000 系列 PTP 精确度保持在 50 纳秒以下。

此解决方案可以与服务器上的软件 PTP 实施相结合，以实现微秒级的精确度。如果需要亚微秒级的精确度，也可以使用服务器上的硬件 PTP。

在Cisco Nexus 9000 系列上，PTP 与 ERSPAN 相结合，为在数据中心内执行延迟监控提供了一个非常强大且易于使用的解决方案。

相关详细信息

有关 IEEE 1588-2008 精确时间协议版本 2 的详细信息，请参阅用于网络测量和控制系统的精确时钟同步协议的 IEEE 标准，网址为：http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?reload=true&punumber=4579757。

有关 Cisco Nexus 9000 系列交换机的详细信息，请参阅位于以下网址的产品主页上的详细产品信息：http://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/index.html。

Cisco Nexus 3100 平台和 9000 系列上的 PTP 和 ERSPAN 的部署指南可在以下网站找到：

- http://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/system_management/configuration/guide/b_Cisco_Nexus_9000_Series_NX-OS_System_Management_Configuration_Guide_7x/b_Cisco_Nexus_9000_Series_NX-OS_System_Management_Configuration_Guide_7x_chapter_0100.html
- http://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/system_management/configuration/guide/b_Cisco_Nexus_9000_Series_NX-OS_System_Management_Configuration_Guide_7x/b_Cisco_Nexus_9000_Series_NX-OS_System_Management_Configuration_Guide_7x_chapter_010001.html

美洲总部
Cisco Systems, Inc.
加州圣何西

亚太地区总部
Cisco Systems (USA) Pte.Ltd.
新加坡

欧洲总部
Cisco Systems International BV
荷兰阿姆斯特丹

思科在全球设有 200 多个办事处。 地址、电话号码和传真号码均列在思科网站 **www.cisco.com/go/offices** 中。

美国印刷 C11-733921-01 05/15