IEEE 1588 Precision Time Protocol (PTP) for Mellanox Onyx® Switches

Design Guide

Rev 1.0

Mellanox Technologies
350 Oakmead Parkway Suite 100
Sunnyvale, CA 94085
U.S.A.
www.mellanox.com
Tel: (408) 970-3400
Fax: (408) 970-3403

Table of Contents

List of Figures

List of Tables

Document Revision History

Table 1. Document Revision History

Revision	Date	Description
1.0	April 21, 2020	First version of this document

About This Manual

As markets transition from bespoke timing solutions towards unified network-based approaches, the IEEE 1588 Precision Time Protocol (PTP) standard has gained significant traction across multiple industries, permitting for a model that supports different levels of precision and accuracy requirements specific to each application.

This guide describes the basis of PTP and how it can be operated in the context of Mellanox Onyx® switches.

Audience

This guide is intended for server administrators and network administrators who are familiar with PTP intending on deploying time transfer based solutions using Mellanox Onyx switches.

1 Why PTP?

Time transfer requirements are evolving towards better precision and accuracy. Network Time Protocol (NTP), which is widely deployed and can run across most infrastructures, provides accuracy within the millisecond range across a LAN. Additionally, in many cases, the NTP client runs on a non-real-time operating system which may further impact the accuracy.

IEEE 1588 PTP is designed to provide time transfer on a standard Ethernet network with a synchronization accuracy at a sub-microsecond level. By leveraging hardware time stamping and PTP-aware network devices such as boundary clocks, achieving synchronization accuracy in the sub-100-nanosecond range is possible.

2 IEEE 1588 Fundamentals

If all nodes in a network must be synchronized according to the principles defined in IEEE 1588, they need to exchange event messages periodically. PTP follows a strict Master-Slave principle for transmitting time information. The synchronization technique relies on a simple principle: The Master transmits synchronization messages (Sync_messages) to all Slave nodes within the respective network on a regular basis (typically at least once every second). The content of these messages is the current time of the Master—the point in time (labelled T_1) at which the Master begins sending the message through the physical channel. Every Slave, in turn, denotes the time at which it receives any such Sync_message on its local time scale (labelled T_2). The difference between these two timestamps is the offset between the two clocks plus the transmission delay of the message through the physical channel.

$$T_2 - T_1 = Transmission\_Delay + Clock\_Offset$$

If the Master is not able to insert a timestamp into the Sync_message with sufficient accuracy while actually sending it (for details on effects deteriorating the accuracy see below), it will merely note the time at which the packet is sent over the network by drawing a timestamp from its accurate local clock (this is referred to as a one-step mode) while actually sending such a message and later on forward this time information by means of a corresponding Follow_up_message again to all its Slaves (this is referred to as a two-step mode). It makes no difference at all for a Slave whether the Master operates in one- or two-step mode; it simply needs to retrieve T_1 from different messages. This is why support for both modes is mandatory for every Slave.

To calculate the transmission delay, the Slave performs a second-time transfer procedure by sending a Delay Request packet (Del_req_message) noting the time when the transmission over the physical medium is initiated (labelled as T_3). The Master, in turn, will record the time when it received such a packet (labelled as T_4) and will relay this data back to the querying Slave by sending a so-called Del_resp_message. This measurement cycle is continuously repeated to allow for filtering and to account for topology changes. The

difference of the two timestamps of the Del_req_message equals the clock offset minus the transmission delay:

$$T_4 - T_3 = Transmission_{Delay} - Clock_{Offset}$$

Now the Slave clock can calculate both the clock offset and the transmission delay using both timestamp differences.

3 PTP Accuracy

The overall accuracy depends on several factors, the most obvious one being the precision with which the timestamps can be taken. The use of PTP-aware network devices, help mitigate packet delay variations (PDVs) which otherwise would further impact the accuracy by introducing jitter in the transmission and timestamping of the PTP messages. Furthermore, a PTP Slave uses complex nonlinear filters within the control loop to adjust its local clock.

4 PTP Message Transport Mechanism

IEEE 1588 PTP allows for PTP messages to be transported over Ethernet_II frames, IPv4 UDP packets, or IPv6 UDP packets. However, there can only be a single transport mechanism per PTP port at any point in time, so co-existence on a PTP port of multiple transports mechanism is not possible.

5 PTP Ports vs. Physical/Logical Ports

IEEE 1588 has been specified as a highly generic time transfer protocol to be deployed on any network architectures supporting at least some flavor of a multicast messaging mechanism (allowing messages to be addressed to more than one receiver). As such, it can be mapped onto different networks using various Ethernet-based transport protocols.

In the most basic form of IEEE 1588, a PTP port is considered an entity capable of processing PTP messages. It has two distinct interfaces: one for processing general PTP messages and the other for dealing with PTP event messages (i.e. messages carrying time information).

Each PTP port runs a single instance of the PTP protocol stack using a specific transport protocol. Note that each PTP port can be mapped to a single physical port. The distinction between a physical port and a logical port becomes useful when multi-port PTP devices, such as Boundary Clocks, are considered. Each PTP port can be configured individually with respect to all PTP parameters such as message rates, PTP domains, or transport-related data. A Boundary Clock can link different PTP subnets to each other. This method facilitates

deploying PTP over multiple VLANs (Virtual Local Area Networks) without any restrictions.

Mellanox Onyx takes this abstraction layer one level further by providing granularity for physical interfaces, logical interfaces (VLAN), Link Aggregation (LAG), and Virtual Routing and Forwarding (VRF) which all support PTP. With Mellanox Onyx, an interface can have multiple PTP ports (per VLAN) while at the same time being part of a LAG that is a member of a specific VRF.

6 Best Master Clock Algorithm (BMCA)

The network autonomously selects one device to become its Grandmaster (GM). It is important to note that only one PTP device can become PTP Grandmaster at a time, while more than one PTP port may assume PTP Master role (in the case of a PTP Boundary Clock).

The selection process is governed by a series of PTP parameters describing the quality of the clock respective to the PTP port it is deriving its time information from. This data is communicated continuously through Announce messages. Every Slave, for example, knows whether the GM is deriving its time information from an external traceable time reference via a Global Navigation Satellite System (GNSS) link such as GPS, Galileo, or GLONASS.

If all PTP ports receive Announce messages at the expected rate without any changes in the parameters contained in these messages, they remain in their respective state. Nevertheless, the BMCA is executed whenever an Announce message is received. Its content is compared with the local data. If these two datasets completely match, no further action is taken.

Unless a PTP port is operating as a PTP Master, its state can change only under two conditions: if the data of the most recent Announce message differs from the previous message or if no Announce message has been received for a predefined amount of time. The latter parameter is defined by the number of consecutive missing Announce messages. Together with the Announce message rate which must be specified by the user for every PTP port, the PTP protocol stack is able to calculate the timeout period.

In most cases, the network must rely on precise absolute time. As such, the network should be provisioned with several GMs that are each linked to one or more GNSS time sources. These are configured in a way where one becomes the GM while the others switch to Passive state. If the active GM fails, the remaining GMs will actively participate in the BMCA, and one of the devices will assume the GM role. Such configurations imply that all other nodes shall not be able to assume the Master role. This can be accomplished by configuring the respective parameters in the Announce messages. In our example, this would be the Clock Class parameter. Boundary Clocks must combine the data gathered by every one of their respective ports during each BMCA round to reach a common conclusion as to which port should switch to Slave state based on the received time information, while all other ports transition to Master state providing time information.

7 PTP Clock Types

A Transparent Clock (TC) acts as normal network devices treating only PTP event messages in a special manner. A TC comprises an accurate clock allowing it to measure the time it requires to forward any given PTP event message. A timestamp is drawn from its clock upon reception of such a message and is stored locally. If the message is re-transmitted via any other port of the TC, another timestamp is drawn. The first timestamp is retrieved and the difference between the two timestamps is calculated, which equates to the residence time of the packet. This information is either inserted into a correction_field within the Sync_message (Del_req_message) or stored and inserted into the respective field of the corresponding Follow_up_message (Del_resp_message). The former method is referred to as one-step and the latter as two-step Transparent Clock.

Boundary Clocks (BC) are intended to partition time distribution within large networks effectively reducing the number of messages a single PTP Master node must process. Rather than simply forwarding PTP messages from a given Master to all ports as TCs do, Boundary Clocks terminate all incoming PTP traffic. The PTP event messages are used to synchronize a highly accurate local hardware clock of the BC to the Master attached to the respective port. Basically, a BC acts as a Slave synchronizing to the Master connected to this port. All other ports will generate Sync_messages using the time information of the local clock. To this end, each port of a Boundary Clock must be capable of acting both as a PTP Master and Slave with all ports sharing the same internal clock. One port will assume the Slave role whilst all other ports will act as PTP Masters (or passive Master, if there is already a better Master in this part of the network). Rather than assuming these roles in a predefined way by means of static configuration, the role of every port will be determined dynamically by the BC itself.

The BMCA Evaluation criteria are done in the following order, lowest value wins:

Priority 1: User-defined field that overrules all other values (use with caution!)

- Clock Class: Clock state derived from current reference clock used by the PTP node
- Clock Accuracy: Derived from the current reference clock used by the PTP node
- Clock Variance: A log scaled statistic representing jitter & wander of clock oscillator

Priority 2: User-defined field generally used to define the GM hierarchy

- Source Port ID: Derived from the MAC address, used as a tiebreaker

There is also an additional “Steps Removed” in case multiple paths to the GM cross BCs, the shortest path to the GM is preferred.

8 PTP Port State

Any PTP port can be operating in any given state at any point in time. While Master and Slave are the most common states, there is a series of them that exist and are part of the

normal operation of a PTP port from its initial initialization until it reaches a stable state. The following table highlights them.

Table 2. PTP Port Scale

Port State	Definition
Initializing	Port initializes its data sets, hardware, and communication facilities
Faulty	Fault state of the protocol, no PTP messages are sent except management messages
Disabled	No messages are on the communication path
Listening	Waiting for the announceReceiptTimeout to expire or waits to receive an Announce message from a master
Pre_Master	Behaves as a master, but no messages are sent, only management messages
Master	Port behaves as master
Passive	No messages sent except signalling or management messages
Uncalibrated	Transient state to allow initialization of synchronization servos, updating of data sets when a new master port has been selected
Slave	Synchronized to the selected master port

9 PTP Profiles

Version 2.0 of the Precision Time Protocol as published in the IEEE 1588-2008 standard has been deliberately defined as a highly generic protocol, leaving ample room to tailor it to specific requirements of different application domains, which are, more often than not, mutually exclusive. Such profiles have been defined for diverse markets covering many industries: telecom, power plants, media production, to name a few.

Among other things, a profile may be used to specify sub-ranges for all message rates, enabling PTP to be deployed on anything from low bandwidth to high performance networks without consuming unacceptably high network resources. The transport protocol (i.e. Ethernet_II frames, IPv4, IPv6, etc.) together with mandatory or suggested network structures and topologies are typically specified in a profile and its related documentation.

10 PTP Grandmaster Architecture

In order to deliver a common timing source to all devices, a Primary Reference Clock (PRC) is required. Typically, this is based on one or more Global Navigation Satellite System (GNSS) such as the United States' Global Positioning System (GPS), Russia's GLONASS, China's BeiDou Navigation Satellite System (BDS) or the European Union's Galileo and fed to the PTP Grandmasters (GM).

Today, sophisticated GMs include the following capabilities:

- Source diversity: Many modern PTP GMs are designed to support multiple GNSS sources—typically, two or three different systems simultaneously. Relying on more than

a single GNSS as a timing source reduces the risk of having all signals jammed in parallel since it would be harder for this to be accomplished. The logic implemented in the GM should compare the different signals and exclude those that may be impacted by an outage or provide distorted timing information. Ideally, three sources are used to better identify, in the case of a disruption, which signal may be "incorrect".

- Frequency diversity: GNSSs use different radio frequency bands for sending their signals. Some are reserved to specific applications (military and aeronautical for example), others are accessible to civilian services. GPS uses the L1 C/A, L2C and L5. Galileo uses the E1-I, E1-Q, E5a, E5b, E6-I, E6-Q. By using systems that are designed to work across multiple bands, signal distortion, it being from the ionosphere, troposphere or due to interferences can be reduced.
- Basic Reception filters: These are used in passive anti-jamming antennas to filter out undesirable noise and energy, including shielding for signals coming from low degrees of elevation which would therefore not be emitted from a satellite-based system.
- Controlled reception-pattern antennas (CRPAs): These are advanced, multi-element antenna solutions that protect a GNSS receiver from jamming sources by making use of spatial diversity. The satellite signals and jamming signals arrive from different directions, therefore, the beam and energy signature are different. To exploit these signal differences, the multi-element antenna model is required. The different signal sources are weighted based on phase, power, direction and other factors in order to "null" the interferences and increase the gain towards the legit GNSS sources by performing an electronic beam steering. The implementations of these techniques are down to individual vendors specific algorithms and antenna designs.

Additionally, the use of Satellite Based Augmentation System (SBAS) can provide additional precision and signal diversity to improve the principal accuracy of the PRC. One or more satellites placed in geosynchronous orbits provide information about the quality of the underlying GNSS signals such as corrections with respect to the GNSS orbits or information about ionospheric disturbances. These systems cover only specific regions of the Earth's surface such as WAAS (Wide Area Augmentation System) for North America, EGNOS (European Geostationary Navigation Overlay Service) for Europe, MSAS (Multi-functional Satellite Augmentation System) and QZSS (Quasi Zenith Satellite System) for Japan, GAGAN (GPS Aided Geo Augmented Navigation) for India, and, finally, SCDM (System for Differential Correction and Monitoring) for Russia.

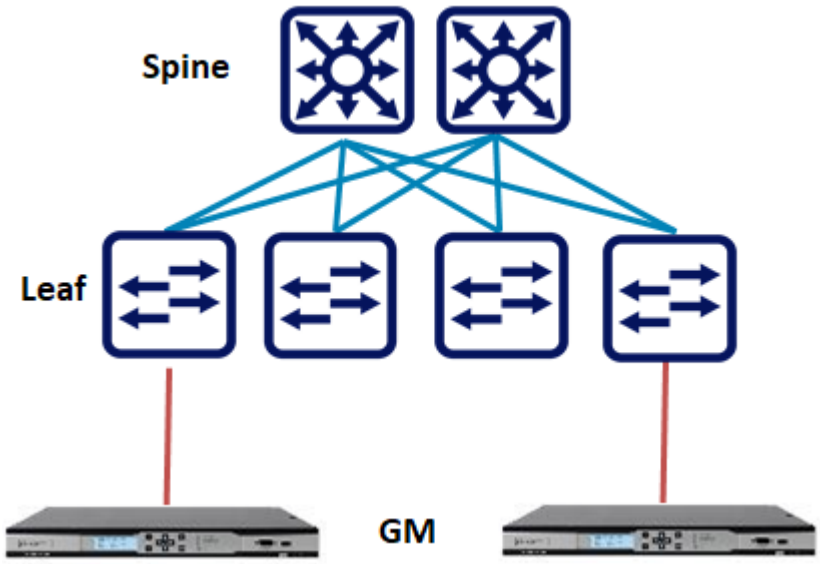
In the case of a service disruption, the GM's holdover performance will be driven by the quality of the local oscillator within the device, defining the accuracy and therefore the maximum holdover time within defined boundaries. Holdover time can range from several seconds in cases where temperature compensated crystal oscillator (TCXO) are used, all the way up to several days in the case of a Rubidium-based oscillator. The operational environment and constraints will determine the level of performance, accuracy, and holdover required.

11 PTP Network Architecture Fundamentals & Topologies

While it is usually assumed that a classic spine-leaf approach is used for distributing time transfer via PTP, some industries may have specific topology requirements. For example, Media based on port density and size of deployment may use a centralized approach (single switch) or dual network fabrics due to all flows being duplicated and sent across two separate paths. Such requirements may stem from historical reasons related to designs prior to the move to time transfer over IP or from other constraints such as the number of endpoints, port density, or environment that are applicable to the specific network deployment.

Whatever the network topology, the switches that serve as the Top of Rack (ToR), will terminate the bulk of the PTP traffic from the PTP end devices. Therefore, hardware processing of PTP at the ToR, combined with running as a Boundary Clock, will reduce the overall PTP message load by limiting PTP message propagation and processing from all PTP nodes. This allows for PTP to scale to a large number of endpoints while minimizing the impact of additional load on all the PTP end devices.

Figure 1. High-Level Network Diagram: PTP Grandmasters Connected to the Leaf/TOR



Another advantage of using PTP Boundary Clocks, when combined with a PTP profile that transmits messages over multicast (224.0.0.129 being the IANA registered address), is that there is no need to add additional multicast infrastructure such as a Rendezvous Point (RP) that would otherwise be required to allow for the multicast topology to converge and be operational.

Additionally, to further reduce the overall PTP message processing by endpoints, use "Mixed mode." In "Mixed mode" all PTP messages originating from the Master are sent as multicast, with the exception of the delay response from the Master to the Slave that is sent as unicast. The unicast response is triggered by a Slave's delay request message sent as unicast. The Master is required to respond in the same manner as the Slave sending the request. The overall number of PTP messages on the network is the same, but now none of the PTP delay request/response messages destined to another PTP device have to be processed by the PTP stack of an endpoint for which that message was not destined for before being discarded since they are not multicasted to all PTP devices listening for PTP messages. Such a case would occur when a number of PTP devices are connected to a single Master port of a Boundary Clock (directly or indirectly) by means of a Transparent Clock or a non-PTP-aware switch.

Additionally, it is worth pointing out that, since PTP message processing is performed in hardware independently of protocols in use on the interface(s), co-existence with other packet flow processing, such as OpenFlow, is supported.

12 PTP Default Message Rates and Values

As mentioned above, IEEE 1588 allows for the definition of different PTP profiles, selecting specific capabilities from the standard based on requirements defined by the entity that developed the specific profile. In Mellanox Onyx, the defaults are aligned with the Society of Motion Picture & Television Engineers (SMPTE) ST 2059-2 profile that is commonly used in the media industry.

Table 3. Default PTP Profile Attributes (SMPTE ST 2059-2)

Name	Range	Default Rate
Announce interval	-3 (0.125s), 1 (2s)	-2 (0.25s)
Announce timeout interval	2, 10	3
Sync interval (logSyncInt)	-7, -1	-3
Delay request interval	logSyncInt, logSyncInt +5	logSyncInt
PTP domain	0, 127	127
Priority 1	0, 255	128

Name	Range	Default Rate
Priority 2	0, 255	128

In addition to the values stated above, Mellanox Onyx supports both PTP over UDP using either IPv4 or IPv6 transport as per the SMPTE ST 2059-2 profile.

These values overlap with several industry profiles permitting use in different scenario and industries.

Below are the message rates for other profiles that overlap with the SMPTE-2059-2 profile include AESr16 and AES67. The values that are in the range of the SMPTE ST 2059-2 profile may be used.

Table 4. Default PTP Profile Attributes (AESr16 - SMPTE 2059-2 & AES67)

Name	Range	Default Rate
Announce interval	0 (1s), 1 (2s)	0
Announce timeout interval	2, 10	3
Sync interval (logSyncInt)	-4, -1	-3
Delay request interval	logSyncInt, logSyncInt +5	logSyncInt
PTP domain	0, 127	0
Priority 1	0, 255	128
Priority 2	0, 255	128

Table 5. Default PTP Profile Attributes (AES67)

Name	Range	Default Rate
Announce interval	0 (1s), 4	1
Announce timeout interval	2, 10	3
Sync interval (logSyncInt)	-4, 1	-3
Delay request interval	logSyncInt, logSyncInt +5	logSyncInt
PTP domain	0, 255	0
Priority 1	0, 255	128
Priority 2	0, 255	128

13 Non-PTP-Aware Devices and QoS Requirements

In cases where non-PTP-aware switches are deployed as part of the PTP infrastructure, it is strongly recommended to take additional measures to limit the impact on time transfer accuracy. Lack of hardware timestamping and specific designs within the switches for PTP

messages and logic may cause inaccuracy in the time transfer and degrade PTP accuracy and performance. This occurs when PTP packets are queued, an asymmetric load is sent across links, and other cases. As a general rule, all PTP messages should be marked for Quality of Service (QoS) with a Differentiated Services Code Point (DSCP) value of 46, meaning that these PTP messages are high priority packets and should be placed in the “Expedited Forwarding” queue. Mellanox Onyx will add this value to all PTP messages that are generated by the switch when running as a Boundary Clock.

14 PTP Redundancy

We must break down the redundancy components into separate blocks. The first has to do with rerouting traffic using well known path restoration techniques that operate either at Layer 2 and/or Layer 3 such as RSTP, OSFP, BGP, or BFD. This is purposefully designed for moving flows around that can accommodate such path rerouting. In the case of media, the SMPTE ST 2022-7 standard uses the RTP headers of the media flows that are duplicated and sent across multiple paths or fabrics to the (multicast) receiver. The first packets with the RTP sequence number of a given flow is stored and any additional copies (duplicates) that arrive over other interfaces/paths to the receiver are discarded. This scales well for media flows and is applicable to other types of flows as well.

PTP messages do not operate in the same way as other types of flows. The messages exchanged (Sync, Delay_Request, and Delay_Response) are sensitive to path asymmetries and delay. Therefore, receiving PTP messages across multiple interfaces on an end node is not simply a function of collecting these across the different interfaces and computing them as part of the PTP stack implementation. Operating separate PTP stacks per port and having a higher layer logic comparing stability and variance is a possible way to address this, but this requires specific implementations that are outside of the scope of the IEEE 1588-2008 standard and there is therefore no cross-industry standard implemented. This will either be dealt with in a future revision and/or within industry-specific implementations.

The other area of importance is GM redundancy. As explained above, the GM itself can be designed to be redundant. From a network perspective, BMCA will be performed with all devices that transmit Announce messages, compare the datasets, and converge to a single GM. The placement of the candidate GMs on the network is either done at the Spine or Leaf. Transmission time across high-speed fabrics does not impact the location choice, nor do modern PTP-aware switch implementations. We typically recommend placing the GM candidates at the Leaf with the other host devices, so that Spine ports are kept free for additional Leaf devices and that these high-speed interfaces are not used with low speed devices (as GMs typically operate at 1G or 10G rates). When connecting the candidate GMs to the Leaf, it is recommended to place them as far apart as possible across the network fabric. This is done to ensure the highest possible physical redundancy, combined with separate antennas and cabling paths to minimize risk.

15 PTP Scalability

In order to accommodate user requirements across a wide range of use cases, the Mellanox Spectrum® switch family has been tested to support up to 1500 PTP slaves operating at the SMPTE ST 2059-2 profile default rates. This ensures that a large number of PTP messages from a number of slaves can be processed across many Ethernet ports.

16 Security

Make sure that all PTP messages are originating from a reliable and traceable source (i.e. the Primary Reference Clock). This is to ensure that the time source is referenceable and traceable back to the frequency and time origination point. By doing so, the PTP messages will contain the Frequency-traceable and Time-Source-traceable flags.

Additionally, it is important that all devices are configured to only accept PTP Announce messages from validated sources. This is done by verifying the ClockID, which is a 64-bit-long value uniquely identifying a GM. The ClockID is derived from the MAC address of the network interface of the PTP port on the GM, or candidate GM that is sending the PTP Announce messages. By filtering these using a feature from the IEEE 1588 standard known as Acceptable Master Table (AMT), all devices that are deriving their time from the GM (i.e. PTP Ordinary Clocks running in a Slave state) can filter out PTP messages that originate from a GM that is not whitelisted via the AMT.

```
switch (config) # ptp amt <Clock ID>
```

In addition to the whitelisting of the valid GM sources, a Boundary-Clock-specific security feature for preventing directly connected devices from being recognized as a potential GM can be applied. This feature ensures the PTP port on which this function is enabled does not switch to Slave state. It typically does so by discarding all Announce Messages it receives, thereby ensuring that a misconfigured or rogue device does not become an authoritative source of time for the Boundary Clock and for the entire network attached to it.

```
switch (config) # interface ethernet x/y ptp enable [ipv6] forced-master
```

In the case where IPv6 is being used as the PTP message transport, the use of Link-Local Addresses (LLA) assigned to a network interface that is only reachable on the local link permits for the use of locally scoped, non-routable addresses. The multicast PTP messages are also locally scoped, therefore, all the PTP traffic is contained within the Layer 2 domain—typically between a Boundary Clock port and a host.

```
switch (config) # interface ethernet x/y ptp enable ipv6 mcast-scope link-
local
```

17 Distributing PTP to Remote Locations

In some cases, there may be a need to distribute PTP over remote infrastructure if used within a metro area or beyond. This could result from building/location constraints at the remote location such as roof access limitations for antennas for the GNSS reception, a lack of sky visibility, or other technical reasons.

Whatever the reason for extending the PTP infrastructure beyond the local network fabric, there are a few key points to keep in mind. The accuracy and stability of PTP is dependent on PTP-aware devices that perform hardware timestamping in order to meet the requirements. As such, networks where the transport is built using an overlay model such as MPLS, VxLAN or other techniques that obfuscate the physical layer from the transport layer, will prevent hardware timestamping at each physical node along the transport path. Therefore, PTP messages carried across these links will not be time stamped as they would on the local network infrastructure.

In order to ensure that PTP accuracy is maintained over the remote connections, this will require access to the Layer 1 transport such as optical wavelengths or dark fibre, so that the active network equipment at each hop along the path is PTP-aware and can perform hardware timestamping.

If running PTP over extended distances, the use of extensive linear and non-linear filtering combined with high message rates and very long settling times is required as an additional means to counteract asymmetries that are common in such environments to reach a high degree of accuracy. Therefore, the use of specific PTP profiles and/or parameters for the long haul vs. the network fabric may be required.

Nevertheless, if roof access is available for GNSS equipment, this should not be a concern in many cases since, as explained above. If the performance targets permit, common and traceable primary reference clocks should allow for the use of a GM per location while still maintaining accurate and stable timing.

18 Troubleshooting

In cases where the overall PTP system may appear to produce unexpected results, a troubleshooting methodology to identify the causes of the possible disruption is typically the best path to detecting a fault. This effort can be broken down into the following steps:

1. Physical/Logical Interface Status
2. PTP Domain Number
3. Message Rates Consistency

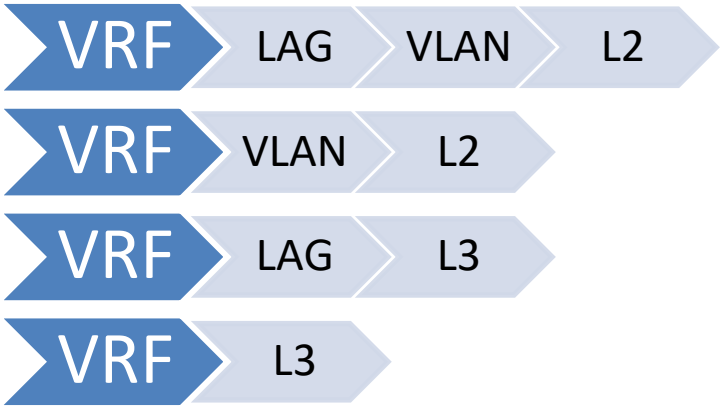
4. PTP Message Counters
5. Announce Message Dataset
6. Slave-Only Flag
7. Forced-Master Log Output
8. AMT Log Output
9. PTP Port State Changes
10. Mean Path Delay Variations
11. Offset from Master Variations
12. Multiple Ports Upstream to the Next Hop

18.1 Physical/Logical Interface Status

The PTP port depends on the underlying physical and, if applicable logical interface(s) being enabled and properly configured. “ptp enable [ipv6]” must be applied to the physical interface (switchport or routed), the corresponding VLAN interface(s) if applicable, the LAG port if applicable and the VRF if not running in the default VRF.

The sum of this logic enables PTP and enables the extended flexibility that the Mellanox Onyx PTP implementation delivers in terms of specifically selecting which (sub) interface(s) are PTP-enabled.

Figure 2. “Nested” PTP Logic for Configuring Interfaces



18.2 PTP Domain Number

All PTP devices are part of a single PTP domain, the default domain ID in Mellanox Onyx is 127, as defined in SMPTE ST 2059-2. Any PTP port that is not a part of the same domain as the configured domain ID will discard the received PTP messages.

18.3 Message Rates Consistency

Within a PTP domain, all PTP message rates should be consistent. If there are discrepancies between PTP devices, the affected PTP ports may find themselves in a form of loop whereby the PTP port state is constantly changing due to announce timeouts for example. Accuracy may also be impacted by having different sync and delay rates configured amongst devices.

Mellanox Onyx reports on Announce and Sync interval rates between what is configured on the Mellanox Onyx Slave port and what is received from the Master port.

```
Nov 19 17:13:27 switch <snip>: [702.322] PTP [Debuggability]: Matched Sync
interval on Eth1/15. Configured -3, Received -3
Nov 19 17:13:27 switch <snip>:: [pm.NOTICE]:<snip>: PTP [Debuggability]:
Matched Sync interval on Eth1/15. Configured -3, Received -3
Nov 19 17:13:27 switch <snip>: [702.322] PTP [Debuggability]: Matched
Announce interval on Eth1/15. Configured -2, Received -2
Nov 19 17:13:27 switch pm[3436]: [pm.NOTICE]: <snip>: PTP [Debuggability]:
Matched Announce interval on Eth1/15. Configured -2, Received -2
```

18.4 PTP Message Counters

To track PTP message communication on any given PTP enabled port, message counters per

PTP message type help further identify discrepancies in message rates between PTP clocks.

For example, identifying that message Delay requests and responses are symmetrical (1:1 ratio) provides guidance about the ongoing message exchange. The example below is for a port running as a Slave (TX Delay_Req, RX Delay_Resp).

```
switch (config)# show ptp interface ethernet 1/15 counters
Eth1/15
RX
3165      Sync message count
0         Delay request message count
0         PDelay request message count
0         PDelay response message count
0         Follow Up message count
3170      Delay response message count
0         PDelay response follow Up message count
1583      Announce message count
0         Signalling message count
396       Management message count

TX
0         Sync message count
3170      Delay request message count
0         PDelay request message count
0         PDelay response message count
0         Follow Up message count
0         Delay response message count
0         PDelay response follow Up message count
0         Announce message count
0         Signalling message count
3         Management message count
0         Forwarded Management message count
```

18.5 Announce Message Dataset

A packet-capture tool, such as PTP Track Hound or Wireshark, can be used for verification to see if the received PTP Announce dataset matches what is expected. This is another means

for ensuring that the BMCA related values are set correctly (Priority 1, Clock Class, Clock Accuracy, Clock Variance, Priority 2). This may occur due to device misconfiguring or erroneous dataset generated by a PTP node.

18.6 Slave-Only Flag

In most cases, PTP devices that do not require to be elected as a Master should have the "Slave only" flag enabled in their configuration. This helps reduce the risk of a misconfigured device from being elected as a Master.

18.7 Forced-Master Log Output

The output of the PTP forced-master logging feature provides visibility into directly connected misconfigured PTP devices connected to the interface that would otherwise trigger a BMCA due to the announce messages generated by those PTP devices.

```
switch (config)# show ptp forced-master log
-------------------------------------------------------------------------------
Clock Identity          Interface VLAN  IP Address      Last Occurrence
-------------------------------------------------------------------------------
00:1E:C0:FF:FE:85:BB:DB Eth1/16   N/A   192.168.211.11  2019/11/19 09:37:48
```

18.8 AMT Log Output

The output of the PTP AMT logging feature provides visibility into misconfigured PTP devices connected to any interface of the switch that would otherwise trigger a BMCA due to the announce messages.

```
switch (config)# show ptp amt log
-------------------------------------------------------------------------------
Clock Identity           Interface VLAN IP Address       Last Occurrence
-------------------------------------------------------------------------------
08:00:11:FF:FE:21:E4:46  Eth1/18   N/A  192.168.12.11    2019/11/19 09:50:05
```

18.9 PTP Port State Changes

PTP port state changes are logged to Mellanox Onyx. This helps identify transient, as well as permanent, port state changes that may locally or globally impact the stability of the PTP infrastructure.

```
Nov 19 12:30:34 switch pm[3435]: [pm.NOTICE]: <snip>: PTP [Debuggability]:
PTP Grandmaster clock has changed from ec0d9a.fffe.fde548 to
080011.fffe.21e446

Nov 19 12:30:34 switch pm[3435]: [pm.NOTICE]: <snip>: port 4: Interface
Eth1/18 state changed from UNCALIBRATED to SLAVE on MASTER_CLOCK_SELECTED
```

18.10 Mean Path Delay Variations

The mean path delay between the Slave port and the upstream Master port should, under normal operation, be stable and only vary within a small range of nanoseconds. This value is partly dependent on the interface speed and in the case of non-PTP-aware infrastructure in the path. Network load will additionally impact this value due to jitter introduced along the path. In Mellanox Onyx, a max mean path delay threshold value in nanoseconds can be set so that it logs each PTP message calculation that crosses the defined value.

```
switch (config) # ptp mean-path-delay <value>
```

18.11 Offset from Master Variations

Under normal operating conditions, the Slave port offset from the master should also be stable and vary within a small range of nanoseconds. In Mellanox Onyx, a max/min offset threshold value in nanoseconds can be set so that it logs each PTP message calculation that results in a value that is outside the user-defined boundaries. It can be defined using the following syntax:

```
switch (config) # ptp offset-from-master <value> <value>
```

The last measured values can be viewed in the CLI via the “show ptp status” command.

```
switch (config)# show ptp status

PTP mode                          : Boundary Clock
PTP Offset Threshold (ns)         : -100000, 100000
PTP Mean Path Delay Threshold (ns): 1000000000

------------------------------------------------------------------------------
Interface  Time                      Offset from Master (ns)  Mean Path
Delay (ns)
------------------------------------------------------------------------------
Eth1/13    2019/12/02 10:42:08.913   -1                       155
Eth1/13    2019/12/02 10:42:08.788   -29                      155
Eth1/13    2019/12/02 10:42:08.663   14                       155
Eth1/13    2019/12/02 10:42:08.538   -9                       157
Eth1/13    2019/12/02 10:42:08.413   -34                      157
Eth1/13    2019/12/02 10:42:08.288   3                        156
Eth1/13    2019/12/02 10:42:08.163   -24                      156
Eth1/13    2019/12/02 10:42:08.038   14                       150
Eth1/13    2019/12/02 10:42:07.913   -9                       156
Eth1/13    2019/12/02 10:42:07.788   32                       156
Eth1/13    2019/12/02 10:42:07.663   11                       159
Eth1/13    2019/12/02 10:42:07.538   -26                      159
Eth1/13    2019/12/02 10:42:07.413   18                       160
Eth1/13    2019/12/02 10:42:07.288   -6                       160
Eth1/13    2019/12/02 10:42:07.163   -35                      161
Eth1/13    2019/12/02 10:42:07.038   16                       161
Eth1/13    2019/12/02 10:42:06.913   -22                      161
Eth1/13    2019/12/02 10:42:06.788   27                       155
```

18.12 Multiple Ports Upstream to the Next Hop

In the case where there are multiple PTP ports on the Boundary Clock that are connected (either directly connected or are multiple hops away) upstream towards the Grandmaster independently of the transport (i.e. Layer 3, Trunk, LAG, etc.), only one port on the BC should be in a Slave state. The BMCA rules described previously apply here, including the “Steps Removed” field so that shortest path (from a PTP hop count perspective) to the GM should be used. Any other upstream ports towards the GM should be running in Passive state.

19 Conclusion

IEEE 1588 PTP provides accurate time transfer capabilities for a number of use cases. Each case is unique in terms of target accuracy and precision. These targets drive the definition of PTP profiles in multiple industries to match the required performance. The PTP implementation on Mellanox Spectrum switches running Mellanox Onyx accommodate for a number of these requirements, both in terms of flexibility through support for PTP across VLAN, Routed, LAG, and VRF interfaces as well as scalability via many slaves per interface at high message rates whilst maintaining accuracy. Network designs requiring the use of PTP should take into consideration scale, topology, PTP message rates, and security requirements when defining their architecture and selecting PTP-aware switches.

20 References

- IEEE 1588-2008, IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems https://doi.org/10.1109/IEEESTD.2008.4579760
- SMPTE ST 2059-2:2015 - SMPTE Profile for Use of IEEE-1588 Precision Time Protocol in Professional Broadcast Applications https://doi.org/10.5594/SMPTE.ST2059-2.2015
- AES-R16-2016: AES Standards Report - PTP parameters for AES67 and SMPTE ST 2059-2 interoperability http://www.aes.org/publications/standards/search.cfm?docID=105
- AES67-2018: AES standard for audio applications of networks - High-performance streaming audio-over-IP interoperability http://www.aes.org/publications/standards/search.cfm?docID=96
- EBU Technical Review: Using PTP for Time & Frequency in Broadcast Applications - Part 1: Introduction https://tech.ebu.ch/files/live/sites/tech/files/shared/techreview/trev_2018-Q2_PTP_in_Broadcasting_Part_1.pdf
- EBU Technical Review: Using PTP for Time & Frequency in Broadcast Applications Part 2: PTP Clock Characteristics https://tech.ebu.ch/files/live/sites/tech/files/shared/techreview/trev_2019-Q4_PTP_in_Broadcasting_Part_2.pdf
- EBU Technical Review: Using PTP for Time & Frequency in Broadcast Applications Part 3: Network design for PTP https://tech.ebu.ch/docs/techreview/trev_2019-Q4_PTP_in_Broadcasting_Part_3.pdf
- Mellanox Onyx User Manual https://docs.mellanox.com/category/onyx
- Mellanox Community: IEEE 1588 PTP on Spectrum Switches Running Onyx https://community.mellanox.com/s/article/ieee-1588-ptp-on-spectrum-switches-running-onyx
- Meinberg PTP Track Hound https://www.ptptrackhound.com/