OPEN POSSIBILITIES.

# Challenges and solutions in PTP Based Time Sync in Hyper-Scale Data Centers : A study from the network's point of view

OCP GLOBAL SUMMIT

NOVEMBER 9-10, 2021

TIME APPLIANCES

# Challenges and solutions in PTP Based Time Sync in Hyper-Scale Data Centers : A study from the network's point of view

Ahmad Byagowi, Research Scientist, Facebook
Amit Oren, Distinguished Engineer, Broadcom
Bhaskar Chinni, Principal Product Line Manager, Broadcom

OPEN PLATINUM™

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

# Agenda

TIME APPLIANCES

- Time synchronization problem is data center networks
- Introduction - Boundary Clock (BC) & Transparent Clock (TC)
- BC, TC deployment scenarios

# Why Time Synchronization in DC Networks?

TIME APPLIANCES

- Tighter distributed clock skewness means faster throughput in distributed databases
- Synchronization across different services
- Easier troubleshooting with precisely timestamped telemetry & logs
- Improved efficiency of reading and writing in warm storage
- SLA/SLO monitoring for high availability
- Advanced precision time-based encryptions and security methods

# Transparent Clock

TIME APPLIANCES

- Transparent Clock from standards perspective
  - End to End Delay , Peer to Peer Delay
  - 1-step, 2-Step
- Transparent Clock Use cases
  - Telecom Transparent Clock (T-TC)
  - Industrial Automation
  - Similarities to 802.1AS Time Aware Bridge
- Syntonization and Processing Latency
- Low Latency Implementation
- Scalability in Deployment
  - TC vs BC

# Transparent Clock & Boundary Clock

- Transparent clock (TC) is defined in the standards as a device that measures the time taken for a PTP event message to transit the device (a.k.a. Residence Time) and provides this information to clocks receiving this PTP event message.

- Unlike a Boundary Clock (BC) which terminates the PTP protocol on its slave port and originates the PTP protocol on its master ports, the TC is a transit device, and does not terminate the PTP protocol, although it modifies PTP protocol headers.

# Transparent Clock in the Standards

TIME APPLIANCES

- Four variations of TC: 1-Step/2-Step, Peer Delay / E2E Delay
- In 1-Step mode, the transit delay (residence time) is communicated to downstream devices by modifying the Correction Field in the Event Message PTP header. Stateless in the sense that does not require correlating the Event message and Followup message
- In 2-Step mode, the transit delay (residence time) is communicated to downstream devices by modifying the Correction Field in the FOLLOWUP Message

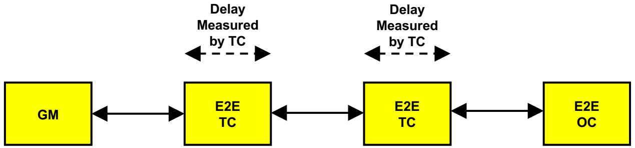
| Timestamping Mode | Delay Mode | What is Measured | Comment |
|---|---|---|---|
| 1-Step | E2E | Residence time | Simplest to implement, stateless |
| 1-Step | P2P | Residence time and incoming link delay | Offers deployment scalability, as scalable as a BC |
| 2-Step | E2E | Residence time | Most complex to implement |
| 2-Step | P2P | Residence time and incoming link delay | |

# Transparent Clock in the Standards

TIME APPLIANCES

- E2E delay measurement requires a packet exchange between the PTP Slave and PTP Master. TCs measure the delay across intermediate nodes, but not the links.

- P2P delay measurement includes intermediate nodes and links. Requires packet exchange to peer node.



Delay Measured by TC

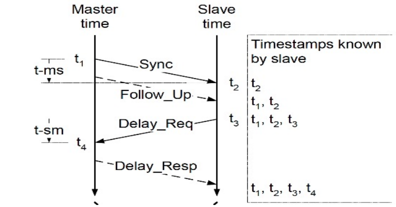
Delay Measured by TC

GM

E2E TC

E2E TC

E2E OC



Master time

Slave time

Timestamps known by slave

t-ms

t-sm

$t_1$

Sync

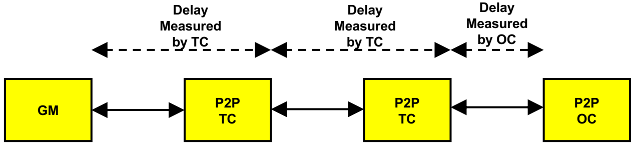
$t_2$ | $t_2$

Follow_Up | $t_1$, $t_2$

Delay_Req | $t_3$ | $t_1$, $t_2$, $t_3$

$t_4$

Delay_Resp | $t_1$, $t_2$, $t_3$, $t_4$



Delay Measured by TC

Delay Measured by TC

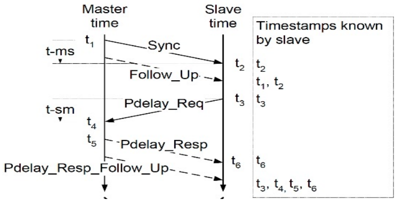
Delay Measured by OC

GM

P2P TC

P2P TC

P2P OC



Master time

Slave time

Timestamps known by slave

t-ms

t-sm

$t_1$

Sync

$t_2$ | $t_2$

Follow_Up | $t_1$, $t_2$

Pdelay_Req | $t_3$ | $t_3$

$t_4$

$t_5$

Pdelay_Resp | $t_6$ | $t_6$

Pdelay_Resp_Follow_Up | $t_3$, $t_4$, $t_5$, $t_6$

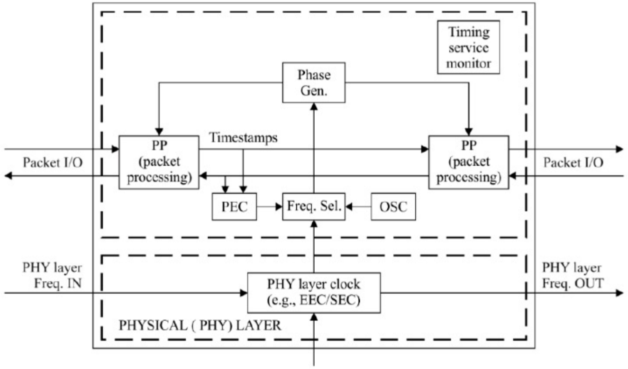
OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT

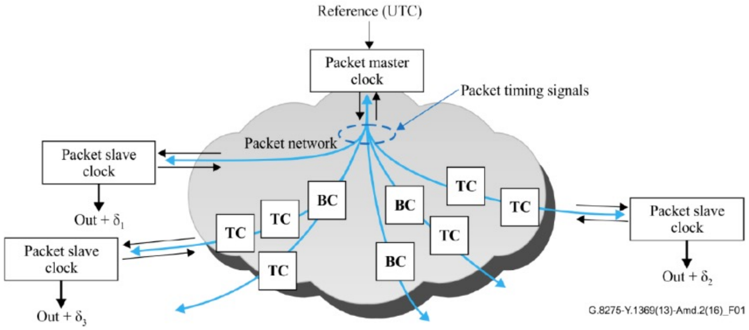
NOVEMBER 9-10, 2021

# Transparent Clock Use Cases

- Telecom Transparent Clock (T-TC) standardized by ITU
- Defined in G.8271, G.8275, G.8273.3
- Operate in E2E mode only
- Uses Synchronous Ethernet (physical layer clock) for syntonization in Full Timing Support (FTS) mode.
- Example for usage – A microwave backhaul link configured as a T-TC between the two link endpoints.



G.8273.3  Clock Model

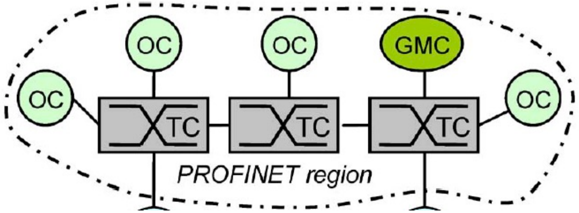


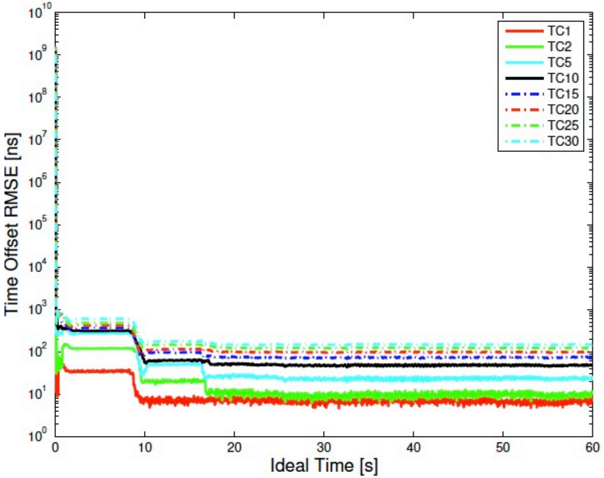
G.8275  Deployment Scenarios

# Transparent Clock Use Cases

TIME APPLIANCES

- Industrial Automation – Profinet
- Ethernet based industrial automation protocol
- Uses PTP for synchronization, TC+OC
- Characterized by long chains (30 or more) of TCs
- Published performance studies indicate about 200nSec RMS time error for 30 hops
- Long chains of TCs do not exhibit accumulated Gain Peaking as do chains of BCs





OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

# Transparent Clock Use Cases

- Time Sensitive Networking (TSN) – IEEE 802.1as
- Defines ”Time-Aware” systems (bridges, routers etc.)
- Mandatory 2-Step, exhibiting very significant processing latencies
- Similar to P2P TCs in the sense that they measure link delay and residence time.
- Additionally implement “Logical Syntonization”, measuring fractional frequency offset to peer nodes, to mitigate effects of processing latencies.
- Usage example: Automotive entertainment and video systems

# Syntonization and Processing Latency

TIME APPLIANCES

- TCs measure residence time by timestamping the PTP Event Message at the ingress and egress ports of the node.
- The measurement relies on a local oscillator that in most cases is not syntonized to the PTP Master frequency.
  - An exception: The Syntonized T-TC, that uses Synchronous Ethernet
- The local oscillator typically exhibits fractional frequency offsets (FFO) ranging from tens of ppb to tens of ppm
- The measurement error of a time interval T due to oscillator skew is equal to T x FFO. E.g., the measurement error of a 1msec interval using an oscillator with 100ppm offset is 100nsec. The measurement error of a 1msec interval using common TCXOs with 4.6ppm offset is 4.6nsec.
- **To avoid the need for syntonization, it is important to reduce residence time and/or use more accurate oscillators**
- Reduction of residence time can be achieved primarily by using 1-Step (no software intervention in forwarding of Event Messages), and higher CoS for PTP Event Messages
- Example – IEEE 802.1as implementations that modify the SYNC messages in software and as a result introduce significant processing delays (>1msec is not uncommon) , require the use of logical syntonization to deal with the measurement errors.

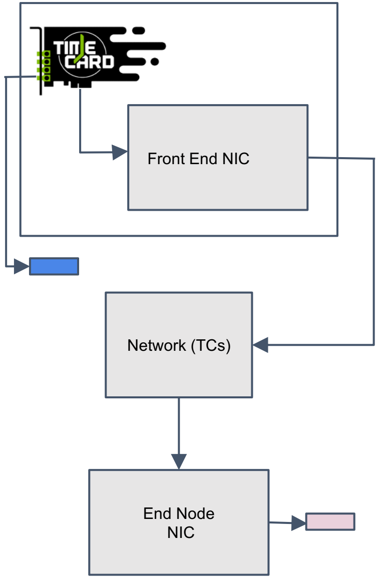
# Low Latency TC Implementation

TIME APPLIANCES

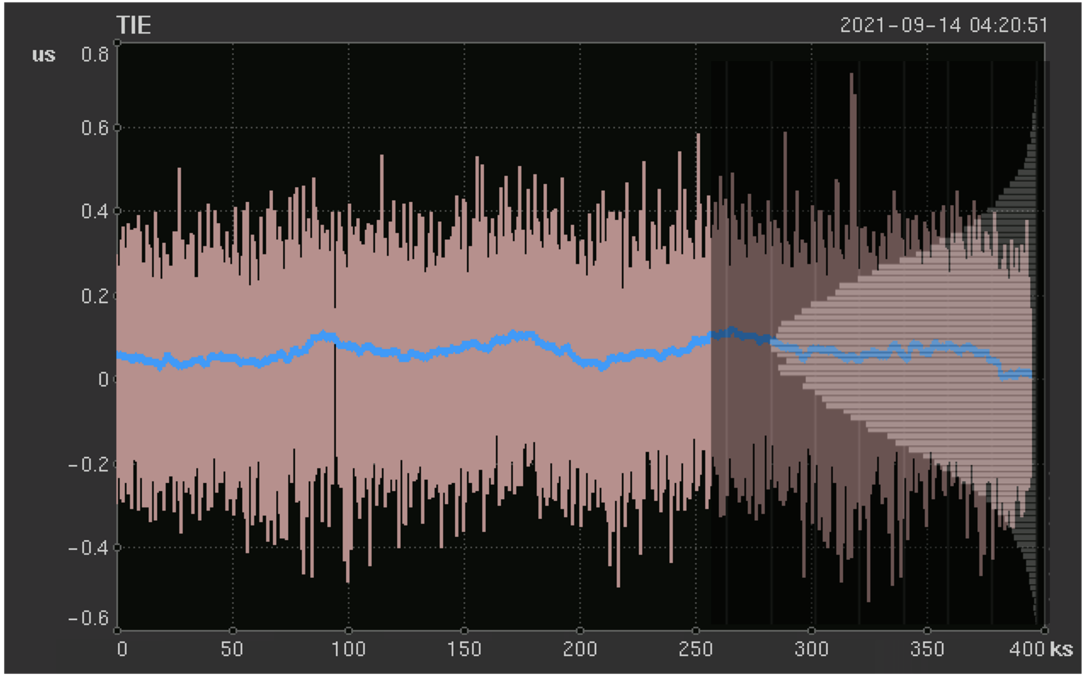
- Given the importance of a low latency implementation, Switch nodes can include hardware support for PTP that reduces latency:
- Hardware update of Correction Field of PTP Event Messages in both unicast and multicast forwarding
- Per port hardware registers for link delay (to assist P2P delay)
- Classification of PTP Event Messages into high CoS queues

# Open Time Server. LAB

TIME APPLIANCES



Ideally!



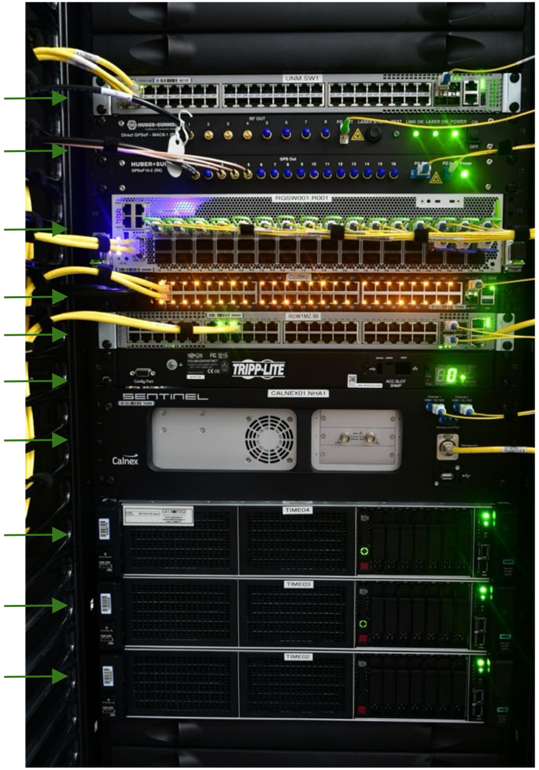
STD[E2E]≃90ns

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

# Open Time Server. NHA1

TIME APPLIANCES



“Dumb” switch

Huber Suhner GPS-over-fiber

RGSW

Console Server

Management SW

Automatic Transfer Switch

Calnex Sentinel

Time Appliance

Time Appliance

Time Appliance

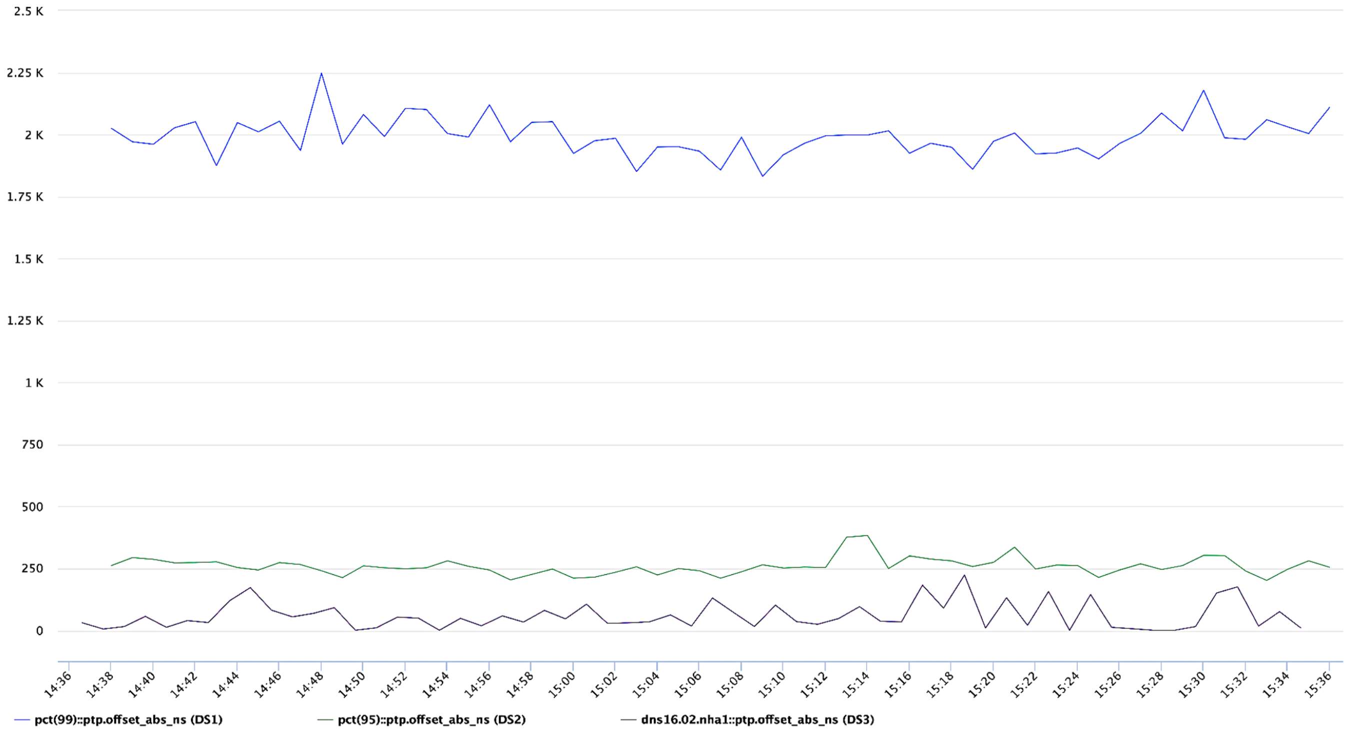
NHA1 PTP Rack

HPE Raiser designed for Time Appliance

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

# Open Time Server



PTP Precision NHA1 clients (ns)
https://fburl.com/ods/4zoz3req

TIME APPLIANCES

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

# Call to Action

- Join us on
  https://www.opencompute.org/wiki/Time_Appliances_Project