Precision Time Protocol Profile for Data Center Applications & Related Network Requirements
PTP Profile for DC Applications & Related Network Requirements

Michel (Meta) and Thomas (NVIDIA)
Agenda

Why Synchronization?
Application & Time Error Requirements
Time Synchronization Service
PTP Profile Aspects
Call to Action
Why Synchronization in Data Centers

- Provide a reliable time synchronization service across the infra of a data center
- Enable set of new applications
- Improve set of current applications
- Using Precision Timing Protocol (PTP), increase the level of accuracy by 2 to 3 orders of magnitude beyond what NTP infra offers today

Nanosecond-level clock synchronization can be an enabler of a new spectrum of timing- and delay-critical applications in data centers
- Stanford / Google paper
OCP-TAP Applications

- Applications discussed within OCP-TAP:
  [https://www.opencompute.org/wiki/Time_Appliances_Project](https://www.opencompute.org/wiki/Time_Appliances_Project)

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributed database systems</td>
<td>- Increase system transactions via stringent time bound guarantees</td>
</tr>
<tr>
<td></td>
<td>- Reduce guard band between read-write operations across machines</td>
</tr>
<tr>
<td>Network monitoring</td>
<td>- Early detection of anomalies (e.g., latency, loss, congestion)</td>
</tr>
<tr>
<td></td>
<td>- One-way delay (OWD) measurements between any given machine</td>
</tr>
<tr>
<td>Telco Cloud Radio Access Network</td>
<td>- 5G air interface synchronization</td>
</tr>
<tr>
<td></td>
<td>- Time distribution across midhaul/fronthaul networks</td>
</tr>
<tr>
<td>Enterprise</td>
<td>- Cloud &amp; enterprise applications</td>
</tr>
<tr>
<td></td>
<td>- Financial services</td>
</tr>
</tbody>
</table>
Distributed Database - Example

- If $t_2 < t_1$ due clock skew, application will see wrong information.
- Ordering of operations is necessary, but not always sufficient.
- Ordering in time leads to improve performance but requires strict clock skew guarantees between machines (e.g., to enable property of linearizability).
Time Synchronization Service & PTP Profile

• Significant advances in time distribution & synchronization across packet networks in the past decade
• Multiple industries rely on high accuracy & reliable time distribution
• Many products and networks around the world run PTP today
• IEEE 1588 (PTP) is a large specification with many capabilities and options to choose from
• A “PTP Profile” defines the capabilities that are required to support a particular use case scenario

A "PTP Profile" is an essential part of a "Time Synchronization Service"
# PTP Profiles in the Industry

<table>
<thead>
<tr>
<th>Industry</th>
<th>Application</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telecom &amp; Mobile</td>
<td>Sync for 2G/3G/4G/5G base stations &amp; fronthaul networks</td>
<td>ITU-T G.8265.1&lt;br&gt;ITU-T G.8275.1, G.8275.2</td>
</tr>
<tr>
<td>Professional Audio/Video</td>
<td>Sync for video/audio feeds between sources and receivers</td>
<td>SMPTE ST 2059-2</td>
</tr>
<tr>
<td>Audio/Video, Industrial, Automation, Automotive</td>
<td>Sync of A/V applications with high QoS/QoE demand and time sensitive networks</td>
<td>IEEE Std 802.1AS-2020</td>
</tr>
<tr>
<td>Industrial Automation</td>
<td>Sync for industrial plants, machine-to-machine real-time control</td>
<td>IEC 62439-3 Annex B&lt;br&gt;IEC 62439-3 Annex C</td>
</tr>
<tr>
<td>Enterprise/Financial</td>
<td>Sync of time tagged and packet latency measurements</td>
<td>draft-ietf-tictoc-ptp-enterprise-profile-21</td>
</tr>
<tr>
<td>Data Center</td>
<td>Sync for time-sensitive applications within data center</td>
<td>OCP DC PTP Profile #1&lt;br&gt;(released Sept 2021)</td>
</tr>
</tbody>
</table>
“Time Sync Service” Reference Model

Time Reference Layer:
- Rootftop antennas, GPS system
- Open Time Server (OTS)

Network fabric Layer:
- Large set of PTP-aware switches
- e.g., Transparent Clock (TC)

Server Layer:
- Very large set of server machines
- End applications requiring time
- HW timestamping

OPEN POSSIBILITIE
“Time Sync Service” Error Requirements
- End Application & Infra

- $|T_{OTS,j} - T_{AI}| \leq 100$ ns
- $|T_{OTS,i} - T_{OC,k}| \leq 2.4$ $\mu$s for $k \neq j$
- $|T_{OC,j} - T_{AI}| \leq 2.5$ $\mu$s
- $|T_{TC,j,k}| \leq 200$ ns
- $|T_{GM,j} - T_{GM,k}| \leq 200$ ns for $k \neq j$
- $|T_{OC,j} - T_{OC,k}| \leq 5$ $\mu$s for $k \neq j$
Software vs. Hardware Timestamping

Software timestamping don’t provide a high accuracy and deterministic behaviour (10 to 100 microseconds) due to system noise, latency, scheduling

Hardware timestamping pulls timestamps as close as possible to the MAC with minimal overhead (sub 10ns in modern implementations)

SW timestamping: TS, Clock & PTP

HW timestamping: TS & PHC vs. PTP
1-step vs. 2-step clock

- Historically (1588v1 era), 2-step SW implementations were available (HW limitations/availability)
- 2-step doesn’t write timestamps to PTP messages as they egress. Older HW couldn’t encode 1-step timestamp messages at higher interface rates (10 Gbps+), this isn’t the case anymore
- 1-step guarantees each sync message is correctly linked up with its timestamp. Especially when there may be multiple possible network routes (Sync vs. Follow_Up) (OoO, packet spraying across links)

OPEN POSSIBILITIES.
Switches & capabilities – Model 1 & 2

**Model 1: Transparent Clock**
- E2E PTP msgs sprayed across links
- 1-step HW TC to avoid OoO msgs
- Network routing for failure recovery

**Model 2: Boundary Clock (proposed)**
- PTP messages processed hop-by-hop
- BC is SW based and handles:
  - PTP dataset analysis
  - BMCA for failure/path recovery
- BC chain may impact overall settling time
- 1-step or 2-step clock may be used
Failure/redundancy mechanism

- An OC Follower can have PTP sessions configured to multiple GMs (Open Time Server)
- Over IPv6 Unicast transport in Model 1 (E2E Transparent Clocks)
- PTP Stack running on the OC compares PTP Dataset via BMCA
- Select ‘Best GM’ based on P1, ClockClass, ClockAccuracy, ClockVariance, P2, SrcID
- Network reachability with candidate GMs relies on IGP
Data Center PTP Profile
Attributes, Parameterization &
Configuration

...putting the end-to-end PTP pieces together to meet performance objectives

<table>
<thead>
<tr>
<th>PTP Attributes</th>
<th>OCP Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCP Company ID (CID)</td>
<td>7A-4D-2F</td>
</tr>
<tr>
<td>Timing awareness</td>
<td>PTP aware in all elements (switch, NICs, open time server, etc.)</td>
</tr>
<tr>
<td>PTP Clock types</td>
<td>GM, TC, OC</td>
</tr>
<tr>
<td>PTP Messages</td>
<td>Announce, Sync, Follow_Up, Delay_Req, Delay_Resp, Signaling, Management</td>
</tr>
<tr>
<td>Network transport</td>
<td>IPv6, IPv4</td>
</tr>
<tr>
<td>PTP path delay measurement</td>
<td>End-to-End</td>
</tr>
<tr>
<td>Domain Number</td>
<td>0</td>
</tr>
<tr>
<td>Clock Operations</td>
<td>One-step and Two-step for GM, OC One-step for TC</td>
</tr>
<tr>
<td>PTP Message rates</td>
<td>Announce {0, -4} Sync {+3, -7} DelayReq/Resp {0, -7}</td>
</tr>
<tr>
<td>Network Communication</td>
<td>Unicast discovery &amp; Unicast negotiation Multicast is prohibited</td>
</tr>
<tr>
<td>Redundancy</td>
<td>Active-Standby Active-Active (future)</td>
</tr>
<tr>
<td>Open Time Server Clock Class</td>
<td>6 (traceable) 7 (holdover, within spec) 52 (holdover, out of spec)</td>
</tr>
<tr>
<td>Open Time Server Clock Accuracy</td>
<td>0x21 (±100 nanoseconds)</td>
</tr>
</tbody>
</table>
Call to Action

• Data Center PTP Profile v1 was released on Sept 21
• Get involved in OCP-TAP workstream #2 to further develop PTP Profile v2
  • Reference Model with Boundary Clock
  • Security aspects
  • Load balancing of PTP unicast sessions
• Project Wiki: [https://www.opencompute.org/wiki/Time_Appliances_Project](https://www.opencompute.org/wiki/Time_Appliances_Project)
• DC PTP Profile v1: [https://github.com/opencomputeproject/Time-Appliance-Project/tree/master/DC-PTP-Profile](https://github.com/opencomputeproject/Time-Appliance-Project/tree/master/DC-PTP-Profile)
Open Discussion