OPEN POSSIBILITIES.

Boundary Clocks vs Transparent Clocks in Hyperscale Network

NOVEMBER 9-10, 2021

TIME APPLIANCES

Boundary Clocks vs Transparent Clocks in Hyperscale Network

Ahmad Byagowi, Research Scientist, Meta
Rohit Puri, Software Engineer, Meta
Dotan Levi, Nvidia

OPEN PLATINUM™

OPEN POSSIBILITIES.

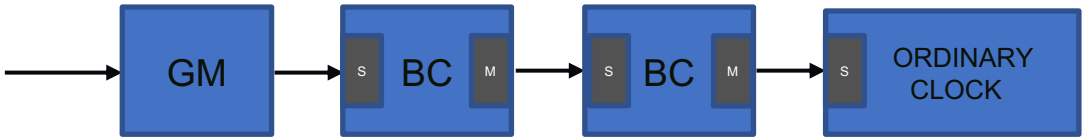
OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

Agenda

- Pros and cons of Boundary and Transparent Clock Deployments in data center
- FBOSS PTP Deployment considerations
- PTP TC Scaling Challenges

Boundary vs Transparent Clocks

BOUNDARY CLOCK

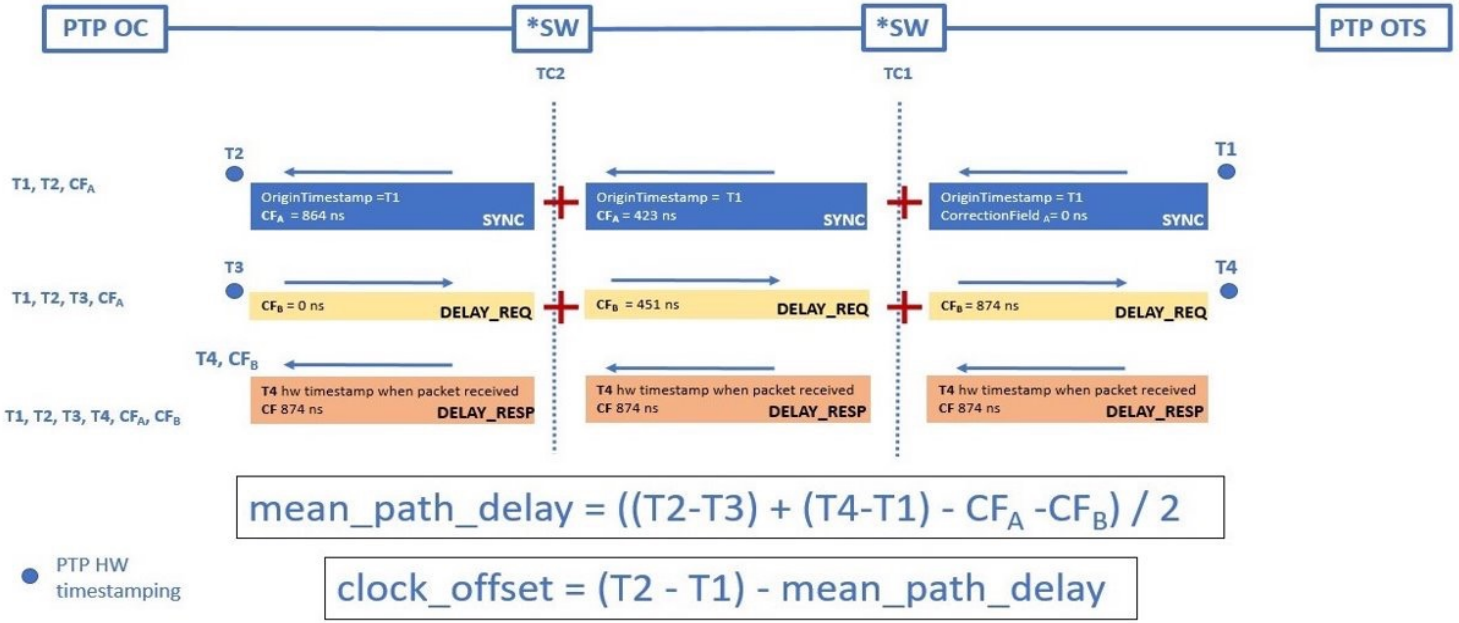


•When considering scalable PTP deployment for data centers, BCs are often viewed as scalable building blocks since they can reduce workload from a PTP GM and distribute the master workload among the BCs. A server OC will conduct Delay Measurement to a nearest BC instead of the GM.

•Point to point synchronization.

•Reduces packet load on GM and scales well as every node acts like master and terminates packets.

Boundary vs Transparent Clocks



$$mean\_path\_delay = ((T2-T3) + (T4-T1) - CF_A - CF_B) / 2$$

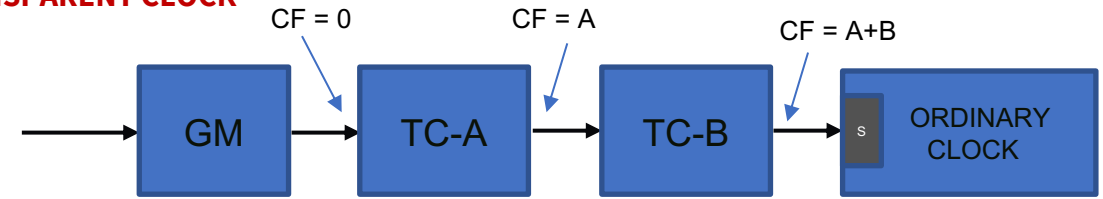
$$clock\_offset = (T2 - T1) - mean\_path\_delay$$

TIME APPLIANCES

Boundary vs Transparent Clocks Continued ..

TIME APPLIANCES

TRANSPARENT CLOCK



- End-to-End synchronization.
- TCs are often viewed as less scalable since a server OC will conduct Delay Measurement to the GM. Every OC needs to talk to GM.
- TCs that do not need to recover time, can use less expensive oscillators, provided they are low latency TCs.
- Ideal for deployment in heterogeneous environment with different network HW capabilities.
- PTP TC is significantly easier to implement in SW and deploy in the network !

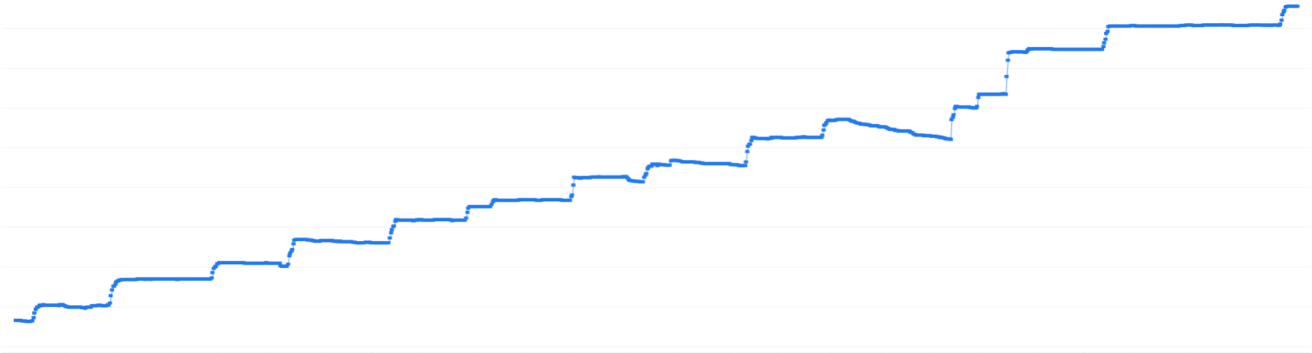
PTP TC Deployment in Meta DC

- FBOSS supports PTP TC in E2E mode. Uses underlying HW timestamping to enable the feature.
- Intermediate nodes in the network are ***not*** required to support PTP TC making deployments simpler.

  One of the DC was safely and gradually upgraded to run PTP TC in under 3 months!
- PTP TC provides clock accuracy which meets our application requirements. 95th percentile accuracy of 400nsecs.

Why PTP TC in Meta DC ?

- 100% of fabric switches in given a given DC has PTP TC enabled across different switch roles (TORs, Spine)
- IPv6 only network

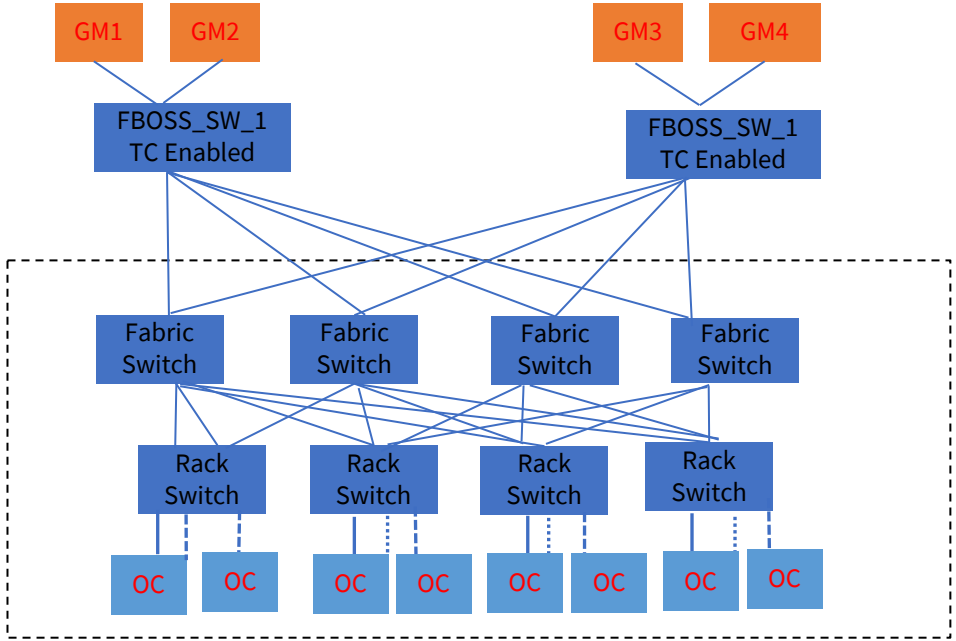


Scaling challenges with PTP TC in Meta DC

TIME APPLIANCES

- One GM cannot scale for the entire DC. So many sessions need to be created per Ordinary Clock.
- Redundancy needed in the DC for clients to move to another clock source if original time source goes down.
- Network should continue to operate unaffected even if we lose 75% of PTP time sources.
- Reliability. There cannot be a single point of failure.
- No multicast in the network.

Scaling challenges Continued ..



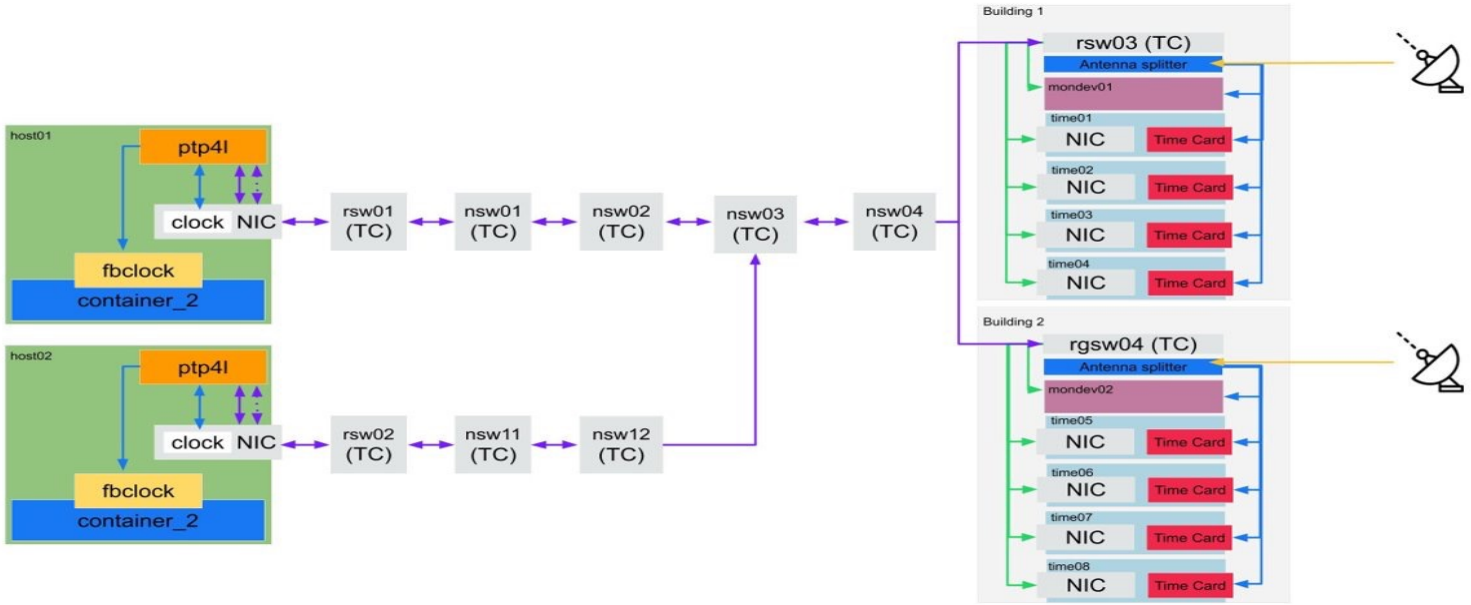
TIME
APPLIANCES

Scaling challenges Continued ..

TIME APPLIANCES

- Improvements in the PTP server (ptp4u) are in works to handle large number of client requests. We are synchronizing ~75K clients per server which are generating 300k requests per second.
- Able to scale to 1M clients per server !
- Effectively 1 GM shown in previous slide can handle the load for all OCs.
- Bigger DCs can have ~500k clients. This will require 16-24 appliances in the given region.

PTP TC in Meta DataCenter



TIME APPLIANCES

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

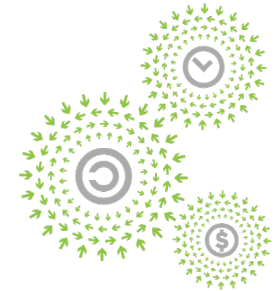
Call to Action

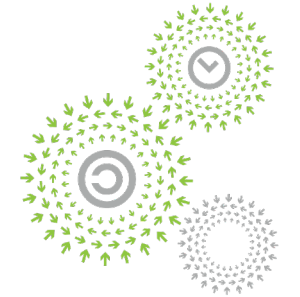
- Join us on
  https://www.opencompute.org/wiki/Time_Appliances_Project

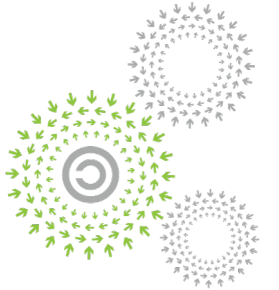
Thank you!

NOVEMBER 9-10, 2021

Please use one of these membership logos to designate your company's membership level.







OPEN POSSIBILITIES.

Please use this logo if you or your supplier is an OCP Solution Provider.

OPEN
Compute Project
SOLUTION PROVIDER®

OPEN
Compute Project
COLO SOLUTION PROVIDER™

OPEN POSSIBILITIES.

Please use this logo if your Facility is an OCP Ready™ facility

Please use if your Product has been recognized as an OCP certified product

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

Track Names

CE (Cooling Environments)
DCF
HW Mgmt
Networking
OSF
R&P (Rack & Power)
Security
Server
Storage
SI (Strategic Initiatives)
TAP
T&E (Telco & Edge)

Please use the appropriate icon representing the Project Group

DATA CENTER FACILITIES

HW MANAGEMENT

NETWORKING

OPEN SYSTEM FIRMWARE

RACK & POWER

SECURITY

SERVER

STORAGE

TELCO

TIME APPLIANCES

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT

NOVEMBER 9-10, 2021

Please use the appropriate icon representing the Sub-Project Group

ADVANCED COOLING FACILITIES

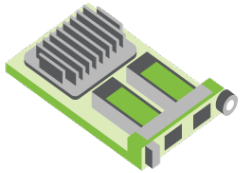
ADVANCED COOLING SOLUTIONS

EDGE

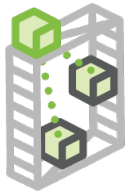
HIGH PERFORMANCE COMPUTING

HW FAULT MGMT

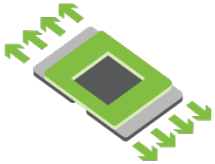
MODULAR DATA CENTER



NIC3.0



OPENRMC



OPEN ACCELERATOR INFRASTRUCTURE

OPEN DOMAIN SPECIFIC ARCHITECTURE

SUSTAINABILITY

onie

Open Network Linux

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT

NOVEMBER 9-10, 2021

Please use the appropriate icon representing the Regional Project Group

Scalable PTP TC Deployment

TIME APPLIANCES

- TCs are often viewed as less scalable since a server OC will conduct Delay Measurement to the GM. This, however, is only true of E2E TCs.
- P2P TCs are as scalable as BCs, since:
- A server OC will conduct P2P Delay Measurement to the nearest TC, not to the GM
- A TC will conduct P2P Delay measurements on each port, 1/sec.
- A downstream SYNC message can be multicasted and delay adjusted across the TC tree to OC endpoints, i.e., no need for unicast SYNC.
- The GM responds to delay measurements to directly attached TCs only
- The TC+OC model can be used on switches that need to recover time for e.g., Telemetry.
- TCs that do not need to recover time, can use less expensive oscillators, provided they are low latency TCs.