Time Transfer requirements: network & data center

Thomas Kernen, Principal Architect | SwiNOG-38, June 21st 2023

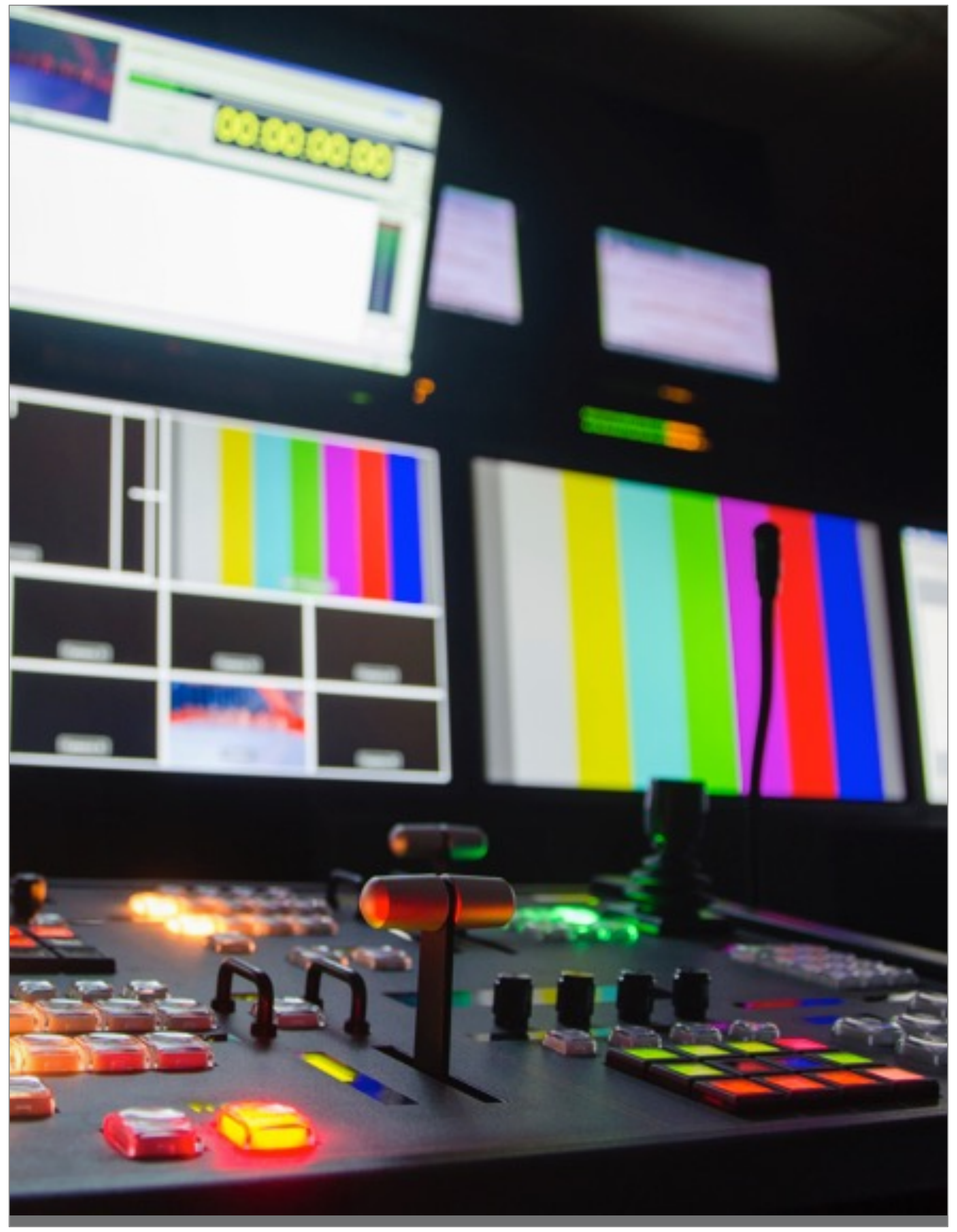
Introduction

- What this presentation is not about – previous topics:
  - Peering policies - SwiNOG #1
  - Designing and deploying a VoIP network - SwiNOG #5
  - Metro Ethernet - SwiNOG #8
  - IPTV/Video over Broadband - SwiNOG #12
  - Video for network engineers: what is relevant to you? - SwiNOG #17
  - 2000-2010: How the Internet has evolved - SwiNOG #20
  - Automatic Multicast without Explicit Tunnels (AMT) - SwiNOG #22
- Today is about:
  - “High precision” Time transfer across networks
  - IEEE 1588 Precision Time Protocol
  - Living in a nanosecond scale world

Agenda

- Use cases for Timing in the Data Center
- Timing 101
- OCP-TAP DC PTP Profile

Media



Telco

Finance

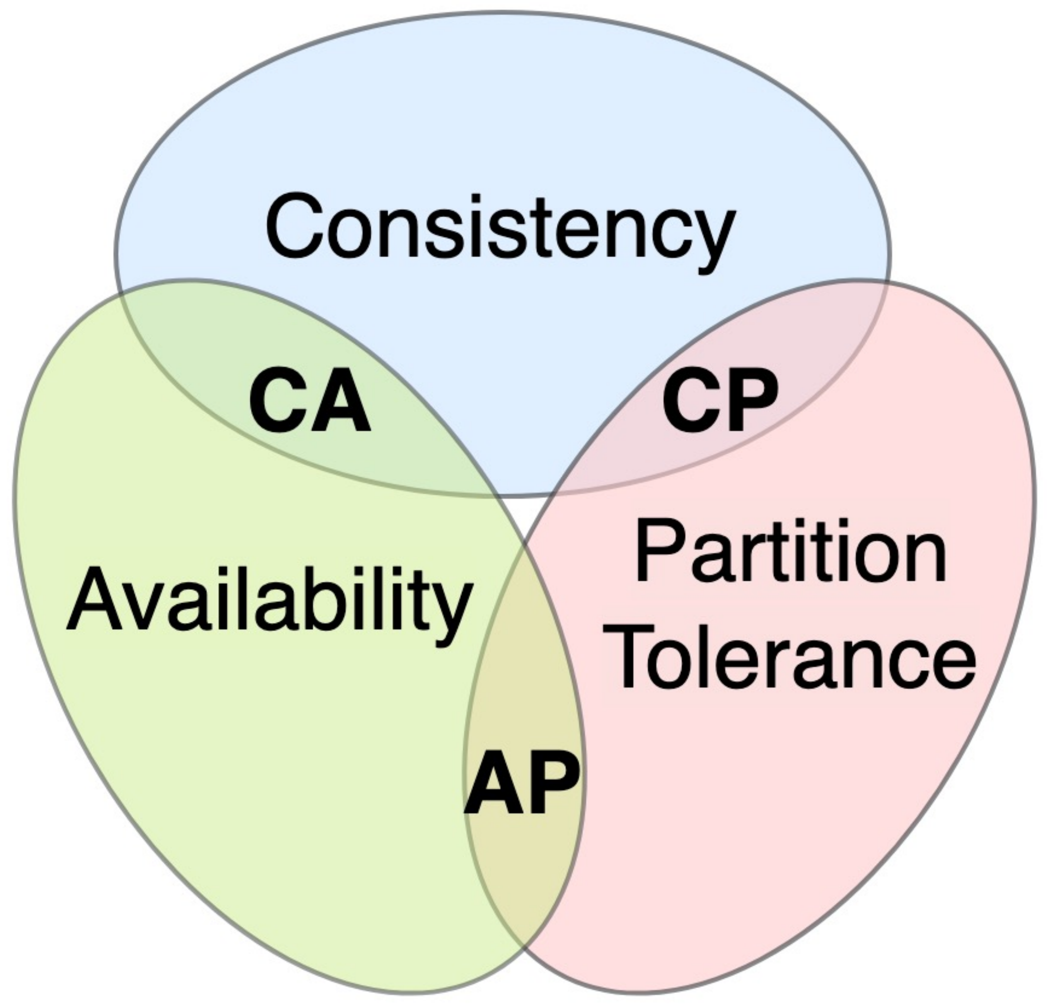
Data Center

Industry specific requirements

> Nanosecond-level clock synchronization can be an enabler of a new spectrum of timing- and delay-critical applications in data centers
>
> — Yilong Geng & *All 2018*

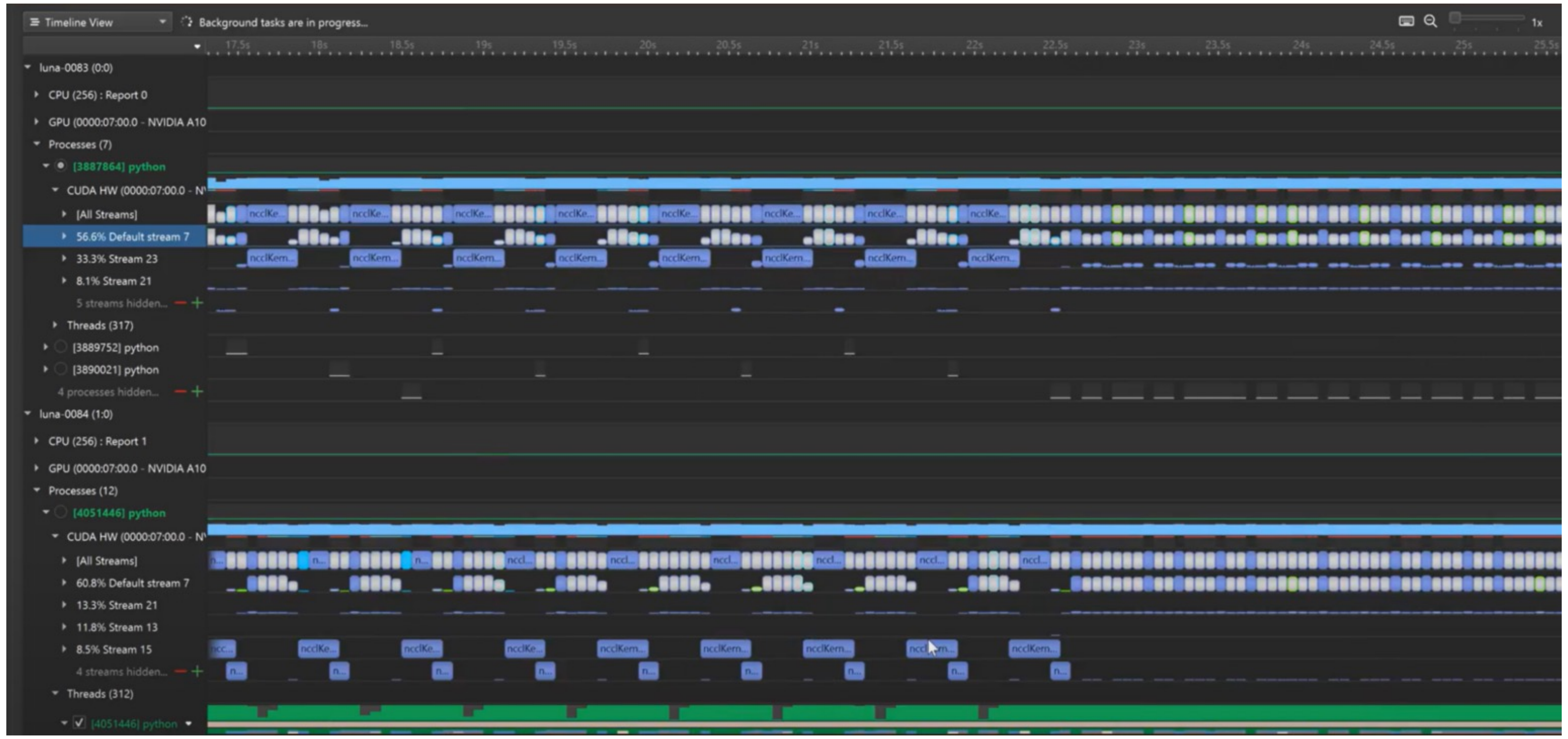
Why Synchronization in Data Centers?

- 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
- Spotlight case: Google Spanner, TrueTime and the CAP Theorem
  - Highly available global-scale distributed database. It provides strong consistency for all transactions. This combination of availability and consistency over the wide area is generally considered impossible due to the CAP Theorem.



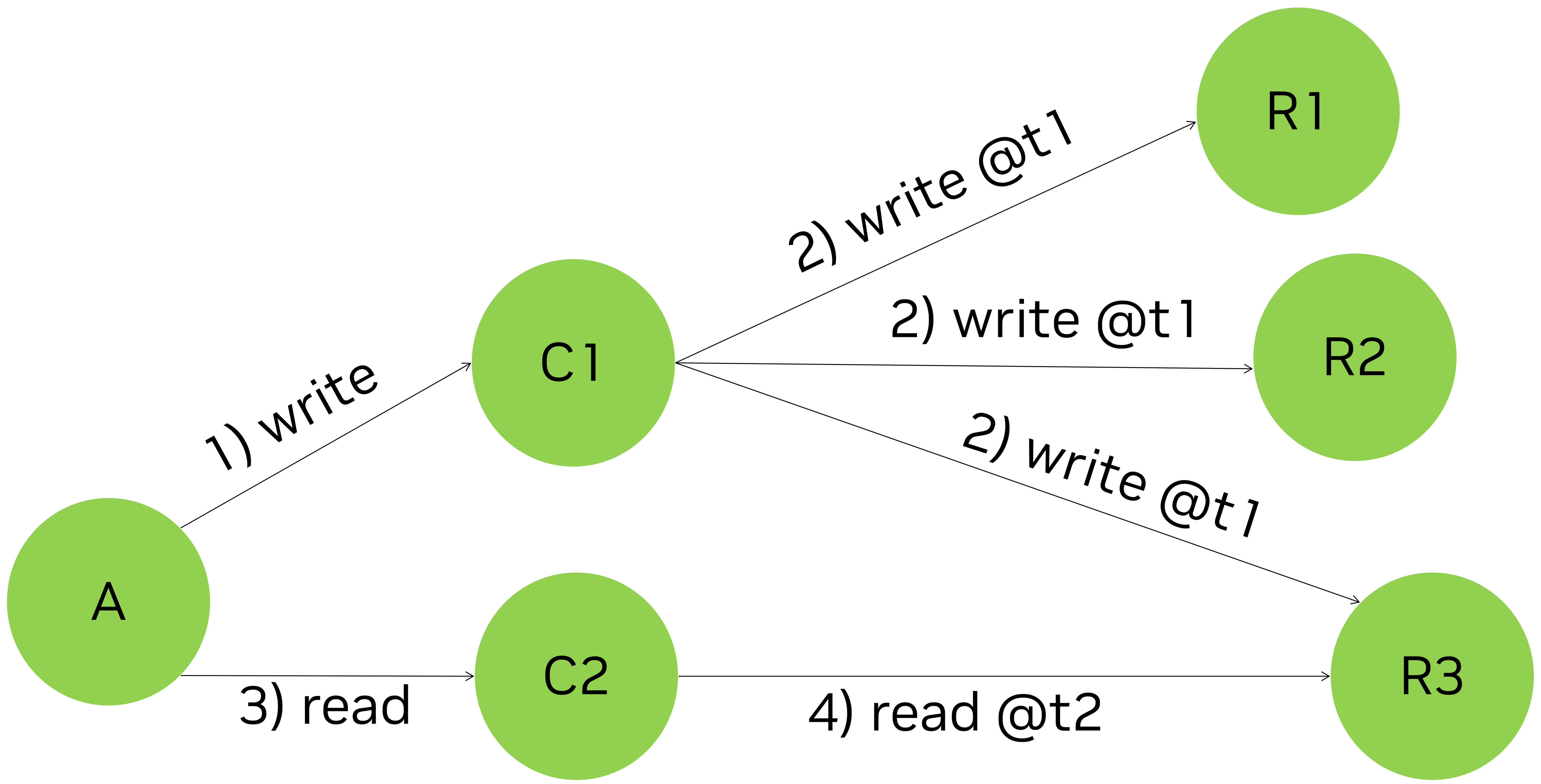
Use cases

- Distributed databases
- One Way Delay (OWD) Measurement
- Network & host based telemetry
  - Microscopic view of bursts, buffer contention, and loss (Millisampler/Syncmillisampler)
- System-Wide Performance Analysis (Nsight Systems)
  - Root cause analysis
  - CPU, GPU interactions and activity
  - Multi-node systems
  - Interrupts, wait states
- Security

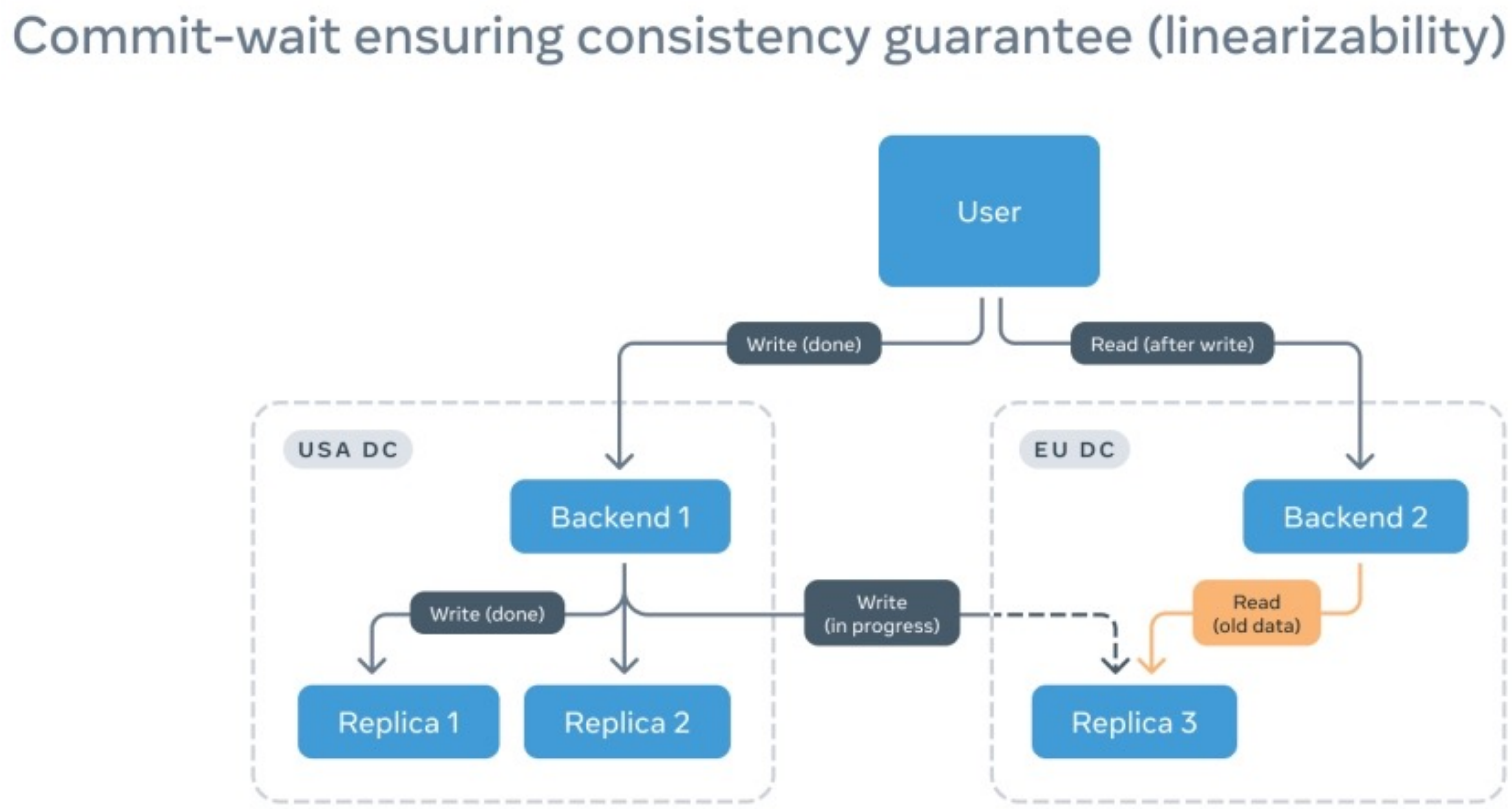


Distributed Database

- Needed to guarantee if a transaction is committed at time T1 (e.g., write operation) before another transaction T2 (e.g., read operation), committed timestamp of T1 is before the committed timestamp of T2 when compared with real-time.
- Aligning the clocks across all nodes in the distributed system ensures that they all display the same time for a given level of accuracy thereby defining a window of time uncertainty (ε)
- Ordering of operations is necessary, but not always sufficient
- Strict serializability (two-phase commit)
- Ordering in time leads to improve performance but requires strict clock skew guarantees between machines (e.g., to enable property of linearizability)

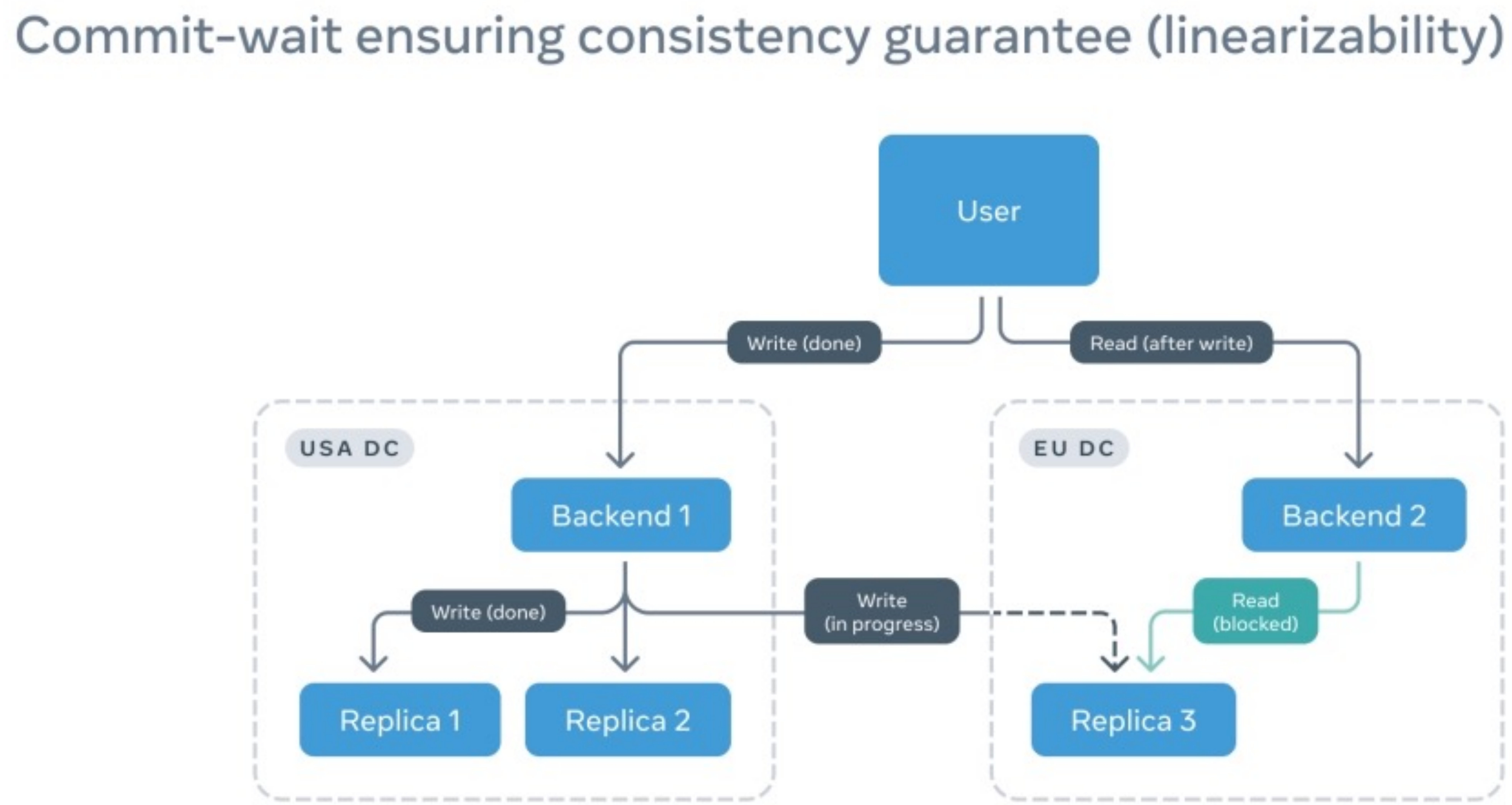


Schematic representation of read returning outdated information



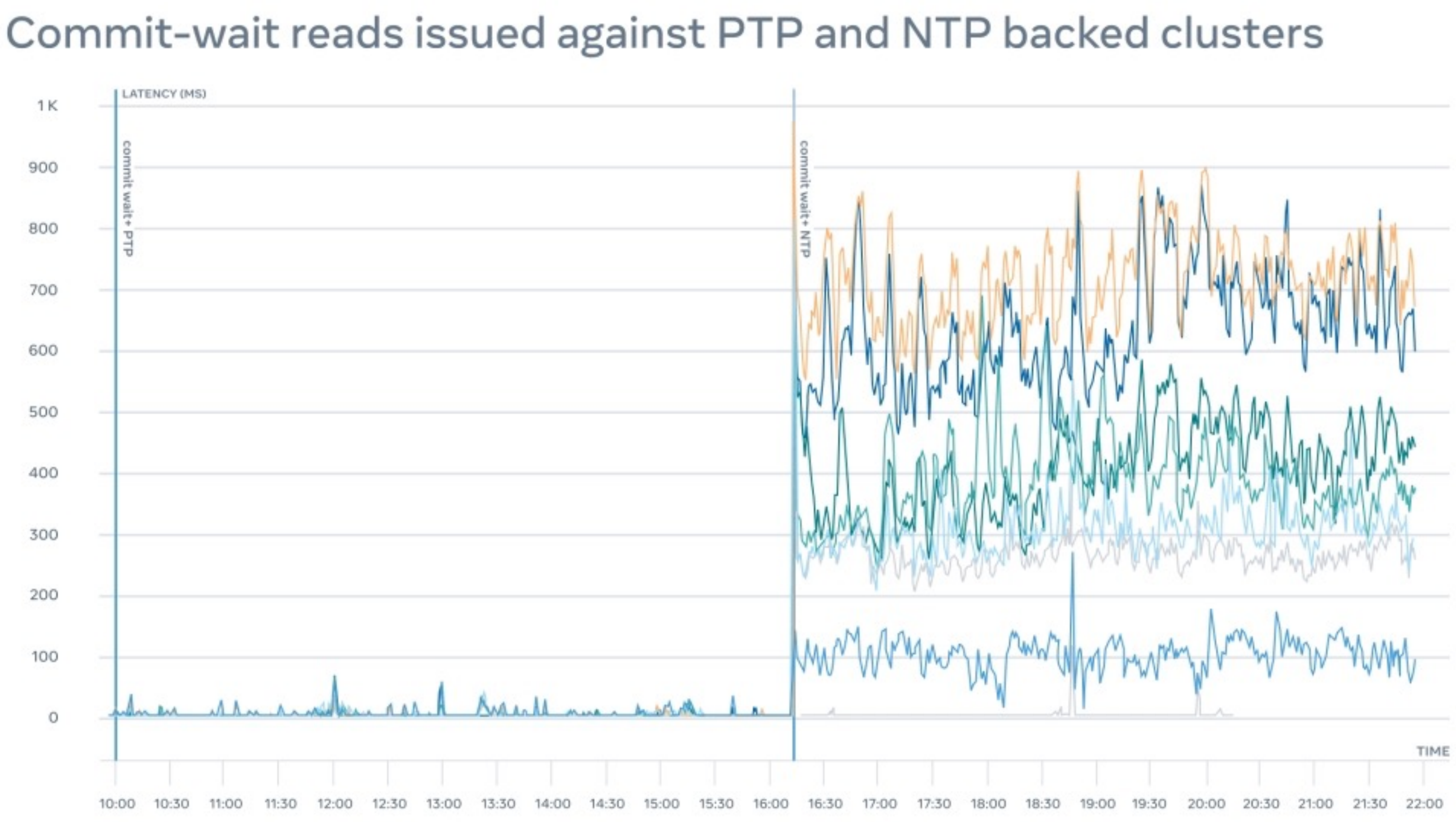
Source: https://engineering.fb.com/2022/11/21/production-engineering/precision-time-protocol-at-meta/

Schematic representation of read returning outdated information



Source: https://engineering.fb.com/2022/11/21/production-engineering/precision-time-protocol-at-meta/

Why is NTP not accurate enough?

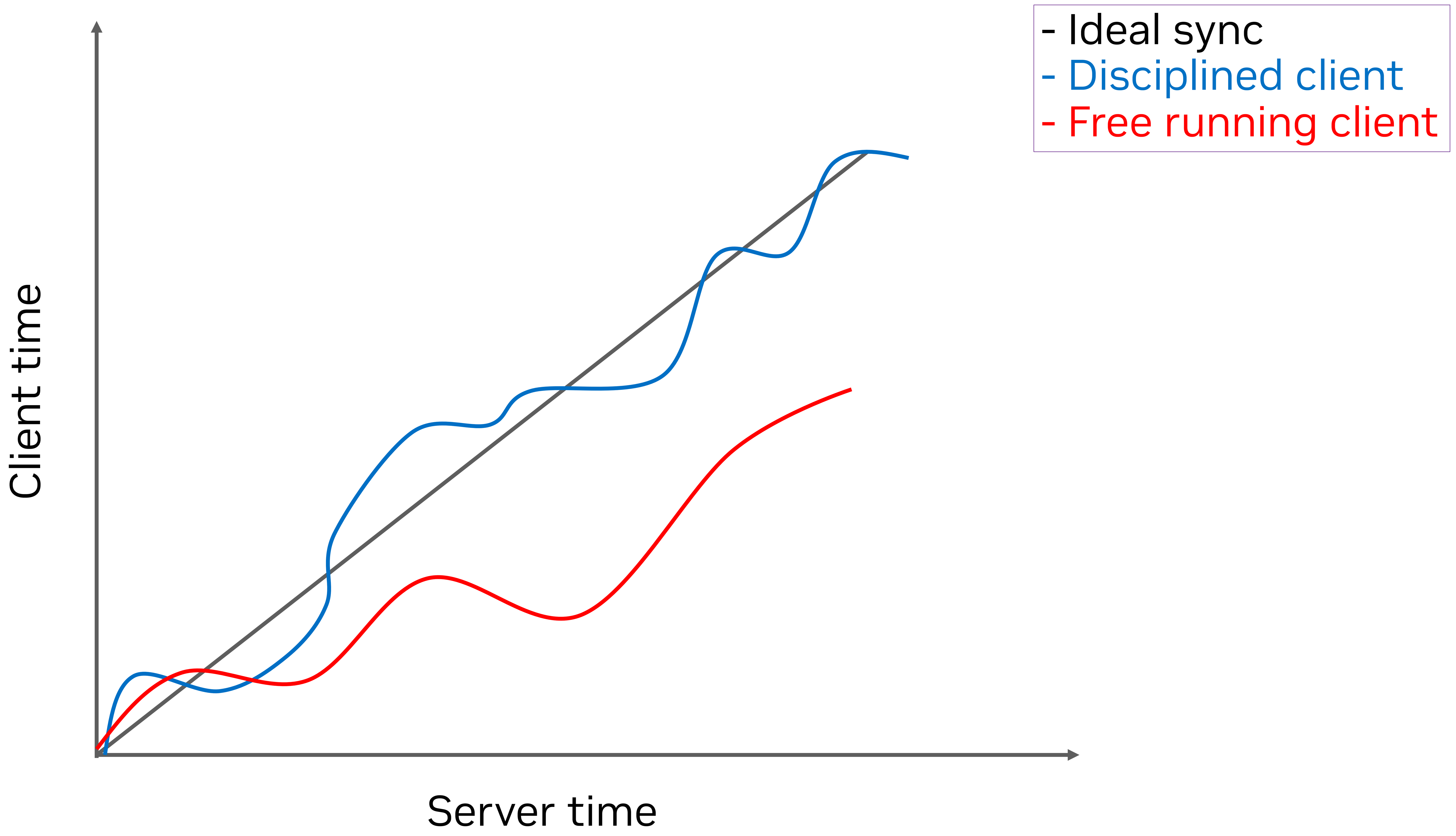


Source: https://engineering.fb.com/2022/11/21/production-engineering/precision-time-protocol-at-meta/

Timing 101

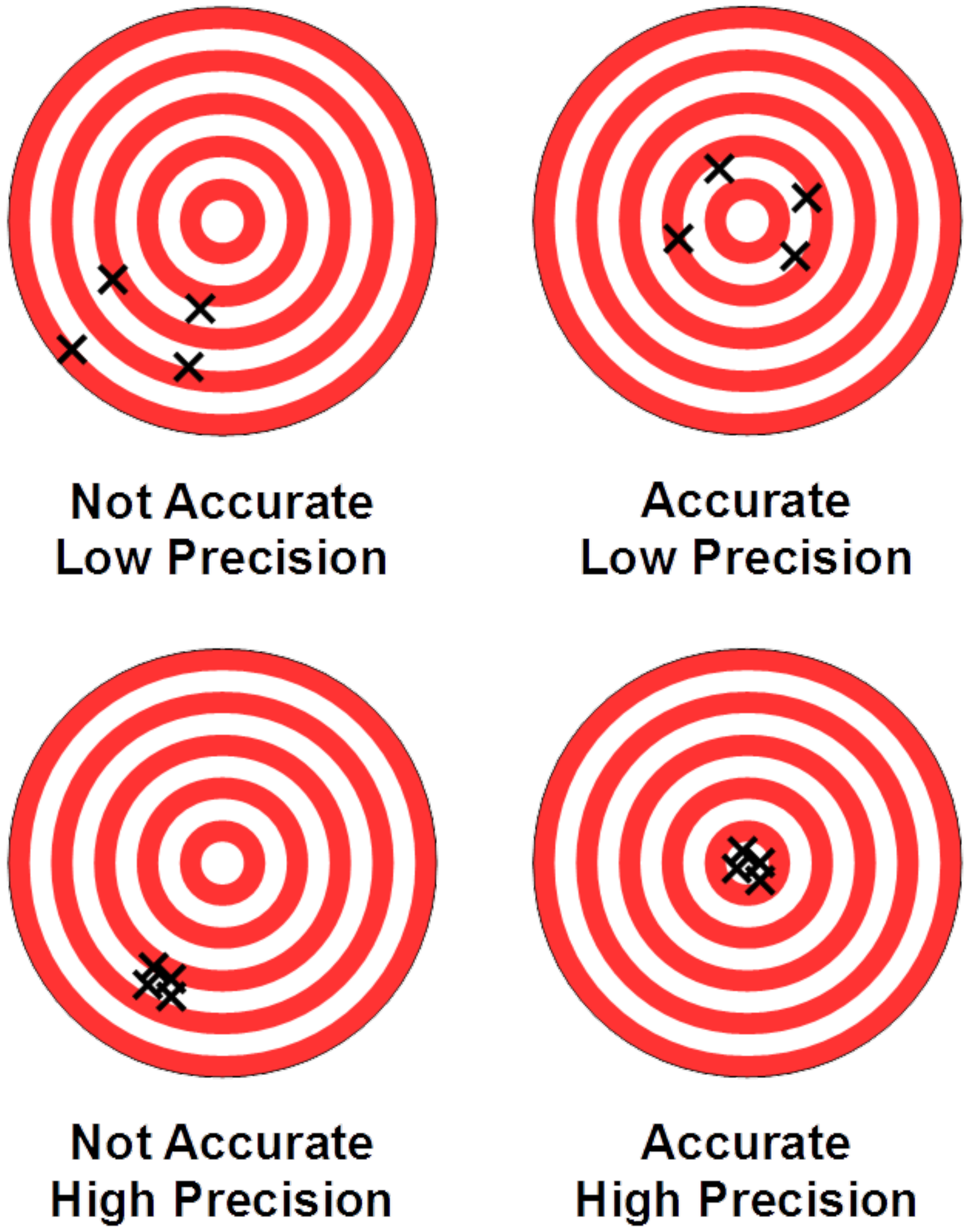
Timing 101

Clock Sync



What is accuracy?

- Node dependent
  - Time stamping resolution
  - Local oscillator quality
- Network related
  - Packet Delay Variations
  - Performance of time aware network devices
  - Different paths upstream and downstream
  - Highly asymmetric network loading
- Configuration dependent
  - Message rates

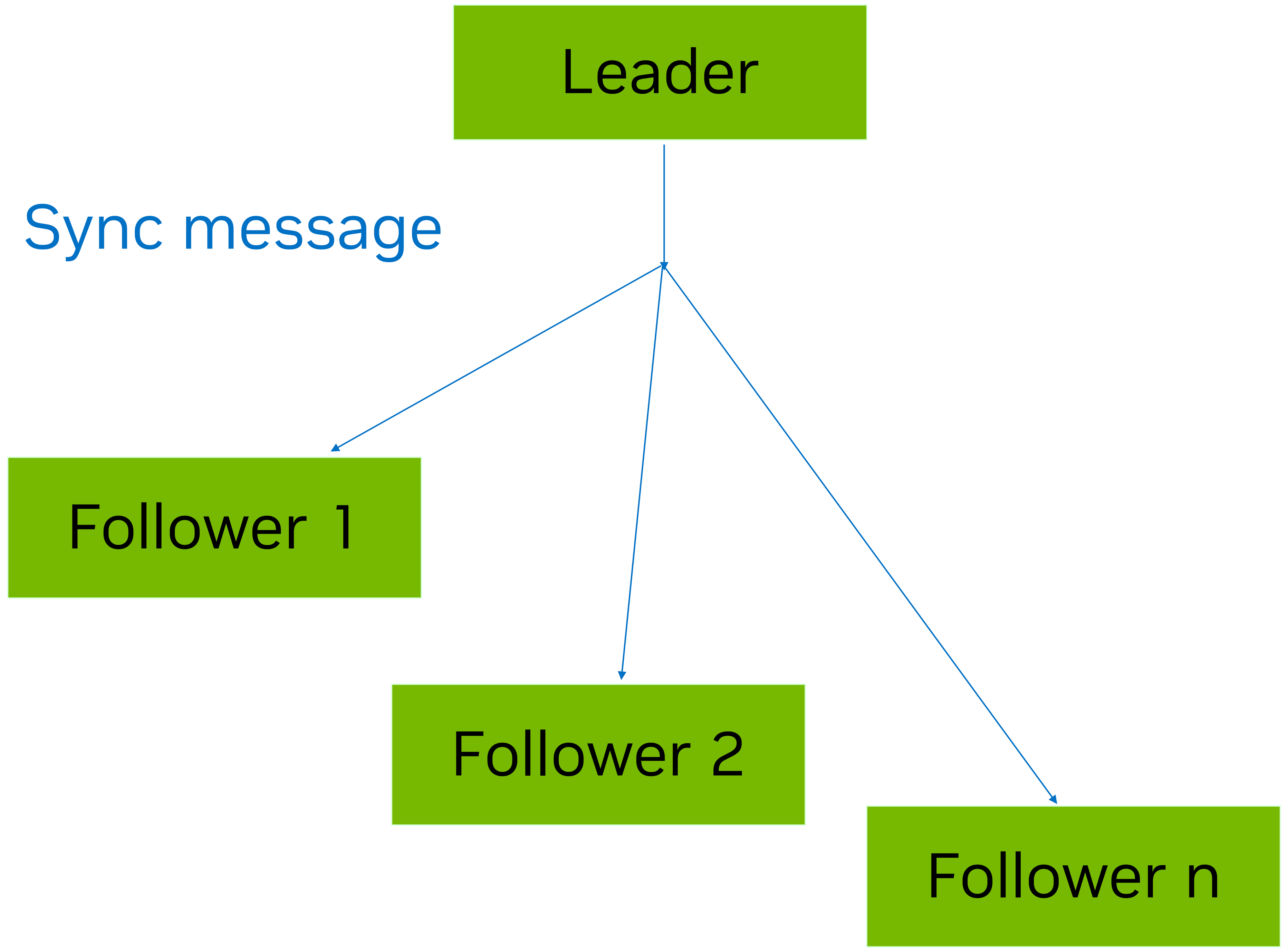


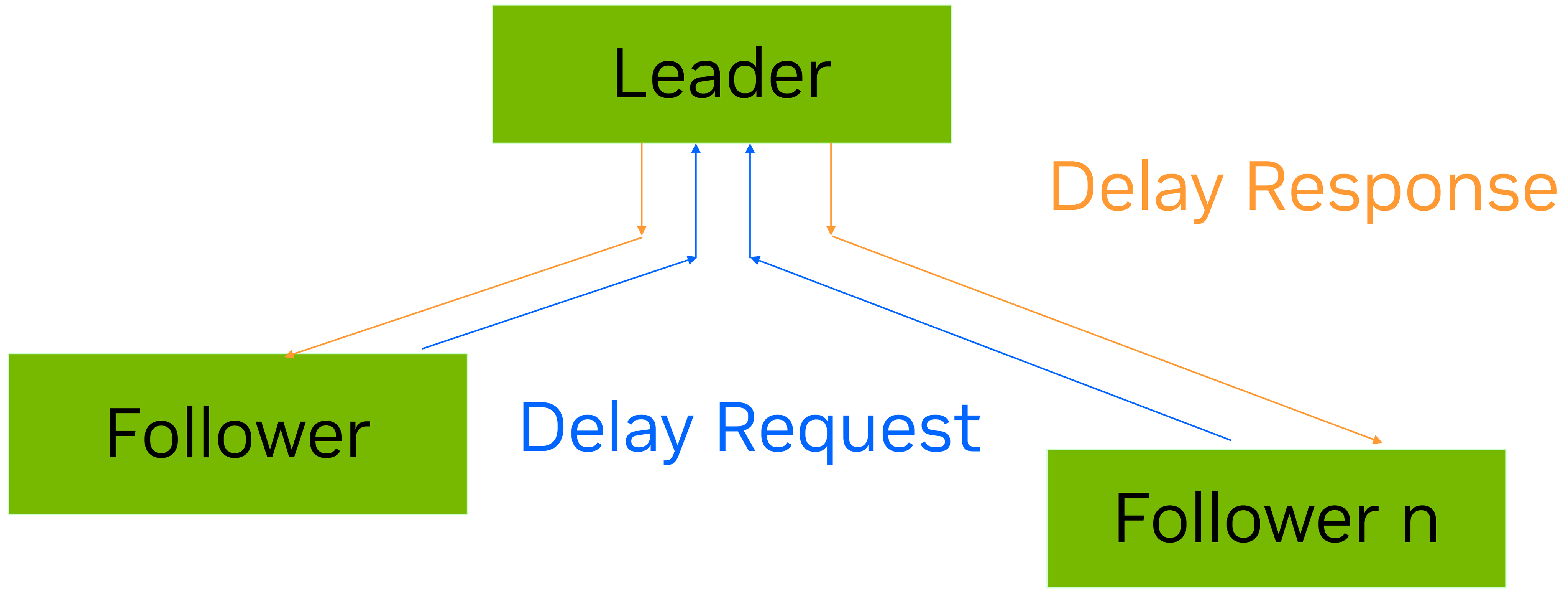
Timing 101

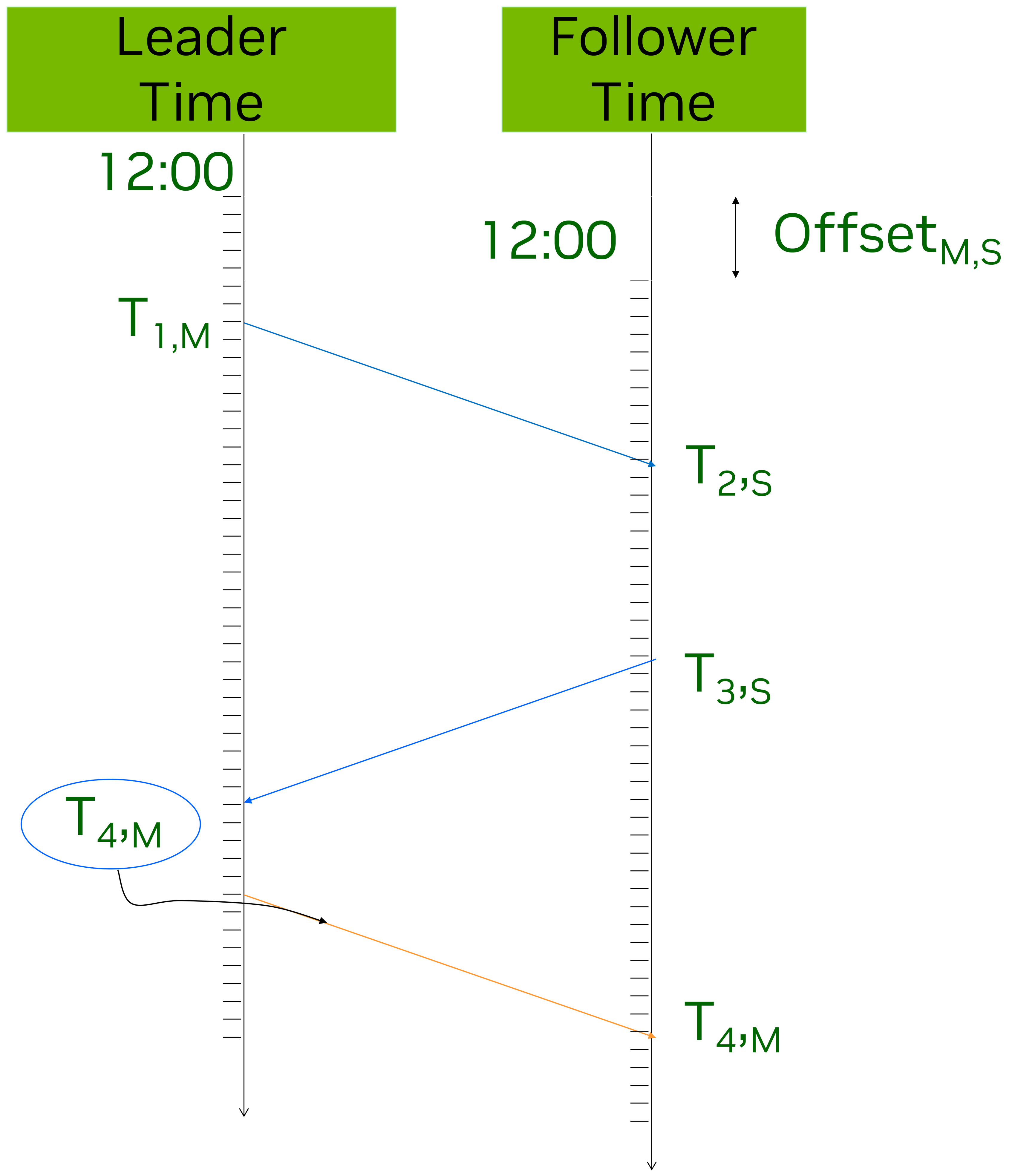
End to End time transfer

Σ time transfer from reference clock to application (userspace) representation

Basic Principles of PTP







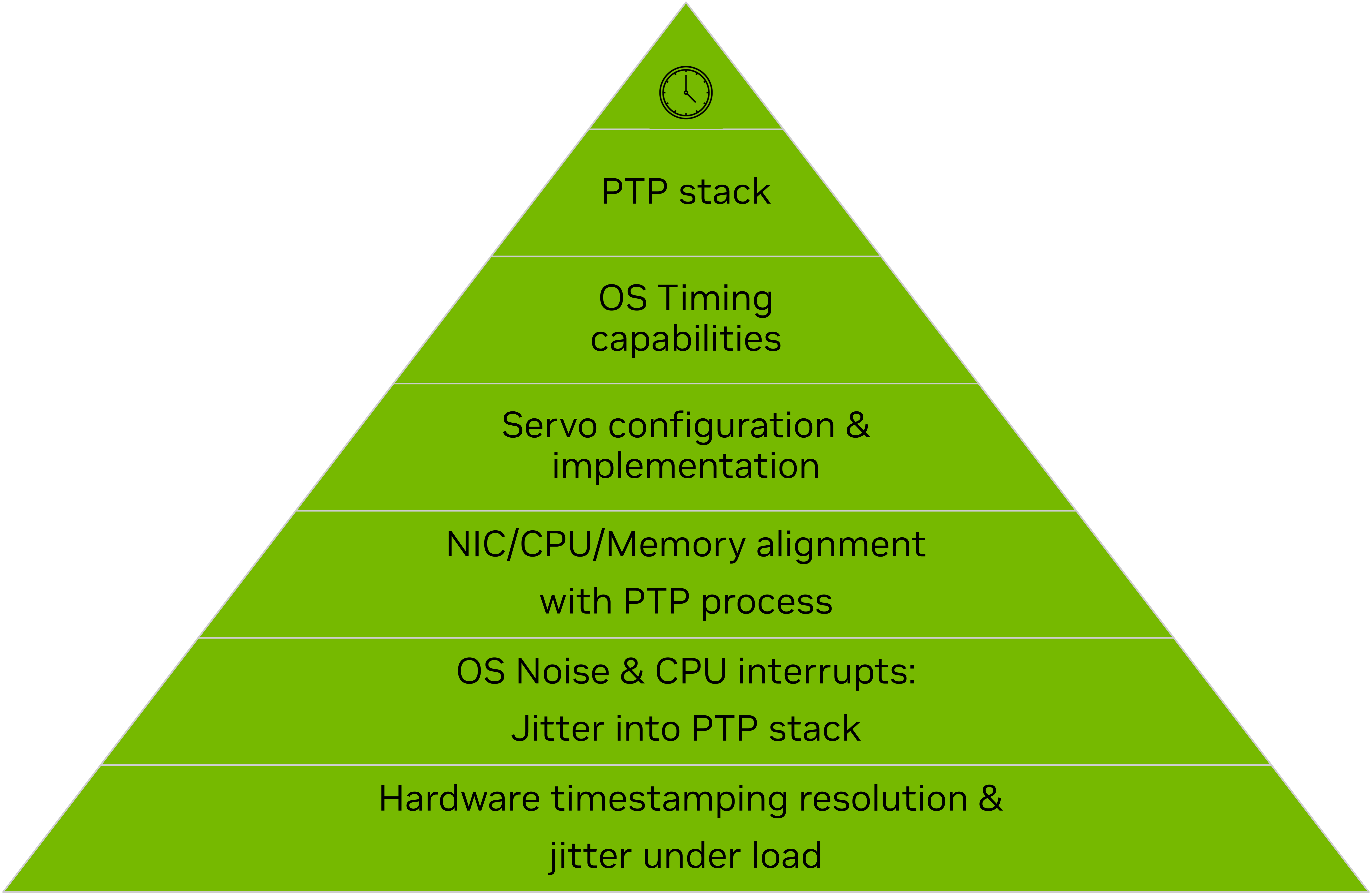
$$T_{2,S} - T_{1,M} = T_{2,1}$$
$$T_{4,M} - T_{3,S} = T_{4,3}$$

$$Offset = \frac{T_{2,1} - T_{4,2}}{2}$$

$$Delay = \frac{T_{2,1} + T_{4,3}}{2}$$

Delivering Consistent Timing

Challenges To Be Overcome

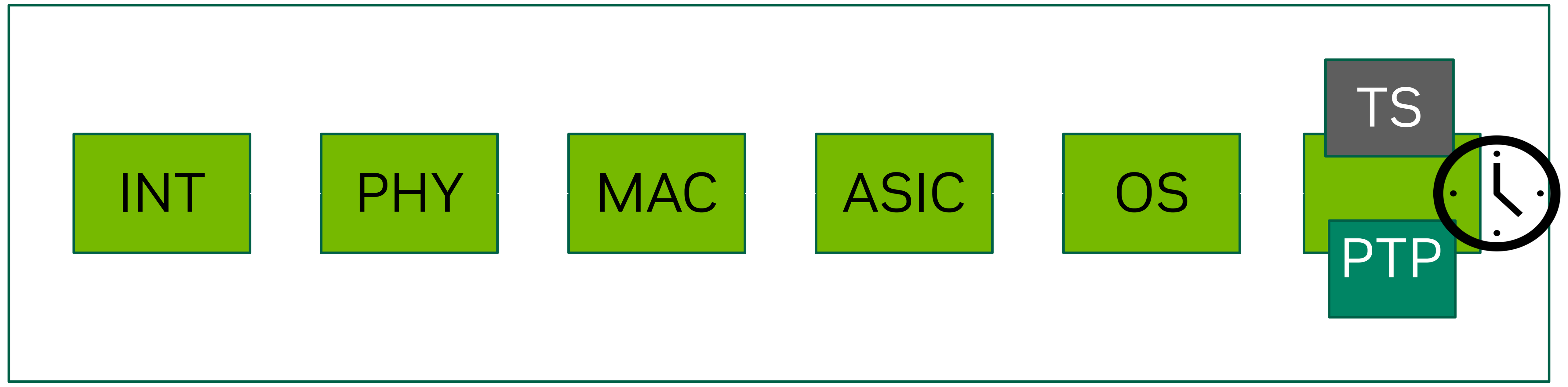


Target is performance dependent (ie: accuracy)

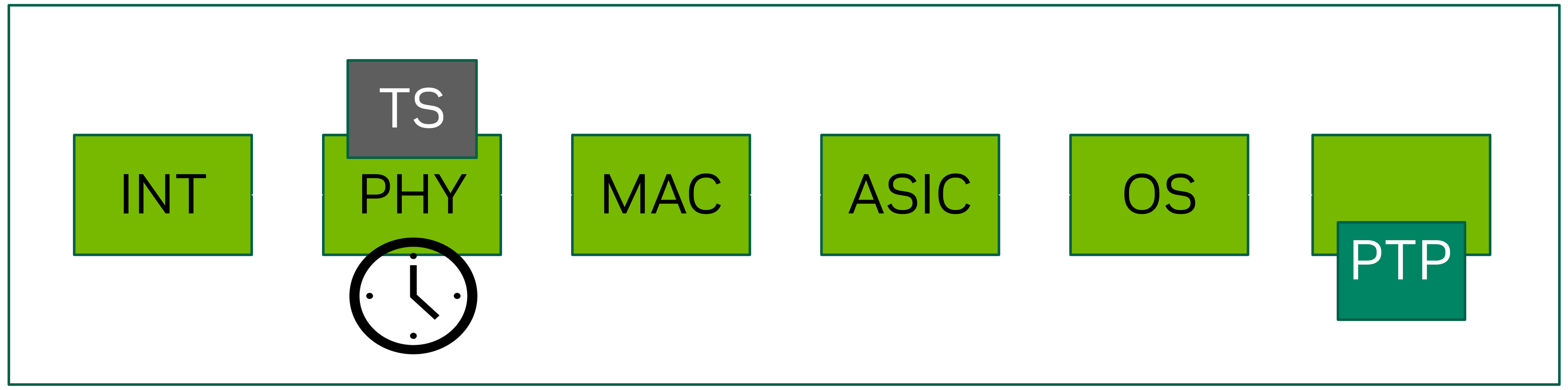
Software vs. Hardware timestamping

Software timestamping doesn'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)



Software timestamping: TS, Clock & PTP



Hardware timestamping: TS, PHC vs. PTP

Timestamping capabilities

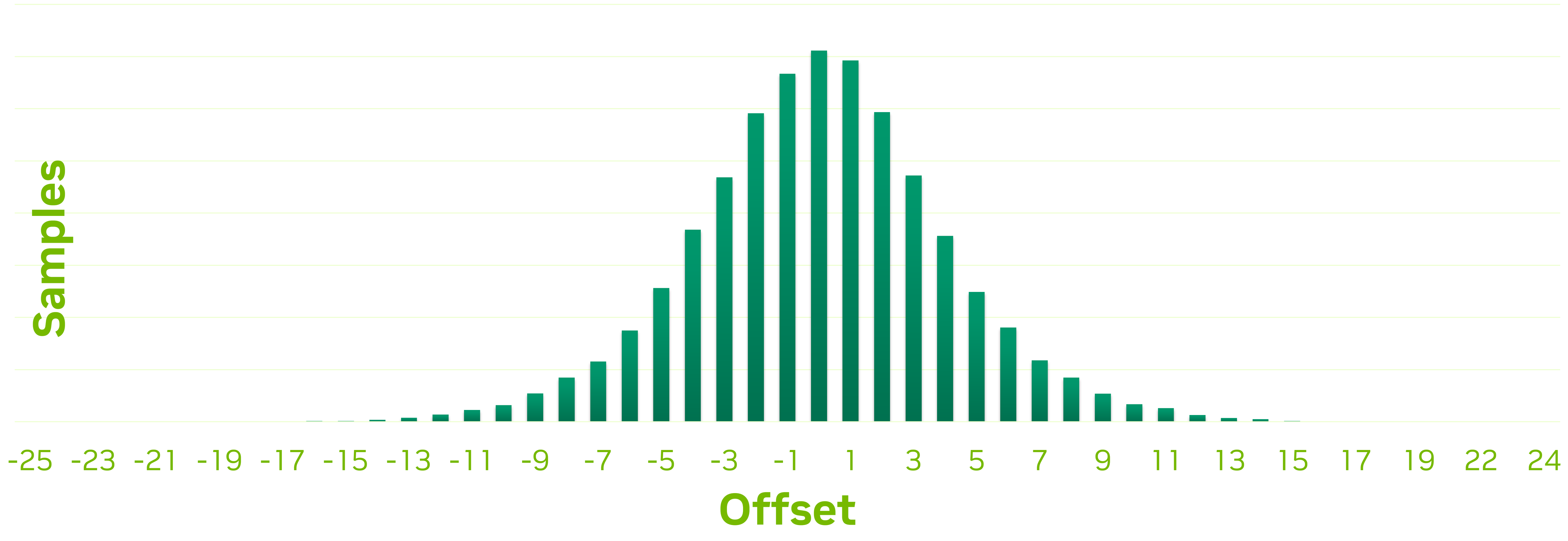
```
Device #1:
----------

  Device Type:      ConnectX7
  Part Number:      MCX713106AS-CEA_Ax
  Description:      NVIDIA ConnectX-7 HHHL Adapter Card; 100GbE; Dual-port QSFP112; PCIe 5.0 x16; Crypto Disabled; Secure Boot Enabled
  PSID:             MT_0000000843
  PCI Device Name:  /dev/mst/mt4129_pciconf0
  Base GUID:        946dae0300088e6e
  Base MAC:         946dae088e6e
  Versions:         Current        Available
     FW             28.37.1014     N/A
     PXE            3.7.0102       N/A
     UEFI           14.30.0013     N/A

  Status:           No matching image found

Device #2:
----------

  Device Type:      ConnectX6DX
  Part Number:      MCX623106TC-CDA_Ax
  Description:      ConnectX-6 Dx EN adapter card; 100GbE; Dual-port QSFP56; Enhanced-SyncE & PTP GM support; PPS In/Out; PCIe 4.0 x16; Crypto and Secure Boot
  PSID:             MT_0000000761
  PCI Device Name:  /dev/mst/mt4125_pciconf0
  Base GUID:        946dae03000abbca
  Base MAC:         946dae0abbca
  Versions:         Current        Available
     FW             22.37.1014     N/A
     PXE            3.7.0102       N/A
     UEFI           14.30.0013     N/A

  Status:           No matching image found
```

Timestamping capabilities

- sudo ethtool -T enp6s0f0np0

```
Time stamping parameters for enp6s0f0np0:
Capabilities:
        hardware-transmit
        hardware-receive
        hardware-raw-clock
PTP Hardware Clock: 2
Hardware Transmit Timestamp Modes:
        off
        on
Hardware Receive Filter Modes:
        none
        all
```

- Offset distribution in nanoseconds



PTP Profiles across Industries

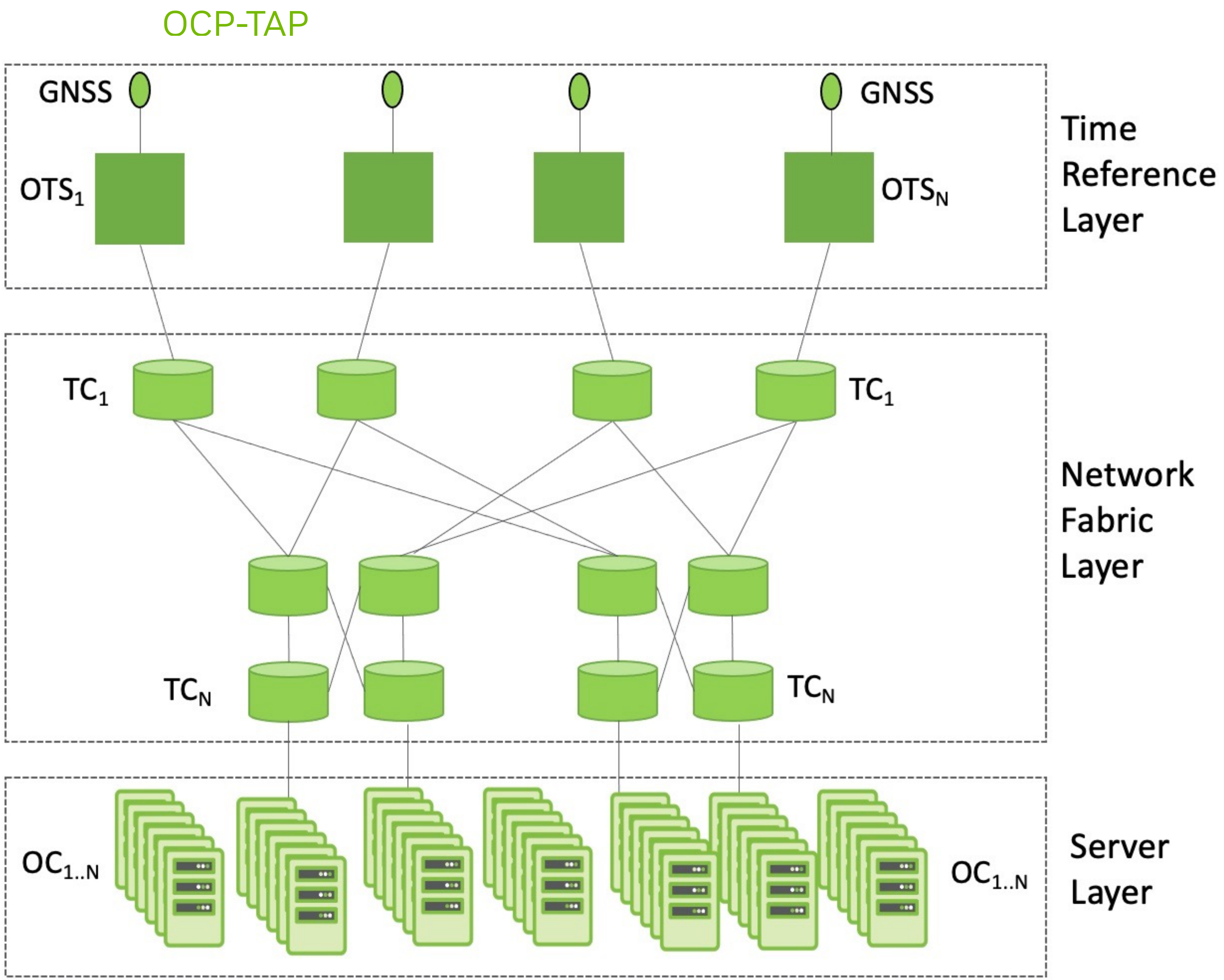
Industry	Application	Specification
Telecom & Mobile	Sync for 2G/3G/4G/5G base stations & fronthaul networks	ITU-T G.8265.1 ITU-T G.8275.1, G.8275.2
Professional Audio/Video	Sync for video/audio feeds between sources and receivers	SMPTE ST 2059-2
Power	Sync for substation sampled values, synchrophasor, power protection	IEEE C37.238-2017 IEC 61850-9-3 & IEC 62493-2 Annex A.2
Audio/Video, Industrial, Automation, Automotive	Sync of A/V applications with high QoS/QoE demand and time sensitive networks	IEEE Std 802.1AS-2020
Industrial Automation	Sync for industrial plants, machine-to-machine real-time control	IEC 62439-3 Annex B IEC 62439-3 Annex C
Enterprise/Financial	Sync of time tagged and packet latency measurements	draft-ietf-tictoc-ptp-enterprise-profile
Data Center	Sync for time-sensitive applications within data center	OCP DC PTP Profile #1

“Time Sync Service” Reference Model

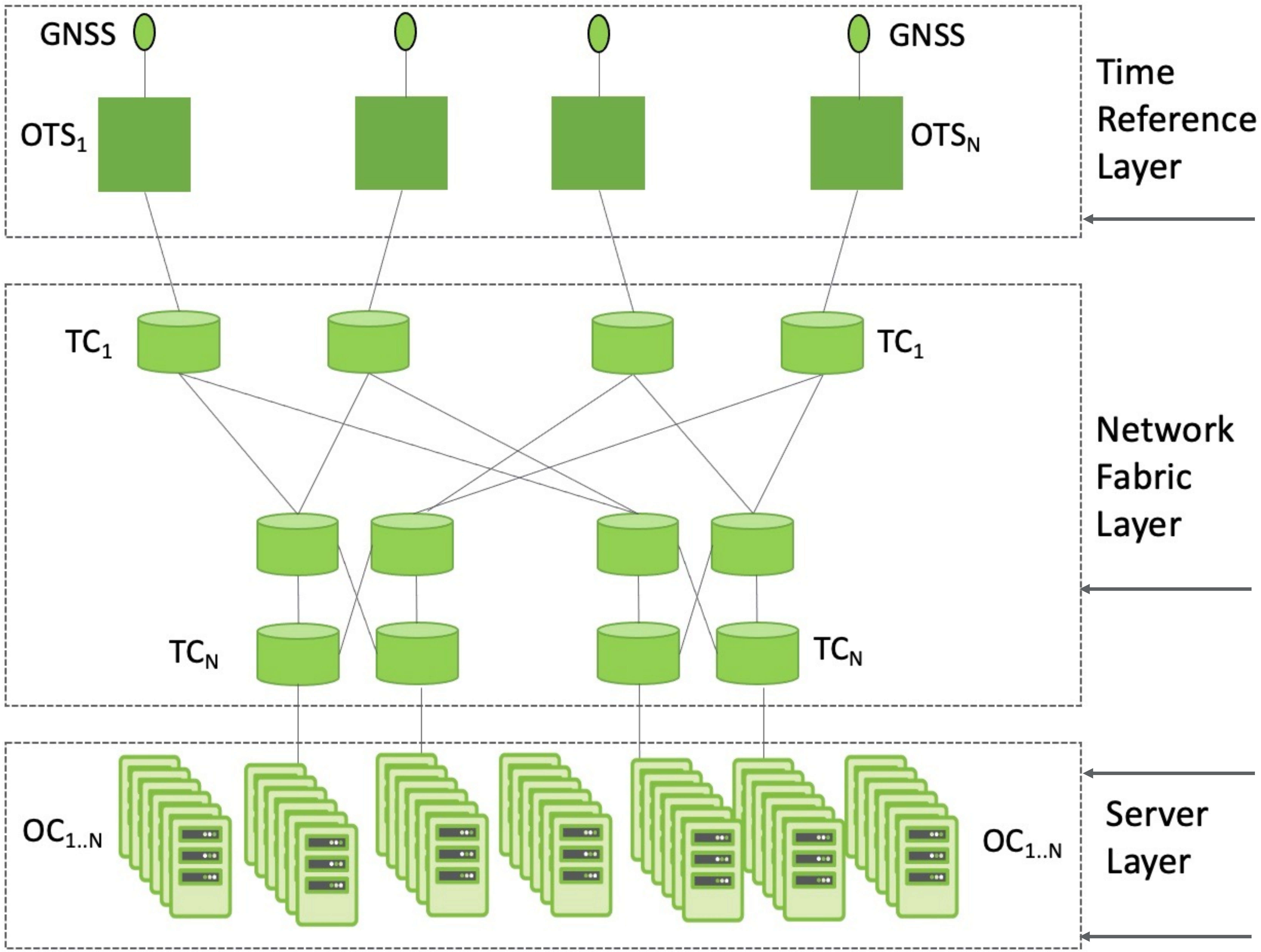
Time Reference Layer:
- Rootftop antennas, GPS system
- Open Time Server (OTS) (aka GM)

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



Time Error Budget



T= Time Error
OTS = Open Time Server
TC = Transparent Clock
OC= Ordinary Clock

$|T_{GM,j} - TAI| \leq 100$ ns

$|T_{GM,j} - T_{GM,k}| \leq 200$ ns for $k \neq j$

$|T_{GM,i} - T_{OC,k}| \leq 2.4$ µs for $k \neq j$

$|T_{TCj,k}| \leq 200$ ns

$|T_{OC,j} - T_{OC,k}| \leq 5$ µs for $k \neq j$

$|T_{OC,j} - TAI| \leq 2.5$ µs

OCP-TAP DC PTP Profile #1

Key values

PTP Attributes	PTP Profile Value
Company ID	7A-4D-2F (OCP)
Clock types	GM, E2E TC, OC
Network transport	IPv6 (mandatory) IPv4 (recommended) Highest class of service
Messages & Rates	Announce {0, -4}, Sync {+3, -7}, Follow_Up, Delay_Req/Delay_Resp {0, -7} Signaling, Management
Path delay measurement	Delay Request-Response mechanism
Domain Number	0
Clock Operations	One-step and Two-step for GM, OC One-step for TC (mandatory) Two-step for TC (not recommended)
Network Communication	Unicast discovery & Unicast negotiation Multicast is prohibited
Clock Class	6 (traceable) 7 (holdover, within spec) 52 (holdover, out of spec)
A-BMCA	Active-Active Active-Standby

In Conclusion

- The nanosecond scale world is fascinating!
- Builds upon IEEE 1588 Precision Time Protocol
- Tuned for DC applications in OCP-TAP
- Enables new applications
- Improves current applications
- Delivers reliable time synchronization as a DC service

NVIDIA®