Bar Shapira, Senior Software Architect, Nvidia, Israel

Bar Shapira is a Senior Software Architect within NVIDIA's NIC and DPU Architecture team.

He leads the SW architecture and development of timing and synchronization solutions, including SyncE and PTP, alongside essential network monitoring and diagnostic features.

He is an innovator in the field, holding multiple patents related to network synchronization systems.

Bar earned his B.Sc. in Computer Science and Physics from Tel Aviv University in 2020."

NVIDIA

Time Synchronization in Data Centre Networks

Bar Shapira | ITU workshop on Future Optical Networks for IMT2030, AI, broadband and more, June 2025

Agenda

- Use Cases and Requirements for Time Sync in Data Centers
  - Distributed Databases
  - Synchronized High-frequency Telemetry and Profiling
  - Congestion Control based on One-way Delay
  - Synchronized Collectives
- G.Sub.DCSync Accuracy levels
- Suggested Synchronization Solutions for Data Centers

“

Time is an accelerator

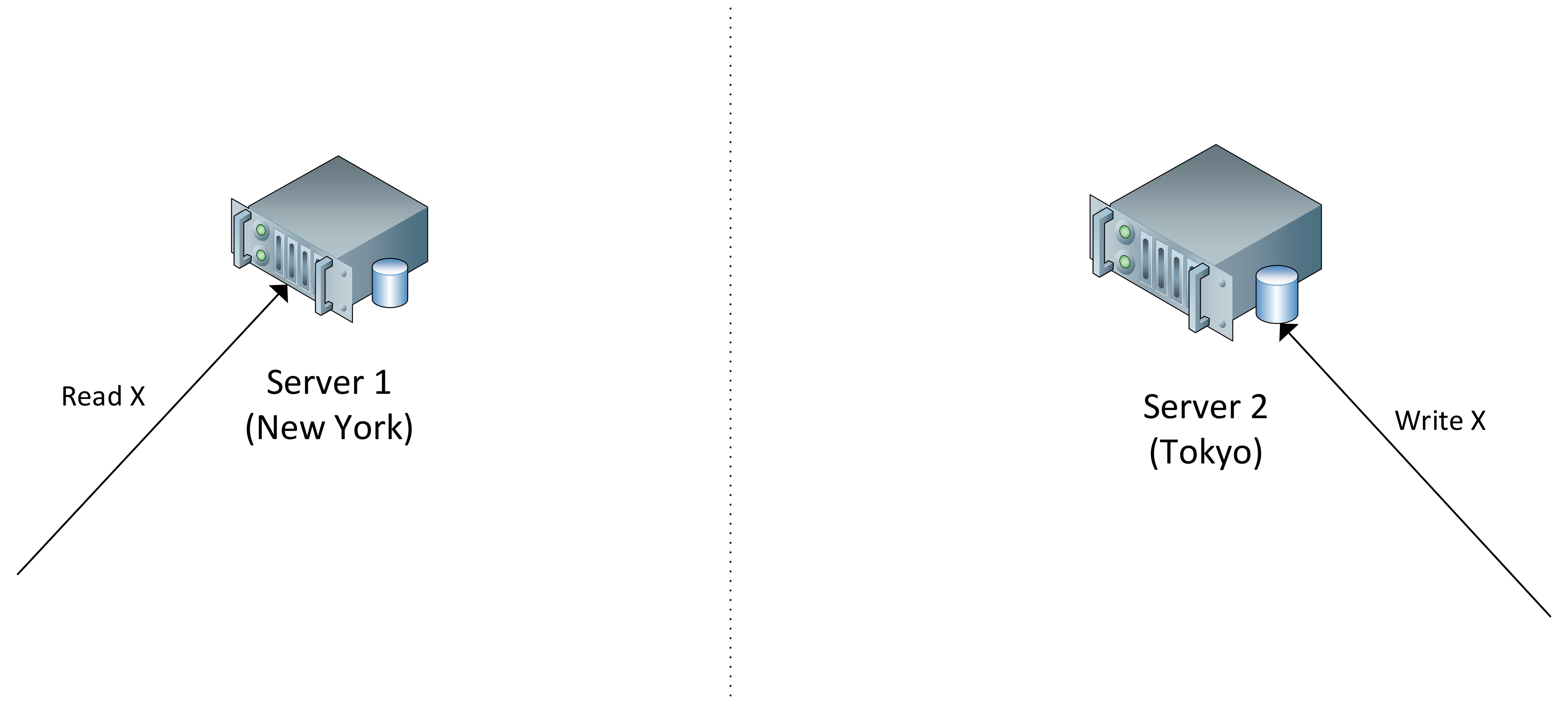
End to End synchronization unlocks the next level of application performance

”

Distributed Databases

Use Cases and Requirements for Time Sync in Data Centers

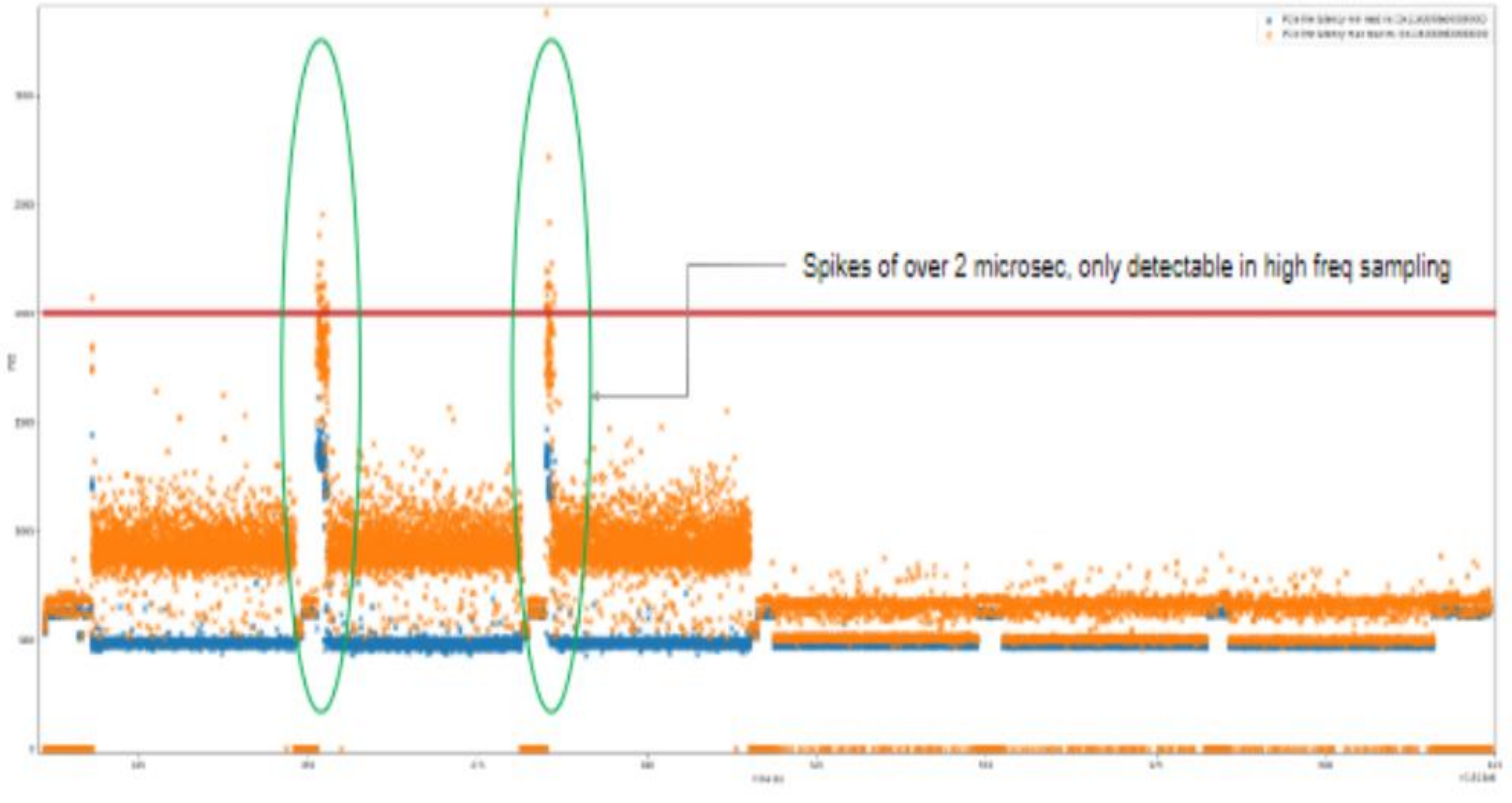
- **Scalable Databases**
  - Scalable Databases with lock-free distributed read transactions based on an API that exposes time uncertainty. Utilize time synchronization to achieve external consistency of distributed transactions.
- **Accuracy requirements**
  - Upon a read transaction, a server must wait a period equal to the time uncertainty. To avoid this being the bottleneck, the time uncertainty should be lower than the network delay and processing time. Absolute time accuracy is a must.



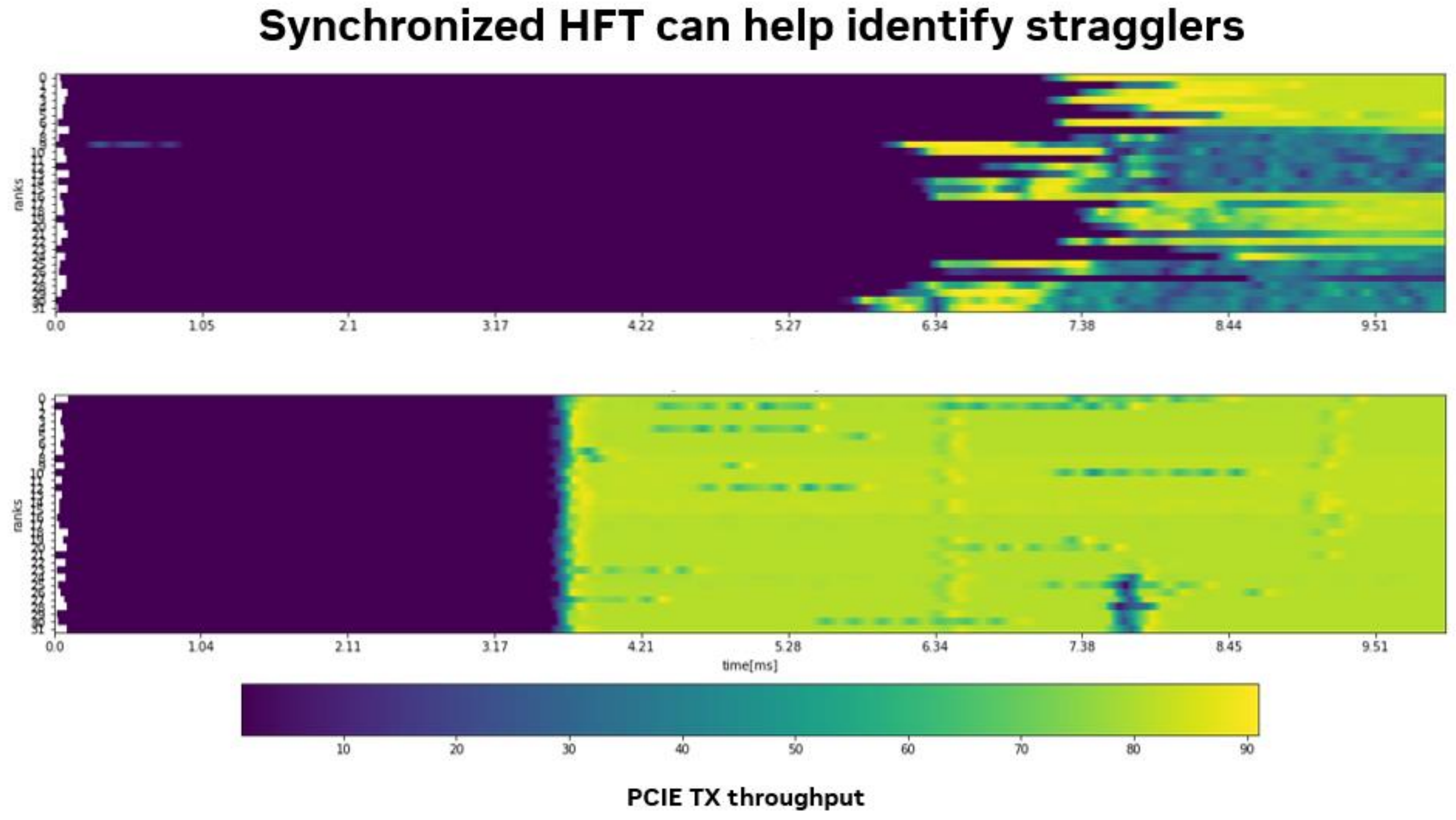
Synchronized High-frequency Telemetry and Profiling

Use Cases and Requirements for Time Sync in Data Centers

- **Synchronized high-frequency telemetry and Profiling**
  - HFT tools sample telemetry data with associated timestamps from elements in the data center at high rates. Accurate synchronization between devices is required to align the data collected from multiple devices on a single timeline.
- **Accuracy requirements**
  - Accuracy between endpoints must be an order of magnitude better than the sampling frequency. The accuracy is needed at the clock that timestamps the events – could be NIC, Switch or CPU clock, depending on the measured events.



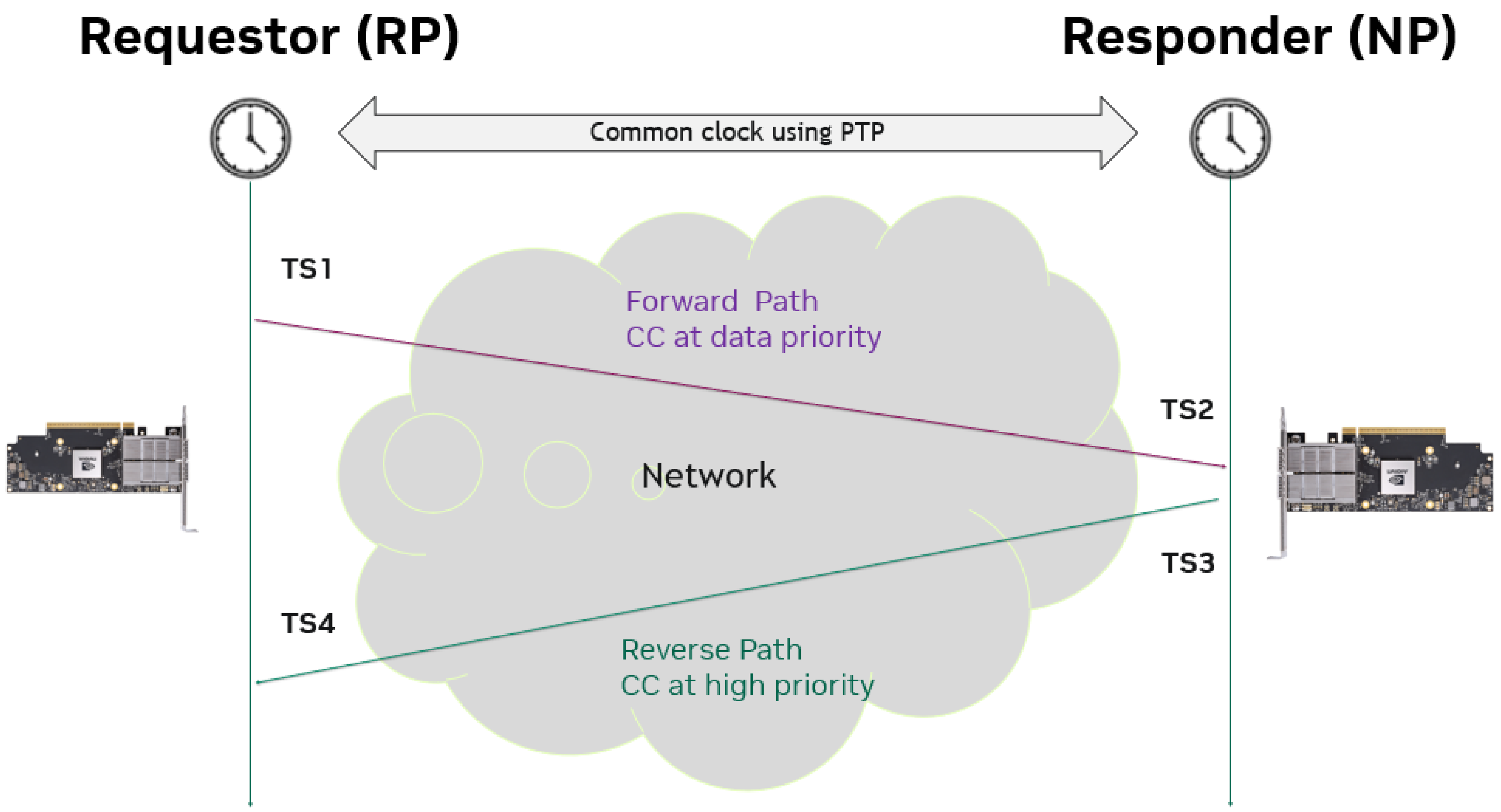
PCIe link latency max read time (ns)



Congestion Control based on One-way Delay

Use Cases and Requirements for Time Sync in Data Centers

- **Congestion control based on one-way delay**
  - Measure the per-path direction (TX vs. RX) delay of packets in the network for the use of congestion control. Can help differentiate between forwarded and reverse path congestion.
- **Accuracy requirements**
  - Accuracy between end points must be an order of magnitude better than the actual delay. Relative time accuracy in the NIC is sufficient.

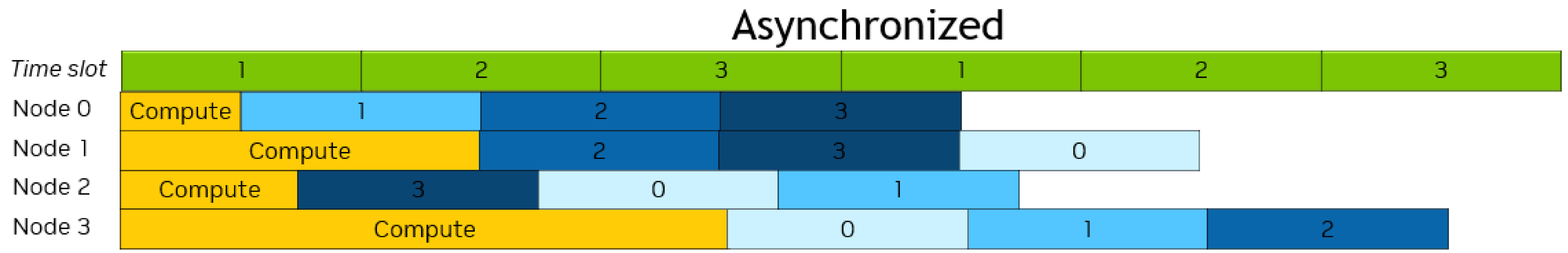


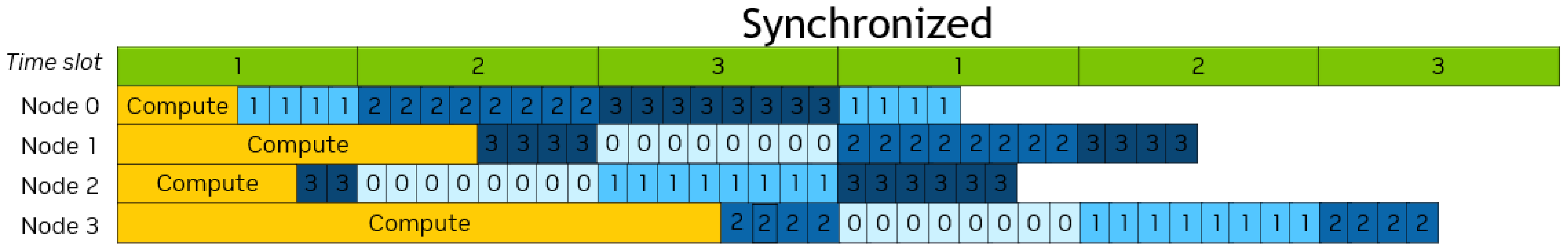
- Forward delay = TS2-TS1
- Backward delay = TS4-TS3

Synchronized Collectives

Use Cases and Requirements for Time Sync in Data Centers

- **Synchronized collectives**
  - Reduce network load and improve completion times by aligning the operations of all nodes to designated timeslots, resulting in reduced network congestion.
- **Accuracy requirements**
  - The accuracy of time should be an order of magnitude smaller than the timeslot duration. The scheduling is done at the NIC using the PHC; relative time accuracy in the NIC is sufficient.

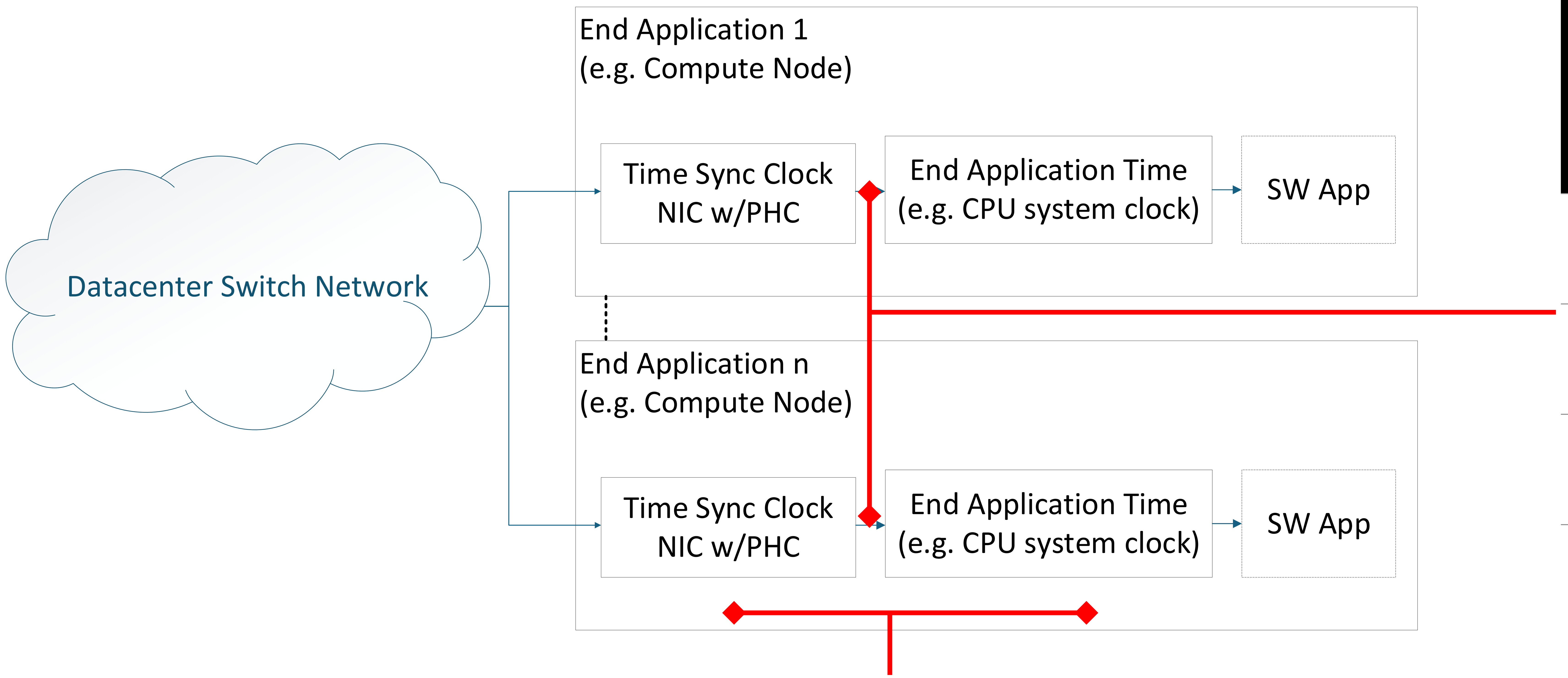




G.Sub.DCSync Accuracy Levels

(WIP)

- Simplified datacentres example



Class level at Time Sync Clocks	Relative Time error between Time Sync Clocks	Typical applications (for information)
1	5µsec	Distributed databases, applications profiling
2	1µsec	High-Frequency Telemetry,
3	200nsec	CC based OWD, Synchronized collective

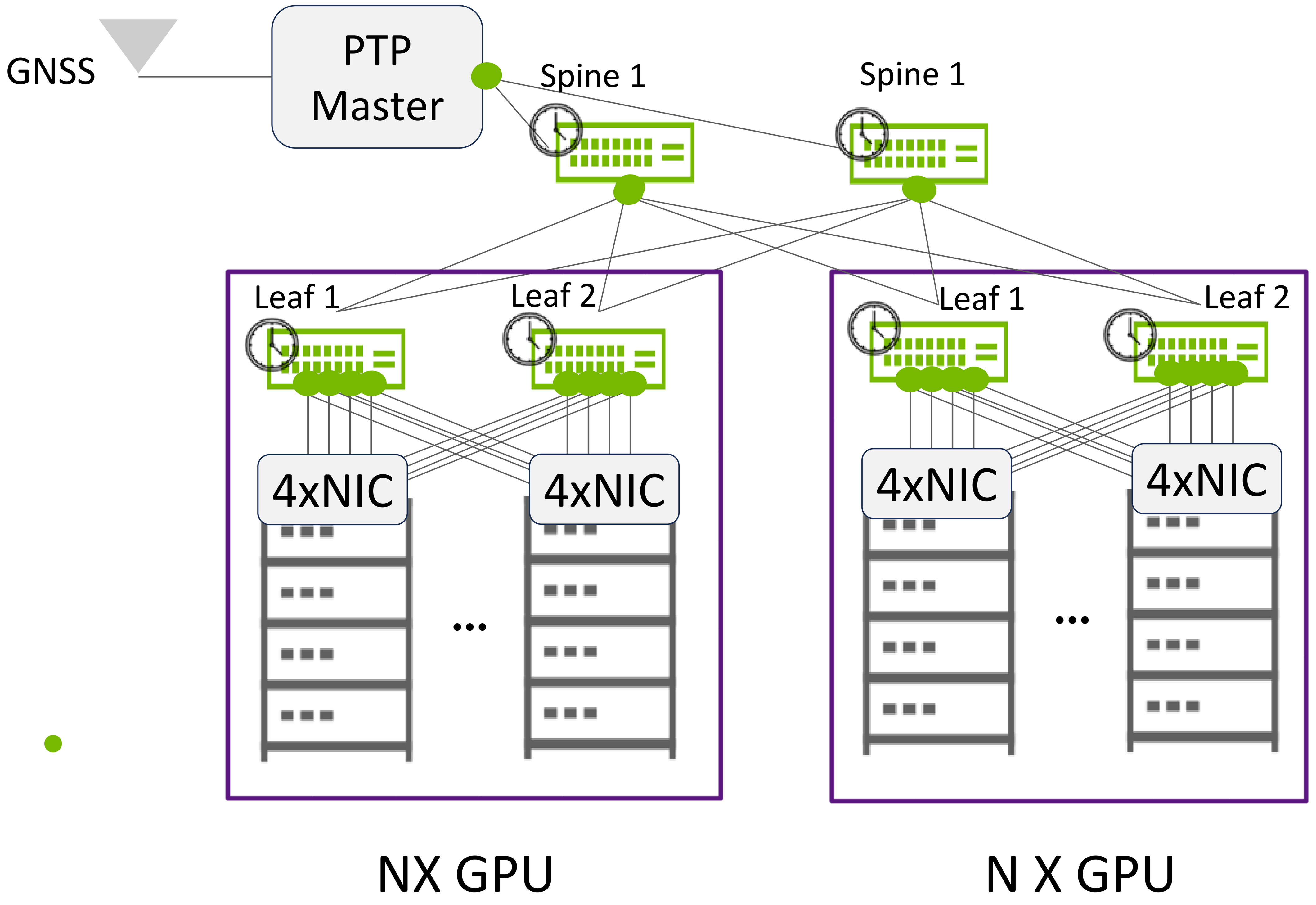
Class levels of End Application	Additional Time Error for End Application	Notes
A	2µsec	Typically, without PTM
B	200nsec	Typically, with PTM
C	50ns	Typically, with physical edge clock signal (e.g. PPS output of the Time Sync Clock to PPS input of the End Application Time Clocks when available).

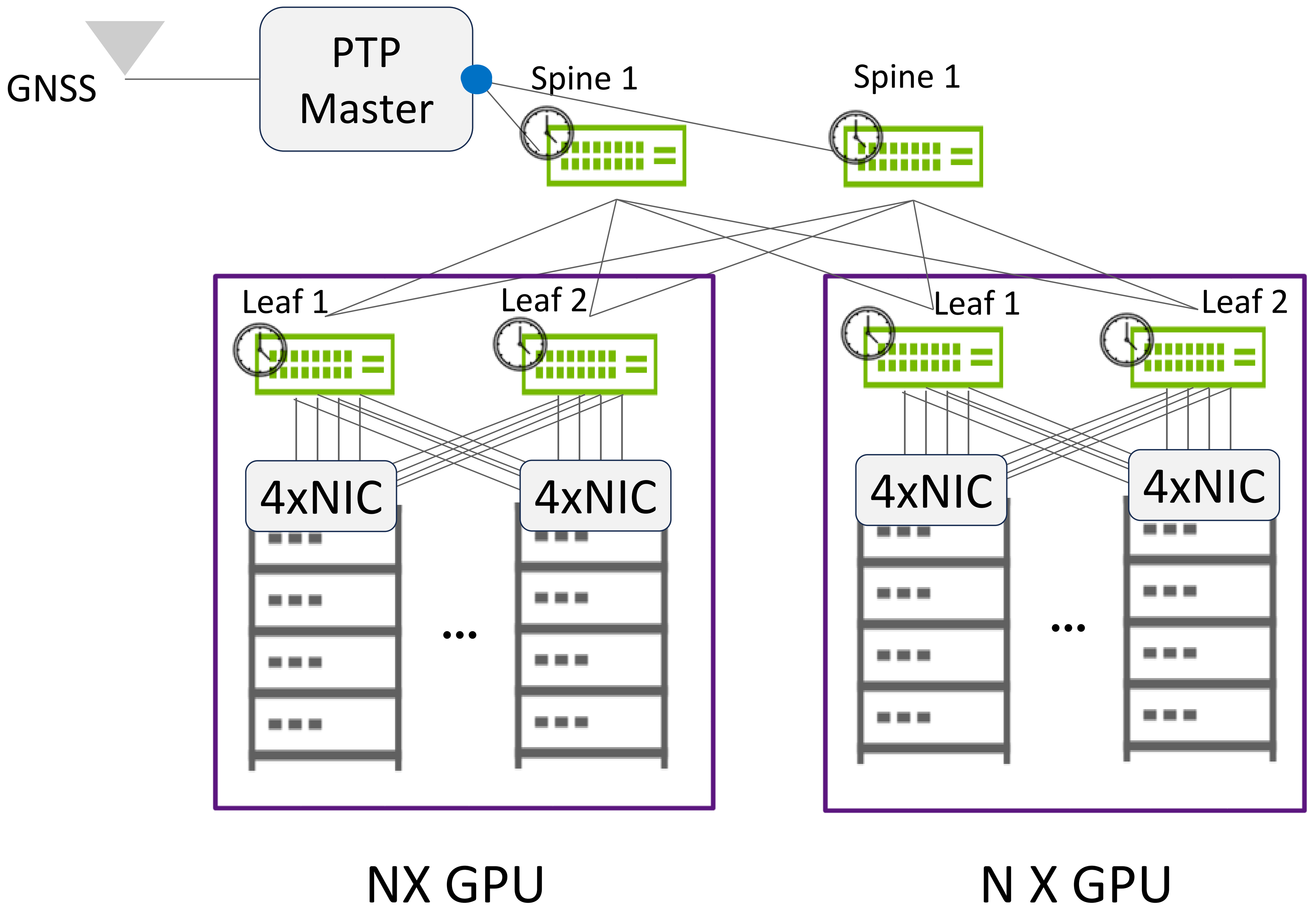
G.Sub.DCSync Suggested PTP Profiles

(WIP)

- Main challenge: **very high scale**
- There are two suggested PTP Profiles for data centers in G.Sub.DCSync. The right profile can be chosen depending on the specific needs and network setups.

Option	Layer	Network support	Notes
G.8275.1	**PTP over Eth -** Local link multi cast	Boundary clocks	G.8275.1 mature standard, used in telecom, easily scalable, no path asymmetry issue, require fully PTP aware network
IEEE 1588.1 CSPTP	**PTP over IP -** Unicast	Transparent clock	Reduce CPU, memory, and network utilization compared to traditional unicast PTP In development (private implementation - Simple PTP by Meta)





NVIDIA