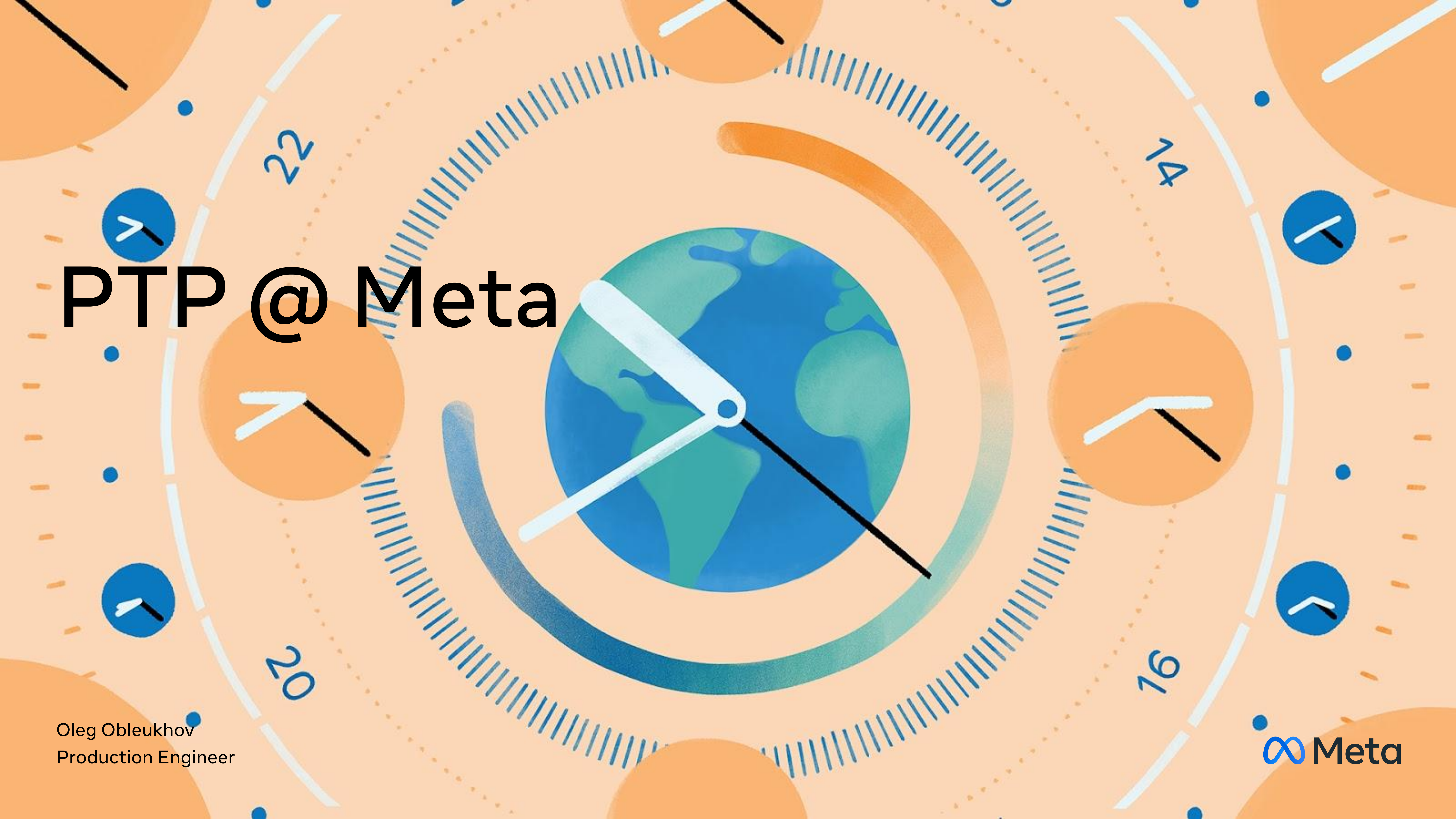


PTP @ Meta

Oleg Obleukhov
Production Engineer

Meta

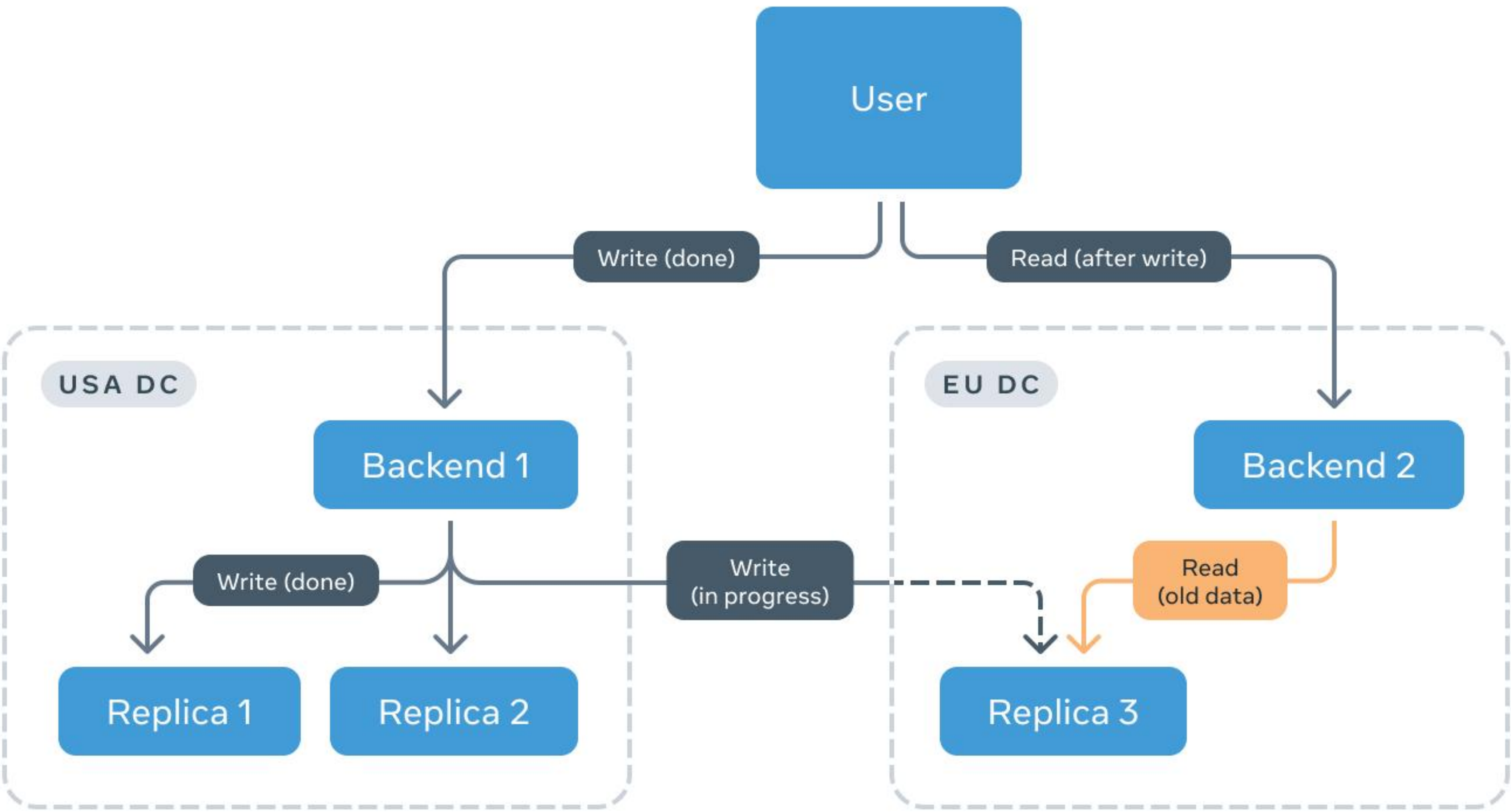
Agenda

- The case for PTP
- The PTP Architecture
  - The PTP rack
  - The PTP network
  - The PTP client
- How we monitor the PTP architecture

The case for PTP

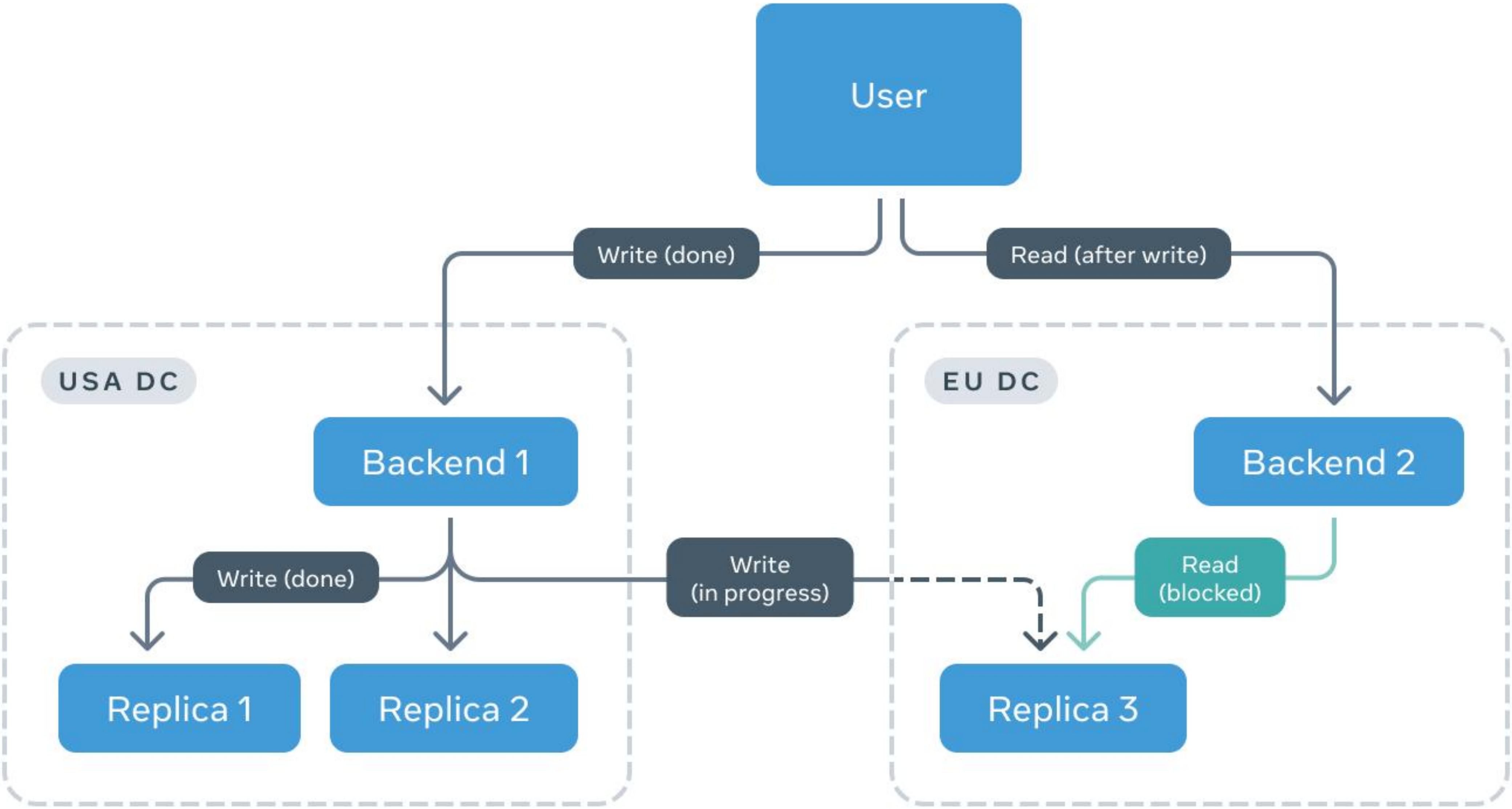
Schematic representation of read returning outdated information

Commit-wait ensuring consistency guarantee (linearizability)

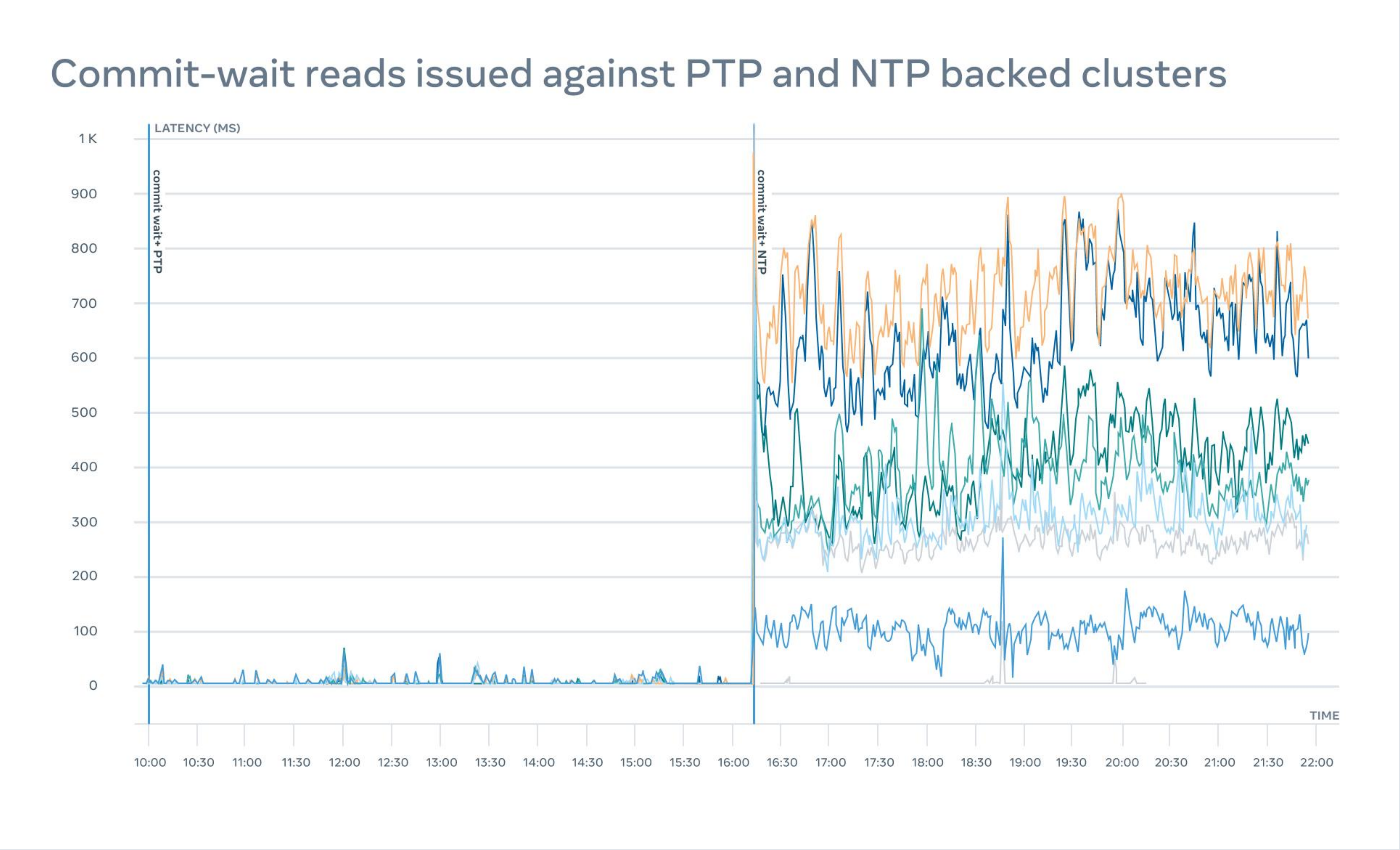


Schematic representation of read returning outdated information

Commit-wait ensuring consistency guarantee (linearizability)

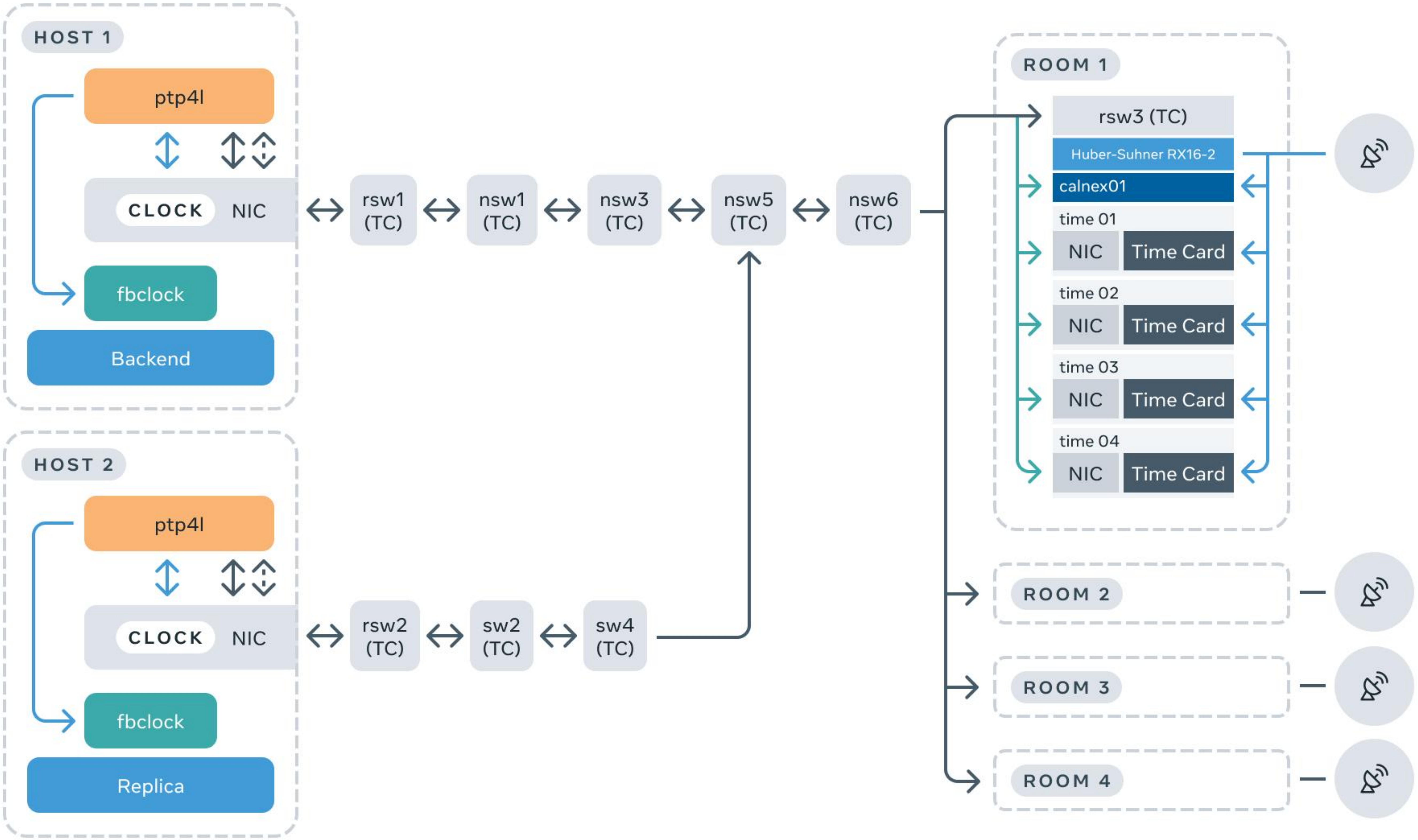


Commit-wait reads issued against PTP and NTP backed clusters



The PTP architecture

Regional PTP architecture



The PTP rack

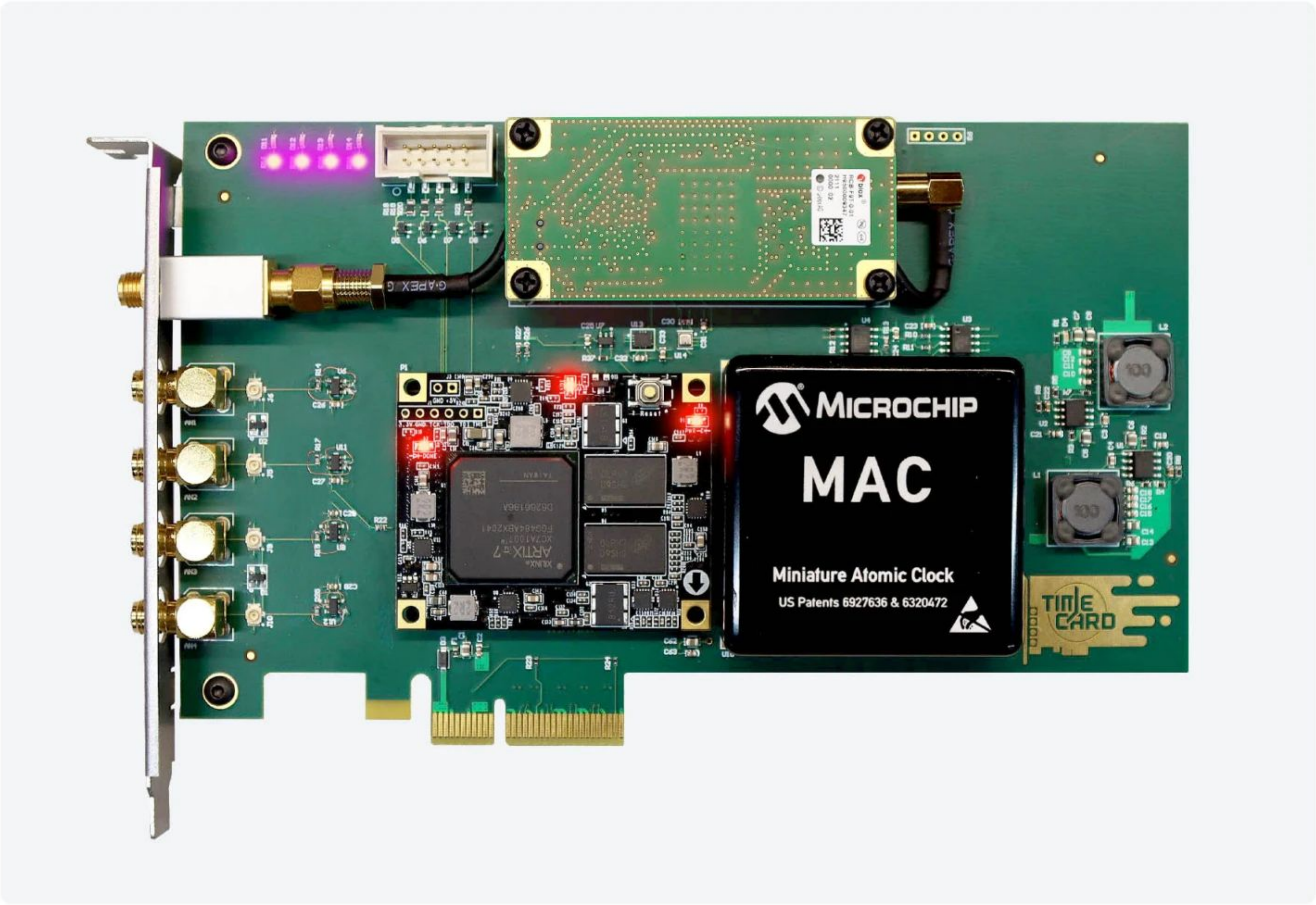
GNSS

GNSS antenna in one of the Meta Data Center location

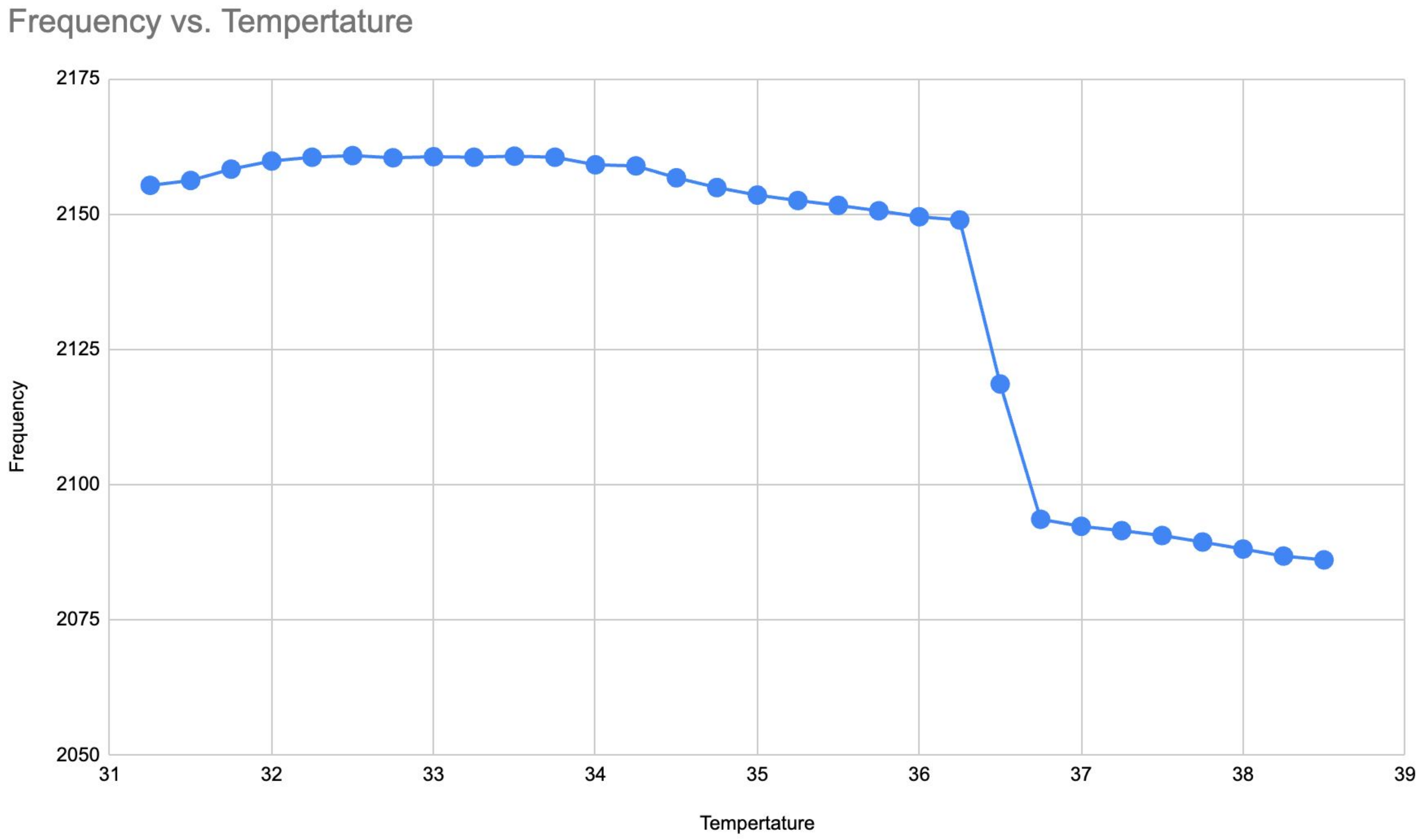
Huber-Suhner GNSS-over-fiber technology tested in Meta Dublin office

The Time Card

Facebook Time Card

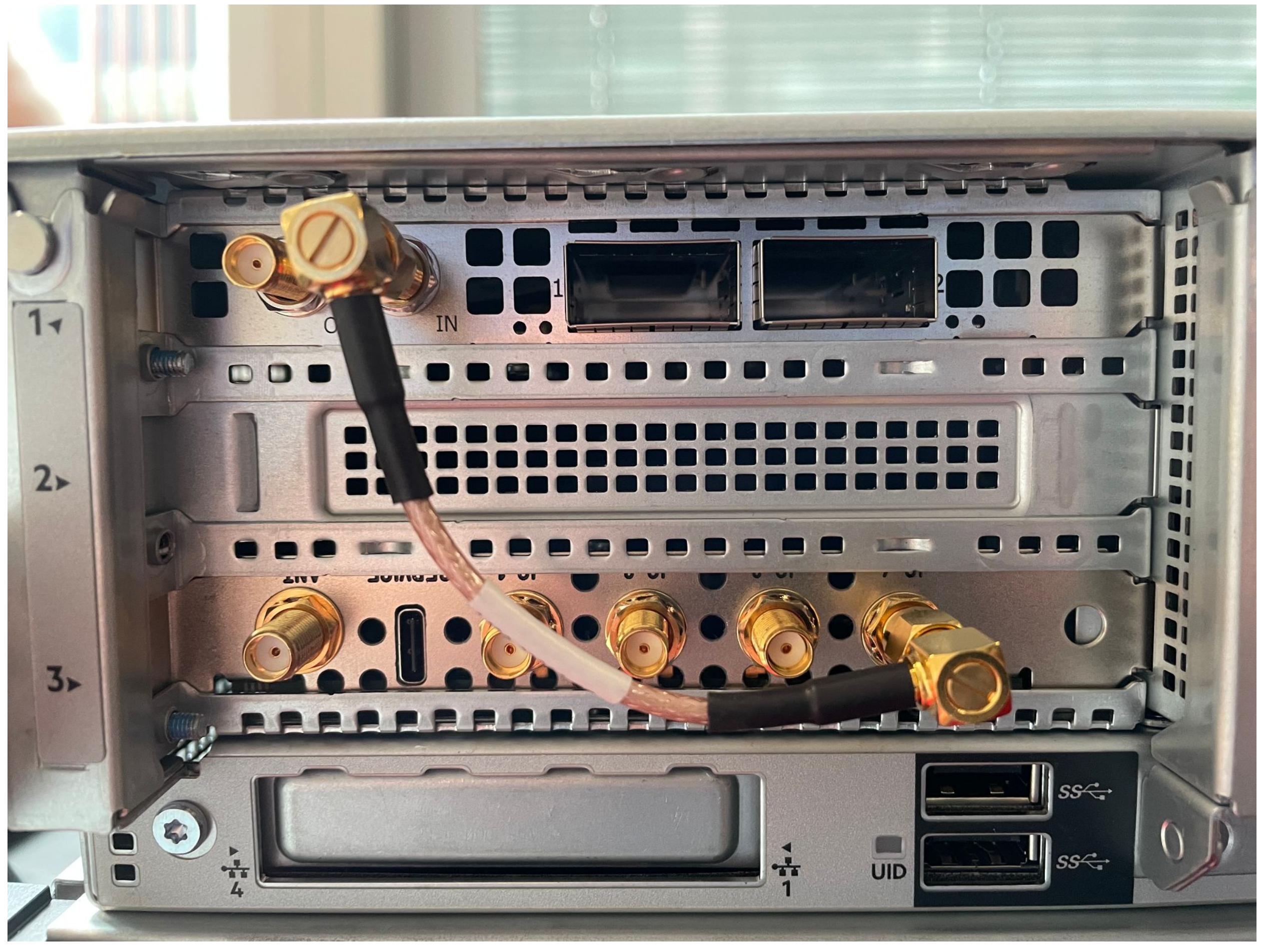


The Time Card

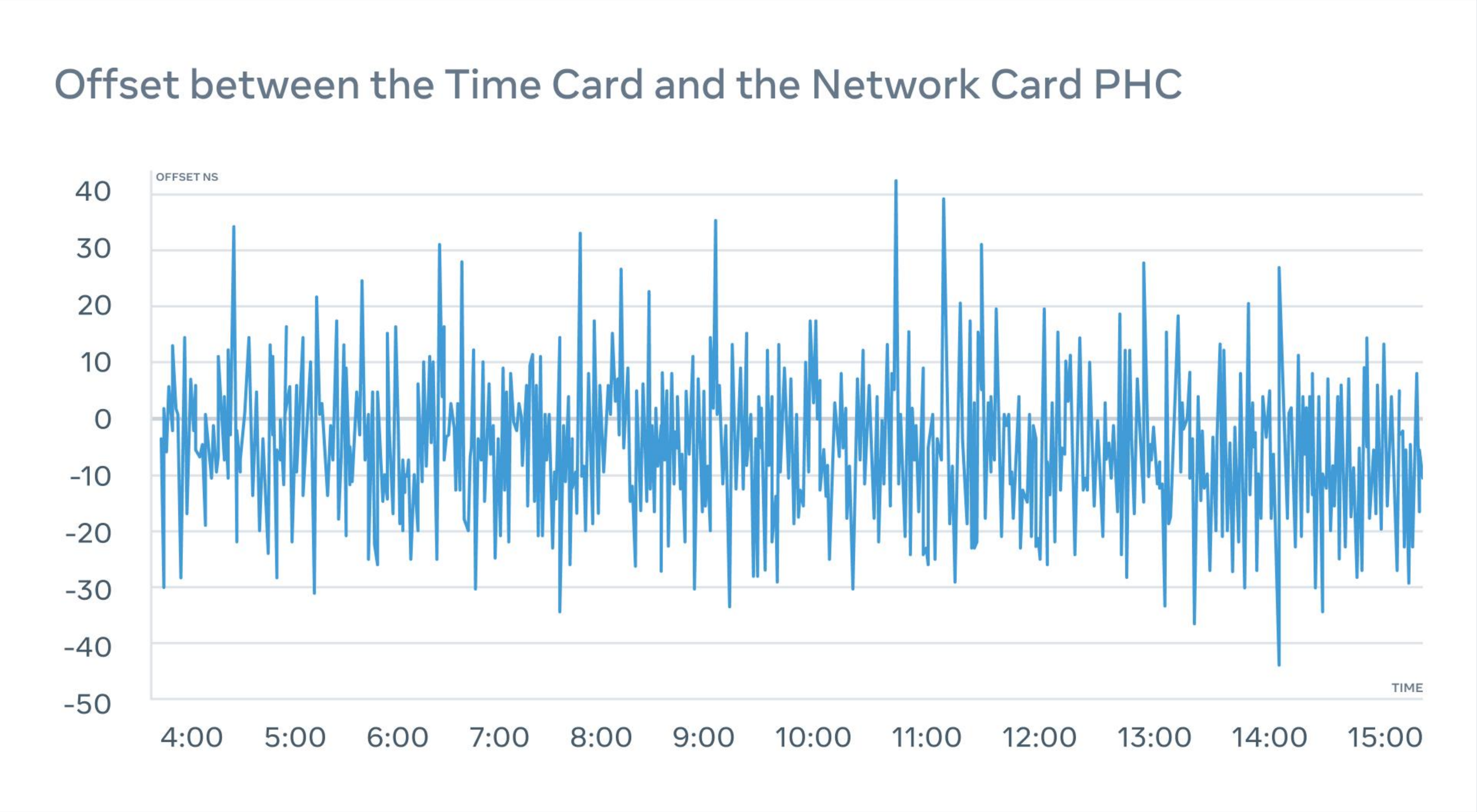


Frequency to temperature ratio of an atomic clock

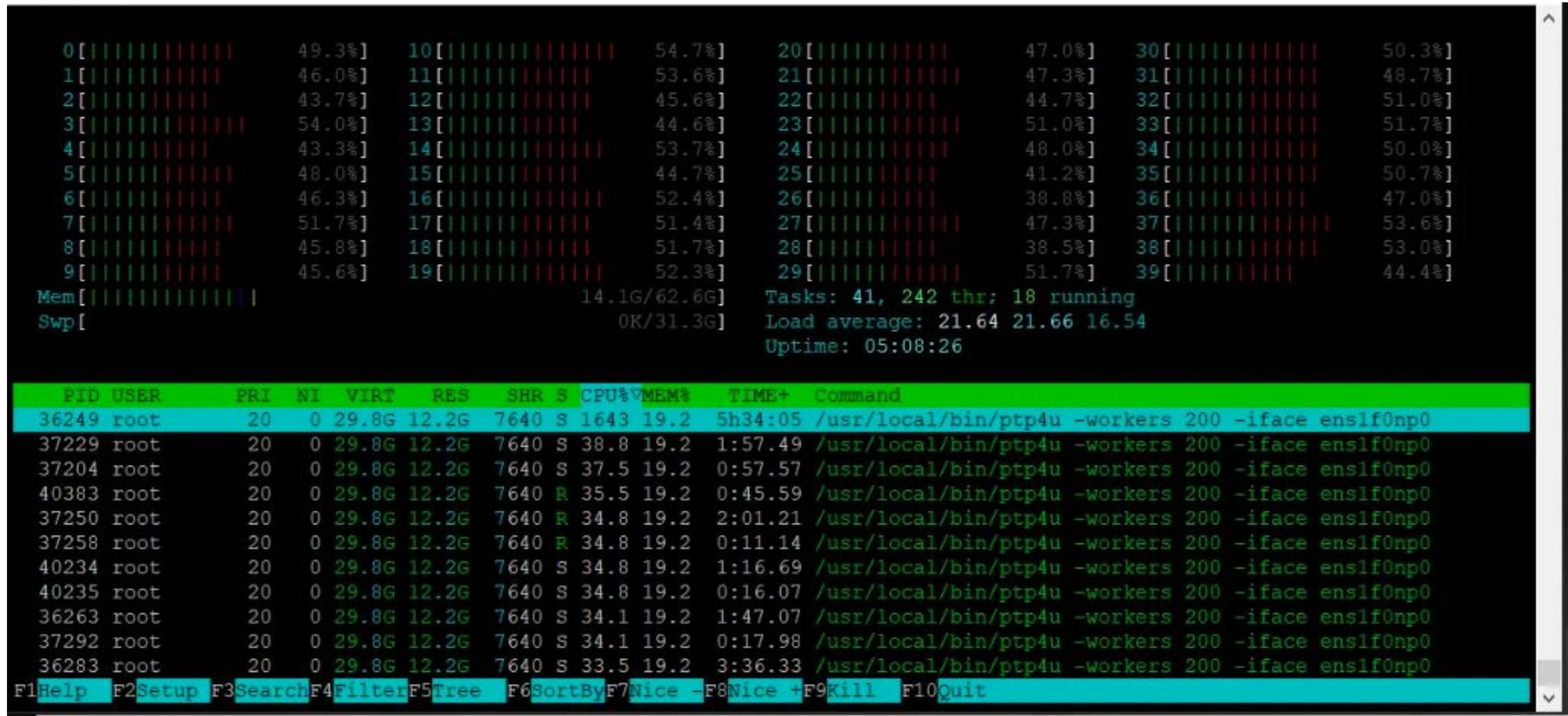
The Time Card



Short cable between PPS-out of the Time Card and PPS-in of the NIC

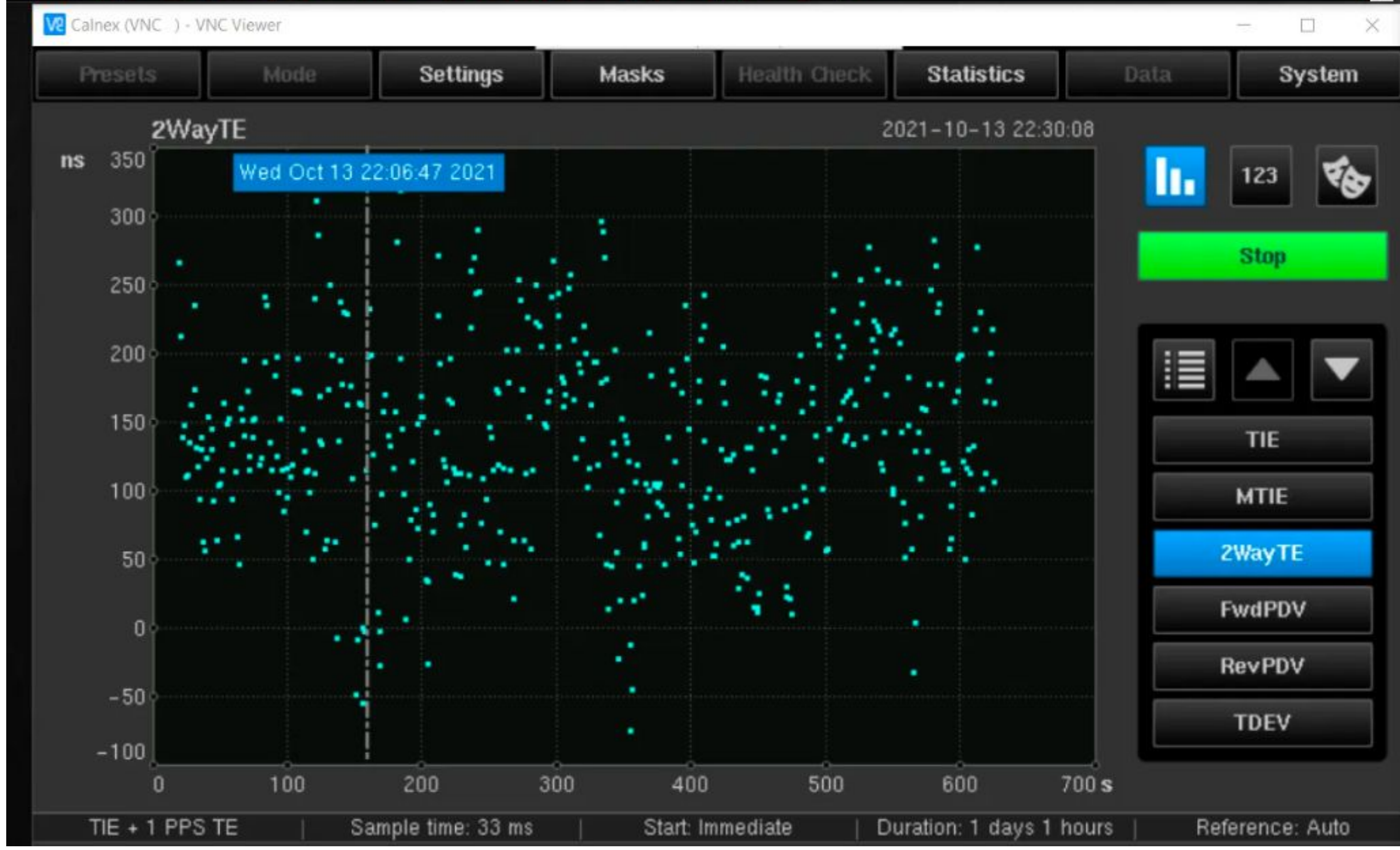


ptp4u

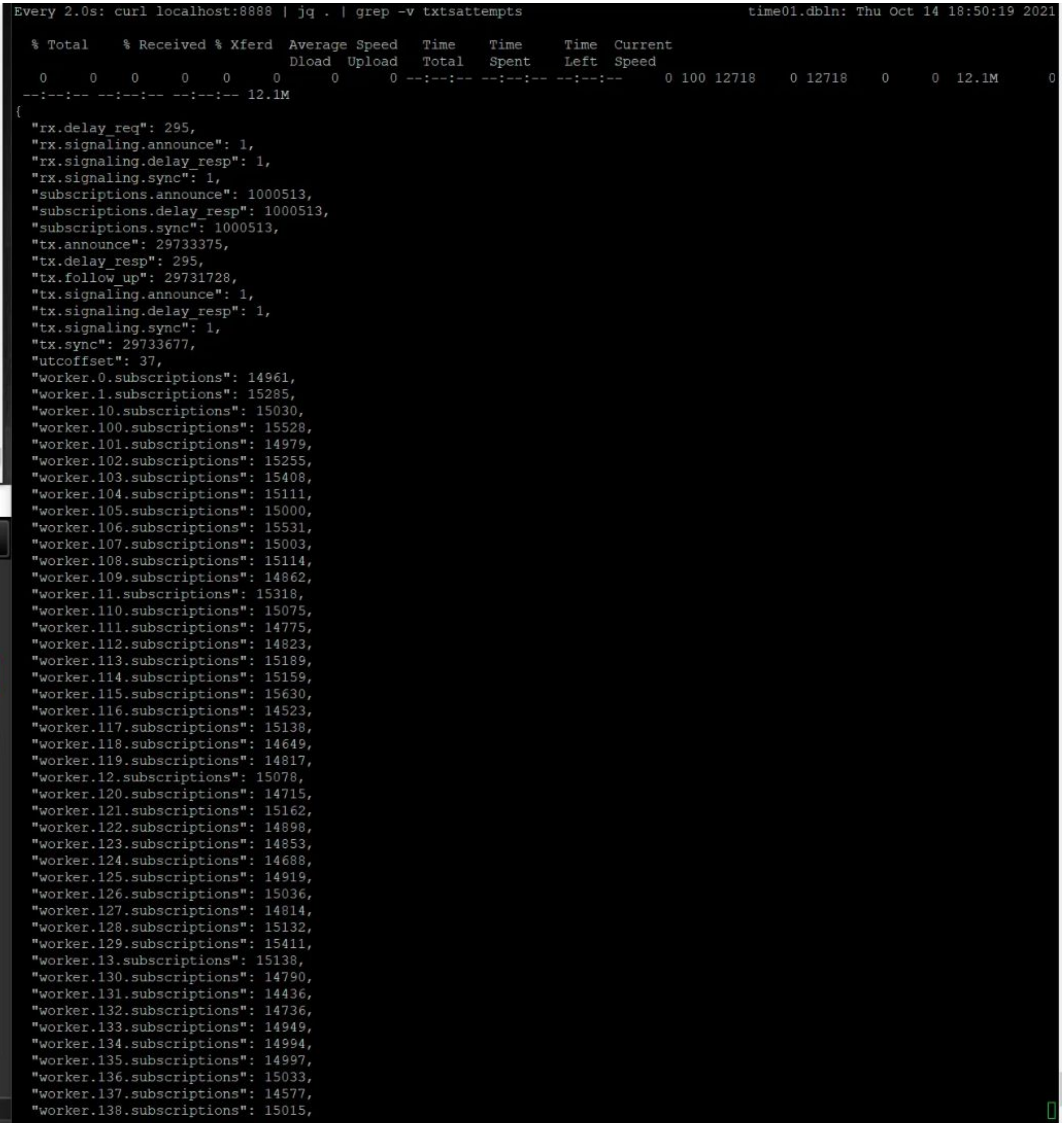


```
  0[||||||||||||||       49.3%]   10[||||||||||||||||     54.7%]   20[||||||||||||      47.0%]   30[||||||||||||||      50.3%]
  1[|||||||||||||        46.0%]   11[|||||||||||||||      53.6%]   21[||||||||||||||    47.3%]   31[||||||||||||||      48.7%]
  2[||||||||||           43.7%]   12[|||||||||||||||      45.6%]   22[|||||||||||       44.7%]   32[||||||||||||||      51.0%]
  3[||||||||||||||       54.0%]   13[||||||||||||||       44.6%]   23[||||||||||||||    51.0%]   33[||||||||||||||      51.7%]
  4[||||||||||           43.3%]   14[||||||||||||||||     53.7%]   24[||||||||||||      48.0%]   34[||||||||||||||      50.0%]
  5[|||||||||||||        48.0%]   15[||||||||||||||       44.7%]   25[|||||||||||       41.2%]   35[||||||||||||||      50.7%]
  6[||||||||||||         46.3%]   16[||||||||||||||||     52.4%]   26[|||||||||||       38.8%]   36[|||||||||||||       47.0%]
  7[|||||||||||||        51.7%]   17[|||||||||||||||      51.4%]   27[||||||||||||||    47.3%]   37[||||||||||||||||    53.6%]
  8[||||||||||||         45.8%]   18[|||||||||||||||      51.7%]   28[|||||||||||       38.5%]   38[|||||||||||||||     53.0%]
  9[||||||||||||         45.6%]   19[||||||||||||||||     52.3%]   29[||||||||||||||    51.7%]   39[||||||||||          44.4%]
Mem[|||||||||||||||                       14.1G/62.6G]   Tasks: 41, 242 thr; 18 running
Swp[                                          0K/31.3G]   Load average: 21.64 21.66 16.54
                                                          Uptime: 05:08:26

  PID USER      PRI  NI  VIRT   RES   SHR S CPU%▽MEM%   TIME+  Command
36249 root       20   0 29.8G 12.2G  7640 S 1643 19.2  5h34:05 /usr/local/bin/ptp4u -workers 200 -iface ens1f0np0
37229 root       20   0 29.8G 12.2G  7640 S 38.8 19.2  1:57.49 /usr/local/bin/ptp4u -workers 200 -iface ens1f0np0
37204 root       20   0 29.8G 12.2G  7640 S 37.5 19.2  0:57.57 /usr/local/bin/ptp4u -workers 200 -iface ens1f0np0
40383 root       20   0 29.8G 12.2G  7640 R 35.5 19.2  0:45.59 /usr/local/bin/ptp4u -workers 200 -iface ens1f0np0
37250 root       20   0 29.8G 12.2G  7640 R 34.8 19.2  2:01.21 /usr/local/bin/ptp4u -workers 200 -iface ens1f0np0
37258 root       20   0 29.8G 12.2G  7640 R 34.8 19.2  0:11.14 /usr/local/bin/ptp4u -workers 200 -iface ens1f0np0
40234 root       20   0 29.8G 12.2G  7640 S 34.8 19.2  1:16.69 /usr/local/bin/ptp4u -workers 200 -iface ens1f0np0
40235 root       20   0 29.8G 12.2G  7640 S 34.8 19.2  0:16.07 /usr/local/bin/ptp4u -workers 200 -iface ens1f0np0
36263 root       20   0 29.8G 12.2G  7640 S 34.1 19.2  1:47.07 /usr/local/bin/ptp4u -workers 200 -iface ens1f0np0
37292 root       20   0 29.8G 12.2G  7640 S 34.1 19.2  0:17.98 /usr/local/bin/ptp4u -workers 200 -iface ens1f0np0
36283 root       20   0 29.8G 12.2G  7640 S 33.5 19.2  3:36.33 /usr/local/bin/ptp4u -workers 200 -iface ens1f0np0
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```



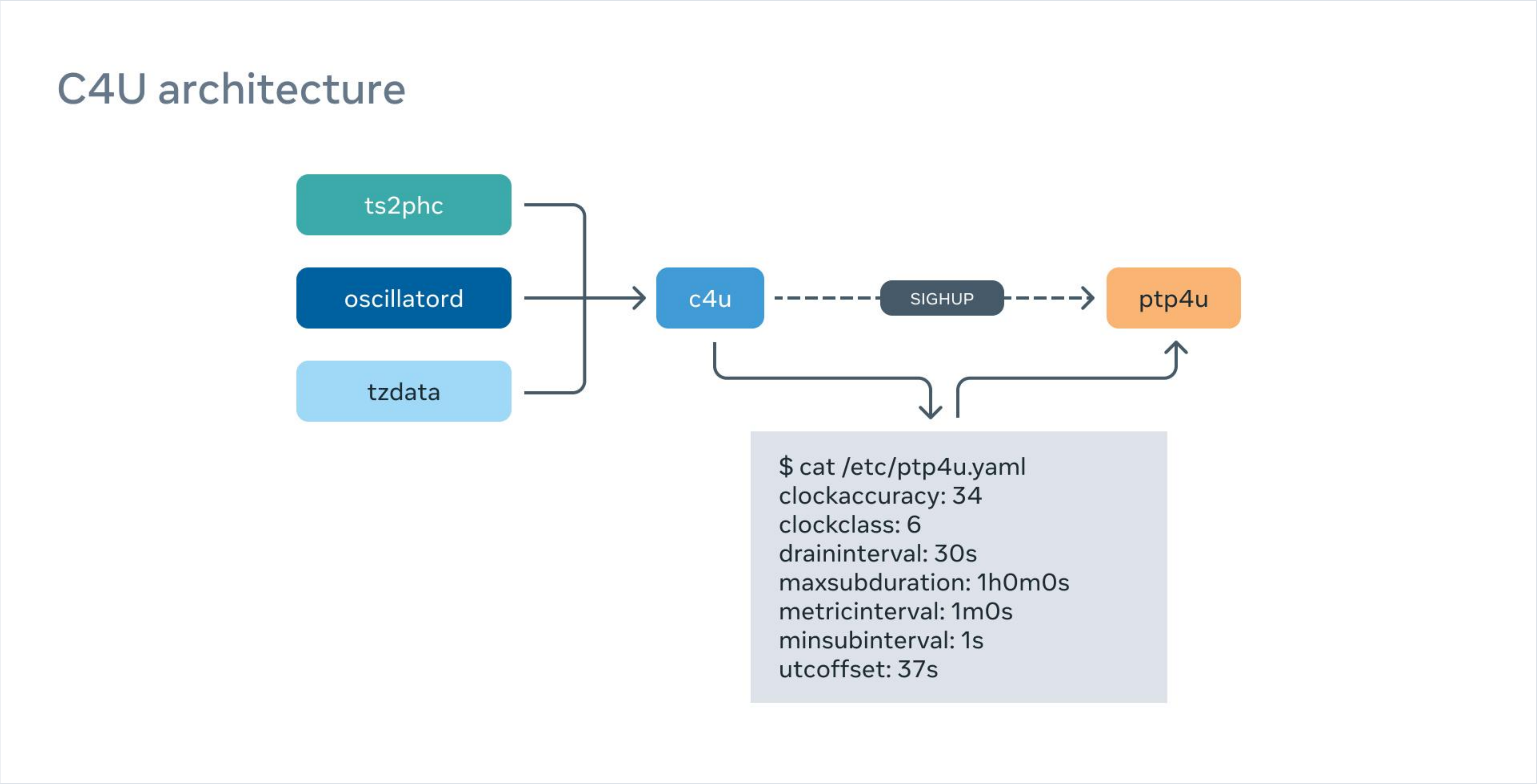
TIE + 1 PPS TE | Sample time: 33 ms | Start: Immediate | Duration: 1 days 1 hours | Reference: Auto



```
Every 2.0s: curl localhost:8888 | jq . | grep -v txtsattempts              time01.dbln: Thu Oct 14 18:50:19 2021

  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
  0     0    0     0    0     0      0      0 --:--:-- --:--:-- --:--:--     0 100 12718    0 12718    0     0  12.1M      0
 --:--:-- --:--:-- --:--:-- 12.1M
{
  "rx.delay_req": 295,
  "rx.signaling.announce": 1,
  "rx.signaling.delay_resp": 1,
  "rx.signaling.sync": 1,
  "subscriptions.announce": 1000513,
  "subscriptions.delay_resp": 1000513,
  "subscriptions.sync": 1000513,
  "tx.announce": 29733375,
  "tx.delay_resp": 295,
  "tx.follow_up": 29731728,
  "tx.signaling.announce": 1,
  "tx.signaling.delay_resp": 1,
  "tx.signaling.sync": 1,
  "tx.sync": 29733677,
  "utcoffset": 37,
  "worker.0.subscriptions": 14961,
  "worker.1.subscriptions": 15285,
  "worker.10.subscriptions": 15030,
  "worker.100.subscriptions": 15528,
  "worker.101.subscriptions": 14979,
  "worker.102.subscriptions": 15255,
  "worker.103.subscriptions": 15408,
  "worker.104.subscriptions": 15111,
  "worker.105.subscriptions": 15000,
  "worker.106.subscriptions": 15531,
  "worker.107.subscriptions": 15003,
  "worker.108.subscriptions": 15114,
  "worker.109.subscriptions": 14862,
  "worker.11.subscriptions": 15318,
  "worker.110.subscriptions": 15075,
  "worker.111.subscriptions": 14775,
  "worker.112.subscriptions": 14823,
  "worker.113.subscriptions": 15189,
  "worker.114.subscriptions": 15159,
  "worker.115.subscriptions": 15630,
  "worker.116.subscriptions": 14523,
  "worker.117.subscriptions": 15138,
  "worker.118.subscriptions": 14649,
  "worker.119.subscriptions": 14817,
  "worker.12.subscriptions": 15078,
  "worker.120.subscriptions": 14715,
  "worker.121.subscriptions": 15162,
  "worker.122.subscriptions": 14898,
  "worker.123.subscriptions": 14853,
  "worker.124.subscriptions": 14688,
  "worker.125.subscriptions": 14919,
  "worker.126.subscriptions": 15036,
  "worker.127.subscriptions": 14814,
  "worker.128.subscriptions": 15132,
  "worker.129.subscriptions": 15411,
  "worker.13.subscriptions": 15138,
  "worker.130.subscriptions": 14790,
  "worker.131.subscriptions": 14436,
  "worker.132.subscriptions": 14736,
  "worker.133.subscriptions": 14949,
  "worker.134.subscriptions": 14994,
  "worker.135.subscriptions": 14997,
  "worker.136.subscriptions": 15033,
  "worker.137.subscriptions": 14577,
  "worker.138.subscriptions": 15015,
```

Config 4 you (c4u)

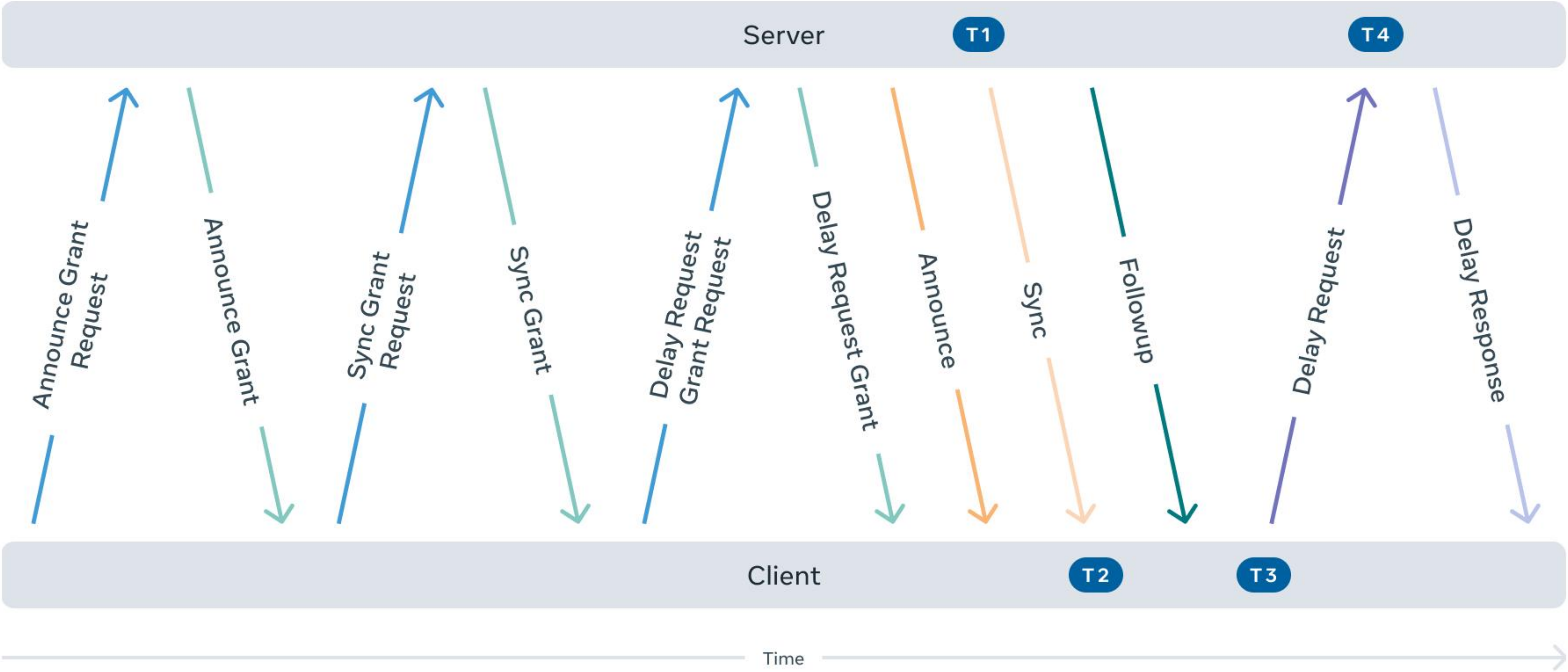


C4U architecture

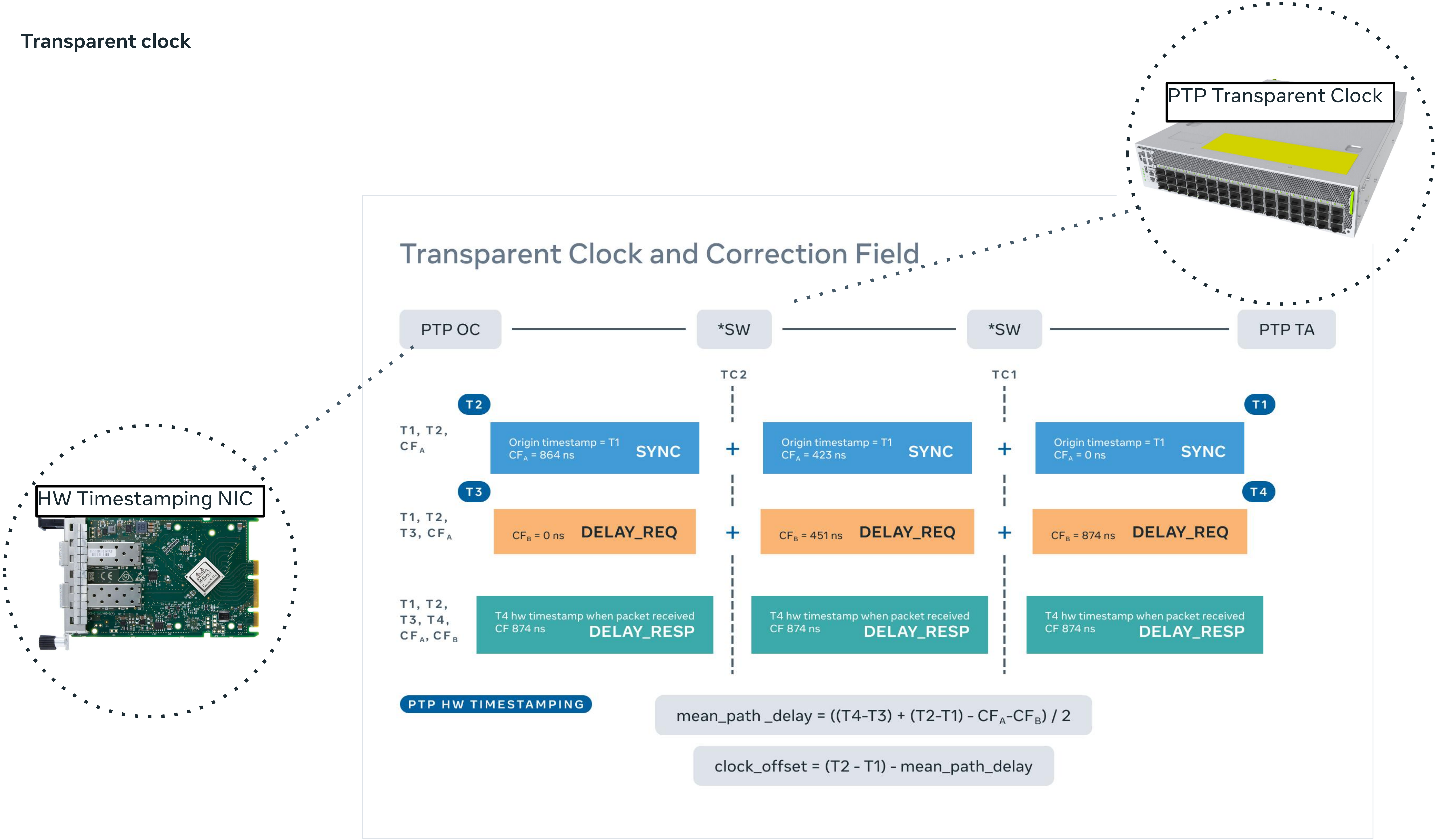
```
$ cat /etc/ptp4u.yaml
clockaccuracy: 34
clockclass: 6
draininterval: 30s
maxsubduration: 1h0m0s
metricinterval: 1m0s
minsubinterval: 1s
utcoffset: 37s
```

The PTP network

Two-step PTP exchange



Transparent clock



PTP Transparent Clock

HW Timestamping NIC

Transparent Clock and Correction Field

PTP OC — *SW — *SW — PTP TA

TC2

TC1

T1

T2

$T1, T2, CF_A$

Origin timestamp = T1 CF_A = 864 ns SYNC

Origin timestamp = T1 CF_A = 423 ns SYNC

Origin timestamp = T1 CF_A = 0 ns SYNC

T3

T4

$T1, T2, T3, CF_A$

CF_B = 0 ns DELAY_REQ

CF_B = 451 ns DELAY_REQ

CF_B = 874 ns DELAY_REQ

$T1, T2, T3, T4, CF_A, CF_B$

T4 hw timestamp when packet received CF 874 ns DELAY_RESP

T4 hw timestamp when packet received CF 874 ns DELAY_RESP

T4 hw timestamp when packet received CF 874 ns DELAY_RESP

PTP HW TIMESTAMPING

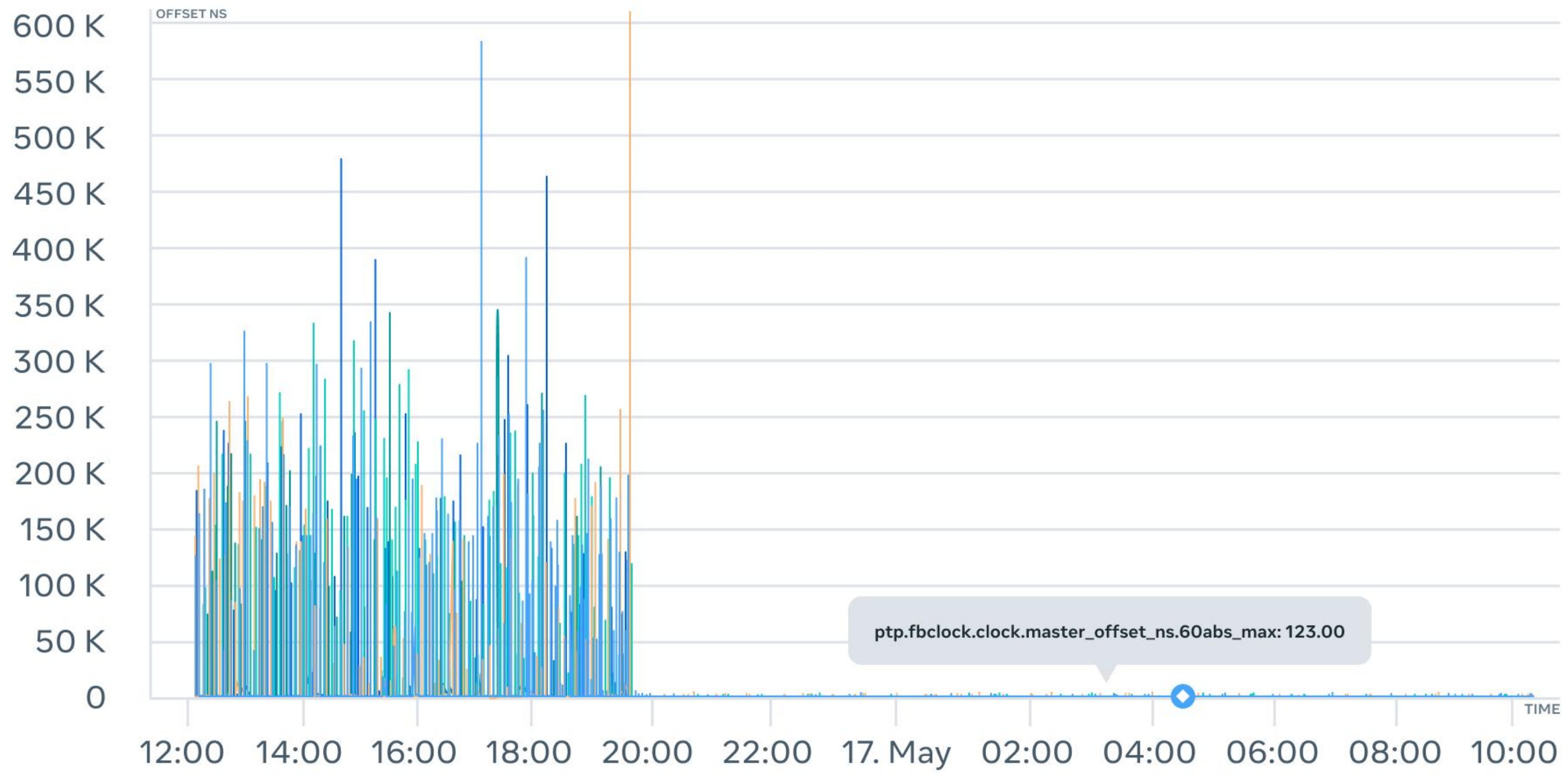
mean_path _delay = ((T4-T3) + (T2-T1) - CF_A-CF_B) / 2

clock_offset = (T2 - T1) - mean_path_delay

Transparent clock impact

```
ptp4l[43.662]: offset         -9 s2 freq  -12372 path delay      4114
ptp4l[44.662]: offset         17 s2 freq  -12349 path delay      4114
ptp4l[45.662]: offset         37 s2 freq  -12324 path delay      4078
ptp4l[46.662]: offset        -70 s2 freq  -12420 path delay      4153
ptp4l[47.662]: offset         95 s2 freq  -12276 path delay      4039
ptp4l[48.662]: offset     266776 s2 freq +254434 path delay      4181
ptp4l[49.662]: offset    -430864 s2 freq -363173 path delay    168255
ptp4l[50.662]: offset     -80141 s2 freq -141710 path delay    168255
ptp4l[51.662]: offset     217086 s2 freq +131475 path delay       408
ptp4l[52.662]: offset      16268 s2 freq   -4217 path delay     57459
ptp4l[53.662]: offset       8101 s2 freq   -7504 path delay     57459
ptp4l[54.662]: offset      55912 s2 freq  +42738 path delay      4776
ptp4l[56.305]: offset     -48984 s2 freq  -45385 path delay     19209
ptp4l[56.662]: offset     -37194 s2 freq  -48290 path delay     19209
ptp4l[57.662]: offset      29964 s2 freq   +7710 path delay    -12022
ptp4l[58.662]: offset       9943 s2 freq   -3322 path delay    -12022
ptp4l[59.662]: offset     -19403 s2 freq  -29685 path delay      8279
ptp4l[60.662]: offset       8560 s2 freq   -7543 path delay     -2377
ptp4l[61.662]: offset      -4906 s2 freq  -18441 path delay      6256
ptp4l[62.662]: offset       4197 s2 freq  -10810 path delay      3249
ptp4l[63.662]: offset        979 s2 freq  -12769 path delay      4917
ptp4l[64.662]: offset       1386 s2 freq  -12068 path delay      4917
ptp4l[65.662]: offset       1741 s2 freq  -11297 path delay      4270
ptp4l[66.662]: offset        509 s2 freq  -12007 path delay      4428
ptp4l[67.662]: offset        395 s2 freq  -11968 path delay      4185
ptp4l[68.662]: offset         -7 s2 freq  -12252 path delay      4185
```

Absolute offset values on hosts connected to the switch without Transparent Clock enabled



The PTP client

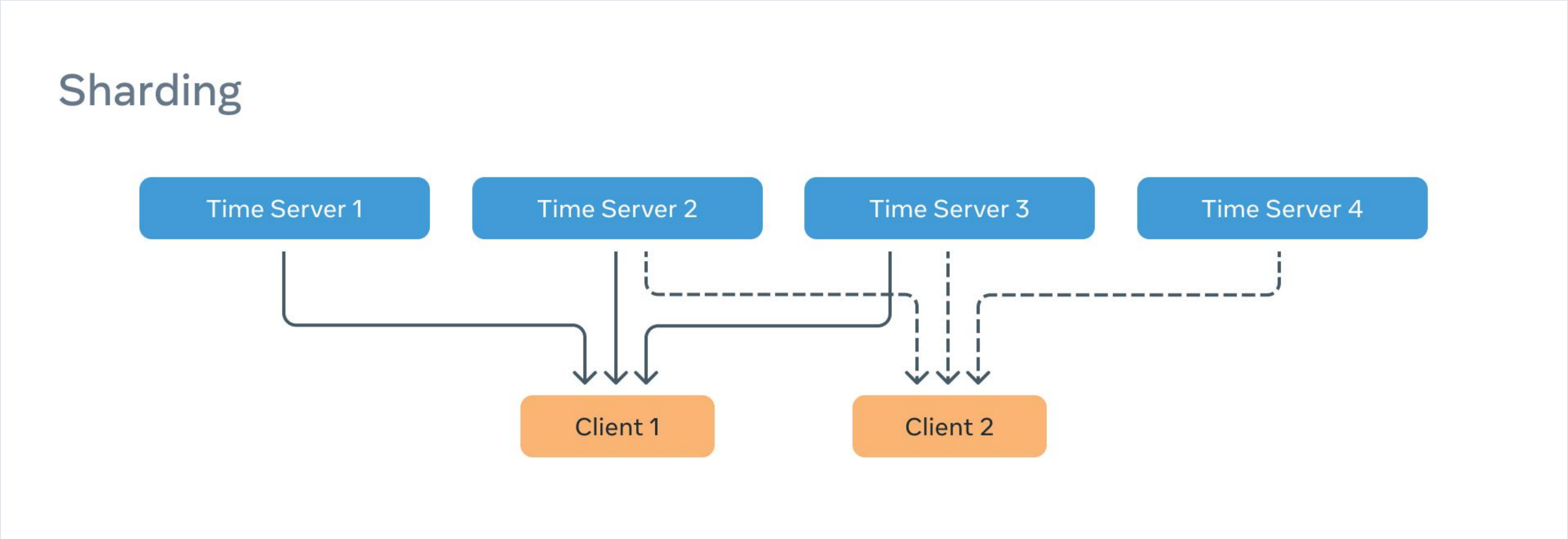
Hardware timestamps

```
$ ethtool -T eth0
Time stamping parameters for eth0:
Capabilities:
   hardware-transmit
   hardware-receive
   hardware-raw-clock
PTP Hardware Clock: 0
Hardware Transmit Timestamp Modes:
   off
   on
Hardware Receive Filter Modes:
   none
   All
```

128 bits	64 bits	64 bits	64 bits
Socket control message header	Software Timestamp	Legacy Timestamp	Hardware Timestamp

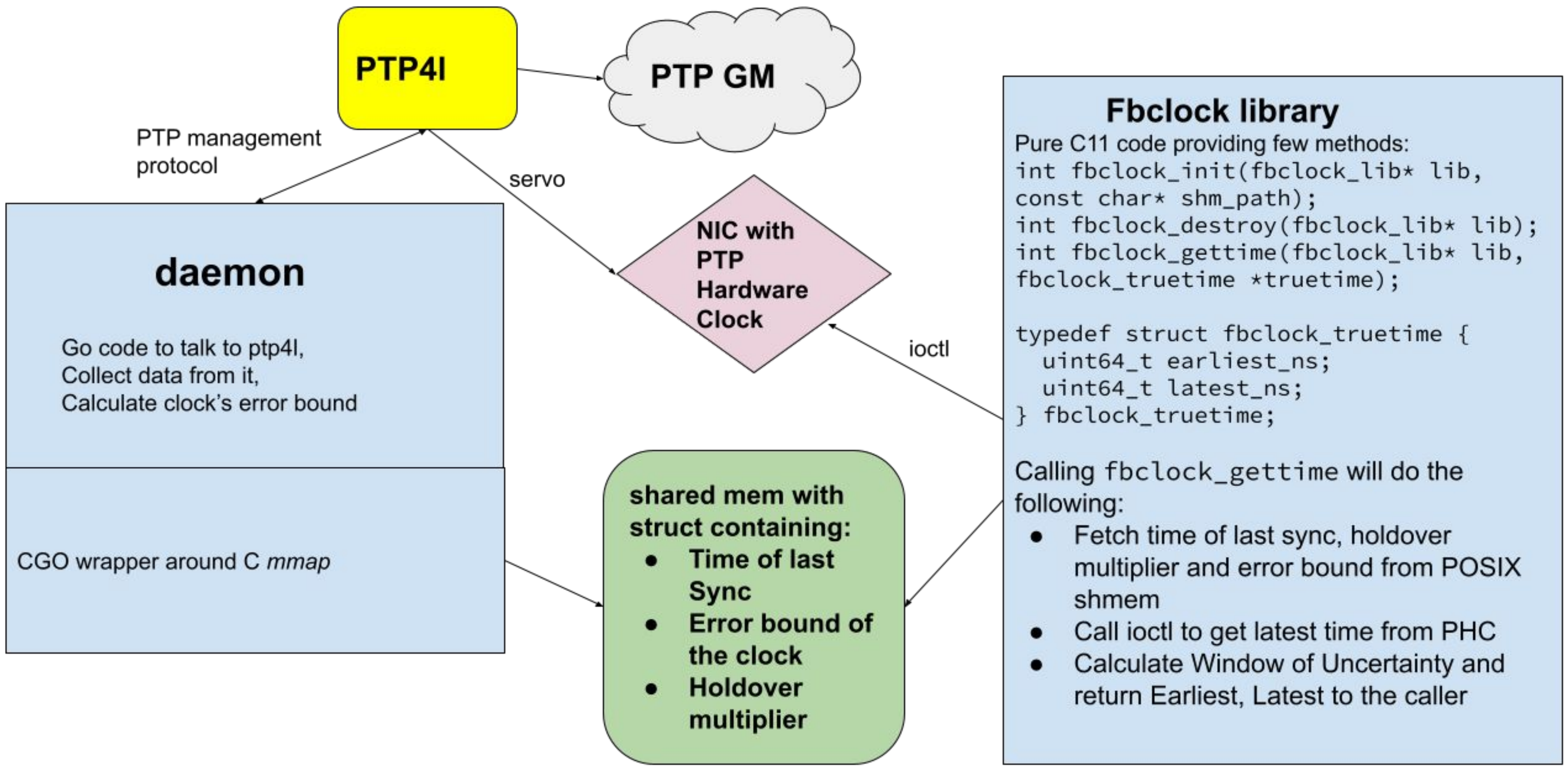
```
ptp4l[40.432]: offset        -16 s2 freq  -13105 path delay       3493
ptp4l[41.432]: offset         -6 s2 freq  -13100 path delay       3493
ptp4l[42.432]: offset          9 s2 freq  -13087 path delay       3493
ptp4l[43.432]: offset         -5 s2 freq  -13098 path delay       3493
ptp4l[44.432]: offset          1 s2 freq  -13093 path delay       3493
ptp4l[45.432]: spike detected => max_offset_locked: 33, setting offset to min_offset_freq_mean: -13065.039314
ptp4l[46.432]: skip 1/15 large offset (>33) 224401
ptp4l[47.432]: offset        -21 s2 freq  -13115 path delay       3493
ptp4l[48.432]: offset          9 s2 freq  -13091 path delay       3493
ptp4l[49.432]: offset         10 s2 freq  -13088 path delay       3493
ptp4l[50.432]: offset         -8 s2 freq  -13103 path delay       3493
```

ptp4l



Schematic representation of sharding

fbclock



daemon

Go code to talk to ptp4l,
Collect data from it,
Calculate clock's error bound

CGO wrapper around C *mmap*

shared mem with struct containing:

- **Time of last Sync**
- **Error bound of the clock**
- **Holdover multiplier**

Fbclock library

Pure C11 code providing few methods:

```
int fbclock_init(fbclock_lib* lib,
const char* shm_path);
int fbclock_destroy(fbclock_lib* lib);
int fbclock_gettime(fbclock_lib* lib,
fbclock_truetime *truetime);

typedef struct fbclock_truetime {
  uint64_t earliest_ns;
  uint64_t latest_ns;
} fbclock_truetime;
```

Calling `fbclock_gettime` will do the following:

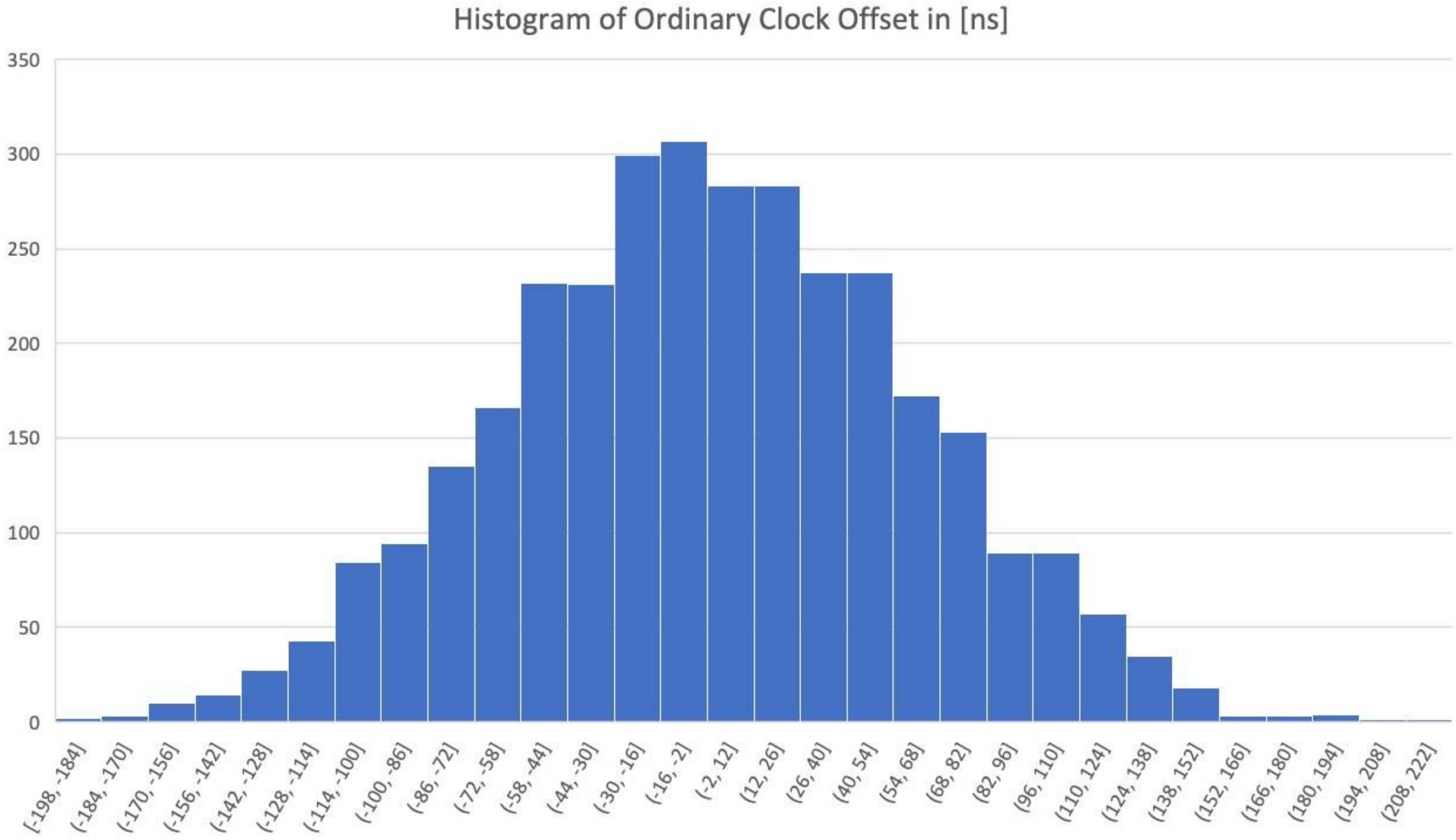
- Fetch time of last sync, holdover multiplier and error bound from POSIX shmem
- Call ioctl to get latest time from PHC
- Calculate Window of Uncertainty and return Earliest, Latest to the caller

fbclock

```
Estimated E2E Variance = [GNSS Variance + MAC Variance + ts2phc Variance] + [PTP4L Offset Variance] = [Time Server Variance] + [Ordinary Clock Variance]
```

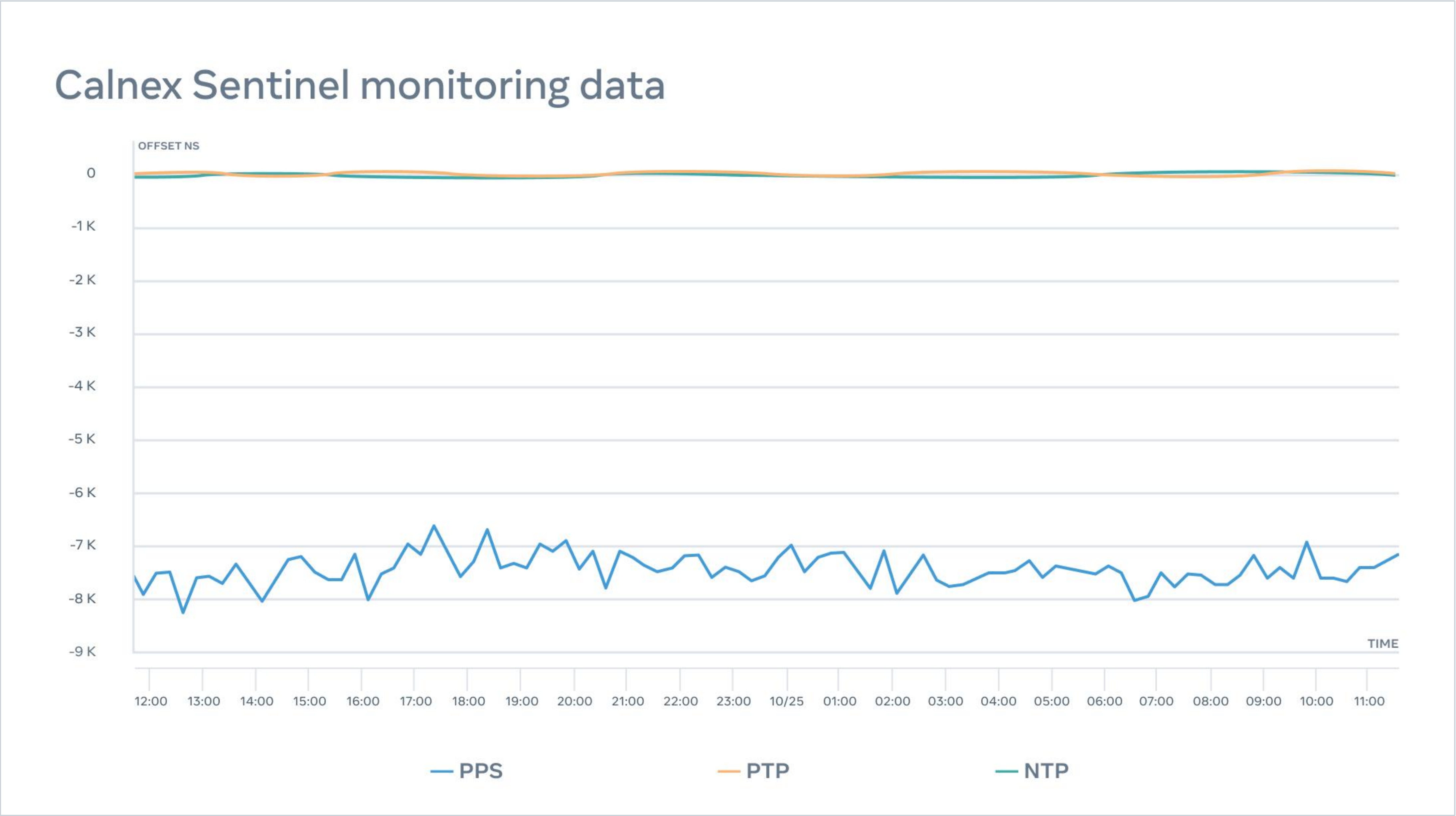
Estimated E2E Variance = (12ns ^ 2) + (43ns ^ 2) + (52ns ^2) + (61ns ^2) = 8418 which corresponds to 91.7 ns

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)$$



How we monitor the PTP architecture

Calnex



Calnex Sentinel monitoring data

ptpcheck

```
$ ptpcheck diag
[ OK ] GM is present
[ OK ] Period since last ingress is 972.752664ms, we expect it to be within 1s
[ OK ] GM offset is 67ns, we expect it to be within 250µs
[ OK ] GM mean path delay is 3.495µs, we expect it to be within 100ms
[ OK ] Sync timeout count is 1, we expect it to be within 100
[ OK ] Announce timeout count is 0, we expect it to be within 100
[ OK ] Sync mismatch count is 0, we expect it to be within 100
```

```
$ ptpcheck fbclock
{"earliest_ns":1654191885711023134,"latest_ns":1654191885711023828,"wou_ns":694}
```

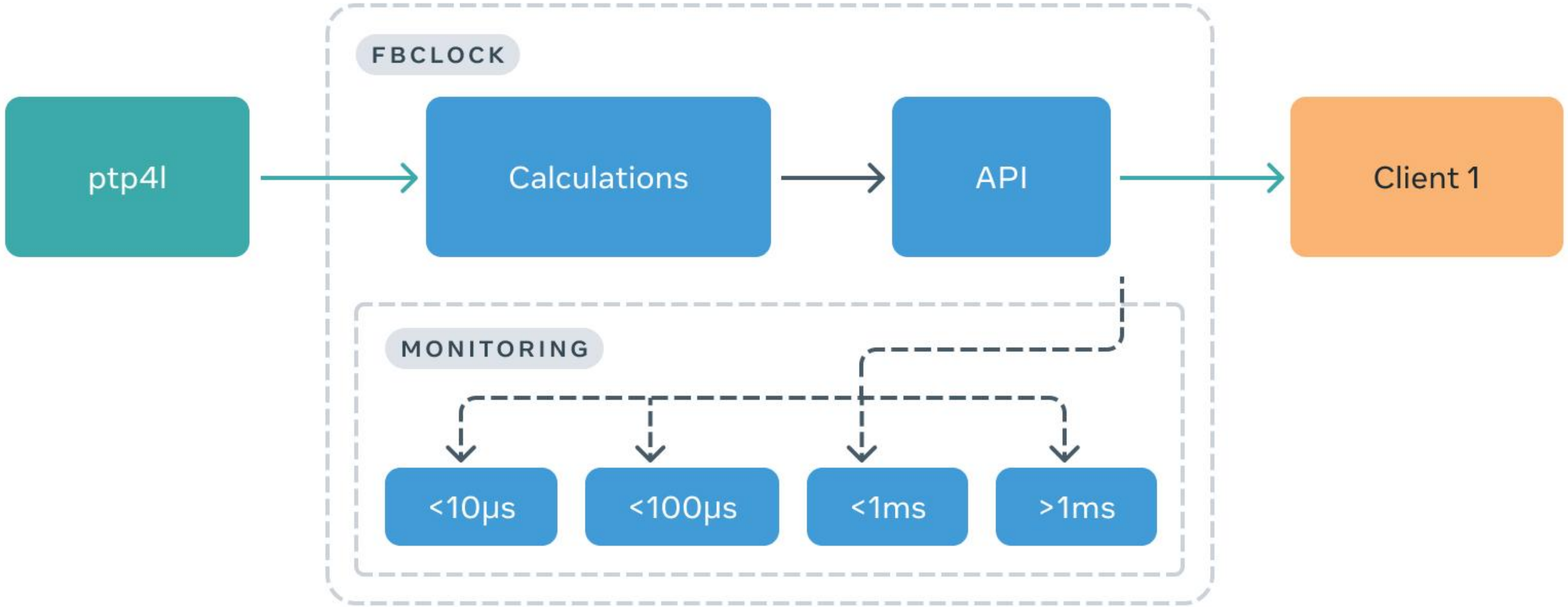
```
$ ptpcheck phcdiff -d /dev/ptp0 -d /dev/ptp2
PHC offset: -15ns
Delay for PHC1: 358ns
Delay for PHC2: 2.588µs
```

```
$ ptpcheck sources
+----------+----------------------+--------------------------+-----------+--------+----------+---------+------------+-----------+--------------+
| SELECTED |       IDENTITY       |          ADDRESS         |   STATE   | CLOCK  | VARIANCE |  P1:P2  | OFFSET(NS) | DELAY(NS) |  LAST SYNC   |
+----------+----------------------+--------------------------+-----------+--------+----------+---------+------------+-----------+--------------+
| true     | abcdef.fffe.111111-1 | time01.example.com.      | HAVE_SYDY | 6:0x22 | 0x59e0   | 128:128 |         27 |      3341 | 868.729197ms |
| false    | abcdef.fffe.222222-1 | time02.example.com.      | HAVE_ANN  | 6:0x22 | 0x59e0   | 128:128 |            |           |              |
| false    | abcdef.fffe.333333-1 | time03.example.com.      | HAVE_ANN  | 6:0x22 | 0x59e0   | 128:128 |            |           |              |
```

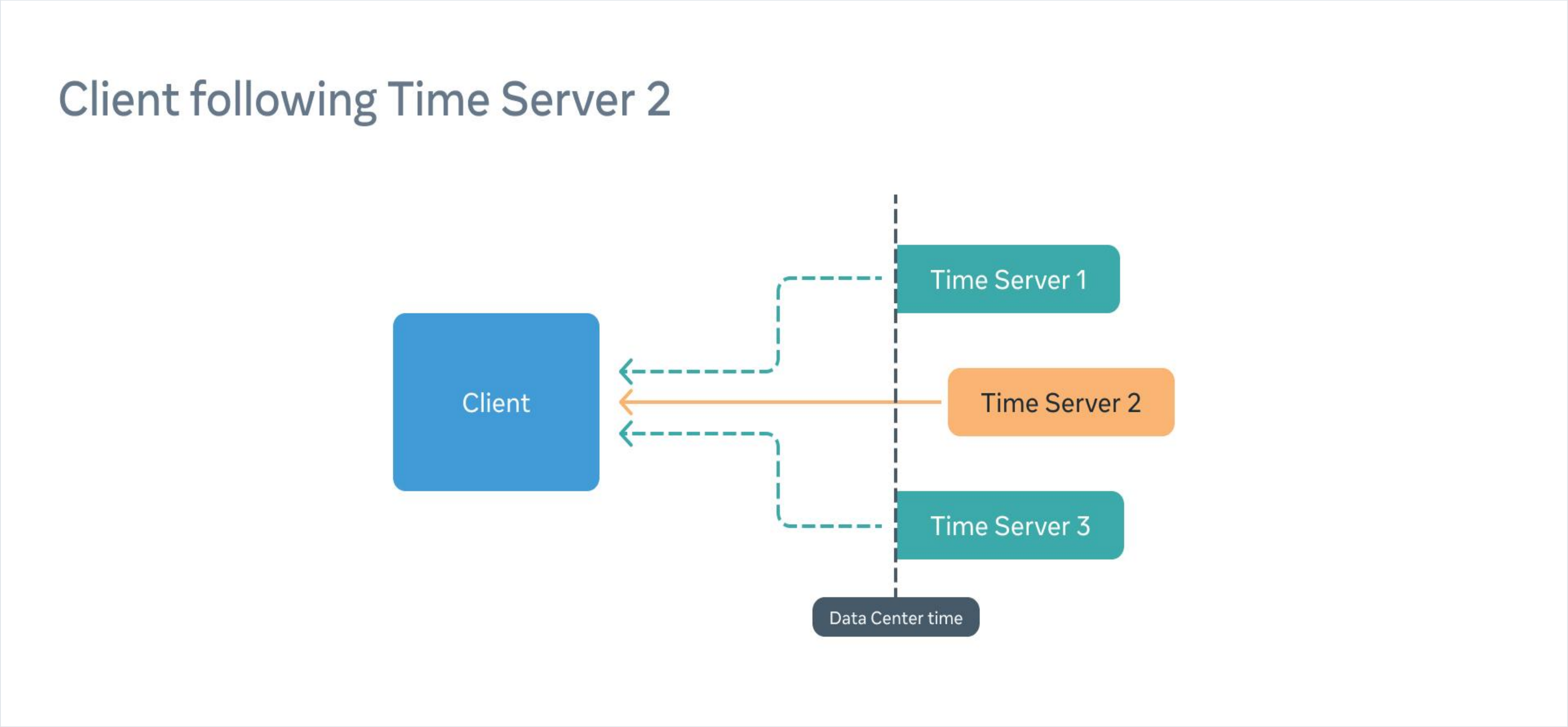
API monitoring

Fbclock API monitoring

µs microsecond   ms millisecond



Linearizability check



Client following Time Server 2

Client

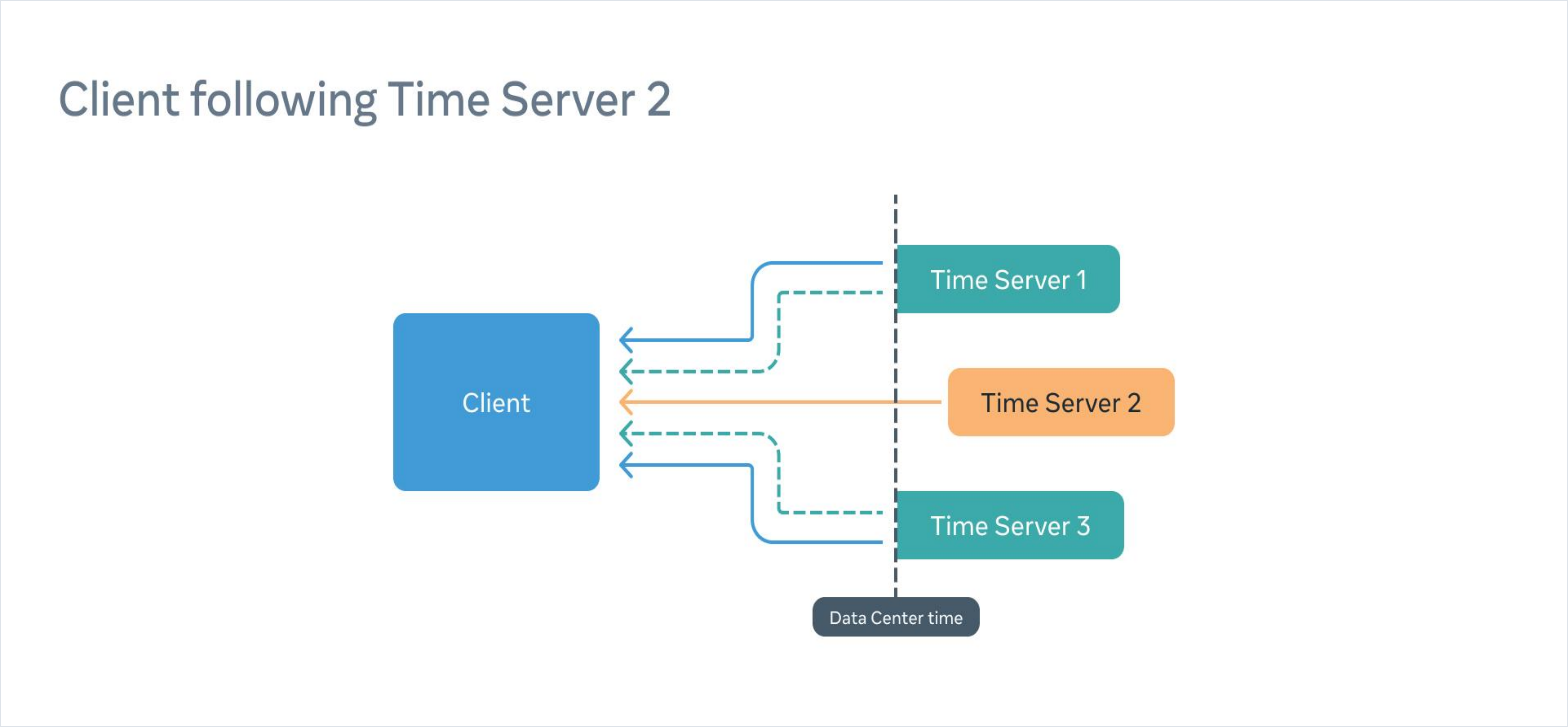
Time Server 1

Time Server 2

Time Server 3

Data Center time

Linearizability check



Client following Time Server 2

Thank you

Meta