DEUTSCHE BÖRSE GROUP

White Rabbit and Beyond

Andreas Lohr and Sebastian Neusüß

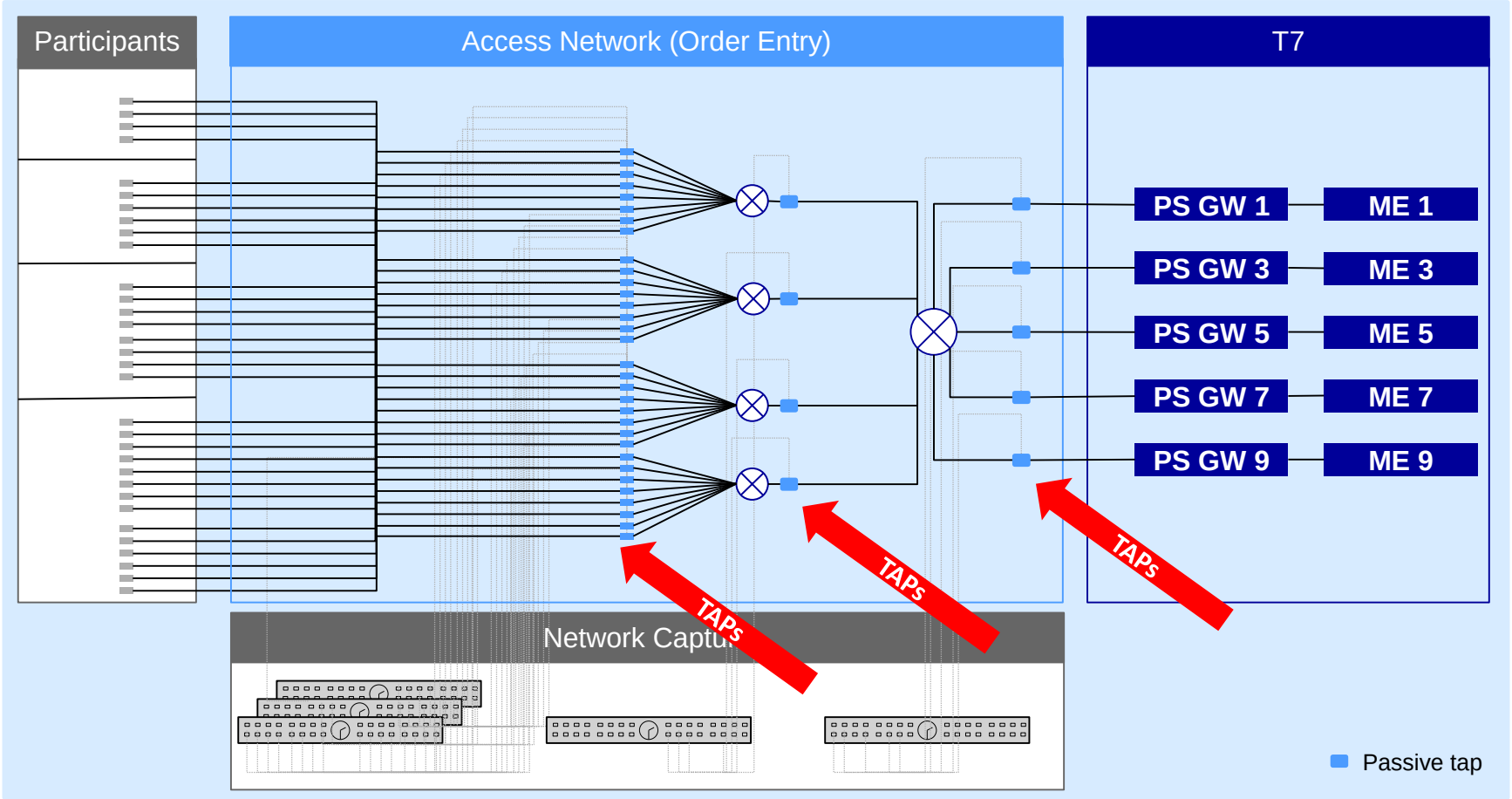
STAC Summit | 15 Nov 2018 | London

Problem Statement

How to capture and timestamp all customer cross-connects in co-location?

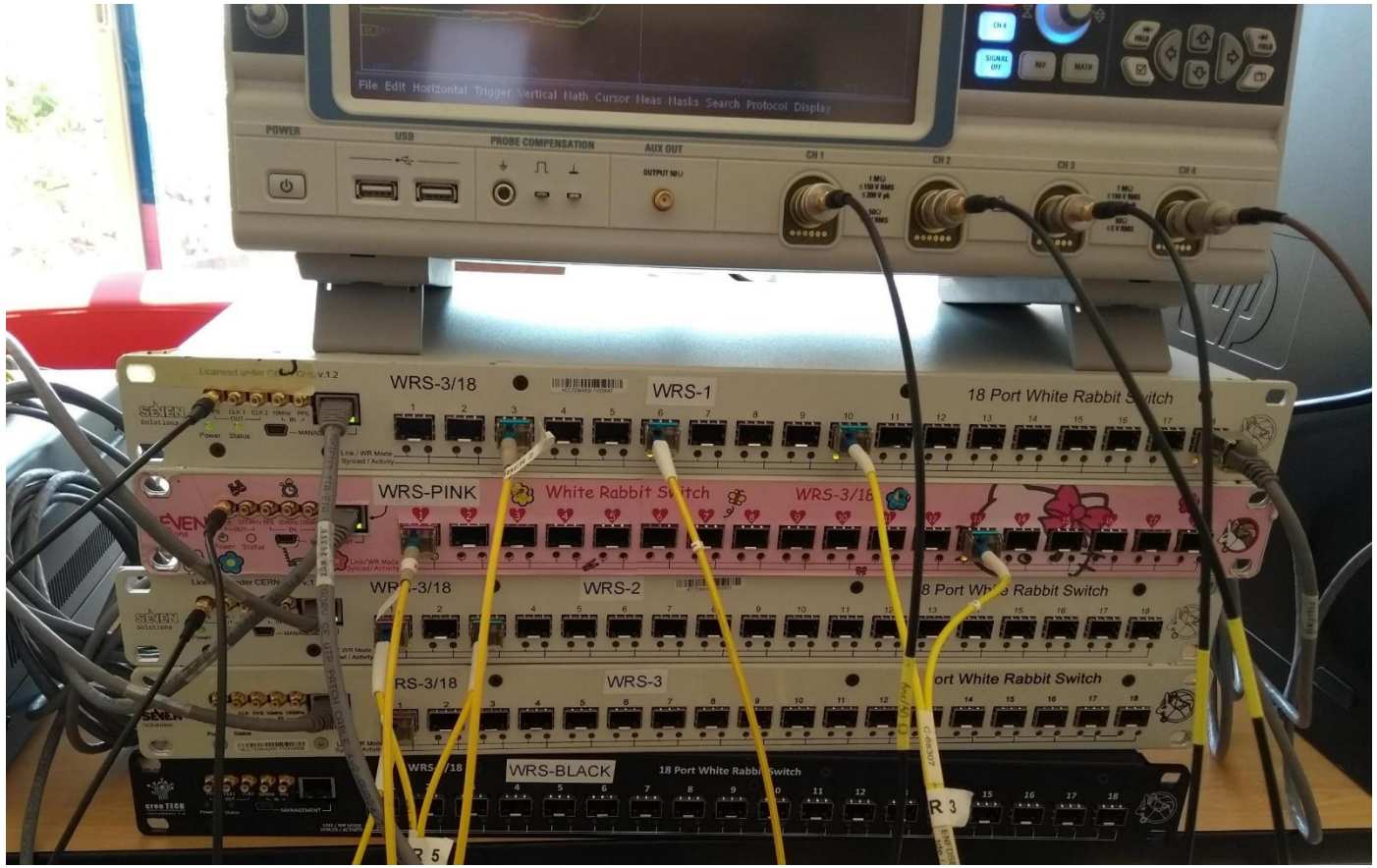
- 500+ capture ports
- 60+ capture devices
- 4 datacenter modules

- PTP +/- 60ns jitter in our infrastructure at best
- Serialization time of order entry message = 120ns
- Goal of sub-10ns precision
- Distances too long for PPS over coax cables



What is White Rabbit?

- White Rabbit is a fully deterministic Ethernet-based network for **data transfer** and **time synchronization**
- Initially developed at CERN
- Provides sub-nanosecond accuracy and picoseconds precision of synchronization
- Tried and tested
- PTP over Synchronous Ethernet
- White Rabbit is the High Accuracy profile of the future standard IEEE1588-20XX
- Commercially available (except the Hello Kitty model)



Issue for us:

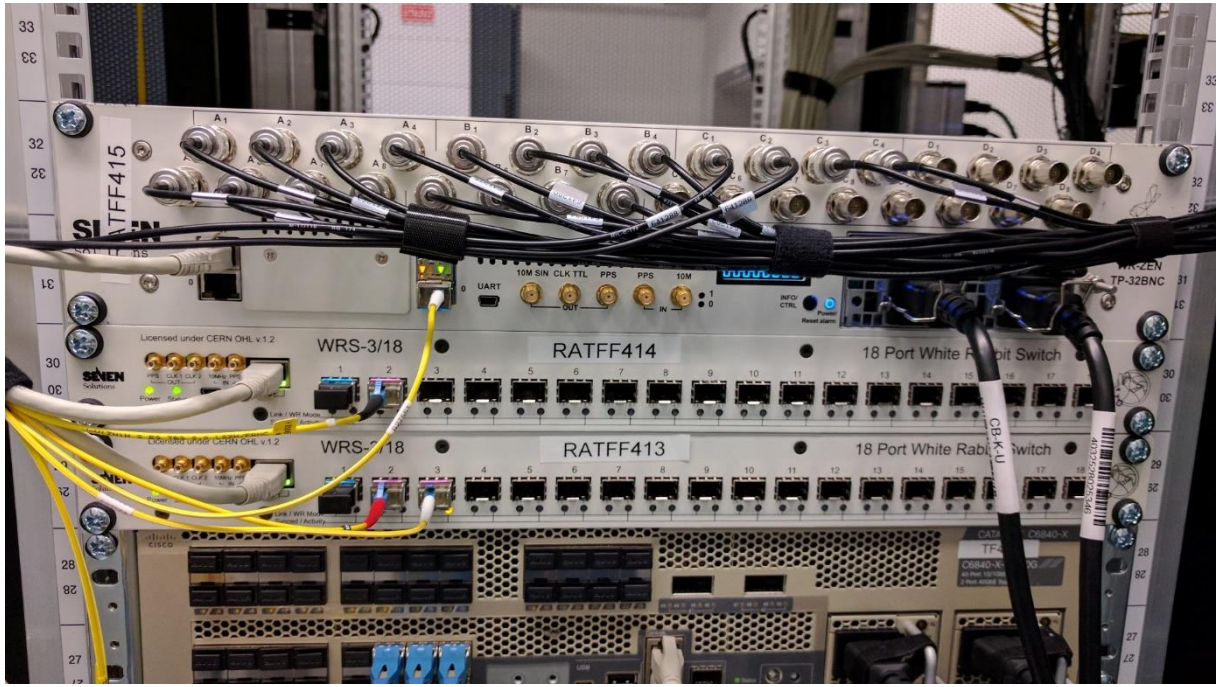
No native White Rabbit support in NICs and switches

Solution:

We use White Rabbit to distribute 1PPS

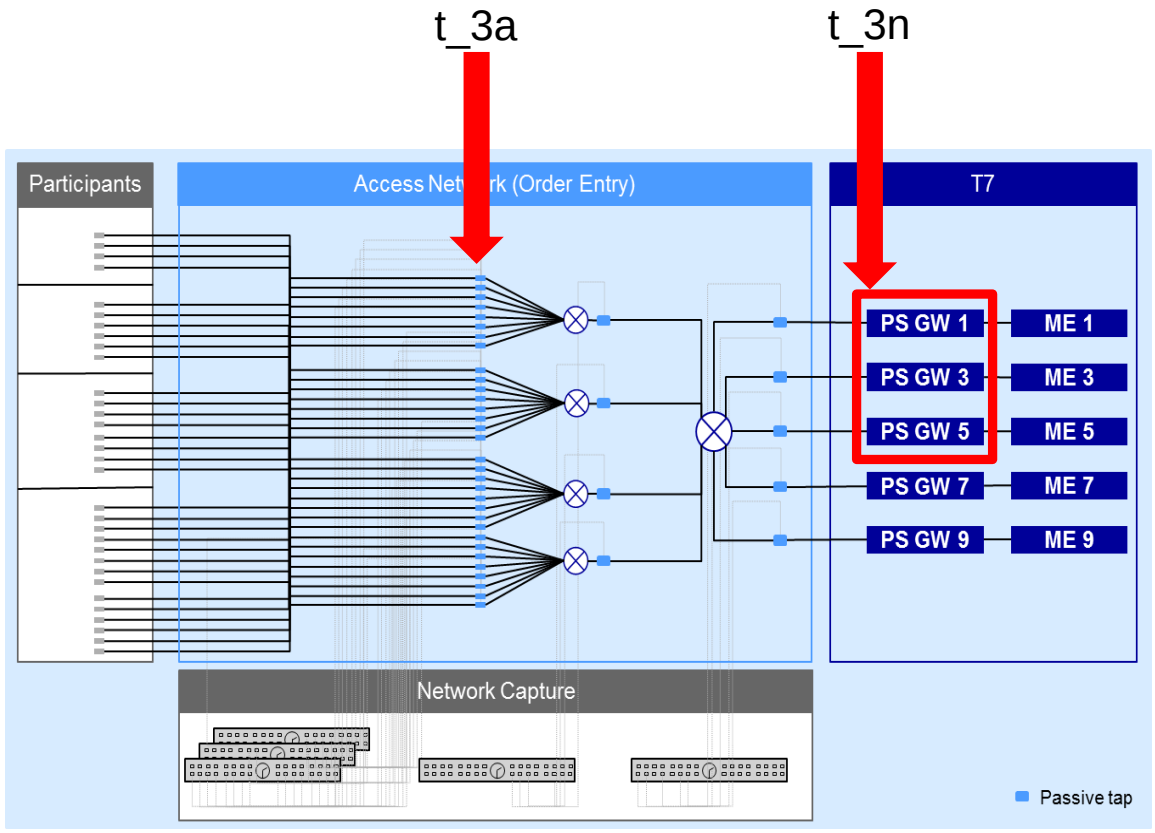
White Rabbit deployed hardware

Timestamping devices synchronized by 1PPS over White Rabbit



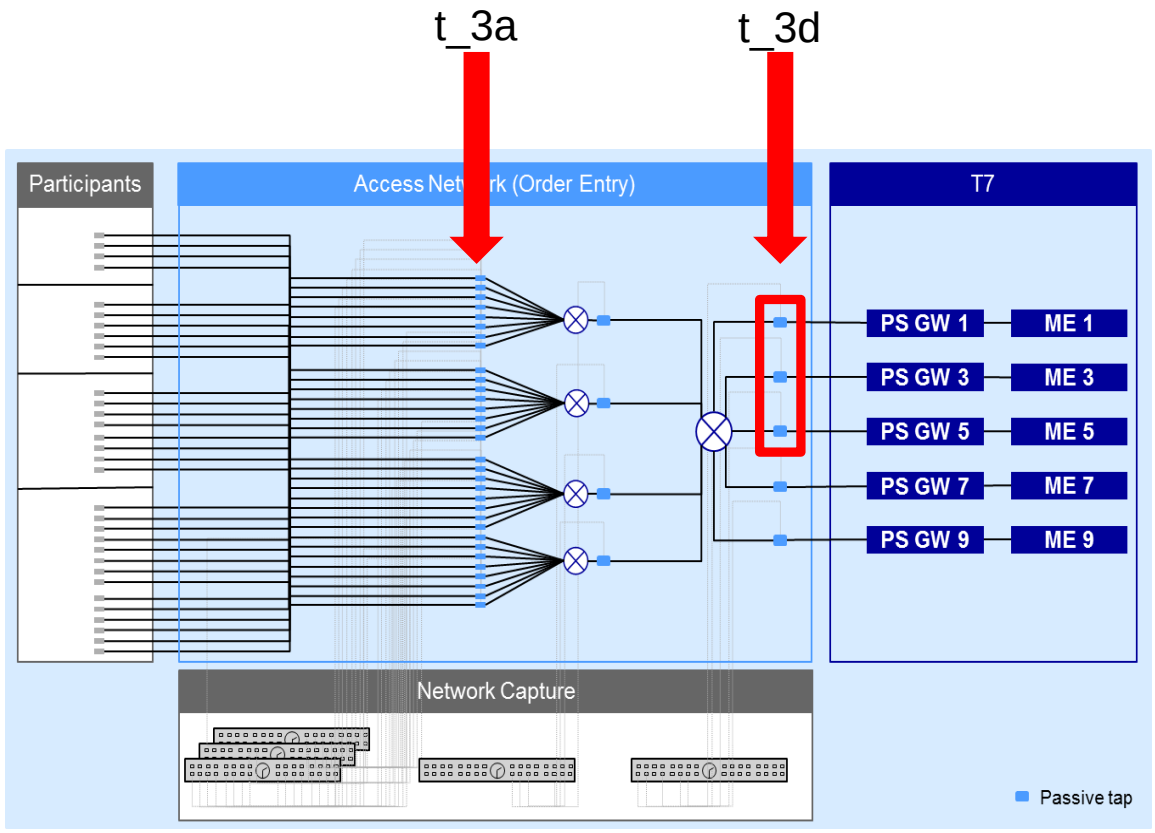
Time Synchronisation

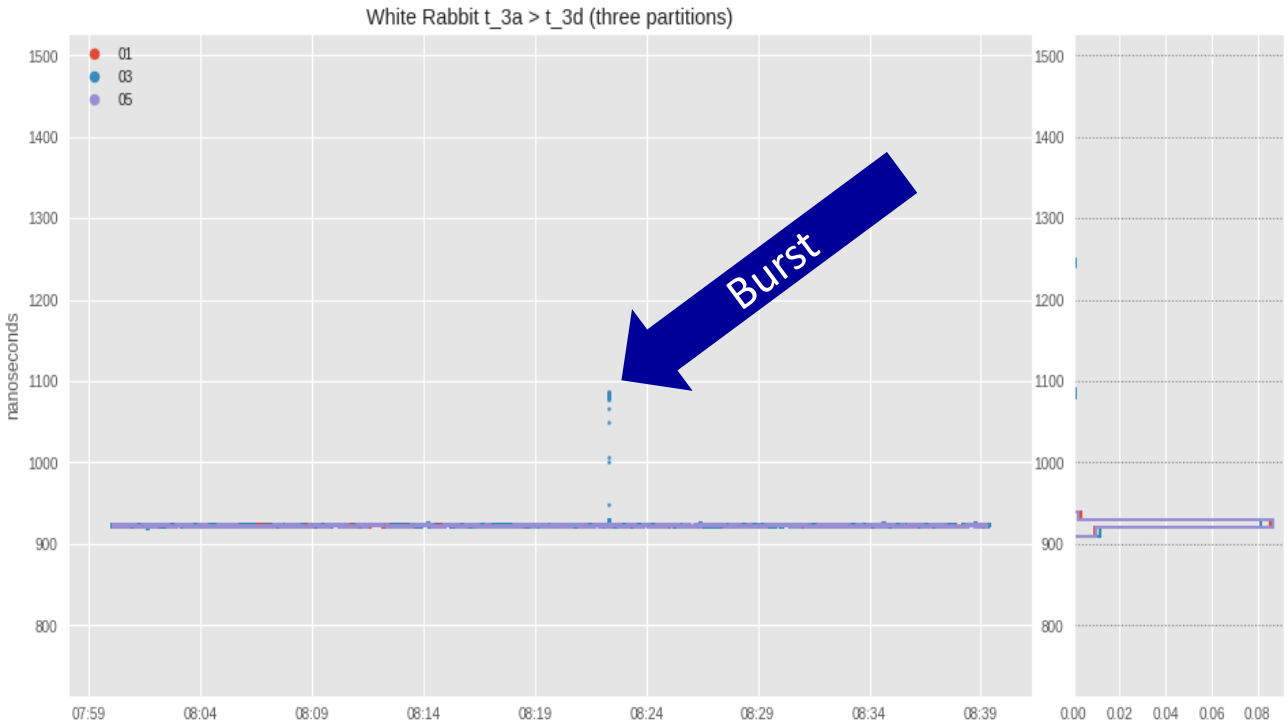
PTP



Time Synchronisation

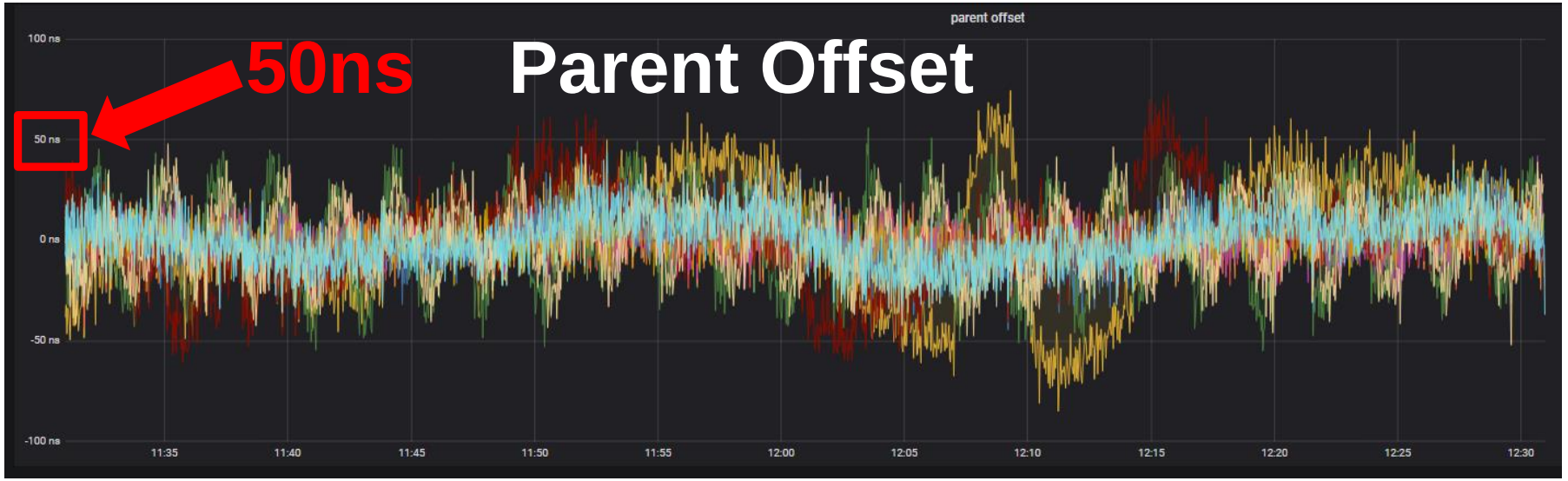
White Rabbit



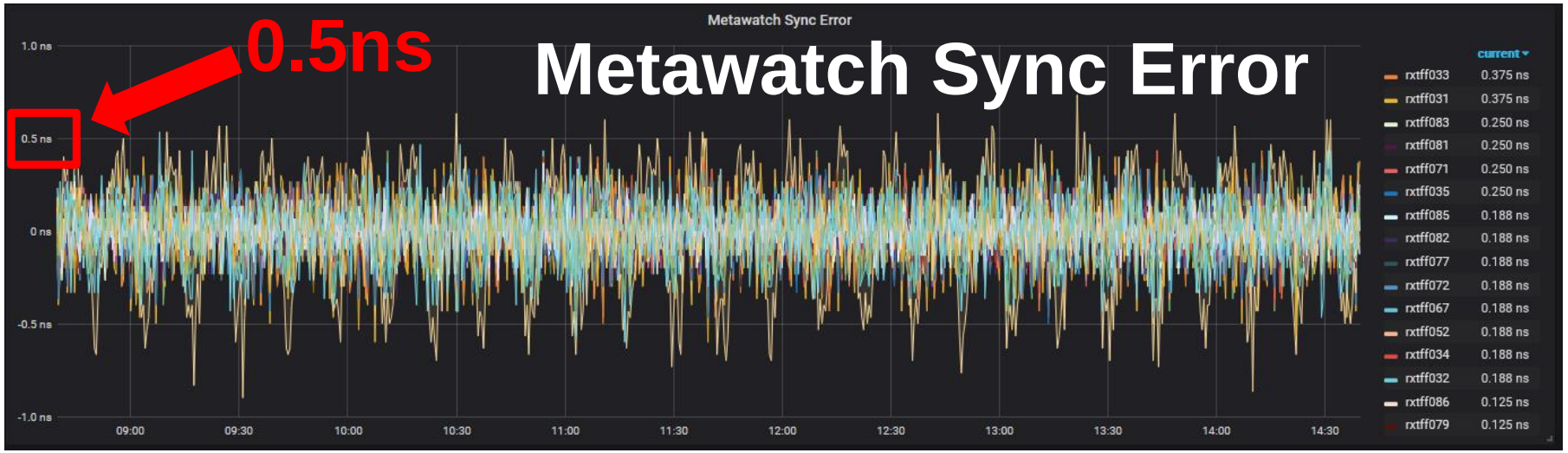


PTP / White Rabbit

PTP



White Rabbit



Precise Timestamping

Burst Analysis

Networking at our exchange is characterised by bursts

The most frequently asked question:

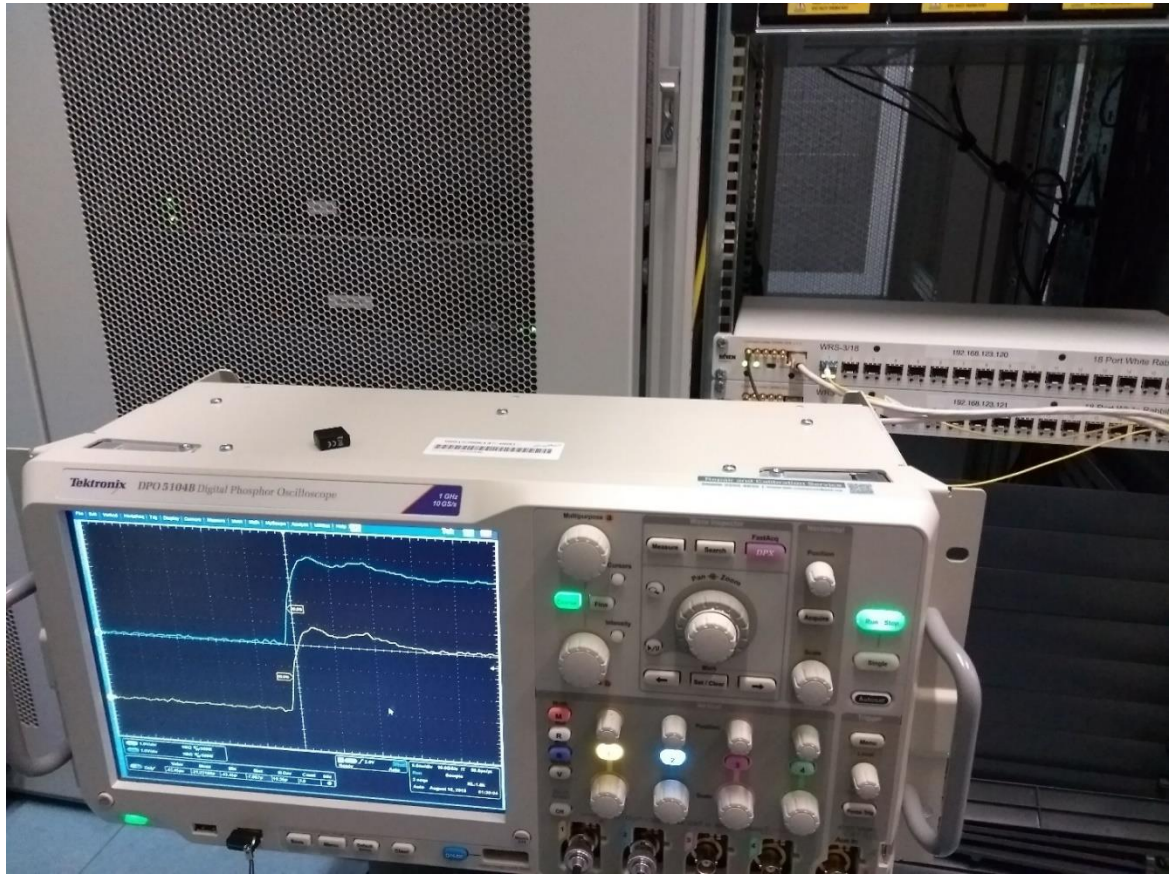
How far behind am I, really?

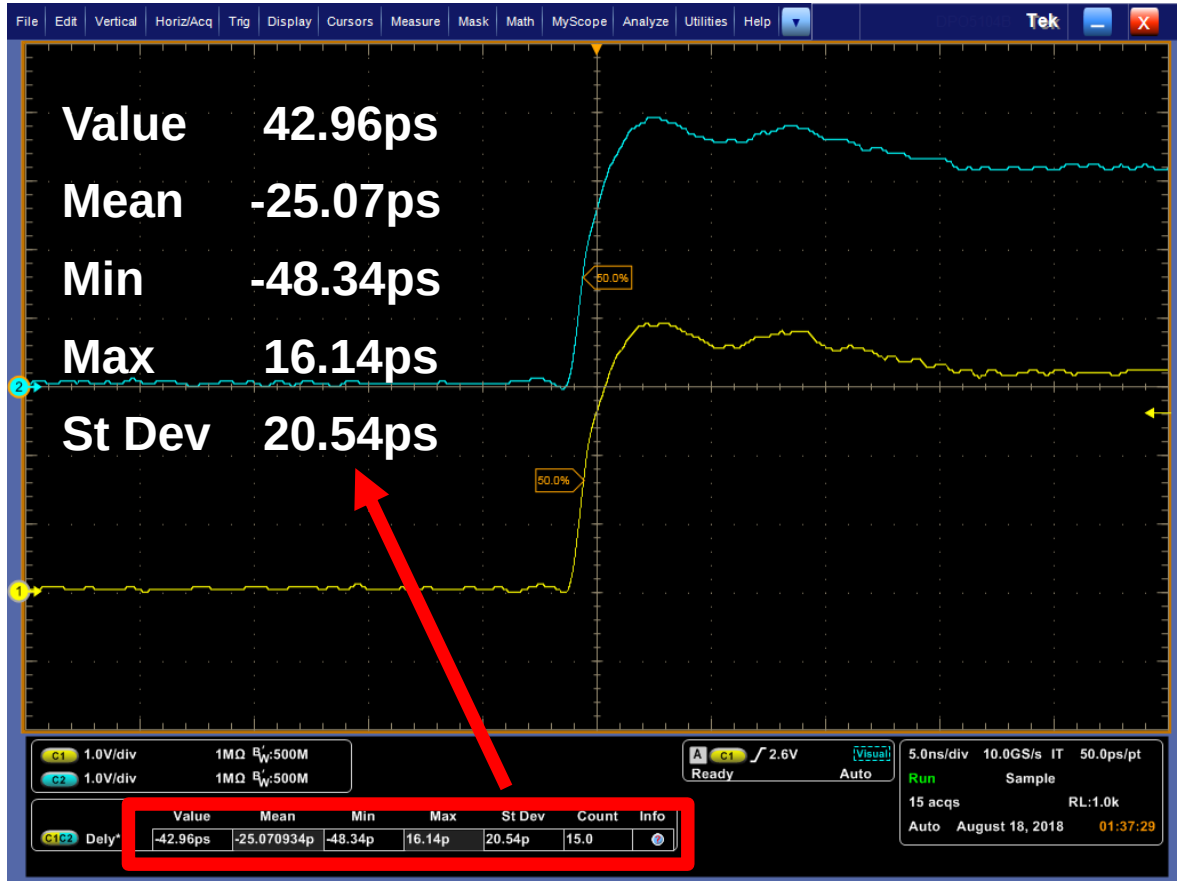
Access Network – White Rabbit					
	switch.in	switch.out	delta.i	delta.o	latency
1	15:38:58.056,303,467	15:38:58.056,303,736	0	0	269
2	15:38:58.056,303,473	15:38:58.056,303,855	6	119	382
3	15:38:58.056,303,477	15:38:58.056,304,095	4	240	618
4	15:38:58.056,303,478	15:38:58.056,303,976	1	-119	498
5	15:38:58.056,303,505	15:38:58.056,304,217	27	241	712
6	15:38:58.056,303,542	15:38:58.056,304,335	37	118	793
7	15:38:58.056,303,548	15:38:58.056,304,457	6	122	909
8	15:38:58.056,303,589	15:38:58.056,304,575	41	118	986
9	15:38:58.056,303,593	15:38:58.056,304,697	4	122	1104
10	15:38:58.056,303,651	15:38:58.056,304,815	58	118	1164
...					
50	15:38:58.056,305,335	15:38:58.056,309,250	44	99	3915
51	15:38:58.056,305,390	15:38:58.056,309,365	55	115	3975
52	15:38:58.056,305,446	15:38:58.056,309,464	56	99	4018
53	15:38:58.056,305,492	15:38:58.056,309,580	46	116	4088
54	15:38:58.056,305,561	15:38:58.056,309,679	69	99	4118
55	15:38:58.056,305,592	15:38:58.056,309,794	31	115	4202
56	15:38:58.056,305,674	15:38:58.056,309,894	82	100	4220

White Rabbit Services for our Trading Customers

Launched two WR based services

1. High-Precision Timestamp (HPT) File (more in a moment)
2. High Precision Time Service (connect to our WR network)





I will be right after the break

Network Design Principles

Determinism

- Single low latency entry point (Partition Specific Gateway)
- Cut Through switches with lowest jitter
- Separation of Market Data and Order Entry network

Fairness

- All cables are created equal

Monitoring Capabilities

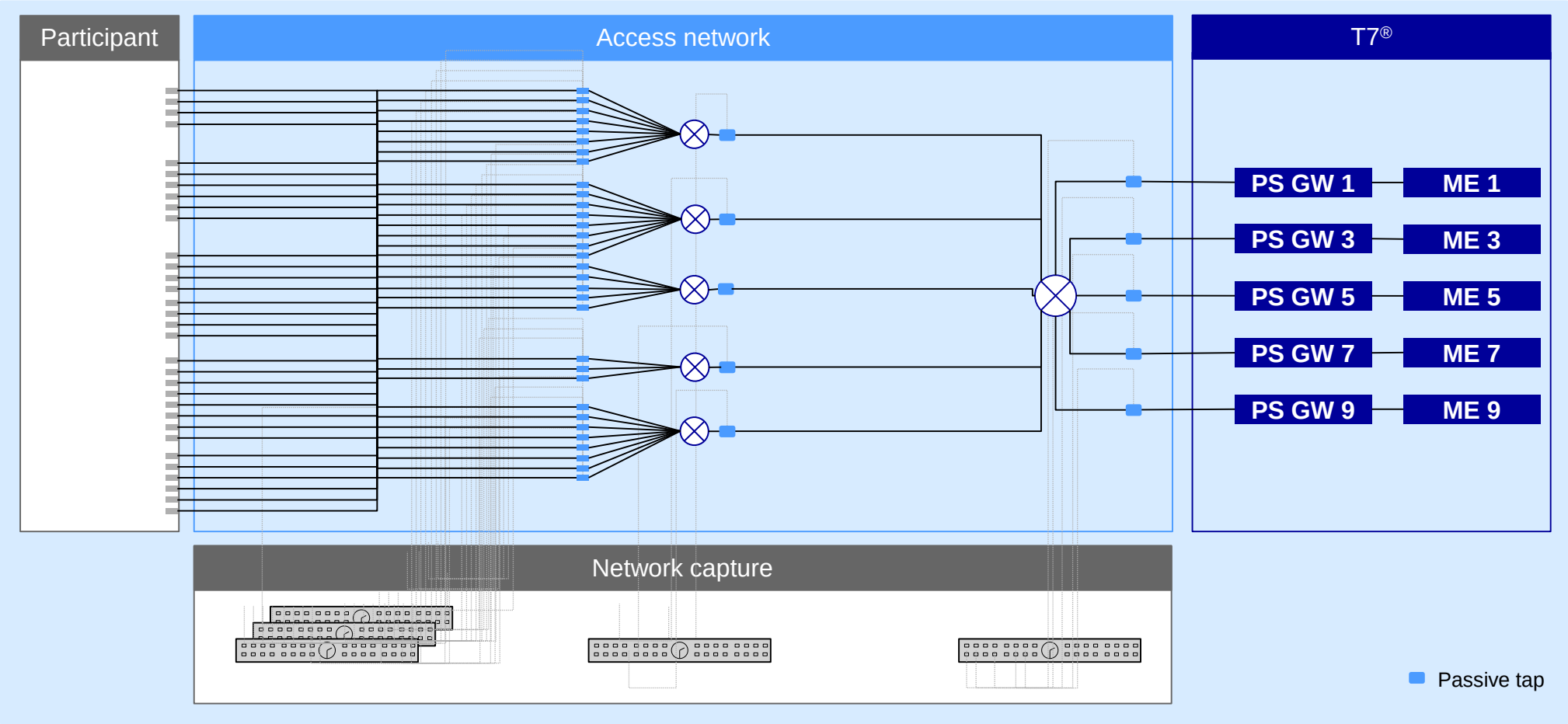
- Full visibility of network dynamics
- Tracing every single packet

Transparency

- Timestamps provided to trading participants via High Precision Timestamp File Service

Network Design

Order Entry Network (Fan-In)



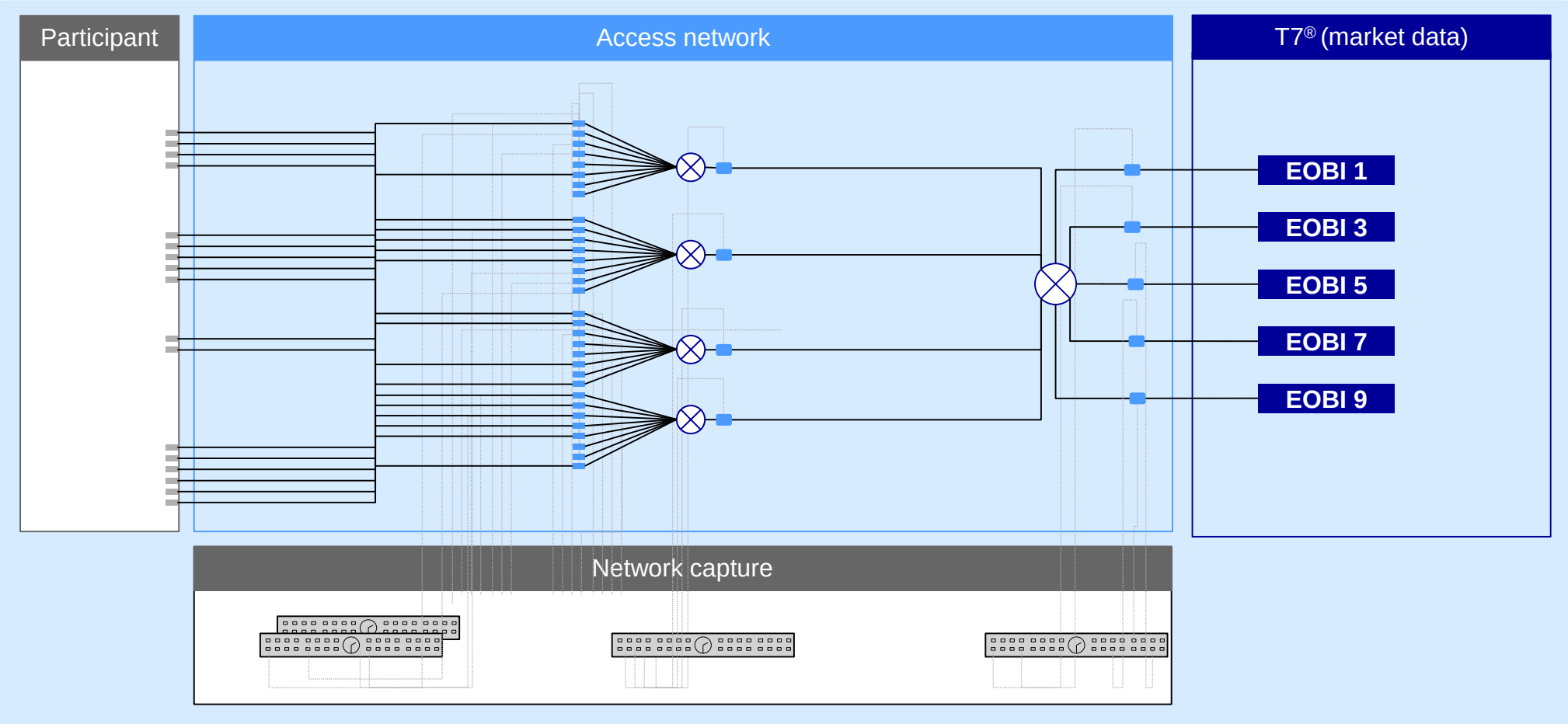
> 260 order entry lines individually captured (> 500 capture ports)

Identical set-up regardless of participant room location and assigned access switch

Only one side of one Market (Eurex) is shown for simplicity

Network Design

Market Data Network (Fan-out)



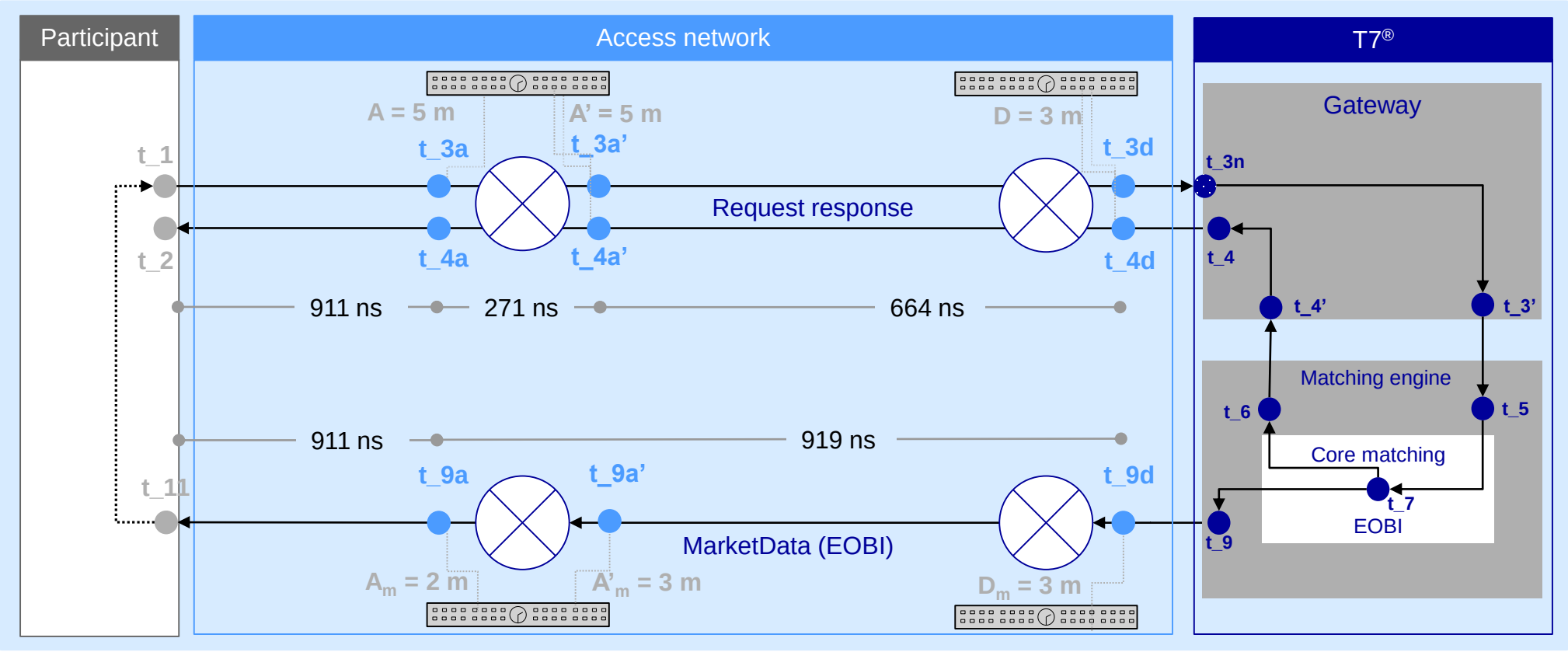
One market data access line per switch captured permanently – others configurable

Identical set-up regardless of participant room location and assigned access switch (differences < +/- 5 ns)

Only one side of one market (Eurex) is shown for simplicity.

Network Designs

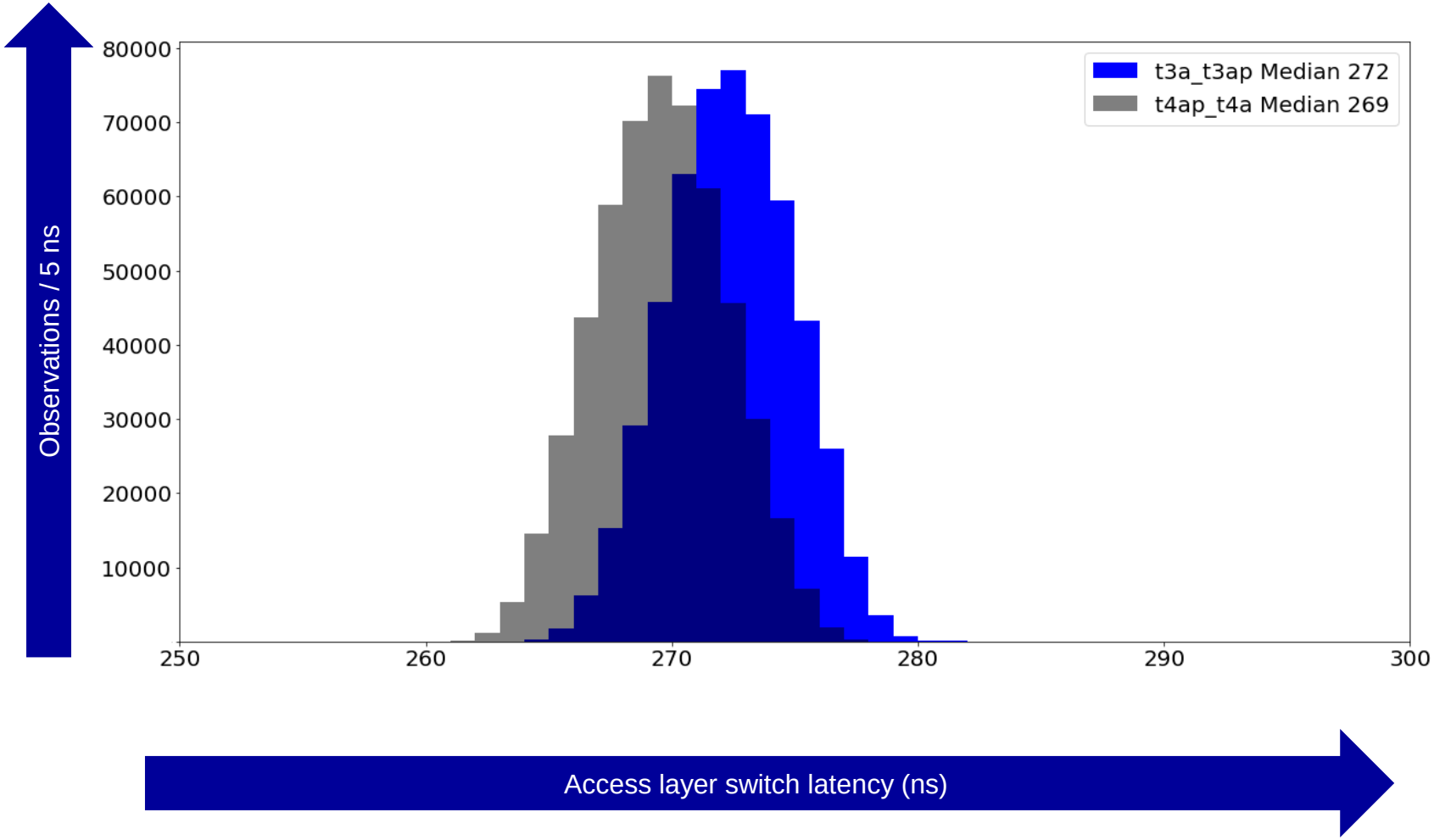
Timestamps



- Timestamps provided in T7 API (in real time) in dark blue (t_3n: taken by network card, other: application level)
- Network timestamps taken using taps and timestamping switches (Metamako)
- Timestamps possibly taken by participants shown in grey

Why we need nanosecond precision and accuracy

Measuring switch jitter and queuing



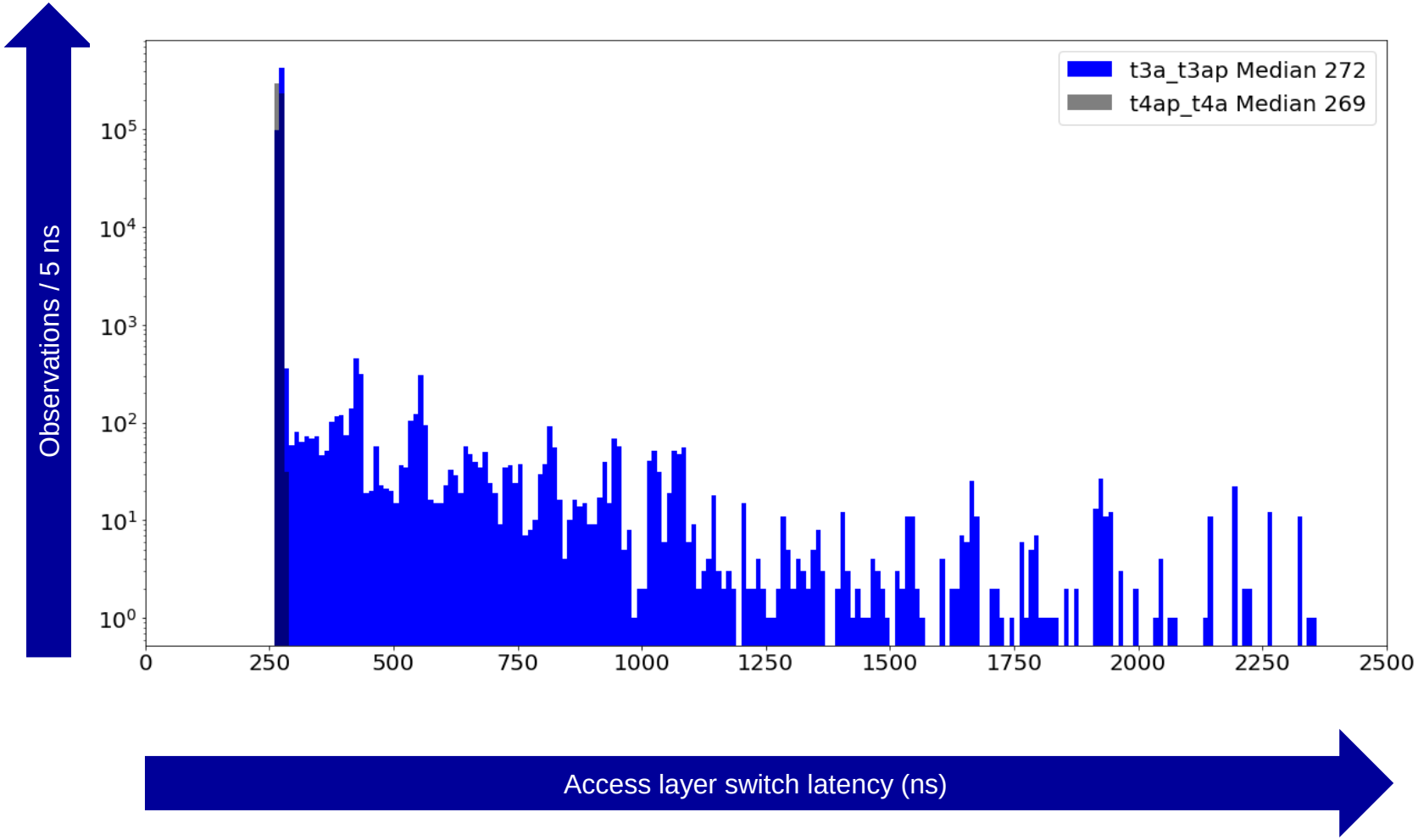
Switch jitter sub 10 ns

➔ Best accuracy and precision needed

Order entry access switch shown

Why we need nanosecond precision and accuracy

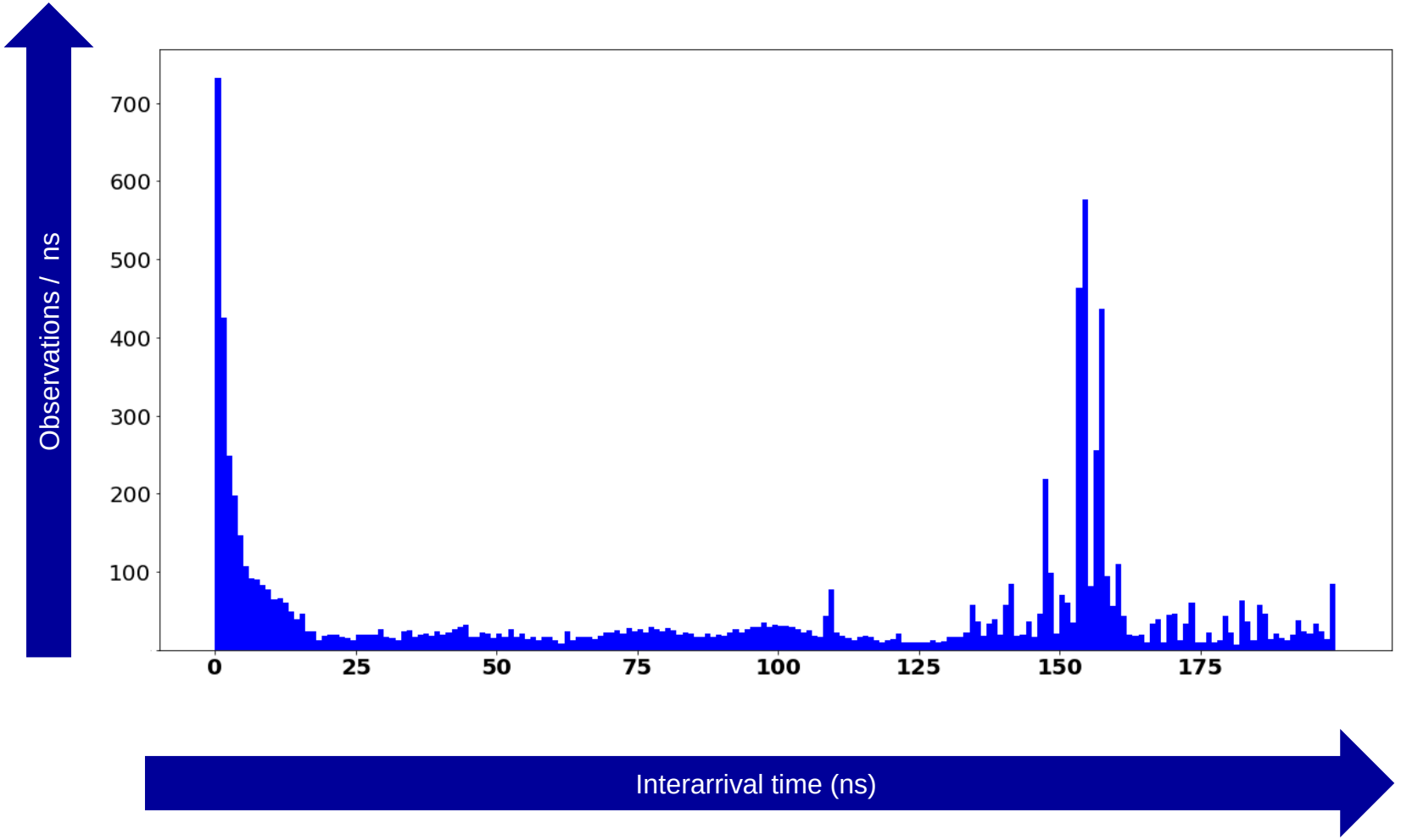
Measuring switch jitter and queuing



Nanobursts lead to queuing depending on number of packets and packet size.

Why we need nanosecond precision and accuracy

Interarrival time at network entry (t_3a)

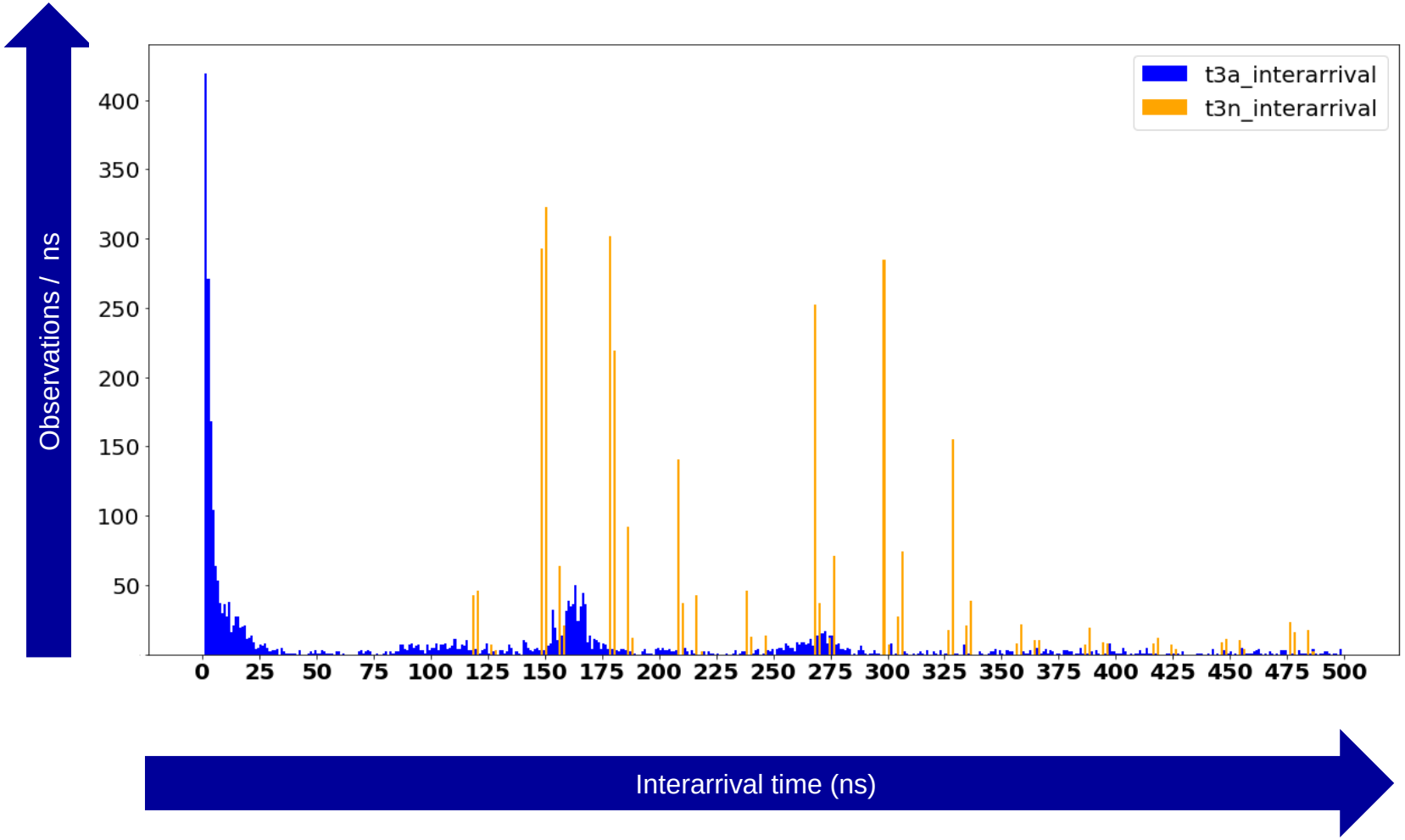


Interarrival time at t_3a arbitrarily small

➔Best accuracy and precision needed

Why we need nanosecond precision and accuracy

Interarrival time at network entry (t_3a) compared to gateway entry (t_3n)



Interarrival time at t_3a arbitrarily small

Interarrival time at t_3n at least serialisation time of preceding packet (> 100 ns)

Network card resulution 8 ns

➔ t_3a for precise inter arrival time

High Precision Timestamp File

Network level timestamps provided for executions

Provides network timestamps to trading participants

- Order Entry timestamps at network entry point t_3a
- Market Data timestamps at common network measurement point t_9a

Takes time synchronization and queuing effects out of the equation

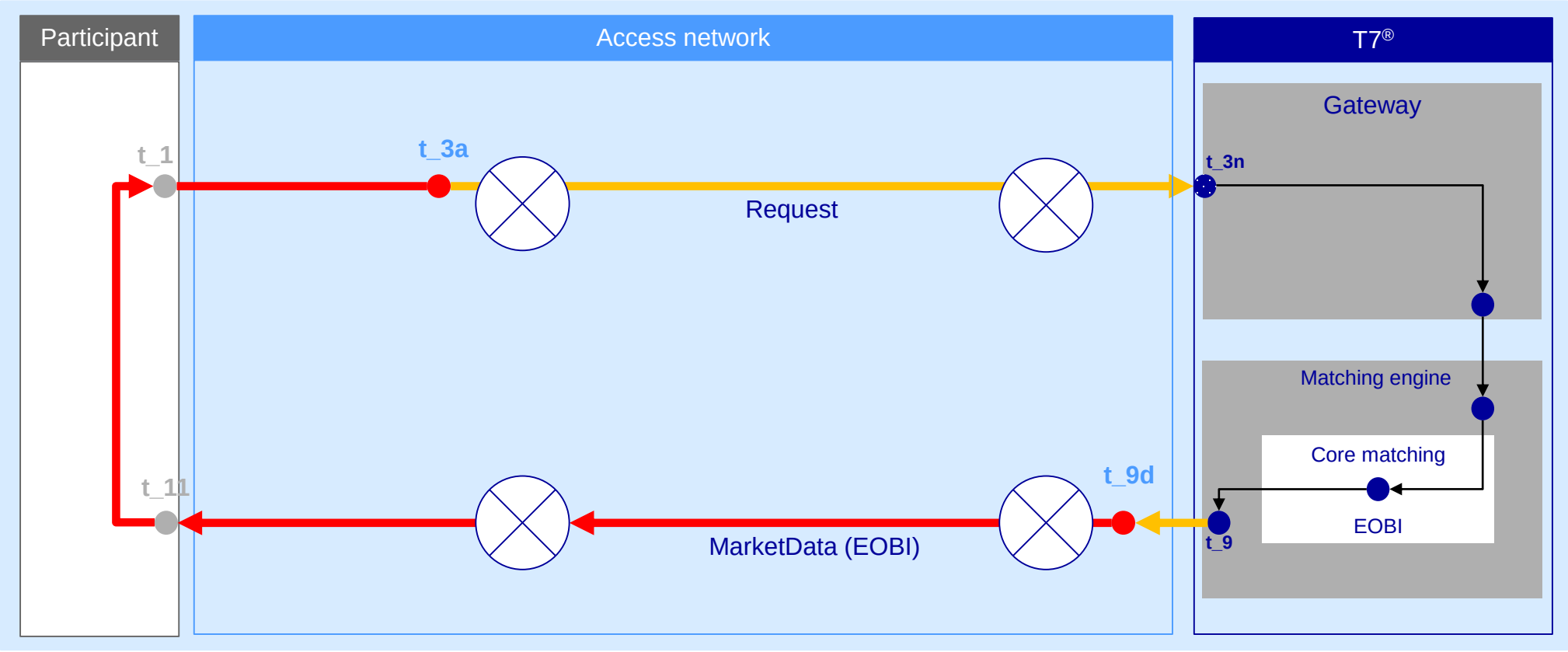
Allows precise judgement on competition ("How far behind am I ?")

Allows back testing with highest timestamp quality

Will be extended to all order book updates early 2019

High Precision Timestamp File

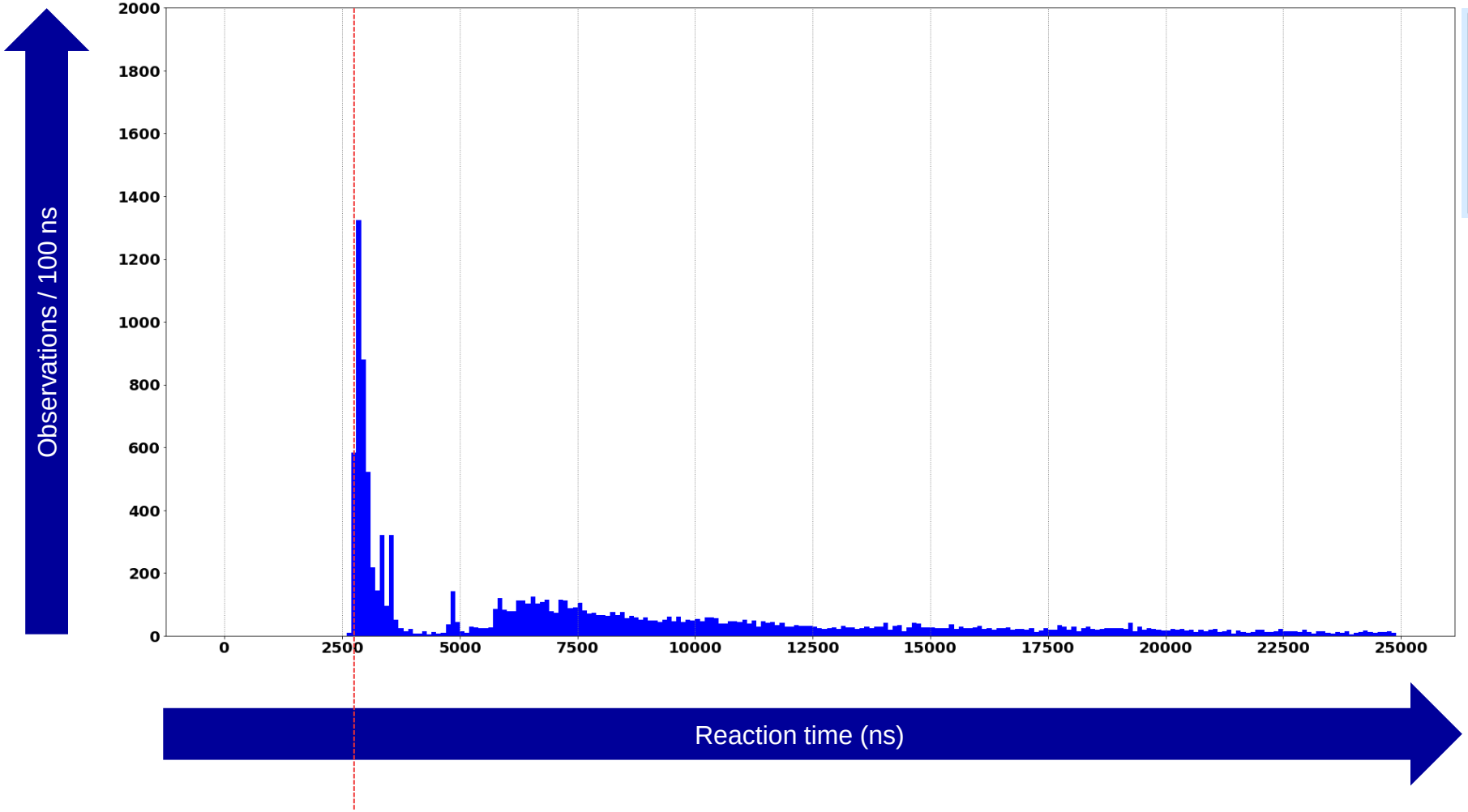
Reaction times of trading participants

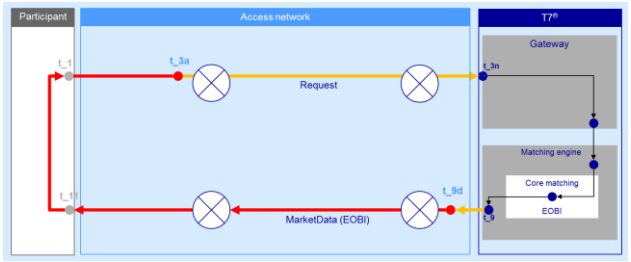


http://datashop.deutsche-boerse.com/High-precision-timestamps

High Precision Timestamp File

FESX => FDAX Reaction time based on T7® times (t_9 to t_3n)*

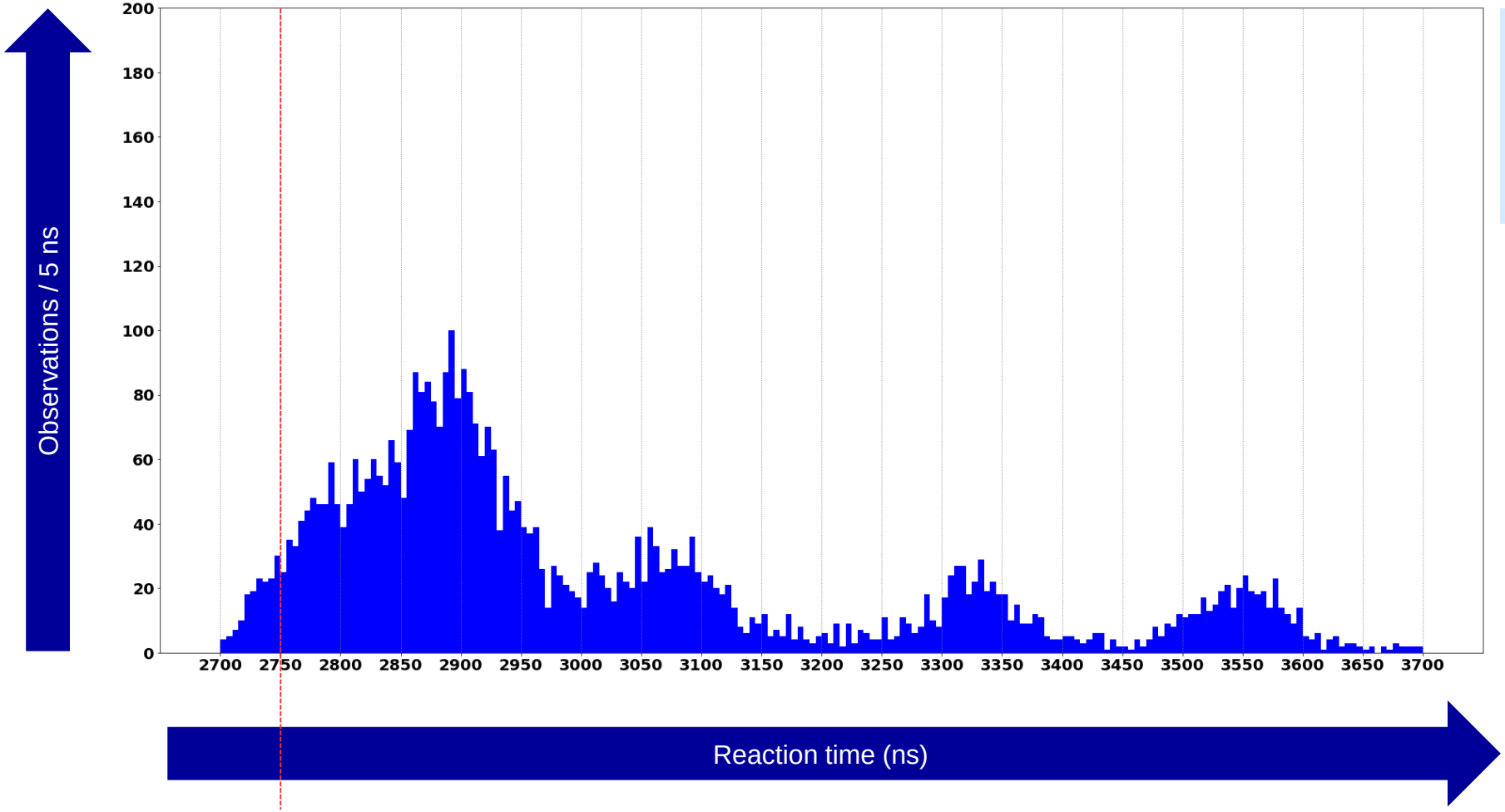


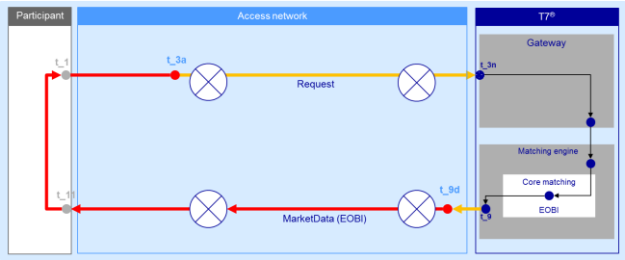


*Distribution of t_3n – t_9 – median (t_9d – t_9) – median (t_3n – t_3a) shown

High Precision Timestamp File

FESX => FDAX Reaction time based on T7® times (t_9 to t_3n)* (close up)



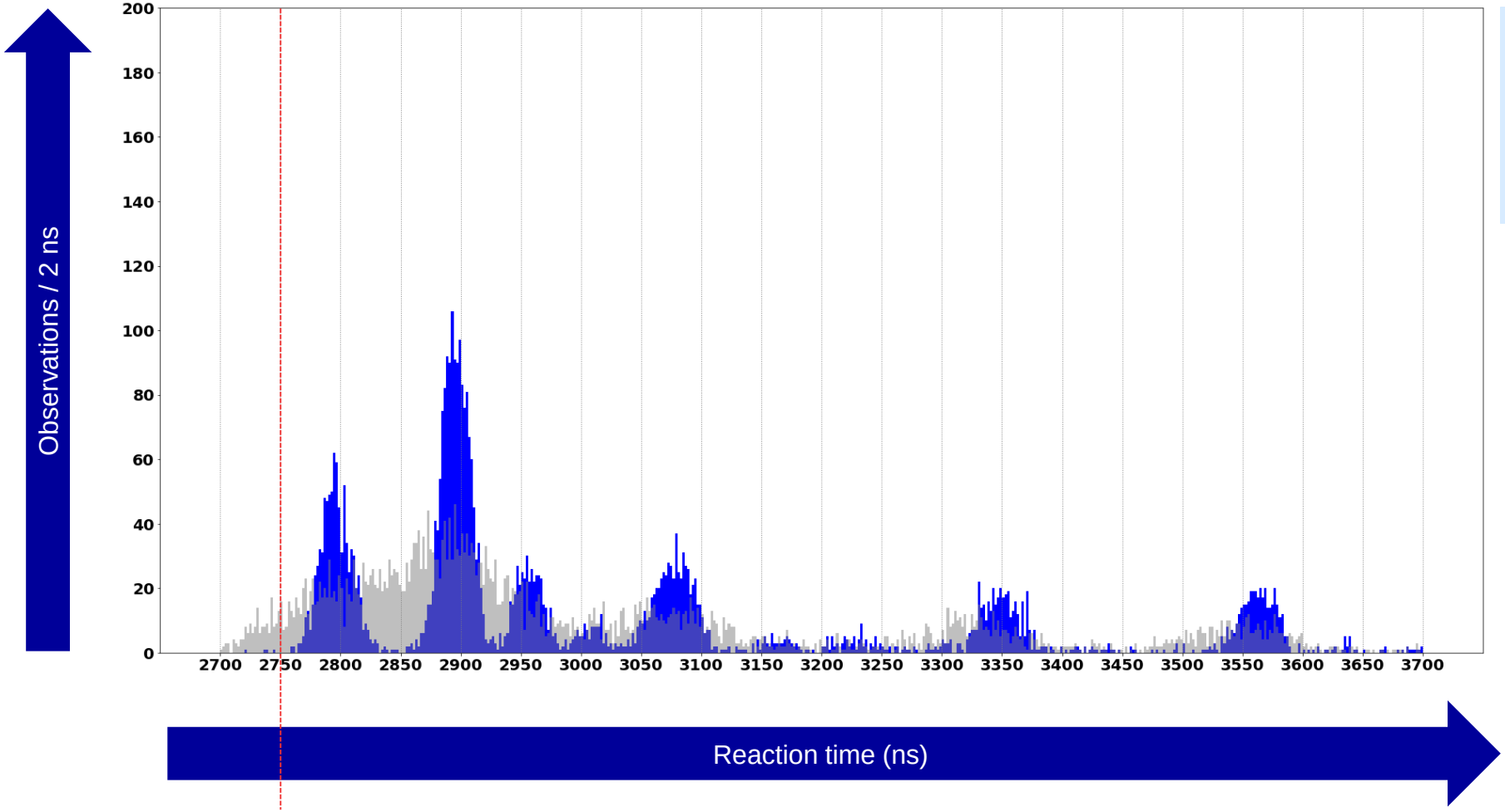


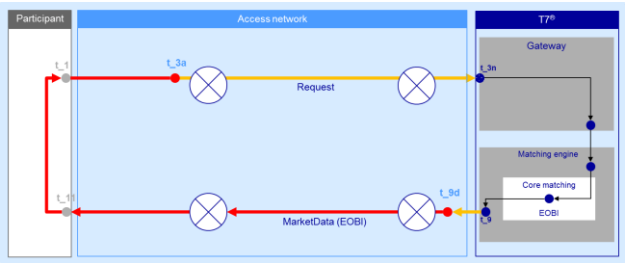
Theoretical minimum (2751 ns)

*Distribution of t_3n – t_9 – median (t_9d – t_9) – median (t_3n – t_3a) shown

High Precision Timestamp File

FESX => FDAX Reaction time based on T7® times (t_9 to t_3n)*

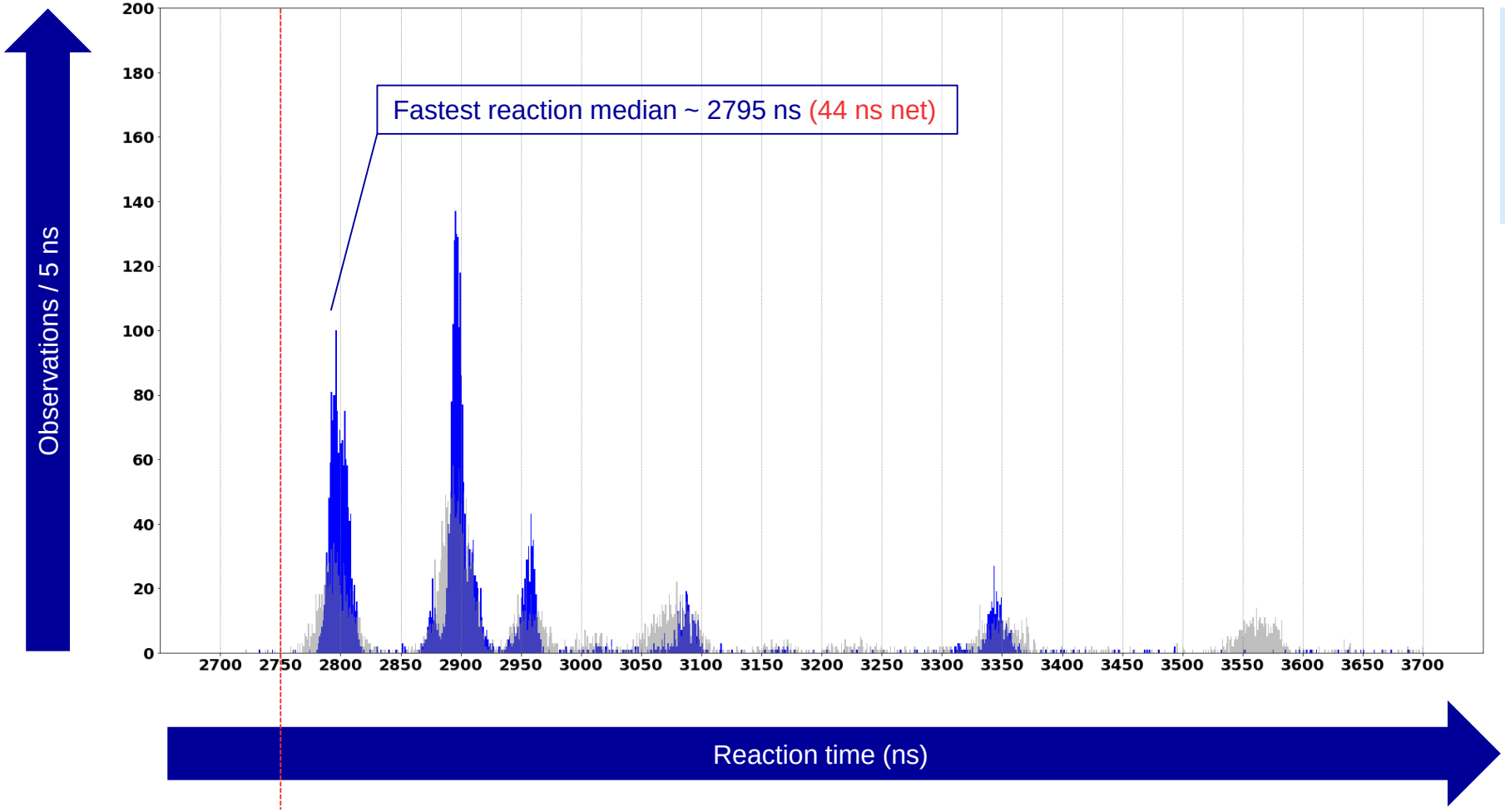


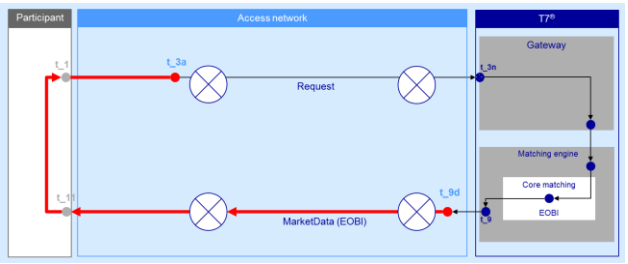


- Time synch error correction using rolling median
- Grey distribution shows correction using overall median

High Precision Timestamp File

FESX => FDAX Reaction time based on network times (t_9d to t_3a)





Grey distribution shows correction using rolling median

I will be right after the break

… and beyond?

- White Rabbit works great
  - PPS does not scale very well (coax cables)
  - No support in NICs (yet)

- Use PTP, but take more care deploying it
  - NICs with lower timestamping granularity
  - Boundary clocks with less jitter or
  - Layer-1 distribution instead of boundary clocks

- Hybrid White Rabbit / Standard PTP Solution
  - It is possible to configure ports on WR switch as standard PTP

- Use something completely different?
  - TickTock Clock Synchronization (TTCS) System
  - Data Center Time Protocol (https://conferences.sigcomm.org/sigcomm/2016/files/program/sigcomm/Session05-Paper01-Global-Ki.pdf)

TickTock Clock Synchronization (TTCS) System

also known as “HUYGENS”

The New York Times

“Time split to the nanosecond is precisely what Wall Street wants.”

- HUYGENS algorithm to synchronise nodes in a network
- Initially developed at Stanford University
- Now commercially developed by start-up Tick Tock Networks, Inc.

Very different from NTP or PTP:

- Exploits network effects, each server probes 10-20 others
- Coded probes
- Machine learning; support vector machines

https://www.nytimes.com/2018/06/29/technology/computer-networks-speed-nasdaq.html

TickTock Clock Synchronization (TTCS) System

- We have conducted an initial test with three nodes
- All components worked right out of the box
- The figures look promising

Exploiting a natural network effect for scalable, fine-grained clock synchronisation
https://www.usenix.org/conference/nsdi18/presentation/geng

Thank you for your attention.

Contact
Sebastian Neusüß
Andreas Lohr
E-mail monitoring@deutsche-boerse.com
Phone +49-(0) 69-2 11-1 86 86

DEUTSCHE BÖRSE GROUP

27 September 2018

Disclaimer

Deutsche Börse AG opens up international capital markets for its customers. Its product and service portfolio covers the entire process chain – from pre-IPO services and the admission of securities, through securities and derivatives trading through the settlement of transactions and the provision of market information to the development and operation of electronic trading, clearing and settlement systems. With its process-oriented business model, Deutsche Börse increases the efficiency of capital markets. Committed employees are the key factor for innovation and further growth: without them, Deutsche Börse Group would not have developed into one of the most modern exchange organisations in the world. More than 5,000 employees work for the Group – a dynamic, motivated and international team.