Open Day 2017

T7® infrastructure and latency

Andreas Lohr and Sebastian Neusüß

5 October 2017

Contents

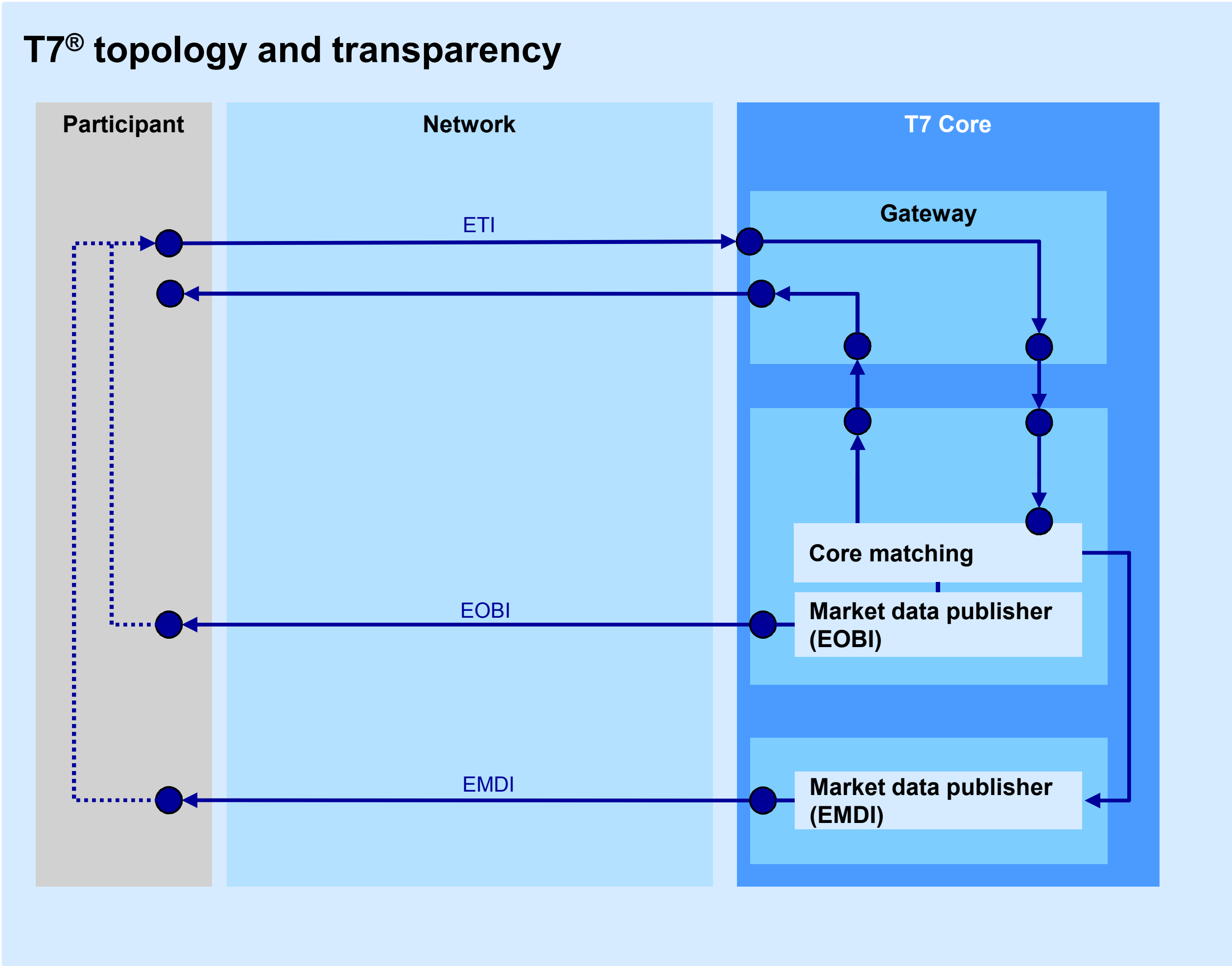
Introduction

Base

Jitter

Queuing

Structure



Aspects of latency

Base	**How fast?** Uncertainty
Jitter	**Always the same?** Predictability
Queuing	**Even under load?** Predictability
Structure	**Latency structure matters!** Fairness, market structure, complexity, transparency

Aspects of latency

Base

Jitter

Queuing

Structure

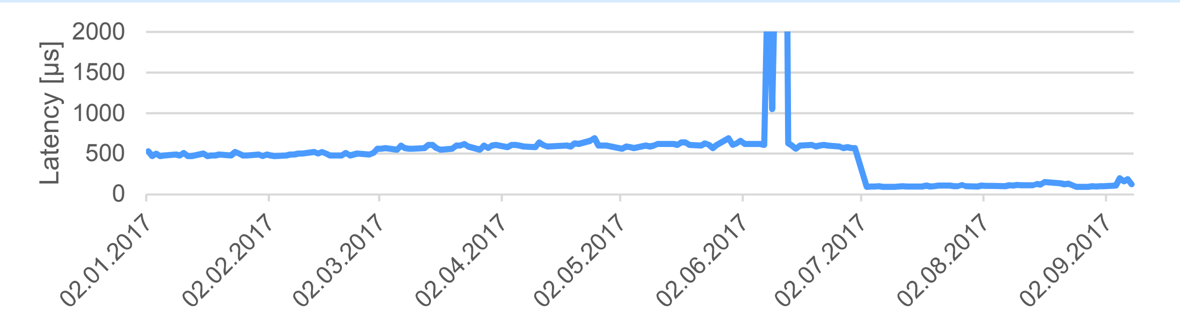
How fast?

Measures used: median, average to describe the latency in one number

Example

Request-response round trip:

Xetra® (average) 140 µs [2016: > 400 µs]



Cash market migration to T7® on 26 June and 3 July 2017

Aspects of latency

Base

Jitter

Queuing

Structure

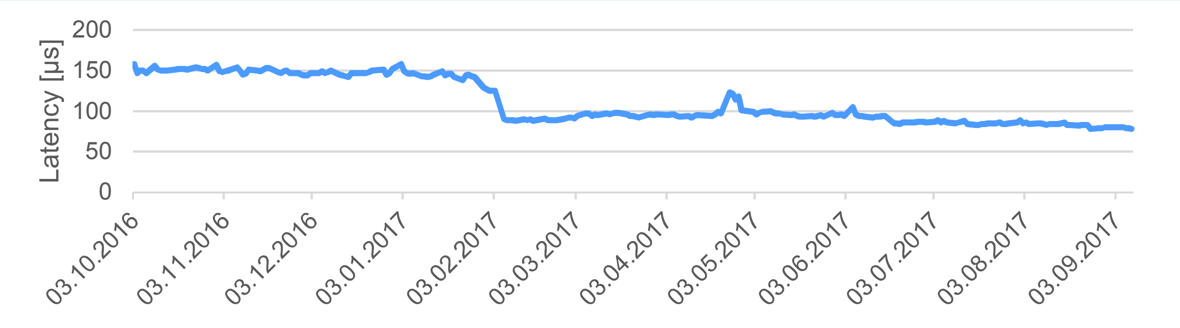
How fast?

Measures used: median, average to describe the latency in one number

Example

Request-response round trip:

Eurex® (median) 76 µs [2016: 150 µs]



New hardware in February 2017
Co-location 2.0 in April 2017
Release 5.0 on 19 June 2017
Cash market migration to T7® in June/July 2017

Aspects of latency

Base

Jitter

Queuing

Structure

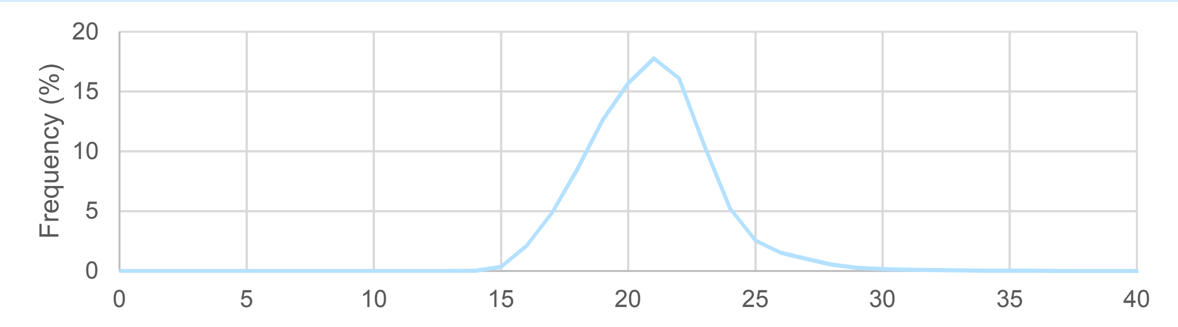
Always fast?

“Random” influences lead to uncertain latency, e.g. cache misses in CPU-based system, “white noise” in scheduler, hardware, cables, switches etc.

Measures used: confidence intervals [e.g. 10–90th percentile]

Example

- Gateway in to matching engine in
- Distribution for “free” order cancel requests (Eurex)



10th percentile: 18.3 μs
90th percentile: 24.2 μs
Confidence interval: 5.9 μs

Aspects of latency

Base

Jitter

Queuing

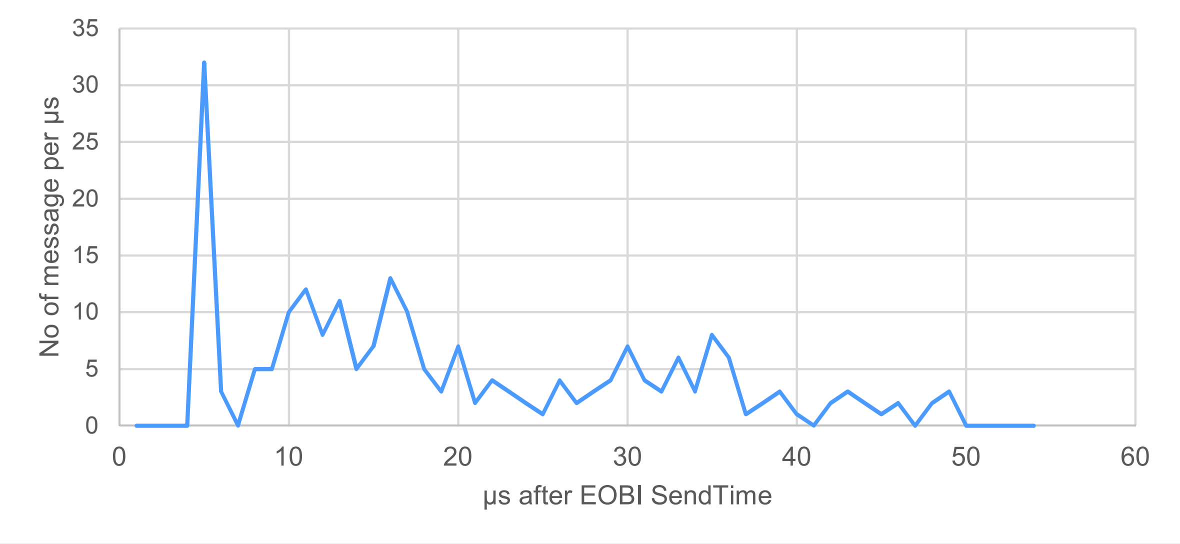
Structure

Always fast?

- Higher input than output rate leads to higher latencies.
- Usually on small timescales (microbursts)

Example

Burst of transactions after trading signal (large trade in FGBM)



Aspects of latency

Base

Jitter

Queuing

Structure

Latency structure matters

- Favour cancel over new?
- Publish order book changes first on public or private?
- Provide equal access to the system? How equal?
- How transparent?

Infrastructure and topology

- Parallel or sequential [FIFO] set-up?

Putting it all together

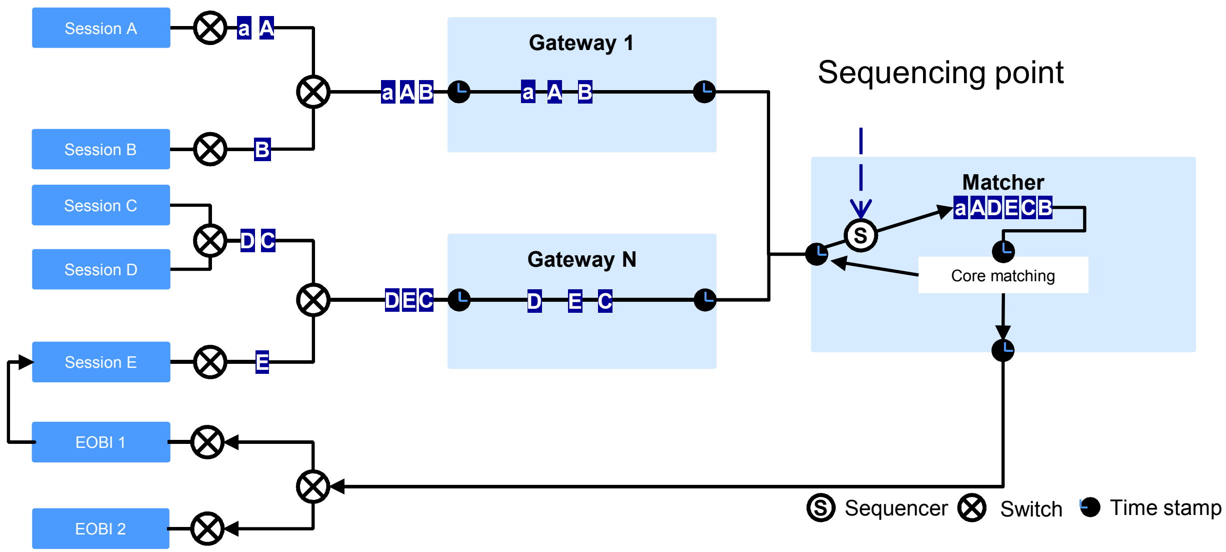
Base

Jitter

Queuing

Structure

T7 topology
Current set-up



Jitter on parallel paths incentivises multiplicity to reduce latency.
Sharp microbursts in turn lead to queuing delay.
FIFO processing has significantly reduced multiplicity.

Putting it all together

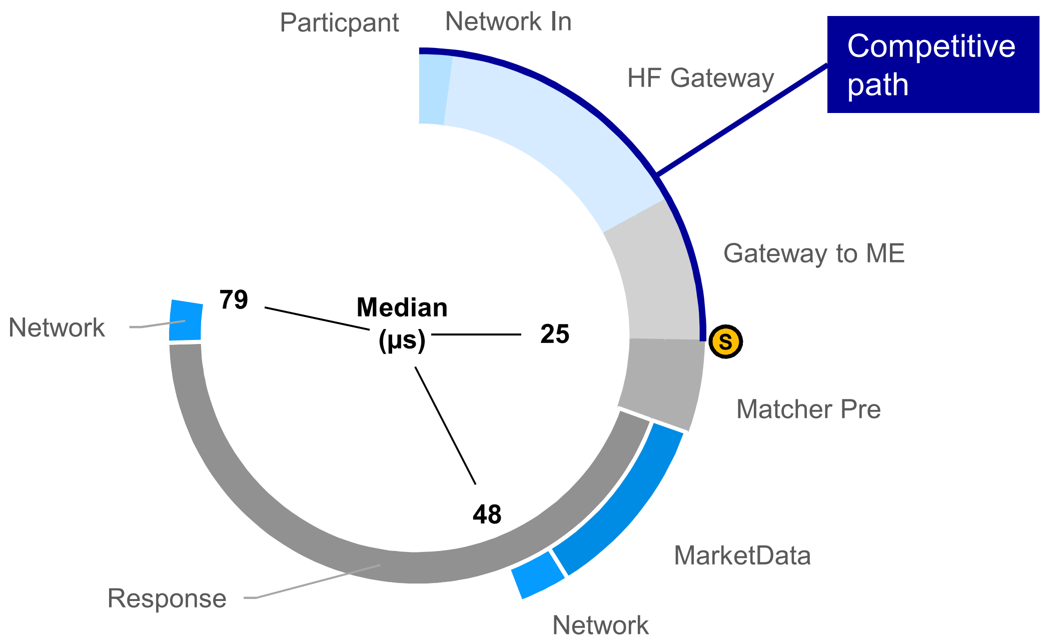
Base

Jitter

Queuing

Structure

Latency composition (1)



Inner circle: Request – response
Outer circle: Request – market data (EOBI)
Full circle: 100 µs

Aspects of latency

Base

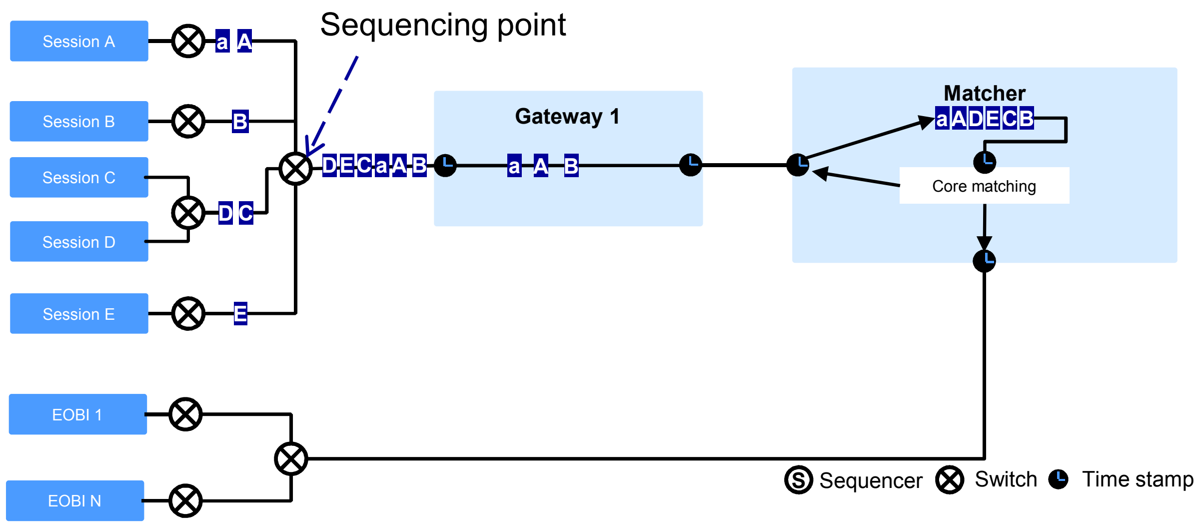
Jitter

Queuing

Structure

Partition-specific gateway

Single low latency entry point means network serialisation determines matching priority.



Putting it all together

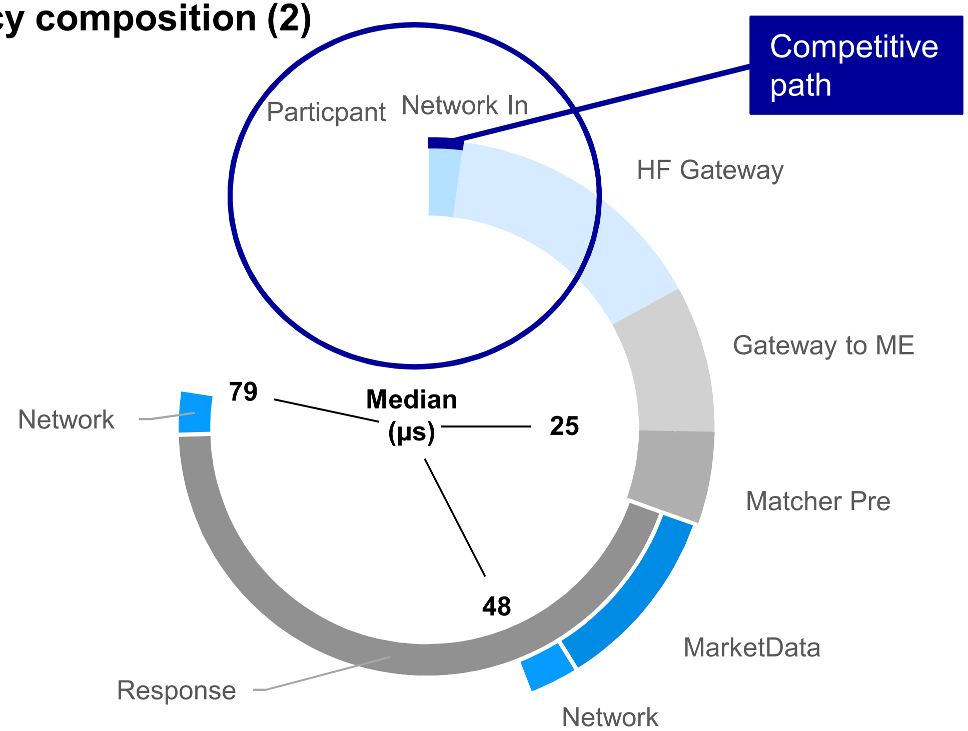
Base

Jitter

Queuing

Structure

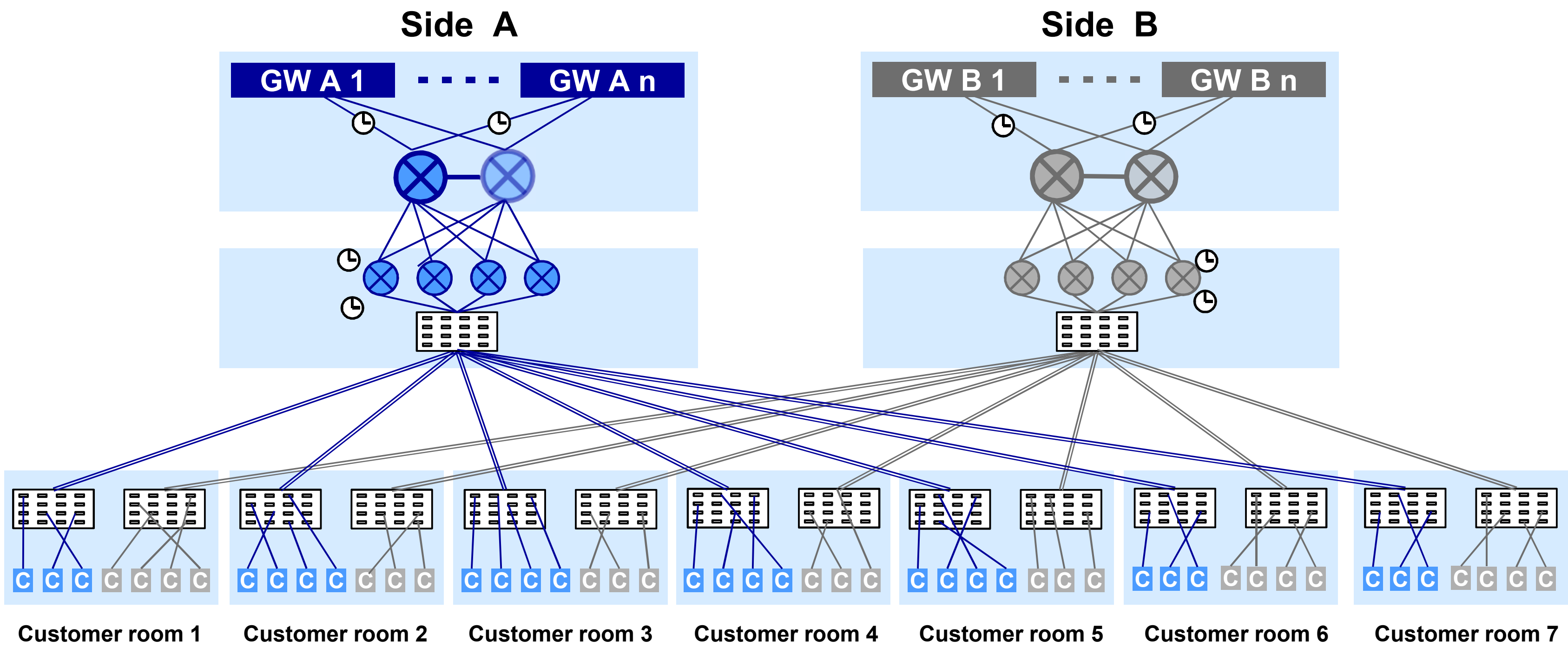
Latency composition (2)



Introduction of PS gateways will shorten the competitive path.
High focus placed on participant to network.

Network topology in 10 Gbit co-location (v2.0)

- 2 switches per gateway room per market (‘distribution layer’, only one market shown)
- Eurex®: 8 centrally located switches (‘access layer’, 4 per side, A and B)
- Xetra®: 4 centrally located switches (2 per side, A and B, not shown below)
- Customers can connect to any access layer switch from any of the 7 co-located rooms
- There is a separate Market Data network with same layout



Co-location 2.0 (1/4)
Equidistant cabling

Tolerances

Co-location 1.0 = +/– 4m

Co-location 2.0 = +/– 1m

Why 1m? Why not 4cm?

- Overview on previous slide is a gross simplification.
- Actual floor layout in Equinix FR2 looks very different.
- There are seven co-location modules of different sizes across two floors.
- Cables have “additional margin” on top of what you order.
- What about all the patch panels, patch cables, SFPs?

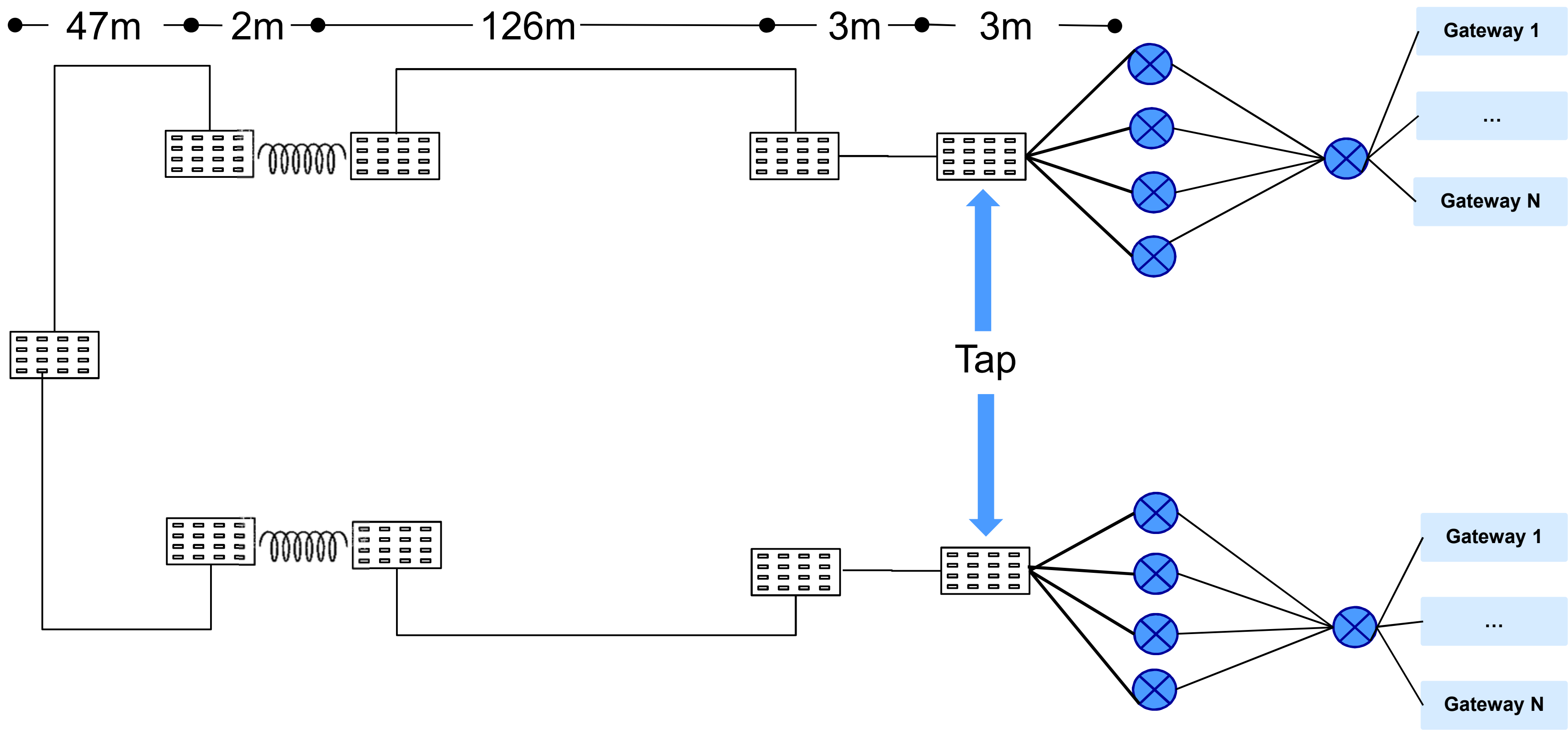
Customers care about cable lengths?

- Some trading participants have sub 200ns response times.[1)]
- Solarflare and LDA Technologies claim 120ns tick-to-trade.[2)]
- 1m fibre optical cable ≈ 5ns

1) http://tabbforum.com/videos/high-performance-timestamping-for-the-enterprise
2) http://www.solarflare.com/solarflare-and-lda-harness-the-power-of-xilinx-fpgas

Co-location 2.0 (2/4)

Equidistant cabling



Co-location 2.0 (3/4)
Equidistant cabling

How did we actually measure?

OTDR (optical time-domain reflectometer)

- Standard practice after physical installation before handing over
- Injects light pulse into cable and uses reflections to characterise cable
- Measures the quality (e.g. attenuation of the signal) and length of cable

Challenges

- Contracted out to a service company ➔ How do we verify their work?
- Accuracy of length measurements unclear
- Found bugs in OTDR analysis software
- Reproducibility issues

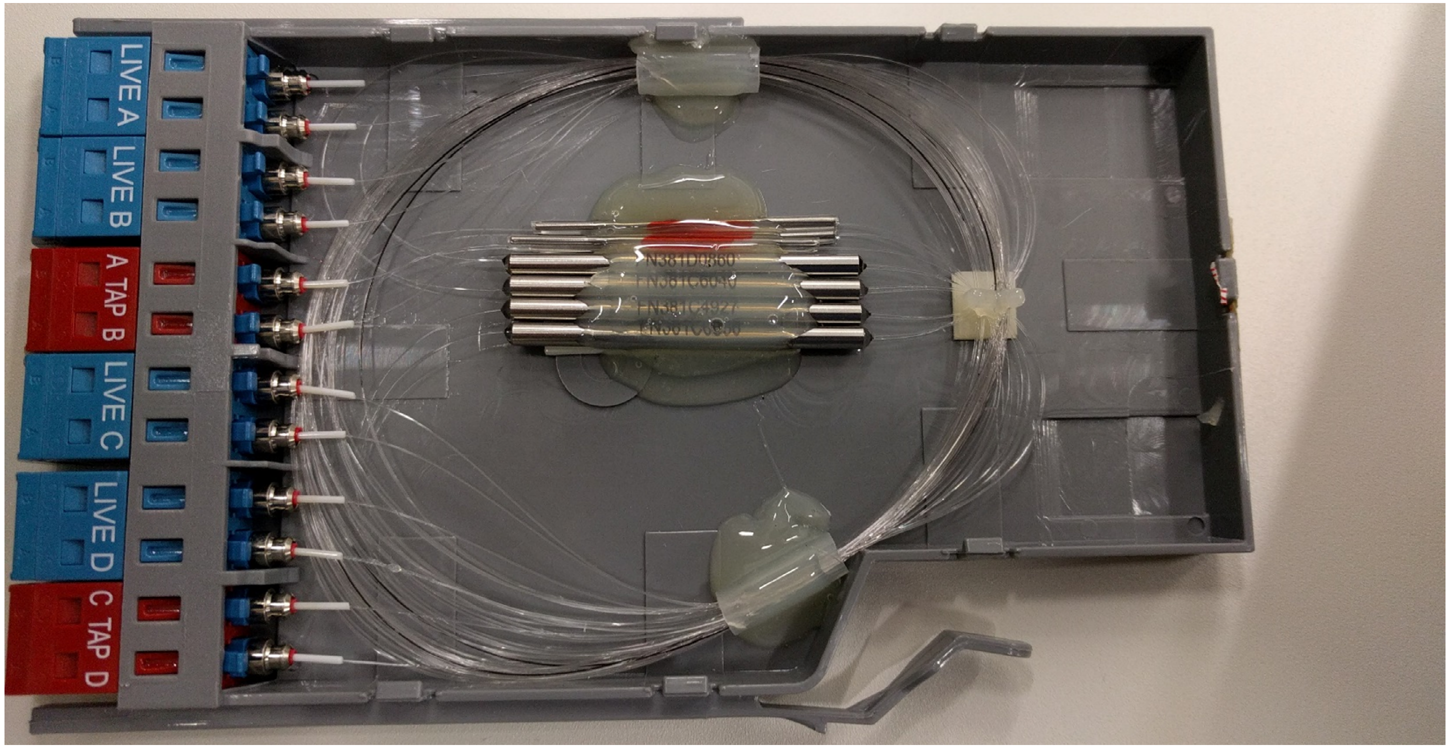
Packet capture to the rescue

- Re-tested using a layer-1 switch with precise timestamping
- Reproducible results
- Solution that worked best for us

Co-location 2.0 (4/4)
Equidistant cabling

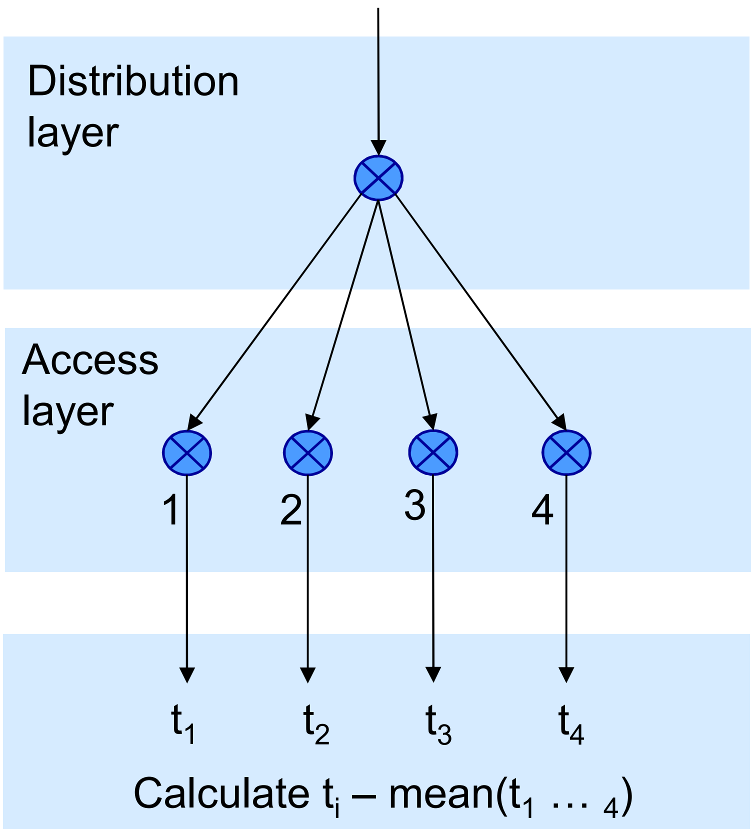
Optical taps

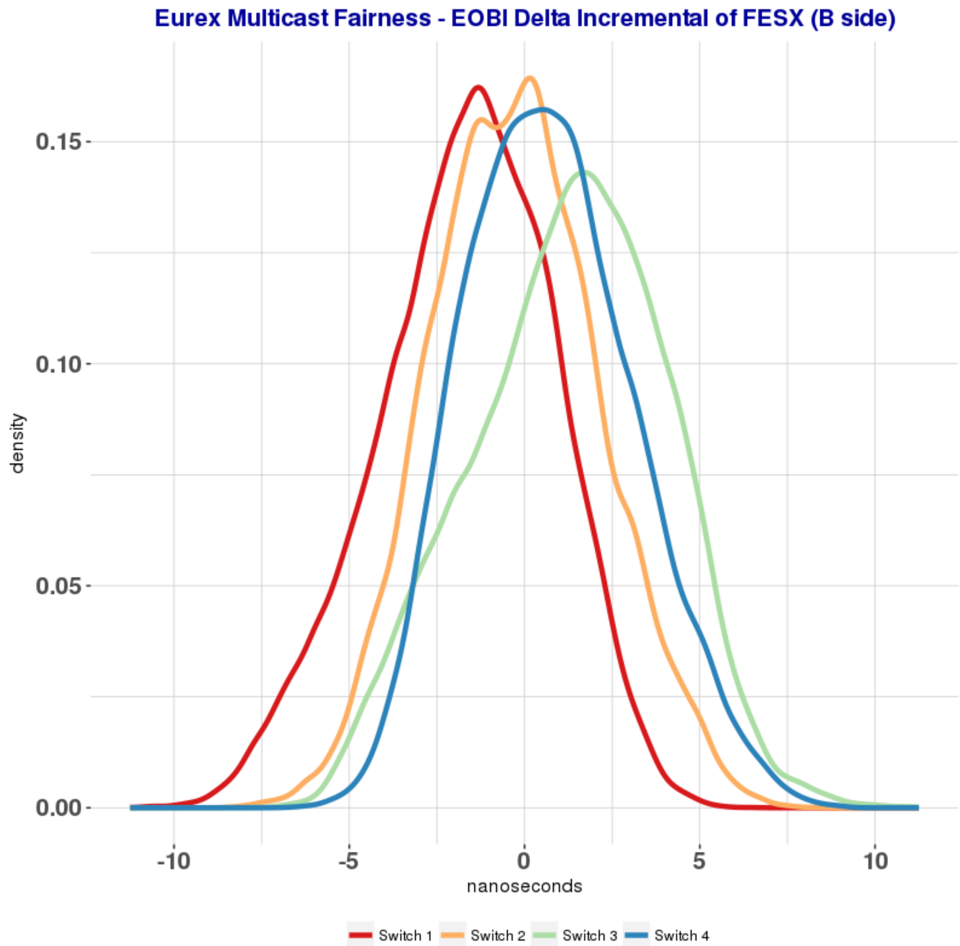
- Introduce negligible latency (?)
- Tap outputs have same latency (?)



Multicast fairness

- By how much do the multicast access layer switches differ?
- Time stamp precision is +/– 4ns

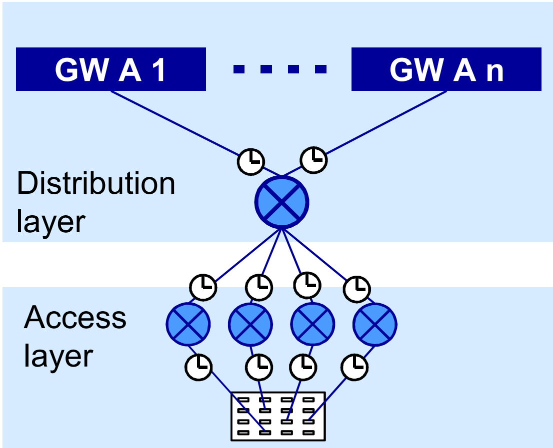




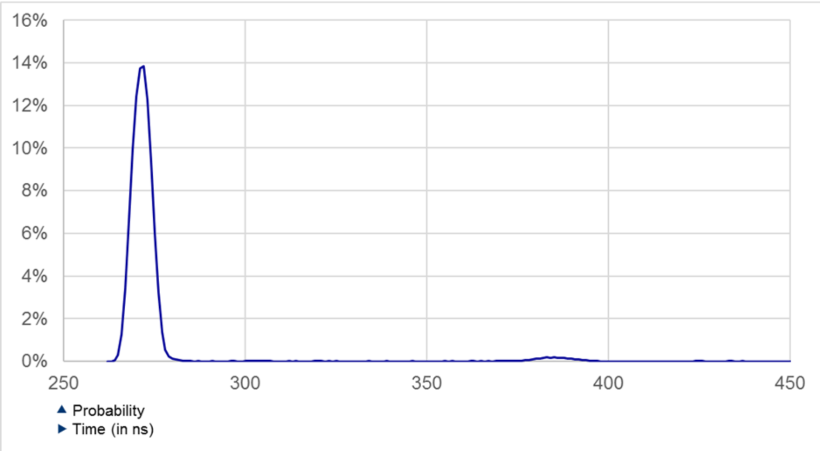
Co-location 2.0

Order Entry latency profile

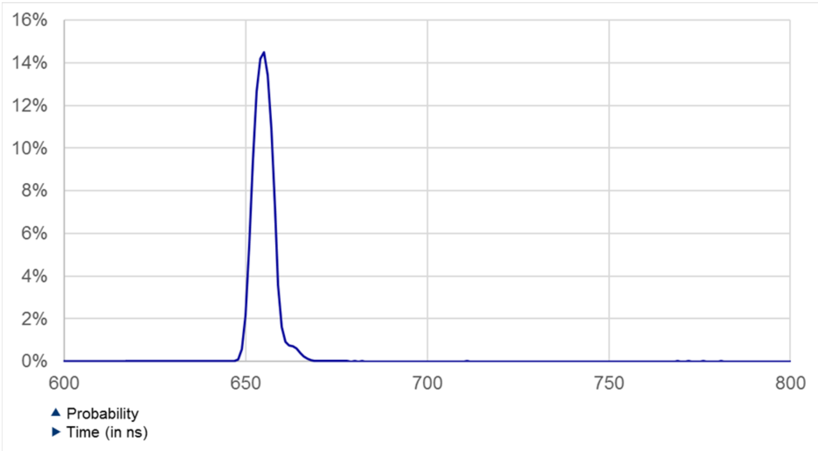
- Highly deterministic access network
- Very tight latency profiles
- Constantly monitored
- Monitoring devices time synched to within single digit nanoseconds



Access layer switch latency



Access to distribution layer latency



Co-location 2.0

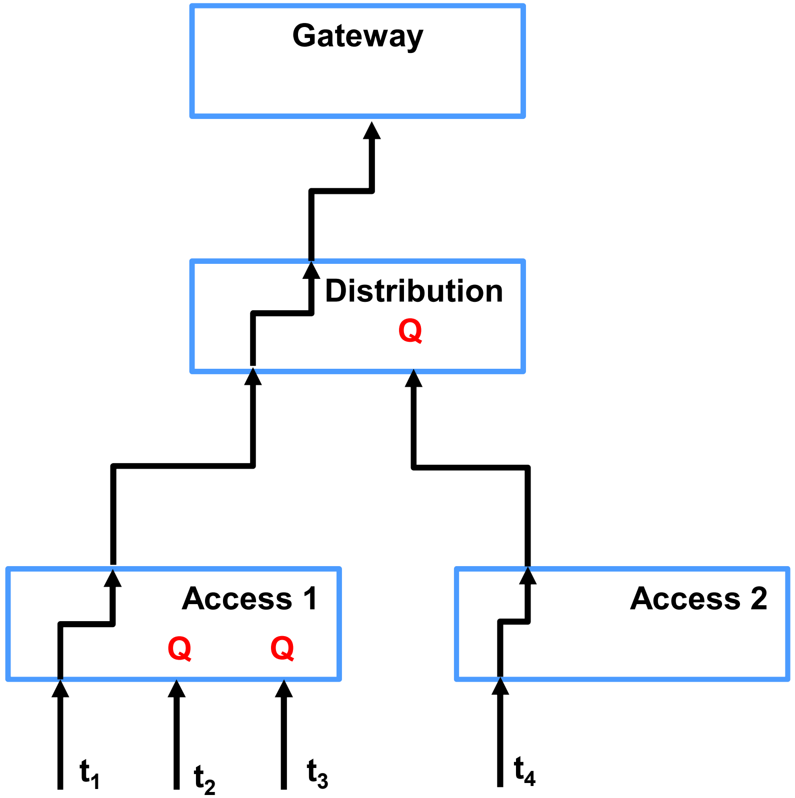
Order Entry

- Cisco 3548X in warp-mode
- Cut-through
- Latency ≈ 200ns

- Message at t_1 will be first in gateway.
- Messages at t_2, t_3 will be queued in Access 1.
- Message at t_4 will be queued in distribution layer.

We observed no overtaking of immediately forwarded frames and less than 1% of queued frames were re-ordered.

The arrival time lag of overtaking frames was almost always within our timestamping precision.



Further information

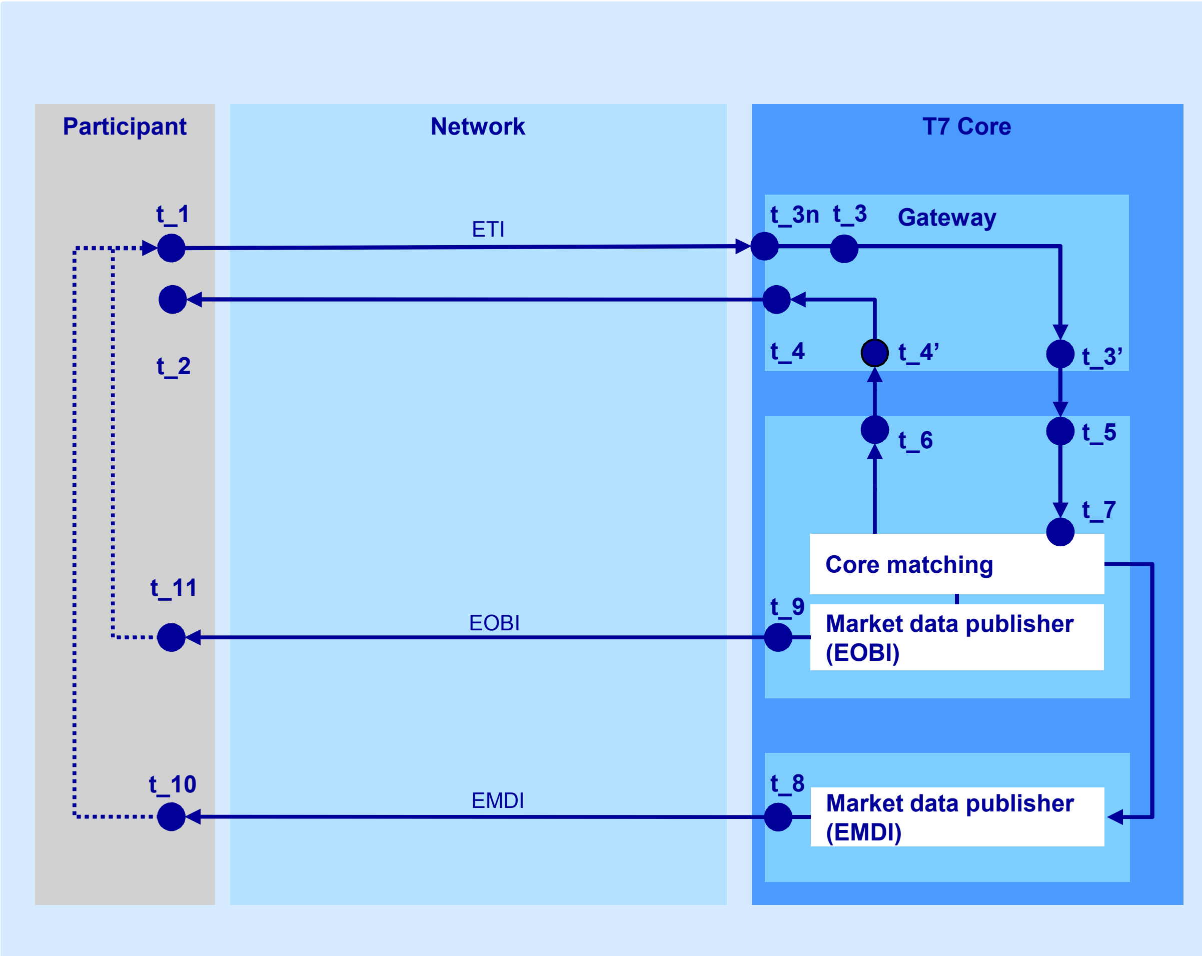
- More details and regular updates are available in the “Insights into Trading System Dynamics” presentation at eurexchange.com > Technology > HFT

- For further questions contact us via monitoring@deutsche-boerse.com.

Thank you for your attention.

Contact
Sebastian Neusüß
Andreas Lohr
E-mail monitoring@deutsche-boerse.com
Phone +49-69-211-18686

T7® system overview – our transparency



Request / inbound

- t_3n: GW in (RequestTime, for HF gateways only)
- t_3: GW application in (RequestTime, for LF gateways only)
- t_3': GW out (RequestOut)
- t_5: Matcher in (TrdRegTSTimeIn)
- t_7: Core matching in (ExecID, MDEntryTime, TransactTime, TrdRegTSTimePriority)

Response / outbound

- t_6: Matcher out (TrdRegTSTimeOut)
- t_4': GW in (ResponseIn)
- t_4: GW out (SendingTime)
- t_8: EMDI out (header SendingTime)
- t_9: EOBI out (header TransactTime)

Further information and regular updates are available in the “Insights into Trading System Dynamics” presentation at www.eurexchange.com/exchange-en/technology/high-frequency_trading.

T7® topology partition-specific gateway and co-location 2.0