How to monitor the DAX

Working with Checkmk @ Deutsche Börse

Jens.Kuehnel@deutsche-boerse.com

About me

- Linux User since 1995
- Since 2000 Freelancing Author, Trainer and System Adminstrator
- Since 2017 System Administrator @ Deutsche Börse
- Started with omd in 2012.
- Checkmk RAW user since day 1.

What people think we do

What we really do

- IT company with 11’000 employees with over 50 locations
- Origins tracing back to 1585
- Develop and Run Trading systems.
- Support the whole value chain

Areas

- Pre-Trading
- Cash (Xetra) & Derivatives (Eurex) Trading
- Clearing
- Post-Trading
- (EEX = Energy Exchange)

How to Explain it to my family

What my Unit does

- OS Infrastructure for Trading, Clearing
- RHEL 7-9 (8 main platform at the Moment)
- Monitoring with Checkmk and grafana (telegraf & Influx)
- Logs managed by graylog & Grafana LOKI
- Hosts Managed by ATIX Orcharino with puppet
- Virtualization done be RHV
- Cloud and OpenShift is handled by our colleges

Red Hat
Enterprise Linux

Grafana

OPENSHIFT

puppet

Checkmk @ Deutsche Börse

- Users: ~100/~90 Service: 90000/75000 Hosts: 1500/1500
- Monitoring Hosts: Clustered with RHEL Cluster
  - 2*2*10*1 @2.5GHz 256GB. Load: 4 CPU: ~10%
- Checkmk Team: 2,5 (down from 3,5) + 2 check Developers
- Started with omd, nagios and Checkmk around 2012
- Checkmk Enterprise since 2015
- One of the first customer to monitoring OpenShift with Checkmk
- Clean Reimplementation with the Help from checkmk in 2019
  - Instead of ~60 Cluster rules only 1

Plugins

- Self developed (agent & check plugin):
  - AMQP Broker
  - Infiniband
  - sfptpd
  - 3 of our internal software
- External:
  - hpe_oneview (SVA)
  - Sslcert (heinlein)
- At the moment: git repos in local, plan to move to mkp
- Wishlist: git2mkp

Target: Clean monitoring

- Seperate Sites: Develoment/Prod (no connection)
- Application vs. OS (tags)
- When someone works on a problem, set a ACK.
- Reboot automatically schedule downtime.
- Check ACKed services and remove monitoring when not handeled in a timely manner.
- Main reason: Someone cares (Greetings Spindy)
- Whishlist: Traffic Light and Tag for Logfiles

Regulatory

- Bank and Stock Exchange regulation
- Air gapped Production network
- Configuration changes only on the Weekend
- Created our own Crypto-Policy
- redundant

T7 Trading System Tuning

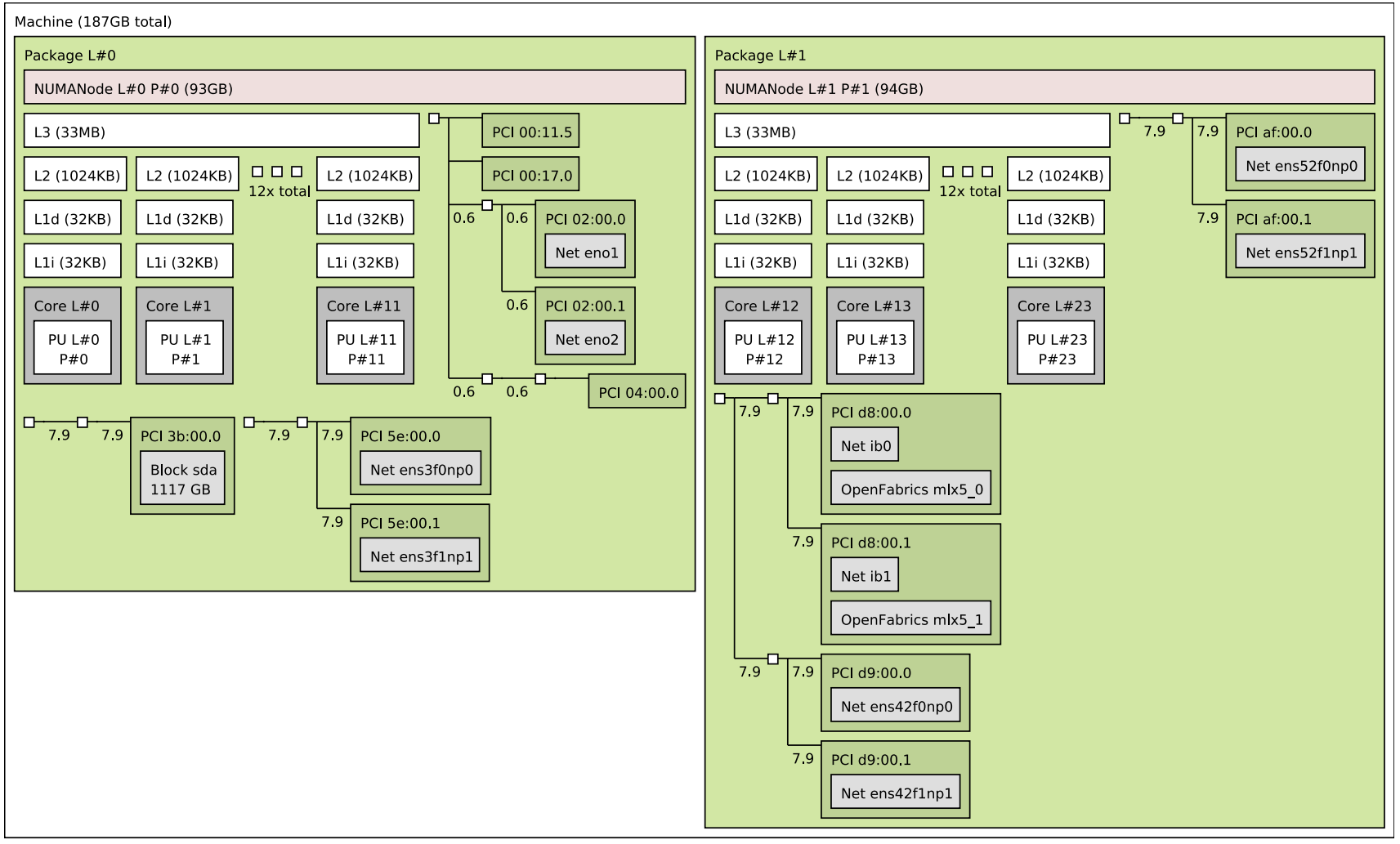
- Tuning Target
  1) Fairness - equal access for all customers
  2) Low Latency - Deterministic latency < 1ms most important

- Used Realtime kernel before → necessary patches now in normal RHEL
- Used infiniband for low latency → moving to ethernet

Tuned

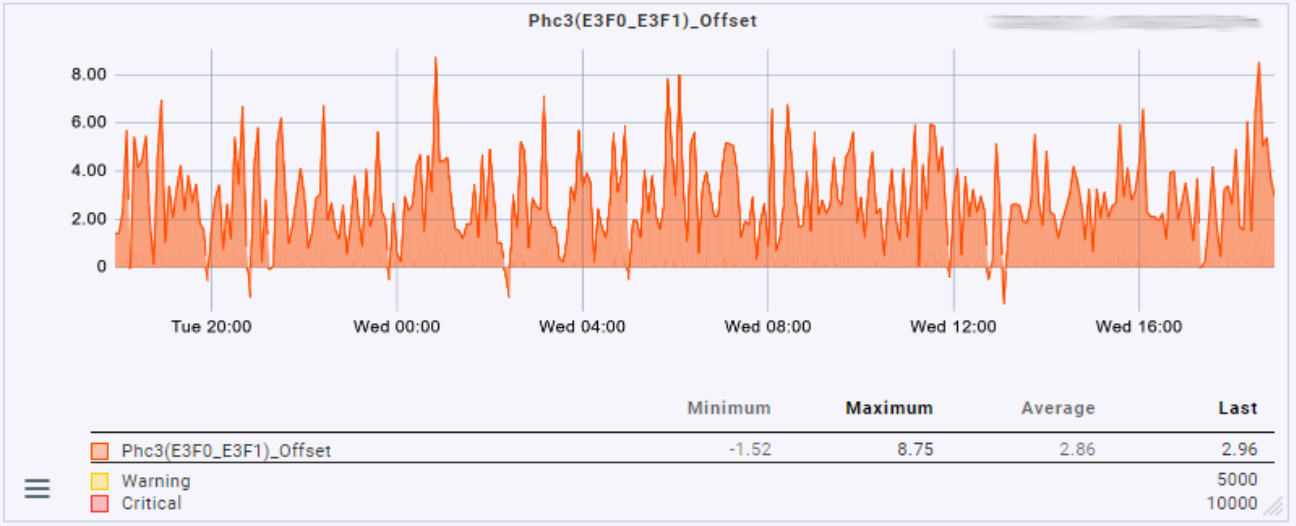
Tuning:

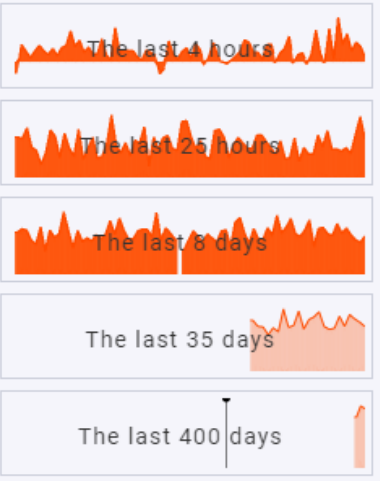
- Selected CPUs and Maschines
- Disable Powersaving state (C states) and HT
- Realtime priority
- Own tuned profiles incl.
  - Limit OS to use only specific CPU cores
  - PIN critical processes to same CPU as the network card

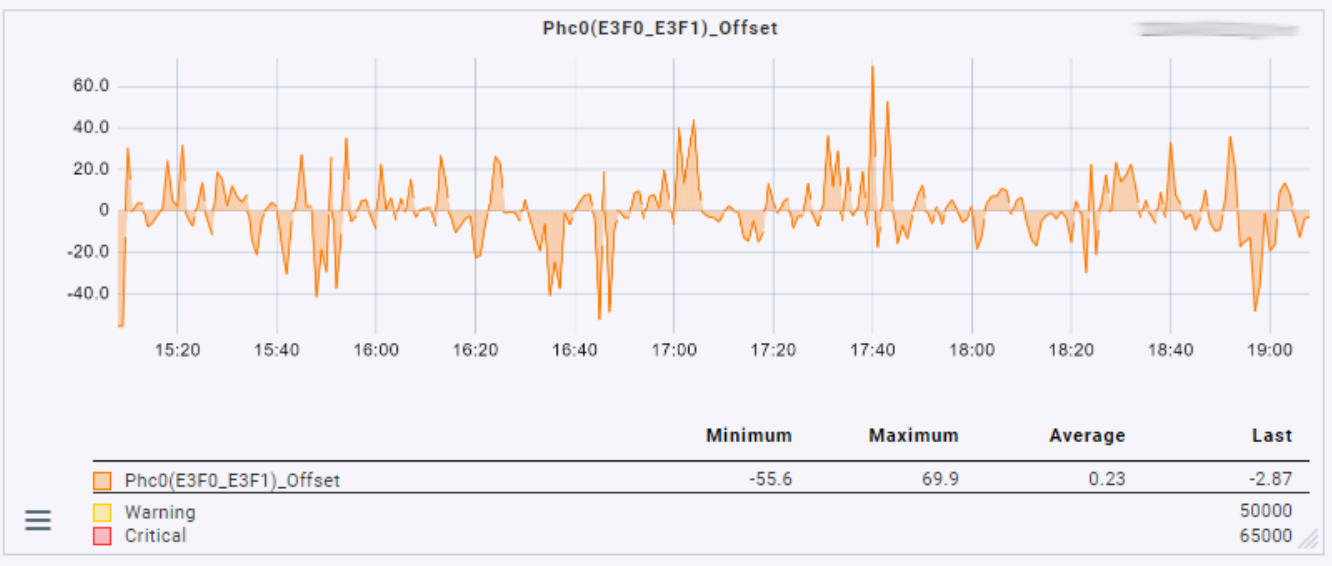

Machine (187GB total)
Package L#0
NUMANode L#0 P#0 (93GB)
L3 (33MB)
L2 (1024KB)
L2 (1024KB)
12x total
L2 (1024KB)
L1d (32KB)
L1d (32KB)
L1d (32KB)
L1i (32KB)
L1i (32KB)
L1i (32KB)
Core L#0
PU L#0 P#0
Core L#1
PU L#1 P#1
Core L#11
PU L#11 P#11
PCI 00:11.5
PCI 00:17.0
0.6
0.6
PCI 02:00.0
Net eno1
0.6
PCI 02:00.1
Net eno2
0.6
0.6
PCI 04:00.0
7.9
7.9
PCI 3b:00.0
Block sda 1117 GB
7.9
7.9
PCI 5e:00.0
Net ens3f0np0
7.9
PCI 5e:00.1
Net ens3f1np1
Package L#1
NUMANode L#1 P#1 (94GB)
L3 (33MB)
L2 (1024KB)
L2 (1024KB)
12x total
L2 (1024KB)
L1d (32KB)
L1d (32KB)
L1d (32KB)
L1i (32KB)
L1i (32KB)
L1i (32KB)
Core L#12
PU L#12 P#12
Core L#13
PU L#13 P#13
Core L#23
PU L#23 P#23
7.9
7.9
PCI af:00.0
Net ens52f0np0
7.9
PCI af:00.1
Net ens52f1np1
7.9
7.9
PCI d8:00.0
Net ib0
OpenFabrics mlx5_0
7.9
PCI d8:00.1
Net ib1
OpenFabrics mlx5_1
7.9
7.9
PCI d9:00.0
Net ens42f0np0
7.9
PCI d9:00.1
Net ens42f1np1

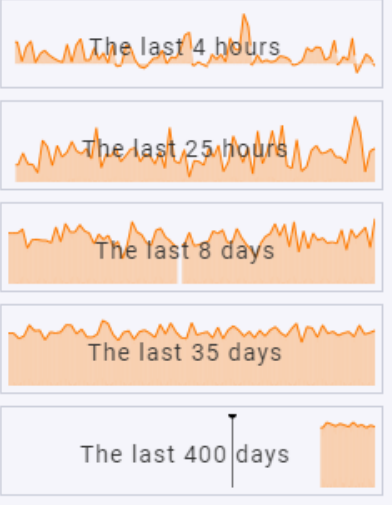
PTP and Cables

- To guaranty fairness every cable is measured (1m ≈ 2ns)
- Dedicated Network for distributing time information
- Network card has buildin Hardware clocks, synchronised with Precition Time Protocoll (PTP)

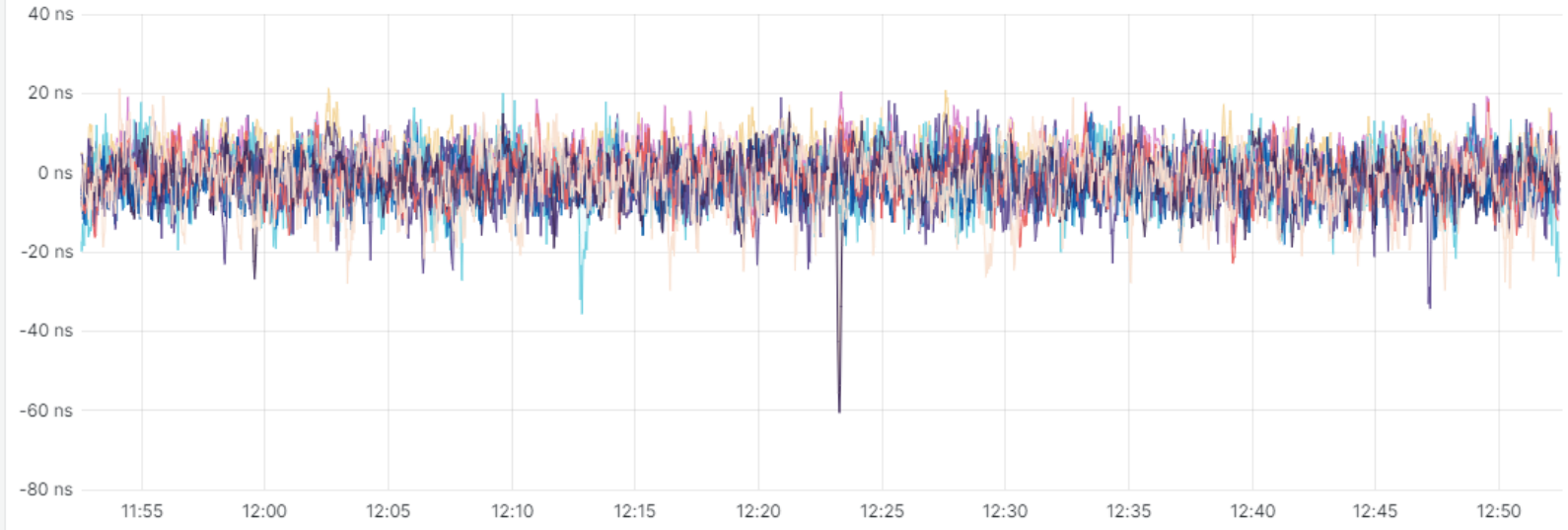

Phc3(E3F0_E3F1)_Offset
8.00
6.00
4.00
2.00
0
Tue 20:00
Wed 00:00
Wed 04:00
Wed 08:00
Wed 12:00
Wed 16:00
Minimum
Maximum
Average
Last
Phc3(E3F0_E3F1)_Offset
-1.52
8.75
2.86
2.96
Warning
5000
Critical
10000


The last 4 hours
The last 25 hours
The last 8 days
The last 35 days
The last 400 days


Phc0(E3F0_E3F1)_Offset
60.0
40.0
20.0
0
-20.0
-40.0
15:20
15:40
16:00
16:20
16:40
17:00
17:20
17:40
18:00
18:20
18:40
19:00
Minimum
Maximum
Average
Last
Phc0(E3F0_E3F1)_Offset
-55.6
69.9
0.23
-2.87
Warning
50000
Critical
65000

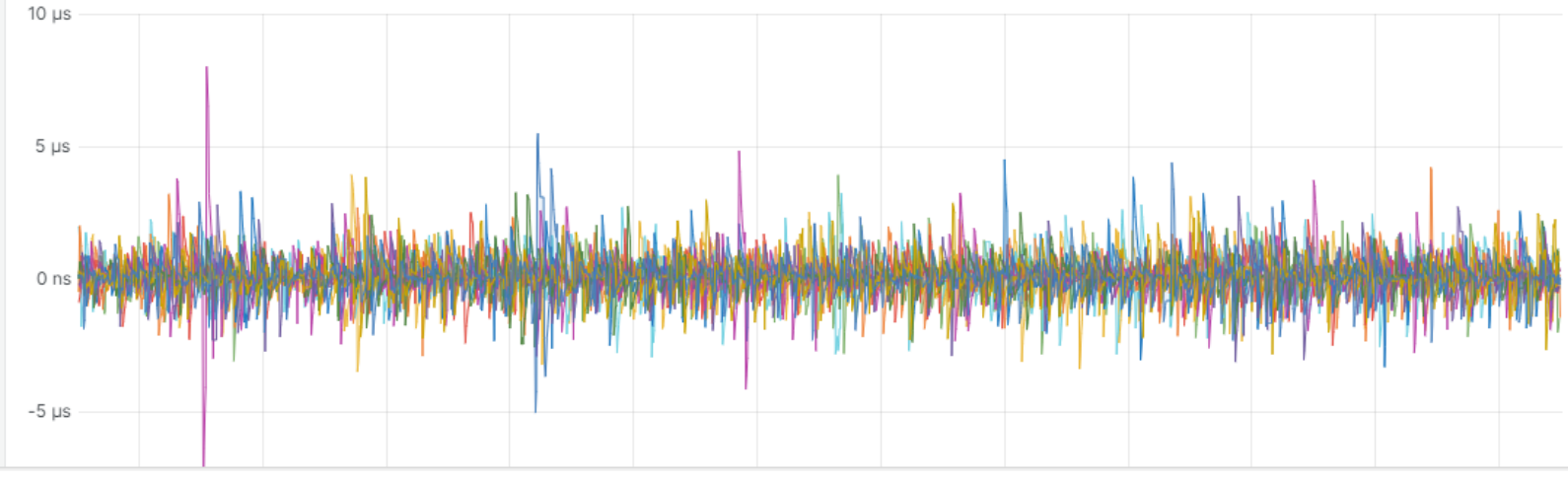

The last 4 hours
The last 25 hours
The last 8 days
The last 35 days
The last 400 days

system offset

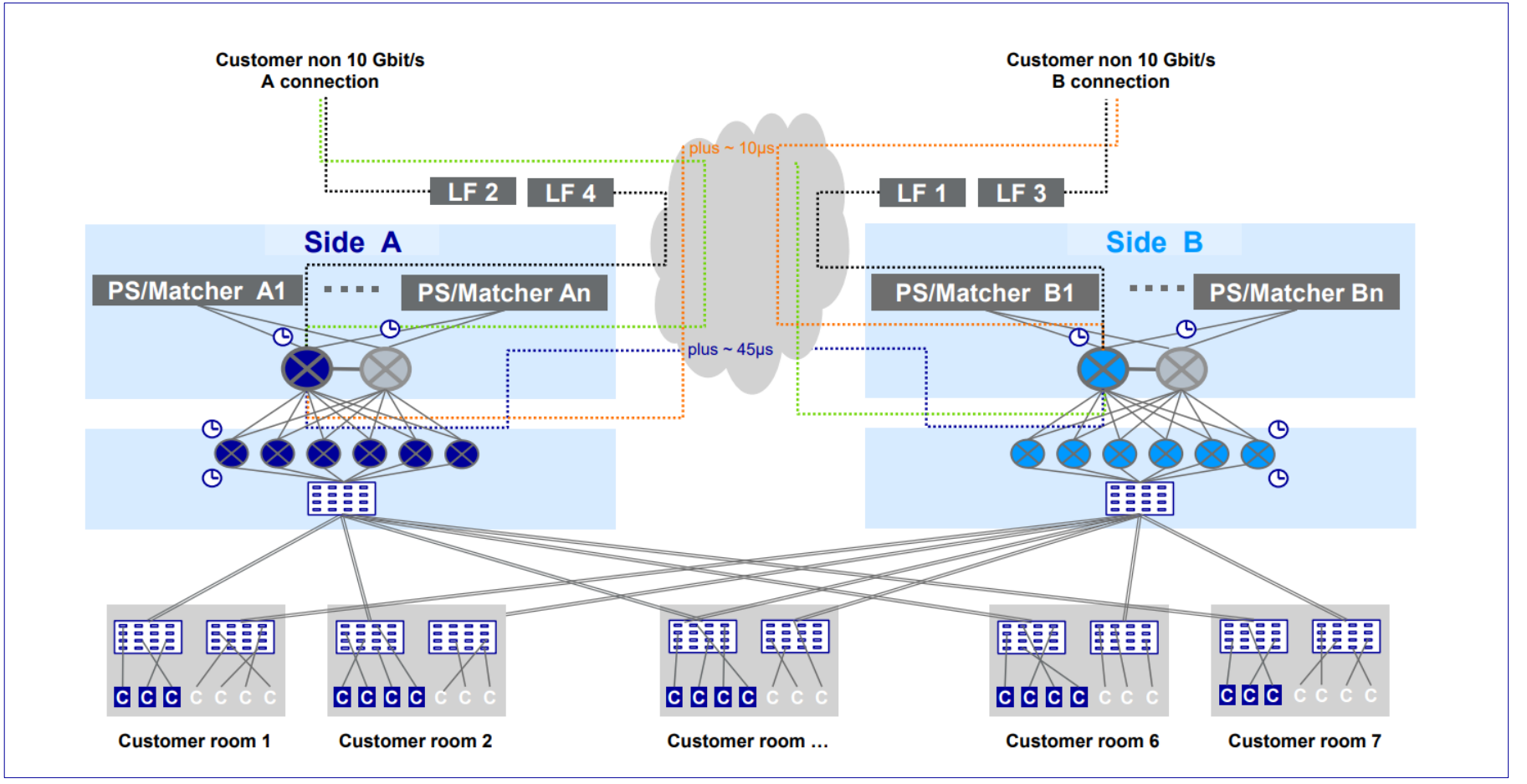


	min	max
[redacted]	-18.2 ns	10 ns
[redacted]	-17.1 ns	14.7 ns
[redacted]	-18.2 ns	14.8 ns
[redacted]	-60.6 ns	15.4 ns
[redacted]	-34.4 ns	17.4 ns
[redacted]	-22.9 ns	17.7 ns
[redacted]	-56.4 ns	18.1 ns
[redacted]	-12.4 ns	18.9 ns
[redacted]	-35.6 ns	19.9 ns
[redacted]	-18.9 ns	20.4 ns
[redacted]	-29.7 ns	21.3 ns
[redacted]	-11.1 ns	21.3 ns

[redacted]



	min	max
[redacted]	-3.10 µs	3.91 µs
[redacted]	-3.49 µs	3.91 µs
[redacted]	-2.96 µs	3.22 µs
[redacted]	-2.91 µs	4.21 µs
[redacted]	-2.52 µs	2.49 µs
[redacted]	-3.31 µs	4.50 µs
[redacted]	-8.47 µs	8.00 µs
[redacted]	-3.14 µs	3.12 µs
[redacted]	-2.58 µs	3.25 µs
[redacted]	-2.84 µs	3.84 µs
[redacted]	-5.06 µs	5.49 µs


Customer non 10 Gbit/s
A connection
Customer non 10 Gbit/s
B connection
plus ~ 10µs
LF 2
LF 4
LF 1
LF 3
Side A
Side B
PS/Matcher A1
PS/Matcher An
PS/Matcher B1
PS/Matcher Bn
plus ~ 45µs
Customer room 1
Customer room 2
Customer room ...
Customer room 6
Customer room 7

Thanks for your attention

Questions?

Sources:

- https://www.eurex.com/ex-en/support/technology/t7