ISPCS 2019 Keynote: Clock Synchronization for the Low-latency Financial Sectors

Pedro V. Estrela, Ph.D.
Senior Performance Engineer
Sep-2019
https://github.com/pestrela/papers

Acknowledgements:
- IMC Financial Markets
- Deustche Börse

Outline

- How electronic markets work

- WAN technologies & PTP monitoring

- LAN technologies & PTP monitoring

- Latest regulations

- PTP robustness

- Comments to other industries

About the presenter

- PhD in IP mobile networks (2007)
- Performance engineer in Financial markets (2008)
  - “Formula 1 engineer”
  - Latency optimization
  - Prototypes and reverse engineering
  - PTPd open source client rewrite

How electronic markets work (Participants / Exchanges)

Electronic Financial Markets

- Think of the airport currency shop, but:
  - Specialized derivative products (futures / options / bonds)
  - Fully automated
  - Global

  - Huge trading + technology needs
  - Fully regulated (in particular, timestamps)

Low-latency Market participants

	Business Model	Price Opinion	Technological Needs
Large Investors	Seek Investment opportunities and Insurance	Yes	Execute orders quickly
Market-makers	Continuously provide liquidity to the market (both Buy *and* Sell prices)	No	Limit the Risk of stale quotes
Exchanges	Continuously sort orders by best price, then best time	No	Ensure a deterministic service

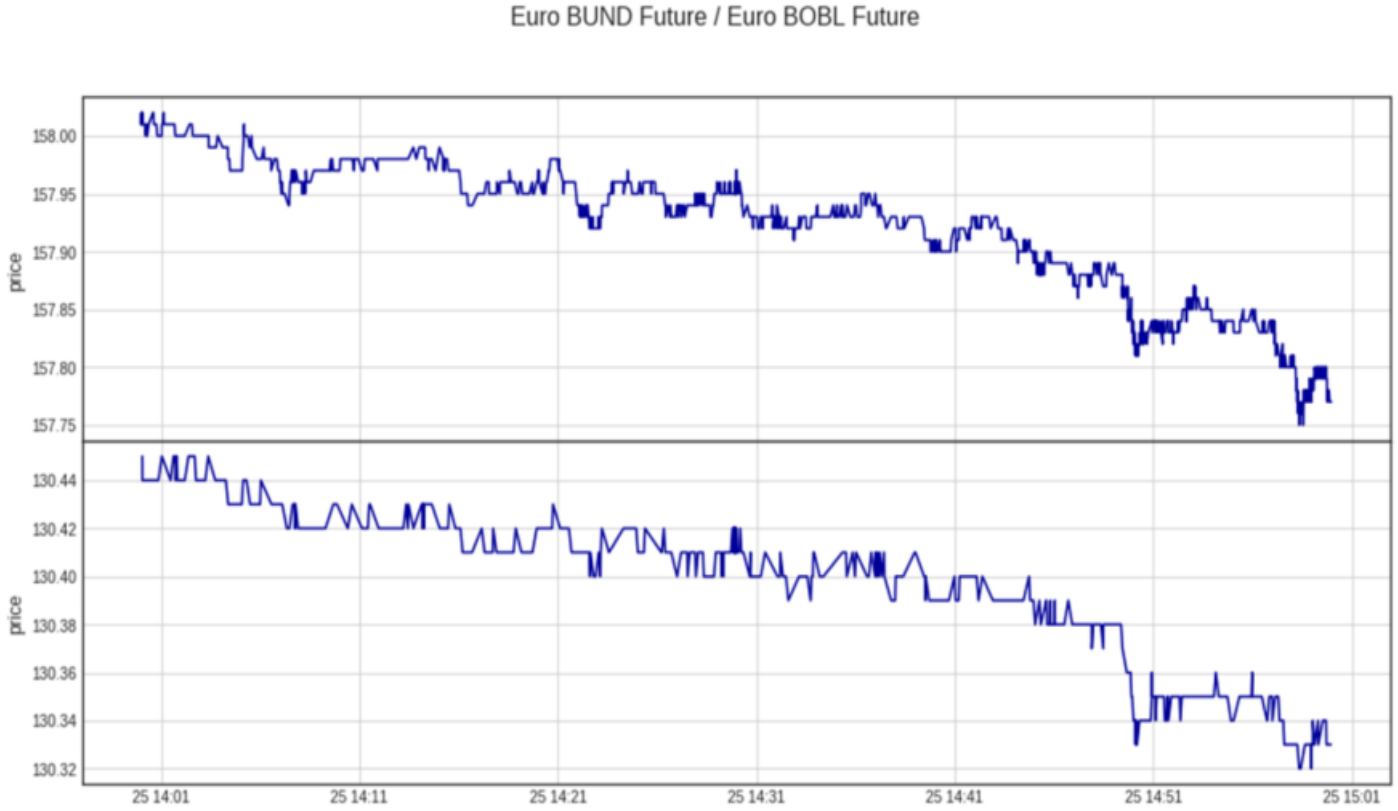
Market Makers

https://www.youtube.com/watch?v=WFsvY_YRhvg

Correlated products

“GOLD”

“SILVER”



Source: DB White Rabbit workshop presentation 2018

Exchanges: Order Book

Cheapest Sellers

Ask	AskVolume	AskCount
73.14	1104	4
73.13	4090	5
73.12	534	4
73.11	1151	5
73.10	1908	11
73.09	6388	15
73.08	4284	14
73.07	5854	17
73.06	2662	15
73.05	3081	5

Ask	AskVolume	QueuePos
73.05	100	5
73.05	791	4
73.05	1368	3
73.05	685	2
73.05	137	1

Fastest Sellers

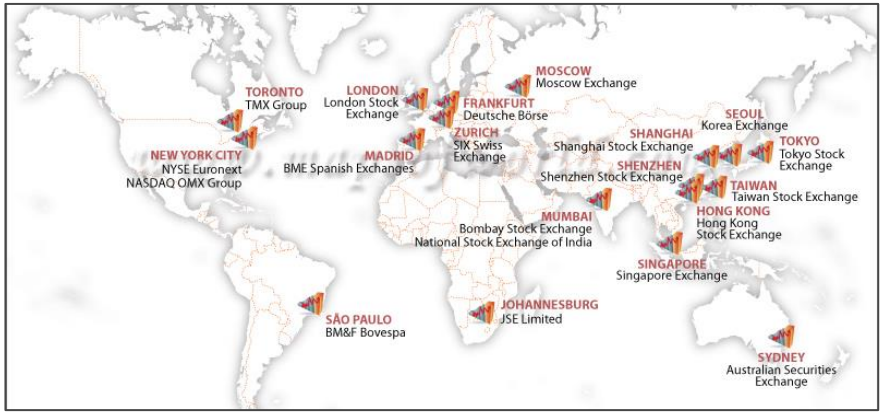
Spendiest Buyers

Bid	BidVolume	BidCount
73.04	624	3
73.03	1474	10
73.02	2505	14
73.01	2843	16
73.00	1925	10
72.99	2328	12
72.98	2814	11
72.97	967	7
72.96	1803	10
72.95	938	8

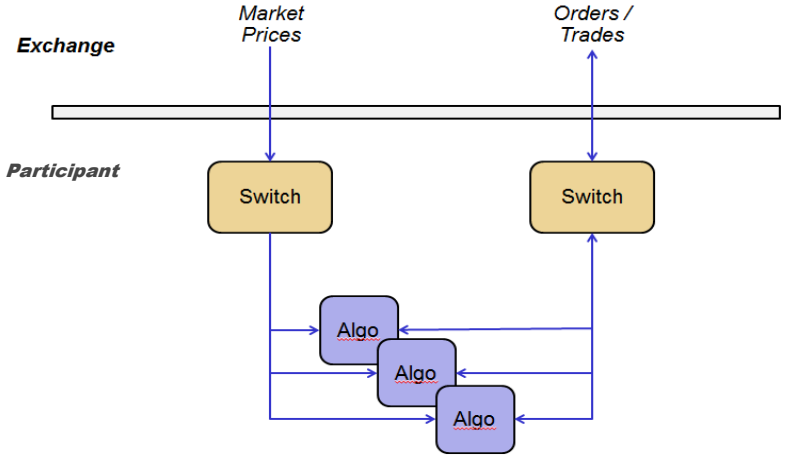
Bid	BidCount	QueuePos
73.04	100	1
73.04	91	2
73.04	433	3

Fastest Buyers

Wide Area



Local Area

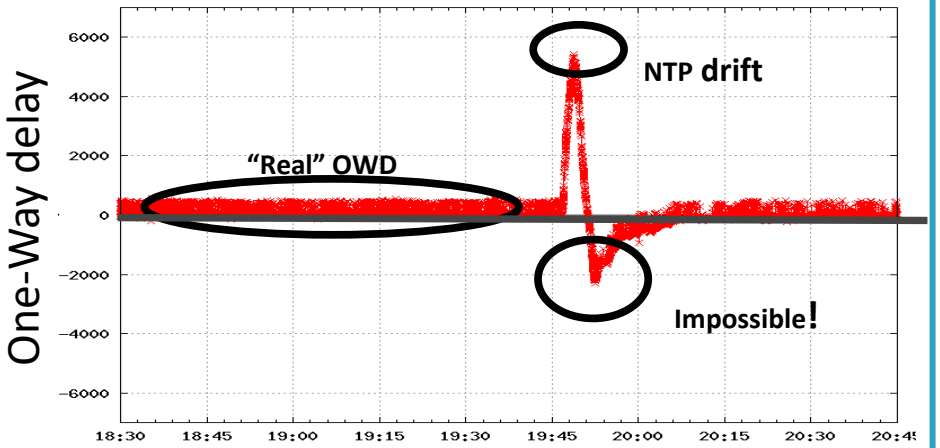


WAN low-latency technologies & PTP monitoring

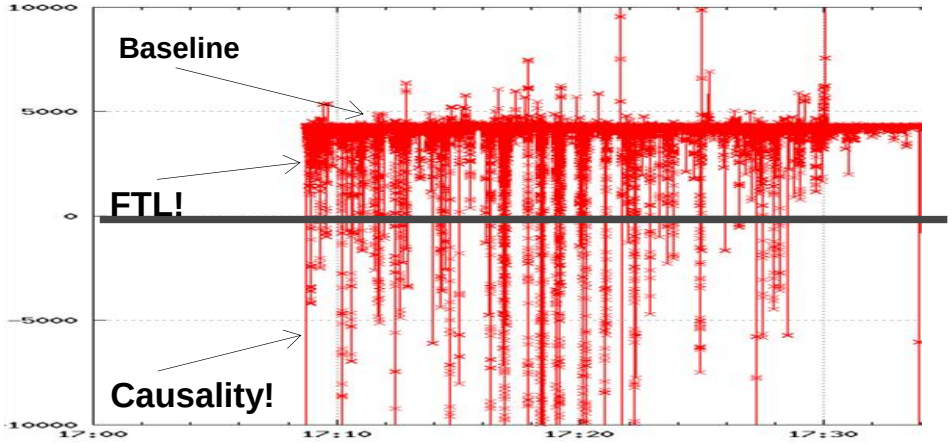
(on a wide area)

NTP starting point (2010)

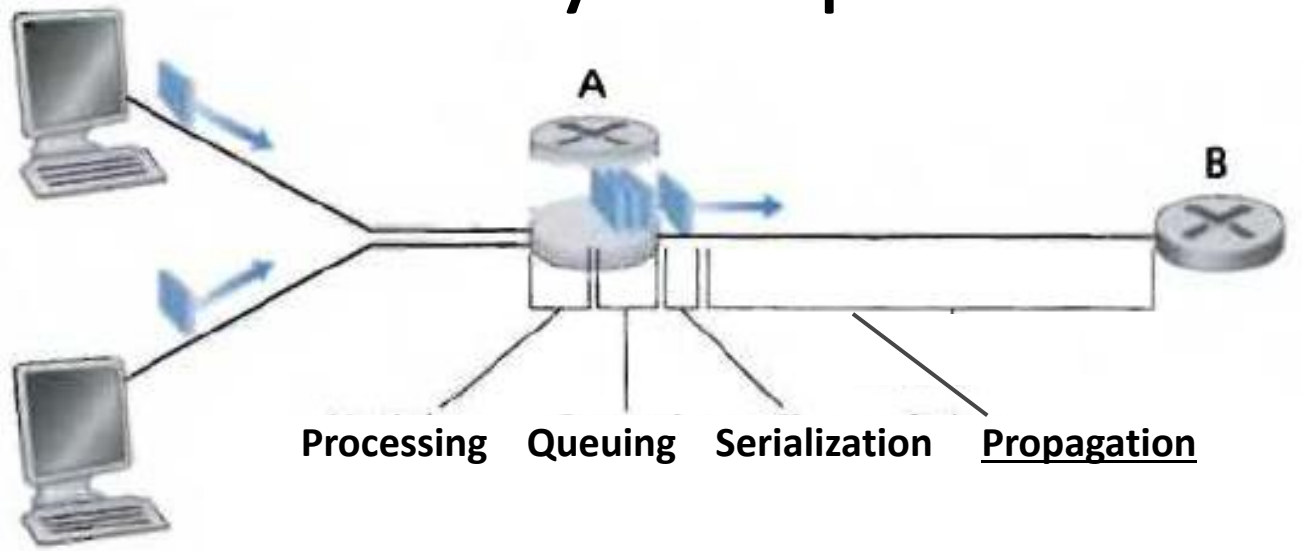
Clock Sync (NTP)



Timestamps (software)



Latency components



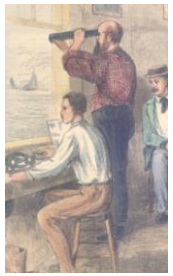
Trend is clearly: Radio / Raw Hardware / More expensive

WAN examples

Year	Technology	Who
1815	Pigeons	*Rothchild knew first that Napoleon lost the war*
1836	Telescopes	*Shore agents knew if cargo was spoilt on Boats*
1897	Telegraph	*Sending horse race results to outside*
…	…	…
1998	DSL	*Digital data over telephone lines*
2005	Satellite	*Geo stationary satellites send market data*
2010	Fiber	*Spread networks drills mountains on NY-Chicago*
2012	Microwave	*McKay jumps the same mountains using Radio*
2015	Fiber	*Hibernia builds new straighter Atlantic cable*

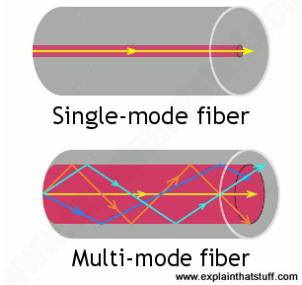
Days / Hours

Milli Seconds



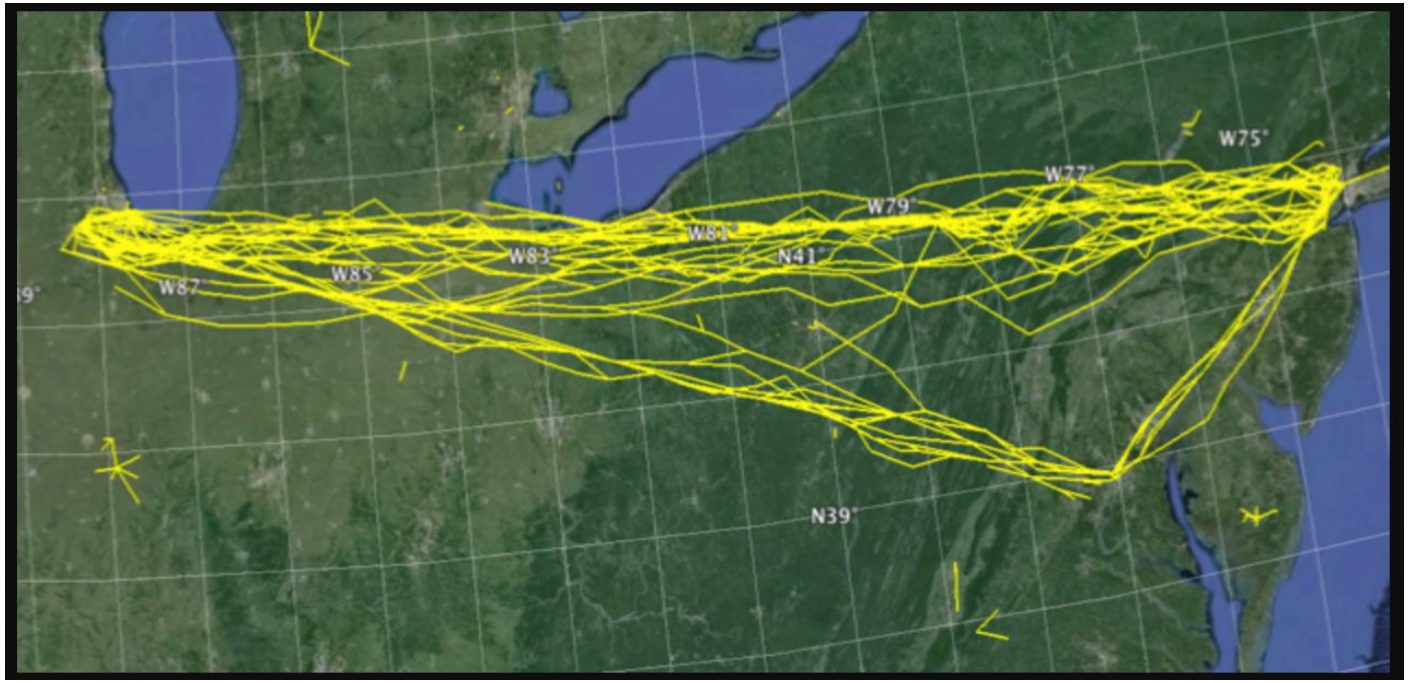
Source: http://www.forbes.com/forbes/2010/0927/outfront-netscape-jim-barksdale-daniel-spivey-wall-street-speed-war.html
Source: https://www.moaf.org/publications-collections/financial-history-magazine/111/_res/id=Attachments/index=0/Plundered_by_Harpies.pdf

Fiber paths in US



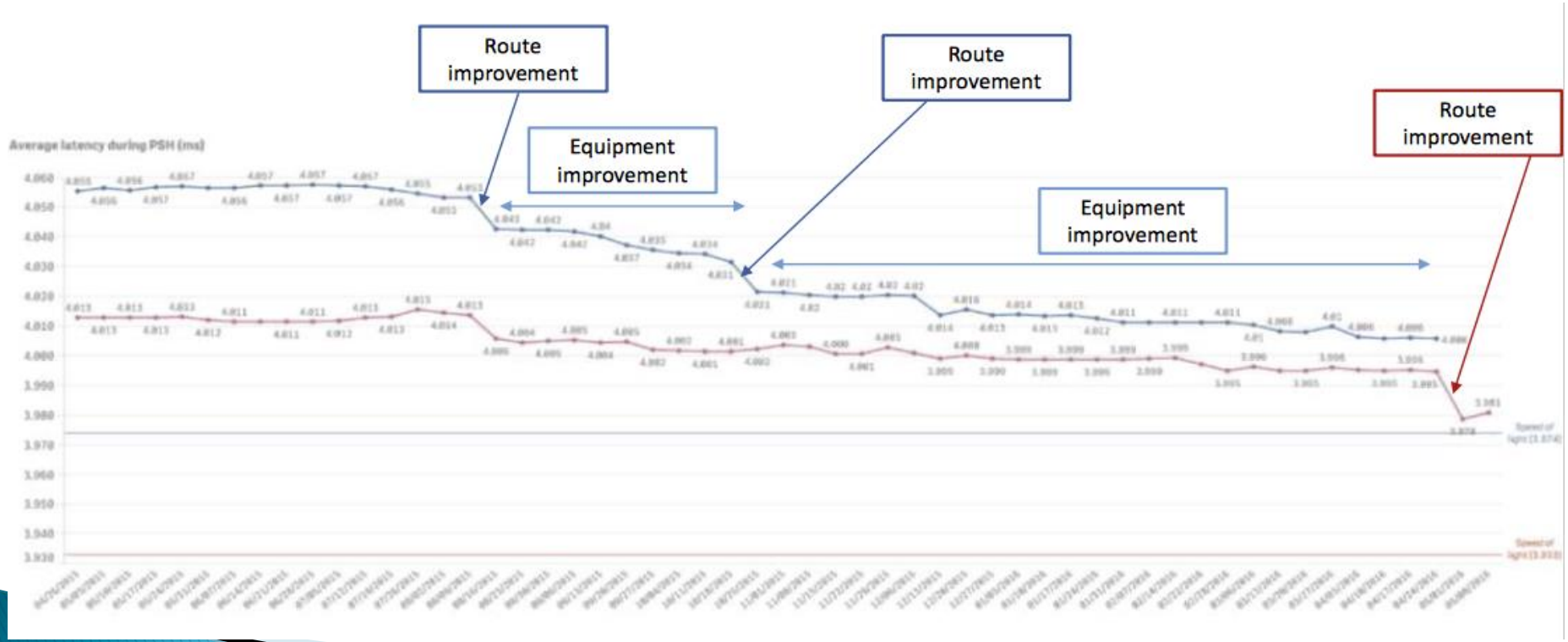
https://www.zayo.com/services/spread-networks/

Microwave paths in US



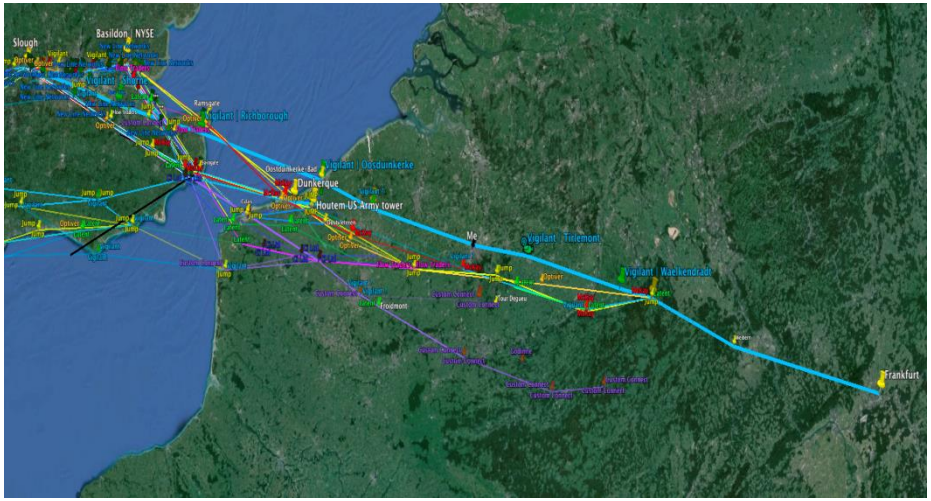
Data analysis: McKay Brothers

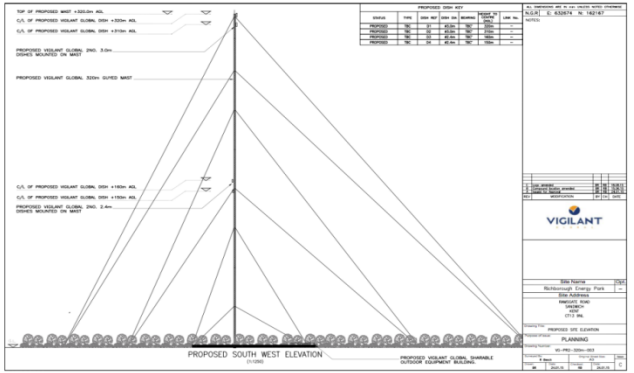
Chicago->NY latency



Data: McKay Brothers

Microwave paths in EU





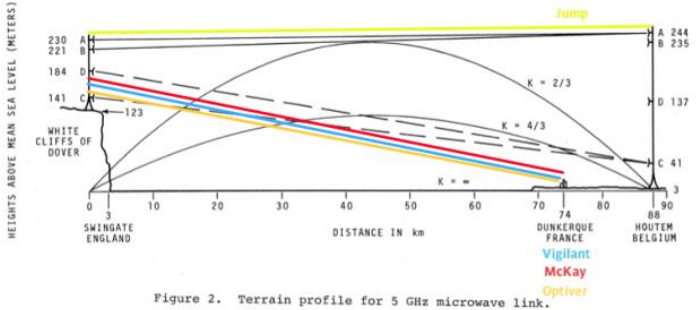
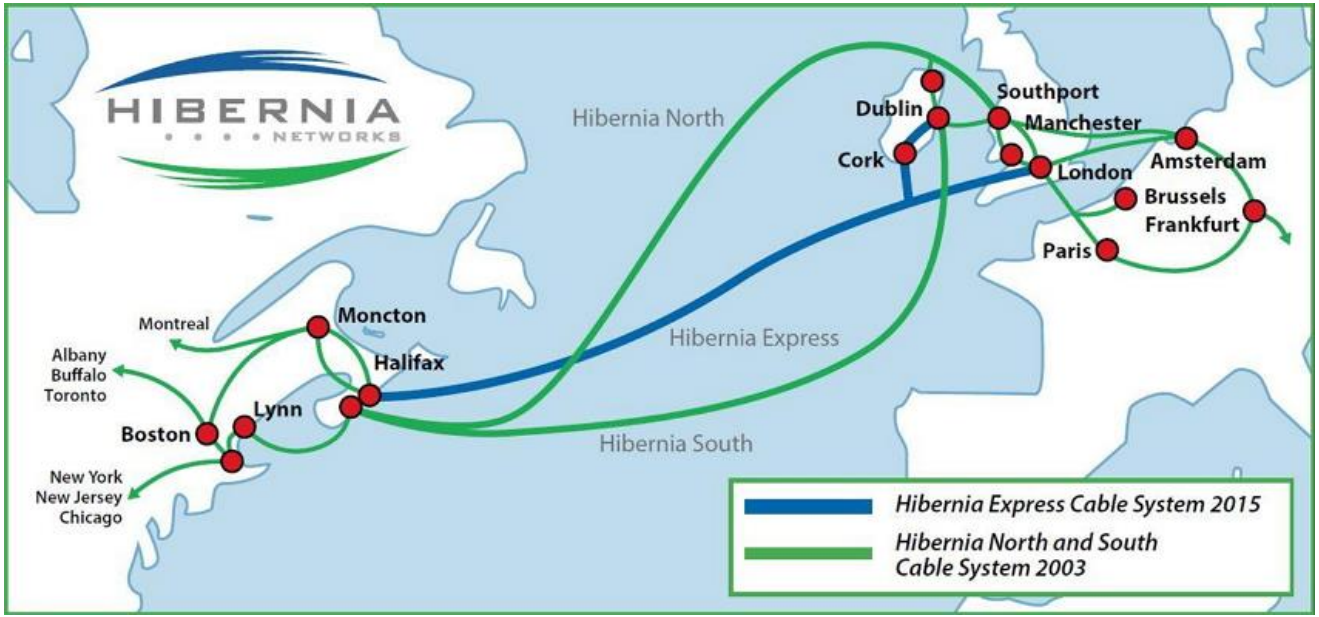
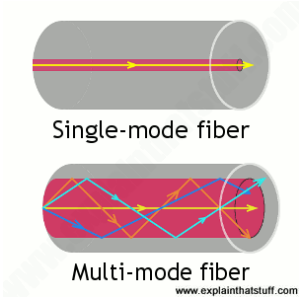


Figure 2. Terrain profile for 5 GHz microwave link.

https://sniperinmahwah.wordpress.com/2016/01/26/hft-in-the-banana-land/

Atlantic fiber paths

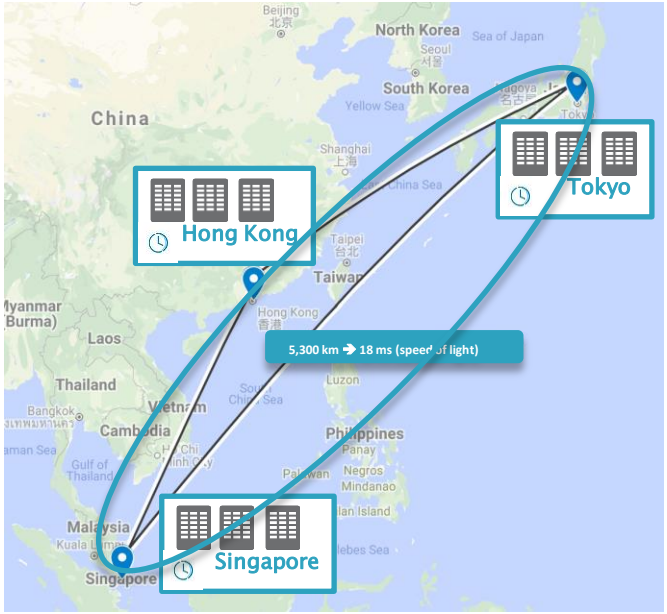




https://www.submarinenetworks.com/systems/trans-atlantic/project-express

Multi-venue Trades

=> Time Perimeter + Hold-and-Release Buffers



1. Hold SG trade in buffer
2. Send it to HK and Tokyo
3. Release it "simultaneously" in SG, HK and Tokyo at future time

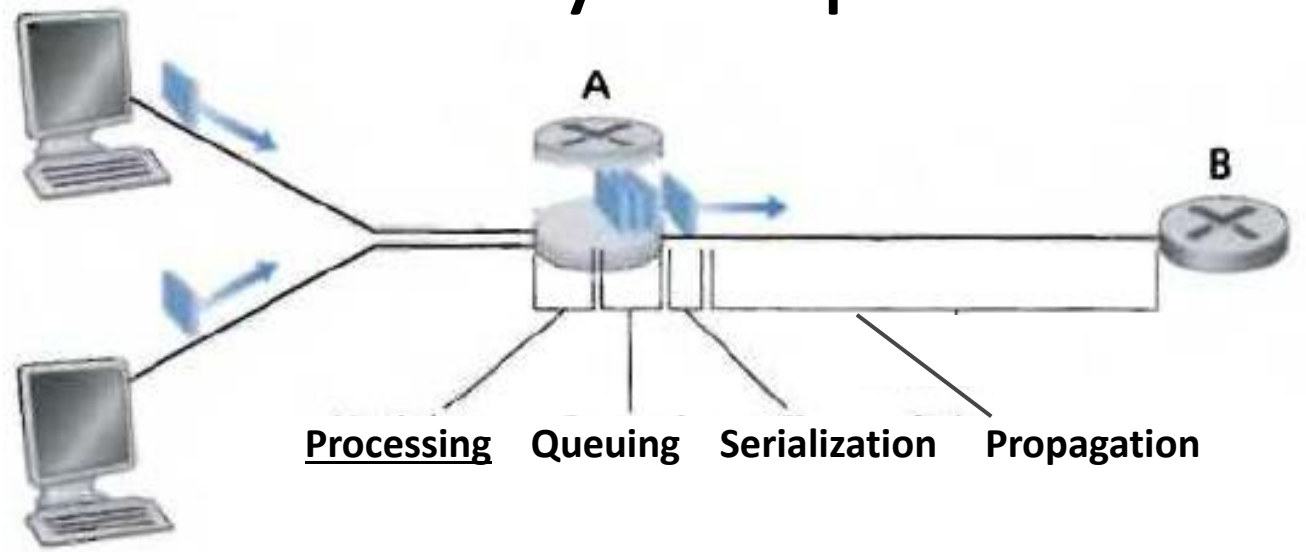
4. Clock sync accuracy <= C

https://stacresearch.com/system/files/resource/files/STAC-Summit-13-Nov-2019-Tick%20Tock.pptx

LAN low-latency technologies & PTP monitoring

(on the same datacenter)

Latency components



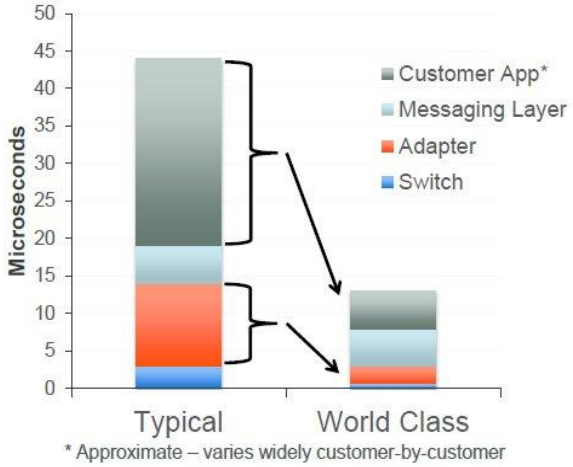
Trend is clearly: Faster / Raw Hardware / More expensive

LAN examples

Year	Device	Latency
2008	Cisco 4900	*2600ns OWD*
2011	Cisco 3064	*1000ns OWD*
2011	Arista 7124	*500ns OWD*
2013	Cisco 3548	*200ns OWD*
2017	Metamux 48	*69ns OWD*
…	…	…

Micro Seconds

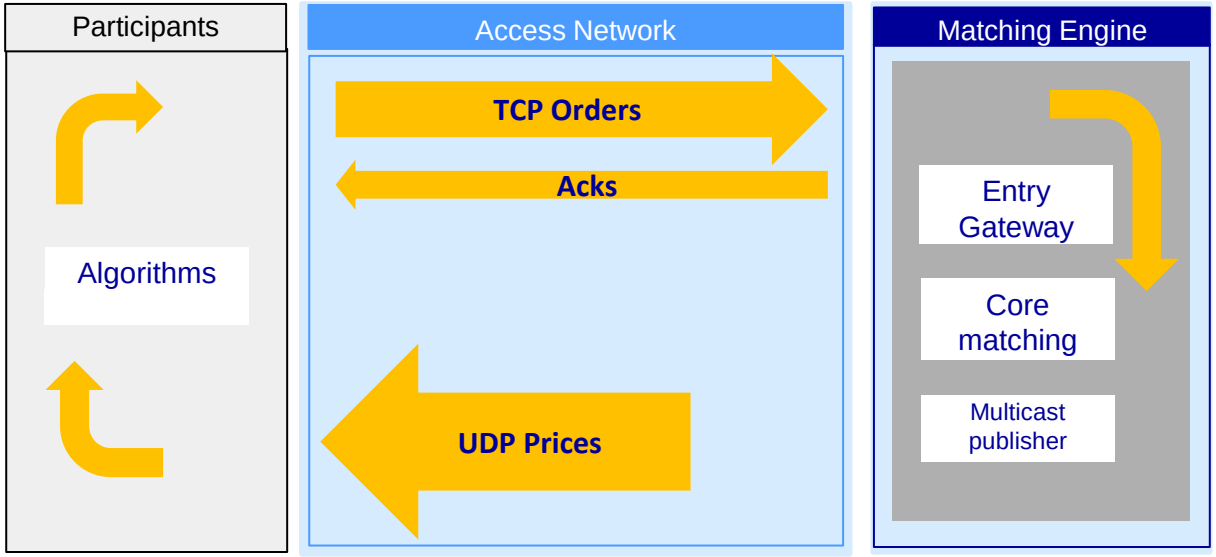
Nano Seconds



(Comparison circa 2012)

Source: https://www.cisco.com/c/dam/en/us/products/collateral/switches/catalyst-4900-series-switches/press_coverage.pdf
Source: https://www.arista.com/en/company/news/press-release/352-pr-20110314-01
Source: https://newsroom.cisco.com/press-release-content?type=webcontent&articleId=1028561
Source: https://newsroom.cisco.com/press-release-content?articleId=362594
Source: https://www.theregister.co.uk/2012/02/08/solarflare_application_onload_engine/
Source: https://meanderful.blogspot.com/2016/11/metamako-69ns-breaks-their-own-record.html

Co-location Loop



Median round-trip from order entry to acknowledgement ≈ **60µs** (TCP->TCP)
Fastest participants have **sub 100ns** response times (UDP->TCP)

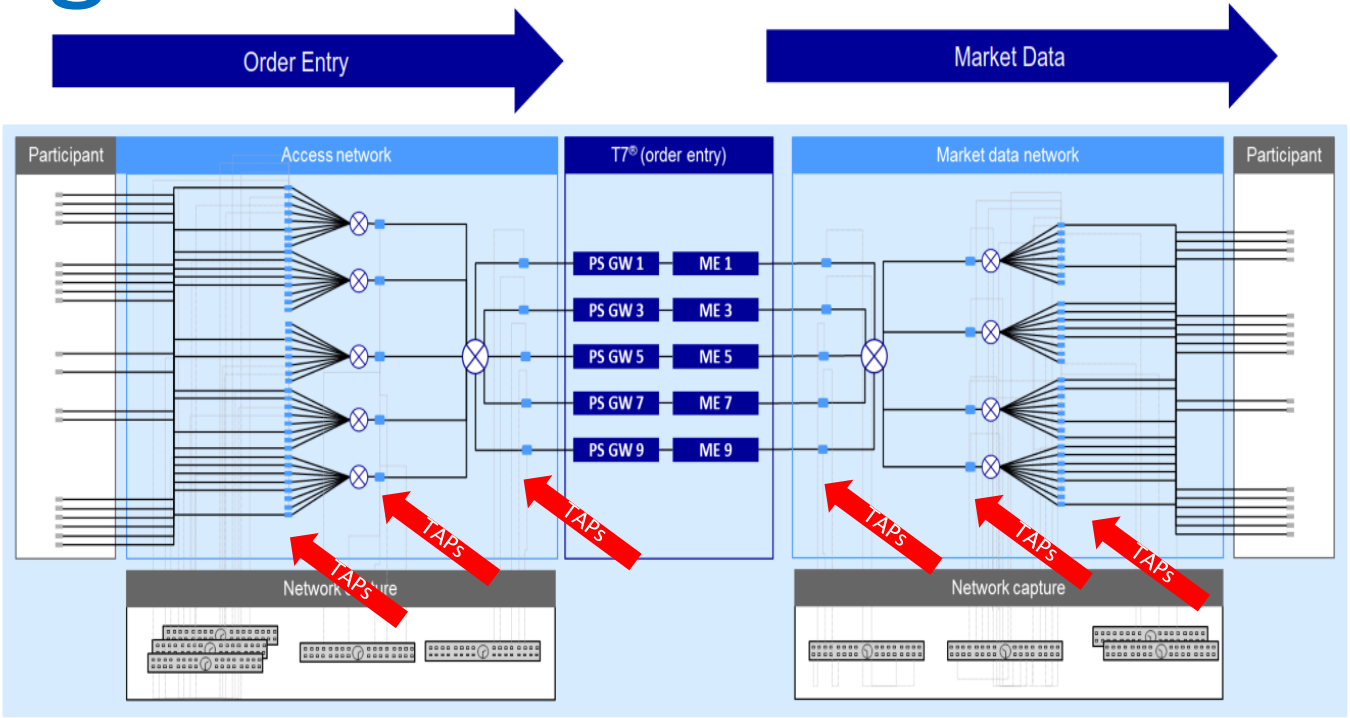
Source: DB White Rabbit workshop presentation 2018

Monitoring Scale

Scale:

- 500+ capture ports
- 60+ capture devices
- 4 data center modules

- PTP +/- 60ns jitter (at best)
- Serialization time: 120ns
- Goal: sub-10ns precision



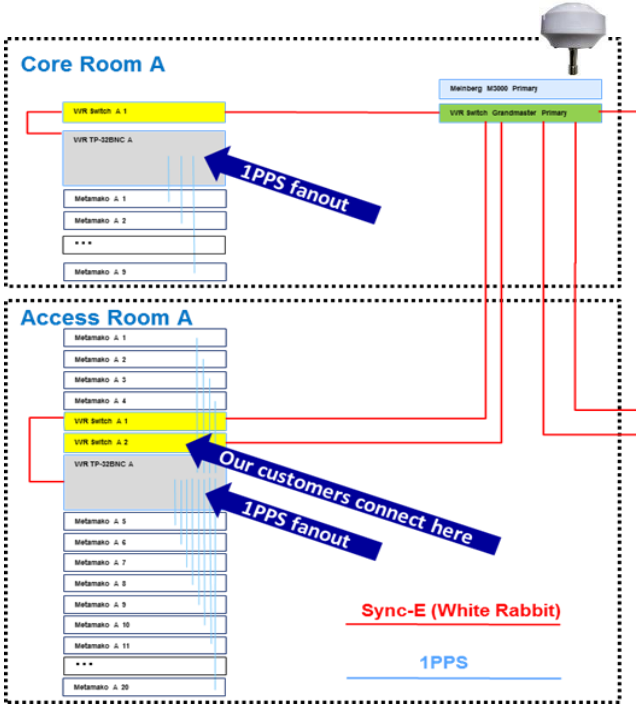
Source: DB White Rabbit workshop presentation 2018

White Rabbit solution

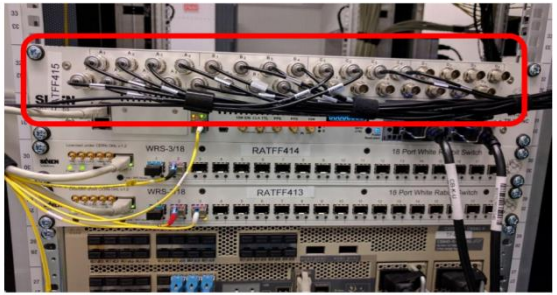
Issues:
- Distances too long for PPS over coax cables
- No white rabbit in NICs and switches

Solutions:
- Used white rabbit to distribute 1PPS to 60x capture devices
- Once timestamped, capture buffers absorb bursts

New Services:
- WR sync to participants
- High accuracy timestamps logs



White Rabbit PPS Distribution Devices



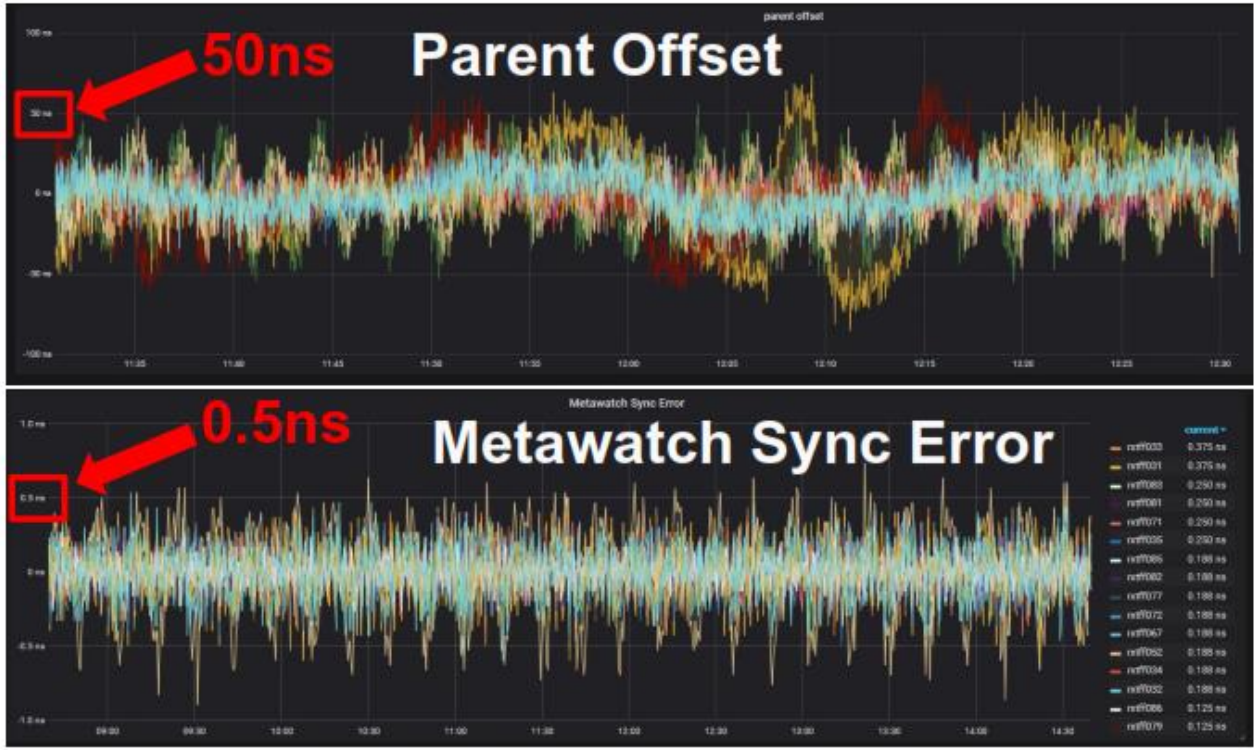
Capture Devices

Source: DB White Rabbit workshop presentation 2018

White Rabbit results #1

PTP

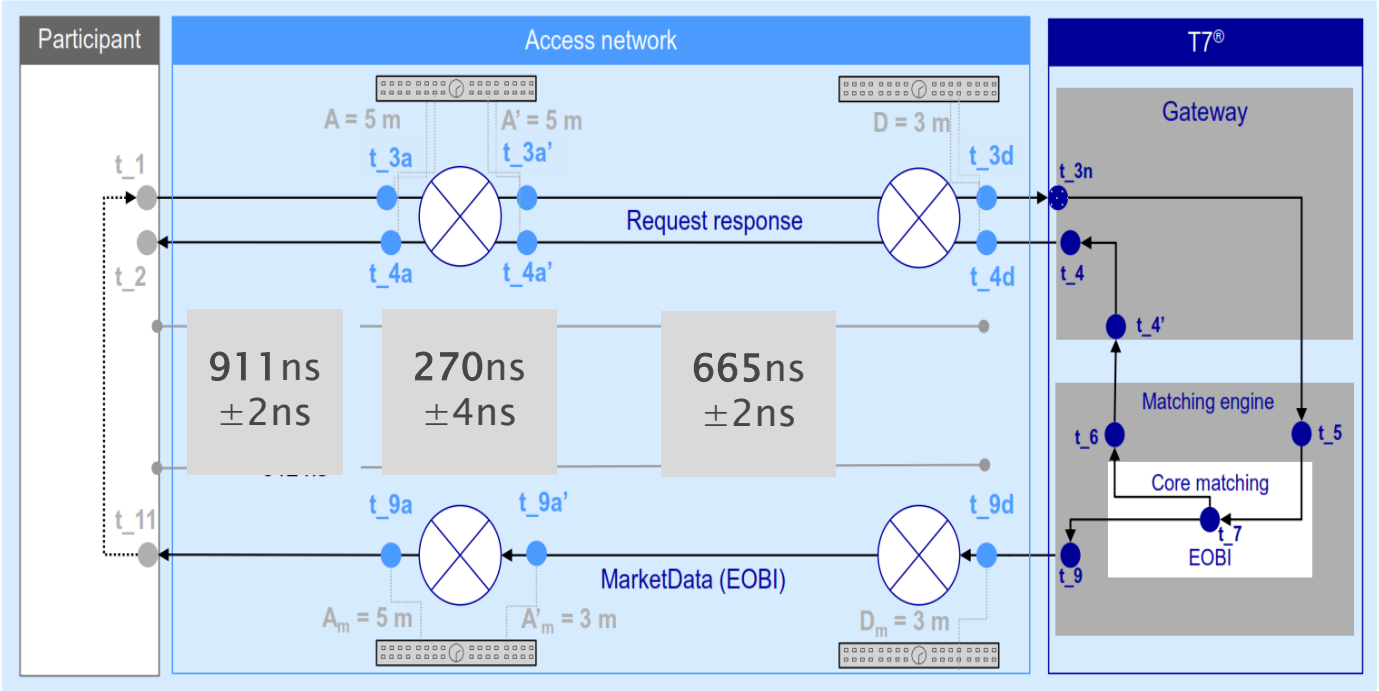
White Rabbit



Source: DB White Rabbit workshop presentation 2018

Co-Location Loop

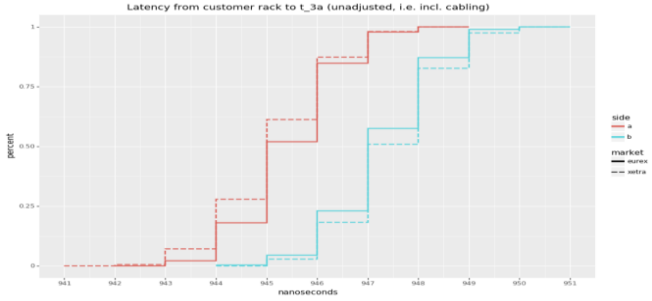
- Cable lengths equalized
- Switch OWDs measured
- Physical location equalized



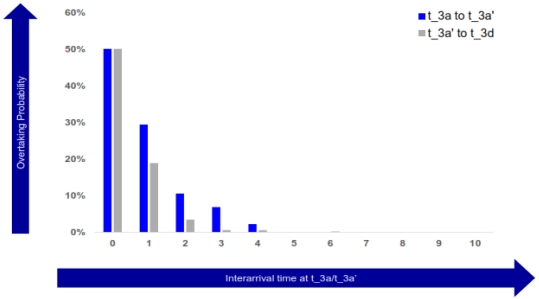
Source: DB White Rabbit workshop presentation 2018

State of the art determinism

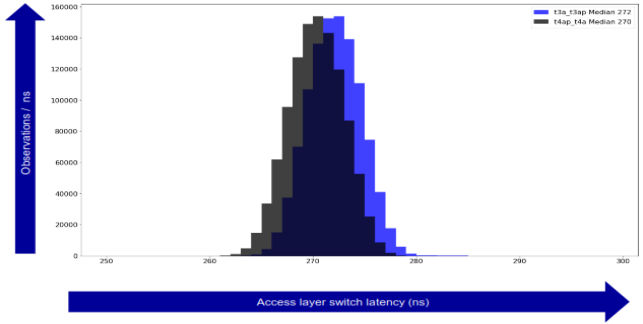
Cable lengths: 911ns ±2ns



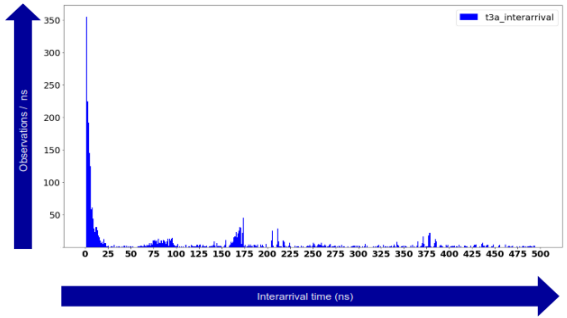
Switch determinism: <5ns



Switch jitter: 270ns ±4ns

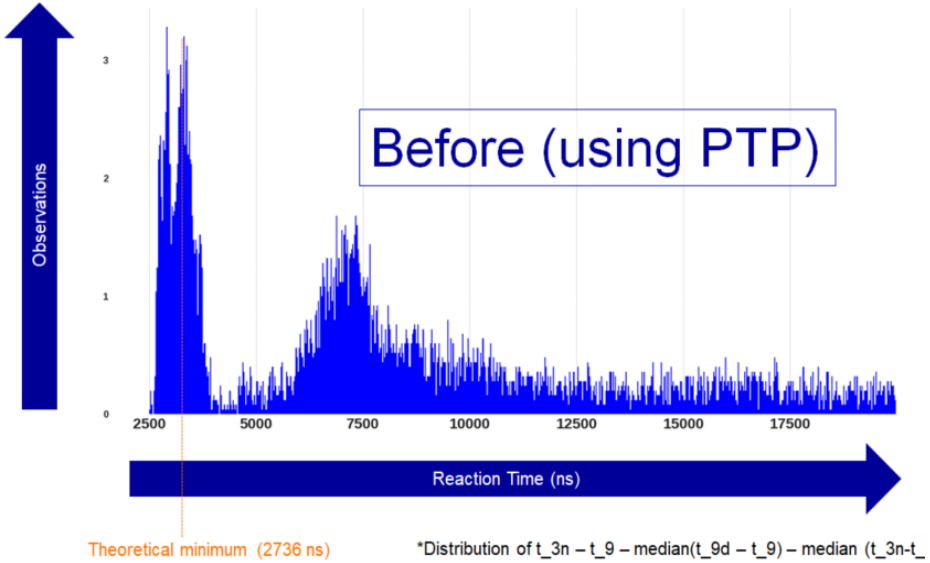


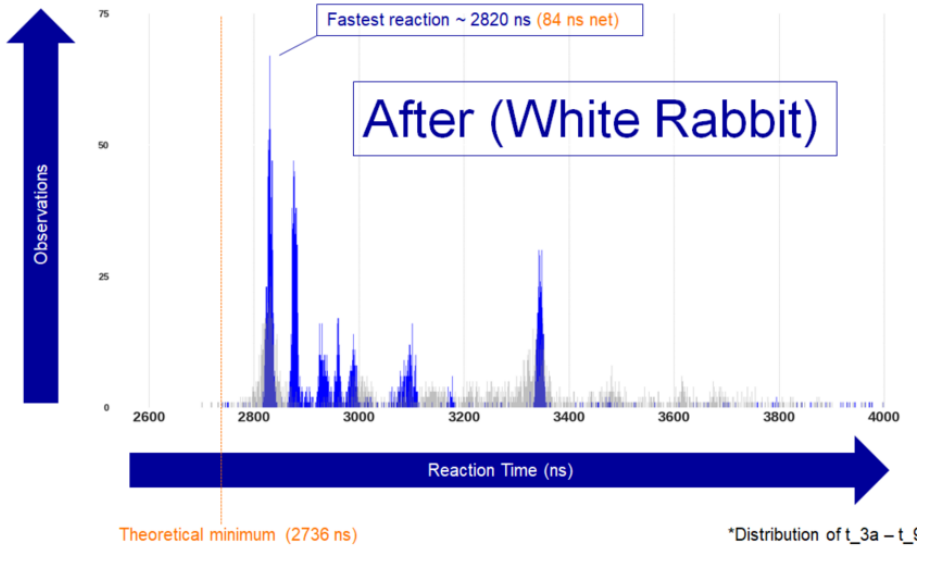
Order InterArrival time



Sources: DB STAC June 2019 presentation

Fastest participants: reaction





Sources: DB White Rabbit workshop presentation 2018

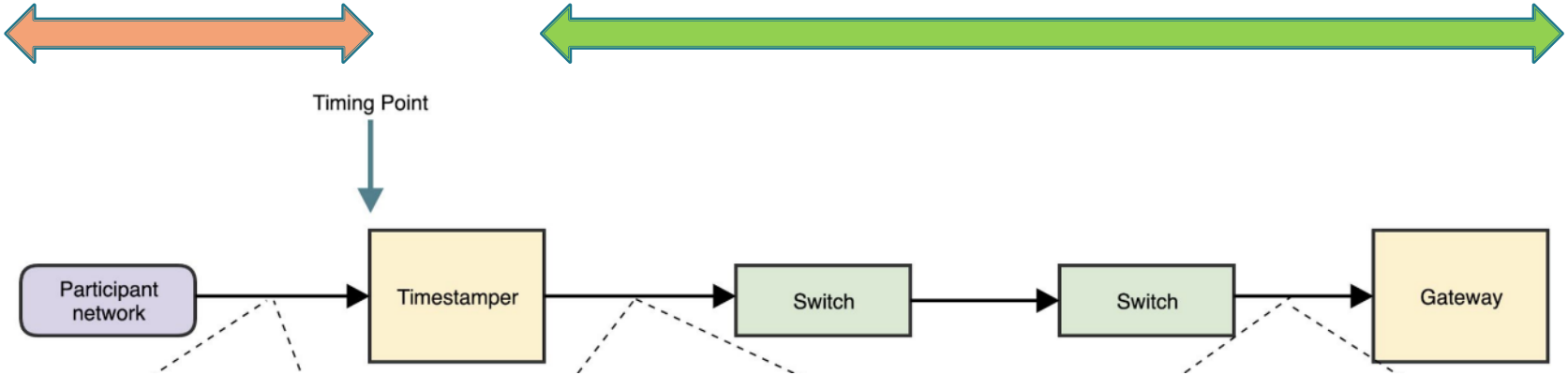
Timestamp at Edge (input)

Latency sensitive:

- Equalized cables = fixed offset
- Single data rate (10G)
- FPGAs, Cut-trough

NOT latency sensitive:

- Deep buffer switches
- Store-and-forward
- Dynamic routing, redundant infra
- VM, cloud, GC
- Buffer reordering = Still needs absolute maximum guarantee!



https://stacresearch.com/system/files/resource/files/STAC-Summit-13-Nov-2019-Arista.pdf

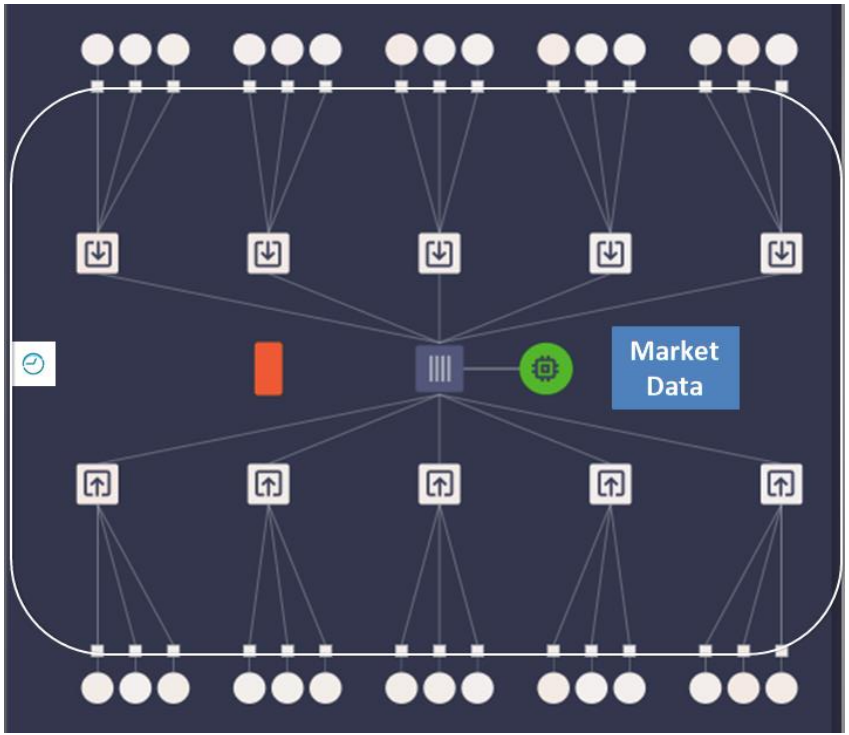
Multicast replacement (Output)

Latency sensitive:

- Release data simultaneously at the edge
  - This is a lot harder than input case!
- Equalized fibers (+latency offset)
- Multicast replication or L1 replication

NOT latency sensitive:

- Same techniques as input case
- Similar comments to absolute maximum case



https://stacresearch.com/system/files/resource/files/STAC-Summit-13-Nov-2019-Tick%20Tock.pptx

Latest regulations
MIFID-II RTS 25

MIFID II RTS 25

- Rule:
  - http://ec.europa.eu/finance/docs/level-2-measures/mifid-rts-25-annex_en.pdf
  - Maximum divergence from UTC: 100 microseconds
  - No provision for outliers!
- Guidelines:
  - https://www.esma.europa.eu/file/20011/download?token=cHI6iMY4
  - "**Relevant** and **proportionate** testing of the system should be required along with relevant and proportional monitoring thereof to ensure that the divergence from UTC remains within tolerance."

Proposal for recursive outliers

- RTS-25 today:
  - <100us
- Idea:
  - X% of business time: >0.1ms outliers
  - 0.X% of business time: >1ms outliers
  - 0.0X% of business time: >10ms outliers
  - 0.00X% of business time: >100ms outliers

A) PTP Deployment: Best practices

- Redundant GPS infrastructure
- Redundant PTP switches
  - Stable internal network
- Custom PTP clients
  - multi-clock robustness
  - WAN filters

B) Monitoring: Self-Health

- Continuous monitoring of:
  - Self-reported clock offsets
  - Self-reported error conditions

- Coverage
  - All GPS servers
  - All PTP Switches
  - All PTP Linux hosts

C) Monitoring: Agreement

- Continuous monitoring that clocks agree to each other on:
  - Delays to/from Exchanges
  - Delays on the Global internal network

- Why does it work?
  - No negative delays
  - No large delays (=> performance issue)
  - Expected delay = length of the cables

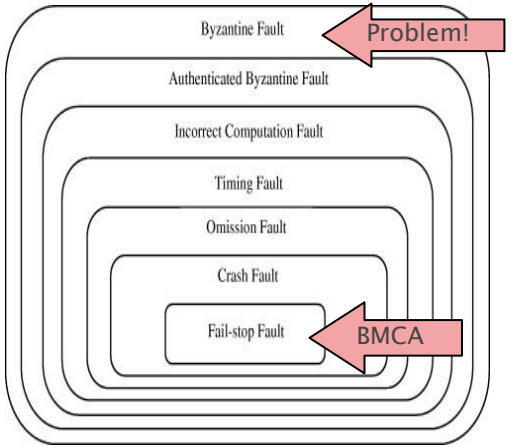
PTP Robustness

PTPv2 byzantine failures

- Deutche Börse, August 2013
  - Active GM sent bad time (Leap seconds=0).
    - Note1: this was *not* on a leap second event itself!
    - Note2: leap seconds are sent in separate messages than originTS!
  - BMCA backup GMs remained passive
  - Slaves jumped 35s => Trading halted
- IMC, July 2011
  - Same issue: Single source
- FIA, June 2015
  - Recommendations for the 2015 leap second event
- ESMA, Jan 2018
  - Regulator requires traceable 100us error

Bits								Octets	Offset
7	6	5	4	3	2	1	0		
header (see 13.3)								34	0
originTimestamp								10	34
currentUtcOffset								2	44
reserved								1	46
grandmasterPriority1								1	47
grandmasterClockQuality								4	48
grandmasterPriority2								1	52
grandmasterIdentity								8	53
stepsRemoved								2	61
timeSource								1	63

Out of Band!!



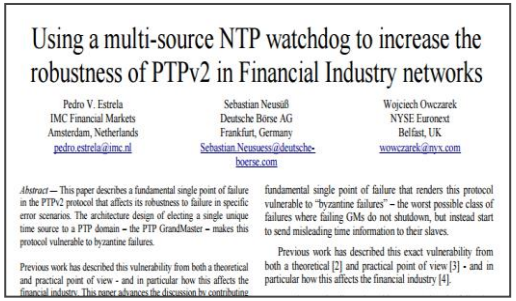
http://fia.org/sites/default/files/content_attachments/FIA%20Leap%20Second%20Exchange.pdf

PTP + NTP watchdog

https://github.com/pestrela/papers

(ISPCS 2014 paper)

- 3x NTP servers queried in parallel to PTP
- Median overrides PTP offset
  - -0.02 ms
  - +0.01 ms
  - +35000 ms
- PTP only touches the clock if allowed



Using a multi-source NTP watchdog to increase the robustness of PTPv2 in Financial Industry networks

Pedro V. Estrela
IMC Financial Markets
Amsterdam, Netherlands
pedro.estrela@imc.nl

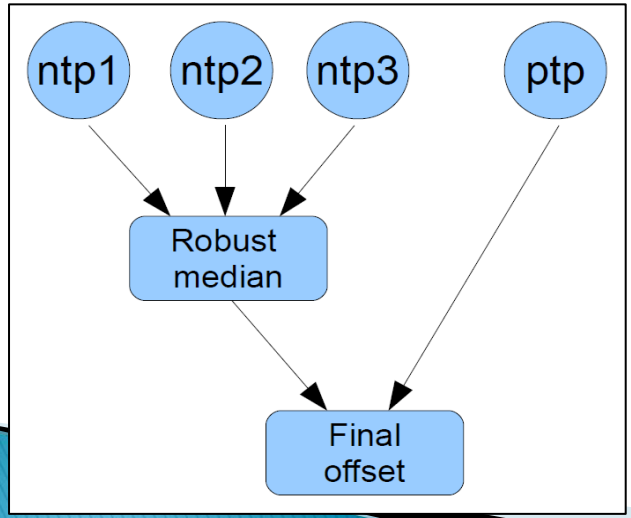
Sebastian Neusüß
Deutsche Börse AG
Frankfurt, Germany
Sebastian.Neusuess@deutsche-boerse.com

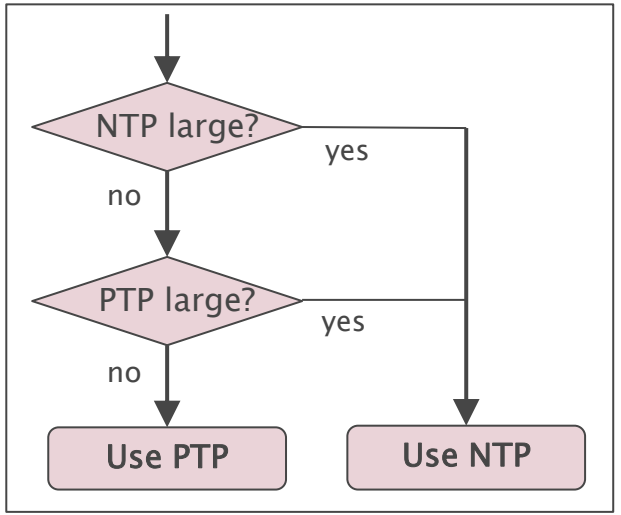
Wojciech Owczarek
NYSE Euronext
Belfast, UK
wowczarek@nyx.com

Abstract— This paper describes a fundamental single point of failure in the PTPv2 protocol that affects its robustness to failure in specific error scenarios. The architecture design of electing a single unique time source to a PTP domain – the PTP GrandMaster – makes this protocol vulnerable to byzantine failures.

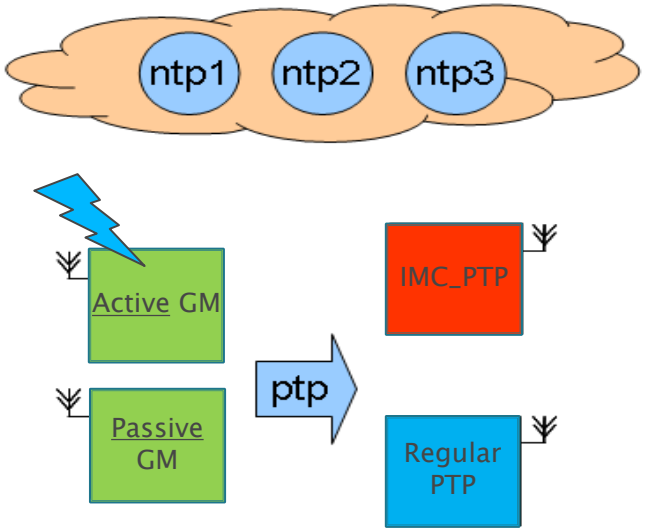
Previous work has described this vulnerability from both a theoretical and practical point of view - and in particular how this affects the financial industry. This paper advances the discussion by contributing fundamental single point of failure that renders this protocol vulnerable to "byzantine failures" – the worst possible class of failures where failing GMs do not shutdown, but instead start to send misleading time information to their slaves.

Previous work has described this exact vulnerability from both a theoretical [2] and practical point of view [3] - and in particular how this affects the financial industry [4].

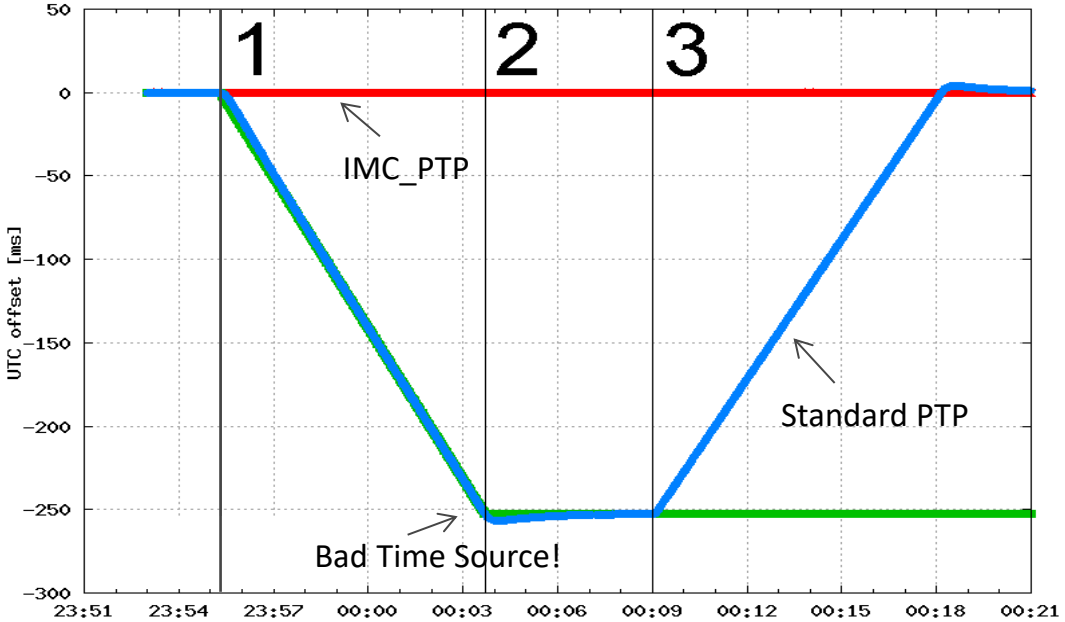




Testbed

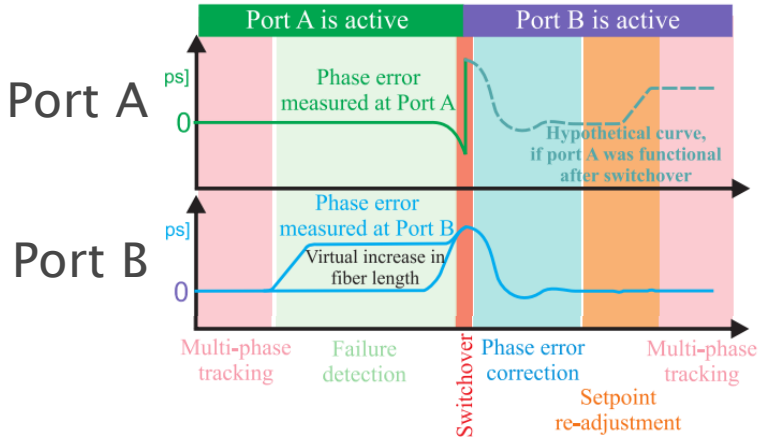


Clock error results



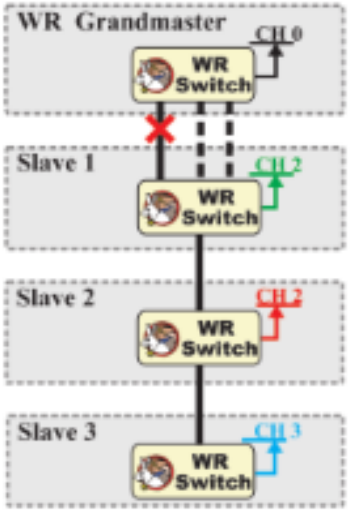
CERN robustness: detection

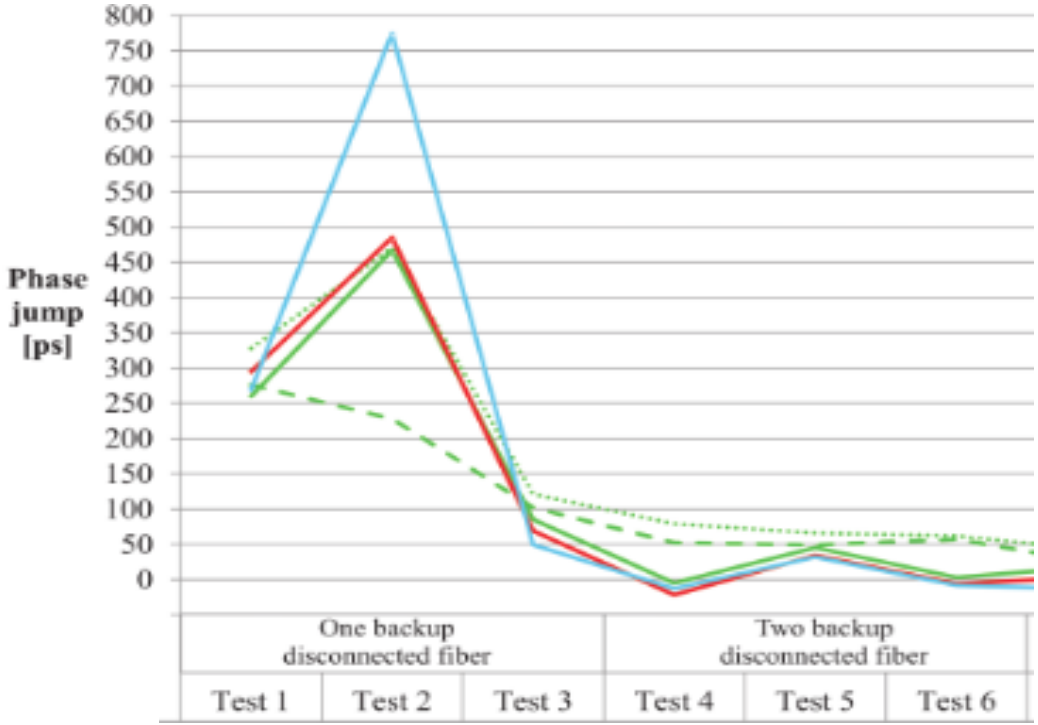
- Supports PTP backup paths
- 1 us sync maintained!
- Majority voting (phase error):
  - ShortTerm moving average;
  - LongTerm moving average;
  - Switch when above averages disagree on the majority of backup ports



Source: Maciej Lipinski PhD thesis

CERN robustness: results



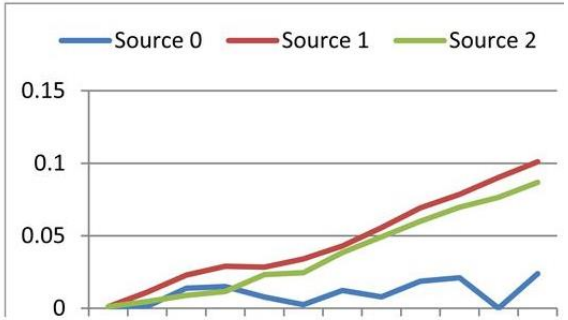


Source: Maciej Lipinski PhD thesis

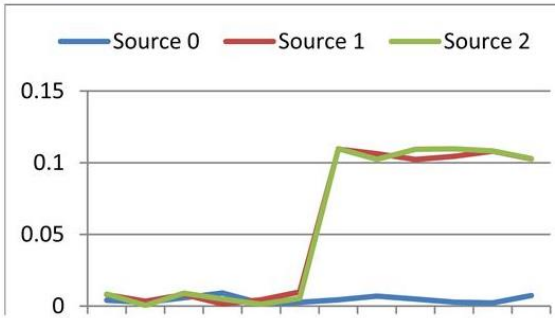
FSMtime TimeKeeper

- Both NTP and PTP backup ports
- Majority voting:
  - both offsets and frequencies
  - Default agreement threshold: 30us
- Robustness:
  - Leap seconds only accepted on “right” times
  - Also has protocol and functional checks

GPS Spoof time offset



GPS Spoof frequency



http://www.fsmtime.com/uploads/whitepapers/TimeKeeper_Sourcecheck_Time_Validation.pdf

PTP v2.1 new features

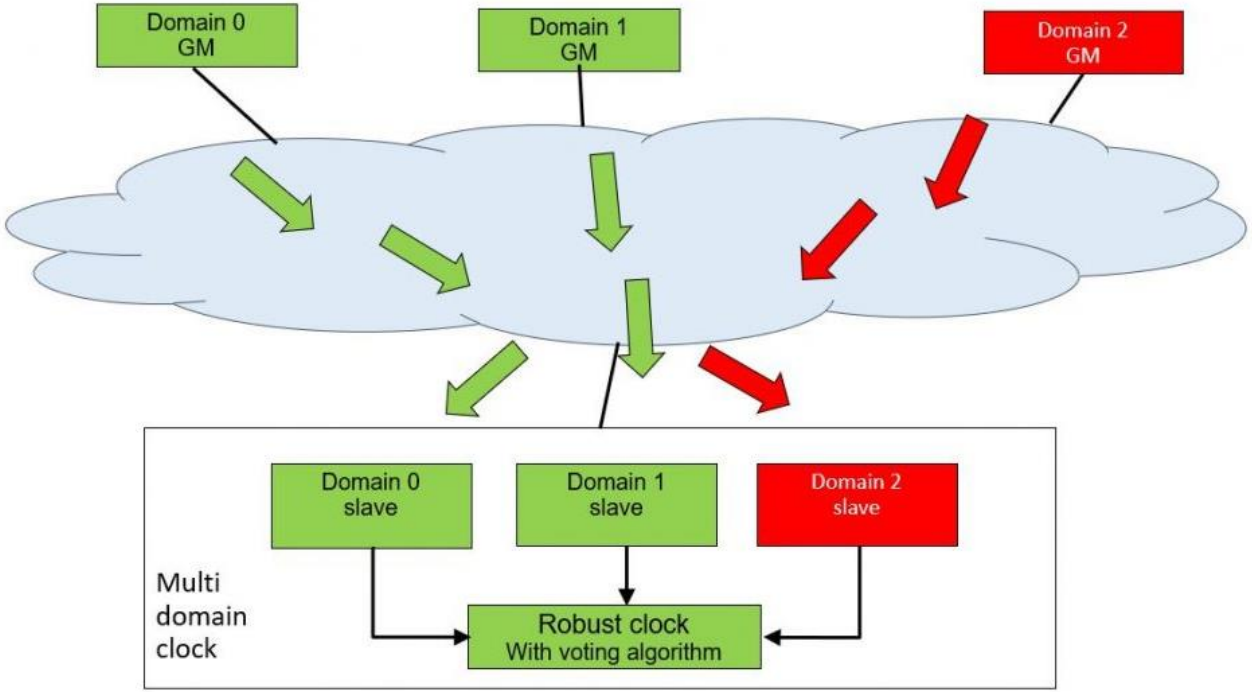
- Robustness:
  - Inter-domain interactions
  - Common Mean Link Delay Service
  - Slave port monitoring
  - Profile isolation
  - Security TLV
    - Both message and source integrity checking
  - Standard performance metrics
    - Mean/Min/max/stddev
    - Both 15 min and 24 hours
- Flexibility
  - Hybrid Multicast/Unicast mode
  - Modular Transparent clocks (incl.SFP)
  - Special PTP ports (eg WiFi)
- Accuracy (White Rabbit):
  - Manual Port Configuration (without BMCA)
  - Asymmetry calibration
  - Physical layer syntonization (eg Synchronous Ethernet)

https://blog.meinbergglobal.com/2017/09/24/whats-coming-next-edition-ieee-1588/

PTP v2.1 InterDomain robustness

Annex R:

- Multi GM PTP
- Multi Path PTP
- Combining algorithm not standardized yet
  - => Enterprise profile

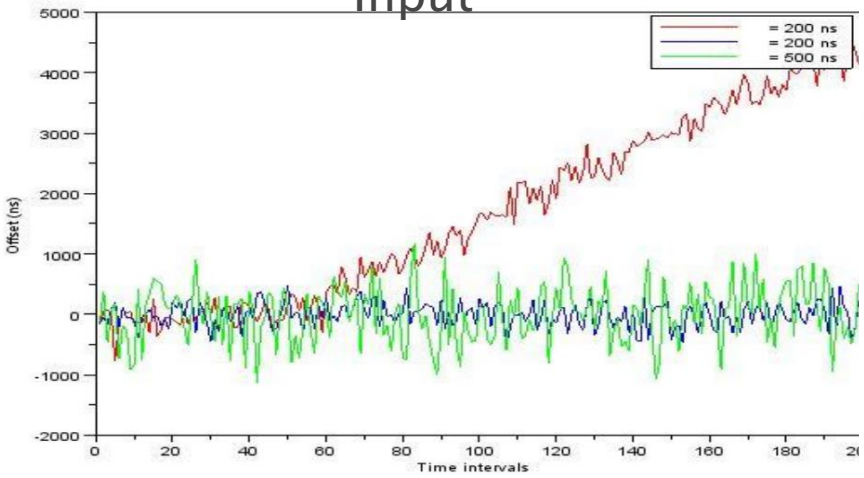


https://blog.meinbergglobal.com/2019/03/01/what-to-expect-in-the-ieee-1588-revision-interdomain-interactions/
http://www.telecom-sync.com/files/pdfs/itsf/2016/day2/0940_Enterprise%20Profile.pdf page 10

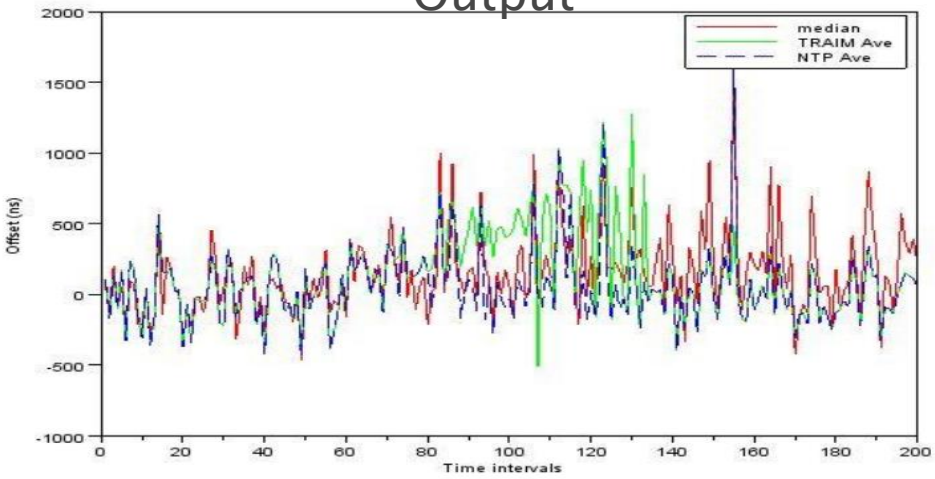
Enterprise profile robustness

https://datatracker.ietf.org/doc/draft-ietf-tictoc-ptp-enterprise-profile/?include_text=1
"Clocks SHOULD include support for multiple domains. Redundant sources of timing can be ensembled, and/or compared to check for faulty Master Clocks."

Input



Output



http://www.telecom-sync.com/files/pdfs/itsf/2016/day2/0940_Enterprise%20Profile.pdf , page 12

TickTock (Huygens)

Focus: userspace scalable synchronization + NIC timestamps

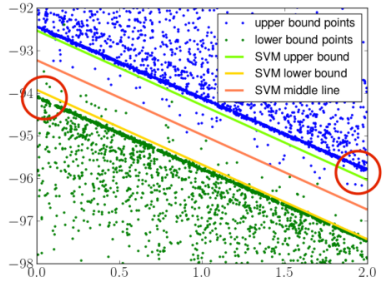
1) Support Vector Machines (ML technique)
2) Coded Probes (rejects impure probes)
3) Network Effect (clocks transitively synced)

Robustness:

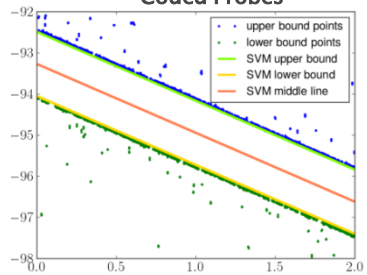
- Support multiple coordinators and Zookeeper
- Support multiple ref clocks per site

https://www.usenix.org/system/files/conference/nsdi18/nsdi18-geng.pdf

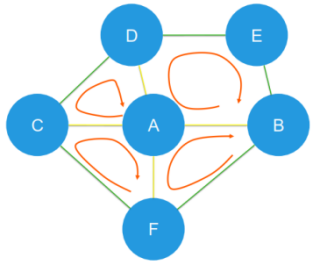
Plain SVM



Coded Probes

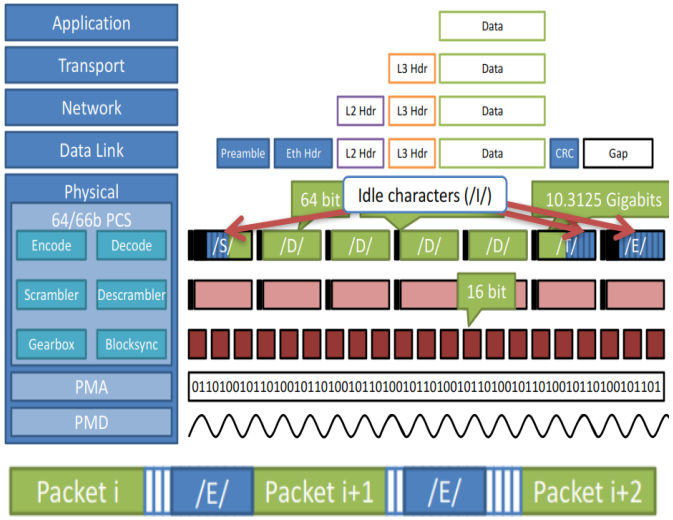


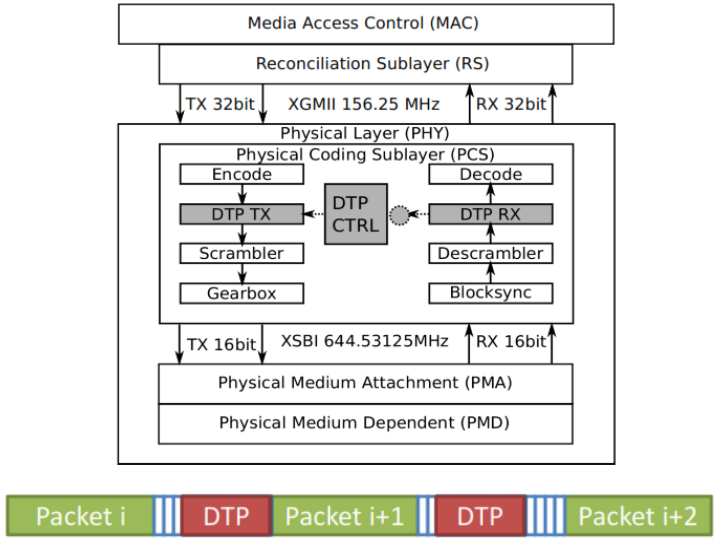
Network Effect



DataCenter Protocol (DTP)

Replace 802.3 10Gb *idle* symbols with transparent time information (taken from IPG or idle frames)





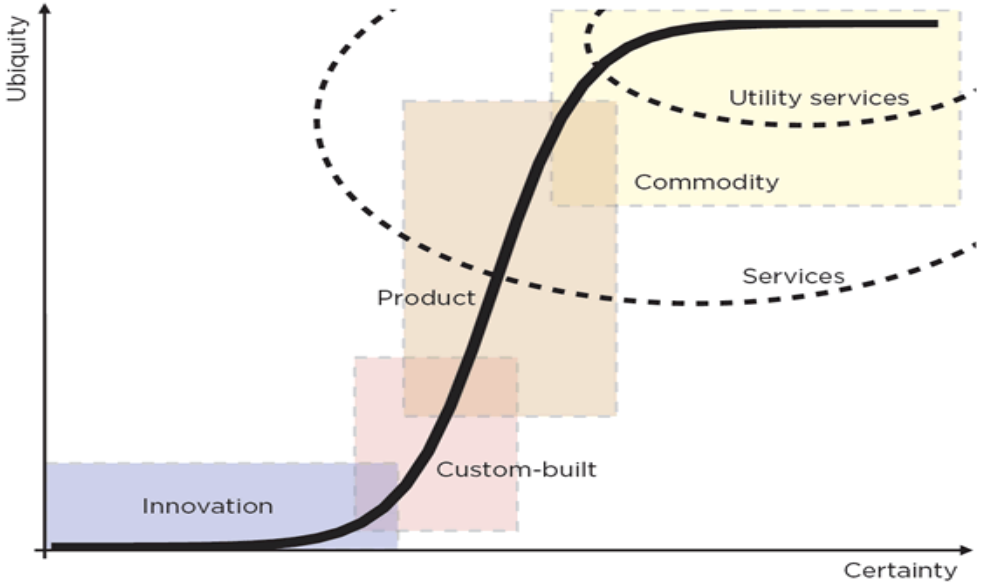
Rate	min IPG
1G	8
10G	5
40G	1
100G	1

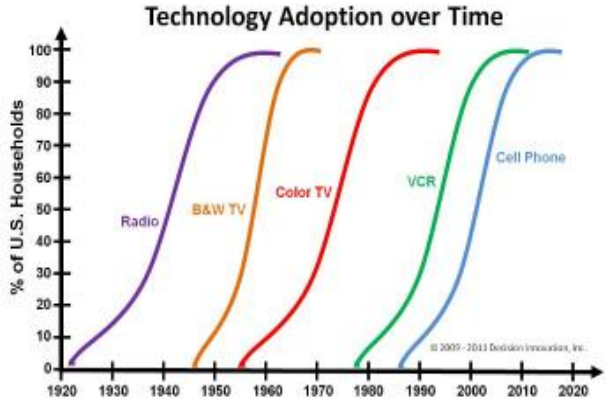
https://www.cs.cornell.edu/~vishal/papers/dtp_2019.pdf

Data Rate	Encoding	Data Width	Frequency	Period	Δ
1G	8b/10b	8 bit	125 MHz	8 ns	25
10G	64b/66b	32 bit	156.25 MHz	6.4 ns	20
40G	64b/66b	64 bit	625 MHz	1.6 ns	5
100G	64b/66b	64 bit	1562.5 MHz	0.64 ns	2

Comments to other industries

Innovation / Good enough





- Success = "Do it Better / Do it cheaper"
- Cheap dominates the end-game

Full slides: https://github.com/pestrela/papers/tree/master/4_Innovation
Recommended Book: https://en.wikipedia.org/wiki/The_Innovator%27s_Dilemma

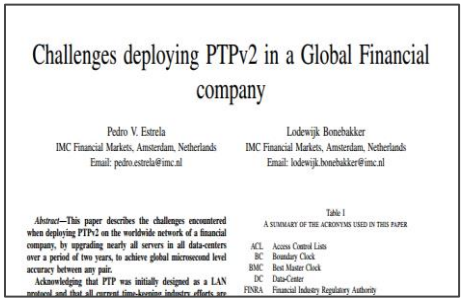
Other industries

- How to replace NTP in large companies?
  - Focus: UTC to software clients
  - Simplicity = Cheap
  - Smooth migration essential. No network changes
  - WANs supported out of the box via unicast
    - Multicast = problems

- IETF Enterprise profile

https://github.com/pestrela/papers

(ISPCS 2012 paper)



Challenges deploying PTPv2 in a Global Financial company

Pedro V. Estrela
IMC Financial Markets, Amsterdam, Netherlands
Email: pedro.estrela@imc.nl

Lodewijk Bonebakker
IMC Financial Markets, Amsterdam, Netherlands
Email: lodewijk.bonebakker@imc.nl

Abstract—This paper describes the challenges encountered when deploying PTPv2 on the worldwide network of a financial company, by upgrading nearly all servers in all data-centers over a period of two years, to achieve global microsecond level accuracy between any pair.

Acknowledging that PTP was initially designed as a LAN

Table I
A SUMMARY OF THE ACRONYMS USED IN THIS PAPER

ACL	Access Control Lists
BC	Boundary Clock
BMC	Best Master Clock
DC	Data-Center
FINRA	Financial Industry Regulatory Authority

IMC public contributions

- 2012: **First paper** on the main PTP Scientific conference. Paper describes multiple issues deploying of PTP worldwide
  - http://tagus.inesc-id.pt/~pestrela/ptp/Challenges_deploying_PTPv2_in_a_Global_Financial_company.pdf

- 2014: **Best paper award** on the main PTP Scientific conference, with Deutsche Börse and ICE/NYSE. Paper describes a solution for the PTP robustness problem.
  - http://tagus.inesc-id.pt/~pestrela/ptp/Pedro_Estrela_-_ISPCS_2014_best_paper_-_Increasing_PTPv2_robustness_-_presentation.pdf

- 2014: Contributed to the FIA/FIA Europe official comments to **ESMA RTS-25**
  - https://epta.fia.org/sites/default/files/content_attachments/ESMA_MiFID2_CP_FIA%20ASSOCIATIONS_REPLYFORM.pdf

- 2015: Contributed to the FIA recommendation on the **2015 Leap Second**
  - https://fia.org/sites/default/files/content_attachments/FIA%20Leap%20Second%20Exchange.pdf

Extra Slides

https://github.com/pestrela/papers