PTPv2 clock synchronization for the financial sector

Pedro V. Estrela, PhD
Sr. Performance Engineer
13-December-2017

- Part #1: Financial markets overview
  - How electronic markets work
  - A brief history of low-latency trading
  - Network Monitoring

- Part #2: Clock synchronization
  - PTPv2 basics
  - State-of-the-art robustness issues
  - MIFID II regulation (RTS-25)

About the presenter

- Pedro V. Estrela
  - PhD in Computer Science (2007 TU-Lisbon)
  - Found Financial industry by luck ☺

- Performance Engineer
  - “Mechanic” of driver-less Formula 1
  - Measure + Remove latency bottlenecks

About IMC

- Think of a currency house, but for:
  - Options / Futures / Stocks / Bonds / ETFs / FX
- Some numbers about IMC
  - All major worldwide Markets, All Timezones, 4 offices, ~500px
  - ~60 datacenters, ~200 links, 10000s equipments
  - ~2000 SW deployments
- Teams' responsibilities
  - Trading team = Find the Price
  - Technology team = Adjust orders Quickly

- Competitive landscape

- Relative latency
  - Total1 = A + **B** + C + D + E
  - Total2 = A + B + C + D + E
  - Trend is clearly: Faster / Raw Hardware / More expensive

Financial Markets overview

WHAT WE DO IN LESS THAN 2 MIN.

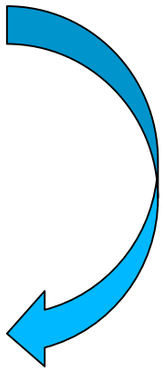
http://www.imc.com/eu/about-us#what-we-do
http://www.economicprinciples.org/

Exchange Price-time priority

- Buyers and Sellers meet at a regulated exchange
- Express their intention to buy / sell
- Orders continuously matched first by price, then time

Imagine this just happened…

London	9.98	9.99	10.00	10.01	10.02	10.03
Frankfurt	9.98	9.99	10.00	10.01	10.02	10.03



London	**9.98**	**9.99**	**10.00**	**10.01**	**10.02**	**10.03**
Frankfurt	**9.96**	**9.97**	**9.98**	**9.99**	**10.00**	**10.01**

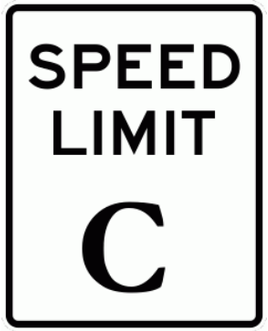
Questions

- Q1: do you see a trading opportunity here?
- Q2: what should the market maker do here?

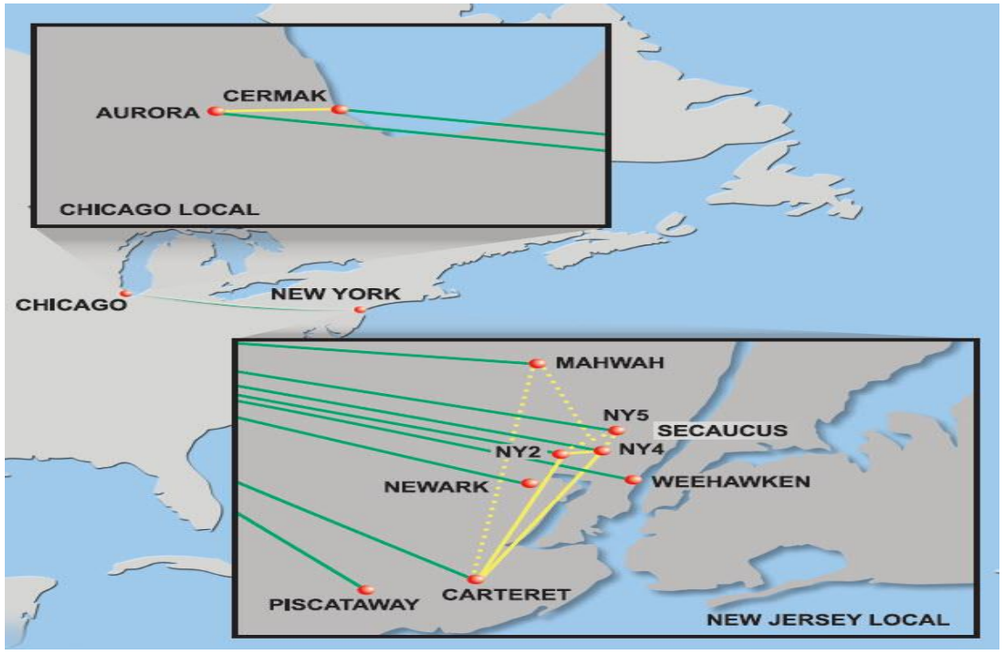
Low-Latency

How long is…

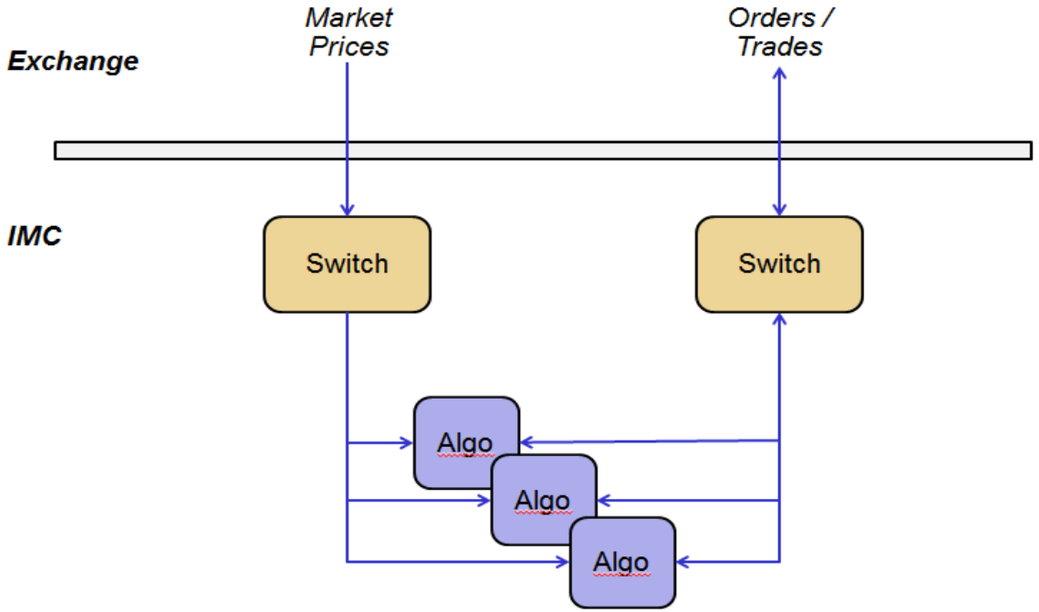
- millisecond (ms)
  - A camera flash illuminates for 1 millisecond
  - Distance between countries
- microsecond (µs)
  - 3 microseconds – Light to travel one Kilometer (1 billion km/h)
  - In and Out a machine, including all processing
- nanosecond (ns)
  - 1 nanosecond – Light to travel 30cm
  - 350ns packet forward in a switch



Wide Area



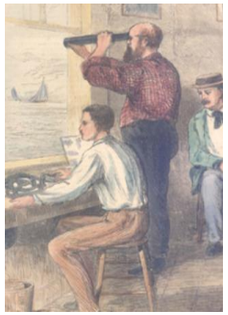
Local Area



Wide area: the story so far

Historical fact: Every new wide area technology was first used for trading

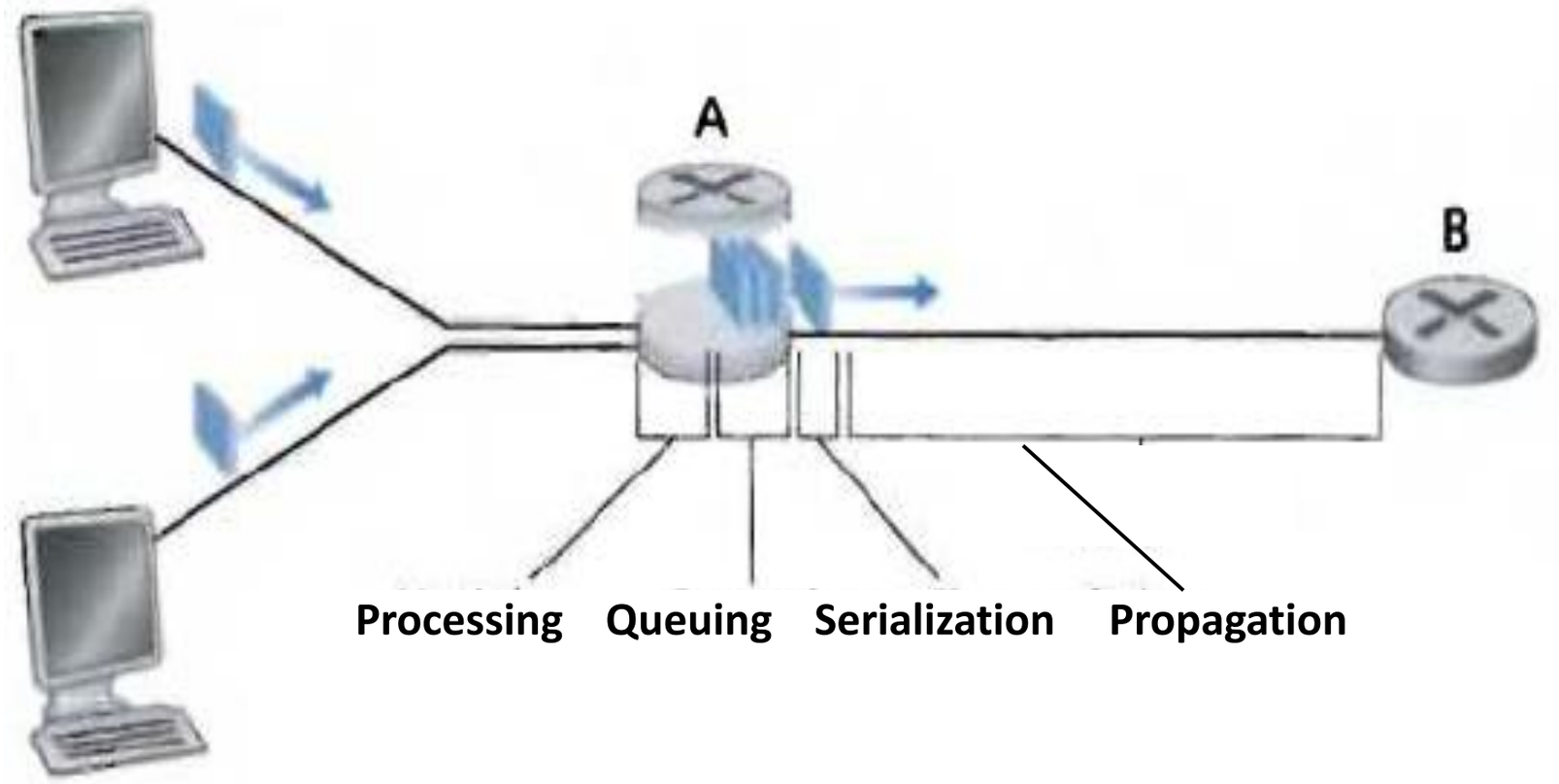
Year	Technology	Who	
1815	**Pigeons**	*Baron Rothchild knew that Napoleon lost the war*	Days/Hours
1836	**Telescopes**	*Shore agents check if coffee was spoilt on Boats*	
1897	**Telegraph**	*Bookies send Horse race results to outside*	
2010	**Fiber**	*Spread networks drills mountains on NY-Chicago*	Milliseconds
2012	**Microwave**	*McKay jumps the same mountains using Radio*	
2015	**Fiber**	*Hibernia builds new straighter Atlantic cable*	

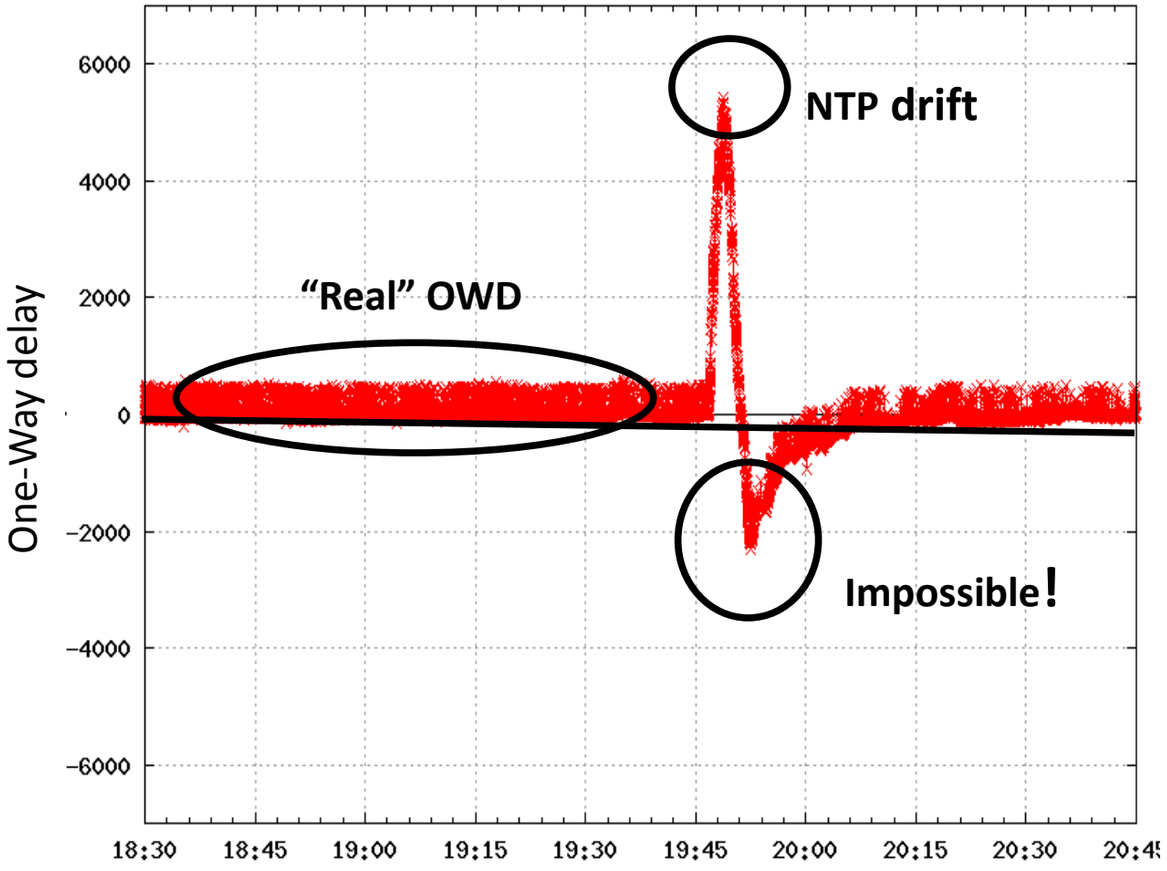
Source: http://www.forbes.com/forbes/2010/0927/outfront-netscape-jim-barksdale-daniel-spivey-wall-street-speed-war.html
Source: https://www.moaf.org/publications-collections/financial-history-magazine/111/_res/id=Attachments/index=0/Plundered_by_Harpies.pdf

Network Monitoring

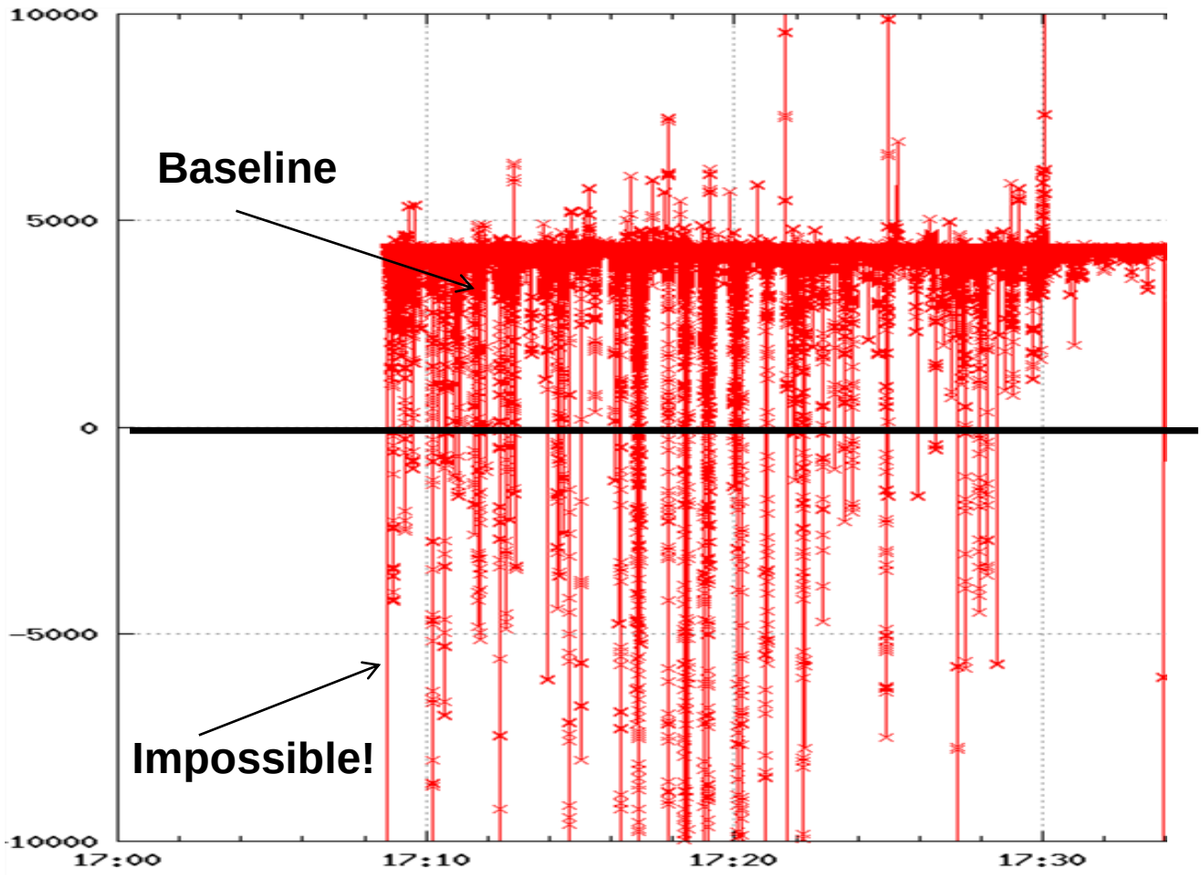
Latency components



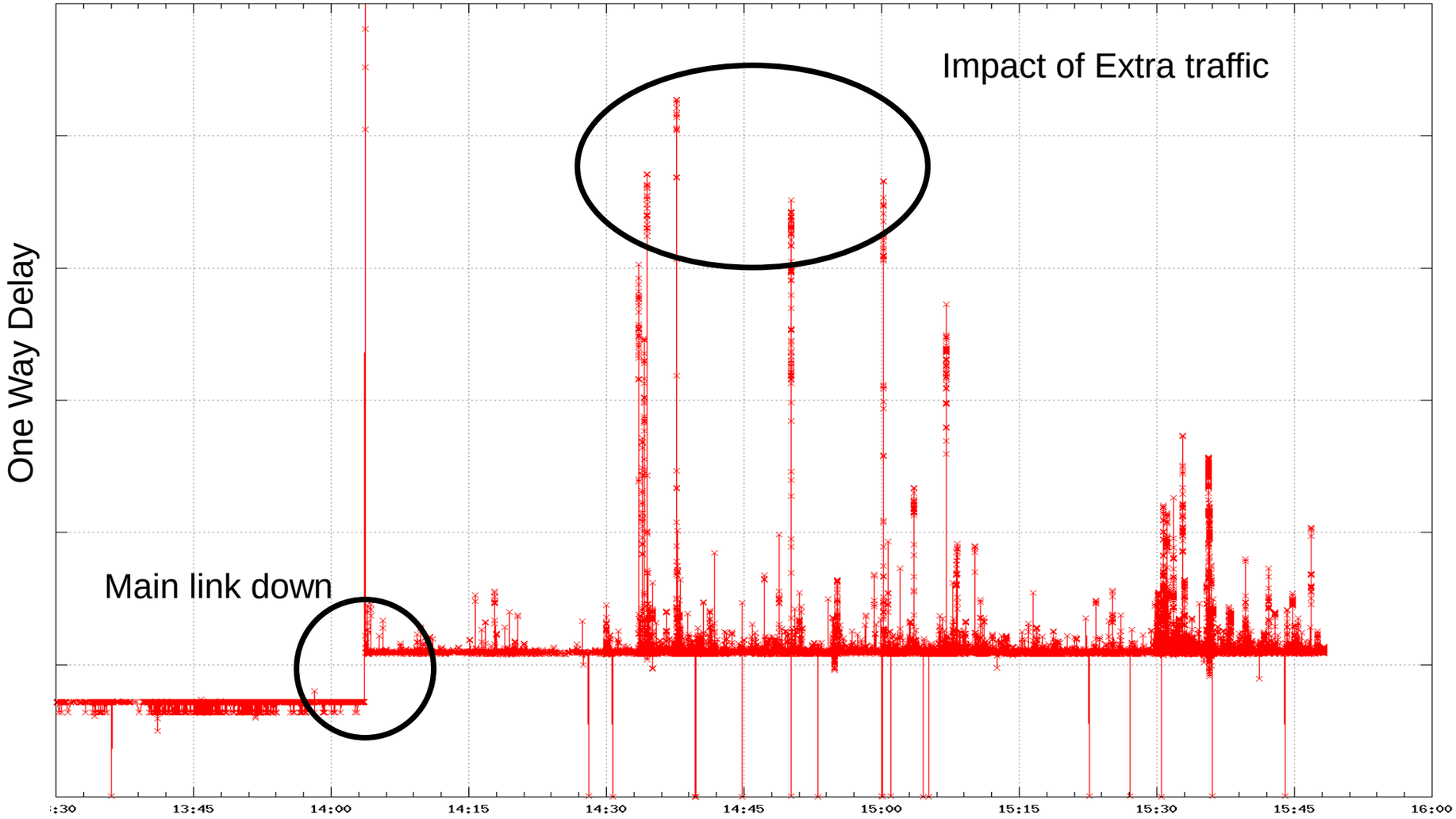
Bad Clock Sync



Bad Timestamps



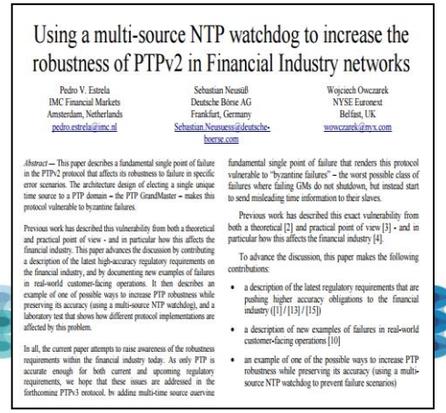
Latency example 2



Part 2: Increasing robustness of PTPv2 Financial networks

25 September
ISPCS 2014
Austin, Texas

(best paper award)



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 a description of the latest high-accuracy regulatory requirements on the financial industry, and by documenting new examples of failures in real-world customer-facing operations. It then describes an example of one of possible ways to increase PTP robustness while preserving its accuracy (using a multi-source NTP watchdog), and a laboratory test that shows how different protocol implementations are affected by this problem.

In all, the current paper attempts to raise awareness of the robustness requirements within the financial industry today. As only PTP is accurate enough for both current and upcoming regulatory requirements, we hope that these issues are addressed in the forthcoming PTPv3 protocol. Its addine multi-time source overvine 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.

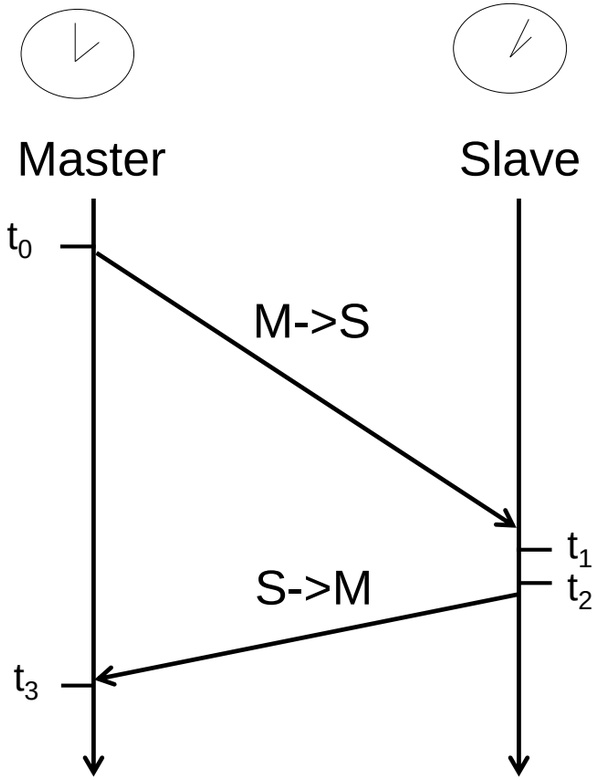
Previous work has described this exact vulnerability from both a theoretical [2] and practical point of view [3] - and in particular how this affects the financial industry [4].

To advance the discussion, this paper makes the following contributions:

- a description of the latest regulatory requirements that are pushing higher accuracy obligations to the financial industry ([1]/[13]/[15])
- a description of new examples of failures in real-world customer-facing operations [10]
- an example of one of the possible ways to increase PTP robustness while preserving its accuracy (using a multi-source NTP watchdog to prevent failure scenarios)

Fundamental challenge

- 3 variables:
  - Clock difference θ
  - Forward delay (δ’)
  - Return delay (δ”)
- 2 equations:
  - Symmetric paths required
  - HW timestamps to remove queuing
  - Dedicated Network for only time distribution



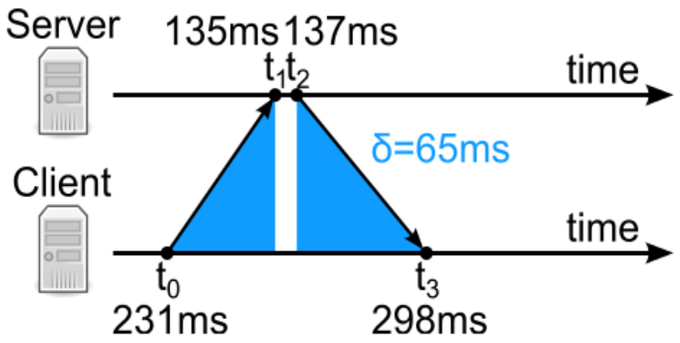
$$\theta = \frac{(t_1 - t_0) + (t_2 - t_3)}{2}$$

$$\delta = (t_3 - t_0) - (t_2 - t_1)$$

Packet-based solutions

Network Time Protocol (NTPv4)

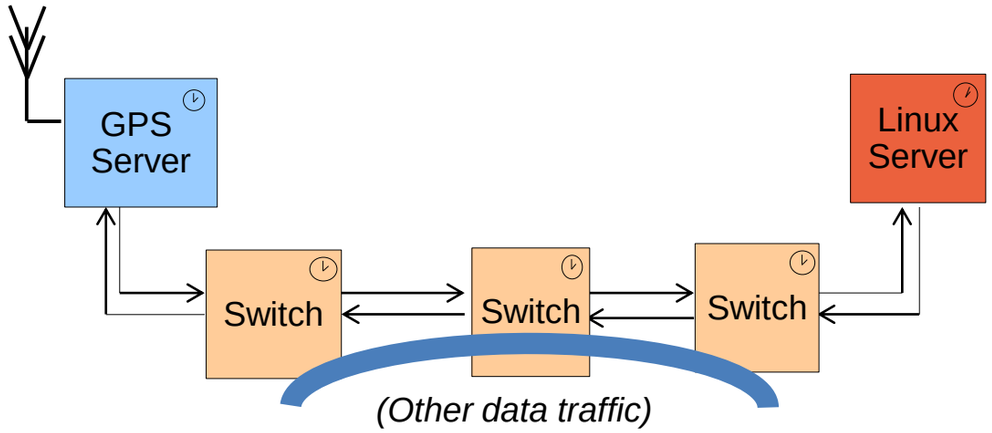
- Mature IETF standard
- Milli-seconds accuracy
- Multiple time sources



https://en.wikipedia.org/wiki/Network_Time_Protocol

Precision Time Protocol (PTPv2)

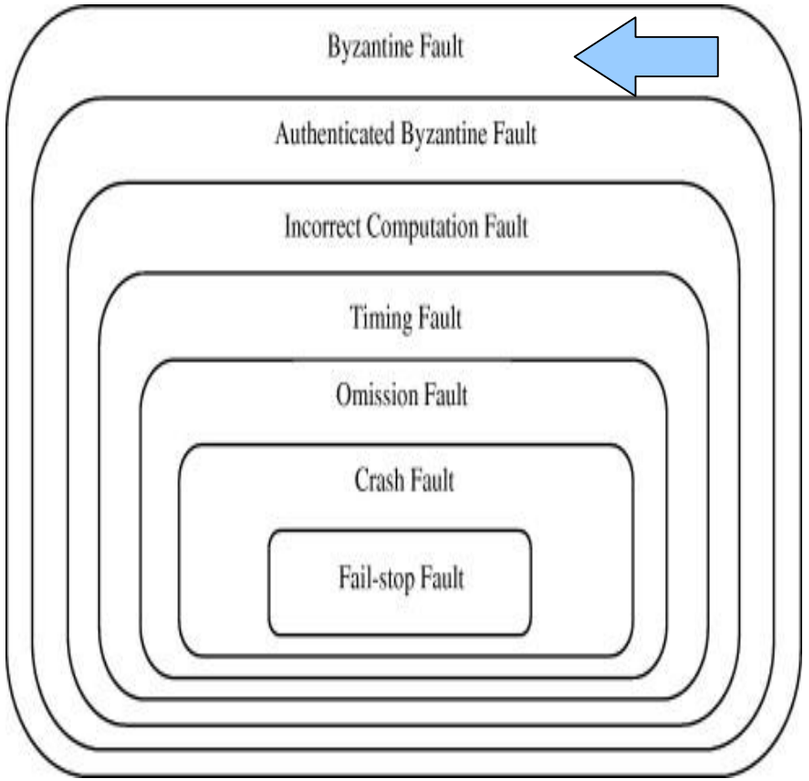
- Recent IEEE standard
- Micro-seconds accuracy
- Single time source



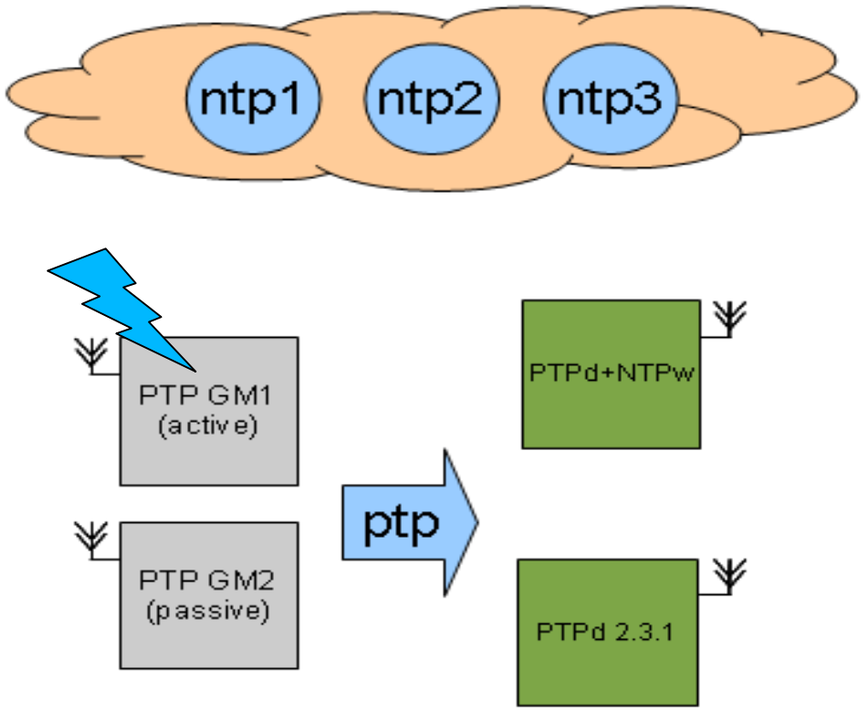
https://en.wikipedia.org/wiki/Precision_Time_Protocol

PTPv2 byzantine failures

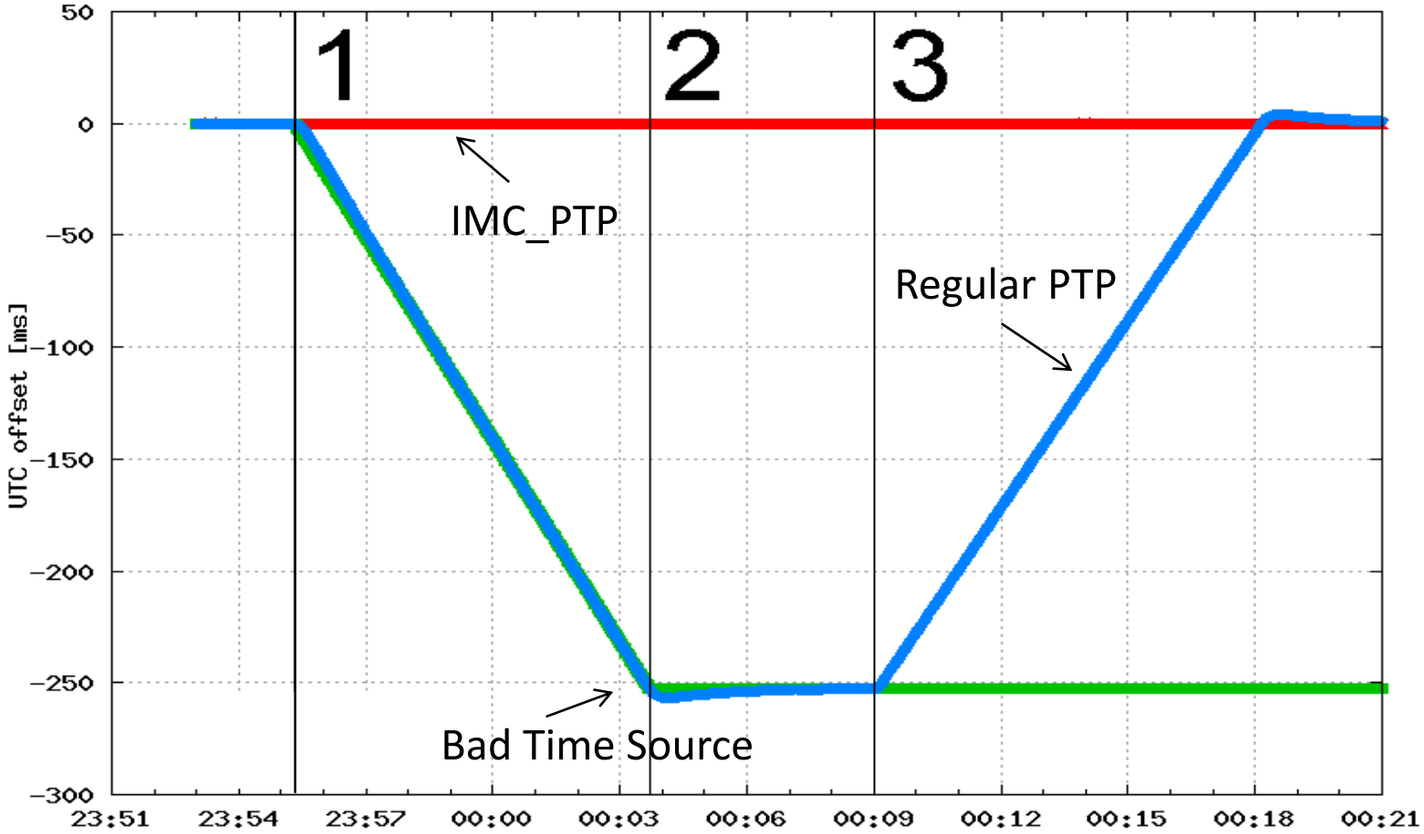
- Eurex, August 2013
  - Active GM sent *bad* time (leap seconds = 0)
  - Backup GMs remain passive
  - Slaves jumped by 35 seconds
  - Trading halted => all customers affected
- IMC, July 2011
  - Same problem as above: Single source



Testbed

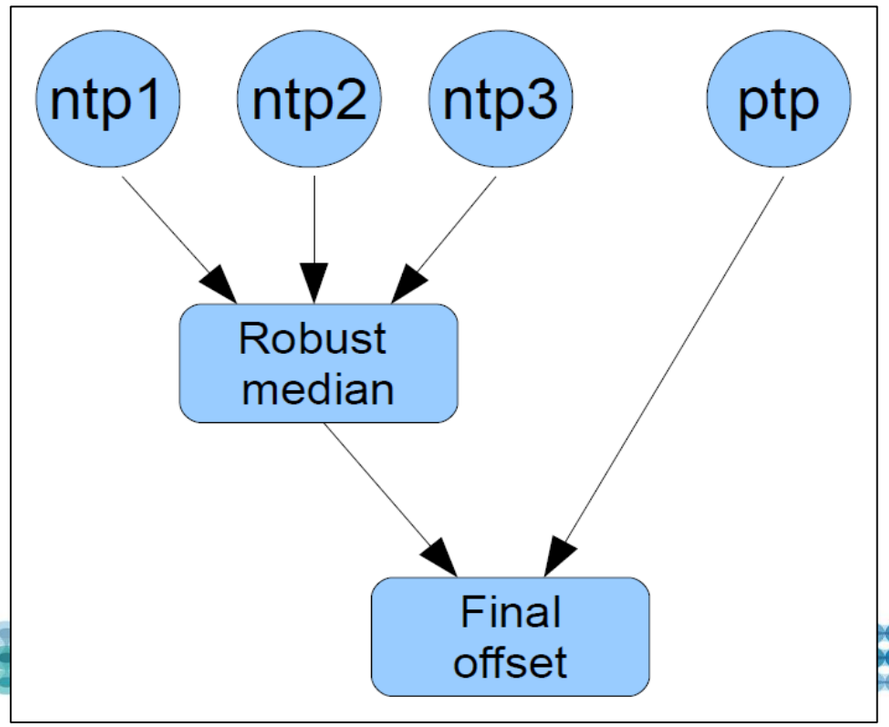


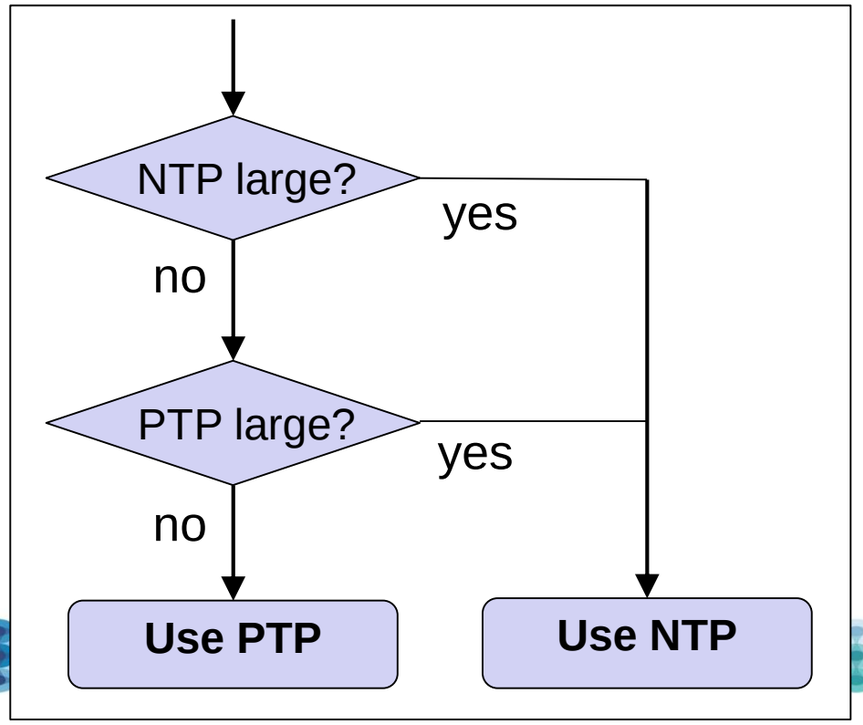
Experiment



Solution, using with NTP watchdog

- 3x NTP servers queried in parallel to PTP
- Robust median offset can override PTP offset:
  - -0.02 ms
  - +0.01 ms
  - +35000 ms
- PTP only touches the clock if allowed by the NTP watchdog





MIFID II RTS-25

http://mifid2017b.executiveindustryevents.com/Event/programme

MIFID II RTS 25 requirements:

- Monitor for <100us accuracy
- Document whole UTC traceability chain
- Identify the precise timestamping point

PTP Deployment - Best practices:

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

Monitoring #1: Self-Health

- Continuous monitoring of:
  - Self-reported clock offsets
  - Self-reported error conditions
- Coverage
  - All GPS servers
  - All PTP Switches
  - All PTP Linux hosts

Monitoring #2: Agreement

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

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

IMC public contributions

- 2012: **Paper** on the main PTP Scientific conference.
  Paper describes multiple issues deploying of PTP worldwide
  - *https://www.researchgate.net/project/PTP-Clock-Synchronization*

- 2014: **Best paper award** on the main PTP Scientific conference, with Deutsche Borse and ICE/NYSE. Paper describes a solution for the PTP robustness problem
  - *https://www.researchgate.net/project/PTP-Clock-Synchronization*

- 2014: Contributed to the FIA EPTA/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

Conclusion

- IMC opportunities
  - Information Technology
  - Quantitative Trading
  - Both Internships, and Full time opportunities

- More questions?
  - IMC: *https://www.imc.com/eu/careers/why-imc*
  - Scientific papers: *https://www.researchgate.net/project/PTP-Clock-Synchronization*

Extra Slides

RTS 25

Rule:

http://ec.europa.eu/finance/docs/level-2-measures/mifid-rts-25-annex_en.pdf

Maximum divergence from UTC: 100 microseconds

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

Leap seconds = Problems

- Heraldsun:
  - “Leap second crashes Qantas and leaves passengers stranded”
- Cnet:
  - “Leap second bug causes site software crashes”
- Globalpost:
  - “Weird Wide Web - Leap second causes flight delays and internet problems”
- Buzzfeed:
  - “How a second brought down half the Internet”
- Wired:
  - “Leap second glitch explained”