The future is PTP

(but it has a Single point of Failure)

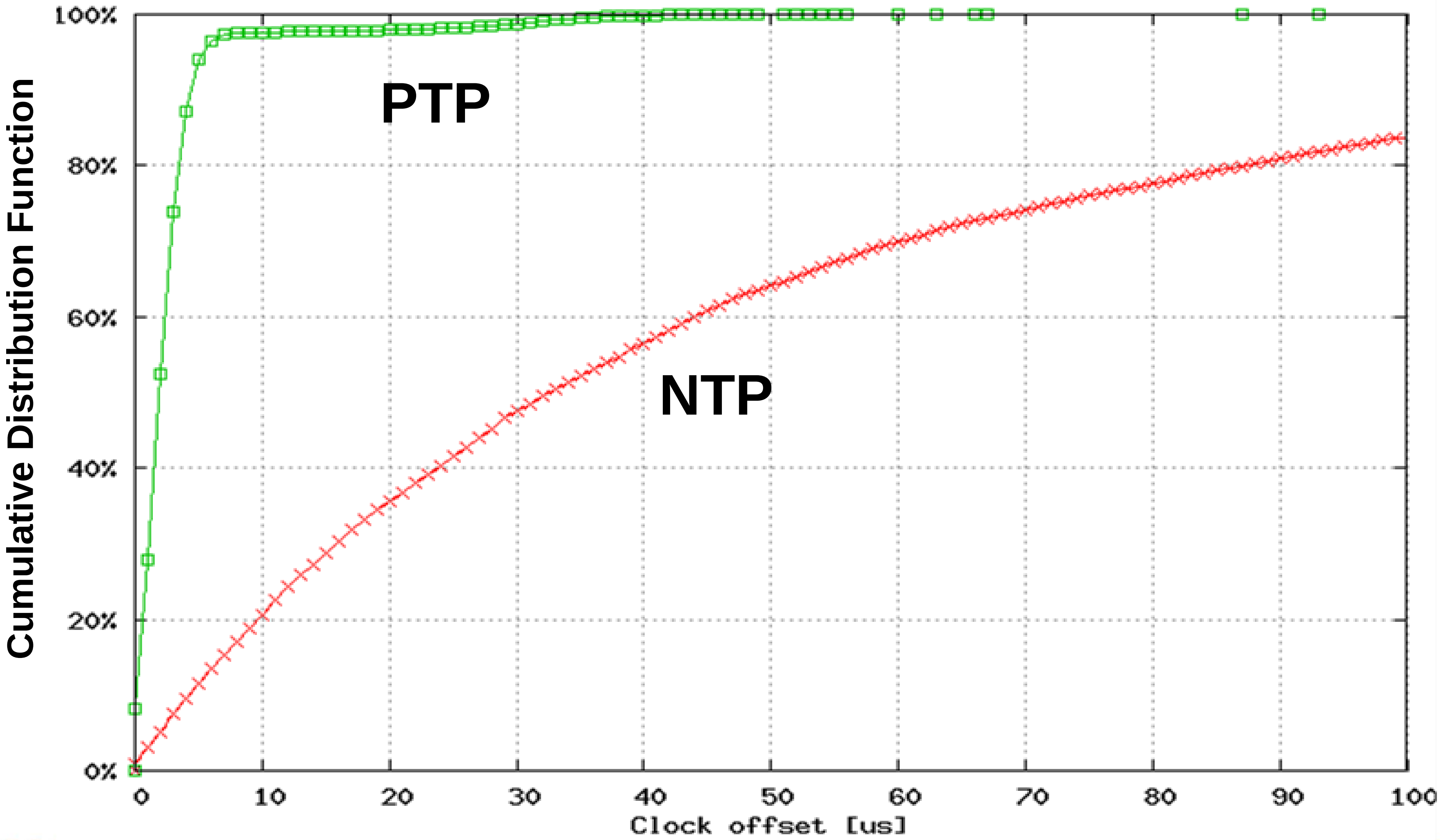
(ISPCS best paper award)

Pedro V. Estrela, PhD
Performance Engineer
26-May-2016

Why PTP network distribution?

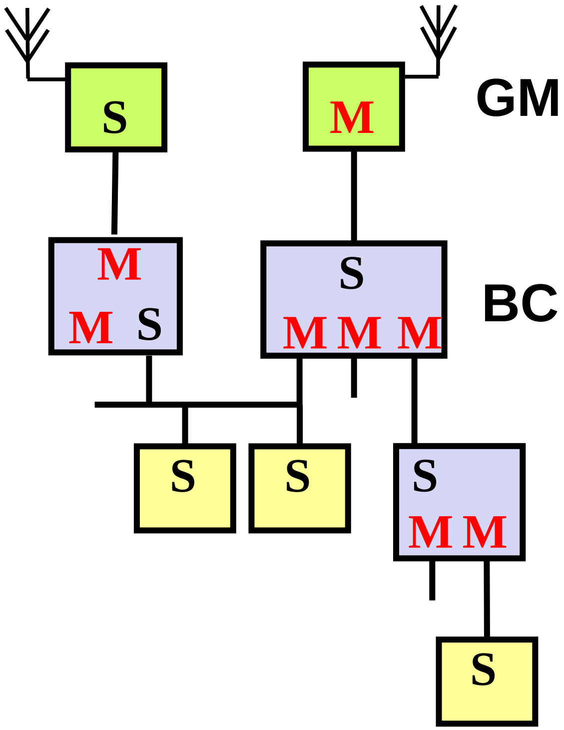
- Cost-effective:
  - Uses your existing IP network to distribute GPS time to thousands of machines
- Very Accurate:
  - Switches and NICs now have PTP-support and HW timestamps to create a symmetric path
- Industry-backed:
  - Only solution with active standardization and research work from the whole Time industry

NTP->PTP Upgrade



PTPv2 failures in Financial networks

- 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 time source

3 independent sources = 1 failure absorbed

- Configuration errors
- Implementation Bugs
- Leap seconds issues

- Human errors
- Multicast / Routing errors
- GPS Jamming

1996 - Mathematical proof
(Fetzer, Christian)

Integrating External and Internal Clock Synchronization

Christof Fetzer and Flaviu Cristian
Department of Computer Science & Engineering
University of California, San Diego
La Jolla, CA 92093–0114*
e-mail: {cfetzer, flaviu}@cs.ucsd.edu
http://www-cse.ucsd.edu/users/{cfetzer, flaviu}

June 4, 1996

Abstract

We address the problem of how to integrate fault-tolerant external and internal clock synchronization. In this paper we propose a new *external/internal* clock synchronization algorithm which provides both external and internal clock synchronization for as long as a majority of the reference time servers (servers with access to reference time) stay correct. When half or more of the reference time servers are faulty, the algorithm degrades to a fault-tolerant internal clock synchronization algorithm. We prove that at least $2F+1$ reference time servers are necessary for achieving external clock synchronization when up to F reference time servers can suffer arbitrary failures, thus the proposed algorithm provides maximum fault-tolerance. In this paper we also derive lower bounds for the best maxi-

is a granular representation of real-time and is typically provided by a standard source of time such as NIST. Clocks can be externally or internally synchronized [1]. A clock is *externally* synchronized if at any point in real-time the distance between its value and reference time is bounded by an a priori given constant called *maximum external deviation*. A set of clocks is *internally* synchronized if at any point in real-time the distance between the values of two correct clocks in the set is bounded by an a priori given constant called the *maximum internal deviation* and each clock runs within a linear envelope of real time. Externally synchronized clocks are always internally synchronized by two times the maximum external deviation, but internally synchronized clocks are not always externally synchronized.

The systems we consider in this paper consist of a

2012 - Failure description
(Estrela, Bonebakker)

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 protocol and that all current time-keeping industry efforts are focused on PTP, the issues can be broadly divided into a) issues on the PTPv2 standard itself, b) issues that have to be addressed when PTP is expanded to work over WANs, and c) issues that caused the biggest operational impact on the (tested) implementations.

In all, this paper contributes concrete examples where PTP's byzantine robustness, scalability and efficiency characteristics range between absent to poor – and attempts to raise awareness on the steps needed to build PTP solutions with the characteristics that global users want.

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
GM	GrandMaster
IGMP	Internet Group Management Protocol
LAN	Local Area Network
MAN	Metropolitan Area Network
NE	Network Equipment
NIC	Network interface controller
NTP	Network Time Protocol
PIM-SM	Protocol Independent Multicast - Sparse Mode
RP	Rendezvous Point
TTL	Time To Live
UTC	Universal Coordinated Time

I. Introduction

This paper describes the challenges encountered when deploying PTPv2 on the worldwide network of a financial company, in order to achieve microsecond level accuracy between any two servers (globally). For this, we will describe the issues discovered over the last two years, while deploying

Taking these considerations into account, this paper divides the encountered issues into a) those that affect PTPv2 as it is defined today (i.e., for LANs), b) the issues that have to be addressed when PTP is expanded to work over WANs and c) the issues that caused the biggest operational impact on the (tested) implementations. In all, this paper attempts to raise

2014 - Proof + Solution
(Estrela, Neusuess, Owczarek)

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

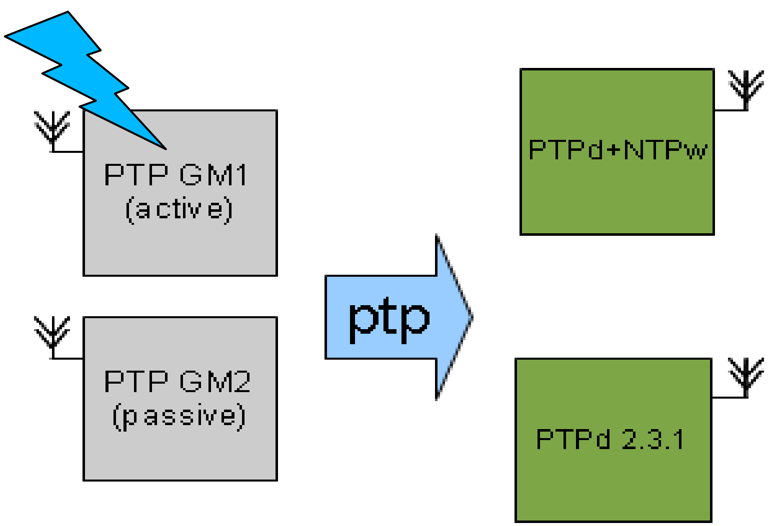
fundamental single point of failure that renders this protocol vulnerable to "byzantine failures" – the worst possible class of failures where failing GMs do not shutdown, but instead start to send misleading time information to their slaves.

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].

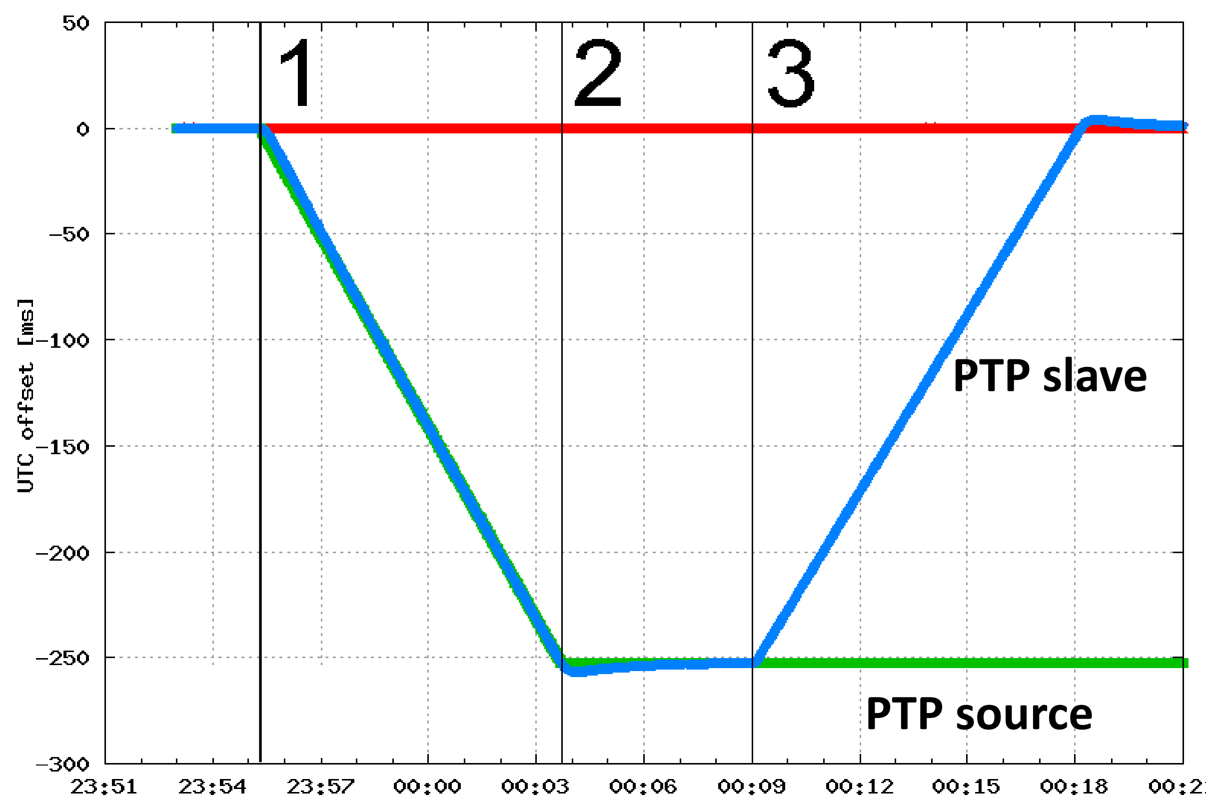
To advance the discussion, this paper makes the following contributions:

- 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)

Testbed

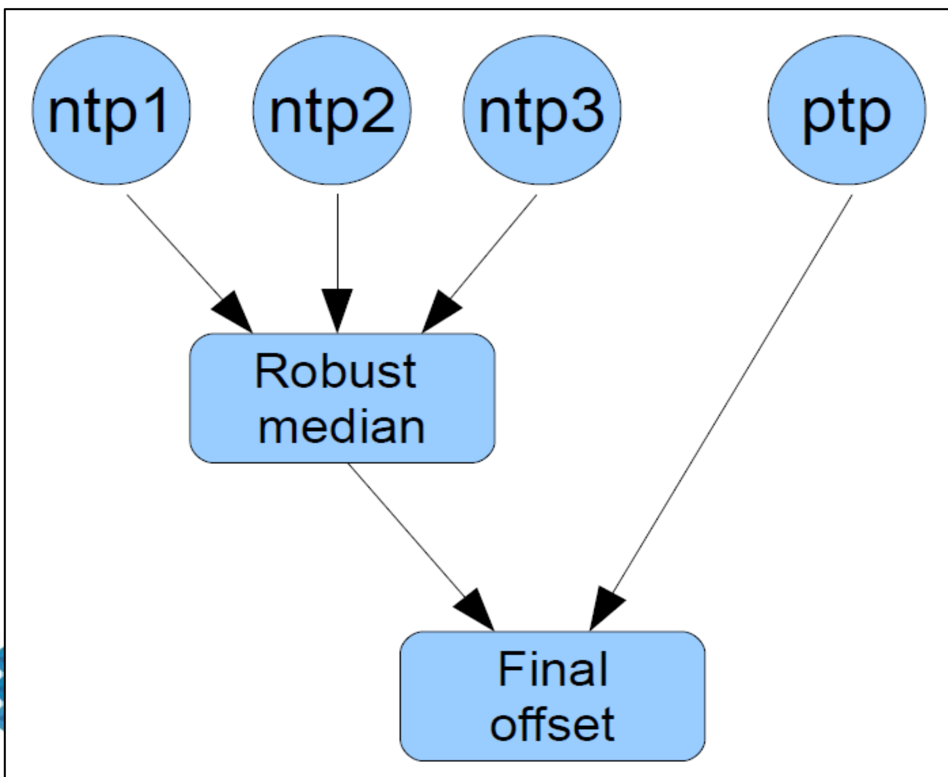


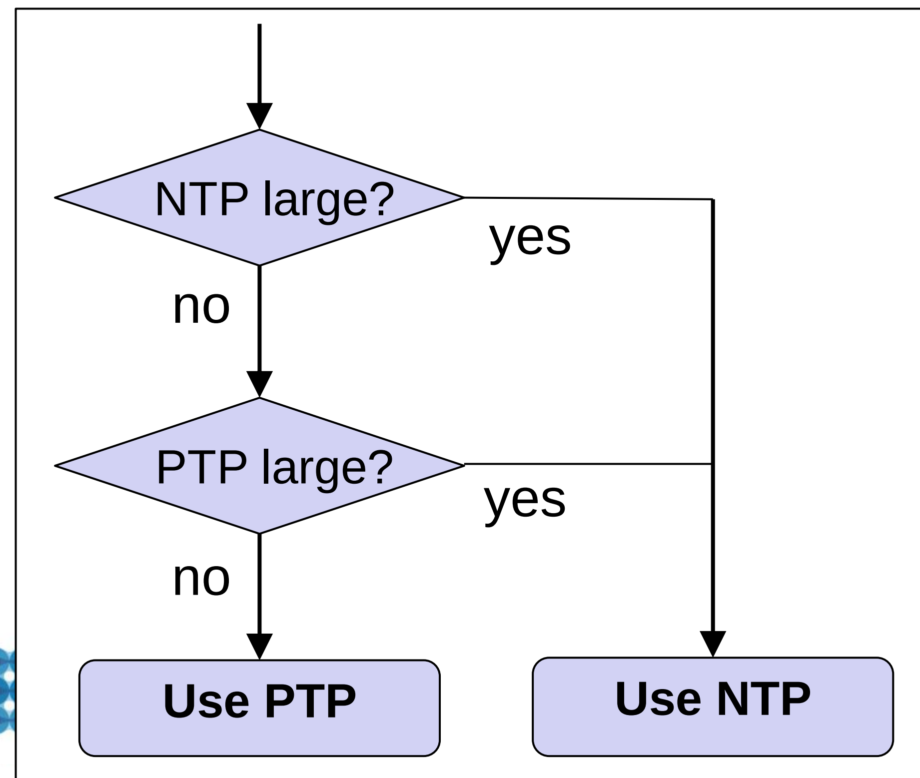
Clock Error attack



Solution, using with NTP watchdog

- NTP servers queried in parallel to PTP
- Robust median offset can override PTP offset:
    - -0.2 ms
    - +0.1 ms
    - +35000 ms
- PTP only touches the clock if allowed by the NTP watchdog





Proposal for recursive outliers

- RTS-25 today:
  - <100us @ 24 hours a day

- Proposal:
  - 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