PTP FOR FINANCIAL NETWORKS

BASICS, PITFALLS, DO’S AND DON’TS – THE VENDOR VERSION

WOJCIECH OWCZAREK, GLOBAL COMMUNICATIONS LAB, INTERCONTINENTAL EXCHANGE, INC.

ITSF 2016

PRAGUE, 2 NOVEMBER 2016

0. OUTLINE

PTP FOR FINANCIAL NETWORKS: OUTLINE

1. **PREPARE**

   What to expect

2. **EDUCATE!**

   Share the (right) knowledge

3. **THE GRANDMASTER**

   Does it fit?

4. **THE SWITCH**

   "Yes, it supports PTP"

5. **THE SLAVE**

   All that matters in the end

6. **WHAT WE DO NOT ALWAYS LIKE**

   …But have to work with

7. **PTP IN FINANCE, ILLUSTRATED**

   Using mostly rectangles

8. **CZECH PLEASE!**

   The inevitable pun

1. PREPARE

WHAT TO EXPECT

PTP FOR FINANCIAL NETWORKS: PREPARE 1/2

WHAT TO EXPECT

You will find networks with a small number of hops, but:

- A mix of high-speed interfaces: 40G / 100G backbone, 10G connected end nodes
- …or a mix of low (er) speed interfaces: 10G, 1G
- Up to several Gbps of multicast load
- Bursty traffic – but users strive not to hit buffers (“upgrade when limit in sight”)
- Some older hardware with inherent delay asymmetry
- Sharing a common network path: timing traffic together with application traffic
- Anywhere from 100 to 3-5k slaves
- No network assistance for PTP, or only some devices with network assistance
- Regulatory requirement of 100 µs faccuracy, but sub- µs for network monitoring tools
- No requirement for frequency (what-so-ever), and not phase alone. UTC time.

PTP FOR FINANCIAL NETWORKS: PREPARE 2/2

WHAT TO EXPECT

You will find users who…

- Know what they want to achieve and who know how to do it
- Expect a solution to fit their network and use case. Have you *really* got one?
- Require continuous slave monitoring, archiving past data

But also users who…

- Are in dire need of education on synchronisation: because life was easy in NTP days
- Know what they want, but got the “how to do it” completely wrong
- Only do this because they have to
- May not provide repeated sales volumes, but good publicity and success stories
  - **We have customers too**
- Are unwilling to make significant investments (yes, really):
  - **MiFID II RTS 25 first consultation: 200+ responses, only 30+ commented on timing requirement**
  - **Of those who commented, only a handful agreed to microsecond level accuracy requirement**

2. EDUCATE!

SHARE THE (RIGHT) KNOWLEDGE

PTP FOR FINANCIAL NETWORKS: EDUCATE!

SHARE THE (RIGHT) KNOWLEDGE

Remember that to some customers PTP is still relatively new, or not 100% clear:

- Review the PTP sync mechanism. Explain one-step, two-step, BC and TC operation
- Assist with end to end time distribution design
- Understand the existing network design
- Explain BMCA in detail: Priority1, ClockClass (how it changes), Accuracy, Variance, **THEN** Priority2
- Consider developing a certification track
- “Sell” the ITU metrics

Educate *yourself*:

- Software application timestamp is the ultimate target
- This is NOT a telecom network. Single-digit microsecond performance can be a challenge!

3. THE GRANDMASTER

DOES IT FIT?

PTP FOR FINANCIAL NETWORKS: THE GRANDMASTER 1/2

DOES IT FIT?

Flickr: ePublicist

- **Make sure you provide:**
  - Default profile and Enterprise profile support
  - Options for future growth: is it modular?
  - Multiple reference support, get ready for GNSS alternatives
  - Copper and optical. Gigabit is a must, forget 100M, have 10GE ready soon
  - Clear specifications: “*n* Delay Request / second”, not “supports *n* slaves”
  - 24h holdover phase offset specification for given oscillator option
- **Prepare to support multiple time source arbitration / combining, not just failover**
- **Consider delayed clockClass degradation in early holdover to prevent BMCA flapping**
- **If not already available, consider BC translating between G.8265.1 and multicast**

PTP FOR FINANCIAL NETWORKS: THE GRANDMASTER 2/2

DOES IT FIT?

A GM is easy to replace. Ensure that:

- PTP output is stable at cold start. Do not announce in unhealthy state!

Do not make mistakes with leap second handling:

- Know about leap second issues before your customer finds out! Provide a test mode
- Also before your GNSS chip vendor finds out. Get a GNSS simulator or go home.
- Ensure that GNSS firmware upgrade is possible as part of GM firmware upgrade
- Be IEEE 1588 compliant. Announce leap second at 12:00:01 PM. **NOT AM NEXT DAY!**

Timing is the minimum. Work on the extras:

- Slave monitoring solution
- Measurement / calibration 1PPS input

Do not forget about security:

- Timing access control lists, ability to disable SET type management messages!

4. THE SWITCH

“YES, IT SUPPORTS PTP”

PTP FOR FINANCIAL NETWORKS: THE SWITCH

“YES, IT SUPPORTS PTP”

A PTP-enabled switch fit for purpose:

- Will not stand a chance with no advanced Layer 3 protocol set and PTP over 802.1Q
- Industrial type switch may be used in the “bolt-on” network variant

TC:

- Make sure you stay IEEE 1588 compliant when interfacing one-step with two-step
- Two-step only TC may work, as long as processing does not limit PTP throughput
- Must support correction when crossing different link speeds

BC:

- Make sure filtering is up to the task: shield the slaves from disturbances
- Must be secure. Implement forced port roles: Master-only, Non-master
- Be careful not to drop PTP timescale or leap second information when sync is lost!
- Provide monitoring capabilities (SNMP?), log, alarm and trap on sync loss events

5. THE SLAVE

ALL THAT MATTERS IN THE END

PTP FOR FINANCIAL NETWORKS: THE SLAVE 1/2

ALL THAT MATTERS IN THE END

The NIC:

- Must support timestamping for VLAN-tagged multicast AND unicast PTP
- Design for 1PPS input and output options

The PTP slave software stack:

- Needs to support bonded interfaces (active/stand-by)
- Needs to keep all NICs and system clock in sync
- Needs to be robust against failures: link, GM change, reloading drivers
- Needs to work with any NIC
- Must protect the system (OS) clock from transient PTP clock instabilities
- Should have an option to prevent from backwards clock step

PTP FOR FINANCIAL NETWORKS: THE SLAVE 2/2

ALL THAT MATTERS IN THE END

The PTP slave software stack, continued:

- Requires advanced filtering – there may be a lot of PDV
- Should protect from the effects of a GM misbehaving (leap second issues)
- Easy to troubleshoot: status overview, message and error counters
- Easy to monitor: remote polling, alarm on failures (LISTENING, no Sync, no Delay)
- Provisions for basic security: access lists, disabling “SET” management messages
- Robust leap second support (be prepared for late or premature leap second insertion)
- Offline leap seconds (leap seconds file)
- Take the NIC vs. OS clock offset into account when providing metrics

6. WHAT WE DO NOT ALWAYS LIKE

…BUT HAVE TO WORK WITH

PTP FOR FINANCIAL NETWORKS: WHAT WE DO NOT ALWAYS LIKE 1/2

…BUT HAVE TO WORK WITH

The fearmongers:

- ***"Only product X meets MiFID requirements. If you use product [non-X], your network will self-combust".*** Most products available today, GM or slave, will allow for full compliance, with careful engineering on our end. Some will make it easier, some will do things for us, but most will work.

- ***"Only multiple time source inspection will guarantee compliance. If you use one source, your neighbour's cat will run into your server room and self-combust, taking your network with it".***

  Multiple time source interrogation is definitely a benefit, but is not essential for initial deployment:

  - Most of today's PTP deployments rely on standard one-active-source, and they are hundreds of times bigger than finance PTP. And yet, it moves.
  - In practice, for finance true multi-source is often not feasible: deploying an extra NIC or extra network path. Three sources over the same path are as good as one source. May end up being a glorified last-resort NTP failover.

- ***"The use of GNSS timing is a health hazard. If you use GPS, a satellite will drop on your DC and self-combust".*** GNSS is prone to jamming – but this is what good oscillators and holdover are for. GNSS can be spoofed – but a co-ordinated attack on multiple antennas is difficult. GNSS vulnerabilities are not to be underestimated, however it still remains the most affordable time source. Start with GNSS and prepare for alternatives: eLORAN and terrestrial timing services from timing labs.

PTP FOR FINANCIAL NETWORKS: WHAT WE DO NOT ALWAYS LIKE 1/2

…BUT HAVE TO WORK WITH

Products “designed for financial services”:

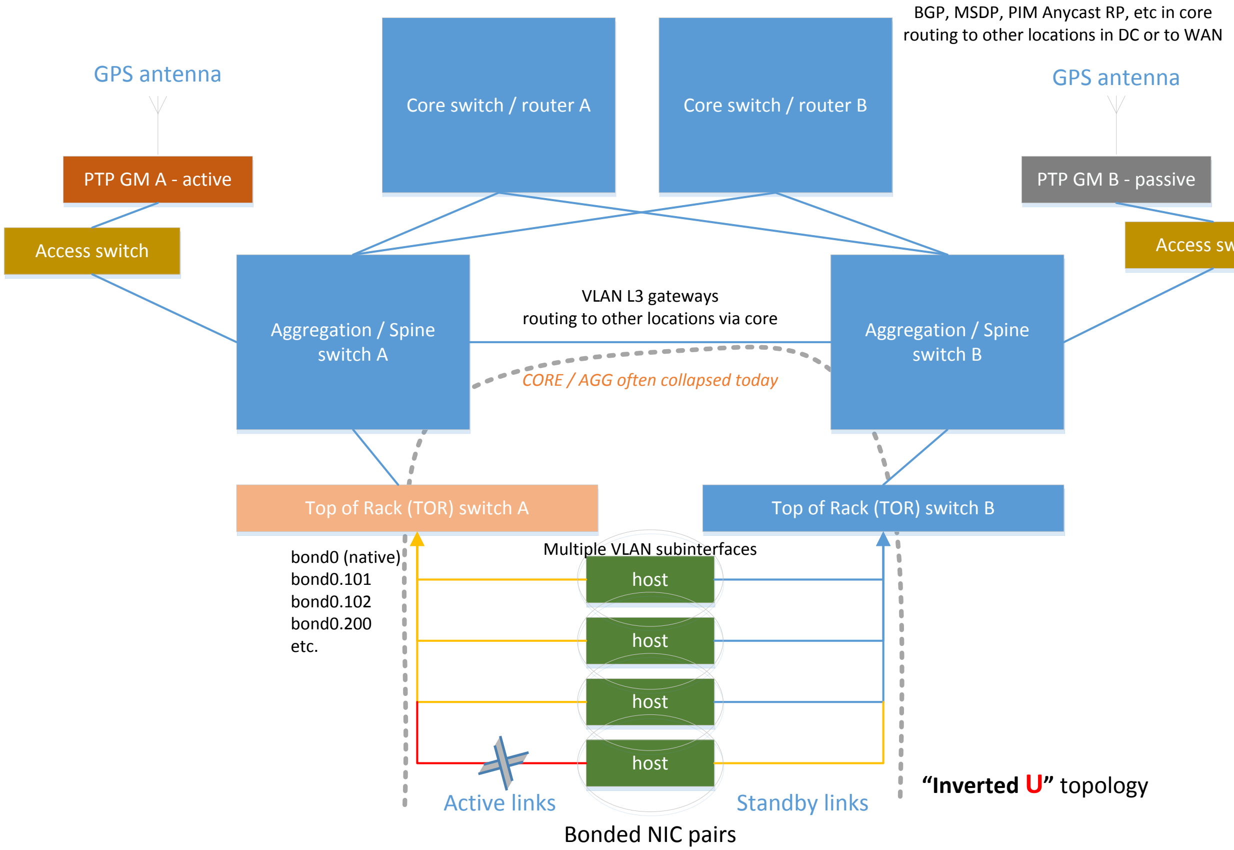
- PTP encompasses a set of capabilities and modes of operation, nothing more
- Our sync requirements are not too different to other industries’
- The key difference is the final time destination, the OS clock
- We need monitoring, but so do other industries

7. PTP IN FINANCE, ILLUSTRATED

USING MOSTLY RECTANGLES

PTP IN FINANCE, ILLUSTRATED 1/4

TYPICAL DESIGNS #1: SHARED NETWORK – PTP AND DATA USE THE SAME PATHS

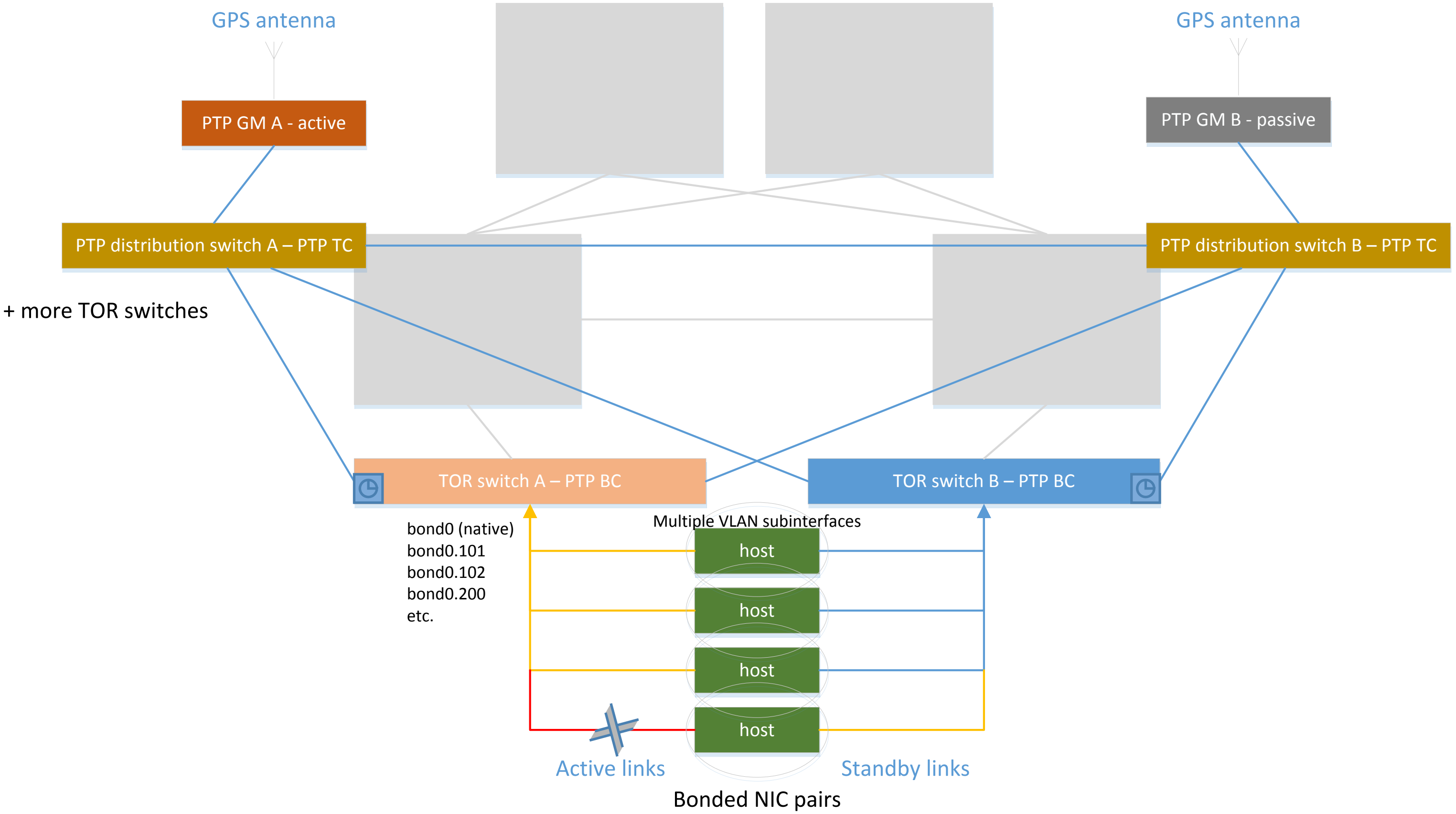


BGP, MSDP, PIM Anycast RP, etc in core
routing to other locations in DC or to WAN
GPS antenna
Core switch / router A
Core switch / router B
GPS antenna
PTP GM A - active
PTP GM B - passive
Access switch
Access switch
VLAN L3 gateways
routing to other locations via core
Aggregation / Spine switch A
Aggregation / Spine switch B
CORE / AGG often collapsed today
Top of Rack (TOR) switch A
Top of Rack (TOR) switch B
Multiple VLAN subinterfaces
bond0 (native)
bond0.101
bond0.102
bond0.200
etc.
host
host
host
host
Active links
Standby links
"Inverted U" topology
Bonded NIC pairs

- At least 3 hops between GM and slave
- No equipment in-between is PTP aware
- PTP traffic contends with other traffic
- Separate management network
- A mix of 10GE, 40GE and GE links

PTP IN FINANCE, ILLUSTRATED 2/4

TYPICAL DESIGNS #2: BOLT-ON TIMING NETWORK



- PTP meets other traffic only at TOR, which is a BC
- Ideally, PTP network switches PTP aware (TC)
- Industrial switches could be used
- Very little load on PTP network
- Almost ideal conditions for time sync
- NIC still shares PTP with other traffic
- Unless PTP TC used, whole PTP network running at the same speed

PTP IN FINANCE, ILLUSTRATED 3/4

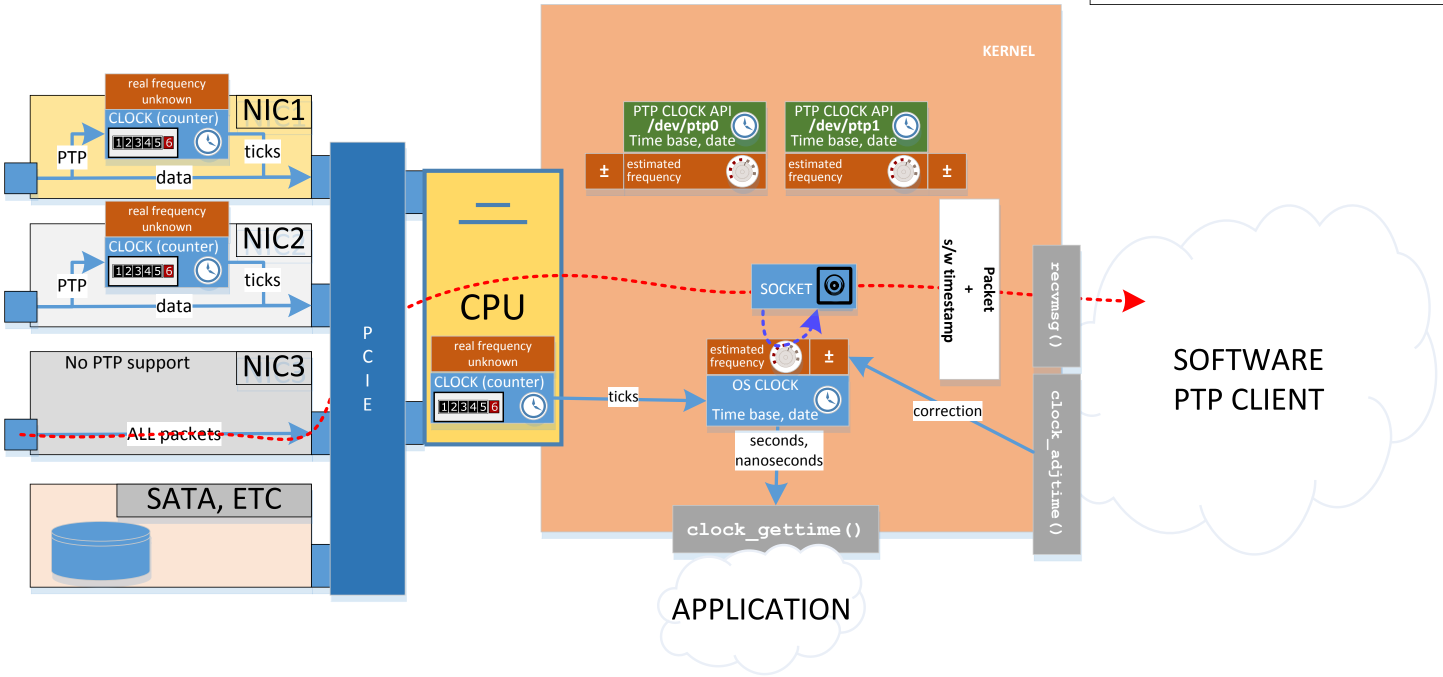
SOFTWARE PTP CLOCK SYNC: STRAIGHT TO OPERATING SYSTEM CLOCK, ROCKY ROAD

PTP clock sync – software timestamps:

1. Frame starts arriving at active NIC

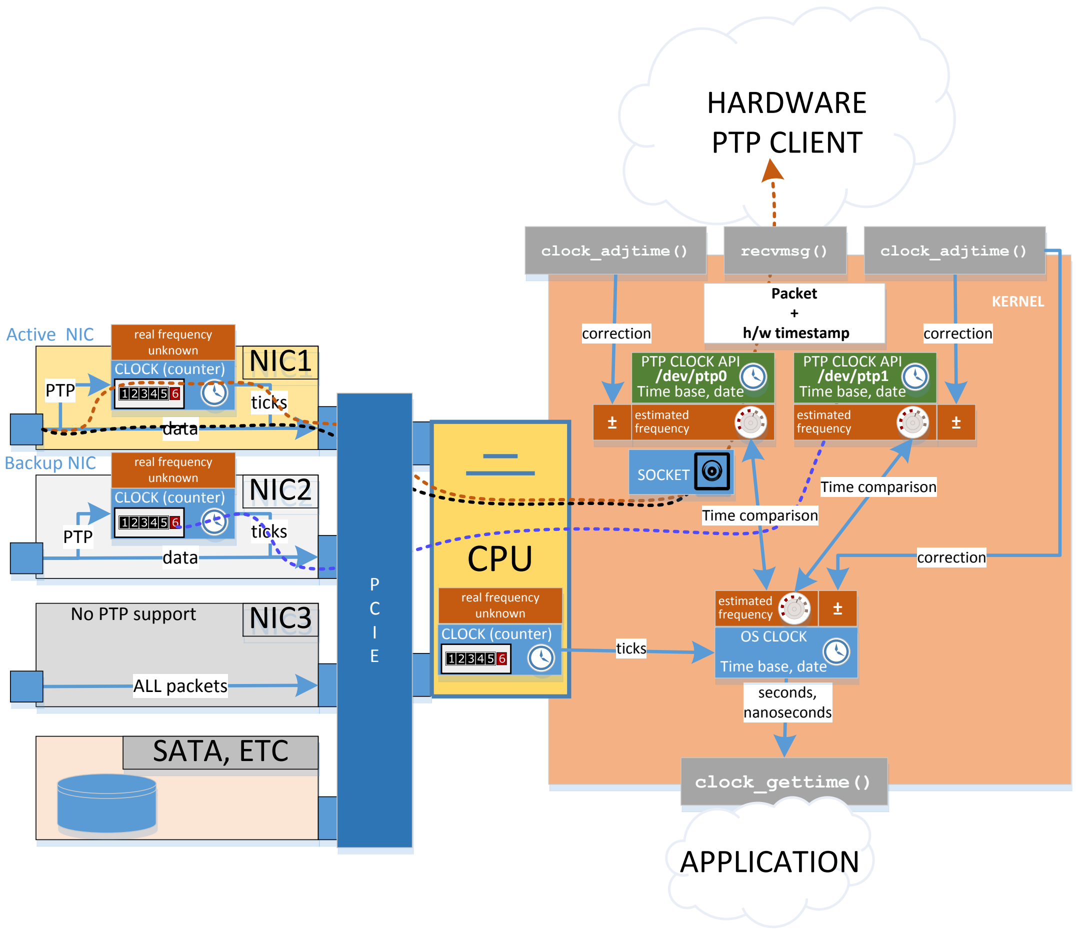
4. RX interrupt, scheduling, CPU busy (jitter)

5. Packet picked up by kernel to socket buffer
6. System time read, attached to packet
8. Packet and timestamp reach application (PTP client)
9. PTP client calculates offset, corrects frequency estimate



PTP IN FINANCE, ILLUSTRATED 3/4

HARDWARE PTP CLOCK SYNC: BUT THE APPLICATION WILL STILL USE THE OPERATING SYSTEM CLOCK



PTP clock sync – hardware timestamps
(one possible variant, simplified):

1. Frame starts arriving at active NIC
2. Counter value grabbed
3. Matched as PTP, counter pushed to timestamp FIFO
4. RX interrupt, packet picked up by kernel to socket buffer
5. Packet matched in kernel as PTP, needs timestamp
6. Counter popped from FIFO (match on seq ID and msg type)
7. Converted to timestamp, based on frequency estimate
8. Packet and timestamp reach application (PTP client)
9. PTP client calculates offset, corrects frequency estimate

OS to NIC clock sync
(PTP_SYS_OFFSET ioctl on /dev/ptpX):

1. Read system time TS_1
2. Read NIC time (counter register), convert to time TS_2
3. Read system time TS_3
4. Run N times, Select lowest $TS_3 – TS_1$
5. Delta = $TS_{2best} – TS_{1best} – (TS3_{best} – TS1_{best}) / 2$
6. Correct OS clock frequency estimate

NIC to NIC clock sync:

1. Get Delta1 = OS vs. NIC1, Delta2 = OS vs. NIC2
2. NIC1 vs. NIC1 = Delta2 – Delta1
3. Correct NIC2 clock frequency estimate

8. CZECH PLEASE!

THE INEVITABLE PUN

PTP FOR FINANCIAL NETWORKS: CZECH PLEASE!

Between the GM and the slave is a familiar territory (can be dragons though)…

…but between the NIC and the application clock…

PTP FOR FINANCIAL NETWORKS: CZECH PLEASE!

Between the GM and the slave is a familiar territory (can be dragons though)…

…but between the NIC and the application clock…

…is a bog…

PTP FOR FINANCIAL NETWORKS: CZECH PLEASE!

DUE HOMAGE TO IVAN MLÁDEK

Between the GM and the slave is a familiar territory (can be dragons though)…

…but between the NIC and the application clock…

…is a bog… on the way to Vizovice…

https://www.youtube.com/watch?v=S4aqM_wu6Ns

Wikipedia: Jožin z bažin

THANK YOU

QUESTIONS?

Wojciech Owczarek
Global Communications Lab

Intercontinental Exchange, Inc.

wojciech.owczarek@theice.com

Anti-spam tip #1: this is an image