Challenges deploying PTPv2 in a Global financial company

Pedro Estrela, Lodewijk Bonebakker
Performance Engineer
22-Sep-2012

Outline

- Paper contributions
  - PTPv2 protocol issues
  - Future WAN support
  - Real-world large deployment tests

Initial global NTP situation

- Truly global private network:
  - All global timezones
  - 10s of datacenters + leased lines
  - 100s of legacy network equipment
  - 1000s of end trading-servers

- Initial Time Sync:
  - PTPv1 inside datacenters
  - NTP between datacenters to far away GPS antennas

Basic WAN support

- Multicast TTL>1:
  - One GM for all timezones !

- Clock separation options:

  a) Fine tuning TTLs

  b) PTPv2 Sub-domain field

  c) Traffic blocking (ACLs)

  d) Separate v1 groups

  e) Multiple roots for the same group (PIM-SP RP)

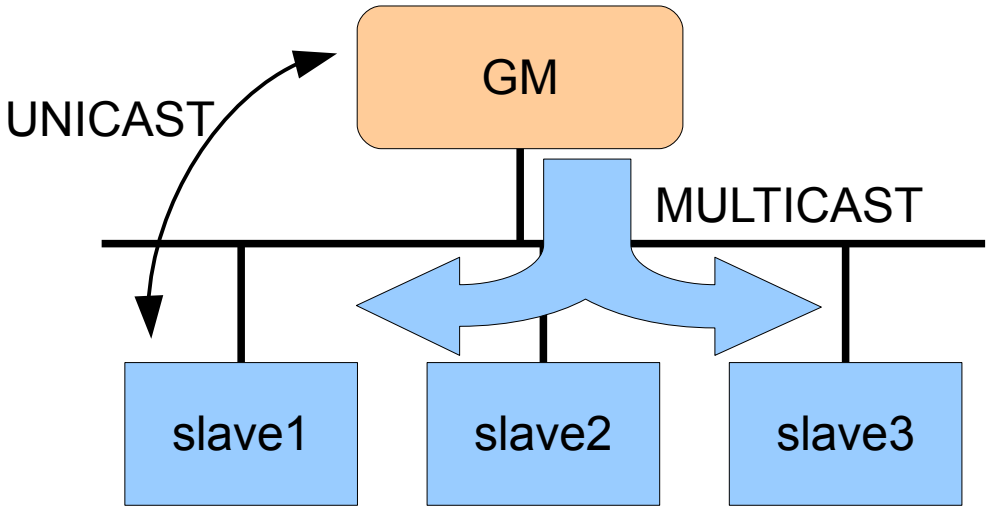
Scalability issues

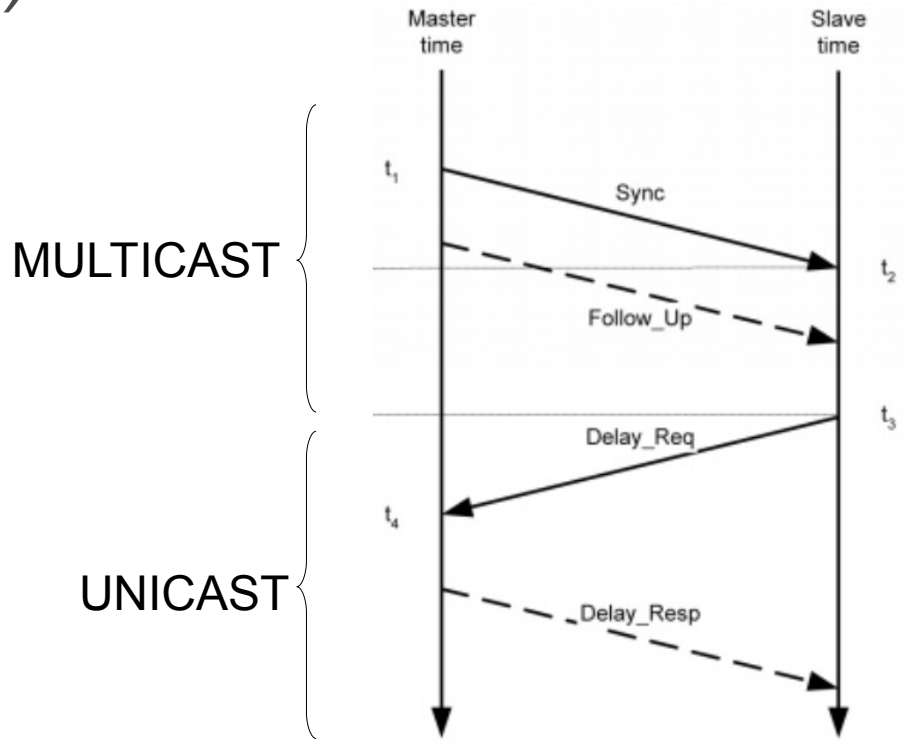
- Multicast profile
  - Upstream messages overhead
    - (think 1000s of clients, no BCs on the network)
  - Endless problems from “All-to-All” semantics

- Unicast profile
  - Downstream messages overhead
  - No implementation support

Hybrid mode solution

- Downstream: regular multicast
- Upstream: unicasted directly to the GM
  - *(already contributed to PTPd)*

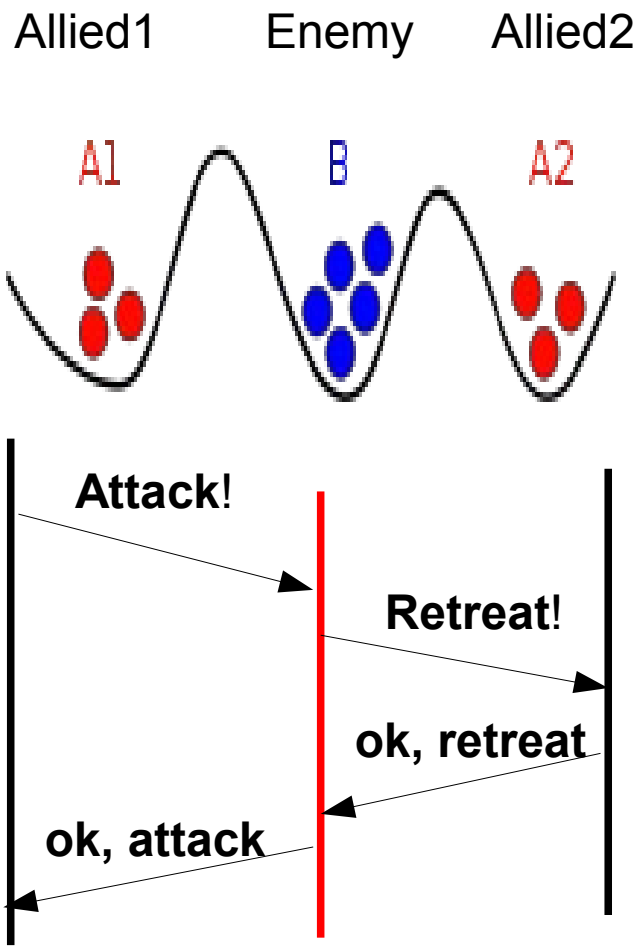


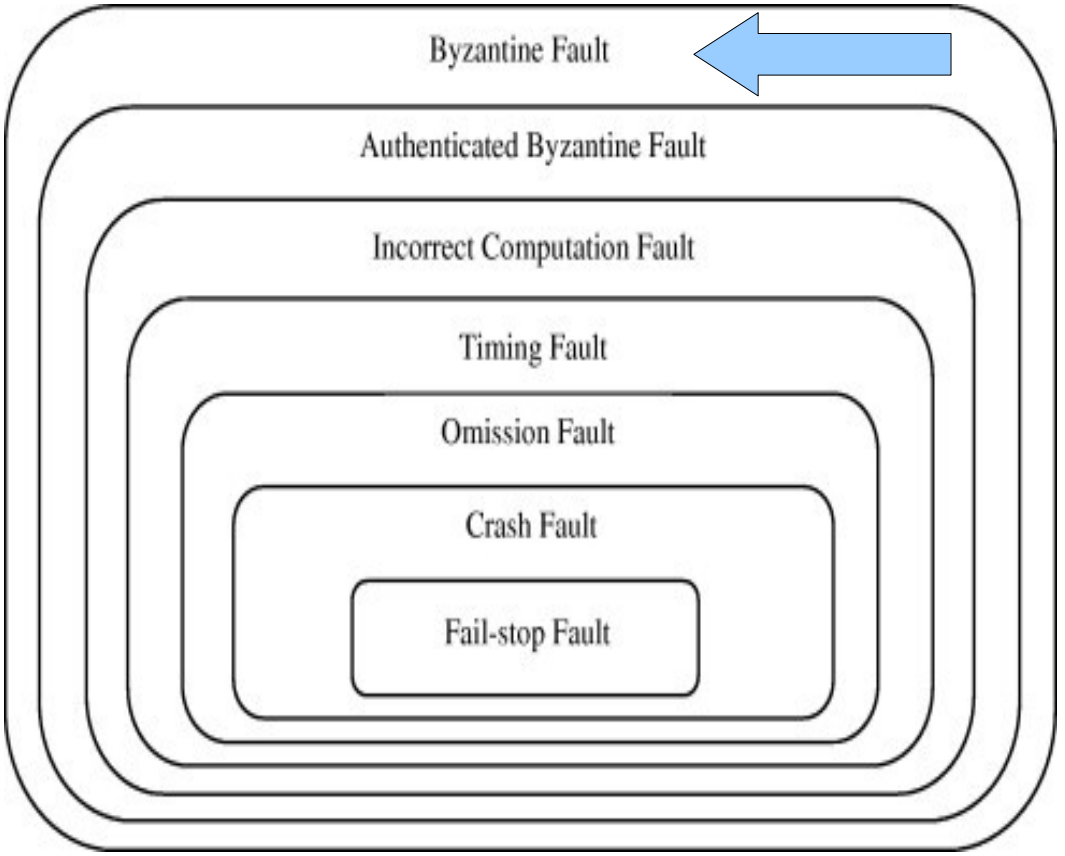


Robustness problems

- Random operational problems
  - GM disconnections, incorrect BMCs, etc
  - “All-to-All” multicast problems = clock drift
  - No mandatory security = trivial to damage PTP
- GM “traitor” scenario:
  - GM sent bad time (leap seconds = 0)
  - Backup GMs stayed passive (same BMC)
  - Clients trusted their single GM = clock jumps

Byzantine Theory recap





Byzantine robustness

- Initial issue mitigation
  - Deprecated clock stepping
  - Deprecated “long” slews
- Complete solution:
  - Always corner cases with single time source
  - Clients must listen to multiple sources
  - 1997 proof: minimum 2*T+1 time sources

Proposal: Enterprise profile

- How to replace NTP in large companies?
  - Focus: UTC to end-user applications
  - Smooth migration is essential
  - WANs to be supported out of the box
- Also requested:
  - Scalability (hybrid mode)
  - Accuracy (jitter filters)
  - Robustness (multiple time sources)

Conclusions

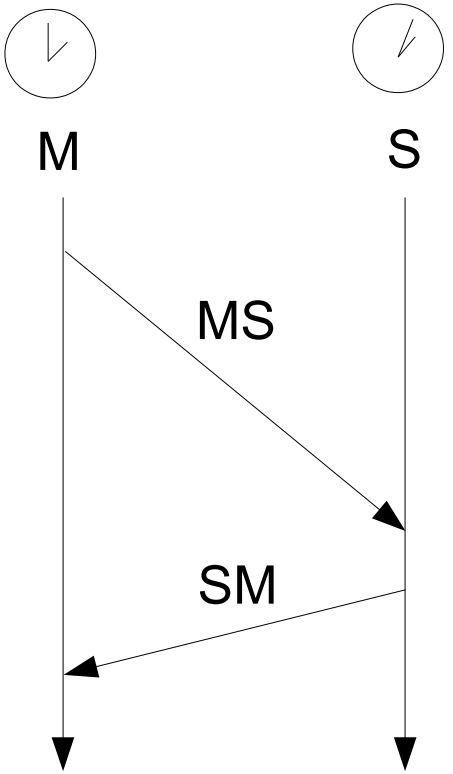
- Issues on PTPv2 itself:
  - Single multicast group for clocks
  - Only “all-to-all” Multicast (or full Unicast)
  - GM = Single point of failure
  - Security is still experimental option

- *(Only covered on the paper)*:
  - Monitoring: no confidence interval estimation
  - Vendor testing: no standardized jitter models
  - Circular delayReq interval dependency

Extra slides

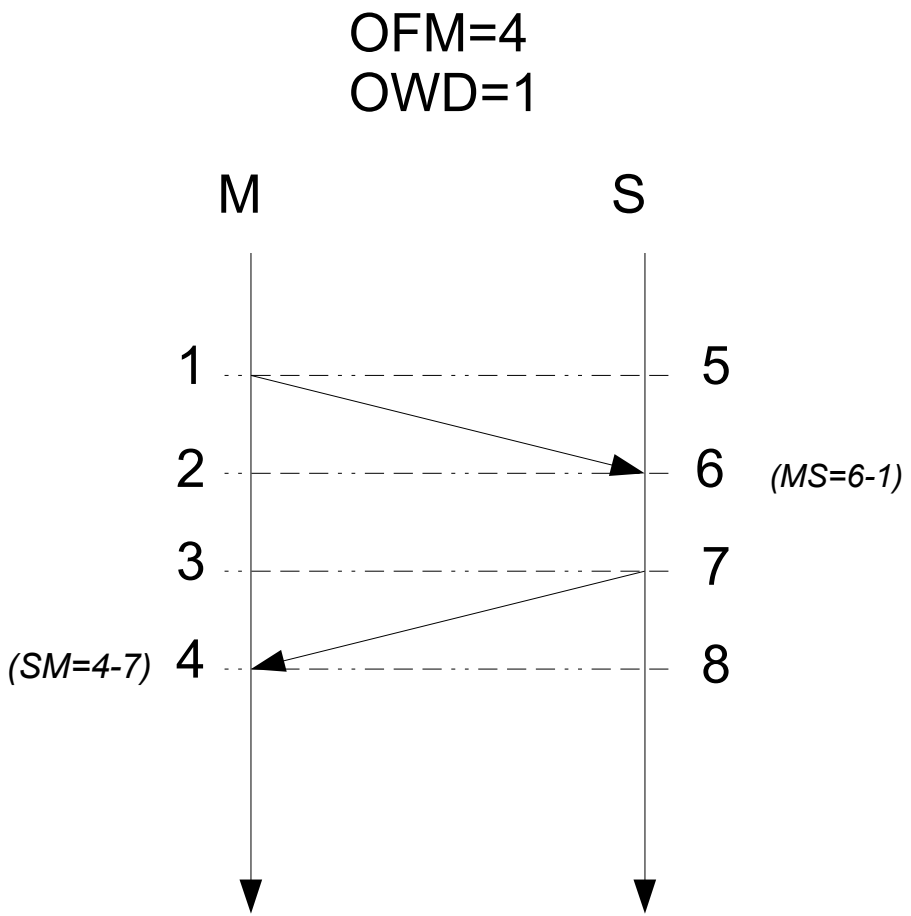
Fundamental challenge

- Always 3 variables...
  - Clock difference
  - Forward delay
  - Return delay
- ...but only 2 equations
  - MS = (OWD_1 + OFM)
  - SM = (OWD_2 – OFM)



- Symmetric paths assumption is inevitable

Detailed example



$$OWD_1 = OWD_2$$

$$OFM = (MS-SM)/2$$
$$= ((6-1)-(4-7))/2$$
$$= (5+3)/2$$
$$= \mathbf{4}$$

$$OWD = (MS+SM)/2$$
$$= ((6-1)+(4-7))/2$$
$$= (5-3)/2$$
$$= \mathbf{1}$$

Promoting symmetric paths

- "Easy" solutions
  - Dedicated paths: separate network
  - Shorter paths: closer GPS
  - Faster paths: 10G serialization time
  - Higher sample rate: multicast distribution
- Reduce buffer queuing
  - Clients: NIC timestamping
  - Routers: Boundary clocks; E2E TC; QoS
  - Lines: P2P transparent clocks