Cornell University

Globally Synchronized Time via Datacenter Networks

Vishal Shrivastav, Cornell University

Joint with Ki Suh Lee, Han Wang, and Hakim Weatherspoon

COSENERS 2016

How can we scalable synchronize clocks with high precision?

Scalable – Entire datacenter

High precision – *bounded precision*; e.g. no clock differs by more than hundreds of nanoseconds

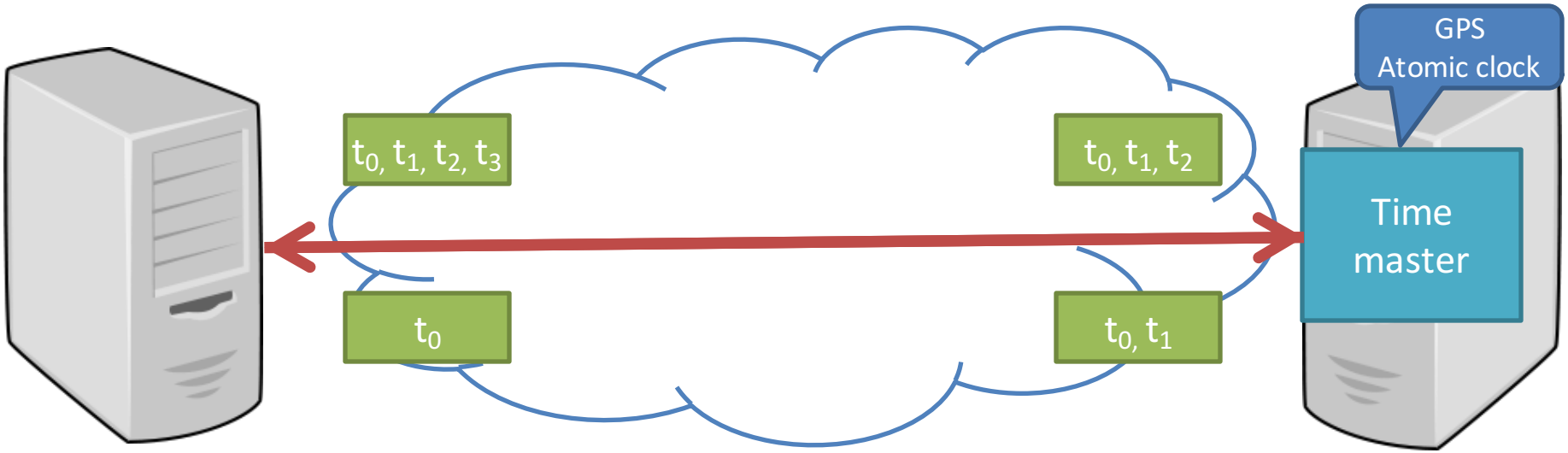
Capability essential for network and distributed applications

Networks – One-way delay, consistent updates, etc

Distributed systems – consensus, snapshots, etc

Problem: Clock synchronization is non-trivial

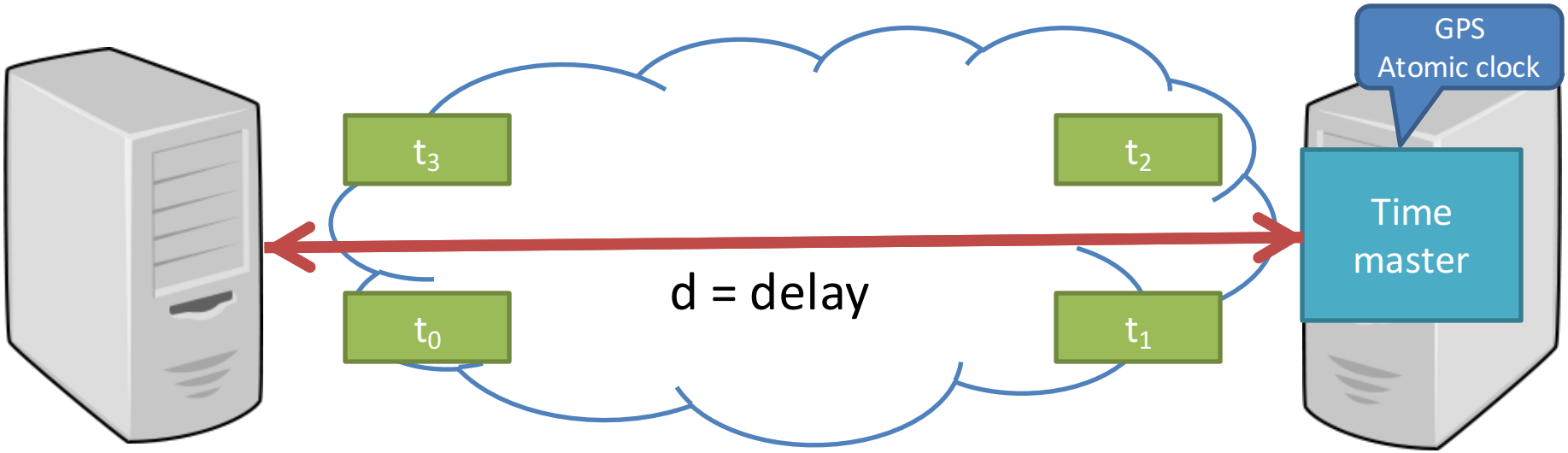
- Precision: difference between any two clocks
- Typical clock offset synchronization
  - Offset



Problem: Clock synchronization is non-trivial

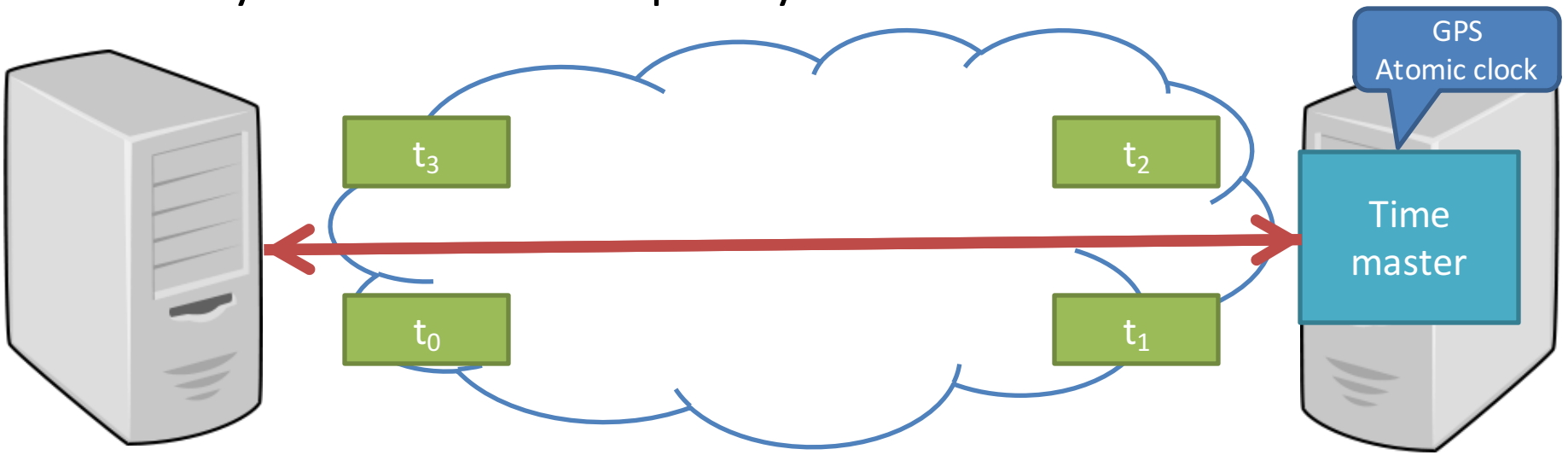
- Precision: difference between any two clocks
- Typical clock offset synchronization
  - Offset = $((t_1 - t_0) - (t_3 - t_2))/2$

[d+offset] [d-offset]



Problem: Clock synchronization is non-trivial

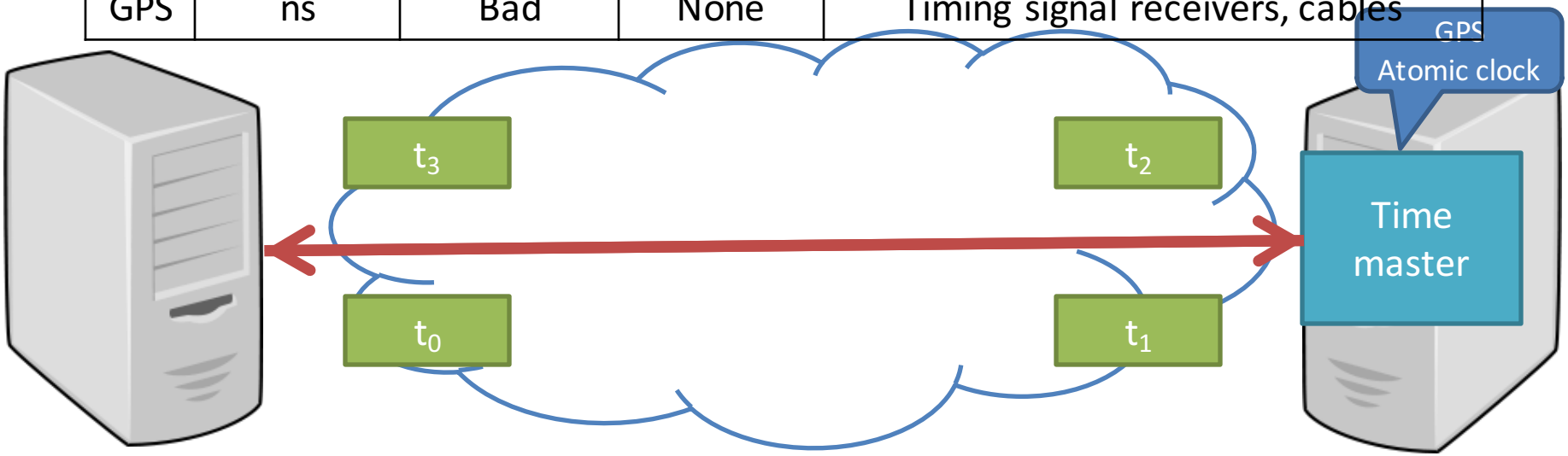
- Precision: difference between any two clocks
- Problems affecting precision
  - Oscillator skew (i.e. frequency of clocks differ)
  - Reading remote clocks: timestamps, network stack, network jitter
  - Resynchronization frequency



Problem: Clock synchronization is non-trivial

- Precision: difference between any two clocks
- Synchronization Protocols

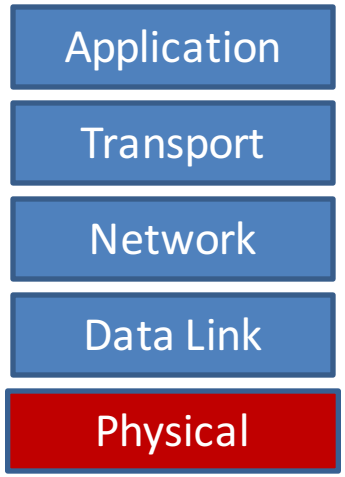
	Precision	Scalability	Overhead	Extra Hardware
NTP	us	Good	Moderate	None
PTP	sub-us	Good	Moderate	PTP-enabled devices
GPS	ns	Bad	None	Timing signal receivers, cables



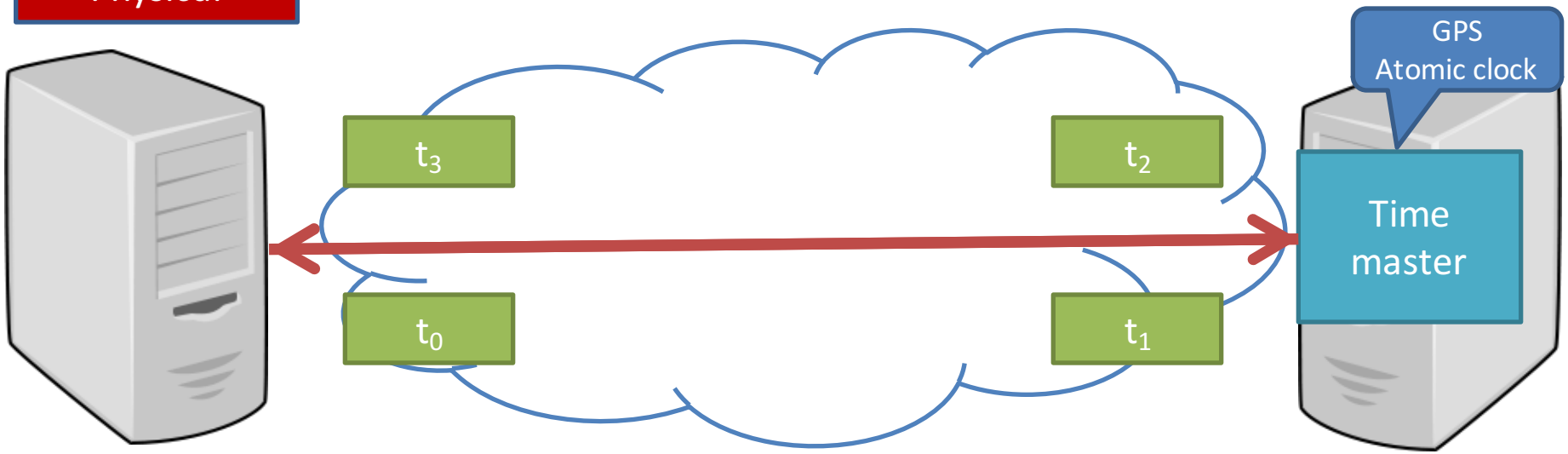
Outline

- Introduction
- Design
- Evaluation
- Conclusion

Use the PHY to synchronize clocks

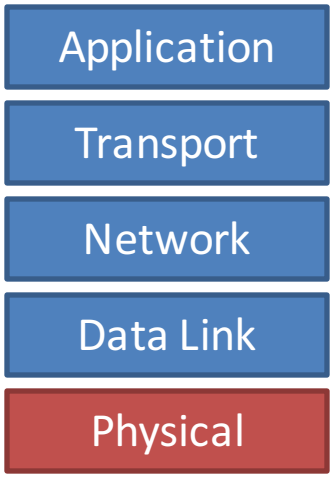


- Protocol in the PHY
  - Each physically link is already synchronized!
  - No protocol stack overhead
  - No network overhead
  - Scalable: peer-to-peer and decentralized



DTP

- 10 Gigabit Ethernet
  - Idle Characters (/I/) and Control blocks (/E/)

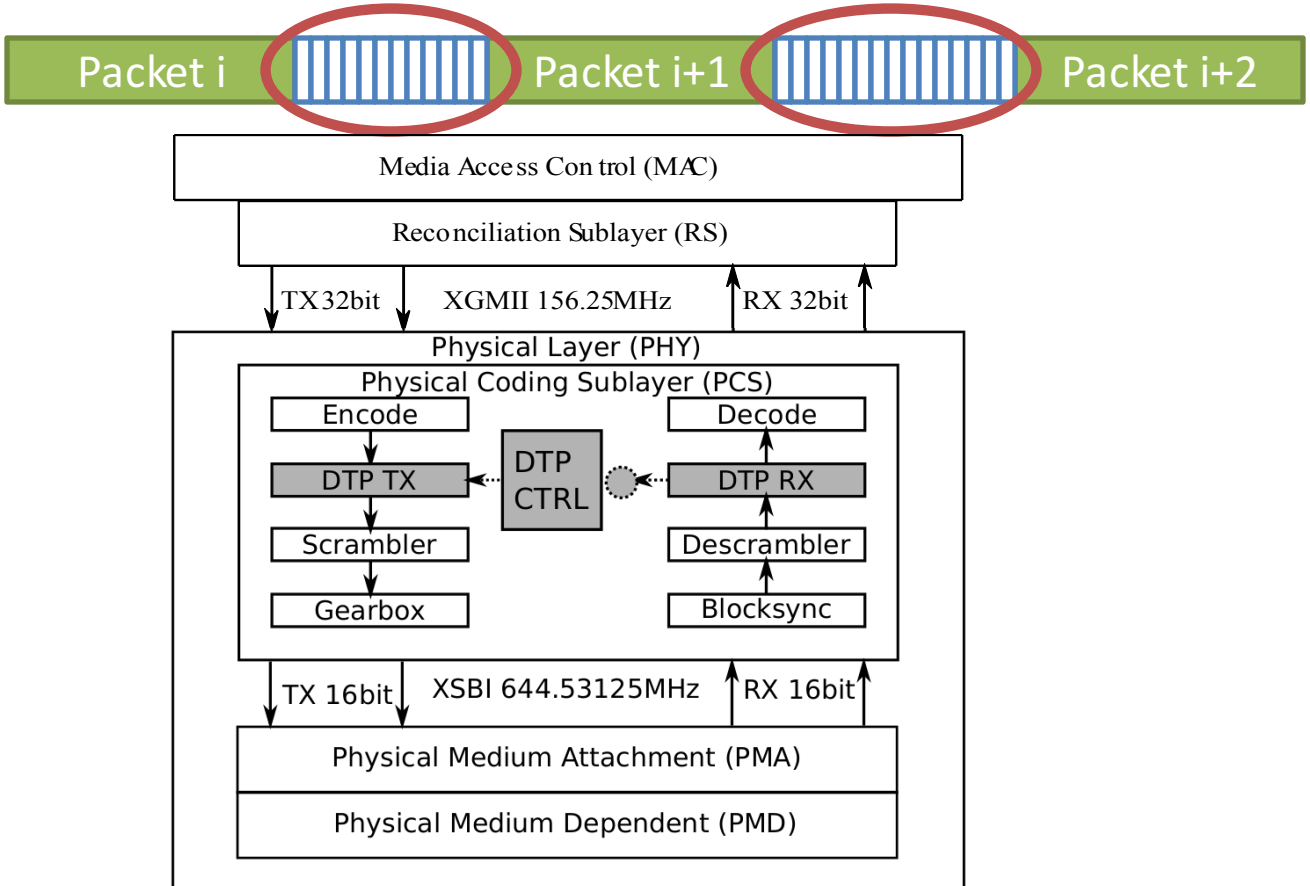


  - Standard requires at least 12 idle characters /I/ between pkts
    - i.e. At least one 64-bit Control Block /E/ between pkts
  - Idle characters / control blocks sent even if no packets to send
  - DTP overwrites idle characters (control block) to send protocol messages

    DTP does not effect standard at all

DTP

- 10 Gigabit Ethernet
  - Idle Characters (/I/) and Control blocks (/E/)

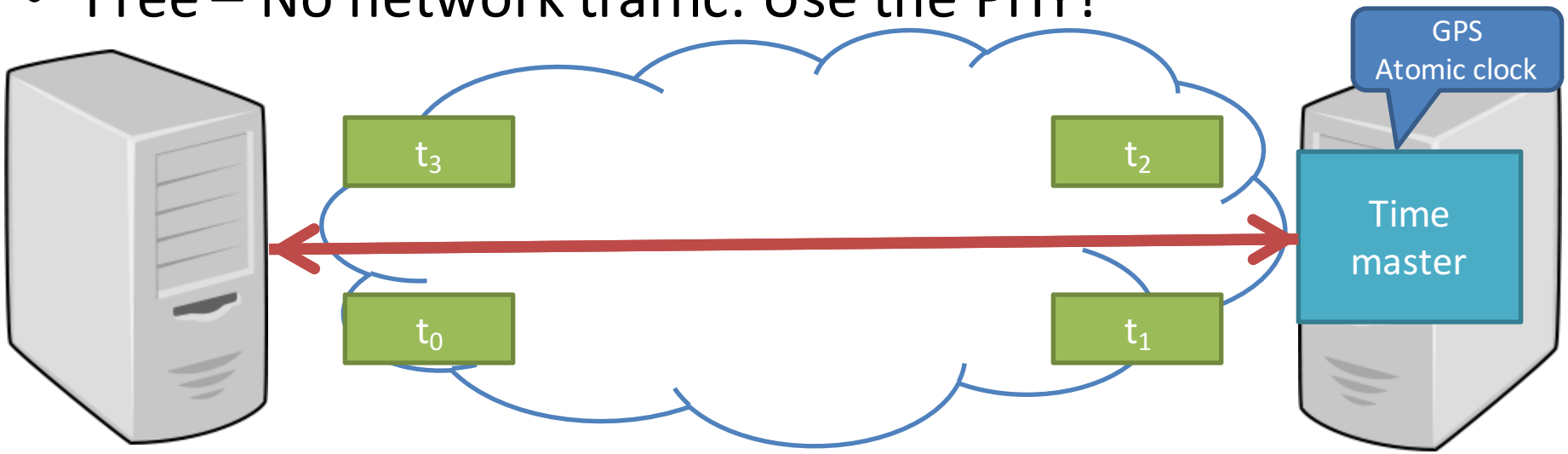


Application
Transport
Network
Data Link
Physical

Datacenter Time Protocol (DTP)

Precise and bounded synchronization

- 4 oscillator ticks (25ns) bounded peer-wise synchronization
- 150ns precision synchronization for an entire datacenter
- No clock differs by more than 150ns
- Free – No network traffic: Use the PHY!

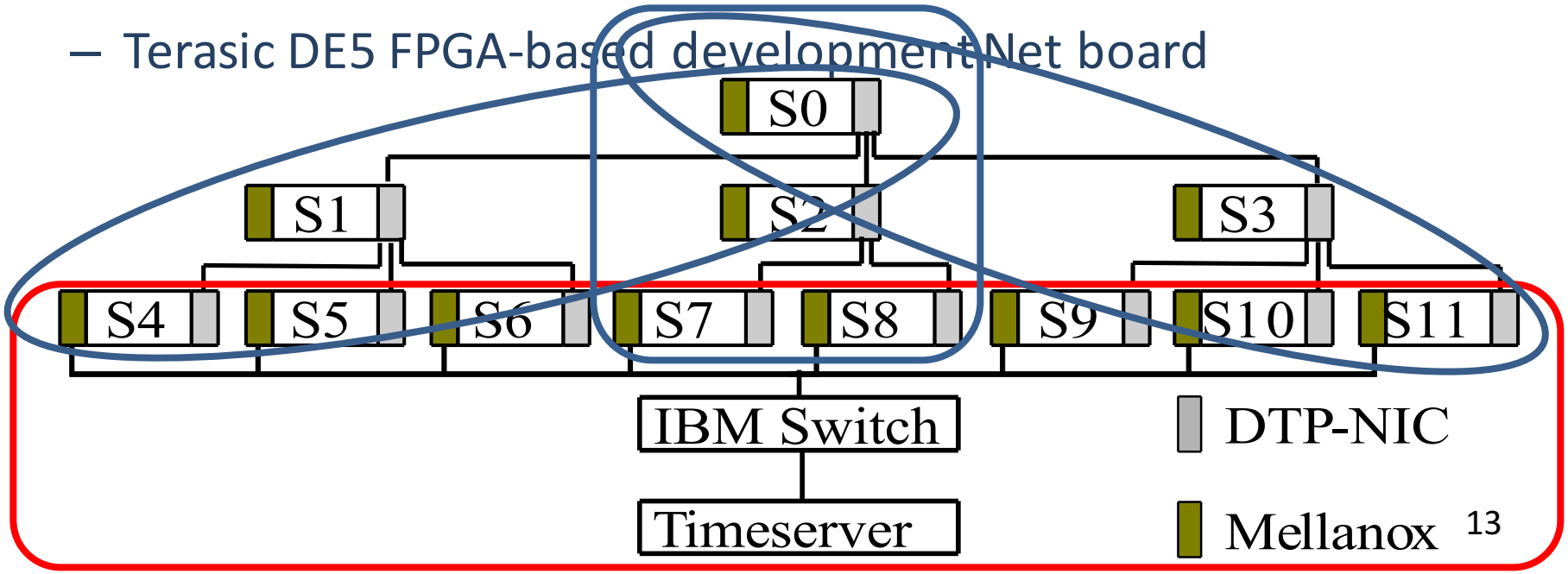


Outline

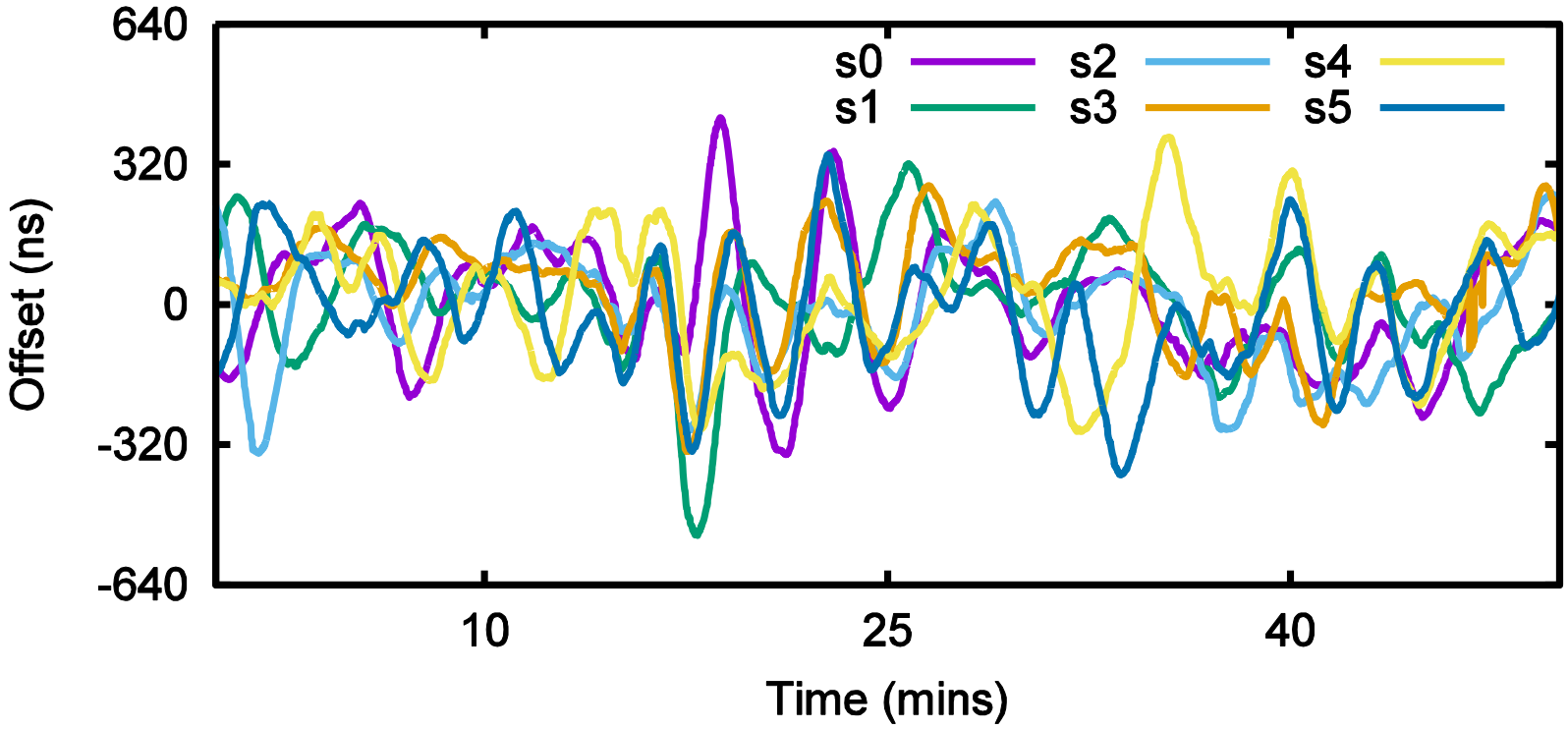
- Introduction
- Design
- Evaluation
- Conclusion

Evaluation

- Compare measured precision of DTP and PTP
  - Measurement and observation period was two days
- PTP: Compare precision between Timeserver and Servers
  - Mellanox NIC (hardware), IBM G8264 Switch, Timekeeper server
- DTP: Compare precision between leaf servers and switches
  - Terasic DE5 FPGA-based development board

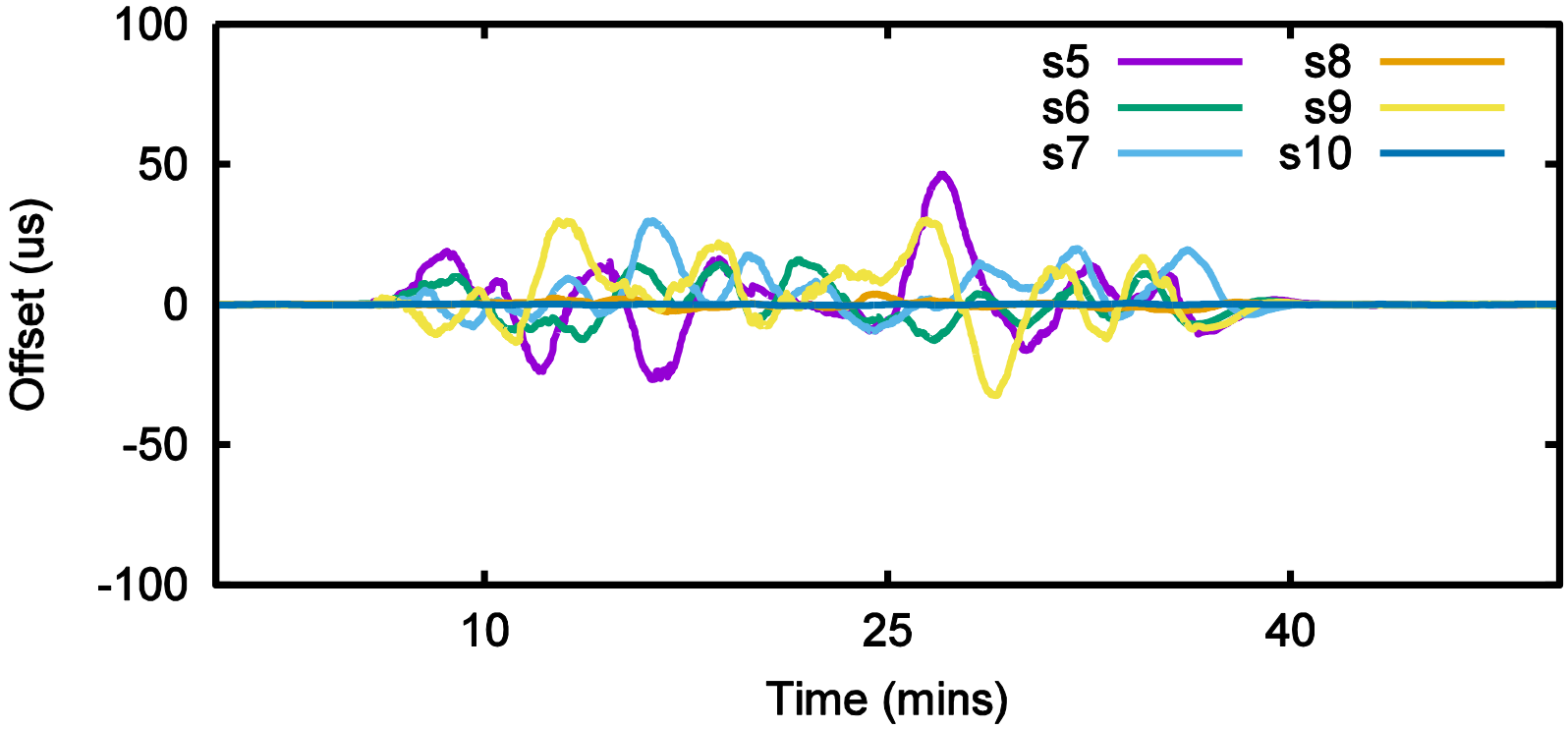


PTP – Idle Network (No Network Traffic)



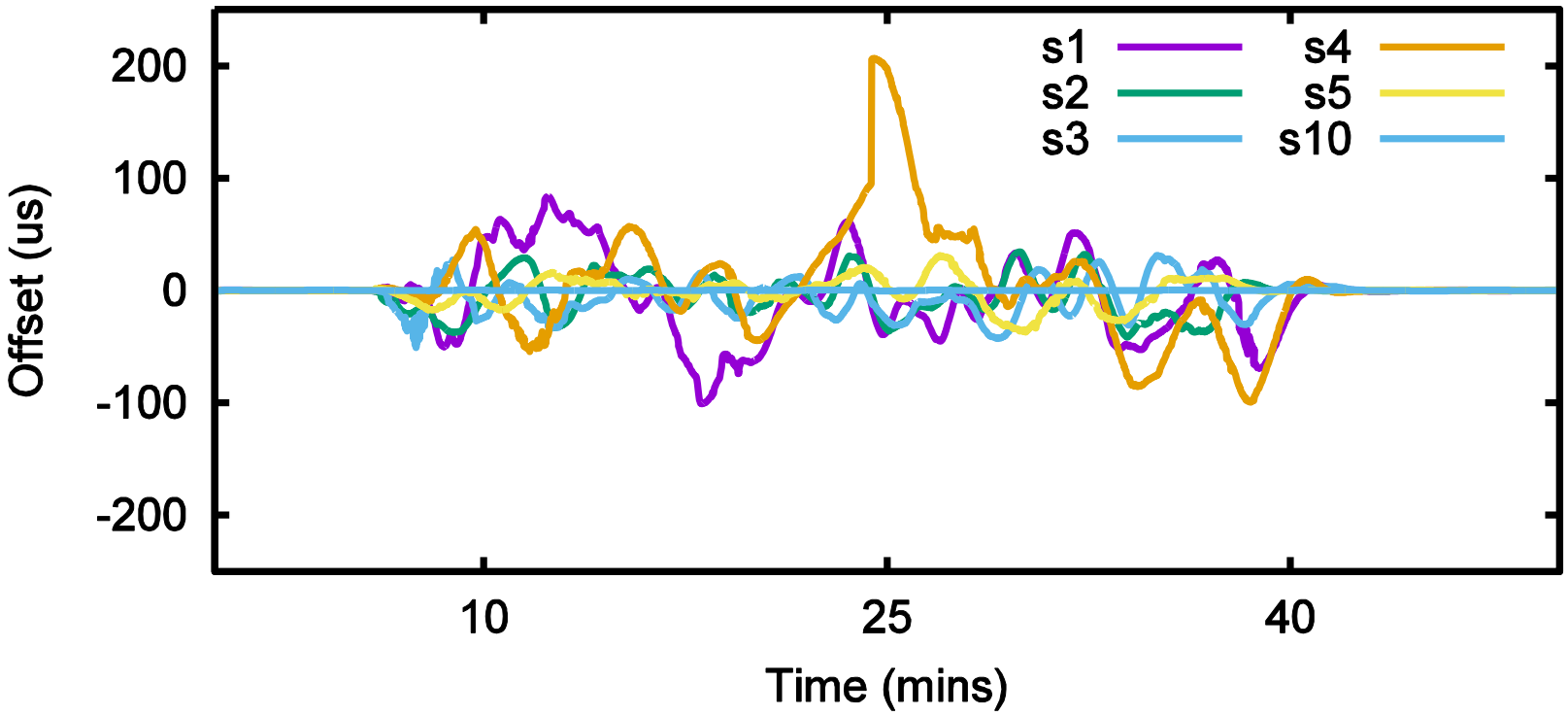
Clocks differ by a few hundred nanoseconds

PTP – Medium Loaded Network (4Gbps Traffic)



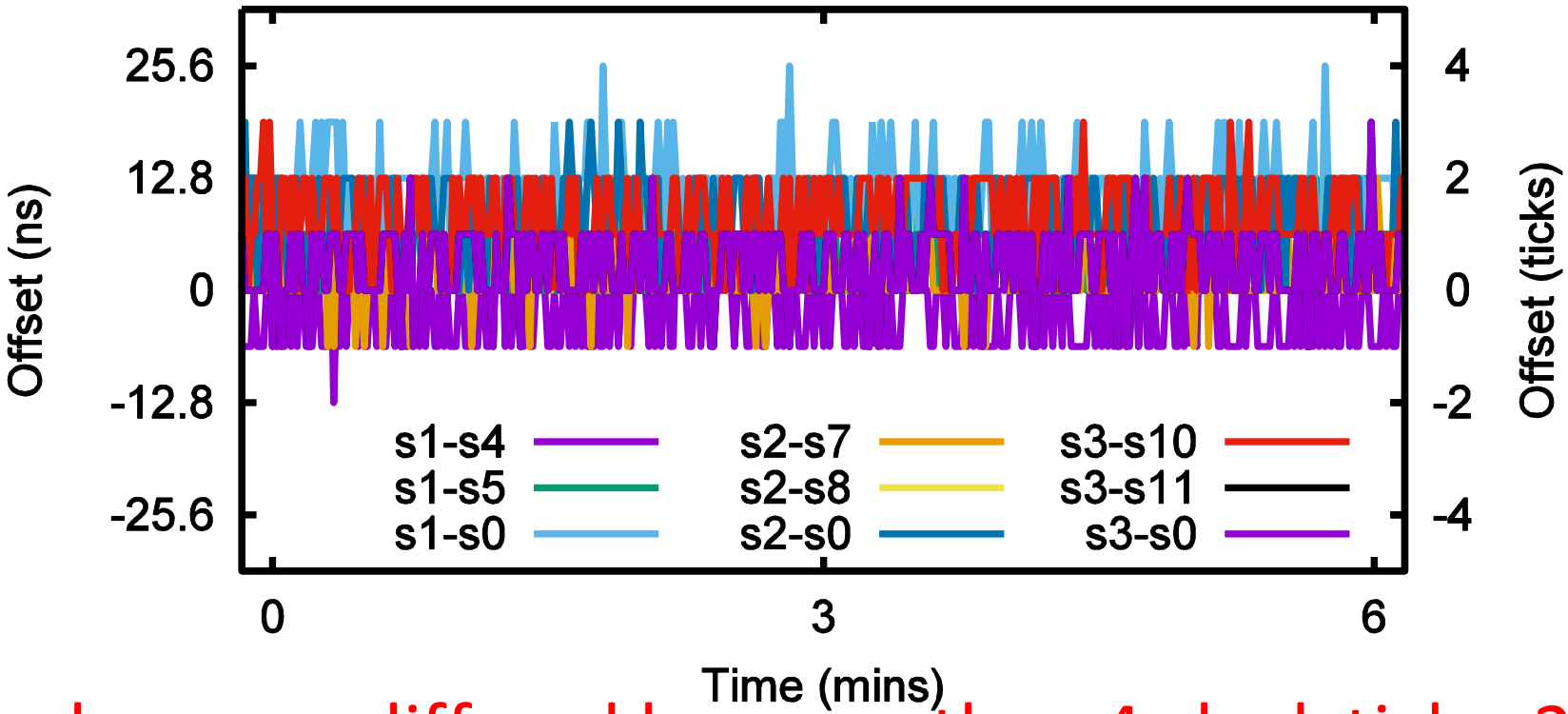
Clocks differ by tens of ***micro***seconds

PTP – Heavily Loaded Network (9Gbps Traffic)



Clocks differ by ***hundreds of micro***seconds

DTP – Heavily Loaded Network (9Gbps Traffic)



Clocks ***never*** differed by more than 4 clock ticks, 25ns

Bounded Precision

Outline

- Introduction
- Design
- Evaluation
- Conclusion

Conclusion

DTP provides

- Bounded precision: 4 oscillator ticks (25ns)
- Scalability: 150ns for entire datacenter
- Free – No network traffic: Use the PHY!
- Needs hardware modifications (just like PTP)

Thank you