Analysis of 10GbE using Hardware Engine for Performance Tuning on LFN

T. YOSHINO, J. TAMATSUKURI, K. INAGAMI,
Y. SUGAWARA, M. INABA, K. HIRAKI

Data Reservoir Project, The University of Tokyo, Japan

Feb/08/2007

Background
Popularization of 10GbE WAN

- State-of-the-art Internet technology
  - 10 Gigabit Ethernet (10GbE)
    - Popularization of
      - Large bandwidth
      - Wide area
      - Packet-exchange (L2)

      Network
  - 10Gbps end-to-end communication
    - World wide grid computing
      - Mass storage synchronization
    - Multimedia distribution

Background
10GbE Debugging Issue

- How can we utilize 10Gbps in end-to-end communication?
  - Technology for large bandwidth and large delay is immature
    - We cannot attain speed by simply replace NIC to 10GbE one
    - Sometimes, throughput becomes lower than 1GbE

      → Cannot be explained clearly
    - We should clarify these problem by precise measurement
    - We need an analyzer for 10GbE era

TCP/IP on LFN (FLD)

- Rate suppressing by congestion avoidance
  - Packet losses lead to low utilization
    - Large delay → Long time for recovery
- Bursty transfer
  - NIC transmits requested data immediately

    → Burst, idle, burst, idle, … in period of RTT
  - Packing of short frame (like ACK) by switches

→ Heavy load to hosts, Buffer overrun in switches

Precise Timestamp (Raw Behavior on Medium)

- Software-based analysis at end-node cannot analyze raw behavior on medium
  - Medium → NIC → Bus → Kernel (→e.g. tcpdump)
  - The result of analysis includes the effect of buffering and scheduling inside PC
- We want to capture flow at point as near to medium as we can
  - Tag precise timestamp right after MAC layer
  - Analyze pressure to NIC, Bus, OS

➔ We need tapping and Hardware support

Research Overview

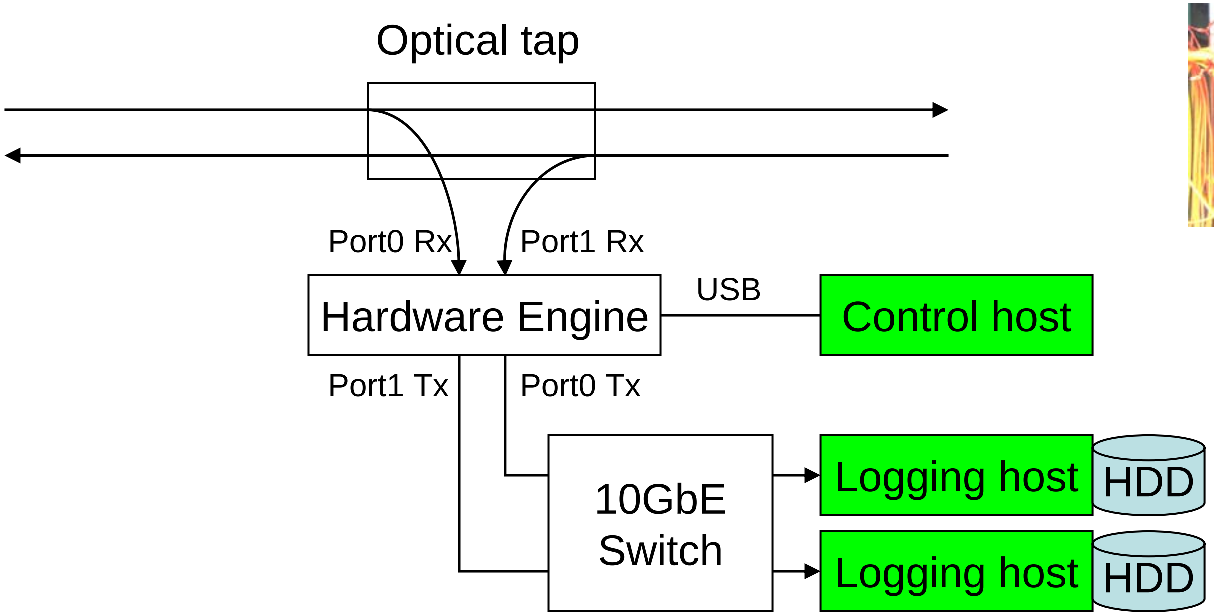
- We developed analysis system for 10GbE
  - Software-hardware coordination

    → We gave commodity PC enough capability
    - Long-time logging
    - Precise timestamp of receiving time
    - Logging of whole headers
    - Flexible analysis on PC

- Evaluated the utility of our equipment
  - Identification of problems of TCP transfer on LFN
  - Analysis of performance and characteristics of our intercontinental 10GbE WAN

TAPEE
(Traffic Analysis Precise Enhancement Engine)

- Duplicate packets by optical tap
- Pre-process packets by hardware engine
- Log and analyze by logging host

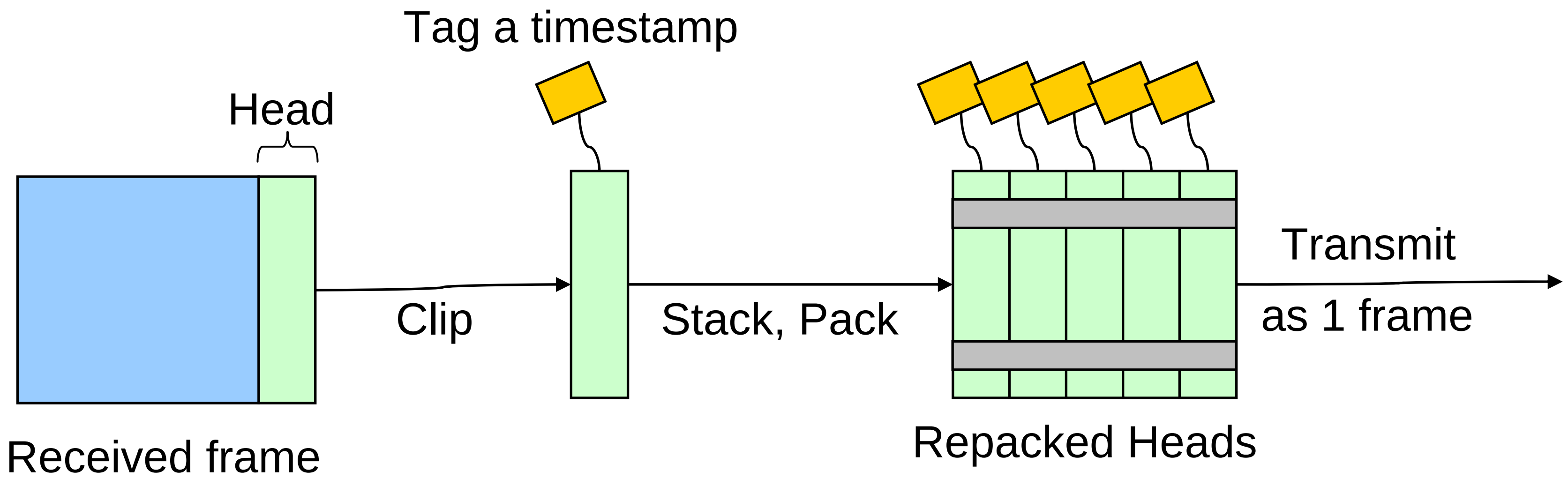


Hardware Engine
TGNLE-1

Logging host
IBM eServer x345

TAPEE's Hardware Engine

- Implemented on our FPGA-based 10GbE testbed *TGNLE-1*
- 3 pre-processing
  - – Clipping → Reduce data input to logging host
  - – Time stamp by hardware → Precise arrival time data
    - 100ns: Enough to distinguish frames in 10GbE
  - – Packing multiple frames → Reduce interruption



Feature of TAPEE

- Long-time logging to HDD
  - We do not have to adjust timing
    - Different from memory-based short-time logging
- Raw header logged to commodity PC
- Sent directly through 10GbE
  - We can start analysis immediately
  - Programmability and flexibility

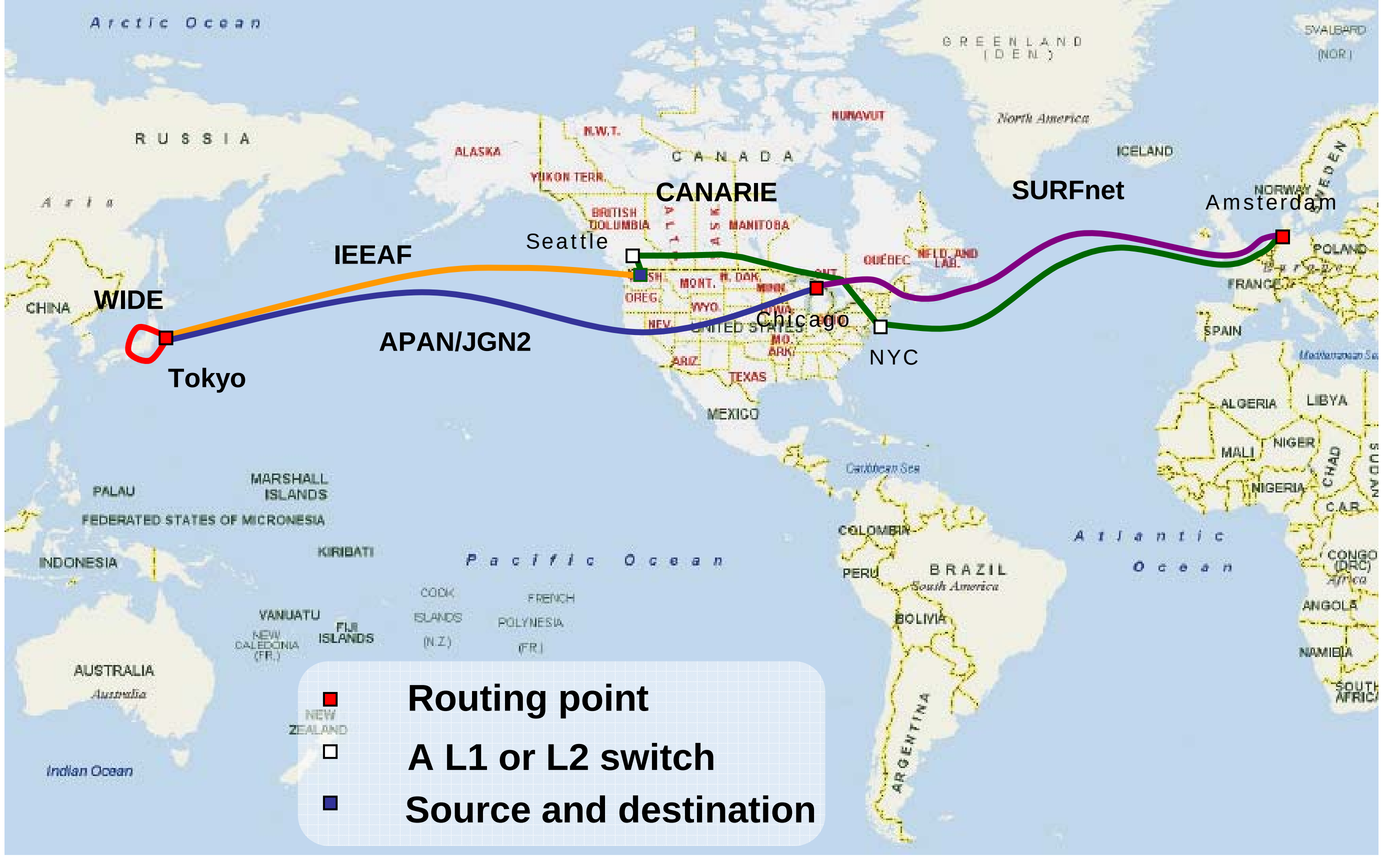
- Programmable hardware
  - Easy to extend
    - Stream embedded into TCP/IP
      e.g. iSCSI

Evaluation

- We performed analysis of TCP transfer on LFN

- We want to show two results
  - Behavior of TCP transfer on LFN in microscopic view
    - Difference between…
      - IPv4 and IPv6
      - W/ and w/o hardware support
  - Comparison of the effect of pseudo-LFN and real-LFN

Real-LFN with 508ms RTT
Round the World Circuit



Network used for LSR Challenge
< 30000km

Pseudo-LFN with 500ms RTT
LFN In Lab by Network Emulator

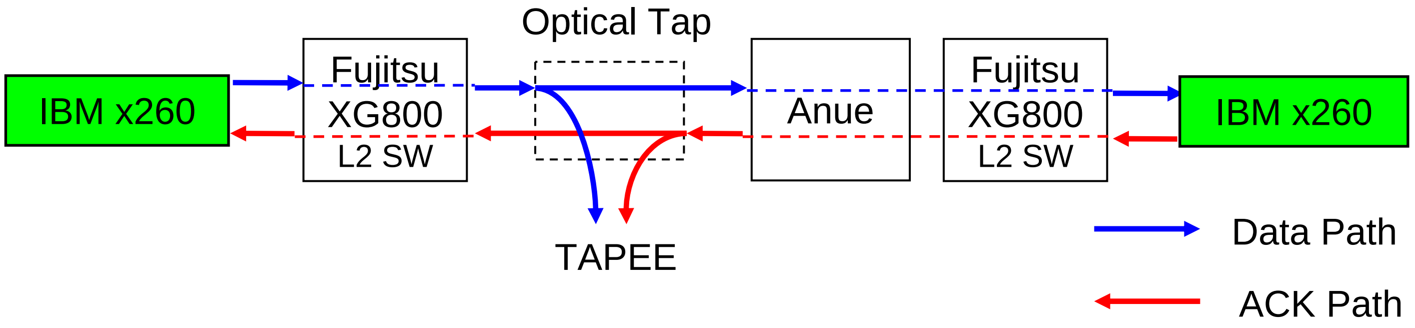
- Analyzing only effect of delay
- Virtually inserts delay
  - Store in DRAMs and forward after X ms.

Anue H Series Network Emulator

[EXP #1] Observing Burstness of Data Packets

- Investigating the effect of large delay by pseudo-LFN
- Compare IPv4 (TOE) and IPv6 (Software) performance

- Transfer on pseudo-LFN
  - Simple single stream TCP/IPv4 and TCP/IPv6
    - Memory to memory transfer using Iperf 2.0.2
  - Chelsio T310 10GbE Adapter on PCI-X 2.0 slot
    - Bus is not bottleneck. CPU bottleneck
  - Pseudo-LFN by Network Emulator
    - Virtually inserts 250ms delay for each channel



Host Spec. & Conf.

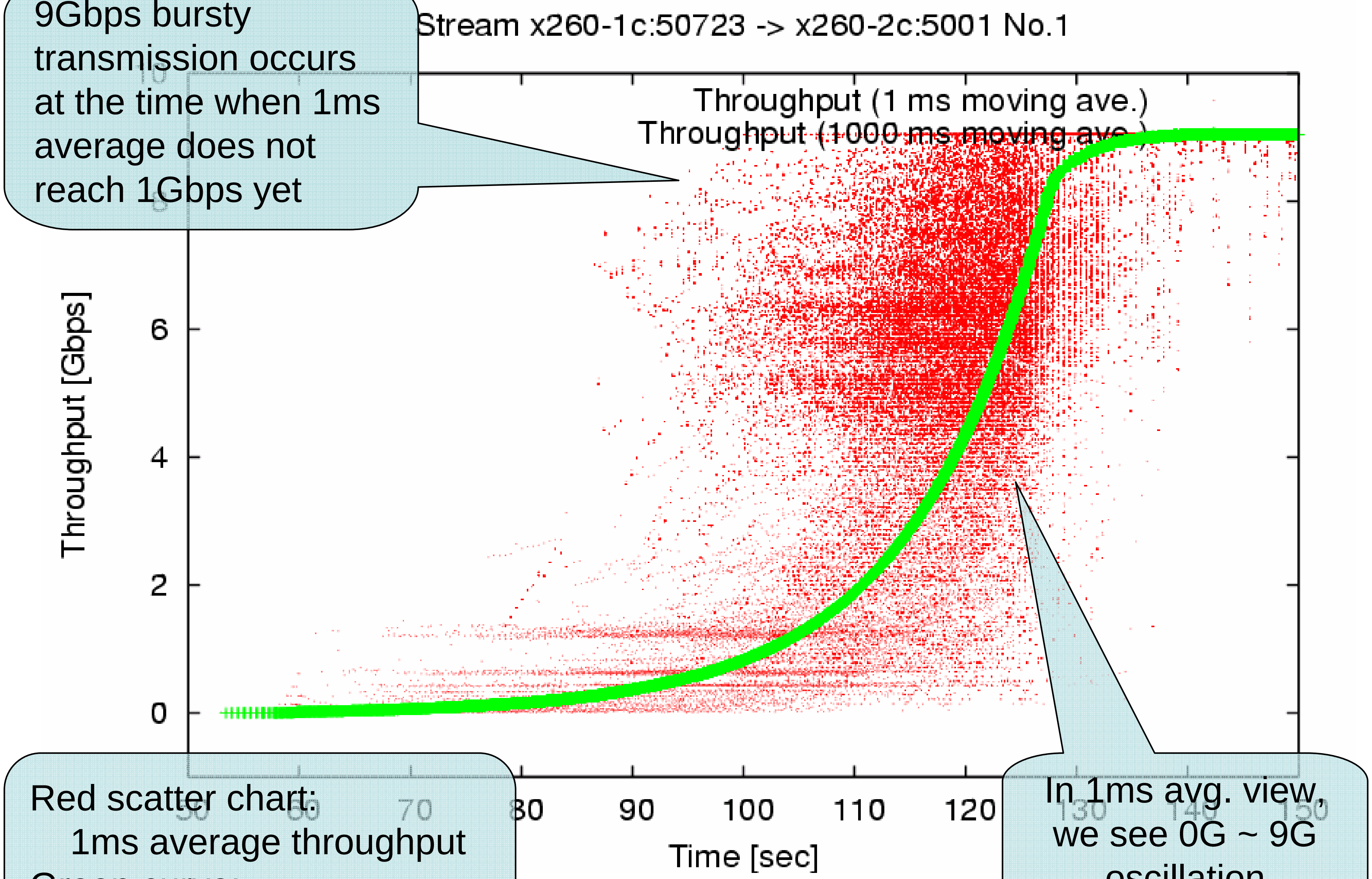
IBM eServer x260

- 9000 octet frame
- BIC-TCP
- Hardware support for TCP
  - For IPv4
    - TOE (TCP Offload Engine) at receiver
  - For IPv6
    - N/A
- Hardware supported packet pacing
  - IPG (Inter-Packet Gap) lower limit is extended to 872 octet

WAN PHY Problem

- WAN PHY bottleneck
  - 9.28Gbps
  - Not avoidable when we use OC-192c based 10GbE WAN
  - Bursty data forwarded from LAN PHY domain (edge network) may collide and be discarded
  - We extended IPG to suppress rate in fine granularity to meet this bottleneck of real-LFN

Transmission rate 1sec average + 1ms average (IPv4)

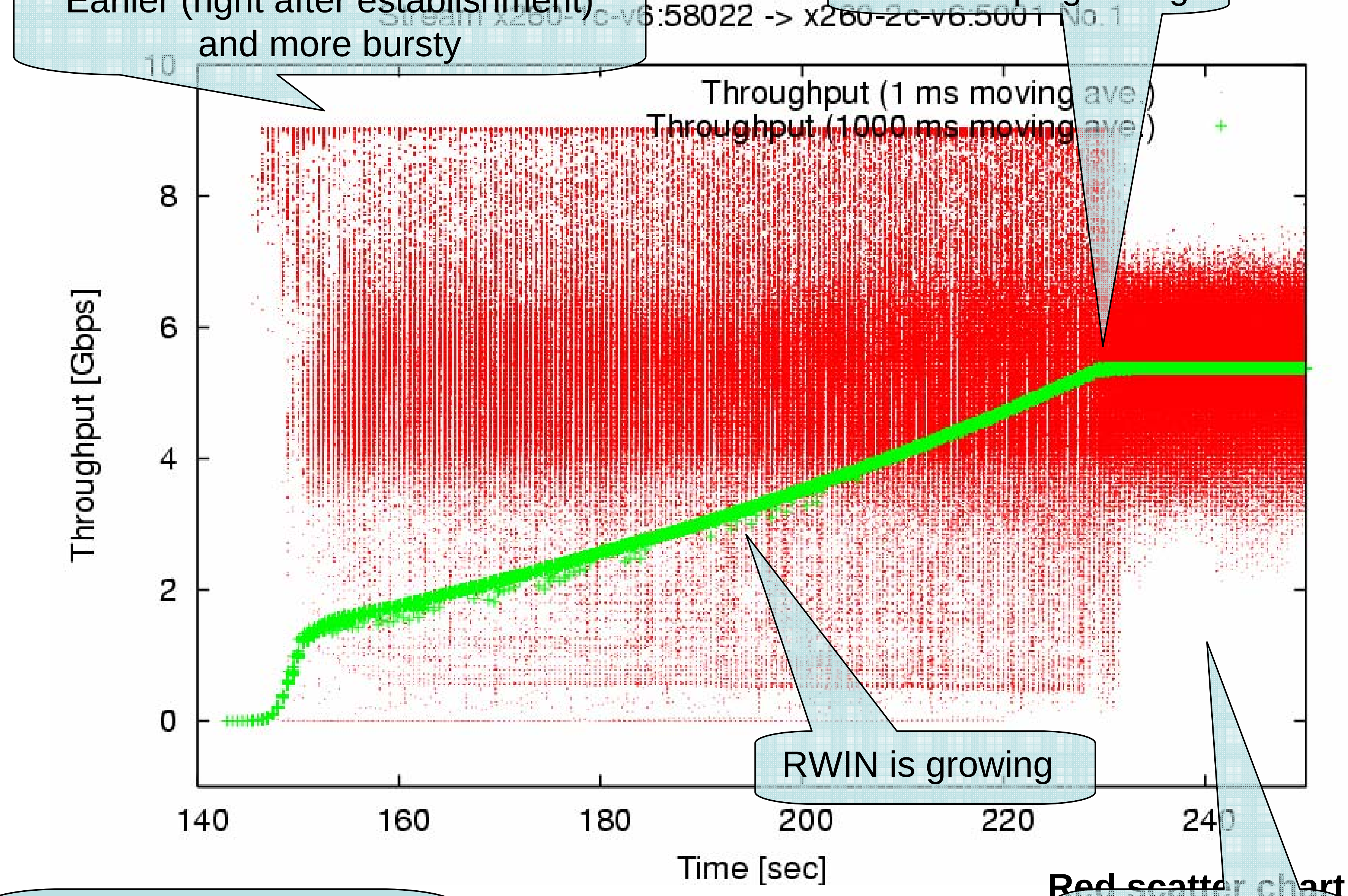


9Gbps bursty transmission occurs at the time when 1ms average does not reach 1Gbps yet

In 1ms avg. view, we see 0G ~ 9G oscillation

Red scatter chart:
  1ms average throughput
Green curve:
  1sec average throughput

1sec average + 1ms average (IPv6)



Earlier (right after establishment) and more bursty

RWIN stops growing

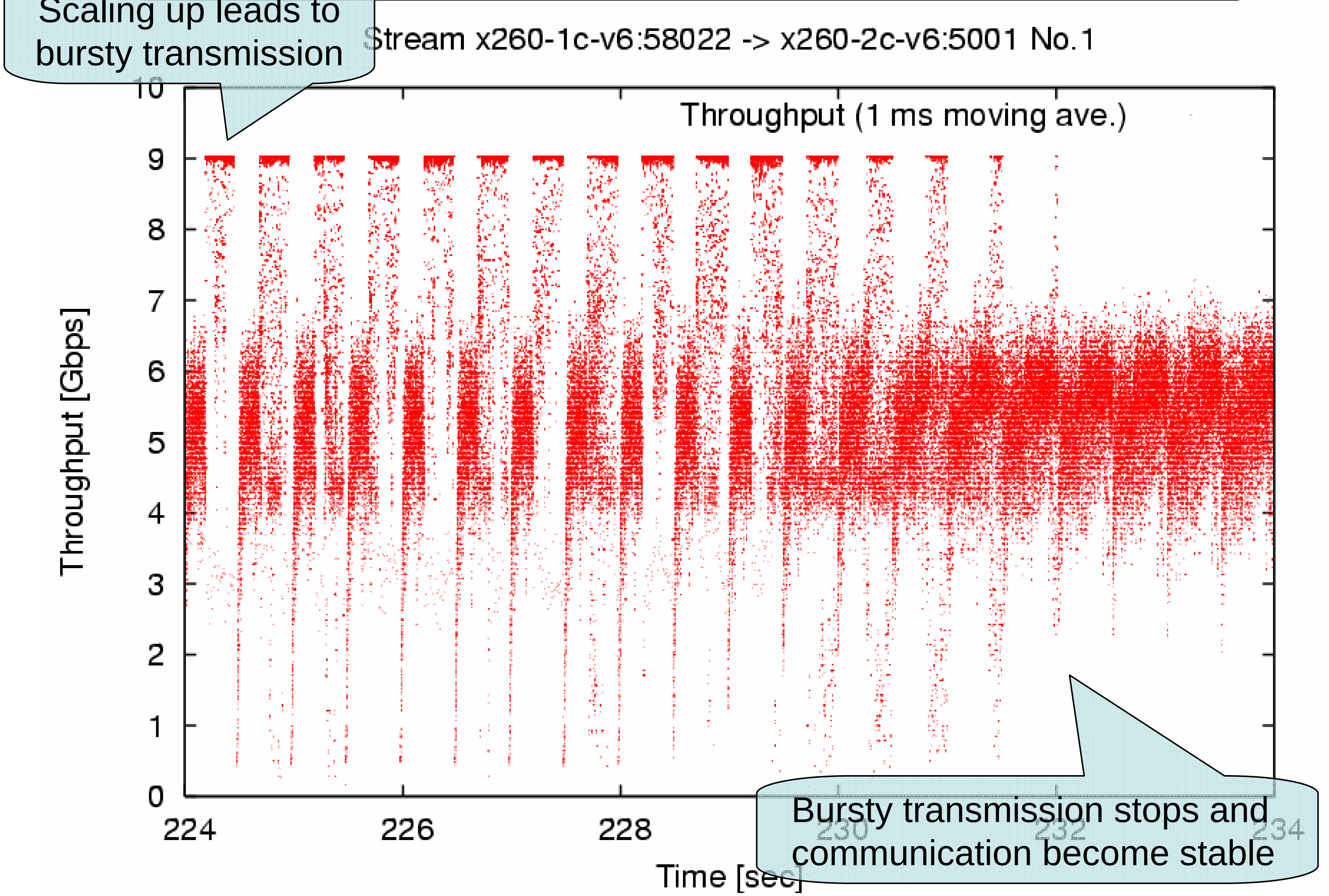
RWIN is growing

After scale up finished, Bursty transfer stopped

Red scatter chart is 1ms ave

CPU power was bottleneck

Magnification of 1ms average at the time scaling finishes (IPv6)



Scaling up leads to bursty transmission

Bursty transmission stops and communication become stable

Discussion: EXP #1

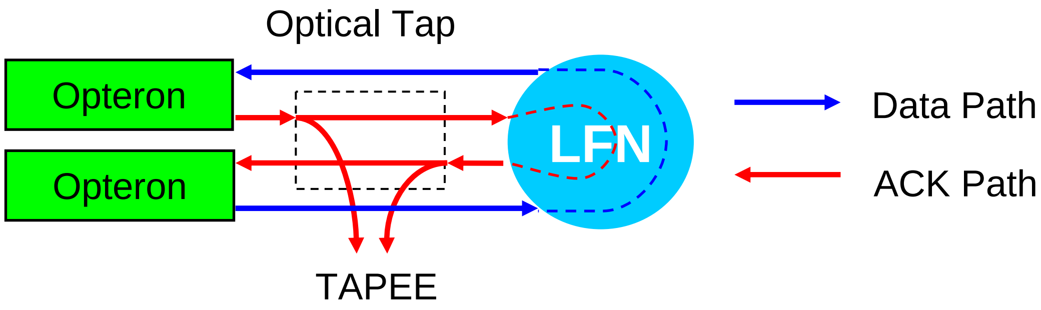
- TOE reduced bursty transfer
  - Offloading assists precise rate suppressing
- Precise pacing is essential for effective congestion avoidance
  - Help congestion point search to work well
  - Not to loss at LAN to WAN conversion point

[EXP #2] Comparison of Real-LFN and Pseudo-LFN

- Clarifying effect of real-LFN by microscopic measurement
  - TCP on real-LFN has lower performance than pseudo-LFN sometimes
  - We have limited chance to perform experiments on real-LFN

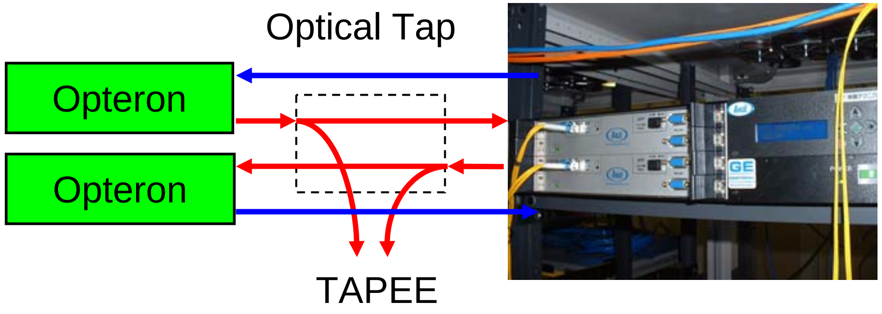
    → Large part of our whole work is done by using pseudo-LFN

    → Clarifying difference between real-LFN and pseudo-LFN is important



Observe and Compare ACK Before and After Passing LFN

Pseudo-LFN



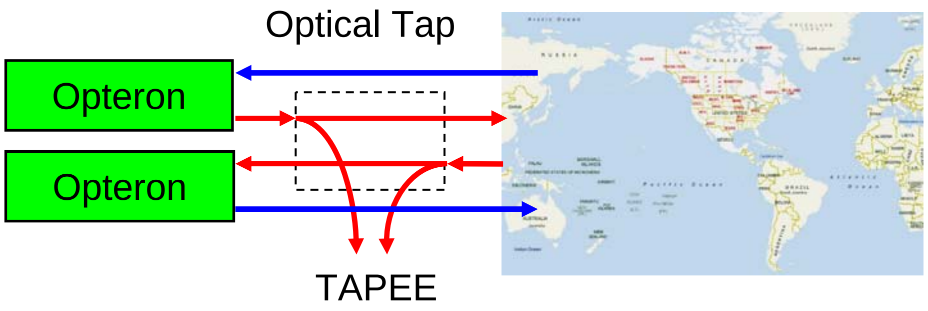
Optical Tap

Opteron

Opteron

TAPEE

Real-LFN



Optical Tap

Opteron

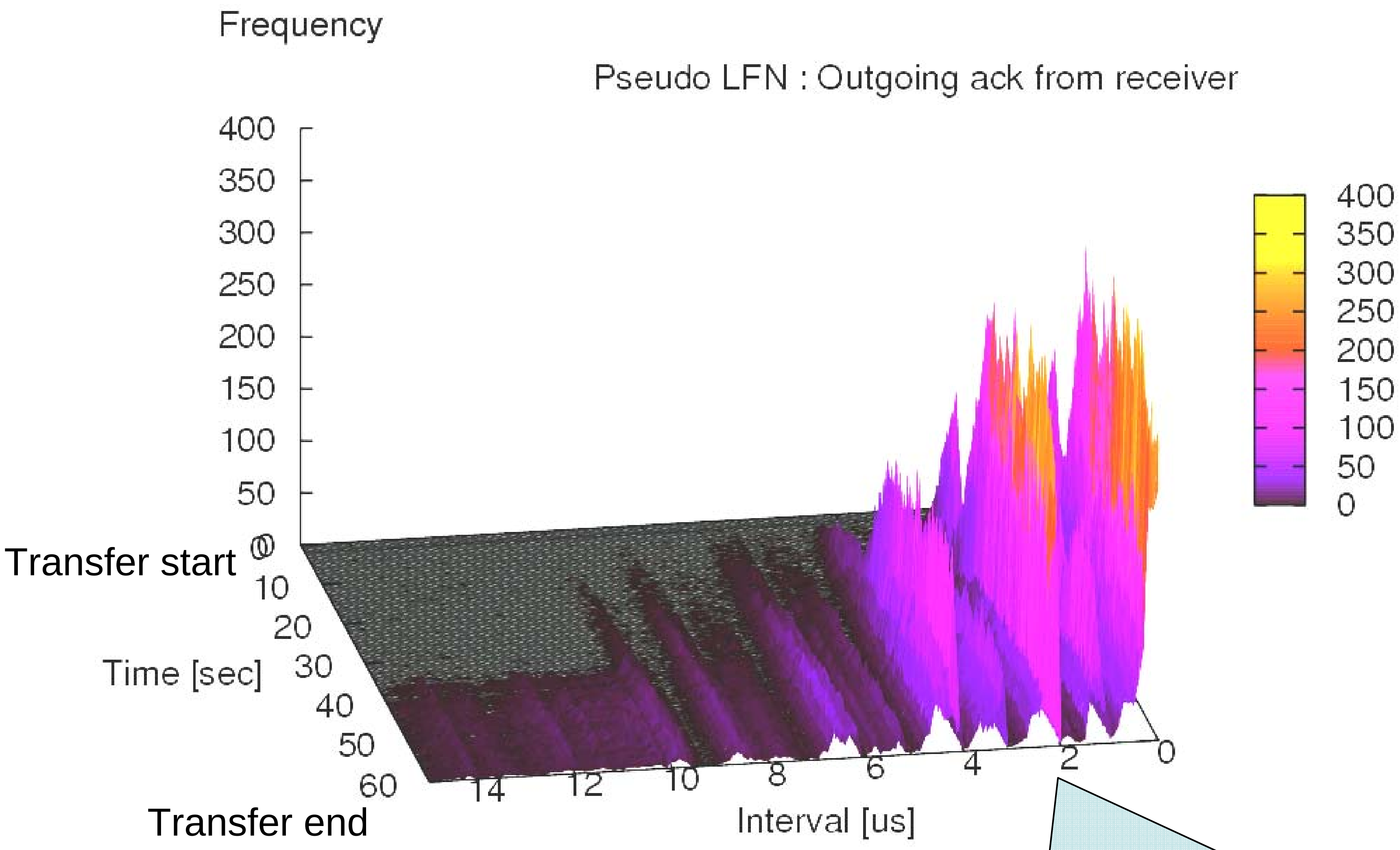
Opteron

TAPEE

→ Data Path

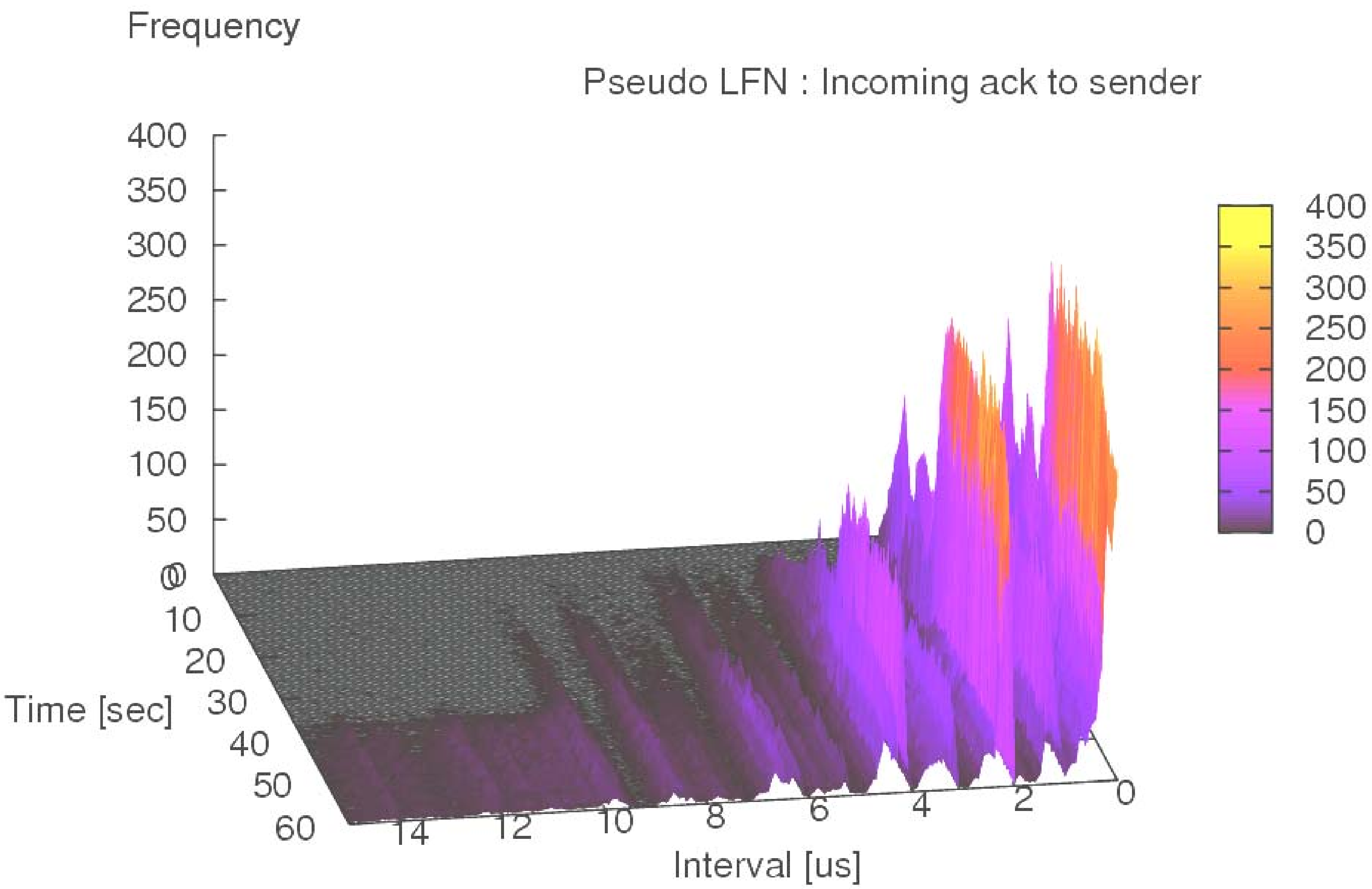
← ACK Path

Packet Interval Histogram of Outgoing ACK from Receiver

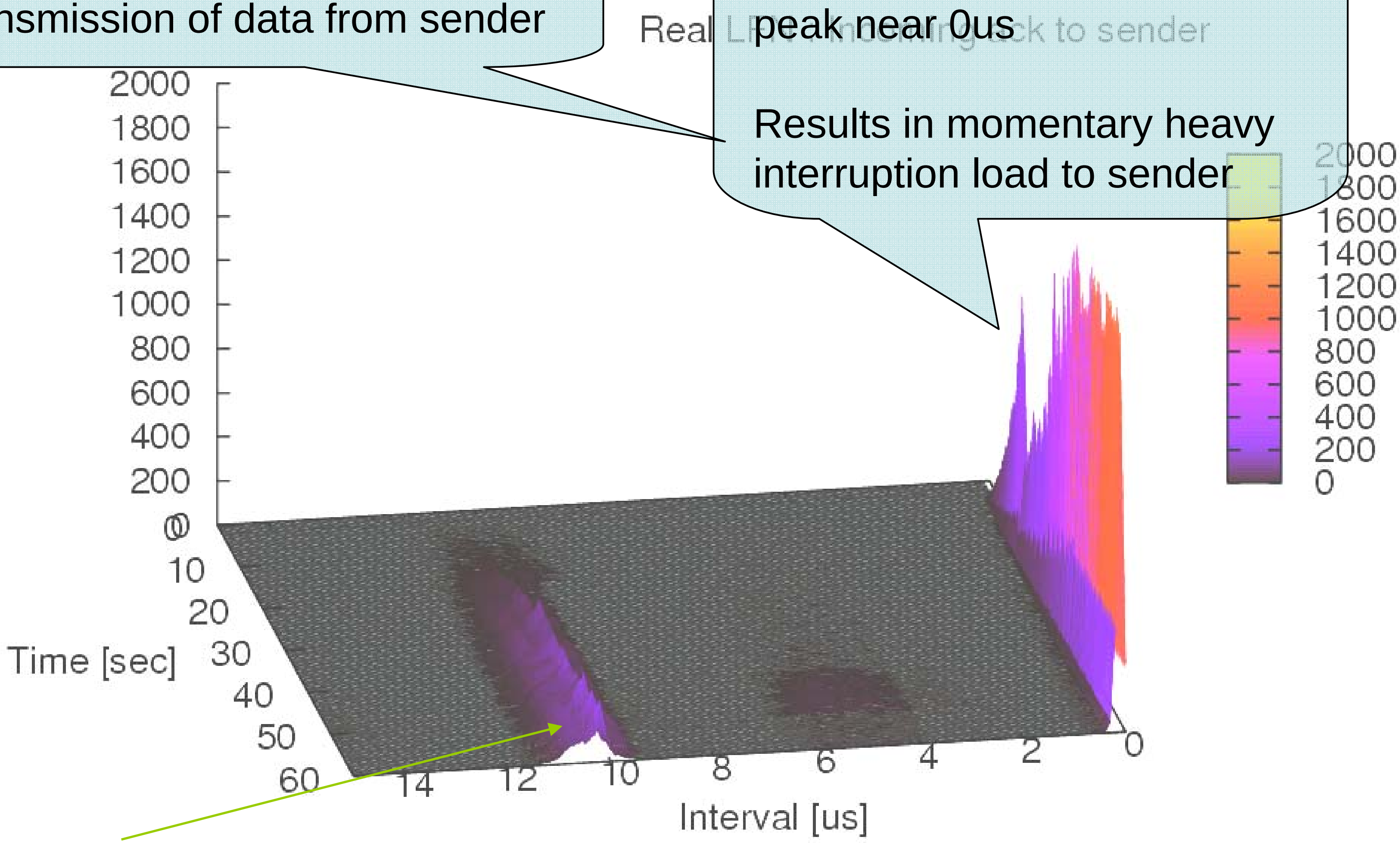


Three peak indicates that most of packets have interval of 0us, 2us, or 4us from the previous packets

Packet Interval Histogram of Incoming ACK to Sender After Passed Pseudo-LFN



Packet Interval Histogram of Incoming ACK to Sender After Passed Real-LFN



Bursty ACK to sender leads bursty transmission of data from sender

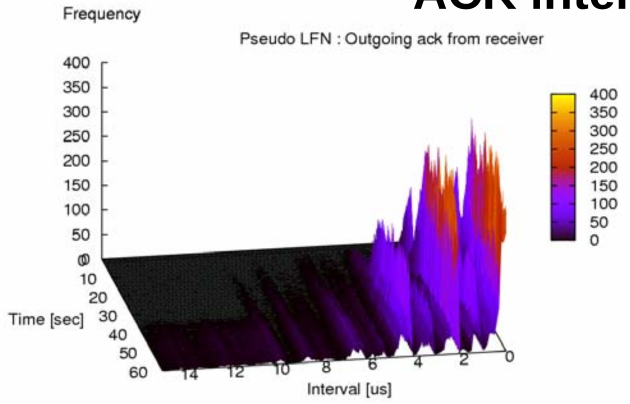
3 peak were merged to be 1 peak near 0us

Results in momentary heavy interruption load to sender

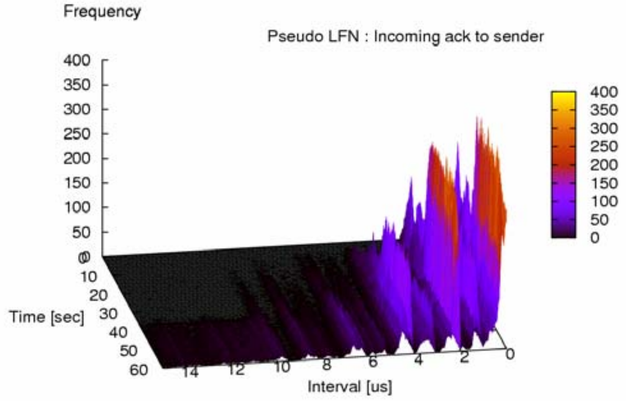
Small peak

Change of Packet Interval

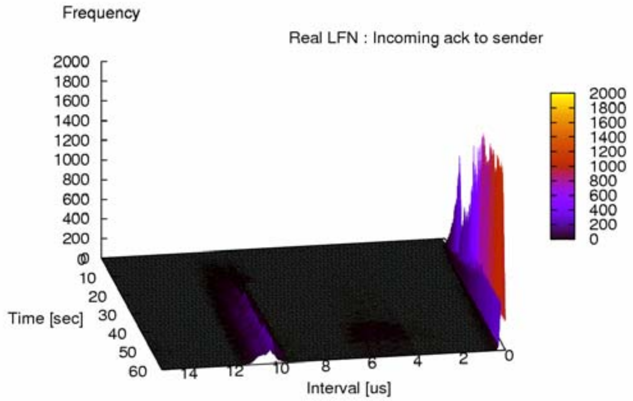
ACK interval at Tx point



After passed pseudo-LFN

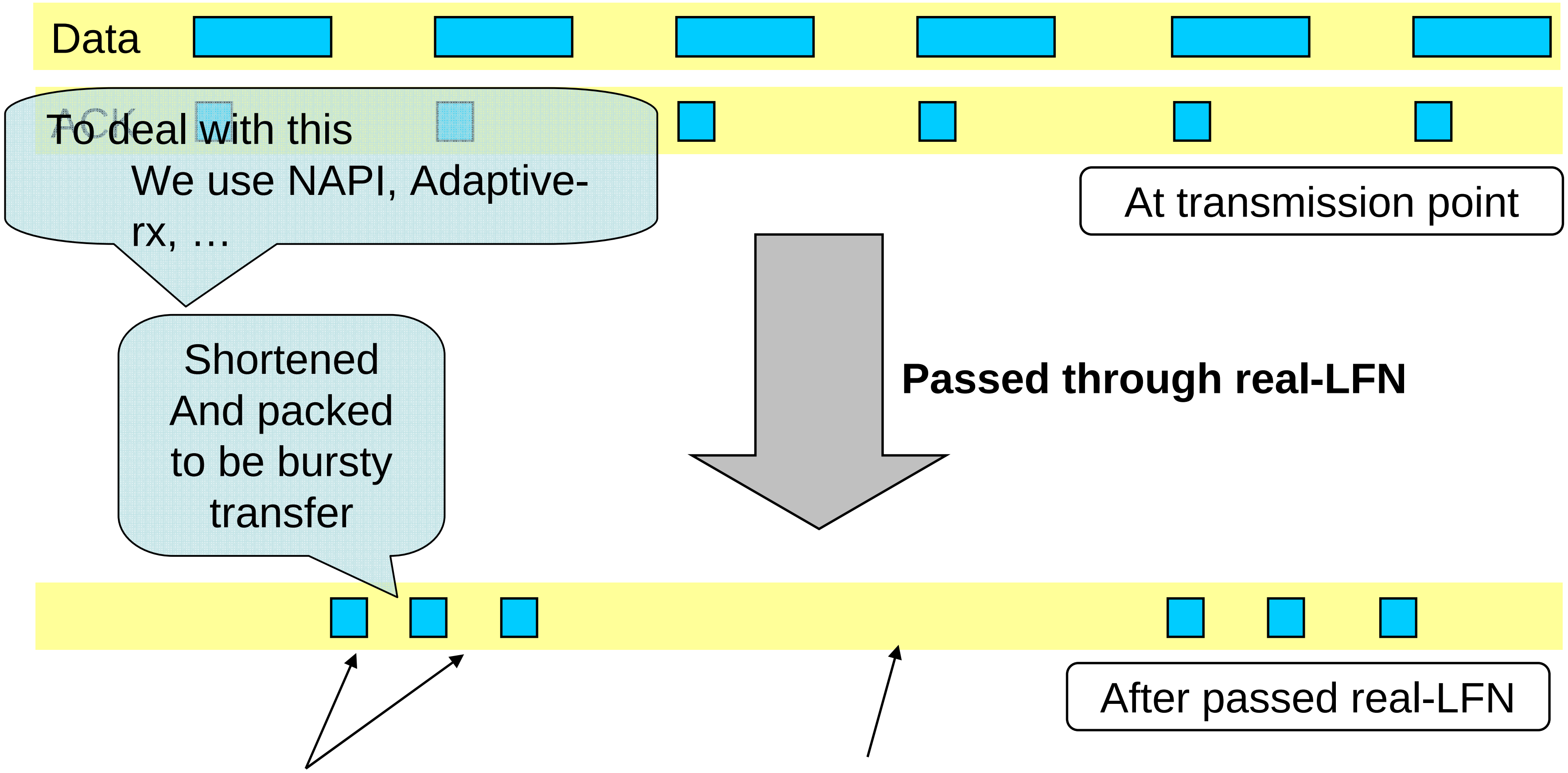


After passed real-LFN



Interval of small packets are shorten

Effect of Real-LFN on Packet Interval



Data

ACK

To deal with this
We use NAPI, Adaptive-rx, …

At transmission point

Passed through real-LFN

Shortened
And packed
to be bursty
transfer

After passed real-LFN

**Peak near 0
in histogram**

**Peak away from 0
in histogram**

Discussion: EXP #2

- Short ACK packets are packet together by some switch in real-LFN
  - It leads to bursty ACK and moreover to bursty DATA
  - ACK pacing has good effect, too

Current Achievement

- Internet2 Land Speed Record (LSR)
  - http://www.internet2.edu/lsr/
  - We are holding records
    - IPv4: 8.80Gbps over about 45 min
    - IPv6: 6.96Gbps over about 30 min
- We are going to update
  - To be continued to next paper

Summary

- We developed long-time precise flexible analysis system for 10GbE
  - Coordination of FPGA-based hardware and commodity PC
- Using our equipment we detected and clarified problems of TCP transfer on LFN
  - The effect of delay in microscopic view
  - Difference between pseudo-LFN and real-LFN
- We have demonstrated the utility for management and development of 10GbE

Acknowledgement

- Experiments supported by
  - Prof. Akira KATO of The Univ. of Tokyo
  - JGN II, IEEAF, WIDE Project
  - Pacific Northwest Gigapop
  - AlaxalA Networks
- Light paths provided by
  - JGN II, WIDE, SURFnet, IEEAF, CANARIE
- Equipments provided by
  - Anue Systems, TOYO Technica
  - PFU, Foundry, Force10, AlaxalA