Linux TCP/IP Performance on Long Fat-pipe Network toward Internet2 Land Speed Record

Junji Tamatsukuri, Takeshi Yoshino,
Yutaka Sugawara, Mary Inaba, Kei Hiraki
Data Reservoir Project / University of Tokyo

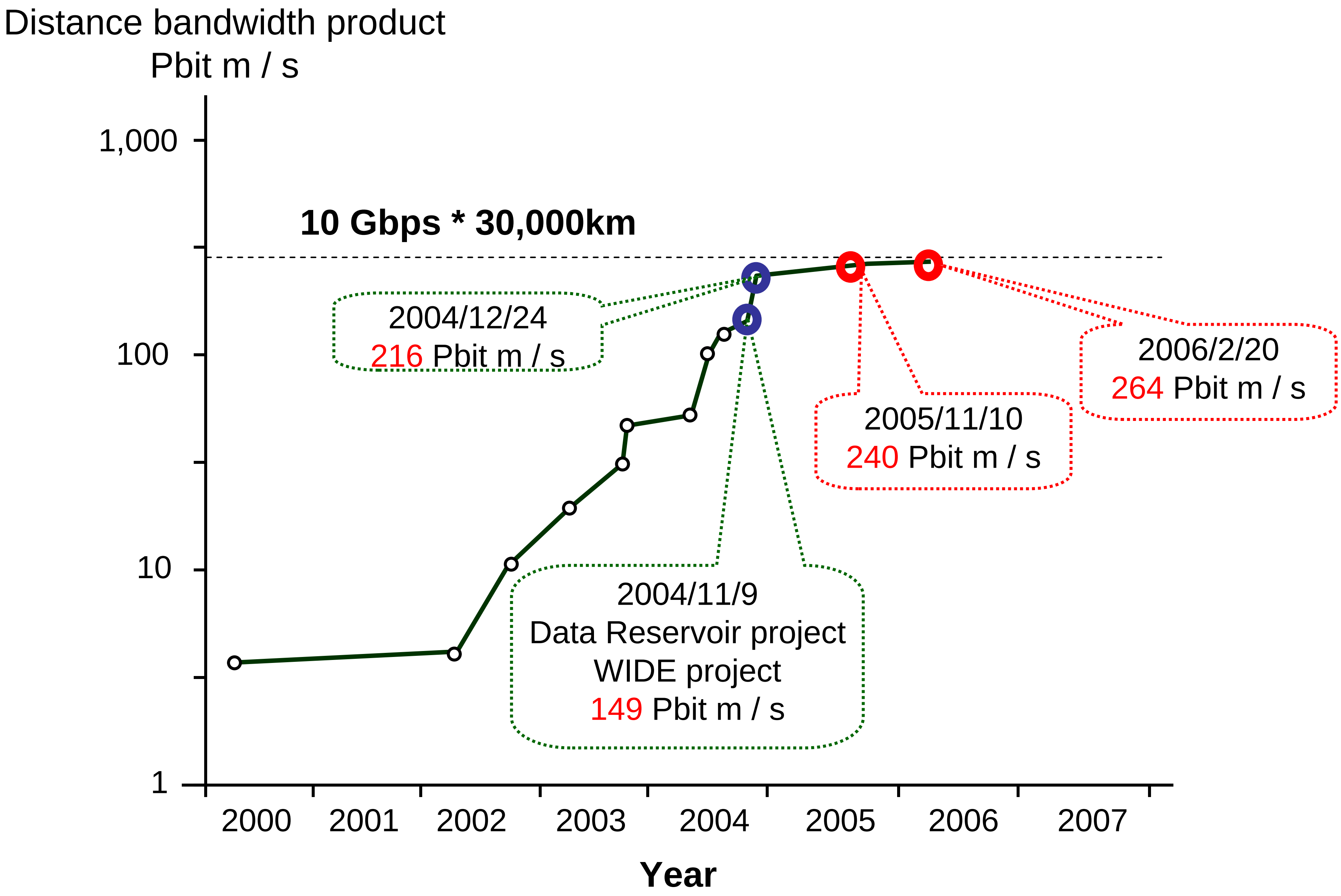
Theme of This presentation

- We present many practical result of the highest performance single stream TCP
  - Linux TCP stacks have essentially high performance rather than other OS.
  - But many problem to get 10Gbps class performace.
    - Many reasons of these problems are unknown now.

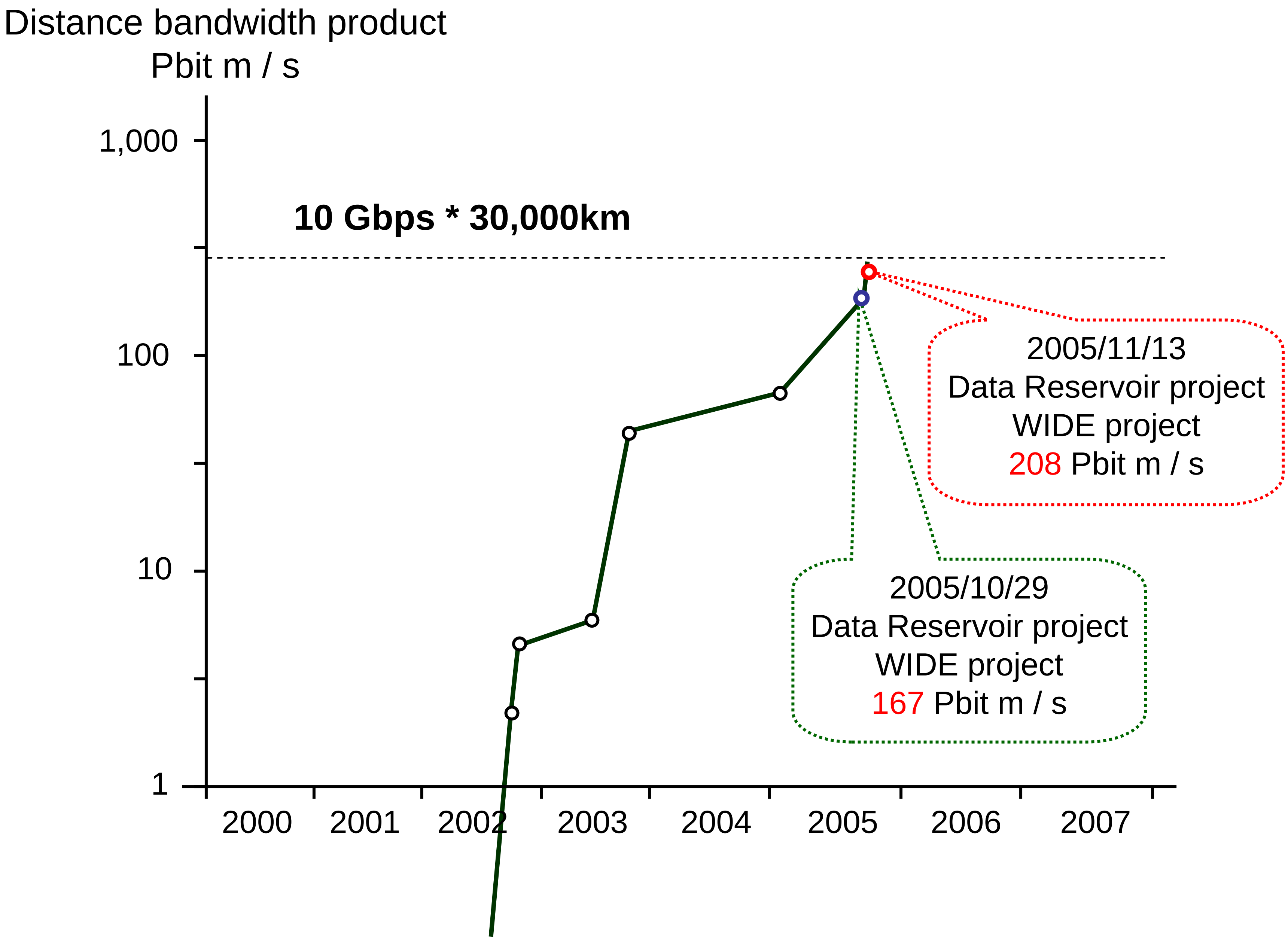
Our Project

- “Data Reservoir” is a Data Grid System for Scientific Data
- Our Goal
  - High performance data site replication between long distance places.
  - We needs high TCP stream performance to realize “Data Reservoir”.

Our Internet2 IPv4 Land Speed Record History



Our Internet2 IPv6 Land Speed Record History



Our TCP/IP result on LFN

- Our project has the most higher experience TCP/IP communication on LFN
- We have 4 points of our tuning approach

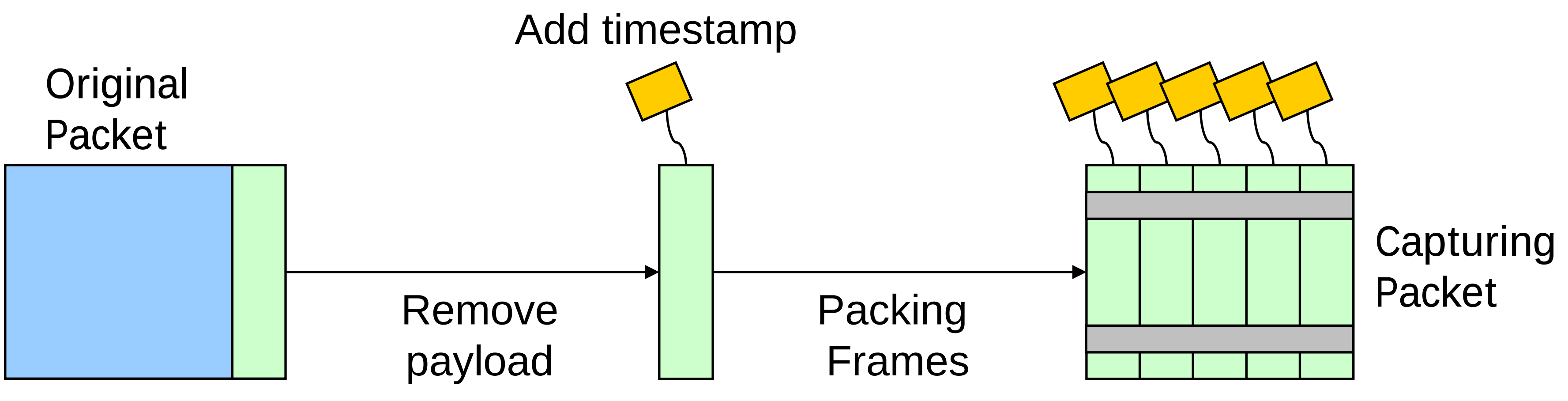
1. Precise logging tools for LFN high speed communication
2. Real LFN over 30,000km and Pseudo LFN environment in our labo.
3. Many result of TCP/IPv4,v6 on both LFN
4. TCP tuning method for LFN

1, hardware logging tool TAPEE

- Packet logging tool with precise timestamp.
  - To analyze TCP stream
  - To view physical layer behavior
- Hardware/Software Solution
  - Packet processing
  - Data capturing/ Data analyzing

1, Function of TAPEE

- Preprocessing by hardware
  - Copy packets by light TAP
  - Remove payload to decrease data size
  - Adding precise timestamps by 100ns
  - Packing Several frames to decrease packet rate.



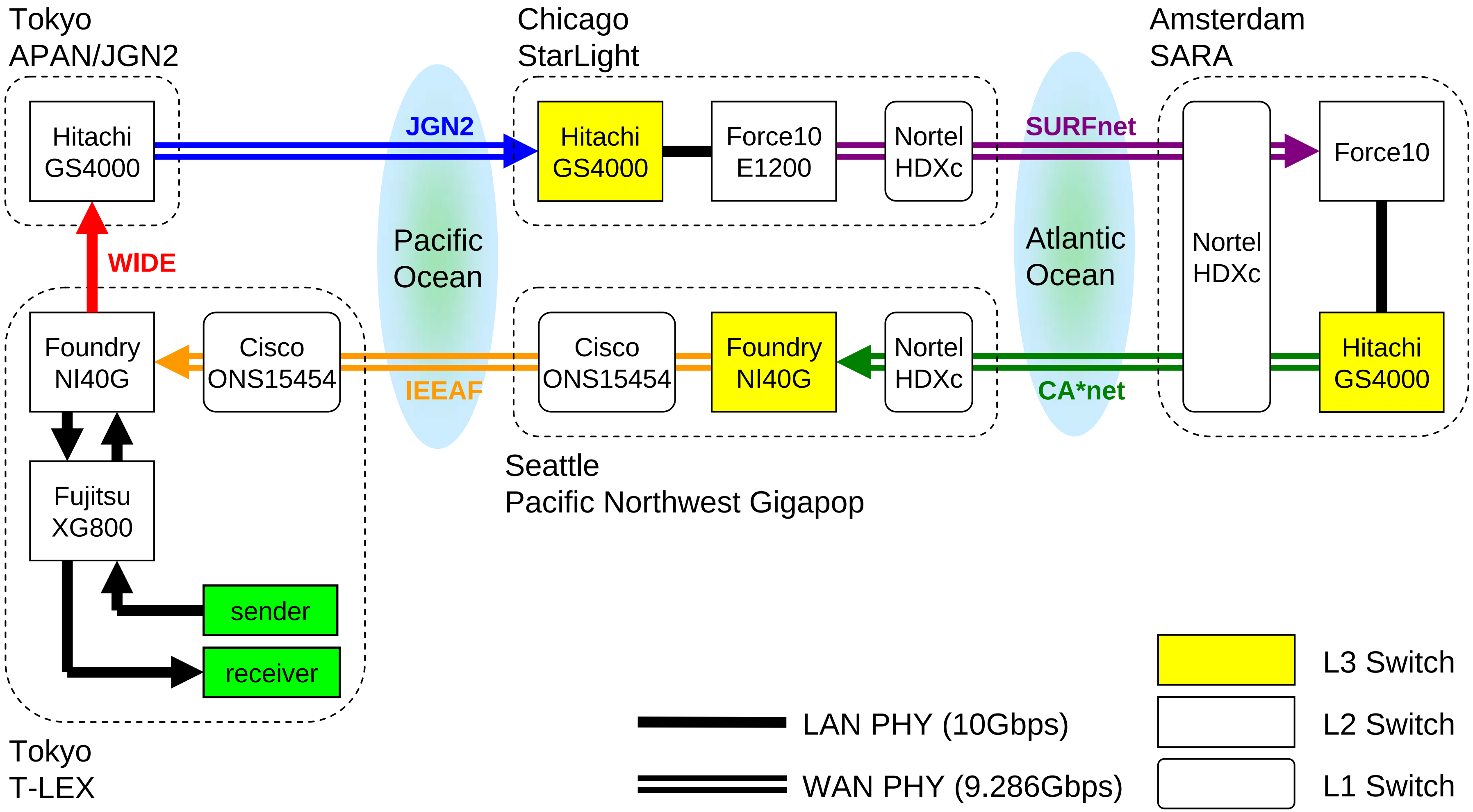
Our Experimental Enviornment

- TCP communication between Linux Servers (Sender → Receiver)
  - Application: iperf-2.0.2
- Servers
  - Opteron / Xeon
- Network
  - Real / Pseudo Network

2, Real Long Fat-pipe Network

- LSR needs 30,000km Network (Our net work is 33,000km)
- Sum of distance among Routing Point
- Oversea Circuit consits of OC-192/SDH
- 10GbE WAN-PHY (9.26Gbps)

2, Our Real LFN Diagram



2, Pseudo LFN Environment

- Insert long latency among servers artifically
- Hardware
  - TGNLE (Our project develop)
    - Upto 1600ms RTT
  - Anue H series Netowork Emulator
    - Upto 800ms RTT
- Test enviornment before Real LFN experiment.

TGNLE-1 (same box of TAPEE)

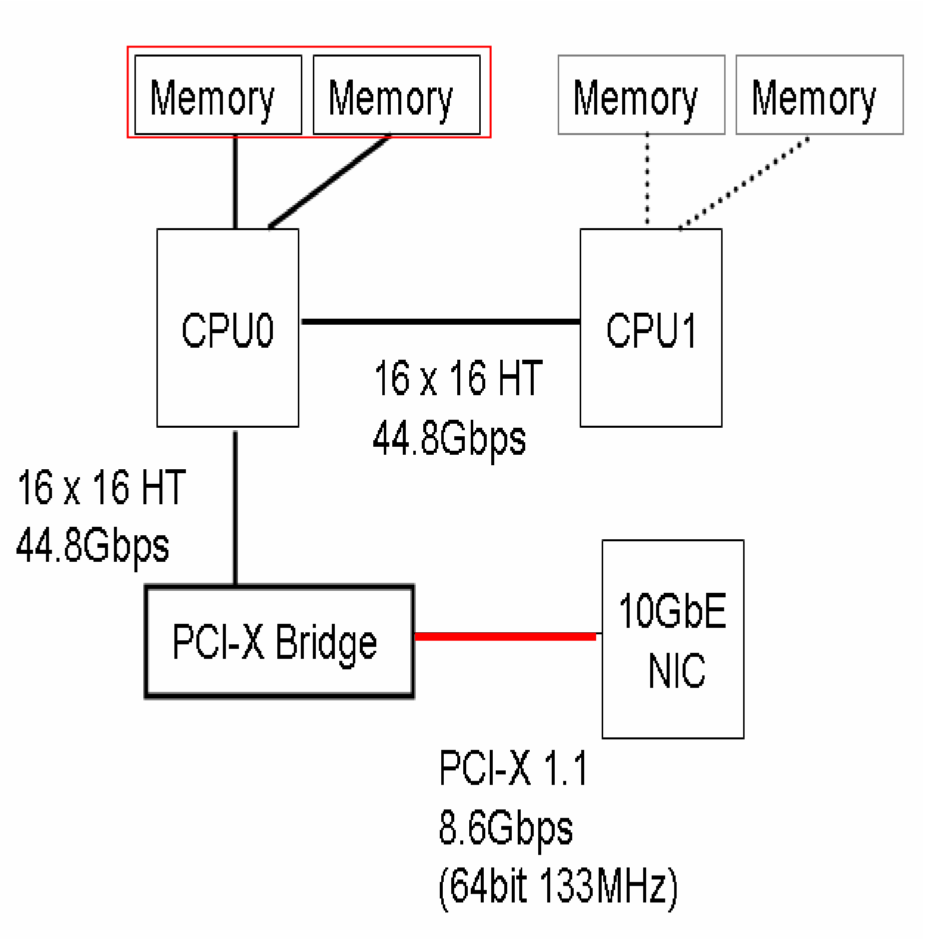
Anue H Series Network Emulator

3, Linux Server Specification Architectural Difference

- PCI-X performance
  - PCI-X 1.0 : Opteron (8.5Gbps)
  - PCI-X 2.0 : Xeon MP (over 10Gbps)
- CPU performance
  - Memory Latency/Bandwidth
    - Opteron With Memory controller
    - Xeon without Memory controller
- Interrupts to CPUs

3, Hardware 1 : Opteron

- Processor: Dual Opteron 250 (2.6GHz)
- MotherBoard: Rioworks HDAMA
- Memory: 2GB (Overclock DDR CL2)
- I/O Performance limitation : PCI-X 1.0 8.6Gbps (133MHz x 64bit)

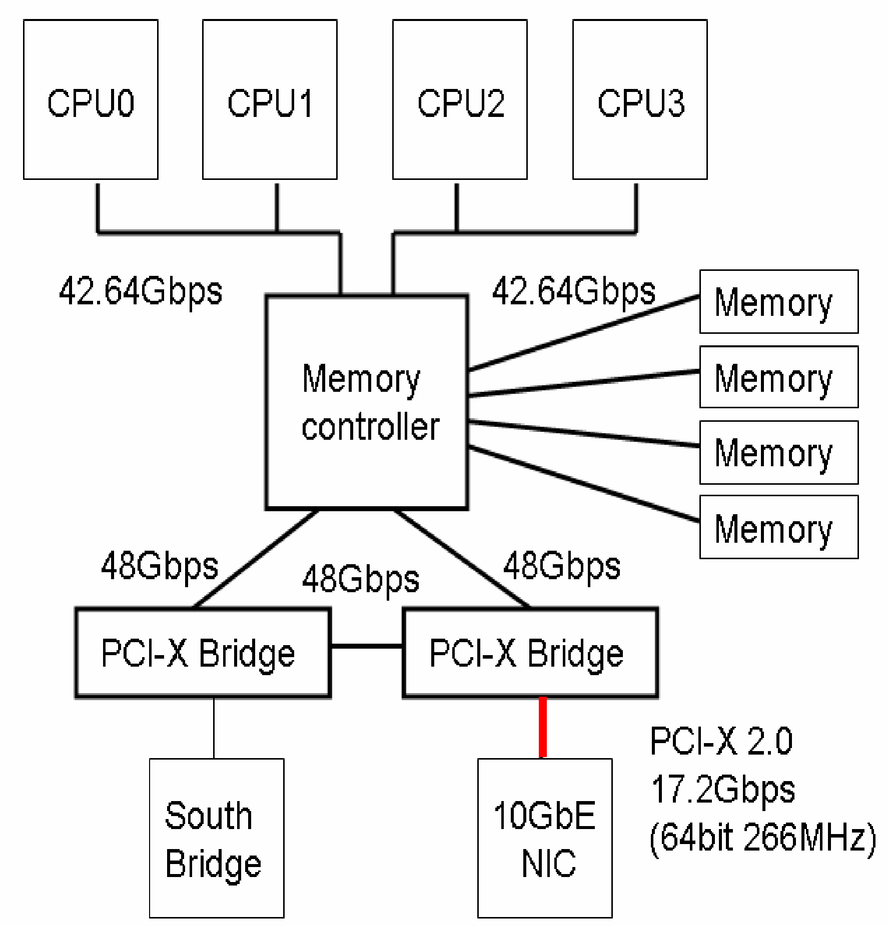


3, Hardware 2 : Xeon MP

- Processor: Quad Xeon MP 3.66GHz (IBM x260)
- Memory：

32GB (DDR2 x 4bank)

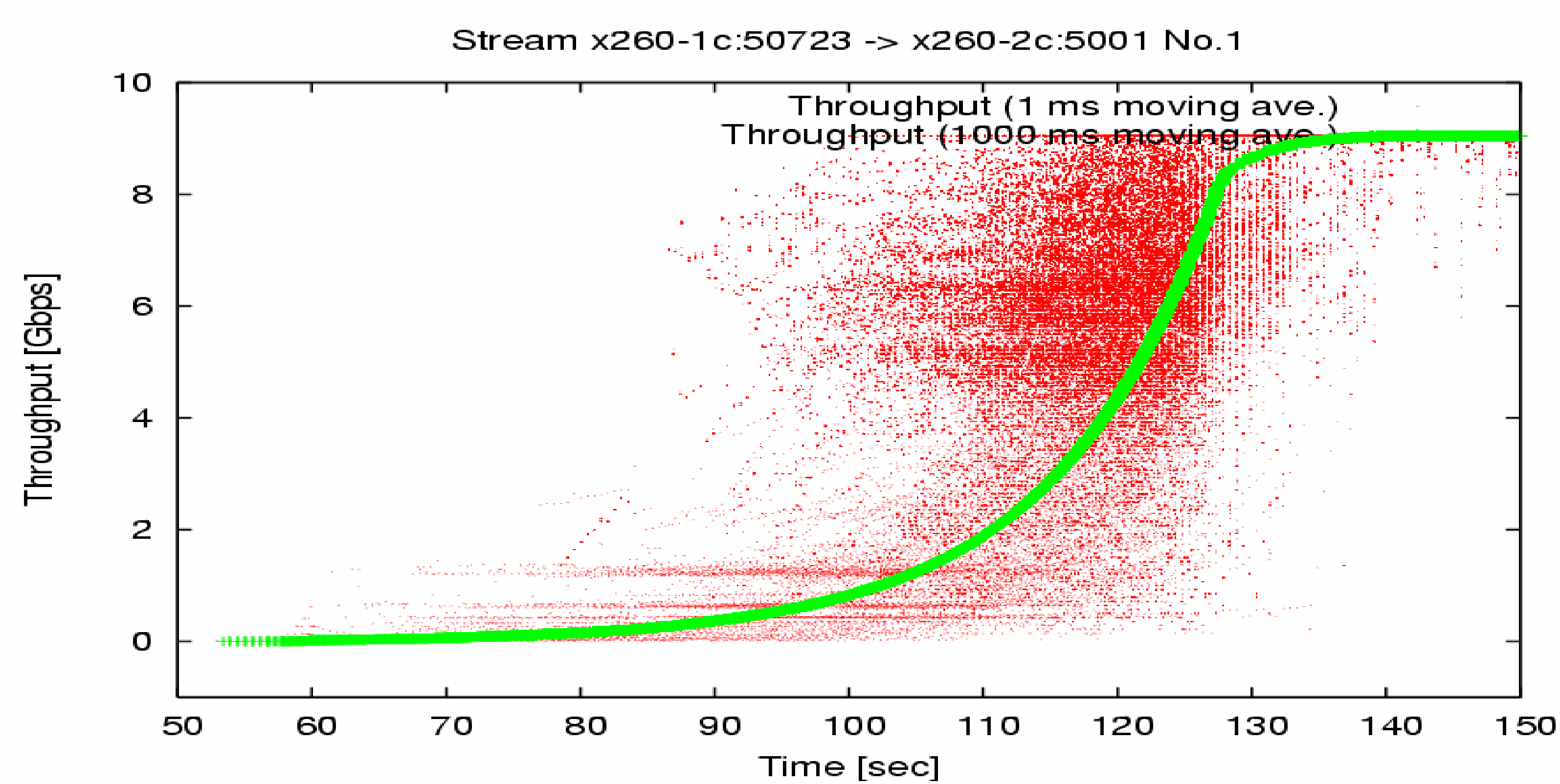
- No I/O Performance limitation : PCI-X 2.0(266MHz x 64bit)



TCP/IP performance matrix

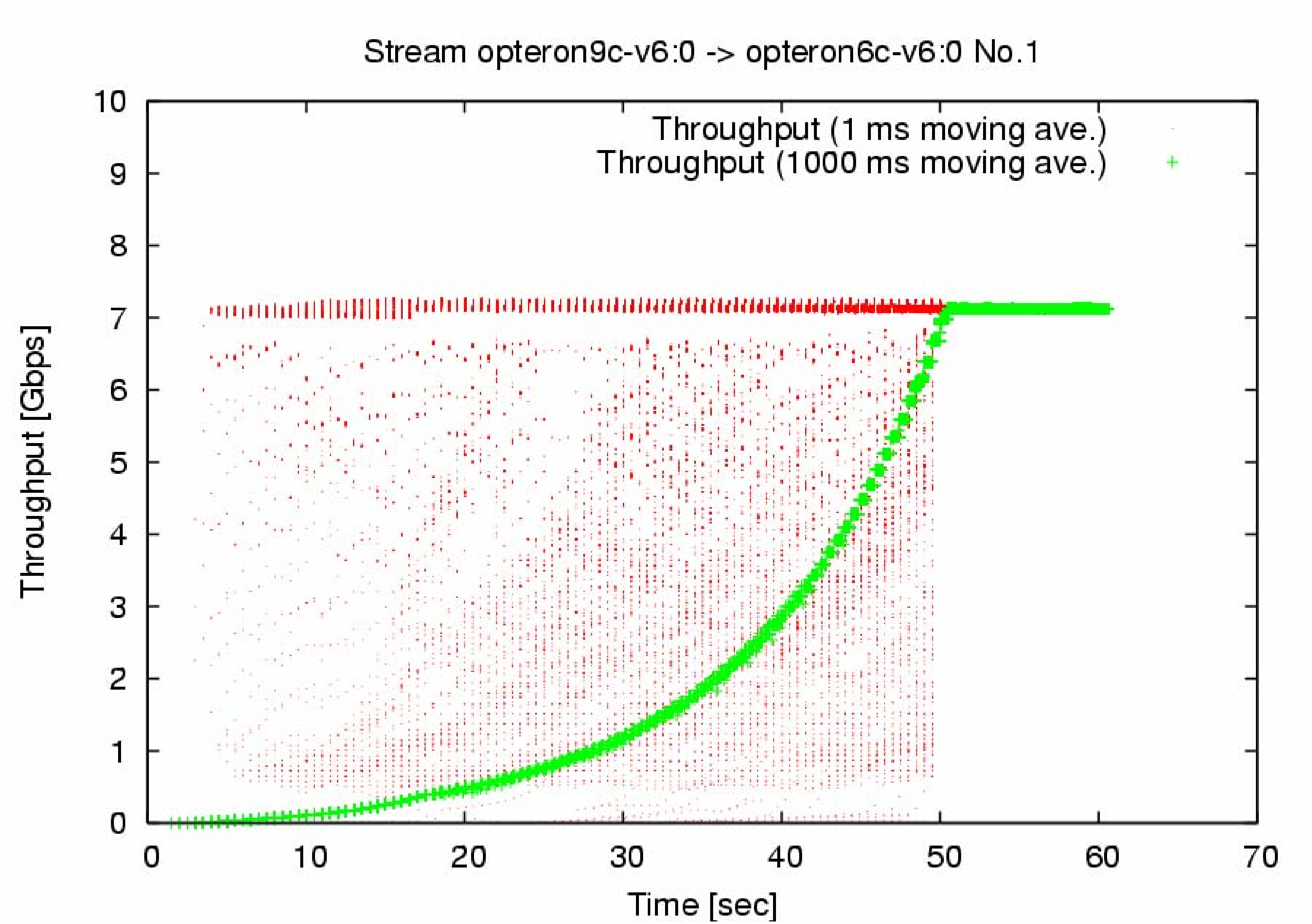
BIC TCP			2.6.12 (chelsio driver)		2.6.17		2.6.18-rc5	
			IPv4	IPv6	IPv4	IPv6	IPv4	IPv6
8G limit	Chelsio T110	opteron	7.2G (90%)	5.9G (75%)	x	x	x	x
	Chelsio N210	opteron	x	x	7.0G (85%)	7.0G (85%)	x	x
		Xeon	x	x	5.75G (75%)	5.75G (75%)	5.43G (65%)	1Gbps (12%)
10G limit	Chelsio T310	Xeon	9.0G (90%)	5.43G (54%)	x	x	x	x

Linux 2.6.12 IPv4 Xeon Performance



- The highest performance stream

Linux 2.6.16 IPv6 Opteron Performance



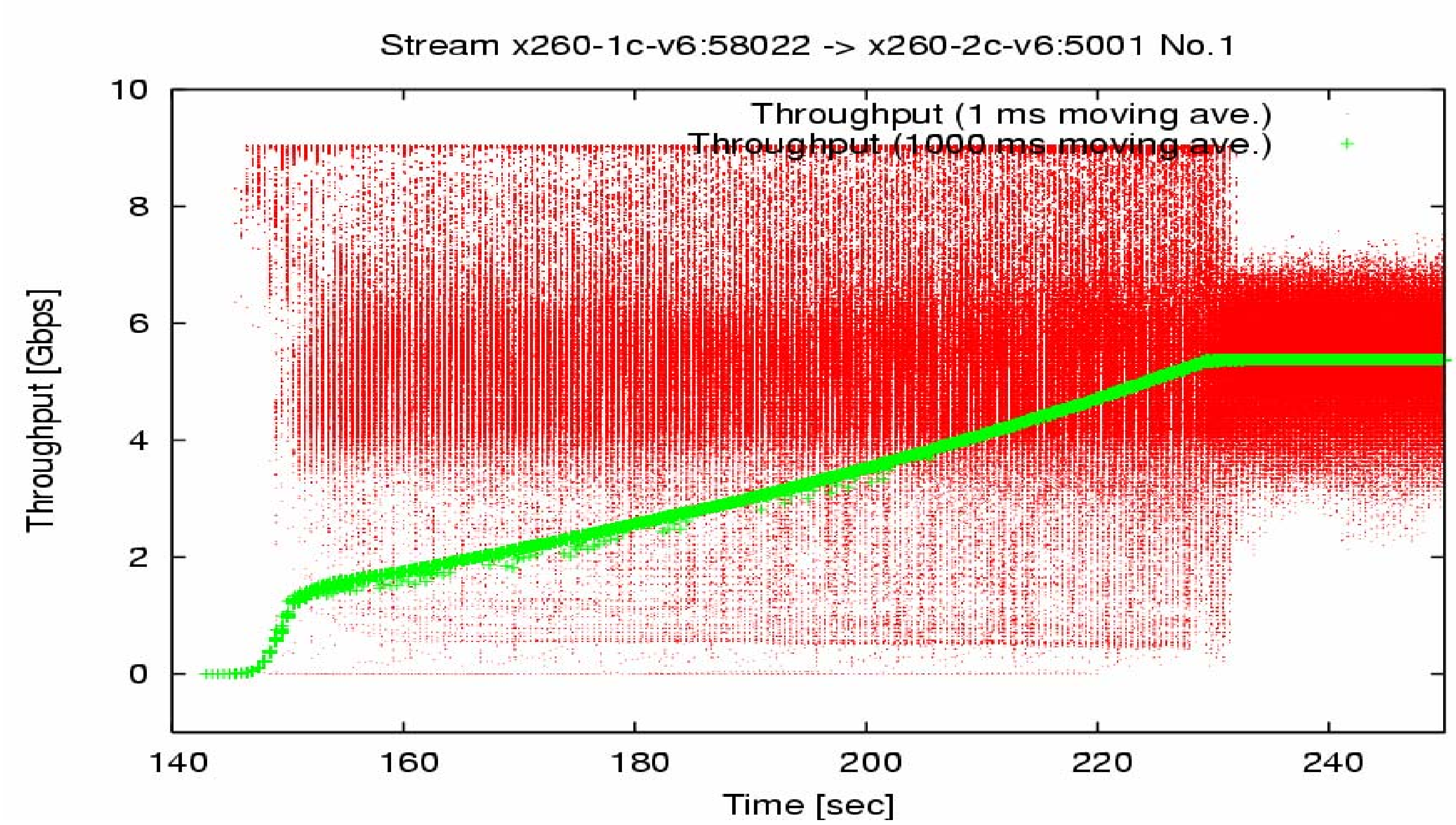
Software TCP performance on Linux 2.6.16 later

- Window Buffer Size
  - Theoretical Value = RTT * Bandwidth
- NAPI
  - Effective for high interrupts from network arrival.
  - We use static optimized interrupt interval.
- TSO
  - Effective for reducing packet checksum calculation
- TCP Scaling
  - Delayed Ack effective for High performance.
  - But longer scaling time is needed.

TCP/IP performance matrix

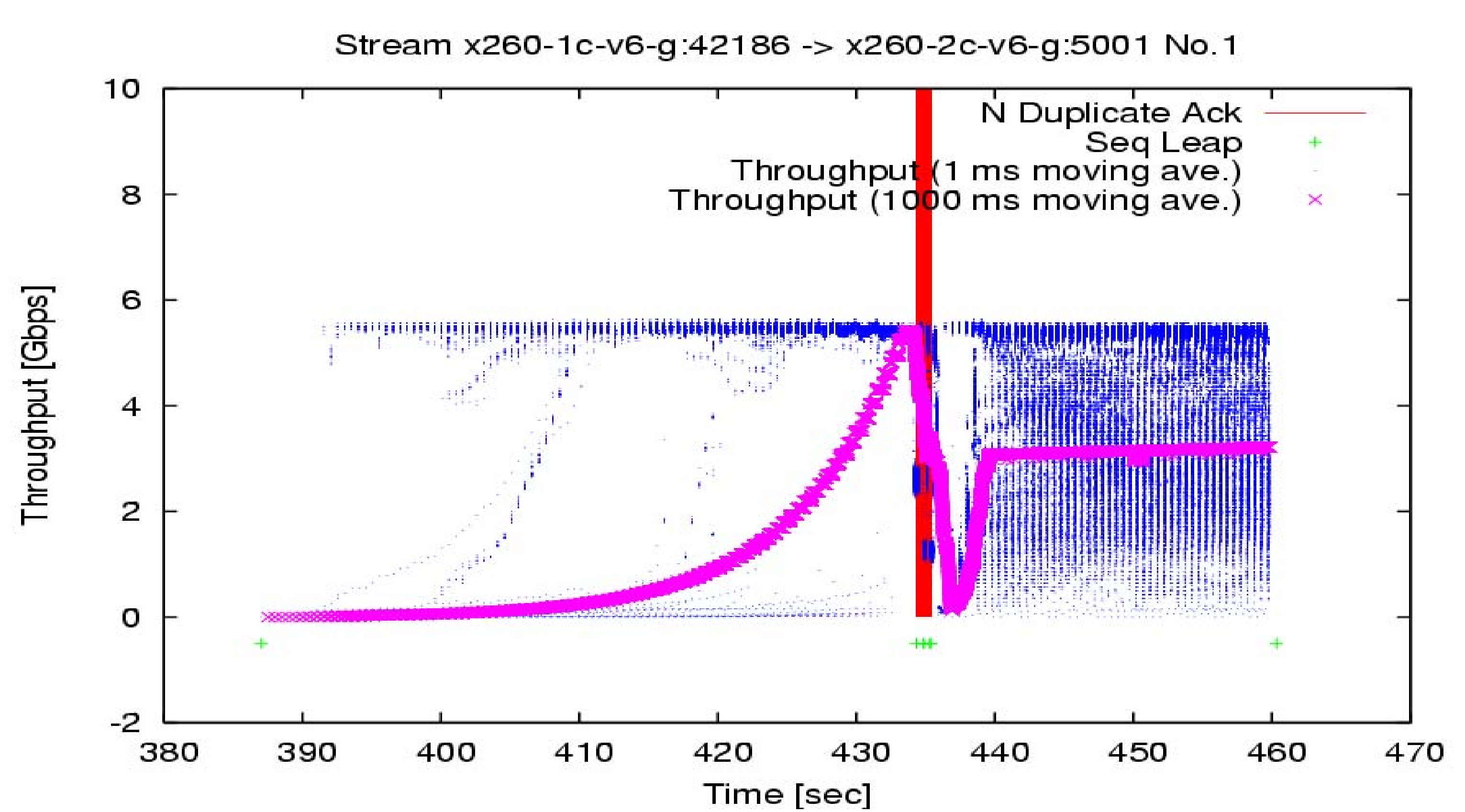
			2.6.12 (chelsio driver)		2.6.17		2.6.18-rc5	
			IPv4	IPv6	IPv4	IPv6	IPv4	IPv6
8G limit	Chelsio T110	opteron	7.2G (90%)	5.9G (75%)	x	x	x	x
	Chelsio N210	opteron	x	x	7.0G (85%)	7.0G (85%)	x	x
		Xeon	x	x	5.75G (75%)	5.75G (75%)	5.43G (65%)	1Gbps (12%)
10G limit	Chelsio T310	Xeon	9.0G (90%)	5.43G (54%)	x	x	x	x

Linux 2.6.12 IPv6 Xeon Performance



- IPv6 result on same host

Linux 2.6.17 IPv6 Xeon Performance

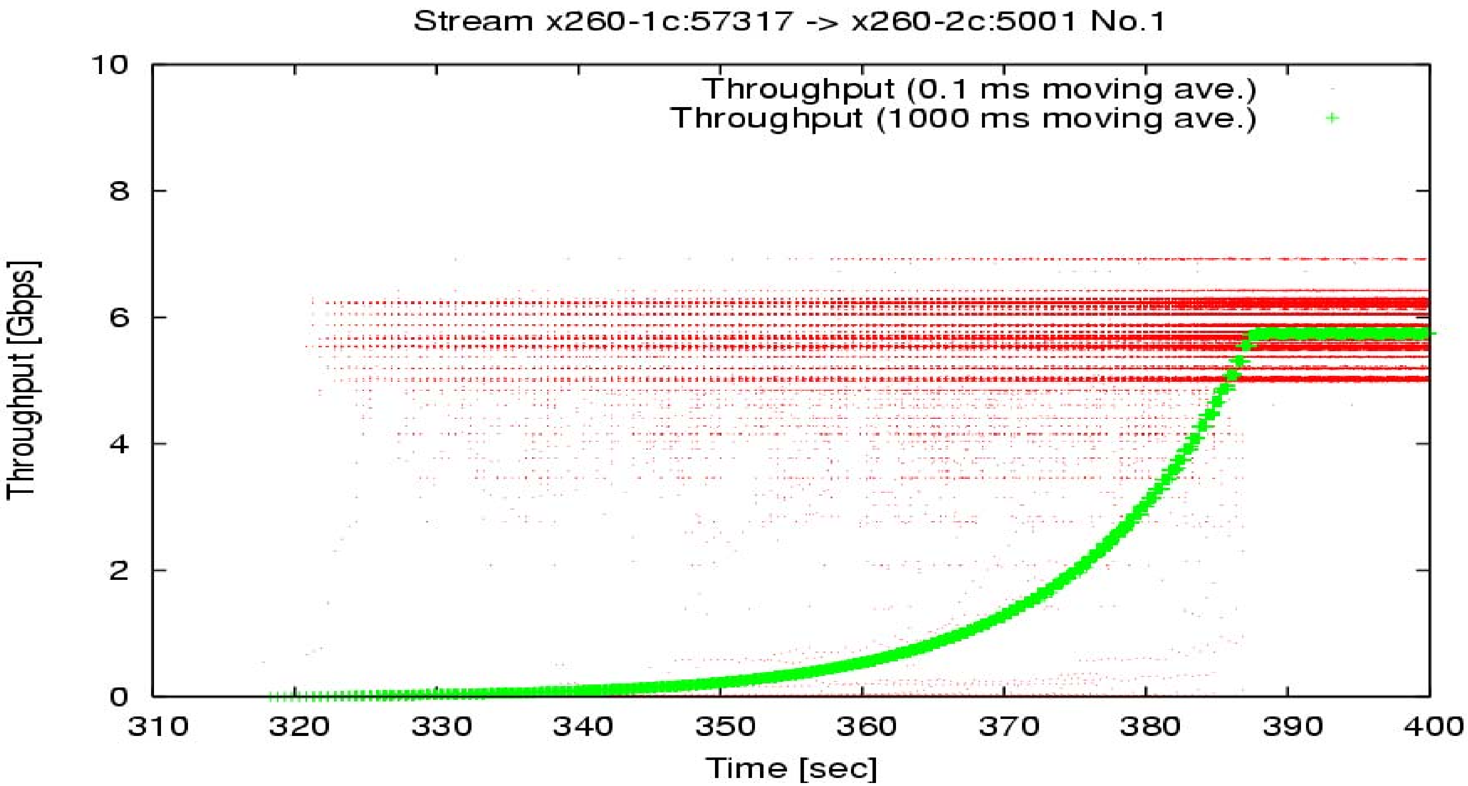


- Current IPv6 performance.
- This result have packet dropping in peak.

TCP/IP performance matrix

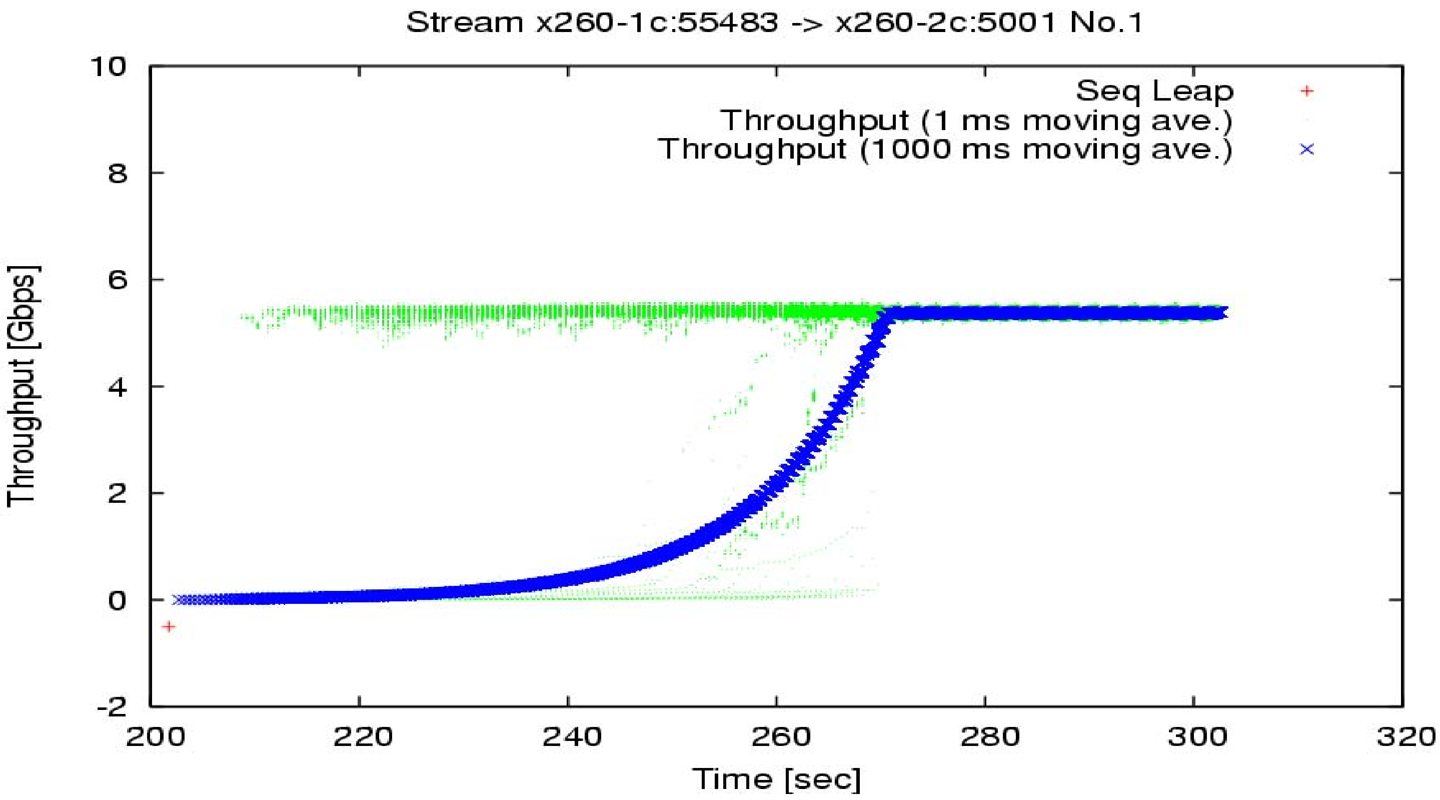
			2.6.12 (chelsio driver)		2.6.17		2.6.18-rc5	
			IPv4	IPv6	IPv4	IPv6	IPv4	IPv6
8G limit	Chelsio T110	opteron	7.2G (90%)	5.9G (75%)	x	x	x	x
	Chelsio N210	opteron	x	x	7.0G (85%)	7.0G (85%)	x	x
		Xeon	x	x	5.75G (75%)	5.75G (75%)	5.43G (65%)	1G (12%)
10G limit	Chelsio T310	Xeon	9.0G (90%)	5.43G (54%)	x	x	x	x

Linux 2.6.18-rc5 IPv4 TSO on



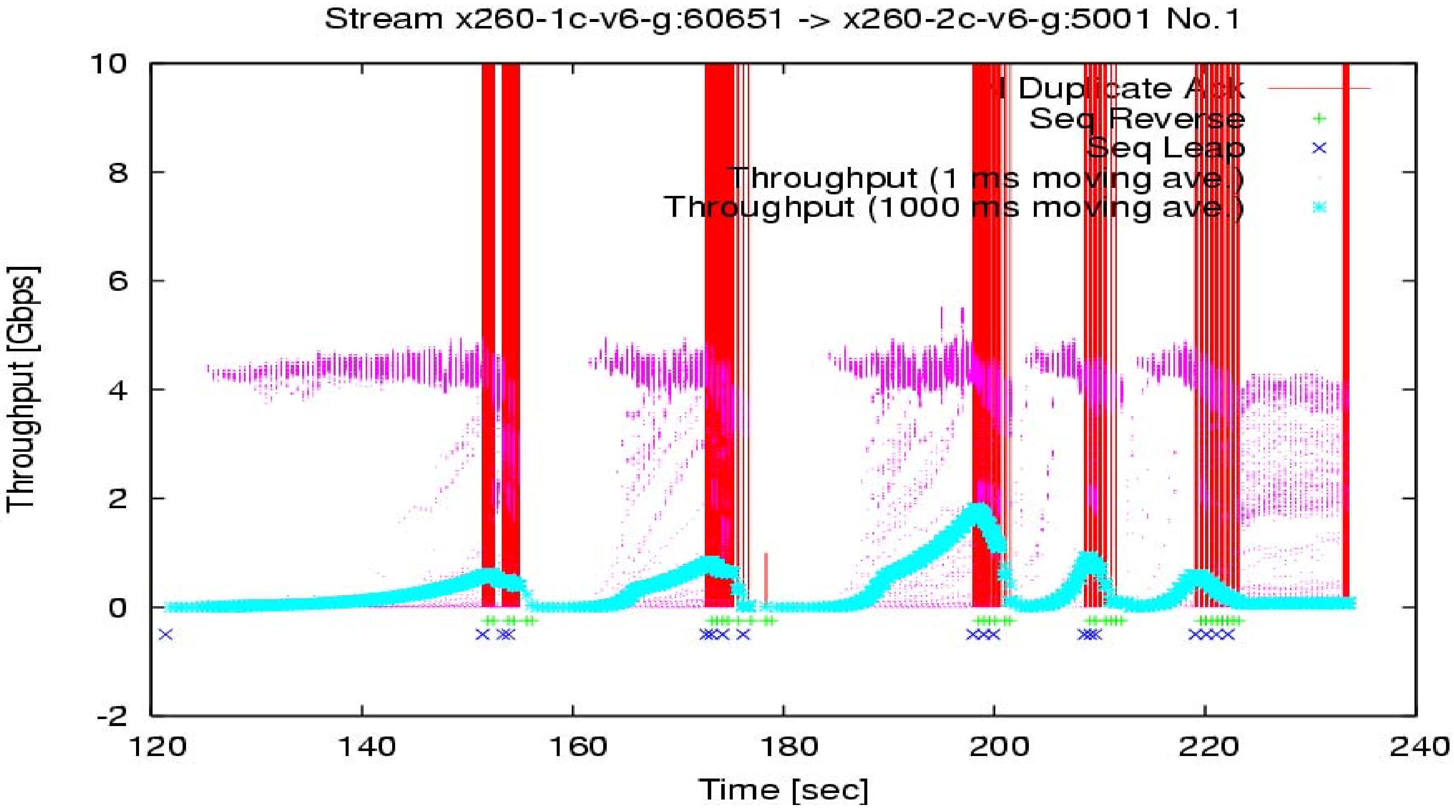
- Only 5.8Gbps on Xeon system
- Relative stable perfomance

Linux 2.6.18-rc5 IPv4 TSO off



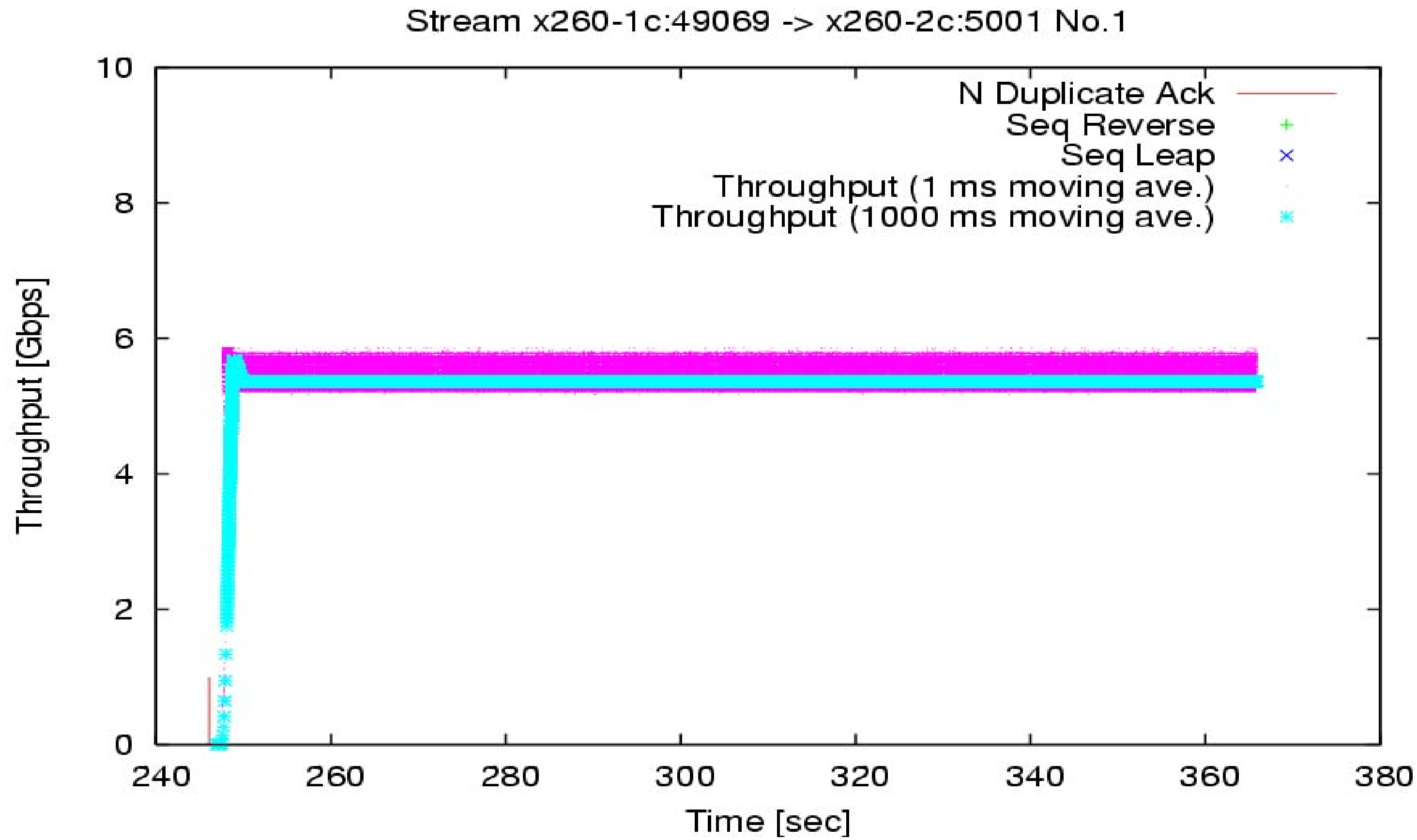
- Only 5.6Gbps on Xeon system
- stable perfomance

Linux 2.6.18-rc5 IPv6 GSO off



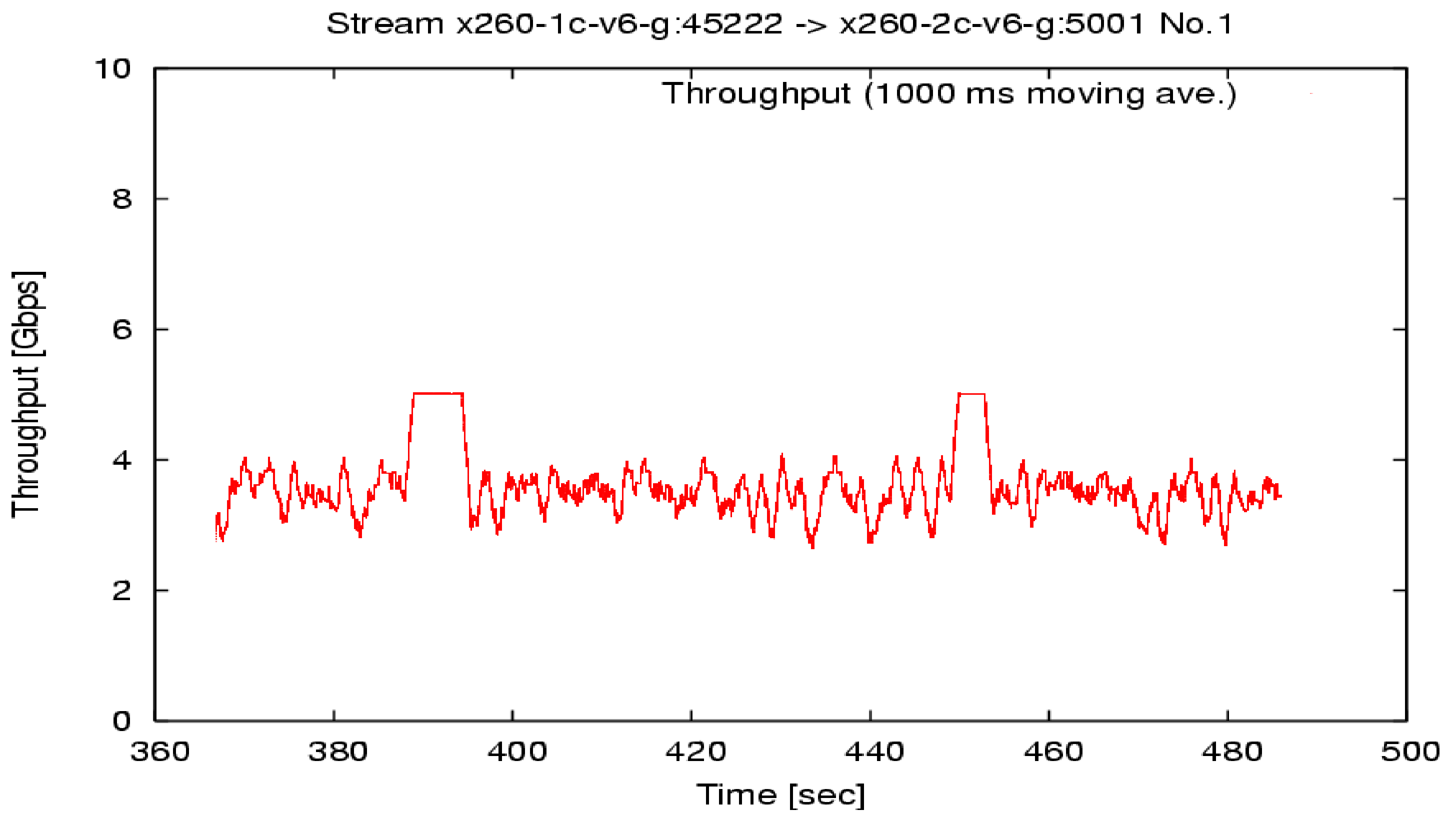
- Unstable performance
- Packet loss happened in the kernel
- TCP stack send duplicate ack for retransmission, but network doesn't drop any packets.

Linux 2.6.18-rc5 IPv4 TSO on RTT=10ms (1s average)



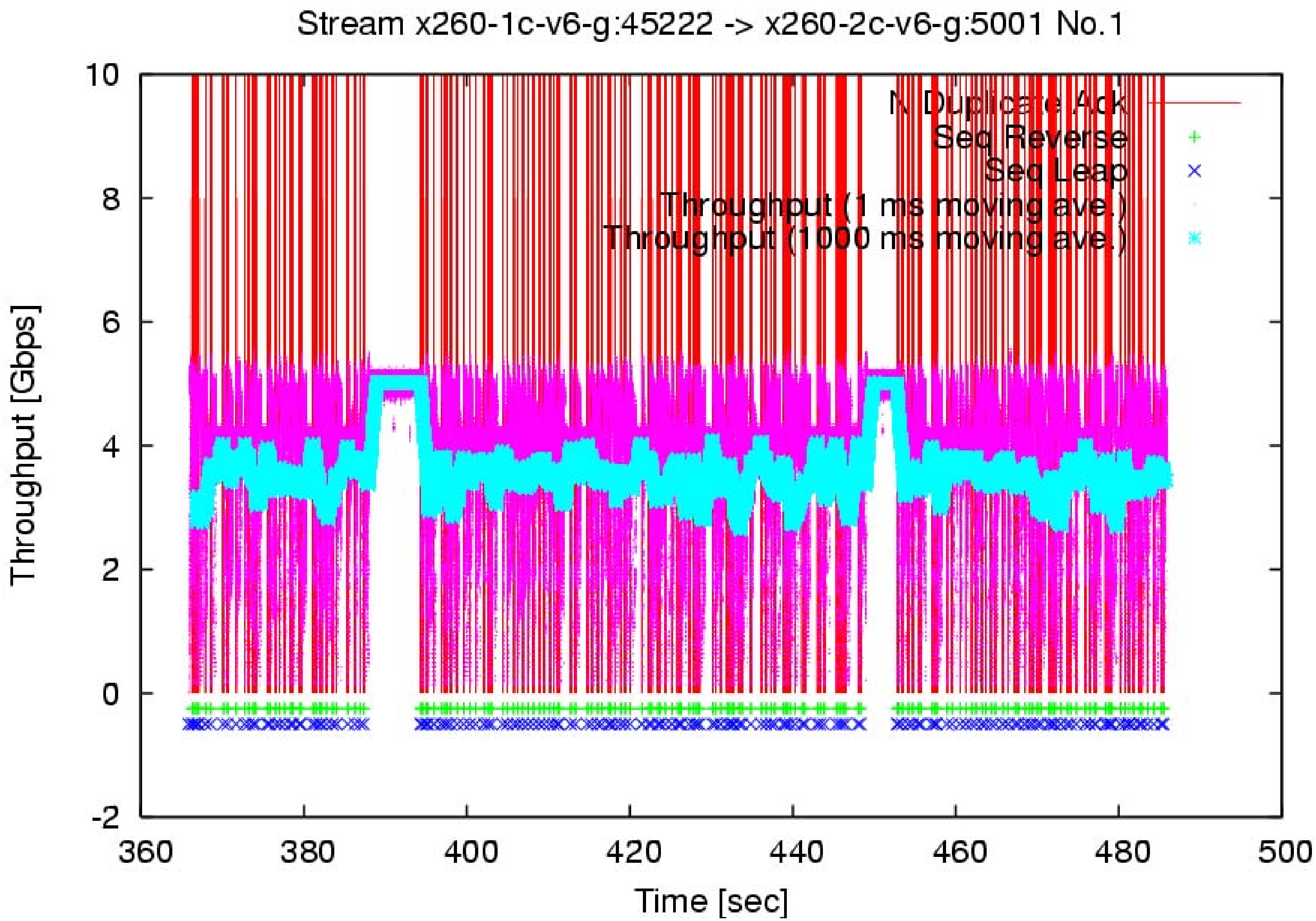
- RTT=10ms peak result is the same of RTT=500ms

Linux 2.6.18-rc5 IPv6 GSO off RTT=10ms (1s average)



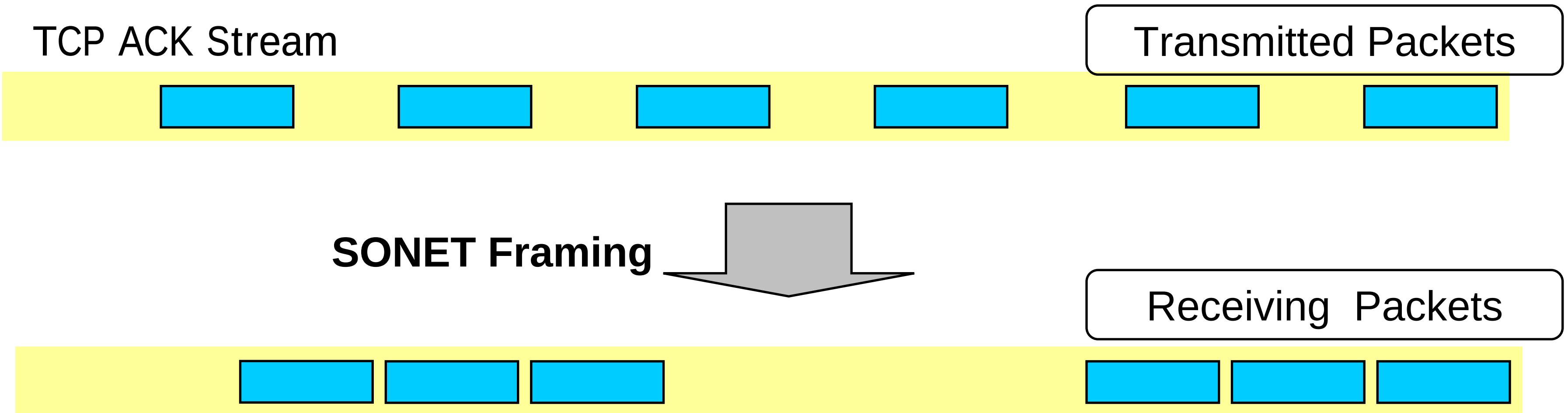
- Average performance is 3.8Gbps.
- This is almost 60% result of IPv4.

Linux 2.6.18-rc5 IPv6 GSO off RTT=10ms (1ms and Stream Info

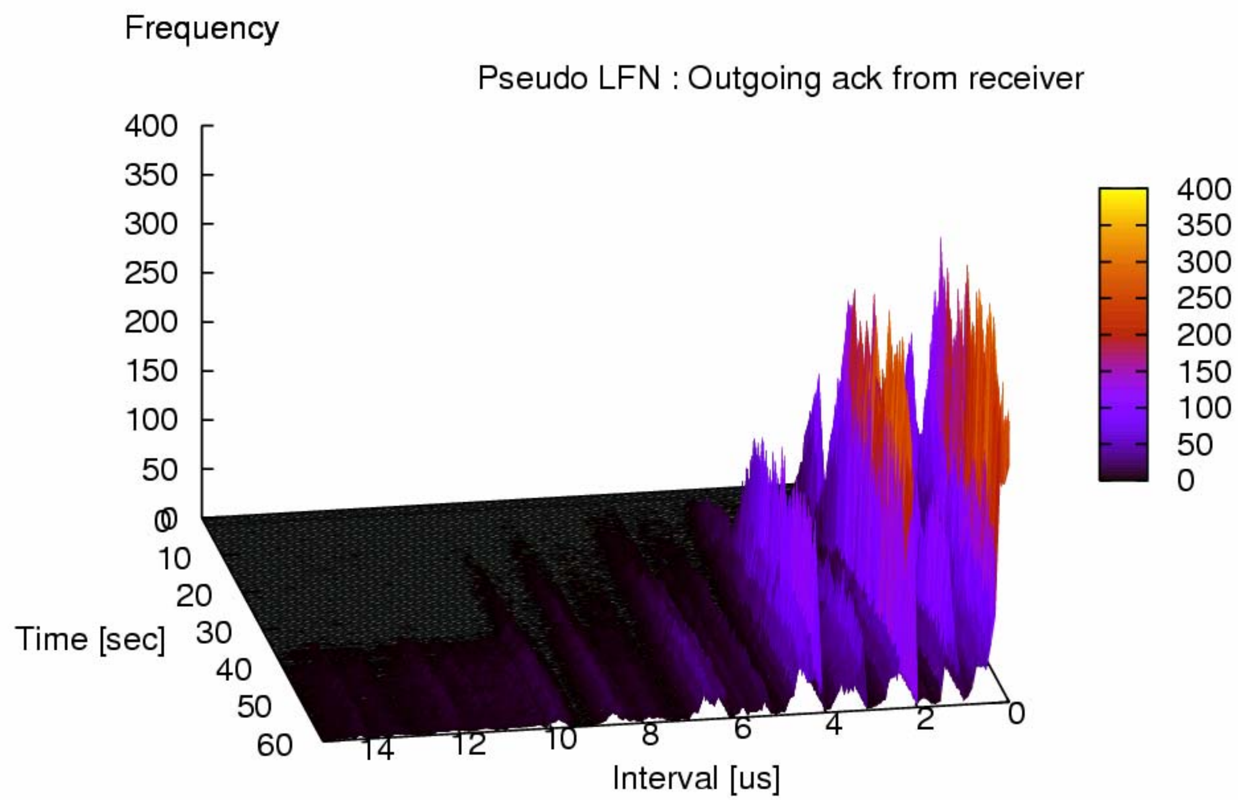


Ack Framing problem

- SONET has frame
- Some network instruments small packet packing into same frame
- Ack packets has no interval or frame interval

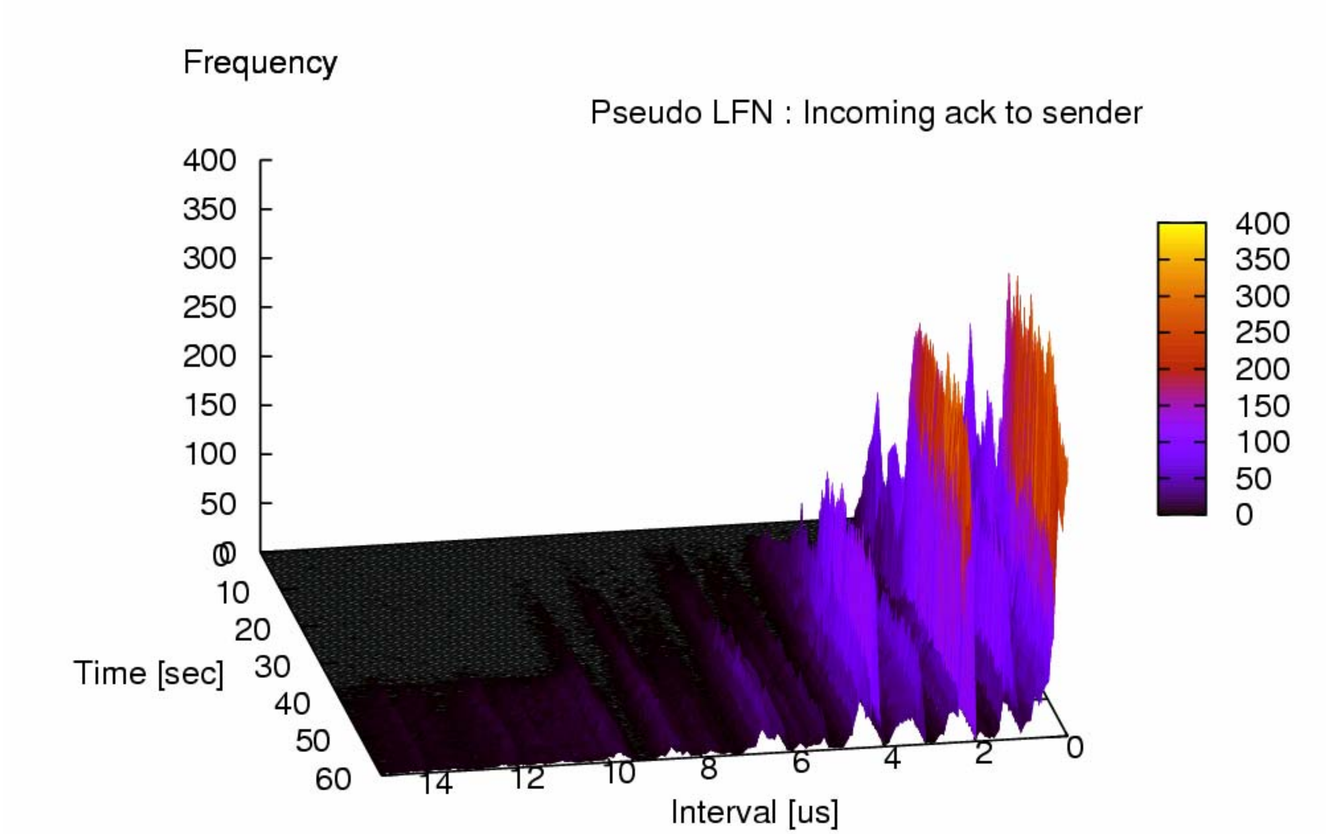


Sending ACK packets



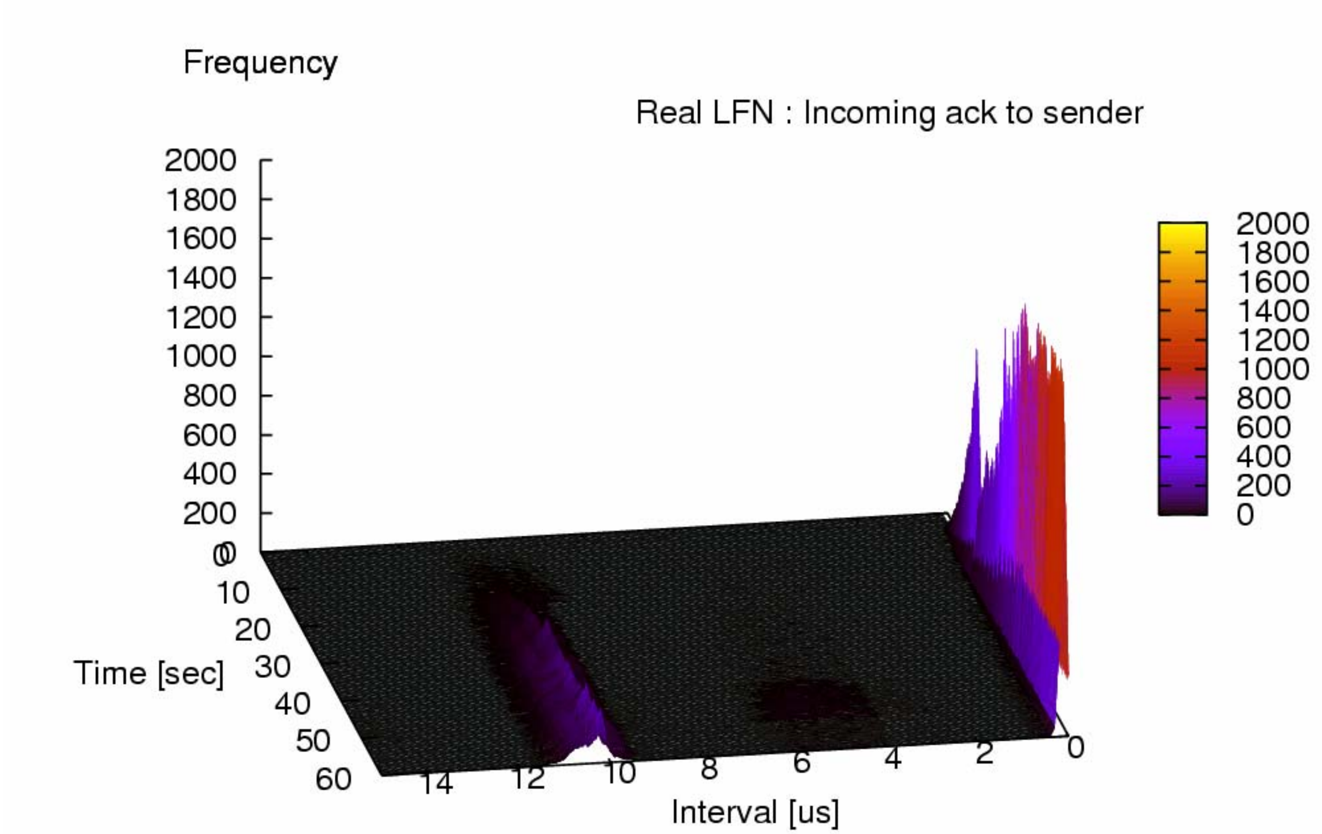
- Ack Sending

Pseudo LFN behavior



- Same packet interval is in Receiver side.

Real LFN behavior



- Almost packet interval push into 0 μ s by framing

Real LFN vs Pseudo LFN

- Both LFN shows the same performance macroscopically
  - 1s average performance is same.
- Real LFN shows the modified packet arrival interval.
  - SONET framing packing Ack packets.
- Receiver side receives short packets burst on Real LFN.
  - Real LFN needs higher packet receiving performance.

Toward the new LSR on IPv6

- We hope GSO stability on IPv6
  - The current performance bottle neck is a CPU performance of checksum Calculation.
- Stable performance on PCI-X 2.0 or PCI-Express x 16
  - There is a performance shield on 6 Gbps

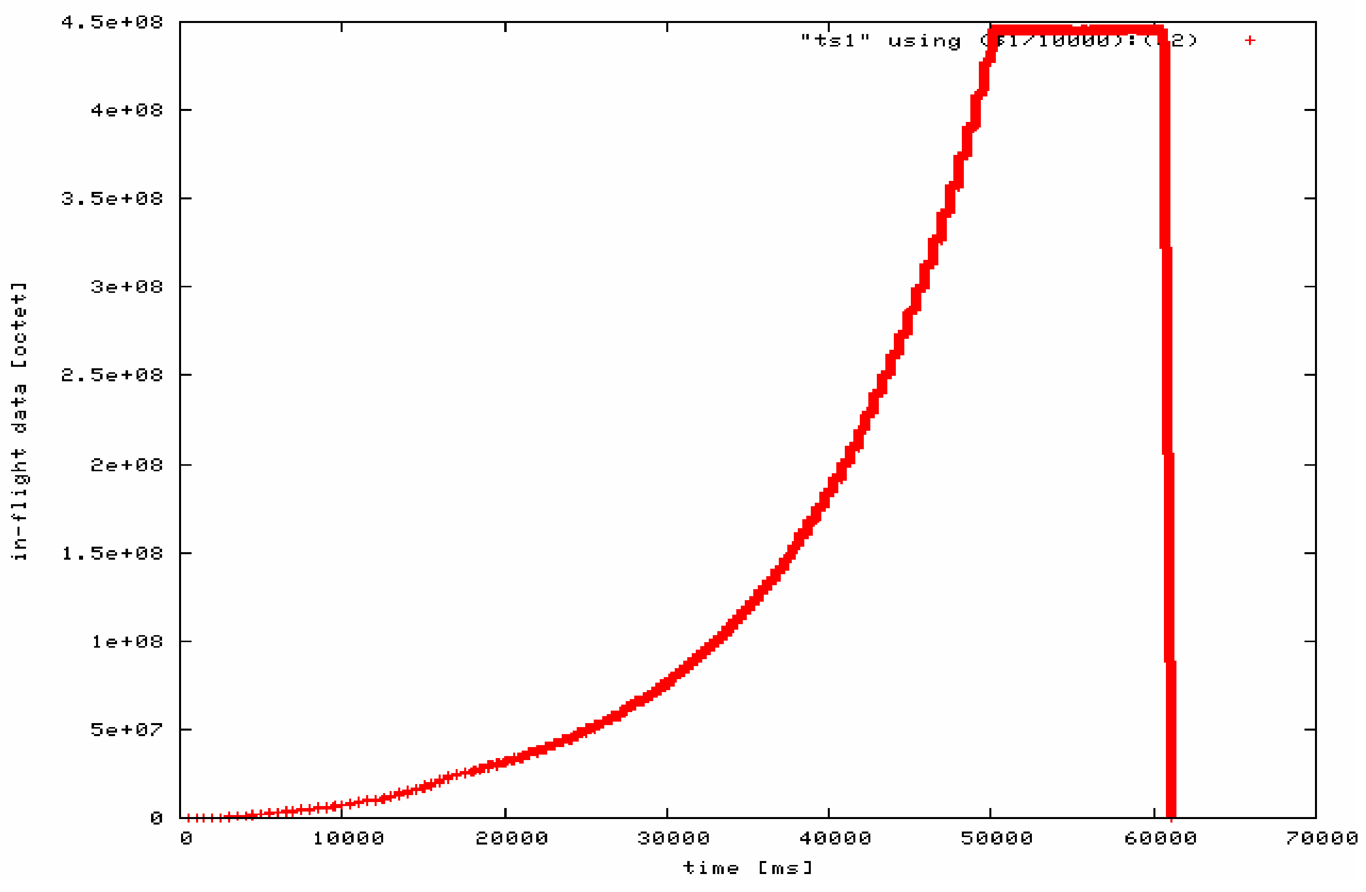
Summary

- Our LSR high performance TCP communication
  - We measured detailed network stream packets and showed many result
  - Feedback tuning for high performance
- TCP communication on LFN is difficult, but we can utilize till the same performance no relation with its latency.

acknowledge

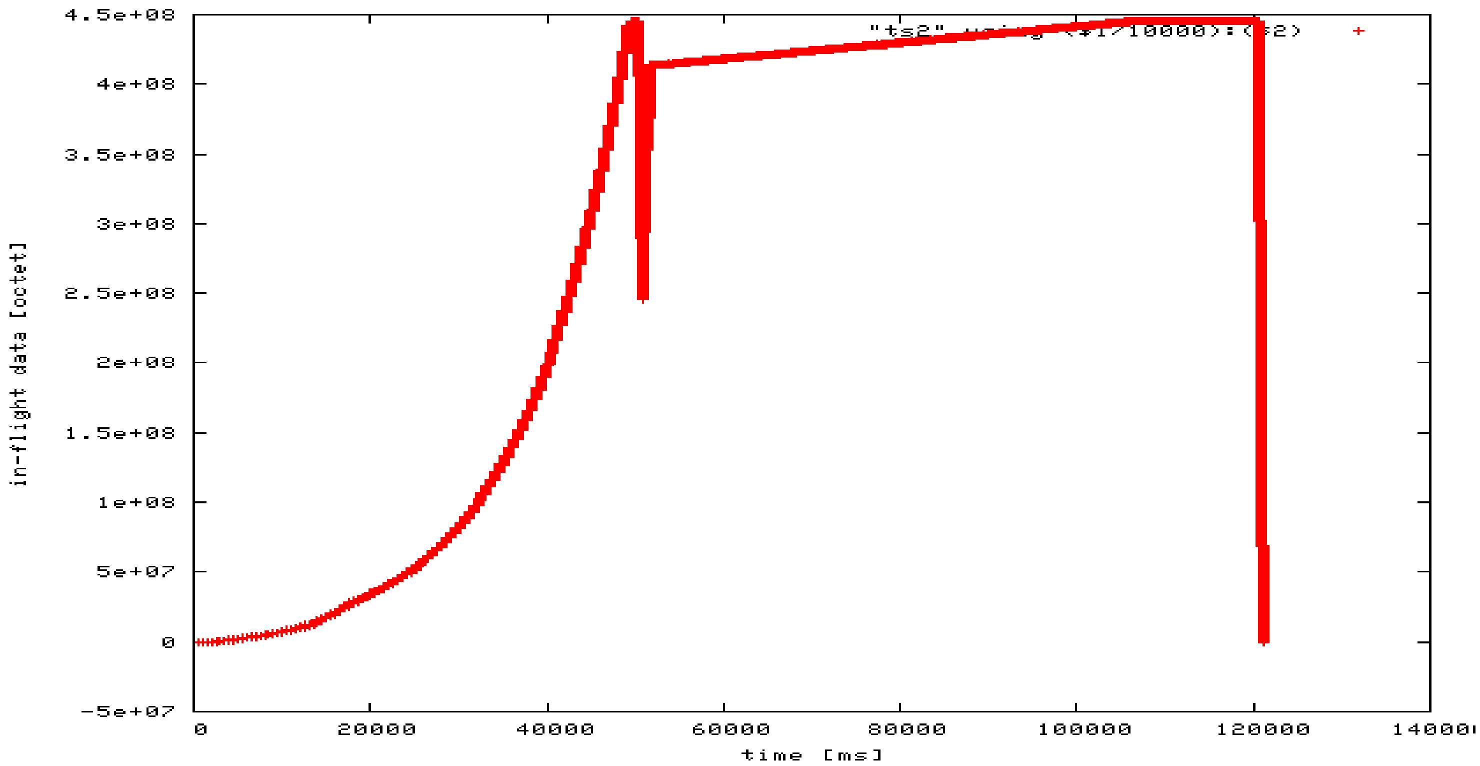
- Thanks for advice and support
  - Prof. Akira Kato University of Tokyo, ITC
  - WIDE Project
  - JGNII, IEEAF
  - Pacific Northwest Gigapop
  - AlaxalA Networks
- Thanks for providing Oversea Network
  - JGNII, SURFnet, IEEAF, CANARIE/CA*net

Linux 2.6.16 IPv6 Opteron Performance



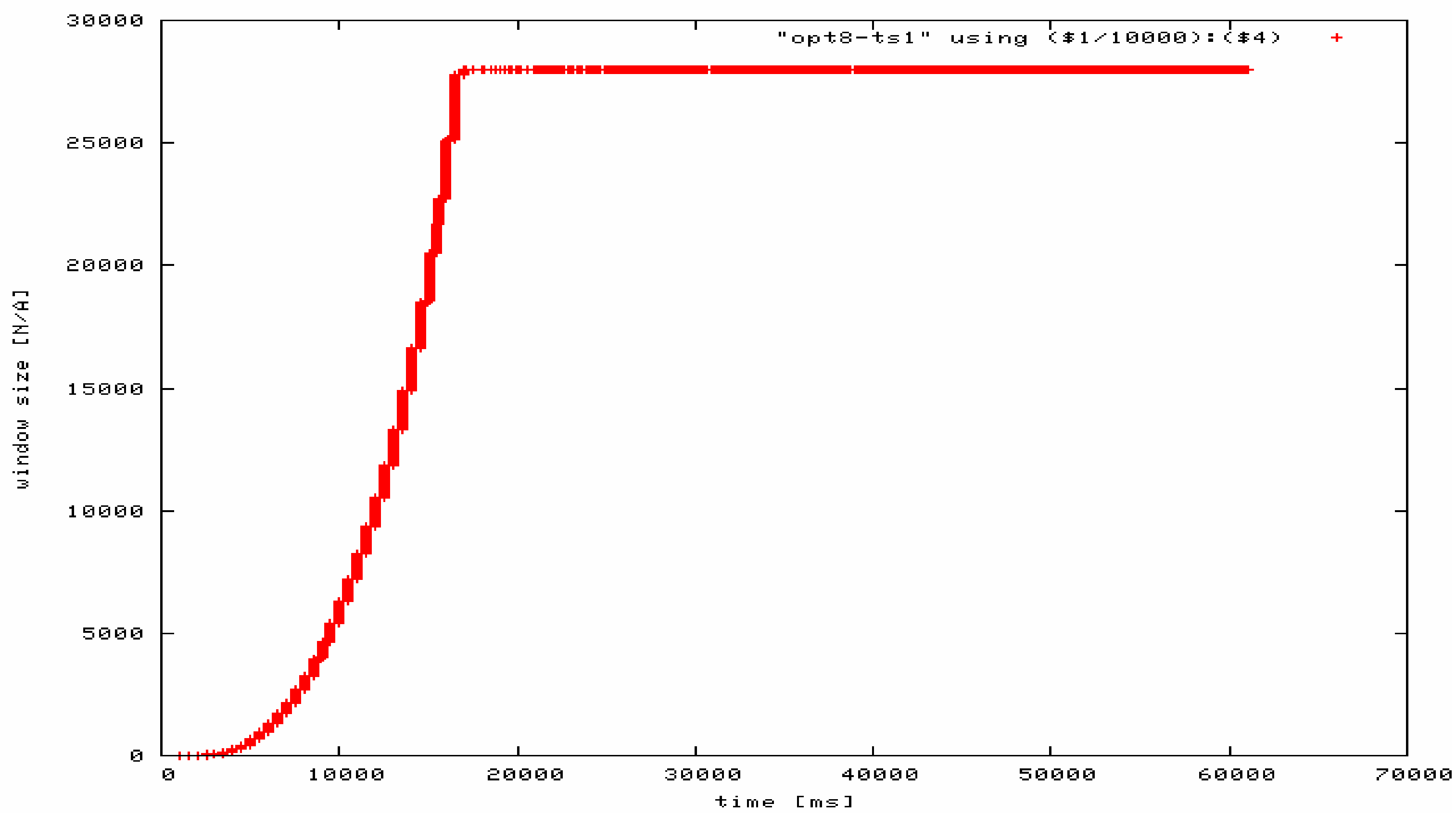
Current TCP stack shows stable window scaling on both IPv4 and IPv6

Larger Window Buffer of TCP



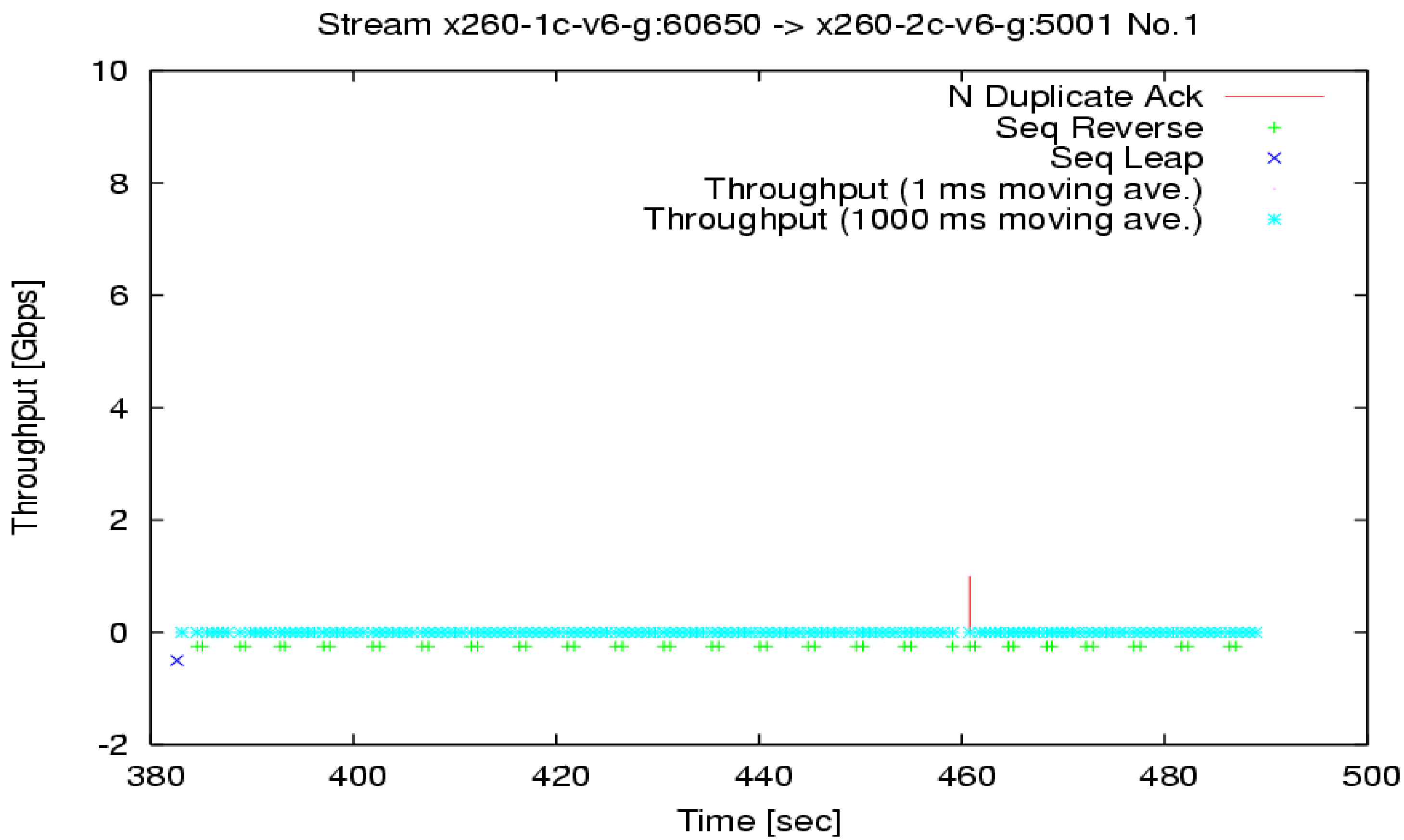
- Large Window buffer occurs packet loss on peak performance.

Linux 2.6.16 IPv6 Opteron Performance



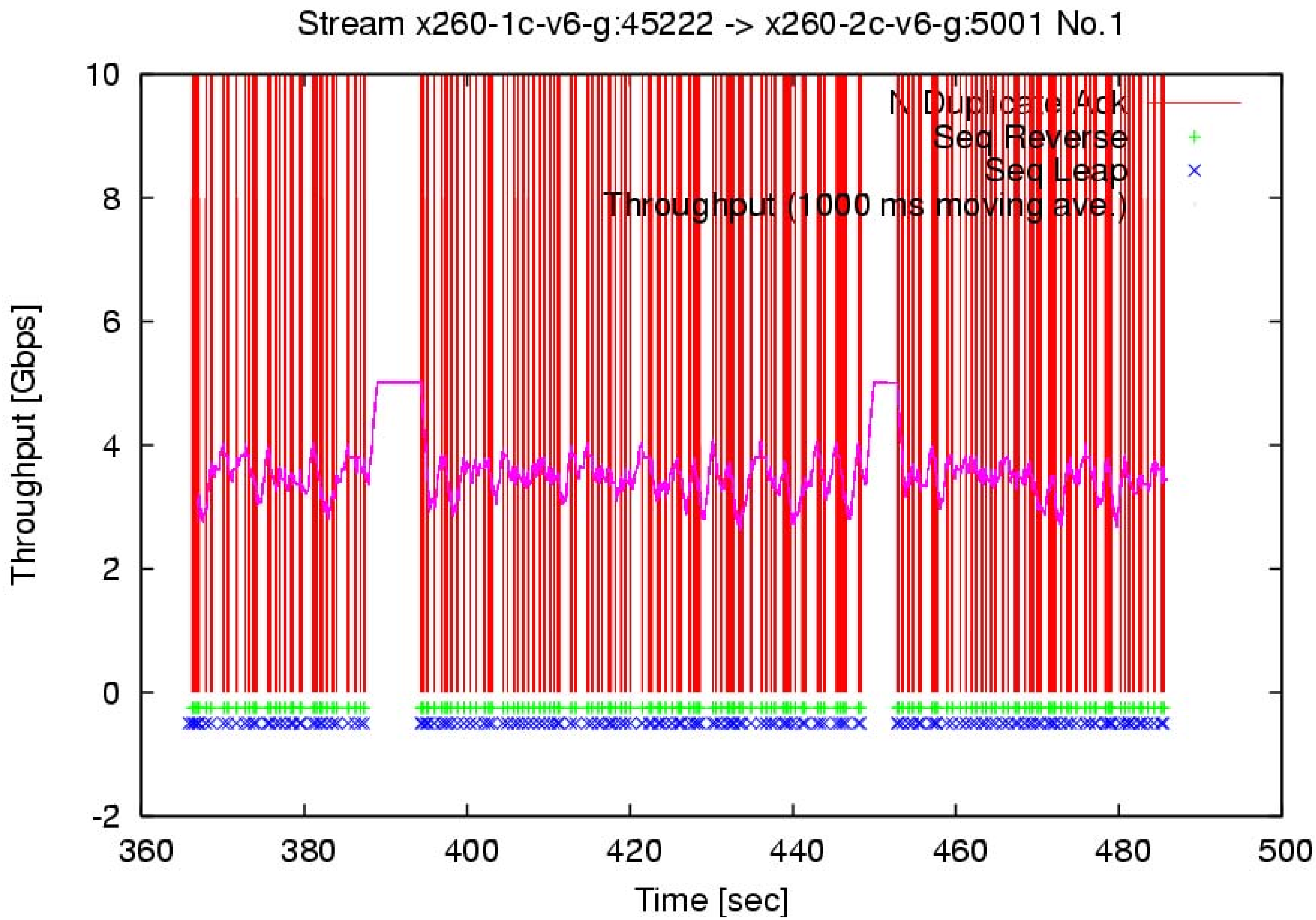
Adversized Window is grown faster than window size.
Slow window scaling is effect of delayed ack.

Linux 2.6.18-rc5 IPv6 GSO on



- Almost 100kbps on same network
- We met same condition on 2.6.12 IPv4 with TSO

Linux 2.6.18-rc5 IPv6 GSO off RTT=10ms (1s and Stream Info)



3, Network Interface Card

- PCI-X 1.0
  - Chelsio N210
- PCI-X 2.0
  - Chelsio T310

Chelsio N210

Chelsio T310

Linux 2.6.12 IPv4 Xeon Performance

usage(%)	function
30.1211	timer_interrupt
10.5991	mwait_idle
6.1435	find_busiest_group
5.7787	apic_timer_interrupt
4.3406	account_system_time
3.8784	scheduler_tick
3.4558	run_timer_softirq
3.2597	t3_intr
2.7998	schedule
2.463	__do_IRQ

IPv4 T310 receiver side

usage(%)	function
39.1652	copy_user_generic
7.1538	tcp_sendmsg
3.7135	tcp_ack
3.592	t3_eth_xmit
3.3121	put_page
2.7089	t3_intr
2.0278	timer_interrupt
1.9771	free_tx_desc
1.8016	skb_release_data
1.6117	kfree

IPv4 T310 sender side

- In IPv4, TSO or TOE is available. This result use TSO on sender side.
- Memory copy spend most of time, both side. From the effect of TSO, packet processing load is relatively small.

Current Performance

- We measured newest kernel 2.6.18-rc5 performance on same pseudo enviornment.
  - Limitation: Chelsio T310 couldnot execute on latest kernel for driver structure change.
  - Chelsio N210 (limited by PCI-X performance, 8.5Gbps)

RTT=10ms Performance

- Same test executed on small latency network.
  - Packet losses decrease the performance smaller than large latency network.
  - same packet loss phenomena shown in short interval
  - But relative higher perfomance than LFN.

Our result

- TCP Stream Behavior
  - Linux 2.6.12, 2.6.17, 2.6.18-rc5
- Behavior difference between Real LFN and Pseudo LFN
- Current Kernel performance

Linux 2.6.12 IPv6 Xeon Performance

usage(%)	function
23.6659	csum_partial_copy_generic
22.9821	copy_user_generic_c
12.8658	csum_partial
3.9911	timer_interrupt
2.1931	kfree
2.1852	process_responses
1.795	tcp_v6_rcv
1.7642	fib6_lookup
1.1321	eth_type_trans
1.1299	memcpy
1.0183	free_block

IPv6 N210 receiver side

usage(%)	function
48.2684	csum_partial_copy_generic
4.0249	timer_interrupt
3.0945	tcp_sendmsg
2.7058	cache_alloc_refill
2.3096	memcpy
1.7153	free_block
1.6977	put_page
1.5748	__rmqueue
1.4846	do_gettimeoffset_pm
1.3065	__mod_page_state

IPv6 N210 sender side

- In IPv6 have no hardware funtion, packet production use most of CPU power.
- CPU load is very high especially in sender side.
- Memory copy load is also high. This is same behavior on IPv4.