Experimental Results of TCP/IP data transfer On 10Gbps IPv6 Network

Junji Tamatsukuri, Katsushi Inagami
Mary Inaba, and Kei Hiraki
University of Tokyo / Data Reservoir Project

Overview

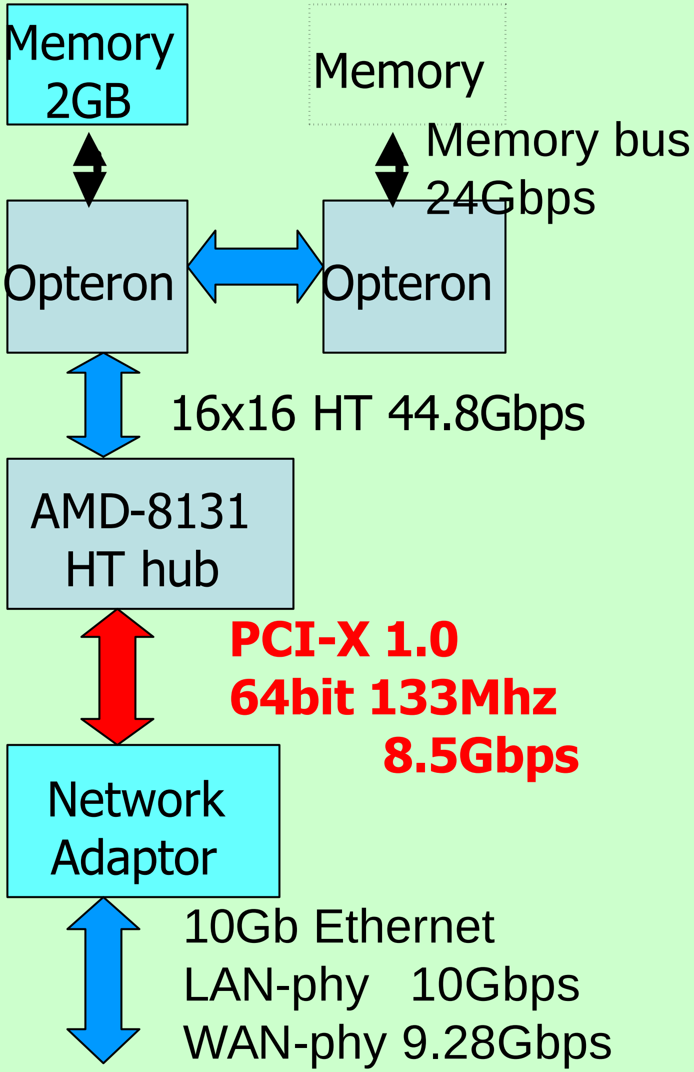
- We show the maximum performance of Single TCP/IPv6 stream on LFN (Long Fat Network)
  - – Pseudo LFN experiments by network emulator
  - – Real LFN experiments (over 30000km)
    - Tokyo – Seattle circuit
    - Tokyo – Chicago circuit

An Important Result

- “We can get same single TCP performance on LFN and local network”.
  - Necessary condition
    - Perfect network condition
    - Sufficient host computer performance
      - CPU for packet processing
      - Memory for data production, TCP window
      - I/O bus for network adaptor, storage
  - These condition mean No bottleneck in path

Current 10Gbps Problem

- Single stream TCP/IP performance is governed by bottleneck.
  - Network
    - 10Gbps Ethernet
  - Host
    - Interconnects (HT)
    - I/O bus (PCI-X)
- Current bottleneck is Host I/O Bus (PCI- X).



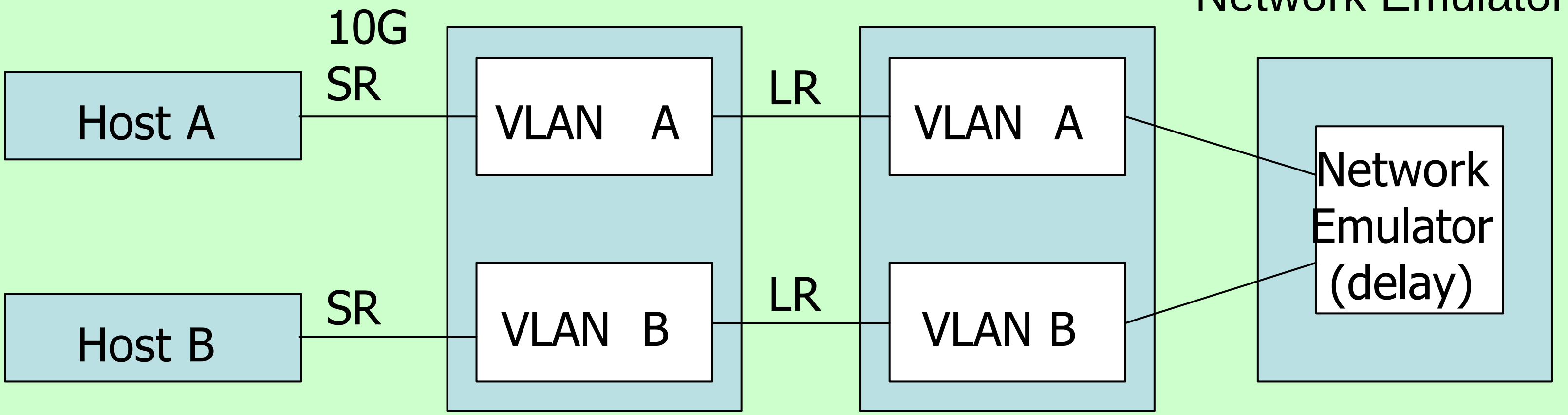
To relax the bottleneck

- PCI-X Bottleneck influence is remarkable at Sender Side.
  - Exploit “**flow control**” on edge port
    - Use back pressure to network
  - **“Transmission rate control”** at sender side
    - Decrease receiving side burst pressure

Pseudo Network Experiment

- LFN is a large RTT network.
  - Insert long delay by network emulator
- We use Anue H series network emulator
  - Anue H series can insert precise delay both direction .
- Flow Control on LFN switches

Anue H Series Network Emulator



Fujitsu XG800

Foundry MG8

Anue H Series

Experiment Equipments

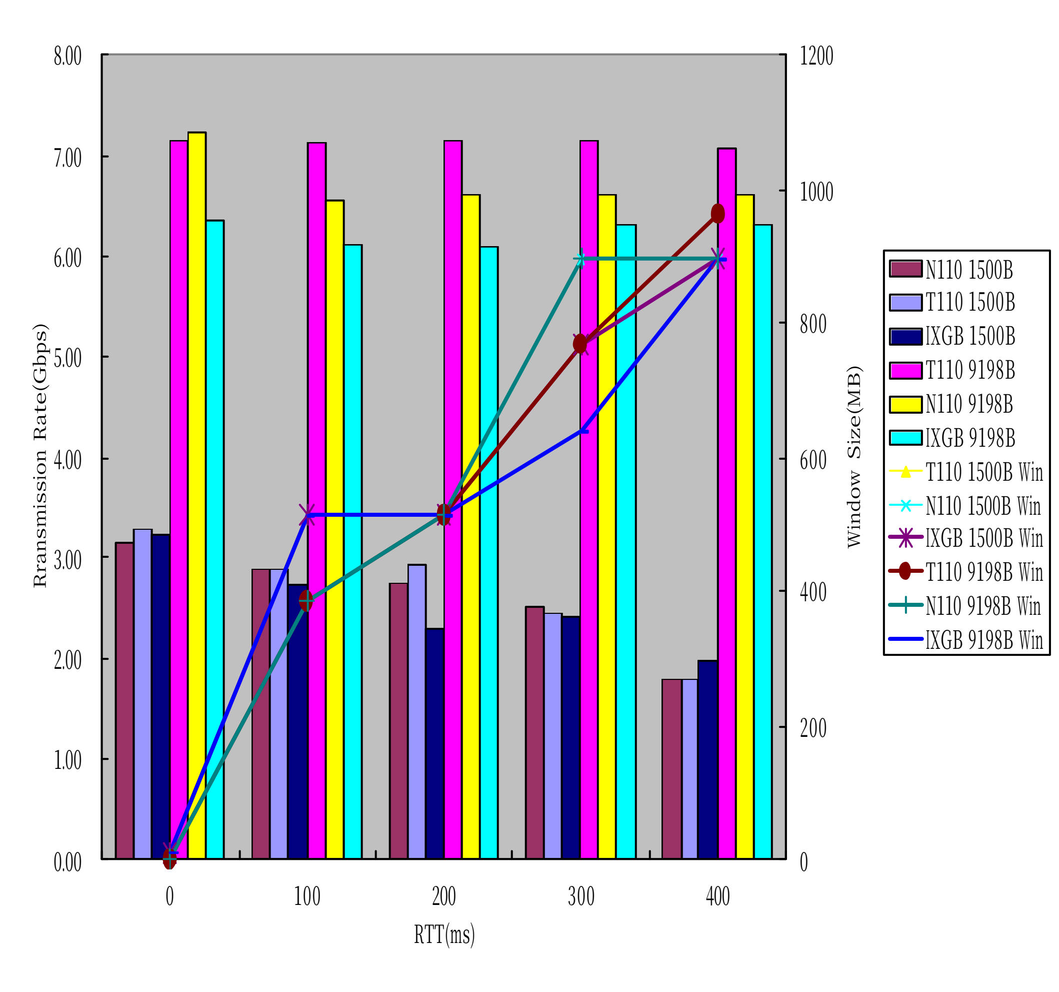
- Dual Opteron 248 (2.2GHz)
  - Rioworks HDAMA
  - DDR3200 CL2 2GB (Only Single Memory Bus)
- OS: Linux-2.6.6 (Linux TCP/IP stack)
- APP: Iperf-2.0.2
- Network Adaptor
  - Chelsio T110 Protocol Engine
    - TOE(TCP Offload Engine)
    - driver: chtoe-t1-1.1.4
  - Chelsio N110 Server Adapter
    - Without TOE
    - driver: cxgb-2.1.1
  - Intel PRO/10GbE
    - NAPI, TSO(TCP Segment Offload)
    - driver: ixgb-1.0.110

Chelsio T110

Intel PRO/10GbE

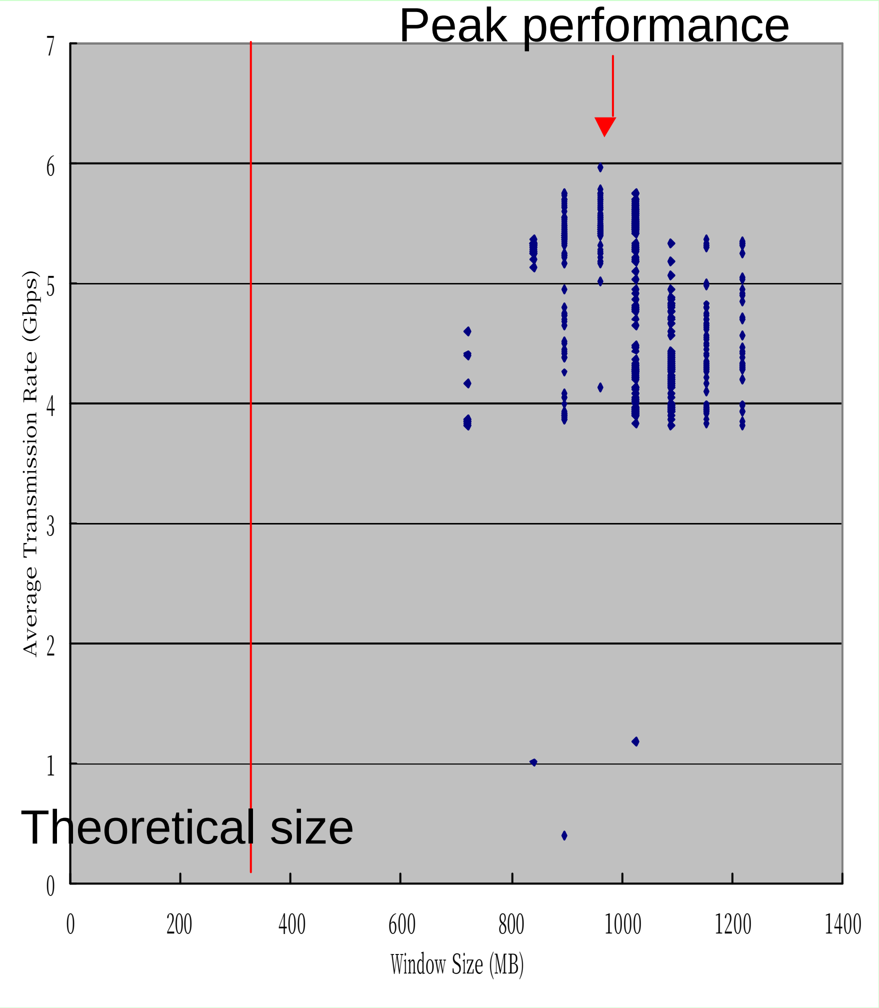
IPv6 Performance on Pseudo Network

- We measure performance from 0ms to 400ms RTT.
  - Standard Frame
  - Jumbo Frame(9198 Byte)
- Good performance on Pseudo LFN.
  - Local: 7 Gbps over
  - 400ms: almost 7Gbps over
  - 3 adaptor show similar peak performance on all RTT.
- Peak performance doesn’t change by RTT.



Window Buffer Size on LFN

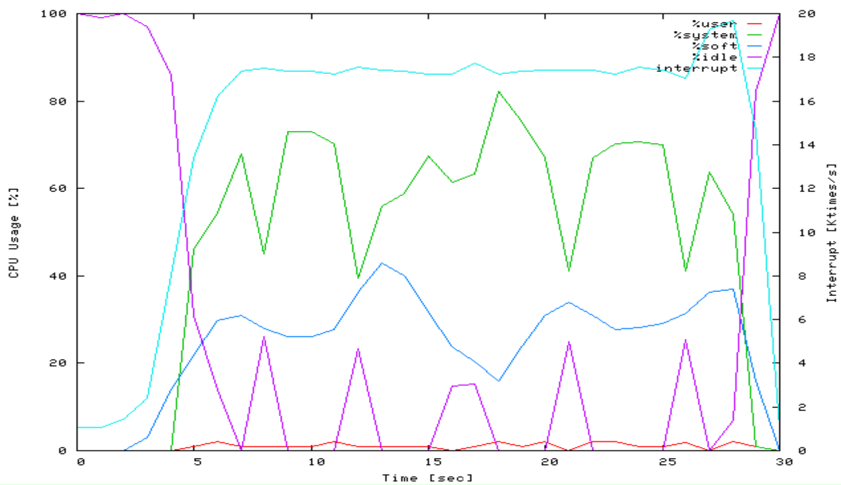
- RTT defines necessary window buffer size
  - Theoretical value
    - Buffer Size = RTT ×Traffic Rate
  - Real value
    - Linux stack needs 3 times larger than theoretical value.
- Proper value shows most stable result on communication


Peak performance

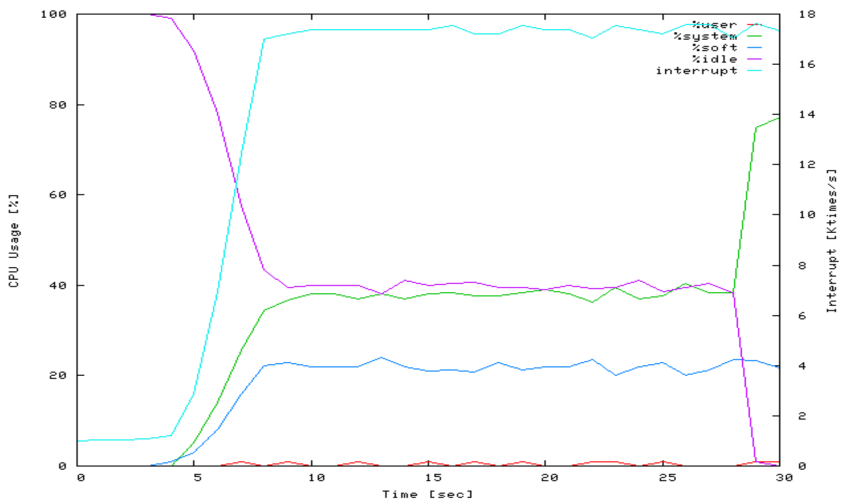
Theoretical size

CPU usage on Host

- Sender Side
  - Almost full use for TCP stack
  - Application use: 1%
  - Unstable behavior
    - Because of heavy CPU load
- Receiver Side
  - 40% idle
  - Stable behavior
    - Only periodical interrupt from network adaptor



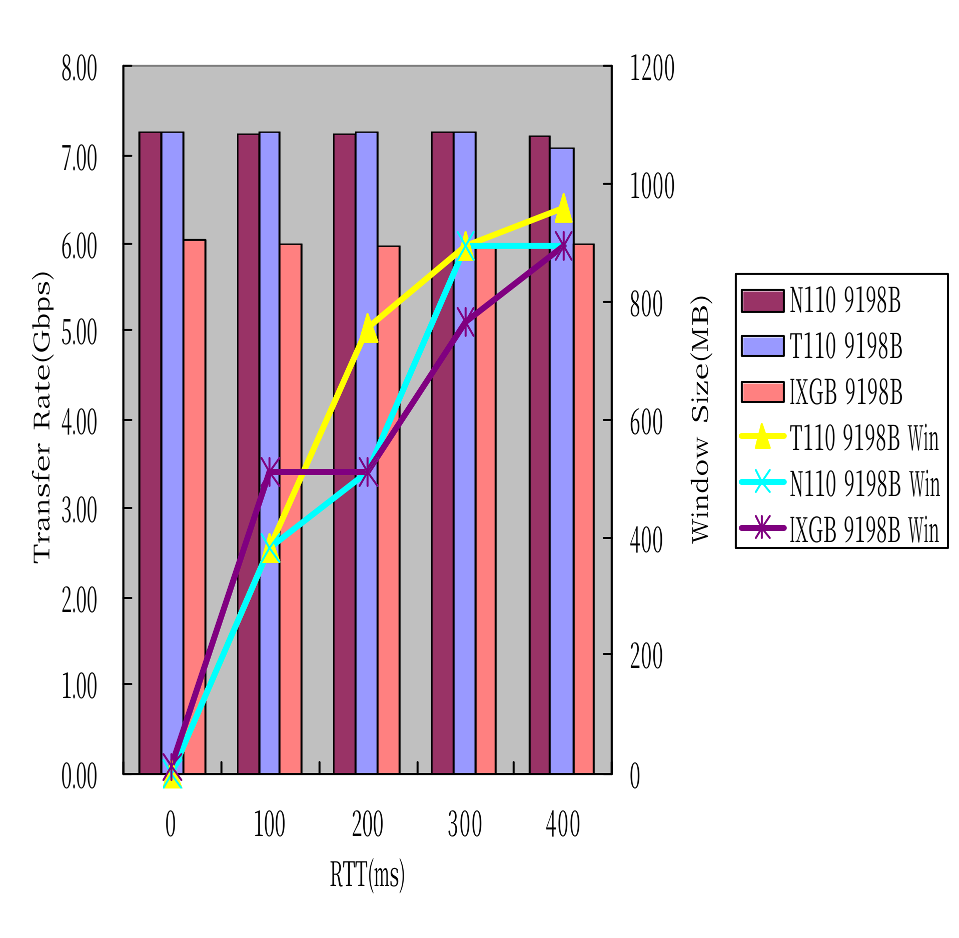
Sender side



Receiver side

IPv4 Performance on Pseudo LFN

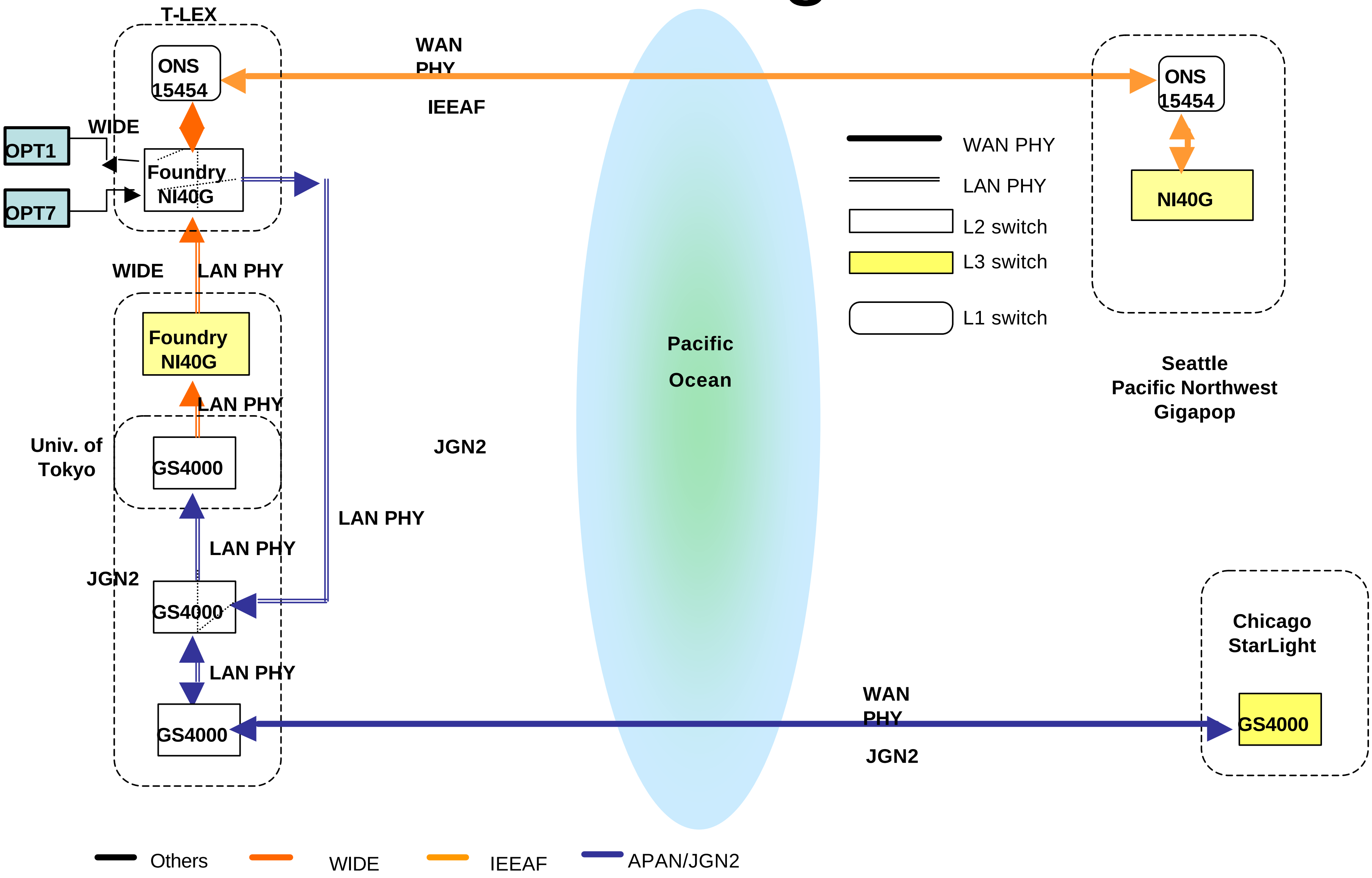
- IPv4 shows same performance as IPv6
  - All result is software performance.
  - Local: 7Gbps over
  - 400ms: 7Gbps over



Real LFN Experiment

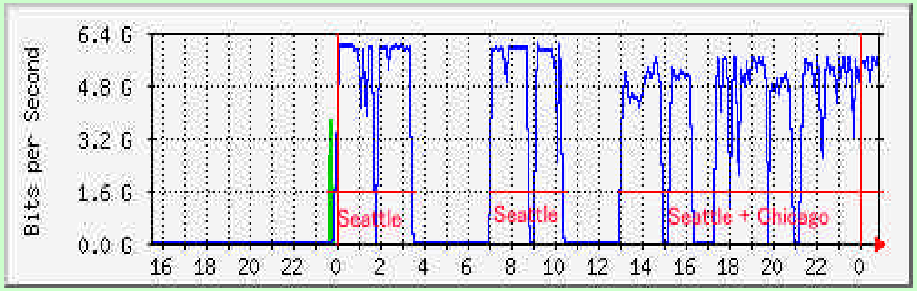
- We tried Real LFN measurement
  - IEEAF Tokyo – Seattle circuit
  - JGN2 Tokyo ー Chicago circuit
- Real LFN has more difficult condition
  - Packet loss, Jitter of Packet arrival
    - By network circuit, network equipment
- All parameters set according to Pseudo LFN Experiment
  - based on 200ms, 400ms result
  - Same Host Configuration with New kernel 2.6.12
  - Only use Chelsio T110 adaptor without TOE

Network Configuration

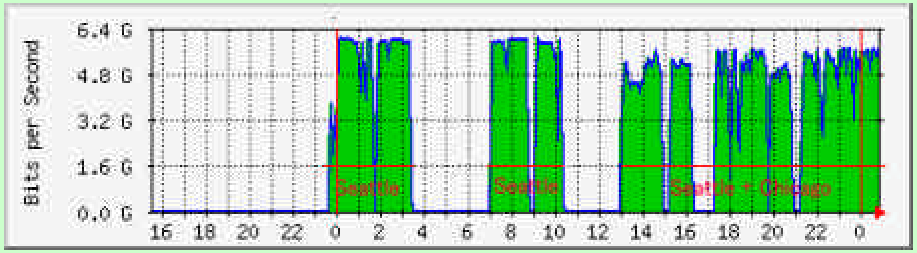


Tokyo – Seattle –Tokyo LFN Experiment

- Tokyo – Seattle Roundtrip
  - 2005/10/28
  - RTT 178ms
  - Distance 15,461km
  - Window buffer 512MB
- LFN routers
  - Foundry NI40G: T-LEX, Seattle, U-Tokyo(NEZU)
  - Hitachi GS4000: U-Tokyo, NTT Otemachi
- Circuit condition
  - Stable but low performance
  - Performance 5.96Gbps



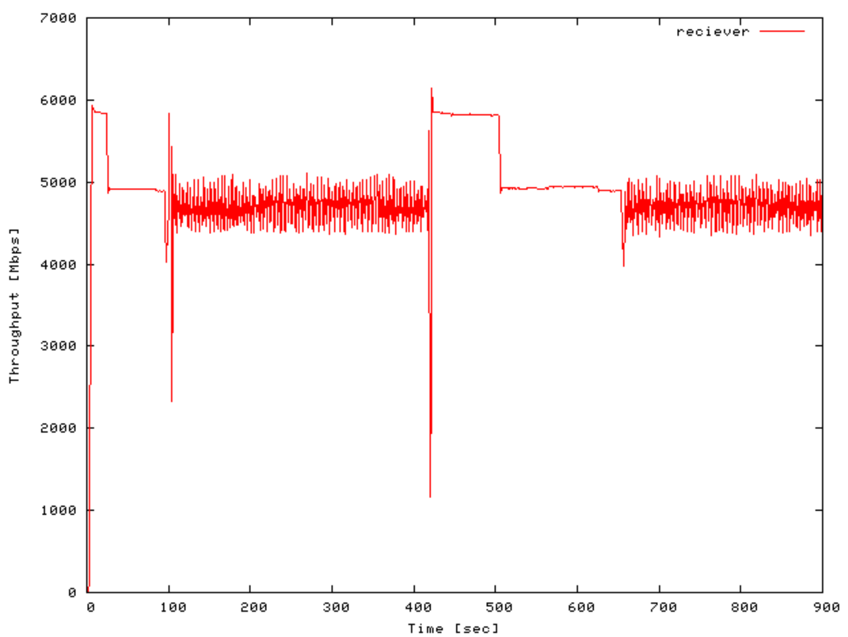
T-LEX NI40G



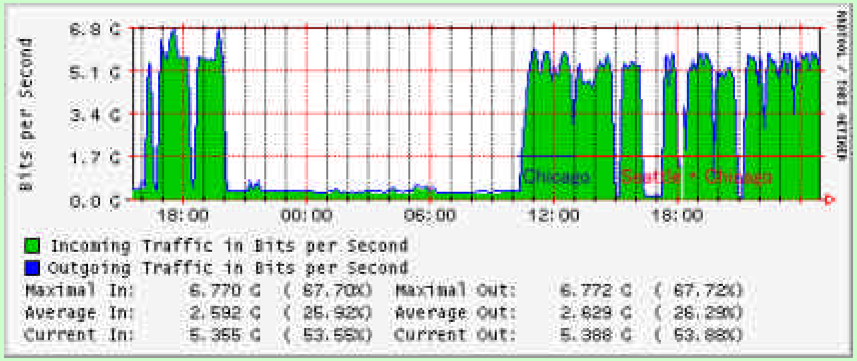
Seattle NI40G

Tokyo – Chicago – Tokyo LFN Experiment

- Tokyo – Chicago Roundtrip
  - 2005/10/28,29
  - RTT 322ms
  - Distance 20,294km
  - Window buffer 896MB
- Route
  - Equipment
    - Foundry NI40G: T-LEX, U-Tokyo(NEZU)
    - Hitachi GS4000: U-Tokyo, NTT Otemachi, Chicago, KDD Otemachi
- Circuit condition
  - Unstable
    - Periodical UP/DOWN condition
  - Performance 5.6Gbps



Periodical UP/DOWN

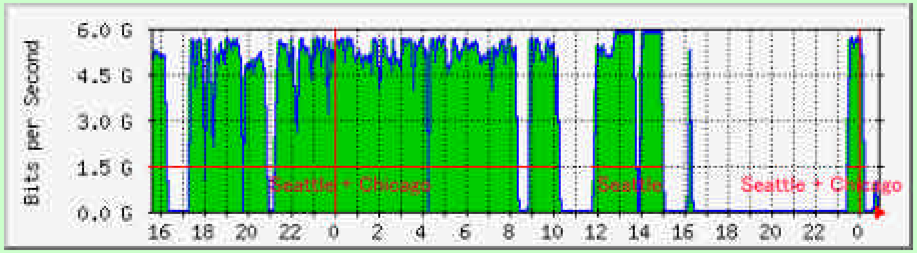


```
🟩 Incoming Traffic in Bits per Second
🟦 Outgoing Traffic in Bits per Second
Maximal In:     6.770 G  ( 67.70%)  Maximal Out:    6.772 G  ( 67.72%)
Average In:     2.592 G  ( 25.92%)  Average Out:    2.629 G  ( 26.29%)
Current In:     5.355 G  ( 53.55%)  Current Out:    5.388 G  ( 53.88%)
```

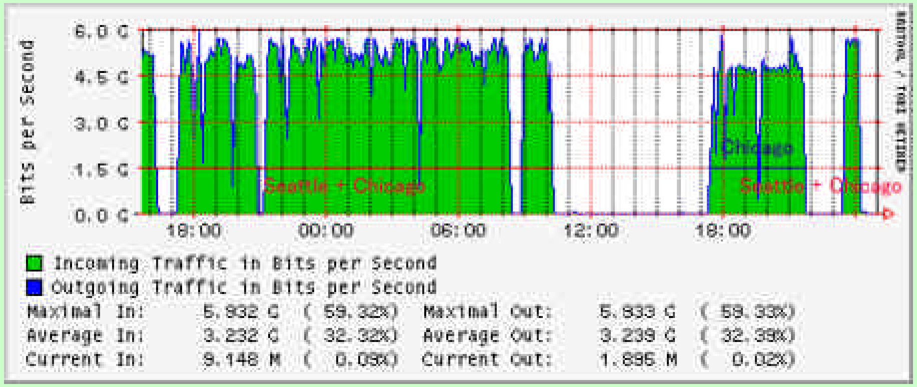
JGN NTT Otemachi GS4000

Tokyo – Seattle –Tokyo – Chicago –Tokyo LFN Experiment

- Tokyo – Seattle – Chicago Roundtrip
  - 2005/10/29
  - RTT 500ms
  - Distance 35,755km
  - Window buffer 896MB
- Route
  - T-LEX -> Seattle -> KDD Otemachi -> Chicago -> U-Tokyo -> T-LEX
  - NI40G, GS4000
- Circuit condition
  - Better than Chicago roundtrip
    - We couldn't observe UP/DOWN condition
  - Performance 5.6Gbps



Seattle NI40G



🟩 Incoming Traffic in Bits per Second
🟦 Outgoing Traffic in Bits per Second

```
Maximal In:      5.932 G  ( 59.32%)  Maximal Out:      5.933 G  ( 59.33%)
Average In:      3.232 G  ( 32.32%)  Average Out:      3.239 G  ( 32.39%)
Current In:      9.148 M  (  0.09%)  Current Out:      1.895 M  (  0.02%)
```

JGN Chicago GS4000

Result on Real LFN

- Network condition has much influence
  - We tried test for preparing of SC2005
    - All the routes have many problem in circuits and equipments.
  - We got 6Gbps level performance on real LFN.
    - We set decreased clock on sender side (6Gbps).
    - For stable receiving.
- Result couldn’t reach pseudo LFN performance.
  - Real LFN has very difficult condition.
  - Except for circuit condition, Real LFN shows same behavior of Large RTT pseudo network

Concluding Remarks

- We show pseudo/real LFN experiment
  - Sender side rate control / Flow Control is effective for Single TCP performance.
  - Real network has many influence elements on circuit, equipment.
- We got Internet2 Land Speed Record
  - IPv6 Single/Multi Stream Category (2005/10/29)
    - 5.6Gbps × 30,000km
- Aimed at more performance
  - We'll try experiments for the result as same as pseudo LFN result.

Acknowledgements

- Thanks to
  - WIDE Project / T-LEX IEEAF staffs
  - JGN2 Domestic / International Operation
  - Hitachi, Alaxala
  - Foundry Networks
  - Chelsio