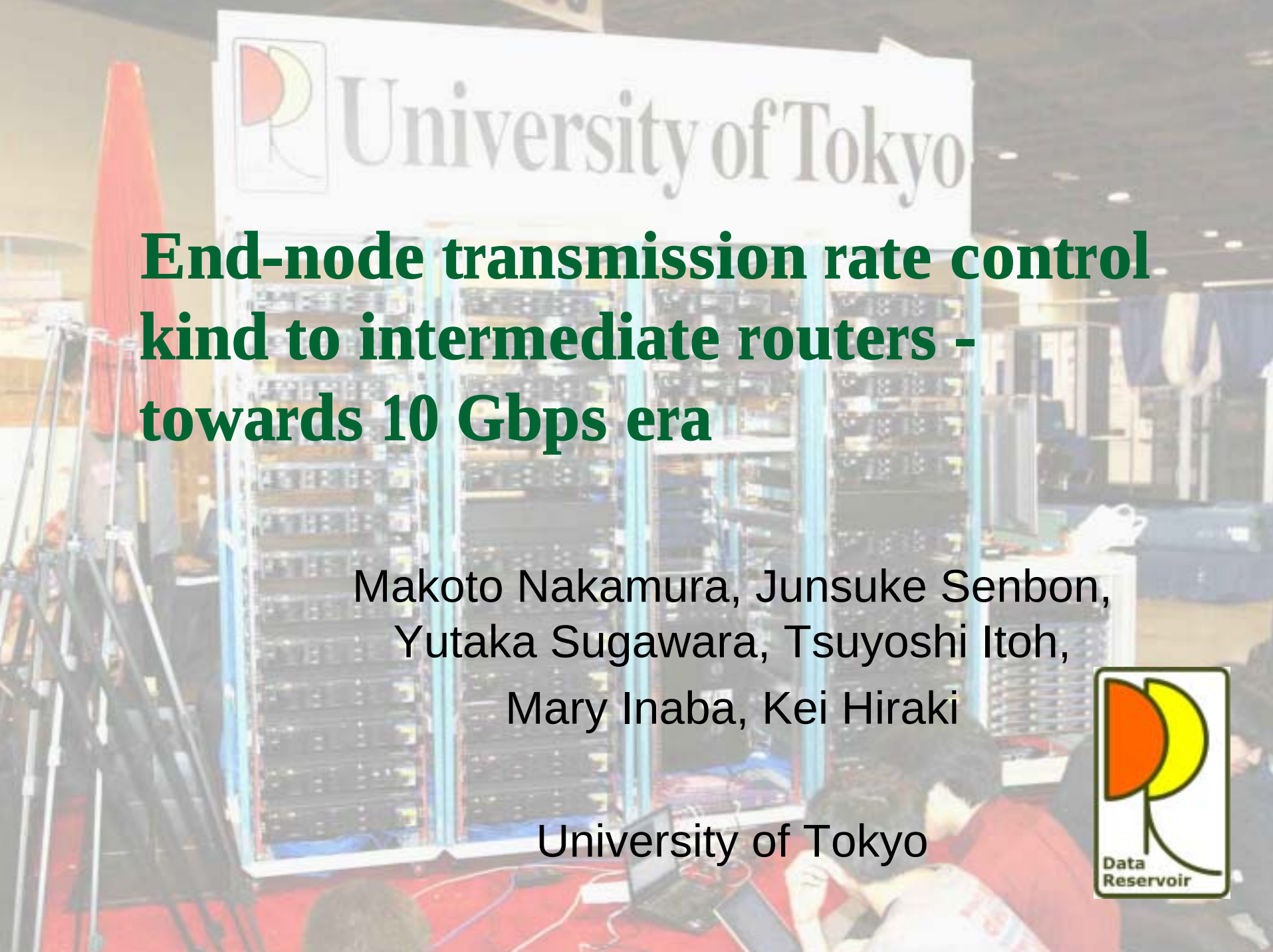


End-node transmission rate control kind to intermediate routers - towards 10 Gbps era

Makoto Nakamura, Junsuke Senbon,
Yutaka Sugawara, Tsuyoshi Itoh,
Mary Inaba, Kei Hiraki

University of Tokyo

Data Reservoir

Outline of this talk

- Background of Data Reservoir Project
- Observations at BWC2002
- Transmission Rate Controlled TCP for DR
  - Software approach
    - IPG tuning
    - Clustered Packet Spacing
  - NIC hardware approach
    - TCP-aware NIC
- Results at BWC2003

Objectives of Data Reservoir

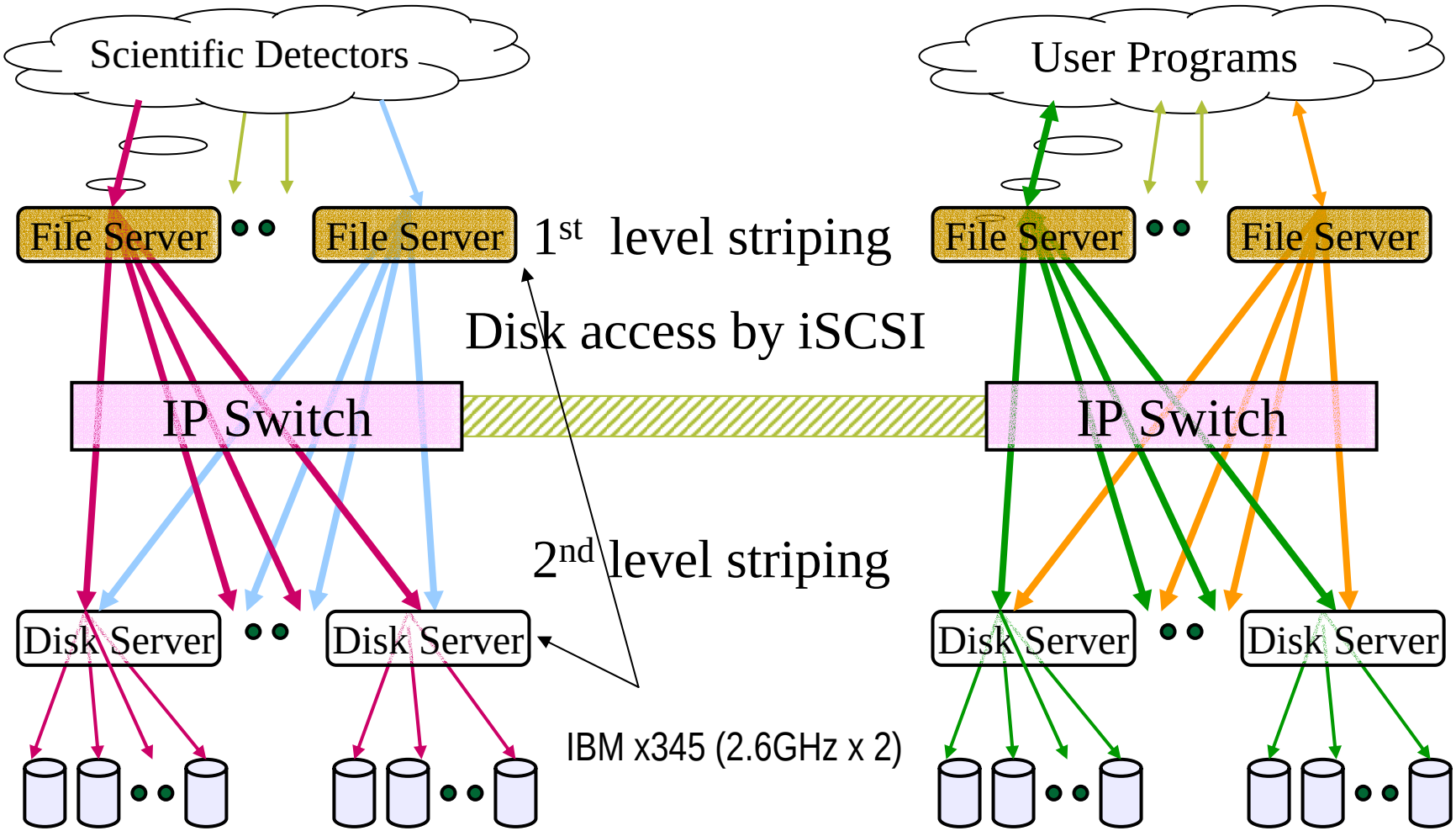
- Sharing scientific data between distant research institutes
  - Physics, astronomy, earth science, simulation data
- Very high-speed single file transfer on Long Fat pipe Network (LFN)

- High utilization of available bandwidth
- OS and filesystem transparency
  - Storage level data sharing
    - High speed iSCSI protocol on TCP

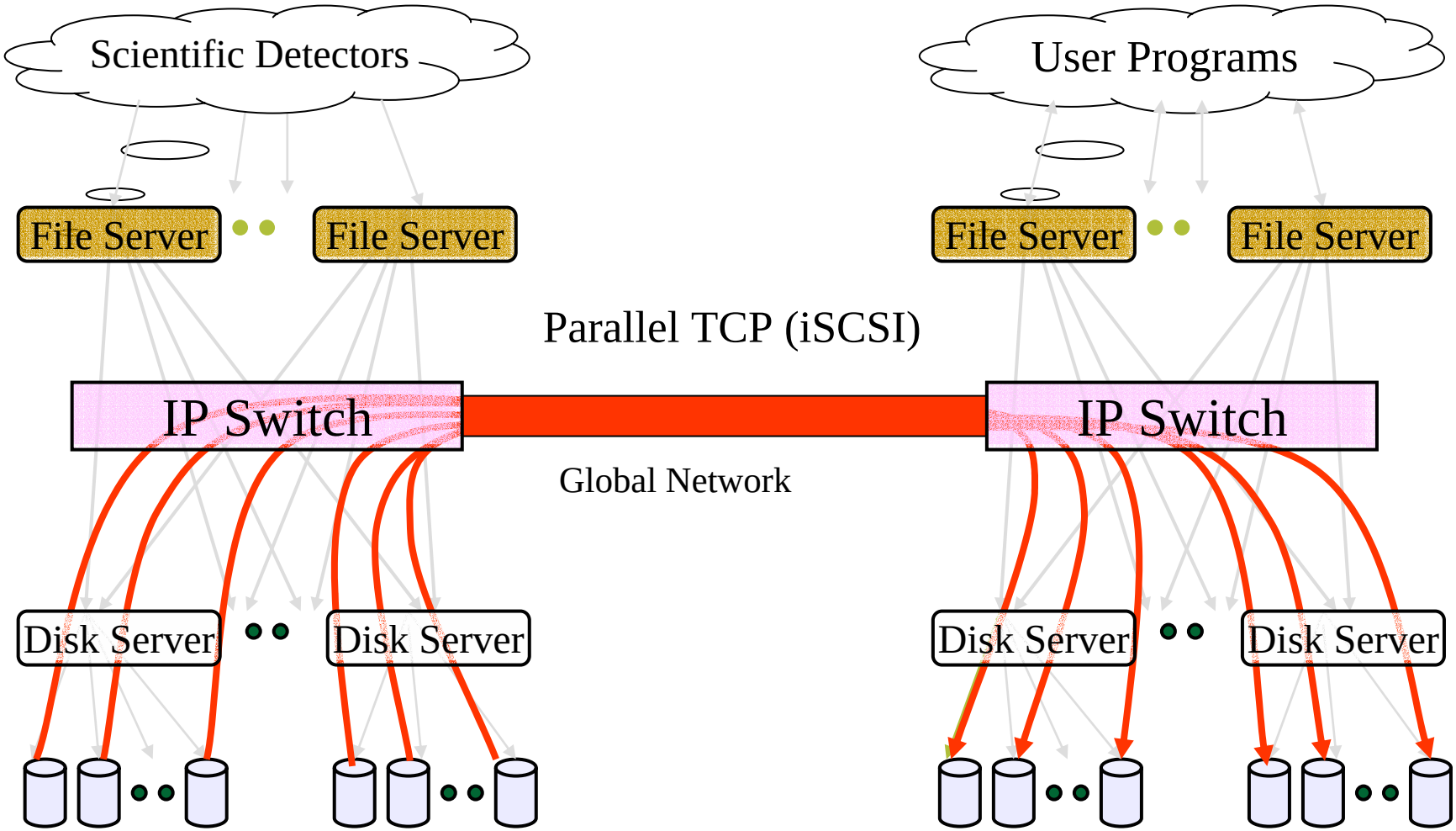
Features of Data Reservoir

- Data sharing in low-level protocol
  - Use of iSCSI protocol
  - Efficient disk to disk data transfer
- Multi-level striping for performance scalability
- Local file accesses through LAN
- Global disk transfer through WAN
  - Unified by iSCSI protocol

File accesses on Data Reservoir



Global disk transfer on Data Reservoir



Observations at BWC2002

Results of SC2002 BWC

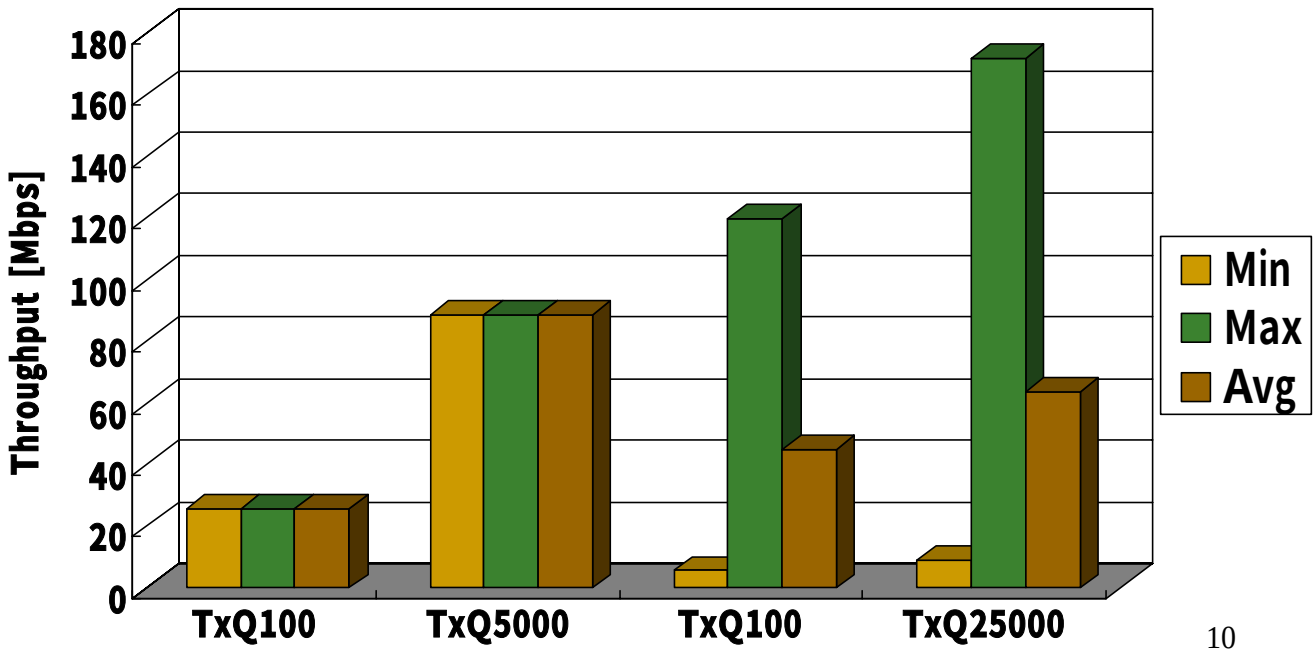
- 550 Mbps, 91% utilization
  - Bottleneck: OC-12, RTT: 200 ms
- Parallel “normal” TCP streams
- 24 nodes x 2 streams
- “Most Efficient Use of Available Bandwidth” award

Observations of SC2002 BWC

- But…
  - Poor performance per stream
    - Packet loss hits a stream too early during slow start
    - TCP congestion control recovers window too slowly
  - Unbalance among parallel streams
    - Packet loss occurs asynchronously & unfairly
    - Slow streams can’t catch up fast streams

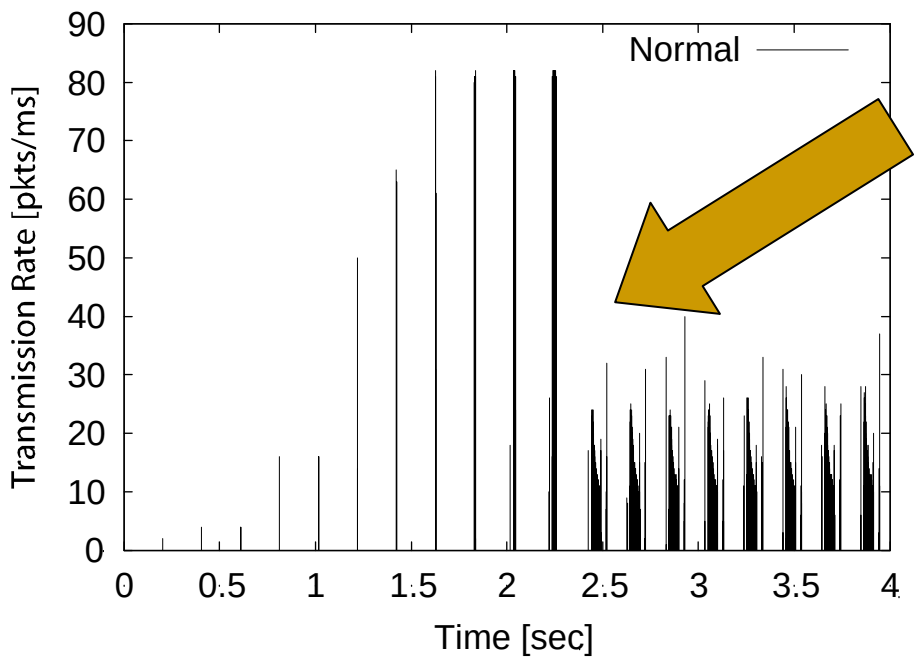
Transmission rate affects performance

- Transmission rate is important
  - Fast Ethernet > GbE
    - Fast Ethernet is “ultra” stable
    - GbE is “too” unstable and poor on average
  - Iperf
  - 30 seconds



Slow start makes burst

- Slow start
  - Double window of data every RTT
  - Send whole window burstly at the beginning of every RTT
  - Packet loss occurs even though huge idle period
    - Packets sent in 20 ms, nothing happen in 180 ms



Packet loss occurred

What's problems to solve?

- TCP/GbE on real LFN is quite unstable
  - Bursty transmission of packets

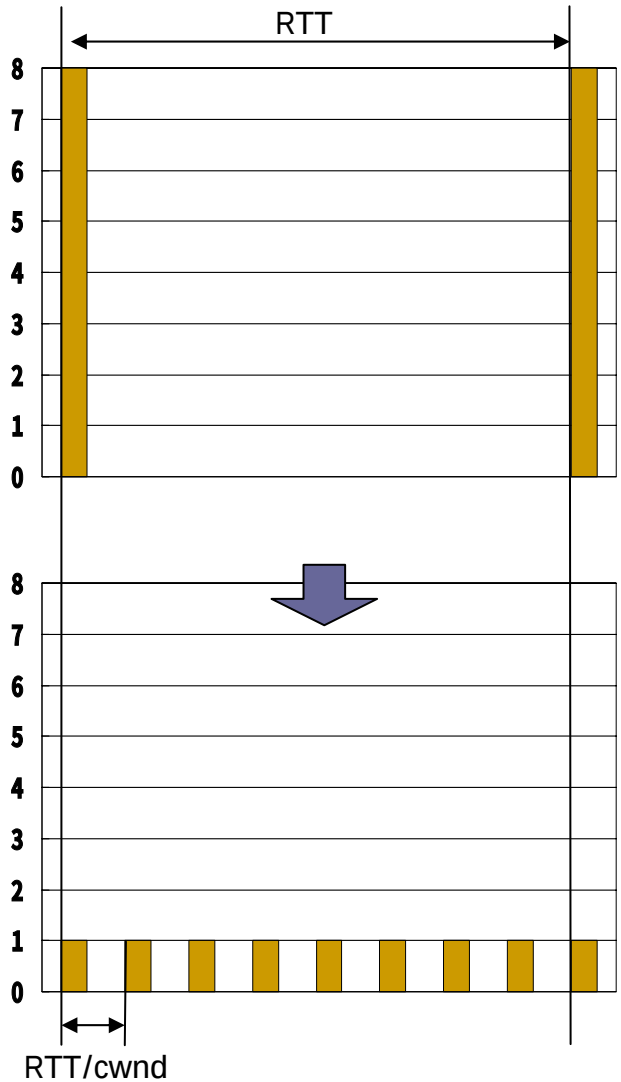
- Next Generation TCP
  - Aggressive but gentle window control algorithm
    - HighSpeed TCP, Scalable TCP, FAST TCP
  - Incorporated "rate control" feature
    - Reducing needless packet loss on underutilized network

Transmission Rate Controlled TCP

Transmission rate control for TCP

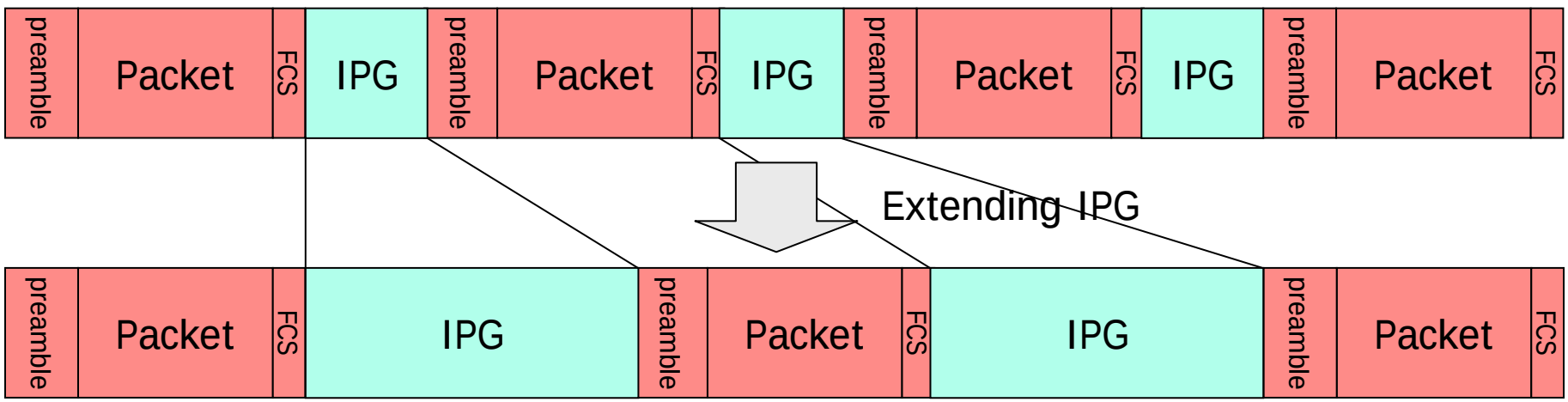
- Ideal Story
  - Transmitting a packet

    every RTT/cwnd
  - 24 us interval for 500Mbps
    - MTU 1500B
  - High load for software only



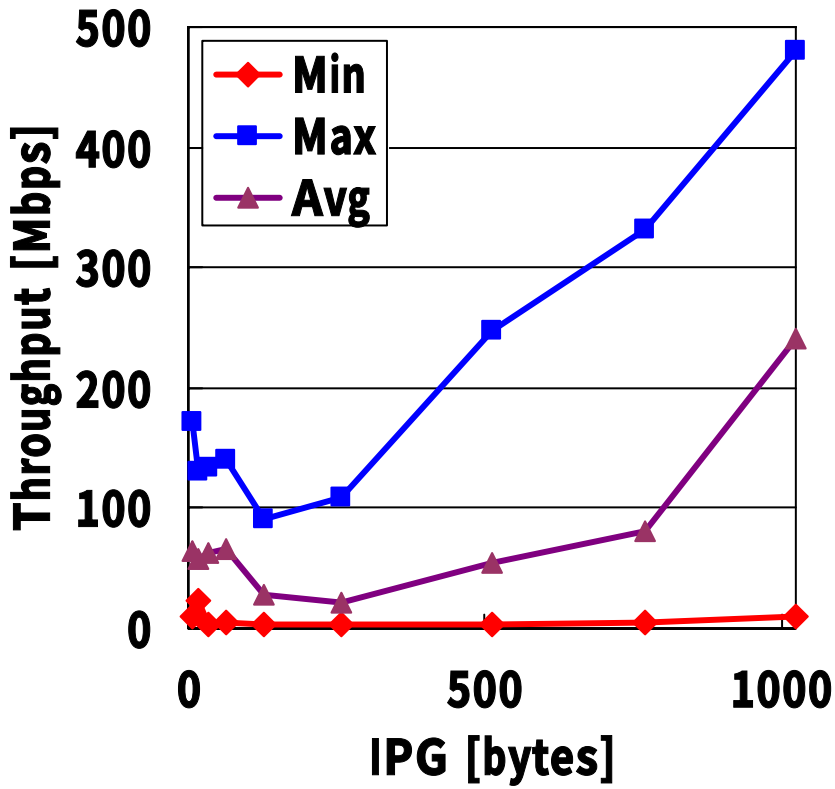
IPG tuning

- Inter Packet Gap (IPG) of Ethernet MAC layer
  - A time gap between packets
  - 8～1023B, 1B (8ns) step in case of Intel e1000
  - TCP stream
    - 941 Mbps ～ 567 Mbps
  - Fine grain, low jitter, low overhead



IPG tuning on GbE

- Bottleneck is 596 Mbps
  - RTT: 200 ms
- Improve in Max/Avg case using IPG 1023B
  - Transmission Rate < Bottleneck bandwidth
- Improve in Max case using IPG≧512B
- No effect in Min case

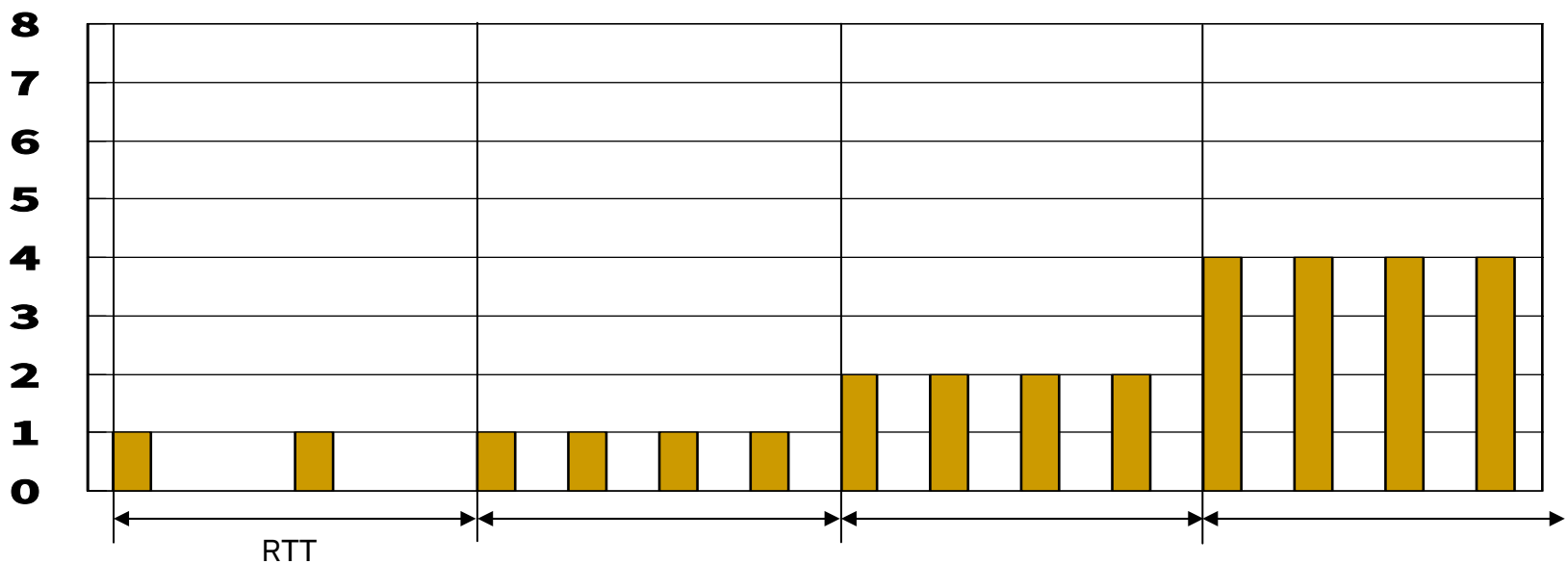


Clustered Packet Spacing (1)

- Insert transmission interval
  - Only during initial slow start
  - Using kernel timer In TCP stack of Linux kernel
    - Resolution: 1ms (Linux 2.6), 10ms (Linux 2.4)
      - Threshold value to transit to normal TCP
  - Coarse grain, low overhead
    - Spacing window of under 500 packets
    - Split burst into small fractions

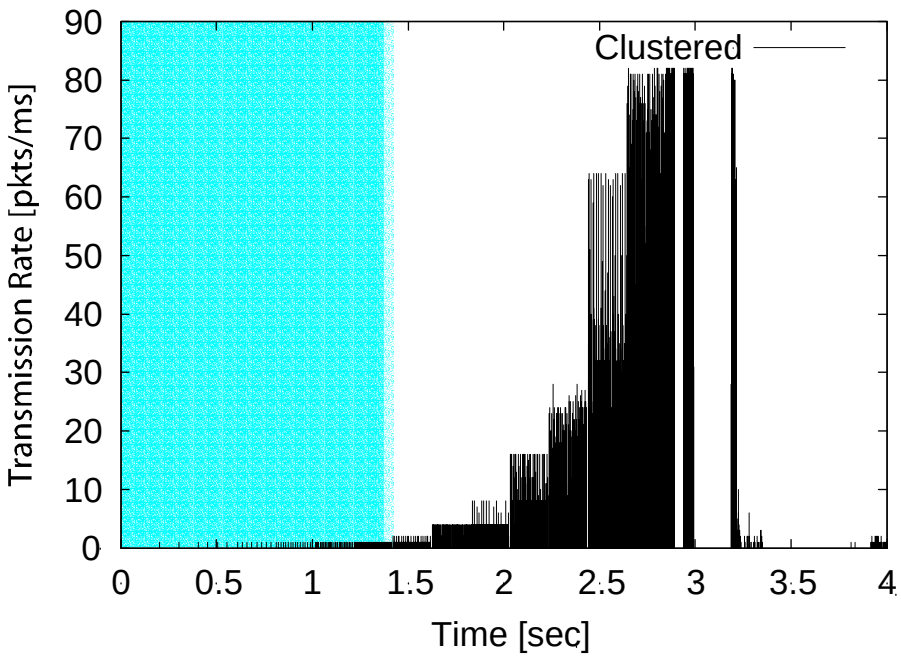
Clustered Packet Spacing (2)

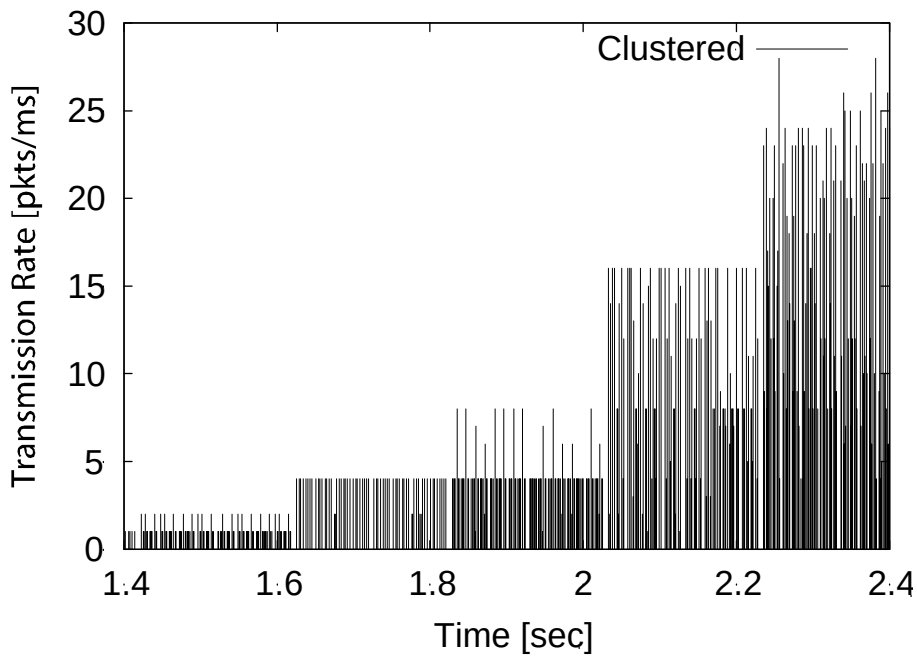
- RTT/cwnd > threshold
  - Rate control rules transmission timing
- RTT/cwnd < threshold
  - Normal TCP congestion control takes over



Slow start of CPS Linux 2.6

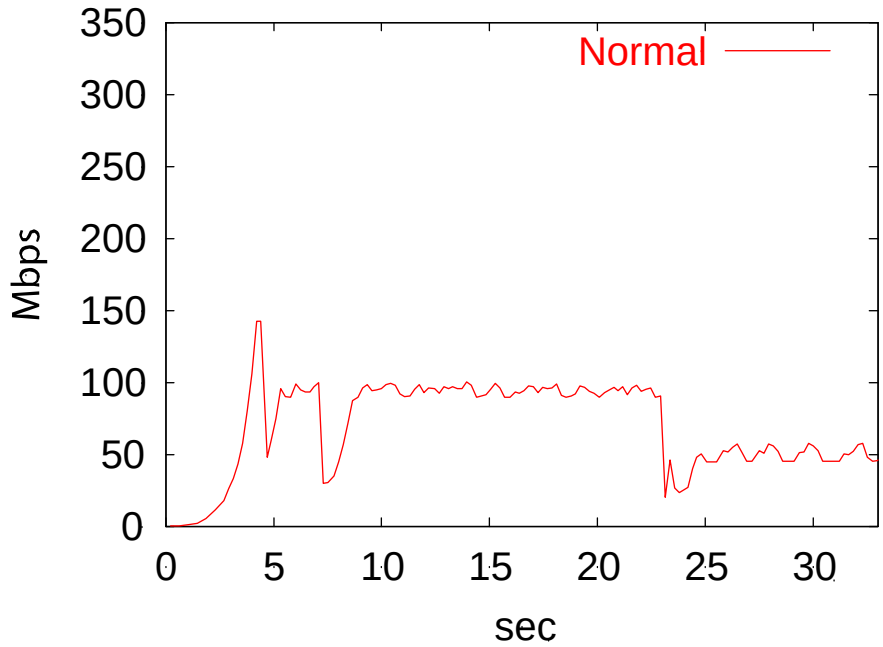
- Rate control while cwnd/RTT > 1ms
  - Blue shaded part
  - Split burst into 200 small bursts
  - Each small bursts is limited up to 80 packets

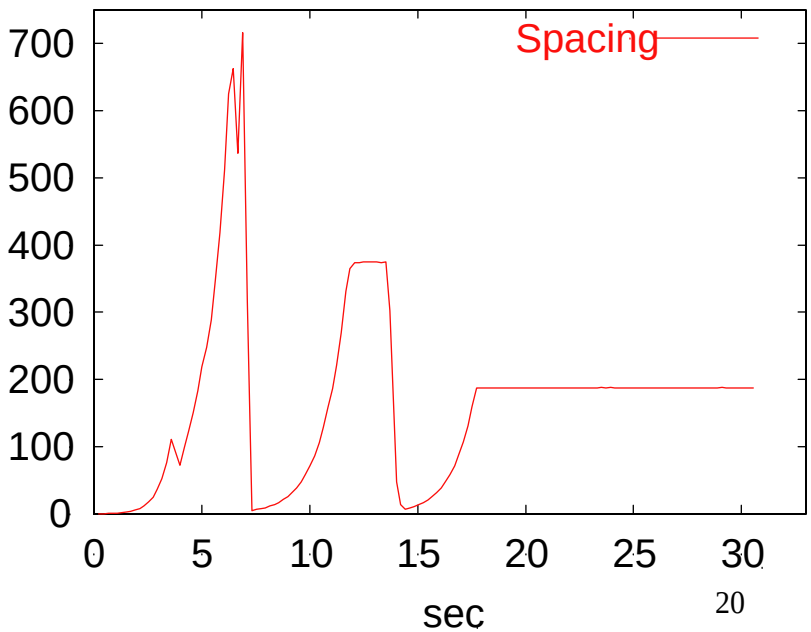




CPS TCP

- CPS can make cwnd bigger when initial slow start
- Current slow start is too aggressive
  - Causing packet losses difficult to recover using Fast Retransmit

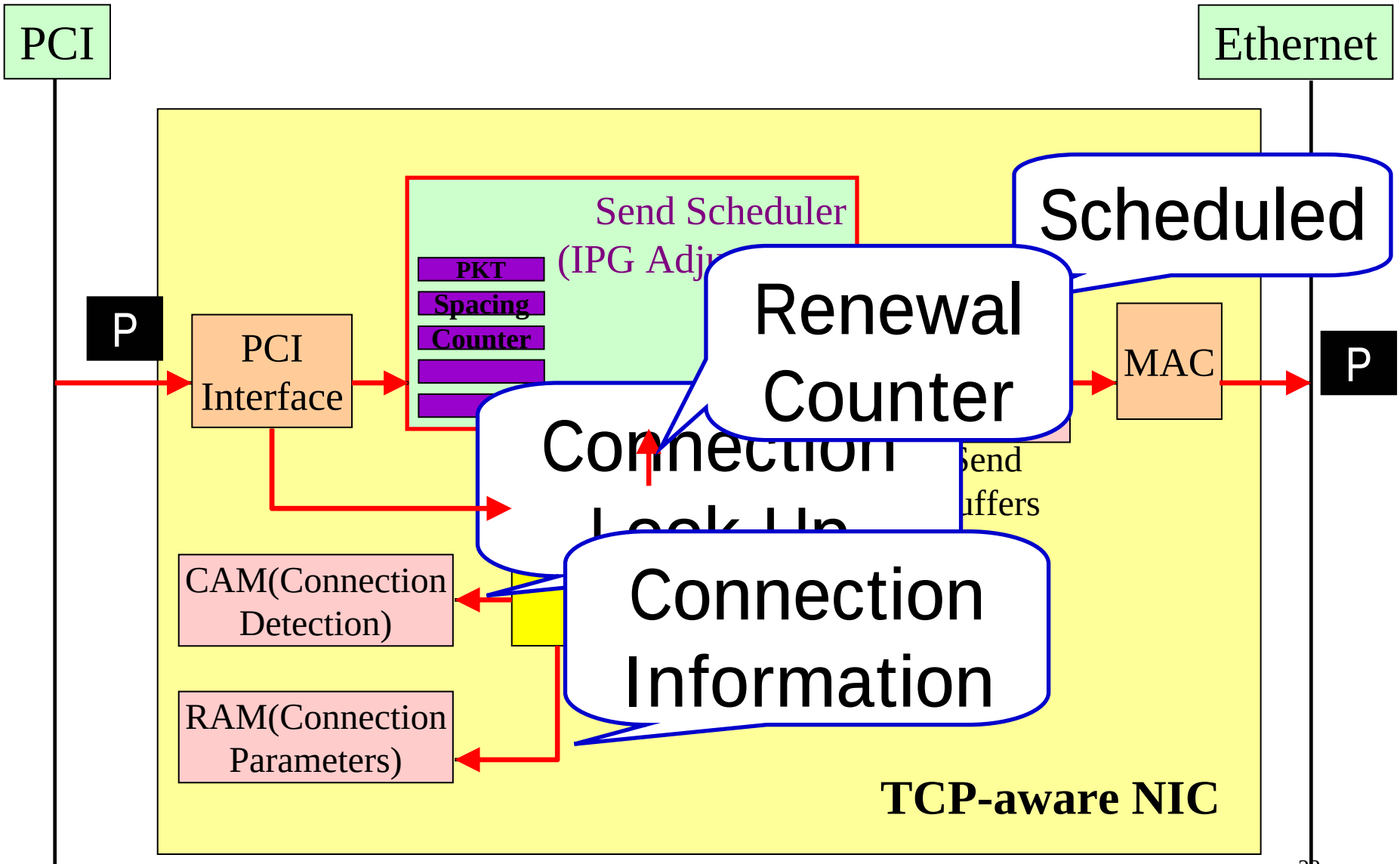




TCP-aware NIC

- Recognizing TCP parameters per stream
  - cwnd
  - RTT
- Adjusting IPG for flow-level rate dynamically
  - IPG = RTT / cwnd – PacketSize / BW
- Mixing multiple streams
  - Advantage against IPG tuning
- Real-time scheduling of packet transmission
  - Resolution of micro-, nanosecond order
  - Multi-interval deadlines
    - Hardware’s matter to deal with

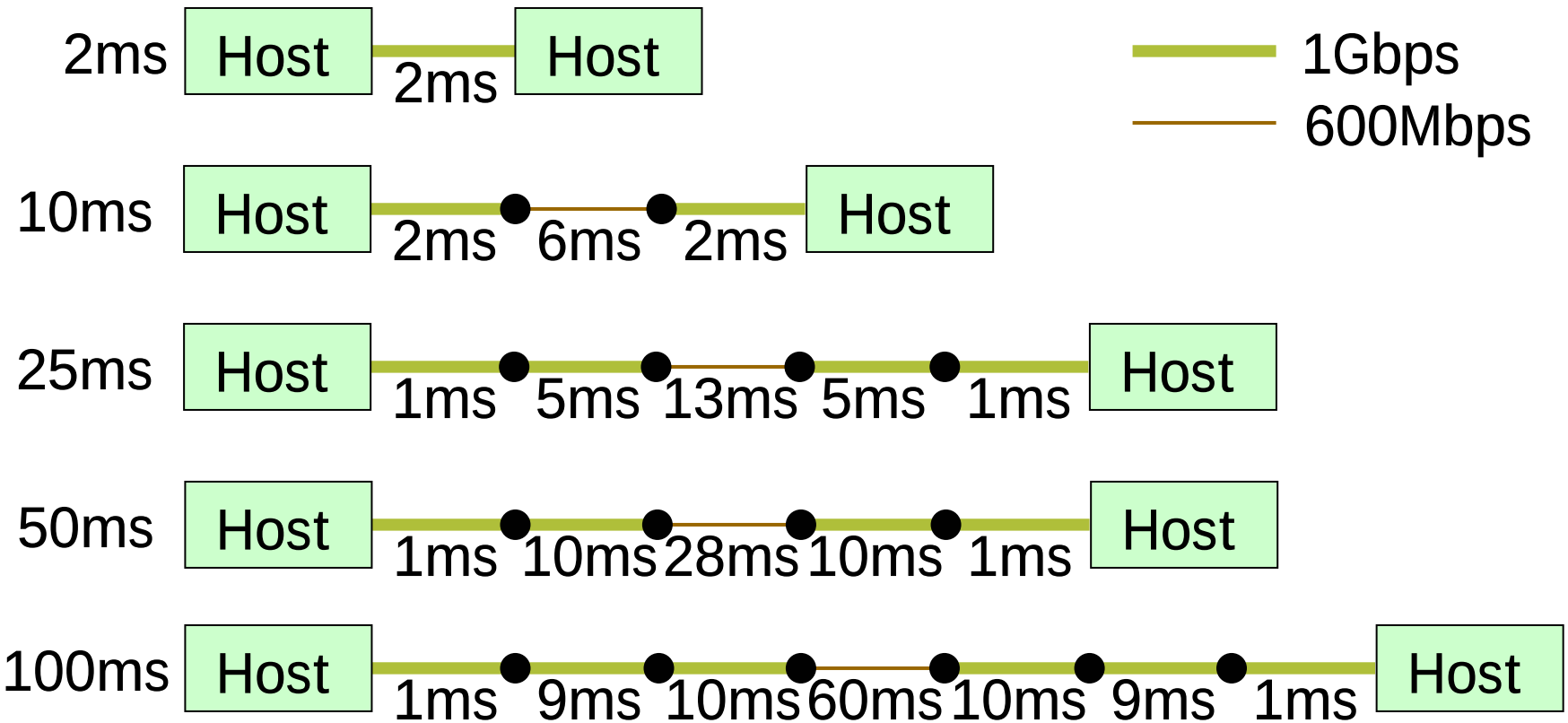
Functional diagram TCP-aware NIC



Simulation evaluation

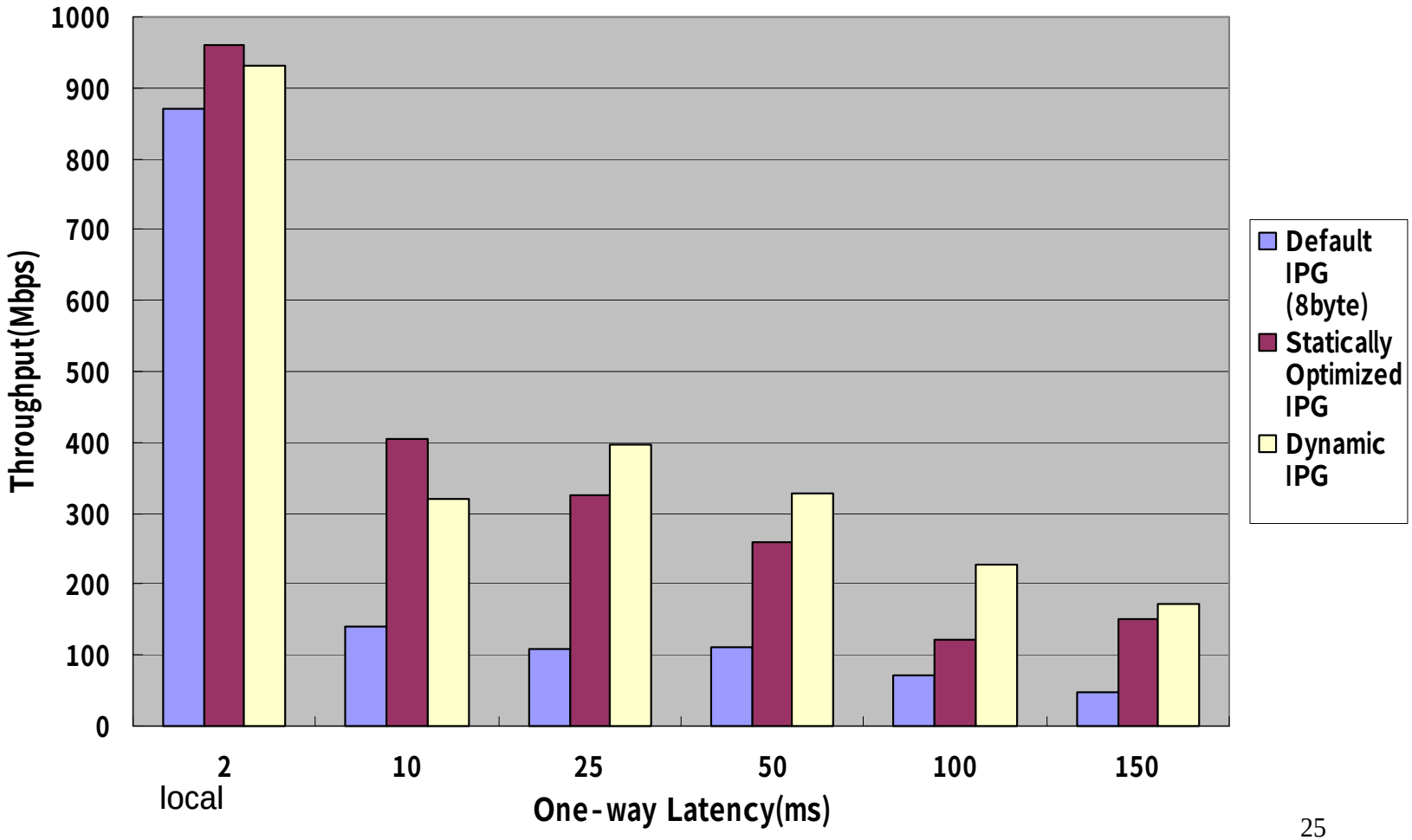
- Parameter
  - One-way Latency (ms)
    - 2, 10, 25, 50, 100, 150
  - IPG
    - 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384
    - Dynamic IPG
- Packet Loss Rate
  - Wire Loss Rate : 0.001%
  - Increased when remaining buffer is small
  - Increased when transmission is bursty
- Single stream

Network configuration



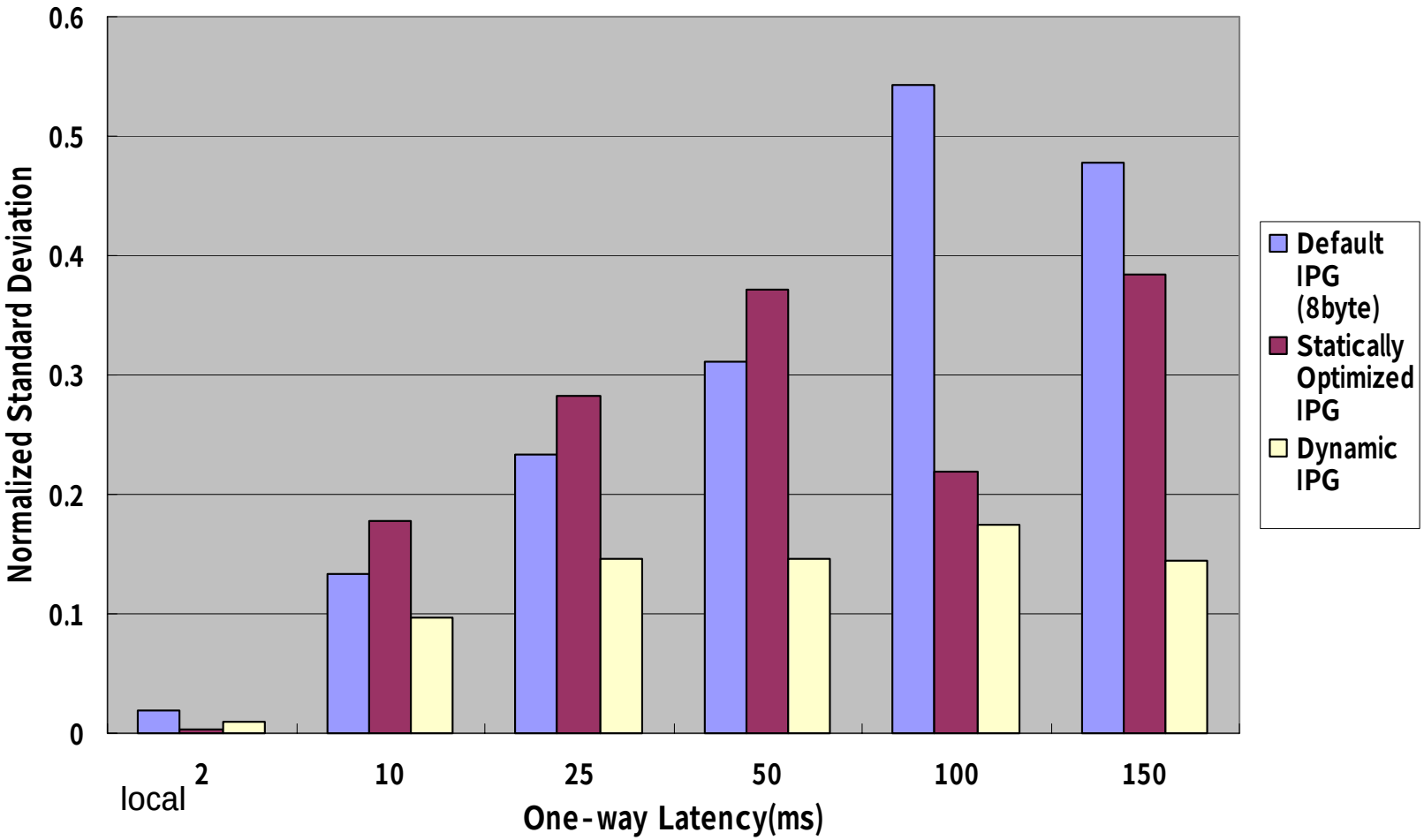
Dynamic / static optimization of IPG (1)

- Throughput



Dynamic / static optimization of IPG (1)

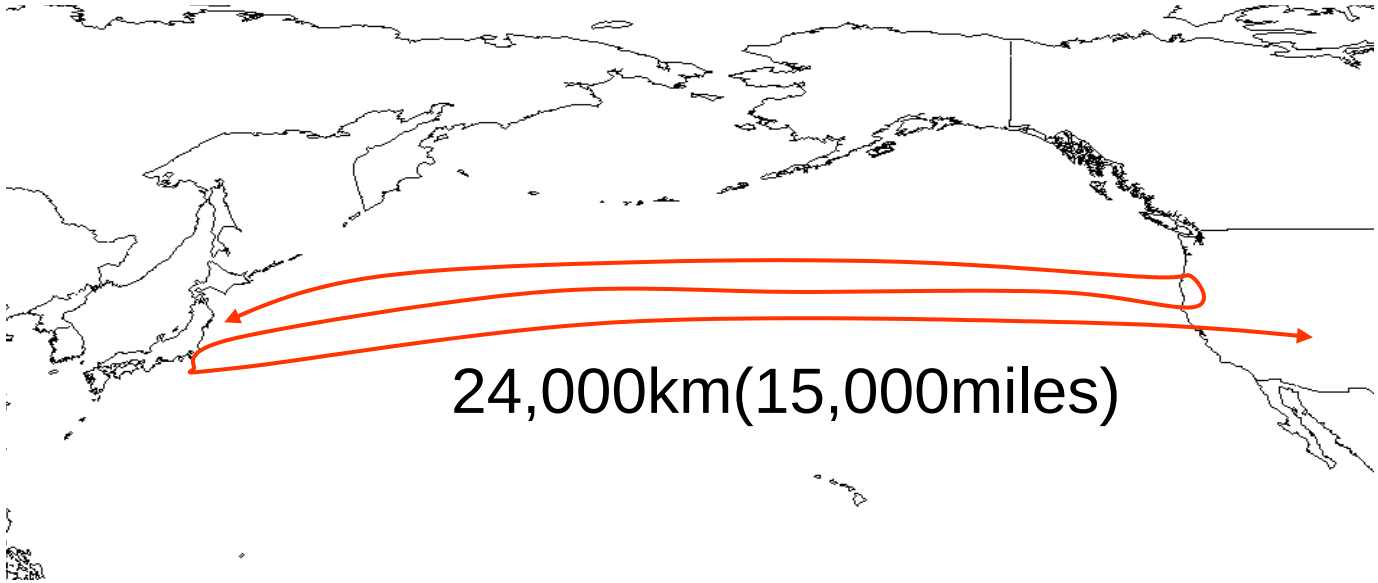
- Normalized standard deviation



Dynamic IPG optimization

- Gains higher throughput
  - When long latency (> 25 ms)
    - than “Default” and “Statically Optimized”
  - When small latency
    - than “Default”
- Gains lower throughput
  - When small latency
    - than “Statically Optimized”
      - Degraded 3% on 2 ms latency, 21% on 10 ms latency
- Can Stabilize throughput
  - Smaller normalized standard deviation
    - than “Default” and “Statically Optimized”

Results at BWC2003



OC-192

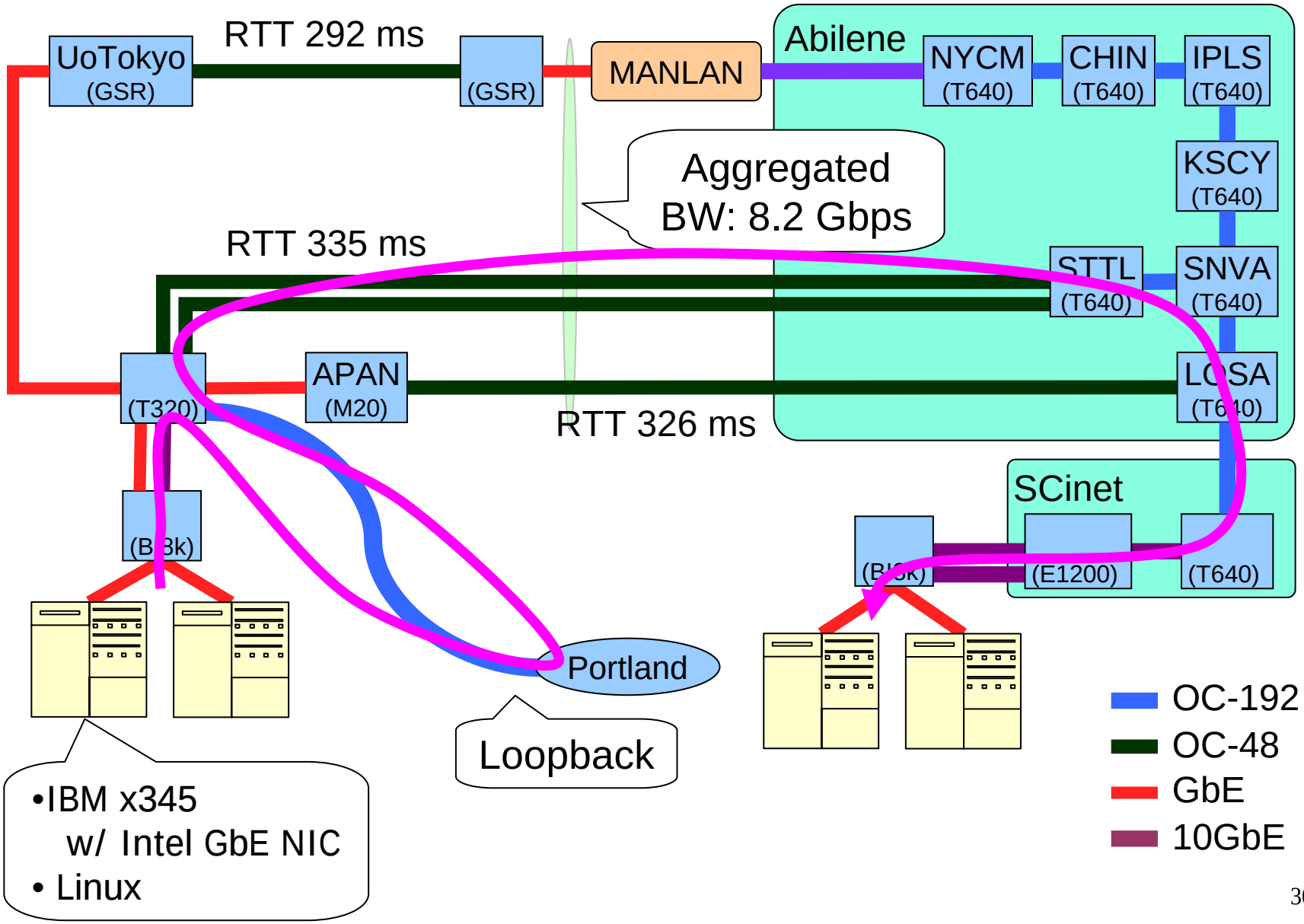
15,680km (9,800miles)

Juniper
T320

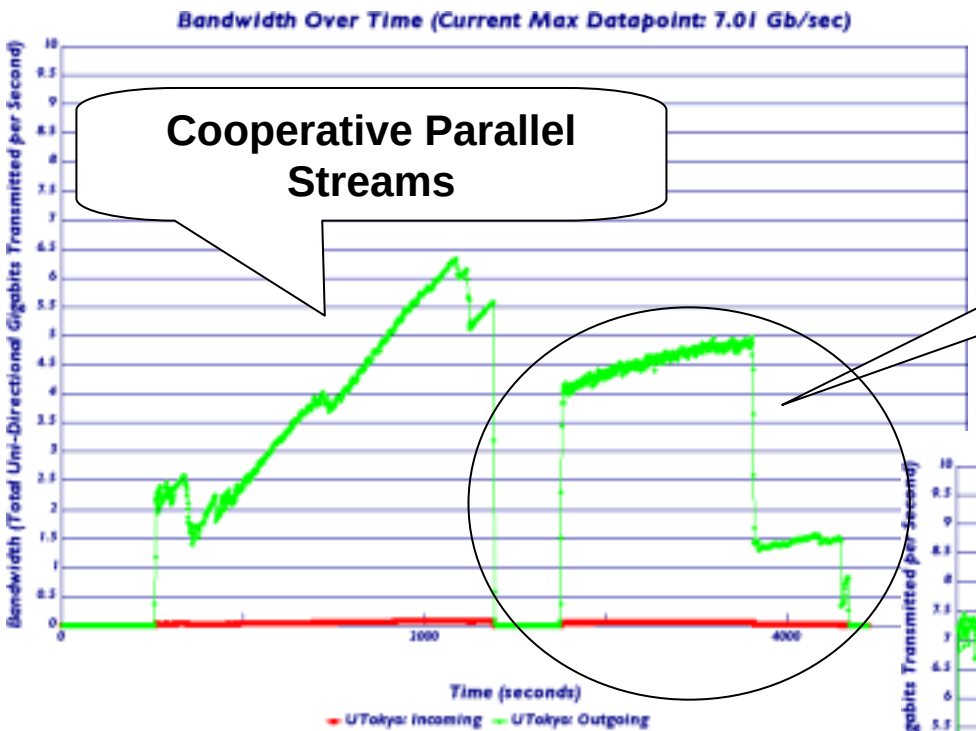
OC-48 x 3
GbE x 1

8,320km
(5,200miles)

Network configuration (During SC2003)

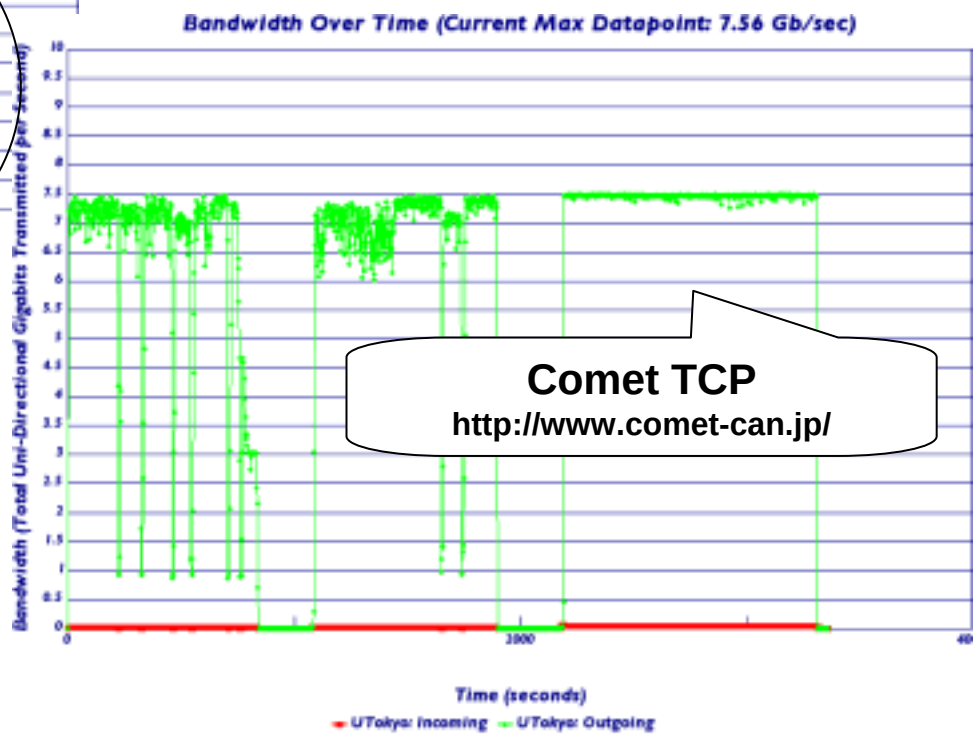


SC2003 BWC



Cooperative Parallel Streams

IPG-tuned Parallel Streams



Comet TCP
http://www.comet-can.jp/

http://scinet.supercomp.org/2003/bwc/results/

Results of SC2003 BWC

- SC2003 BWC
  - Bottleneck: 2 x OC-48, OC48 (3 x GbE), GbE
  - RTT: 335 ms, 326 ms, 292 ms
- Parallel IPG-tuned TCP streams
  - 16 nodes x 4 streams
  - Maximum throughput: 5.42 Gbps
- “Distance x Bandwidth Product and Network Technology” award

Our contribution

- Highlight importance of rate control for TCP
  - Alleviating bursty behavior on GbE
  - Reducing needless packet loss on underutilized network
- Demonstrate real data transfer on real LFN
  - Disk-to-Disk file transfer
    - Utilization of low-level data sharing
      - Using iSCSI protocol
    - Transmission rate control in TCP stack
    - Parallel streams

Conclusion

- Transmission Rate Controlled TCP
  - Stabilize and improve performance
- IPG tuning
  - Static, low overhead, easy to use
- Clustered Packet Spacing
  - Flexible, feasible with little overhead
- TCP-aware NIC
  - Dynamic, low overhead
  - Adaptable to heterogeneous streams simultaneously

BWC 2003 Experiment is supported by