Packet Switch Architectures for Very Small Optical RAM

Onur Alparslan, Shin'ichi Arawaka, Masayuki Murata

Advanced Network Architecture Laboratory
Graduate School of Information Science and Technology

Osaka University

26 August 2009

Outline

- Problem Statement
- Objective
- Proposed Solutions
- Switch Architecture
- Simulations
- Conclusions
- Future Research

Problem Statement

- **Major differences and limitations between Optical packet-switched (OPS) networks and electronic packet-switched (EPS) networks.**
- **In EPS networks, contention is resolved by**
  - Storing the contended packets in a random access memory (RAM)
- **Limitations in optical domain,**
  - Optical to electronic domain in order to use electronic RAM is not a feasible solution, because of the processing limitations of EPS.
  - Processing and switching in the optical domain is necessary.
- **Buffering in the optical domain**
  - Fiber Delay Lines (FDL)
    - FDLs require very long fiber lines, which cause signal attenuation, inside the routers.
    - There can be a very limited number of FDLs in a router due to space considerations, so they can provide a small amount of buffering
  - Optical RAM
    - Still under research
    - Not expected to have a large capacity, soon
- **TCP has low throughput due to burstiness, when buffer is very small**

Objective

- **Designing an all-optical OPS network architecture that can achieve high utilization and low packet drop ratio by using very small Optical RAM buffers**
- **Show and compare the buffer requirements**

Advantages

- **Decreasing the buffer requirements in the core**
- **Realizing all-optical high-speed OPS networks**

TCP Pacing

- Evenly spacing transmission of a window of TCP packets over a round-trip time (RTT)
  - **Packets are injected into the network at the desired rate of W/RTT when W is congestion window size.**
  - **Smoothing the traffic**
- It is shown that O(logW) router output buffer size is enough for high utilization when Paced TCP is used
  - **Aggregate paced TCP traffic converges to poisson**
- Requires changing the TCP senders

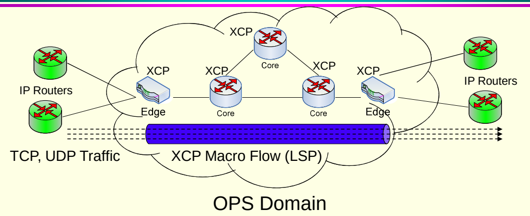
M. Enachescu, Y. Ganjali, A. Goel, N. McKeown, and T. Roughgarden, "Part III: Routers with very small buffers," ACM SIGCOMM Computer Communication Review, vol. 35, pp. 83–90, 2005.

XCP-Based Proposed Solutions 1/2

- **Preventing wavelength over-utilization**
  - Apply XCP-based congestion control
    - XCP is a new congestion control algorithm specifically designed for high-bandwidth and large-delay networks.
    - Network layer control
    - Nodes exchange probe packets in order to learn link information
    - Uses an efficiency controller for high link utilization and fairness controller for high fairness among flows
  - Carefully select XCP parameters
  - Control maximum wavelength utilization ratio by XCP

D. Katabi, M. Handley, and C. Rohrs, "Congestion control for high bandwidth-delay product," in *Proceedings of ACM SIGCOMM*, 2002, pp. 42-49.

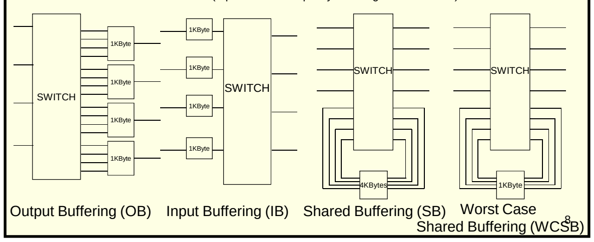
XCP-Based Proposed Solutions 2/2



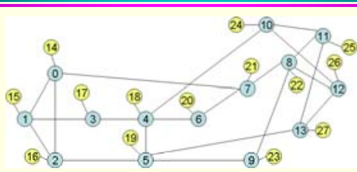
- **Burstiness**
  - Establish macro flows between edge nodes
  - Assign incoming TCP, UDP traffic to macro flows (similar to XCP-CSFQ, TeXCP)
  - Apply leaky bucket pacing to macro flows according to XCP flow rate at edge node
  - Possible to use LSPs for controlling macro flows if GMPLS is available

Buffer and Switch Architectures

- Shared Buffering
  - Total buffer size in a node increases linearly with the number of links
  - For example, when buffer size per link is 1KByte, a node with 4 links has 4Kbytes Shared Buffer
  - Total buffer size inside the switch is the same as OB and IB. Only buffer
  - placement is different
- Worst Case Shared Buffering
  - Total buffer size is constant (equal to buffer capacity of a single OB or IB link)

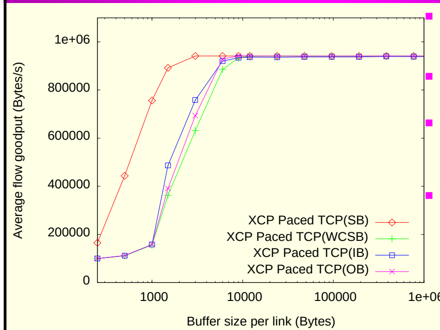


NSFNET Simulations



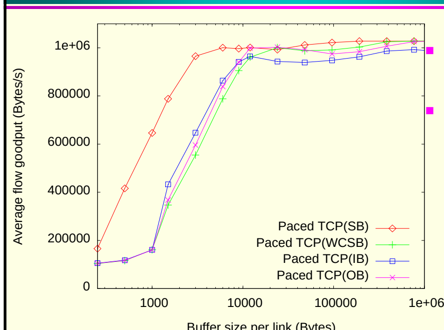
- 28 nodes (14 edge + 14 core) and 35 links (21 core + 14 edge)
- Wavelength speed 1Gbit/s
- 40 seconds simulation (use last 5 seconds for results)
- 1587 TCP Reno flows (Poisson flow arrival)
- TCP maximum congestion window size is 20 packets
- Data packet size (MSS) is 1500 Bytes
- Optical RAM
- Cut-through optical packet switching and buffering
- Evaluate average goodput of TCP flows

XCP Pacing (separate ACK macro wavelength)



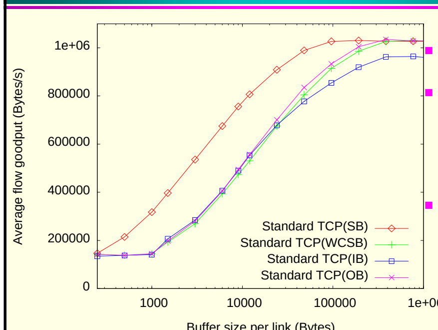
- **More than 3 times higher goodput with SB**
- **IB has higher goodput than OB**
- **WCSB goodput is close to OB with a much smaller buffer**
- **IB, OB, WCSB have the same goodput when buffer size is less than MSS**

TCP Pacing



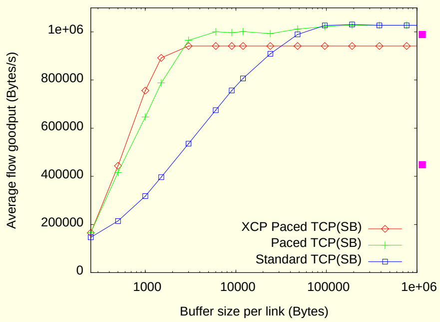
- **Similar to XCP Paced TCP**
- **When buffer is large, IB has the lowest goodput due to head of line blocking**

Standard TCP



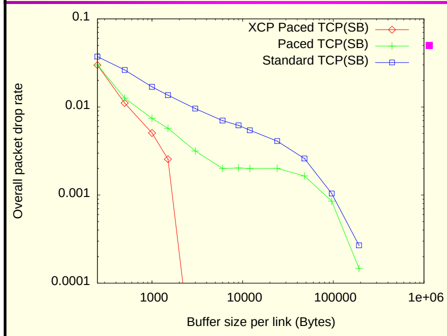
- **Lowest goodput**
- **OB, IB and WCSB give almost the same throughtput when buffer is small**
- **When buffer is large, IB has the lowest goodput due to head of line blocking**

Goodput Comparison of Pacing Methods



- **XCP Pacing gives the highest goodput when buffer size is very small (less than MSS)**
- **XCP Pacing has a better fairness, so maximum average goodput is lower**

Packet Drop Rate inside Core Network



- **XCP Pacing has a much lower packet drop rate**

Conclusions

- When buffers are very small, XCP-based paced standard TCP flows can achieve higher goodput and lower packet drop rate than TCP pacing
- When the total buffer capacity in a node is the same, the shared buffering with XCP pacing has much better performance than the input and output buffering
- The performance of worst case shared buffering is close to the output buffering even though worst case shared buffering uses much less buffer per node

Future Work

- NSFNET nodes mostly have a small nodal degree of 3 to 4, so worst case buffering shows good performance
  - Simulate topologies with a higher nodal degree like Abilene topology
- More realistic traffic models
- Buffer requirements of WDM

Thank you