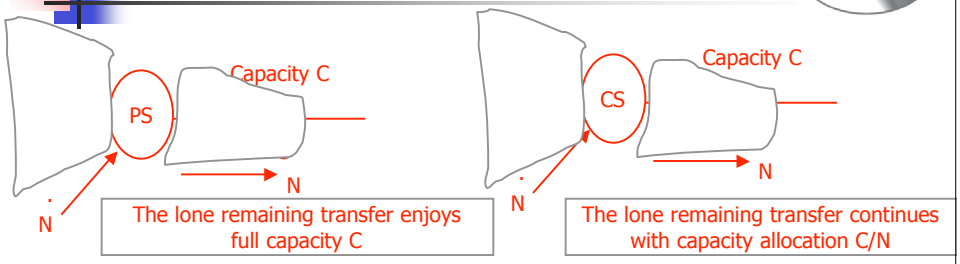
Scheduling and transport for file transfers on high-speed optical circuits

Authors: M. Veeraraghavan & Xuan Zheng (University of Virginia)
Wu Feng (Los Alamos National Lab)
Hojun Lee (Polytechnic University)
Edwin Chong & Hua Li (Colorado State University)

Outline

- Problem statement
- Varying-Bandwidth List Scheduling (VBLS)
- Varying-Bandwidth Transport Protocol (VBTP)
- Conclusions and future work

Drawback of using circuits for file transfers



- PS: Packet switch
- CS: Circuit switch
  - Fixed bandwidth scheme

3

VBLS: A Lambda-Scheduling Algorithm for File Transfers

- End host applications request lambdas for file transfers by specifying a three-tuple $(F, R_{\max}, T_{req})$
  - F : file size
  - $R_{\max}$: a maximum bandwidth limit for the request
  - T_{req}: the desired start time for the transfer
- The scheduler assigns a Time-Range-Capacity (TRC) vector $\{(B_k, E_k, C_k), k = 1,2,...,\tau\}$ for each transfer
  - B_k : the start of the kth time range
  - E_k : the end of the kth time range
  - C_k : the capacity allocated for the transfer in the kth time range.

4

VBLS: an example

Assume the available capacity of a 4-channel link is as shown below

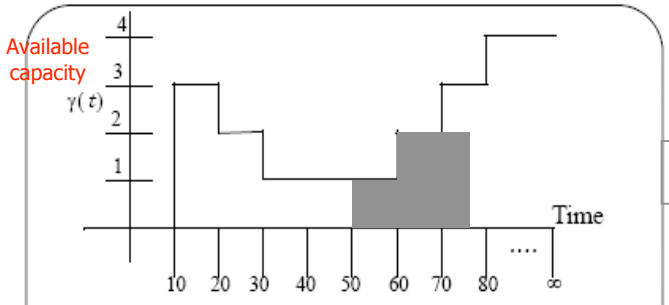


Figure 1. Example of $\gamma(t)$, $P_1 = 0$, $\gamma(0) = 0$, $P_2 = 10$, $z_{max} = 9$, and $P_{z_{max}} = P_9 = 80$.

F: 5GB
Rmax: 2 channels
Treq: 50

Per-channel rate: 10Gbps
Time unit: 100ms

In 10 time units can transfer 1.25GB

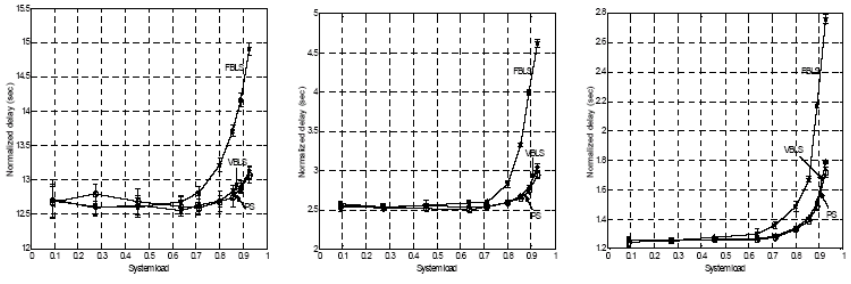
TRC allocated:
(50, 60, 1)
(60, 70, 2)
(70, 75, 2)

5

VBLS: Simulation comparison of VBLS against FBLS and PS

- Normalized delay (D)
- System load (ρ)

$$D = \frac{\sum_i F^i d^i}{\sum_i F^i} \qquad \rho = \frac{\lambda \overline{F}}{C}$$



(a) R^i_{max} of 1 channel (b) R^i_{max} of 5 channels (c) R^i_{max} of 10 channels

6

Discussion

- Advantages of VBLS
  - Achieves close to idealized PS (infinite buffer) performance
- Disadvantages of VBLS
  - Complexity: reprogram switches multiple times within one file transfer time
  - Currently not for heterogeneous traffic
- Apriori knowledge of "available bandwidth"
  - run-time discovery of optimal sending rates not needed as with TCP enhancements

7

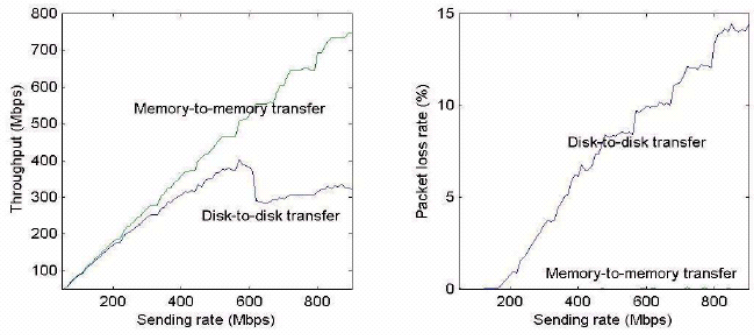
VBTP: overview

- Flow control: rate based scheme
  - easier said than done!
  - indeed rate available through the network is not an issue – it remains constant in circuit env.
  - but, sending and receiving hosts are general-purpose hosts with generic OSs that schedule not only networking tasks but other tasks as well
- Error control: selective-ARQ scheme
- Congestion control: not required during the data transfer

8

VBTP: flow control

- The problem with rate-based flow control (ideal rate?):
  - Play it safe and set a low rate: avoid/eliminate receive-buffer losses
  - Or send data at higher rates but have to recover from losses
- Experiments with SABUL implementation



(MTU=1500B, UDP buffer size=256KB, SABUL data block size=7.34MB)

9

VBTP: error control

- Selective ARQ to recover from losses due to link errors and receive-buffer overflows
- Will negative ACKs suffice since circuits offer in-sequence delivery?
  - No, if disk access rates are low - performance better if a retransmission buffer is used
    - Implication: Need positive ACKs to keep removing data from retransmission buffer
- Be utilization obssessed:
  - Drop circuits immediately after completion of transfer
  - Implication: Errors identified after the last block is sent are handled by retx. on TCP/IP path (CHEETAH paper from last PFLDN workshop)

10

VBTP: VBLS-induced effects

- TRC allocation should be determined not just for the initial file transfer but also for retransmissions
  - Errors from receive-buffer overflows should be allowed to achieve high rates (Solution: F+ε)
- The sender may not be able to send the data at exactly the rates specified in the TRC vector
  - Due to OS scheduler at end hosts not "honoring" application-set data rate at which blocks are passed to the Ethernet driver for transmission

11

Conclusions and future work

- VBTP overcomes a well-known drawback of using circuits for file transfers in which with a fixed-bandwidth allocation mode fails to allow users to take advantage of bandwidth that becomes available subsequent to the start of a transfer
- Simulations showed that VBLS can improve performance over fixed-bandwidth schemes significantly for file transfers
- The transport protocol that works in conjunction with VBLS should be a rate based, flow-control scheme along with a selective-ARQ based, error-control scheme
- Future work: to include a second class of user requests for lambdas, specifically targeted at interactive applications such as remote visualization and simulation steering

12