96-0177R1

TCP/IP over ABR

[Was: TBE and TCP/IP Traffic]

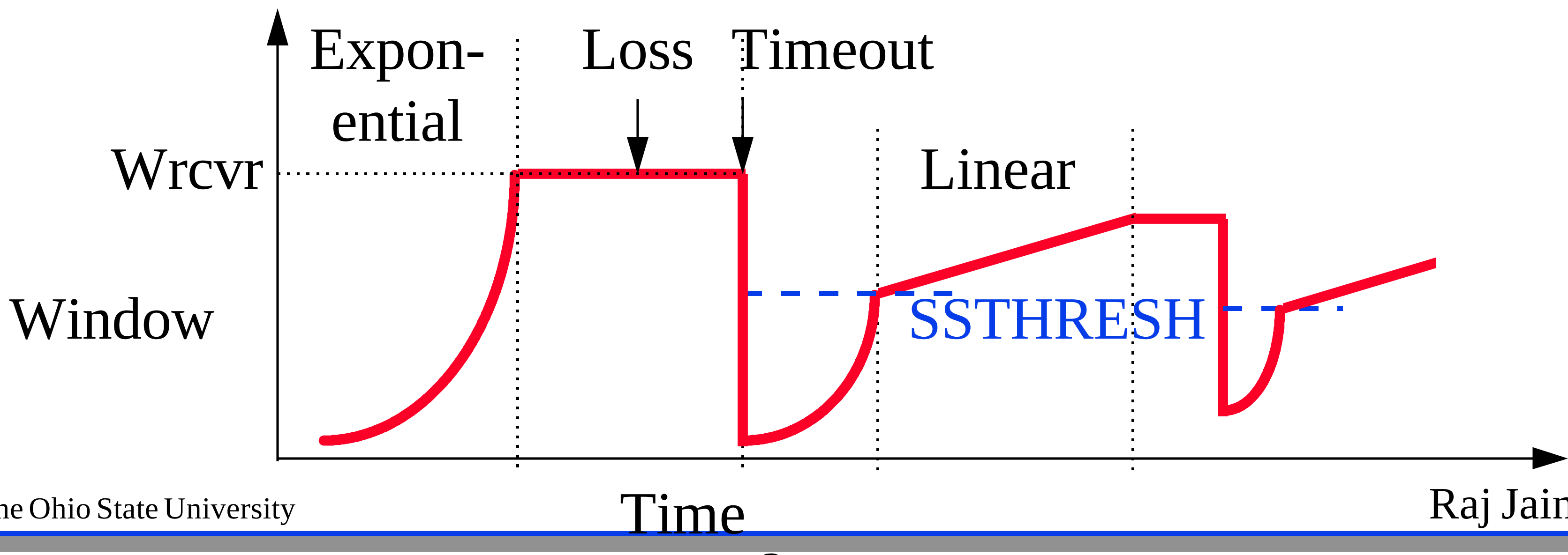
Raj Jain, Shiv Kalyanaraman, Rohit Goyal, Sonia Fahmy, Fang Lu
The Ohio State University
Saragur M. Srinidhi
Sterling Software and NASA Lewis Research Center

Raj Jain is now at Washington University in Saint Louis,
jain@cse.wustl.edu http://www.cse.wustl.edu/~jain/

- TCP/IP's load control mechanisms
  Slow-start, Timeout, Retransmissions
- Simulation Results
  ABR + Finite buffers + 100 ms granularity + WAN
- Effect of TBE and finite buffers
- Effect of timer granularity, tail drop, VBR

TCP/IP Slow Start

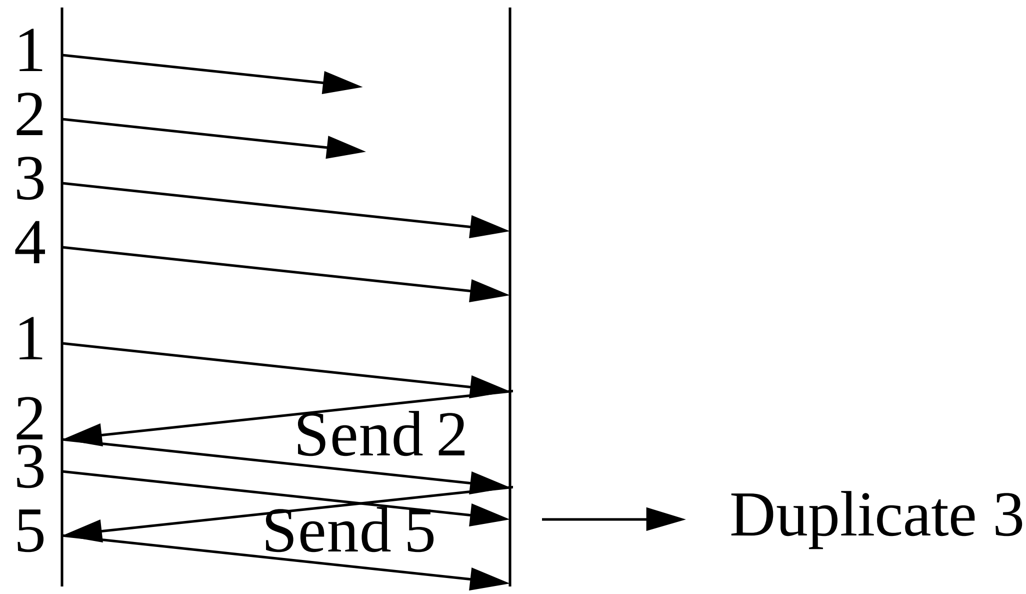
- Maximum Segment Size (MSS) = 512 bytes
- Congestion Window (CWND)
- Window W = Min{Wrcvr, CWND}
- Slow-Start Threshold = max{2,min{CWND/2,Wrcvr}}
- Exponential until SSTHRESH: W = W+1 for every ack
- Linear afterwards: W = W + 1/W for every ack until Wrcvr



Timeout and Timer Granularity

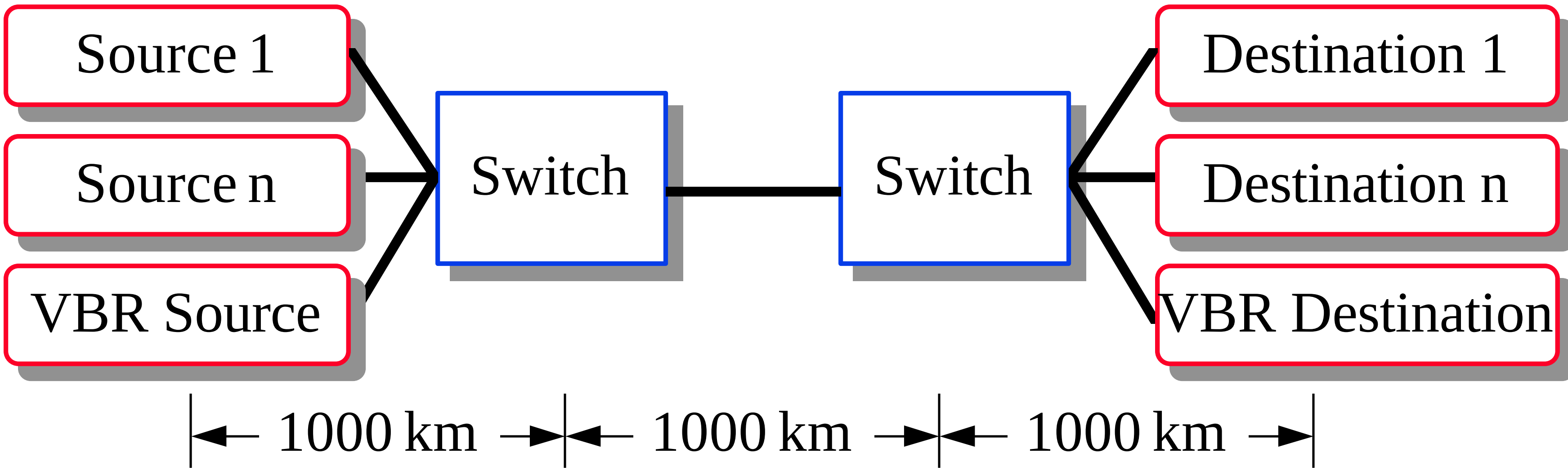
- Remember segment # and Send_time
- Upon acknowledgment: RTT = Now - Send_time
- Keep an exponential average of mean and std. dev. of RTT
- Retransmissions ⇒ Ignore the measured value
  Cumulative Ack ⇒ Use it as usual
- Timeout = Mean + 4 × Std. Dev.
- Only one packet is timed
- All times are measured using a granularity of 100 ms
  (500 ms in Solaris and all BSD implementations)
- RTT < 100 ms ⇒ RTT = 100 ms
- Upon retransmission: Timeout = 2 × Timeout until 128 ticks

Packets Dropped at Destination



- On every loss of n packets, time lost = Timeout + fn(n) RTT

n Source + VBR Configuration



- All links 155 Mbps
- If VBR background , 100 ms on, 100 ms off, start at t = 2ms
- All traffic unidirectional, Large file transfer application
- Parameters: # sources={2, 5}
  Buffer size = TBE × # sources × {1, 2, or 4}

Simulation Parameters

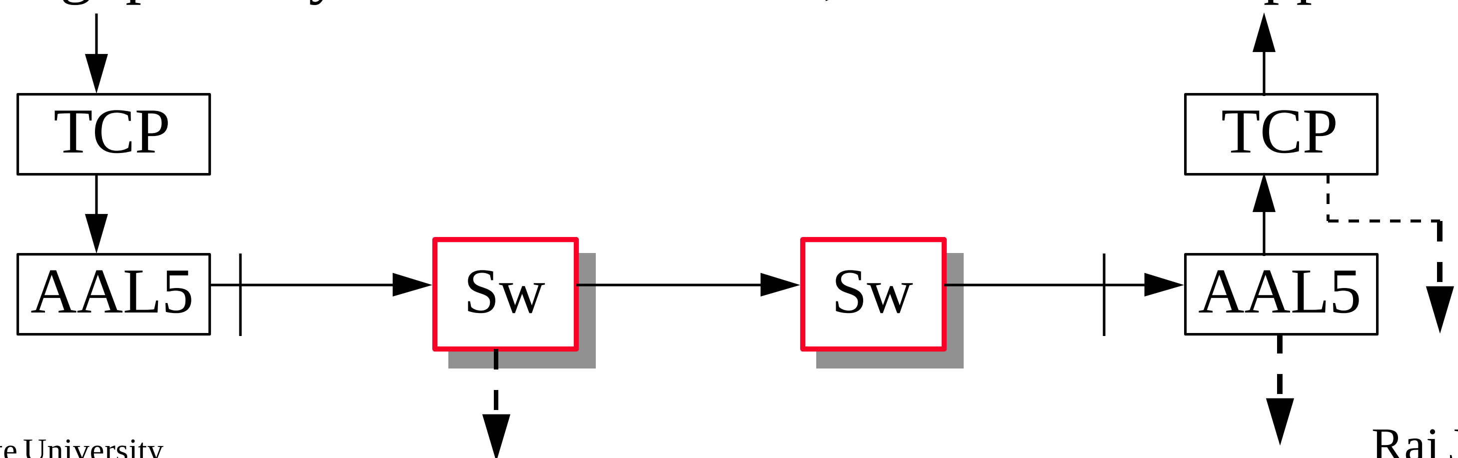
- Source: Parameters selected to maximize ACR
  TBE = 128, 512
  CDF (XDF) = 0.5
  ICR = 10 Mbps
  CRM (Xrm)= ⌈ TBE/Nrm ⌉
  ADTF = 0.5 sec
  PCR = 155.52 Mbps, MCR= 0, RIF (AIR) = 1,
  Nrm = 32, Mrm = 2, RDF = 1/512, Trm =100ms, TCR = 10 c/s
- Traffic: TCP/IP with Infinite source application
- Switch: ERICA modified
  Target Utilization = 90%
  Averaging interval = min{100 cells, 1000 μs}

TCP/IP Parameters

- Maximum Segment Size = 512 bytes
- Timer granularity = 100 ms
- Fast retransmit/recovery not completely experimented
- Early packet drop (EPD) not yet experimented
- No TCP processing time
- Max window = 16×64 kB,
  One-way delay = 15 ms = 145 kB
- No ack delay timer

Performance Metrics

- Sequence numbers at the source, Congestion window
- ACR, Link utilization, Queue length in the switch
- Bytes sent = Sent once + Retransmitted
  = Bytes delivered to application
  + data bytes dropped in the switch + bytes in the path
  + Partial packet bytes dropped at the destination AAL5
  + duplicate packet bytes dropped at the destination TCP
- Throughput = Bytes delivered/Time, CLR = Cells dropped/sent



Infinite Buffers & Fixed Capacity

- Buffer size = 4096, TBE = 512
- CLR = 0
- Maximum TCP throughput = 103.32 Mbps
- Throughput = 155 Mbps
  × 0.9 for ERICA Target Utilization
  × 48/53 for ATM payload
  × 512/568 for protocol headers
  (20 TCP + 20 IP + 8 RFC1577 + 8 AAL5 = 56 bytes)
  × 31/32 for ABR RM cell overhead
  × 0.9 TCP window startup period
- Fair
- ABR Rate limited

Finite Buffers & Fixed Capacity

- Buffer size = 2048, TBE =512
- CLR = 0.18%
- TCP throughput = 34.16 + 31.70 = 65.86 Mbps
  = 64% of Max
- 0.18% of CLR but 36% throughput loss
- Window limited
- Time lost in retransmissions
- With TCP, you don't loose cells but you loose time.

Simulation Results: Summary

# srcs	TBE	Buffer Size	T1	T2	T3	T4	T5	Throughput	% of Max	CLR.
2	128	256	3.1	3.1				6.2	10.6	1.2
2	128	1024	10.5	4.1				14.6	24.9	2.0
2	512	1024	5.7	5.9				11.6	19.8	2.7
2	512	2048	8.0	8.0				16.0	27.4	1.0
5	128	640	1.5	1.4	3.0	1.6	1.6	9.1	15.6	4.8
5	128	1280	2.7	2.4	2.6	2.5	2.6	12.8	21.8	1.0
5	512	2560	4.0	4.0	4.0	3.9	4.1	19.9	34.1	0.3
5	512	5720	11.7	11.8	11.6	11.8	11.6	58.4	100.0	0.0

- CLR has high variance
- CLR does not reflect performance. Higher CLR does not necessarily mean lower throughput
- CLR and throughput are one order of magnitude apart
- Bursty losses are less damaging than scattered losses

Observations I

- TCP's slow-start does reduce network load
  Most of the queues are at the source
  Not much queue in the switch
- CLR in the switch is low
  But, throughput is also low
  - TCP does not use all the available bandwidth
  - Many packets are dropped at the destination
  - Much time is lost due to timer granularity
- Lower CLR does not mean higher throughput

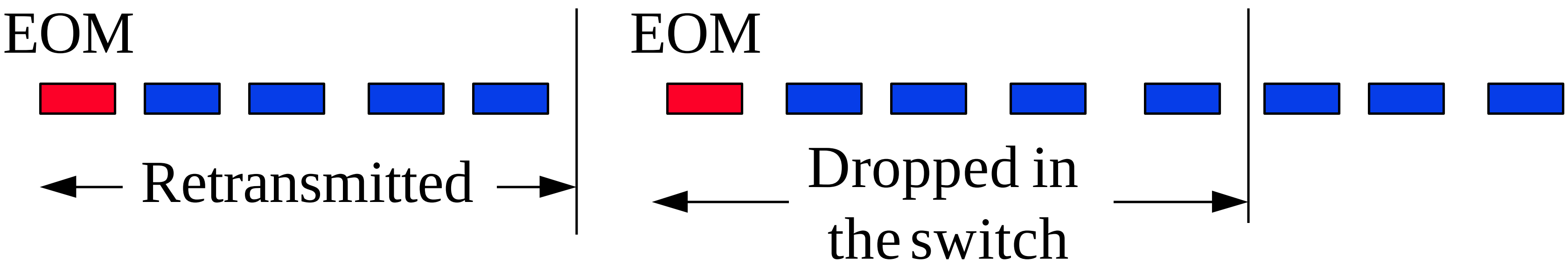
Observations II

- Larger buffer size ⇒ Higher throughput
- Effect of buffers on CLR is mixed.
  Larger buffer ⇒ CLR may be lower
  or may be higher (if loss occurs at a higher window)
- TBE's effect on throughput is mixed
  Lower TBE ⇒ Rule 6 ⇒ Less CLR ⇒ Higher throughput
  Lower TBE ⇒ Rule 6 ⇒ Rate limited ⇒ Lower throughput
- Only very low values of TBE's produce different result.
- In general, TBE of 512 or higher has no effect in this configuration

Observations III

- As the number of sources is increased, generally the total throughput increases
- TCP sources are generally window limited.
  Five sources with small windows pump more data than two sources with small windows
- Interaction among: TBE, buffer size, and number of sources

Tail Drop

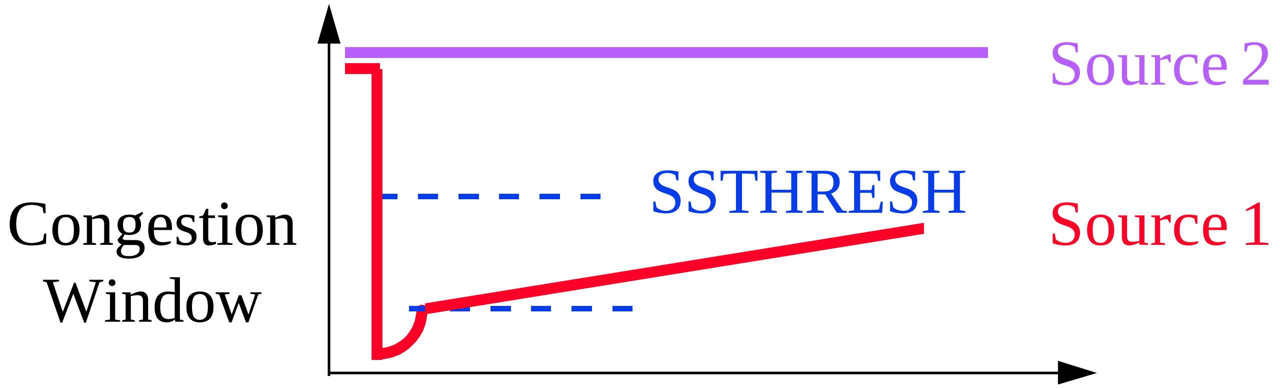


- AAL5 marks the last cell by End-of-Message (EOM)
- If EOM is dropped
  Retransmitted packet gets merged with previous partial packet.
  Fails CRC and is dropped at the destination by AAL5

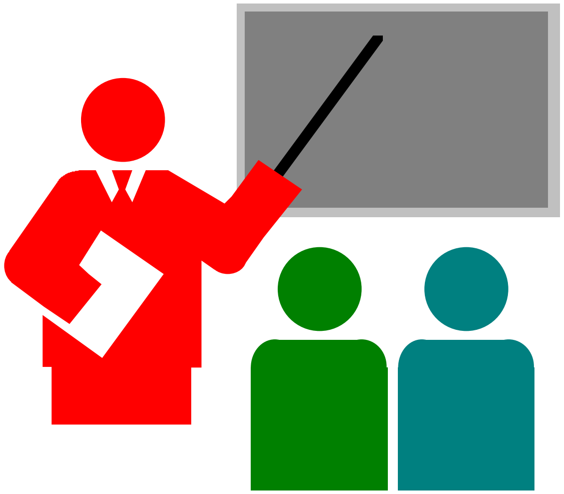
⇒ Two retransmissions in a row

Tail Drop (Cont)

- Two retransmissions in a row
  - On 1st Retransmission: SSTHRESH = W/2; W = 1
  - On 2nd Retransmission: SSTHRESH = 2, W = 1
    ⇒ Window is increased linearly
    ⇒ Very low throughput
    ⇒ Unfairness
- Intelligent Tail Drop: Do not drop EOM ⇒ Improved fairness



Summary



- TCP's slow-start + ABR's Load Control = Overcontrol
- With TCP, you may not lose cells but you lose time
  ⇒ Lower CLR but also lower throughput
- Time lost depends upon timer granularity.
- Buffers help. TBE and number of sources interact.
- Indiscriminate cell drop may cause unnecessary retransmissions and unfairness ⇒ Try not to drop EOM cells

Fast Retransmit and Recovery

- Idea: Don't wait for time-outs. Duplicate Acks indicate loss.
- Upon 3 duplicate acks, assume loss:
  - Set SSTHRESH = max{2, min{CWND/2, Wrcvr}}
  - Retransmit one packet
  - Set CWND = SSTHRESH + 3
  - For every duplicate ack: CWND = CWND + 1
  - At new ack: CWND = SSTHRESH
    This results in a sudden burst
- Reset duplicate ack count on piggybacked acks
  Intermingled duplicate and piggybacked acks $\Rightarrow$ No action

Effect of Fast Retransmit

- Fast retransmit helps only if occasional losses
  Mild congestion or errors
- With n packet loss, SSTHRESH is reduced to half after each retransmission. Window enters the linear-increase zone even when the window is small $\Rightarrow$ Low throughput.
- Even with fast retransmits, there are time-outs when the losses are bursty. These time-outs are more damaging than if there is no fast retransmit since SSTHRESH is low.

	Bursty Loss	Scattered Loss
With Fast-Retransmit Fast-Recovery	×	√
Without Fast-Retransmit Fast-Recovery	√	×