Design Issues in Traffic Management for the ATM UBR+ Service for TCP over Satellite Networks: Report II

Raj Jain

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

Overview

- Statement of Work: TCP over UBR Issues to Study
- Task 2: Drop Policies
- Task 6: TCP Implementation Issues
- Task 7: SACK Optimization
- Task 4a: GFR

Why UBR?

- Cheapest service category for the user
- Basic UBR is very cheap to implement
- Simple enhancements can vastly improve performance
- Expected to carry the bulk of the best effort TCP/IP traffic.

Goals: Issues

1. Analyze Standard Switch and End-system Policies
2. Design Switch Drop Policies
3. Quantify Buffer Requirements in Switches
4. UBR with VBR Background
5. Performance of Bursty Sources
6. Changes to TCP Congestion Control
7. Optimizing the Performance of SACK TCP

Non-Goals

- Does not cover non-UBR issues.
- Does not cover ABR issues.
- Does not include non-TM issues.

Status

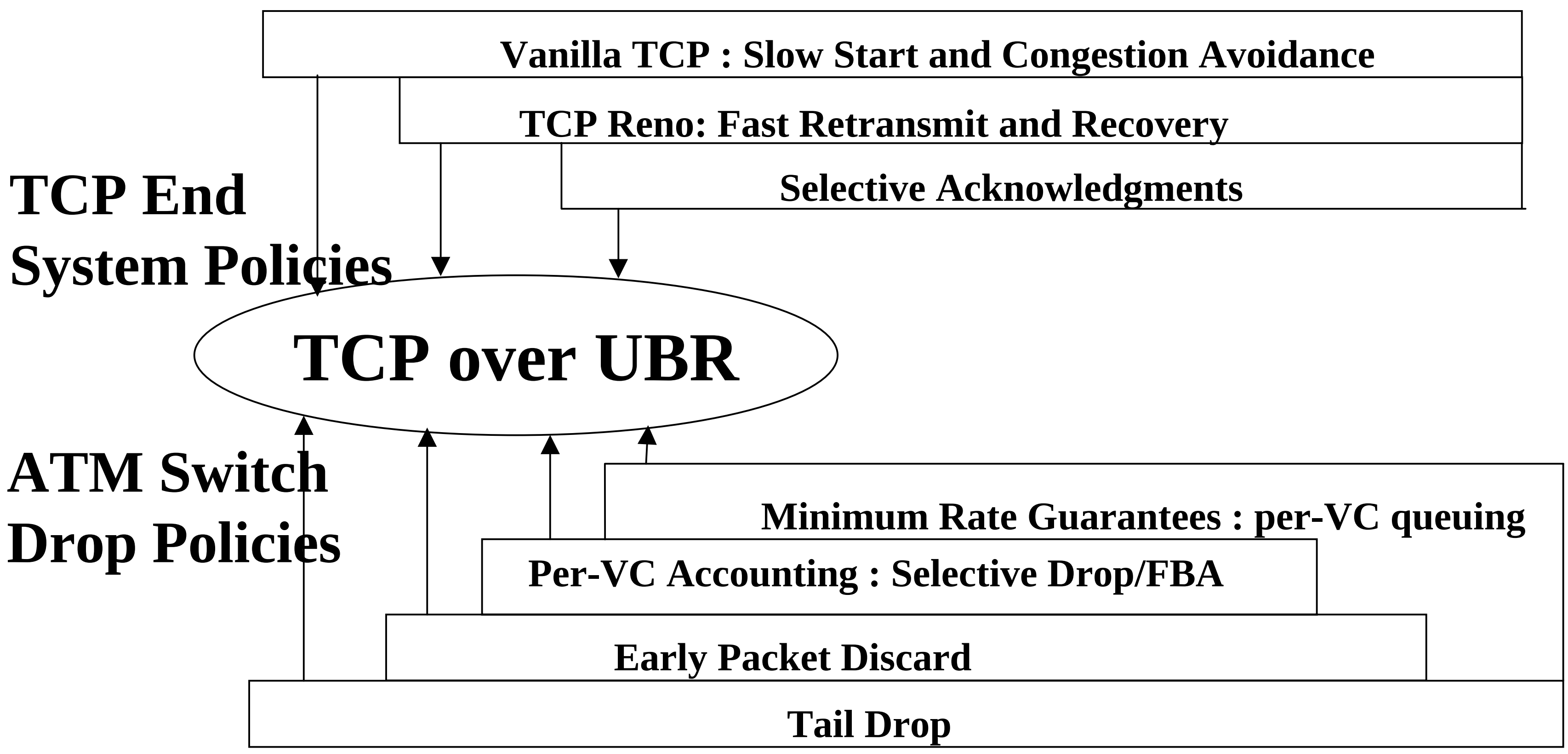
1. Analyze Standard Switch and End-system Policies[1]
2. Design Switch Drop Policies[2]
3. Quantify Buffer Requirements in Switches[1]
4. UBR with VBR Background
    4a. Guaranteed Frame Rate[2]
    4b. Guaranteed Rate[1]
5. Performance of Bursty Sources
6. Changes to TCP Congestion Control[2]
7. Optimizing the Performance of SACK TCP[2]

Status: [1]=Presented at the last meeting, [2]=Presenting now

1. Policies

Switch Policies		End-System Policies: No FRR	FRR	New Reno	SACK + New Reno
No EPD					
	Plain EPD				
EPD	Selective Drop				
	Fair Buffer Allocation				

Policies


Vanilla TCP : Slow Start and Congestion Avoidance
TCP Reno: Fast Retransmit and Recovery
Selective Acknowledgments
TCP End
System Policies
TCP over UBR
ATM Switch
Drop Policies
Minimum Rate Guarantees : per-VC queuing
Per-VC Accounting : Selective Drop/FBA
Early Packet Discard
Tail Drop

1. Policies: Results

- In LANs, switch improvements (PPD, EPD, SD, FBA) have more impact than end-system improvements (Slow start, FRR, New Reno, SACK). Different variations of increase/decrease have little impact due to small window sizes.
- In satellite networks, end-system improvements have more impact than switch-based improvements
- FRR hurts in satellite networks.
- Fairness depends upon the switch drop policies and not on end-system policies

Policies (Continued)

- In Satellite networks:
  - SACK helps significantly
  - Switch-based improvements have relatively less impact than end-system improvements
  - Fairness is not affected by SACK
- In LANs:
  - Previously retransmitted holes may have to be retransmitted on a timeout
    $\Rightarrow$ SACK can hurt under extreme congestion.

4b. Guaranteed Rate: Results

- Guaranteed rate is helpful in WANs.
- For WANs, the effect of reserving 10% bandwidth for UBR is more than that obtained by EPD, SD, or FBA
- For LANs, guaranteed rate is not so helpful. Drop policies are more important.
- For Satellites, end-system policies seem more important.

Past Results: Summary

- For satellite networks, end-system policies (SACK) have more impact than switch policies (EPD).
- Fast retransmit and recovery (FRR) improves performance over LANs but degrades performance over WANs and satellites.
- 0.5*RTT buffers provide sufficiently high efficiency (98% or higher) for SACK TCP over UBR even for a large number of TCP sources
- Reserving a small fraction for UBR helps it a lot in satellite networks

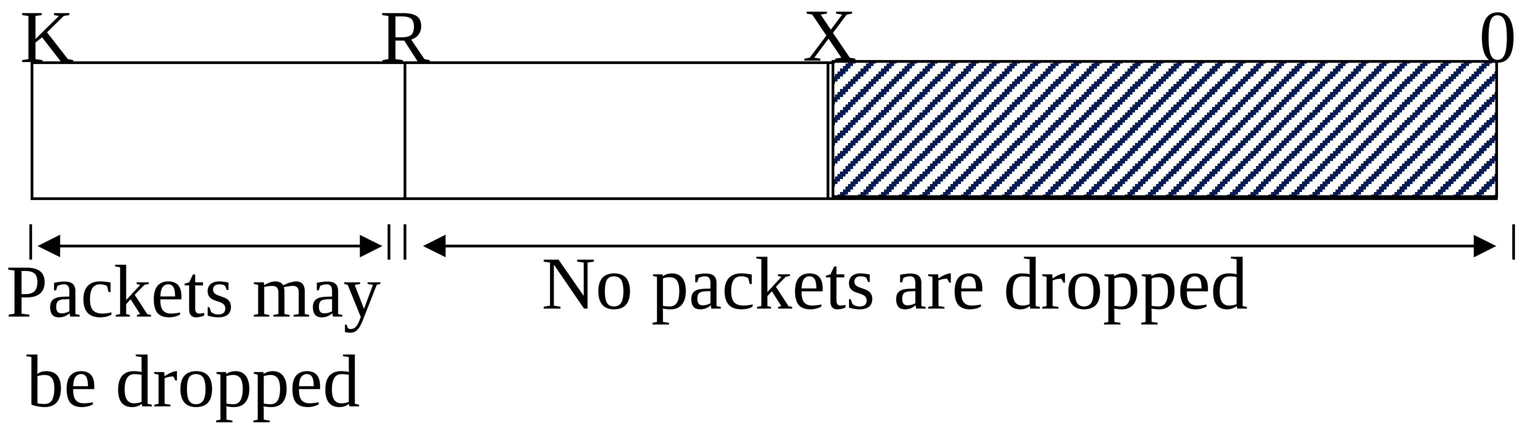
TCP over UBR: Past Results

- For zero TCP loss, buffers needed = Σ TCP windows.
- Poor performance with limited buffers.
- EPD improves efficiency but not fairness.
- In high delay-bandwidth paths, too many packets lost $\Rightarrow$ EPD has little effect in satellite networks.

2. Switch Drop Policies

- Selective Drop
- Fair buffer allocation

UBR: Selective Drop



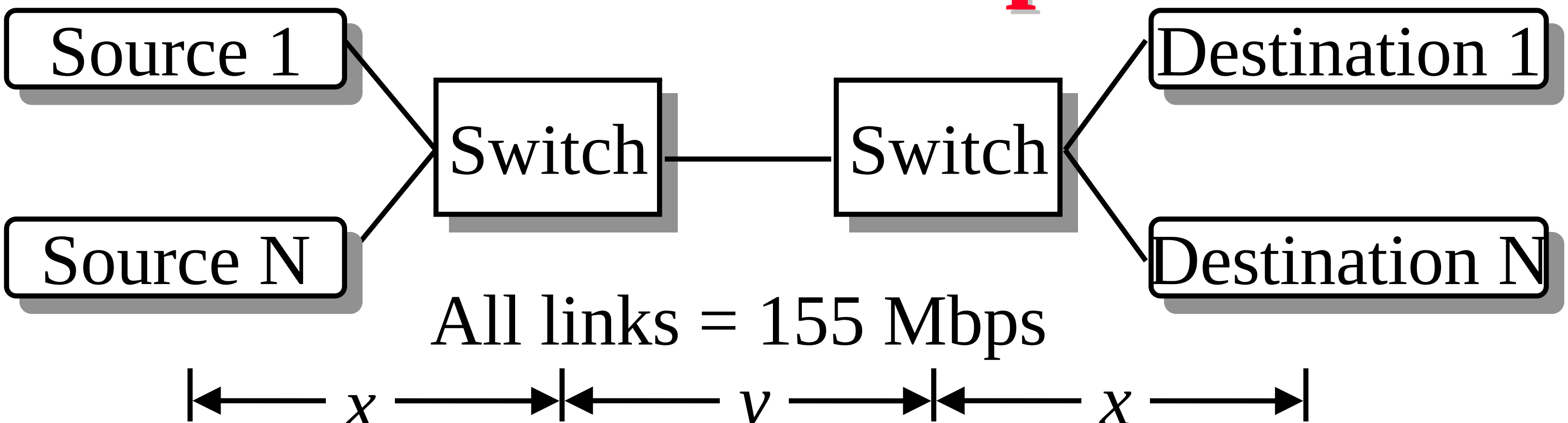
- K = Buffer size (cells).
- R = Drop threshold.
- X = Buffer occupancy.
- EPD: When (X > R) new incoming packets are dropped. Partially received packets are accepted if possible.

Selective Drop (Cont)

- N_a = Number of active VCs in the buffer
- Fair Allocation = X / N_a
- Per-VC accounting: X_i = # of cells in buffer
- Buffer load ratio of $VC_i = X_i /(X/ N_a)$
- Drop complete packet of VC_i if:

  Selective Drop: $(X > R)$ AND $(X_i/(X/N_a) > Z)$

The Simulation Experiment



- Buffer size (cells): LAN: 1k,3k. WAN: 12k,36k.
  Satellite: 200k, 600k
- RTT: LAN: 30 μs, WAN: 30 ms, satellite (y = satellite hop): 570 ms
- Efficiency: Σ throughputs / max possible throughput
- Fairness: $(\Sigma x_i)^2 / (n \Sigma x_i^2)$, x_i = throughput of *i*th TCP
- MSS (bytes): 512 (LAN,WAN), 9180 (satellites)

TCP Parameters

- TCP maximum window size, LAN: 64 Kb.
  WAN: 600,000. Satellite: 8.7 million bytes.
- MSS = 512 Bytes (LANs and WANs),
  9180 (Satellites)
- No TCP delay ack timer
- All processing delay, delay variation = 0
- TCP sources are unidirectional
- TCP timer granularity = 100 ms

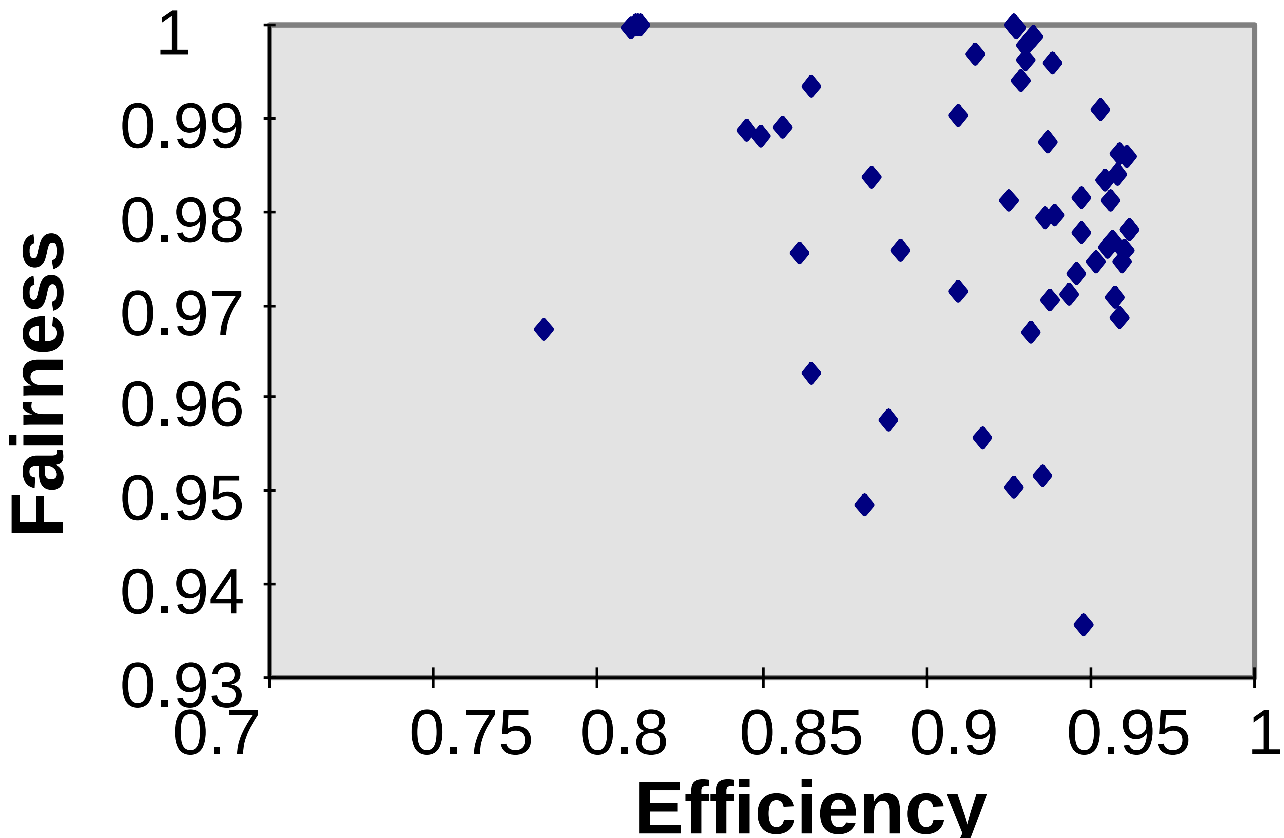
Efficiency

Configuration	TCP	UBR	EPD	Selective Drop
LAN	SACK	0.79	0.89	0.95
	Vanilla	0.34	0.67	0.84
	Reno	0.69	0.97	0.97
WAN	SACK	0.94	0.91	0.95
	Vanilla	0.91	0.90	0.91
	Reno	0.78	0.86	0.81
Satellite	SACK	0.93	0.80	0.86
	Vanilla	0.79	0.77	0.78
	Reno	0.57	0.16	0.17

Fairness

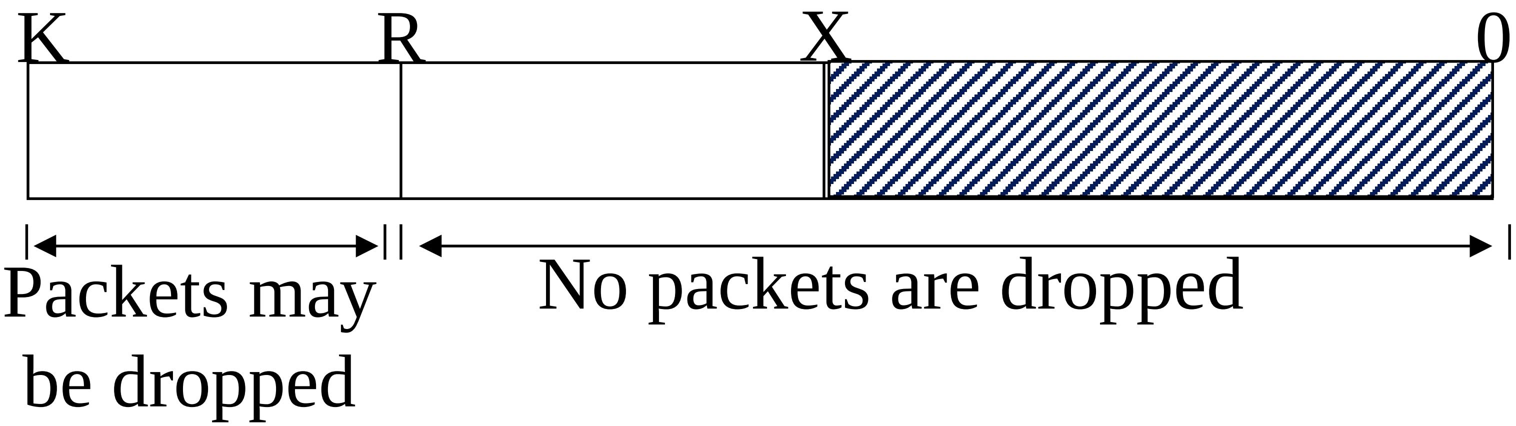
Configuration	TCP	UBR	EPD	Selective Drop
LAN	SACK	0.54	0.84	0.97
	Vanilla	0.69	0.69	0.92
	Reno	0.71	0.98	0.99
WAN	SACK	0.98	0.97	0.97
	Vanilla	0.76	0.95	0.94
	Reno	0.90	0.97	0.99
Satellite	SACK	1.00	0.92	0.97
	Vanilla	1.00	0.94	0.95
	Reno	0.98	0.99	0.99

SD: Effect of Parameters

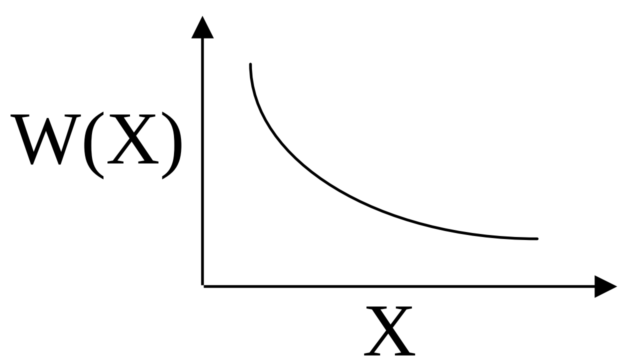


- Tradeoff between efficiency and fairness
- The scheme is sensitive to parameters
- Best value for Z = 0.8, R = 0.9*K

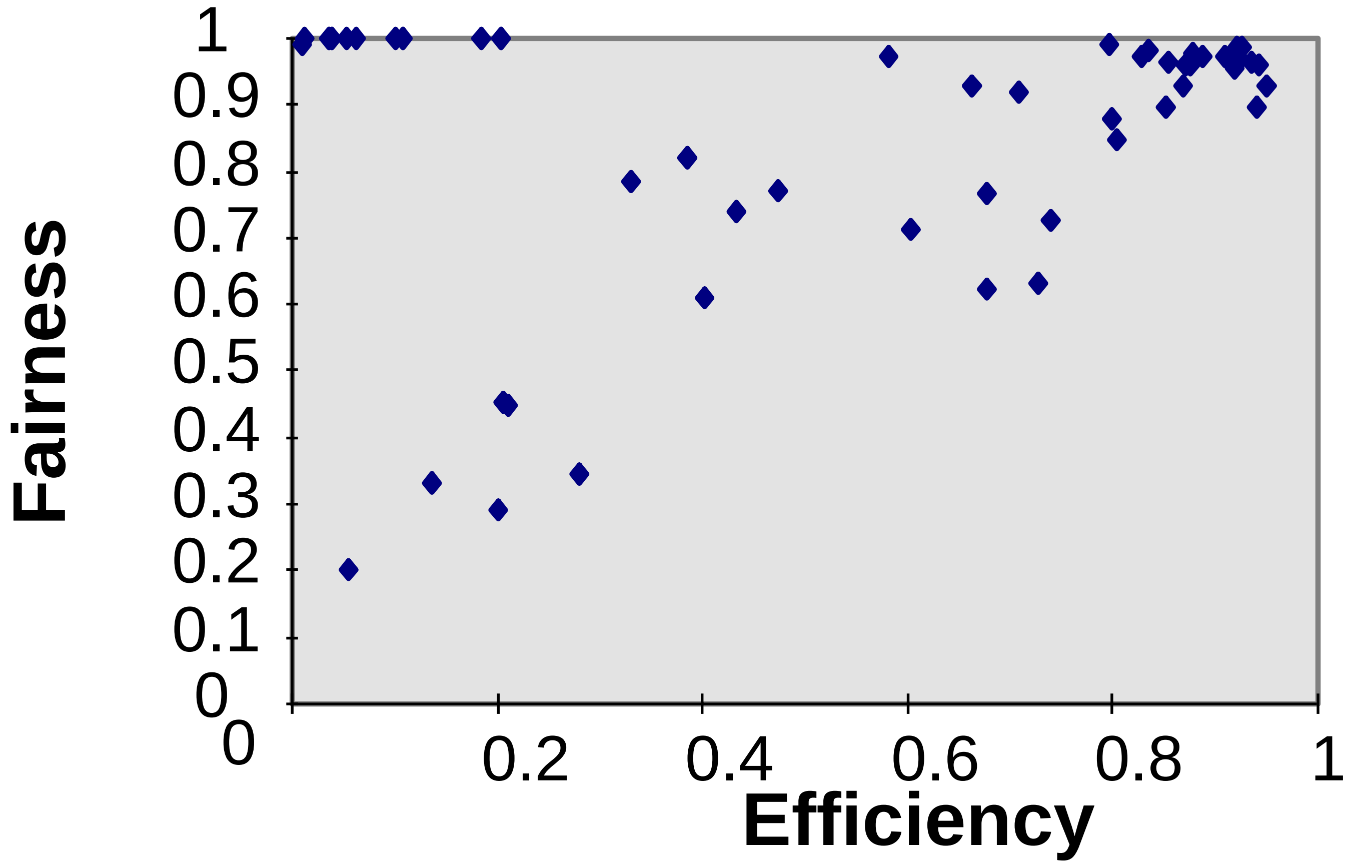
Fair Buffer Allocation (FBA)



- Drop complete packet of VC_i if
  $(X > R)$ AND $(X_i * N_a / X > W(X))$
  $W(X) = Z*((K - R)/ (X - R))$



FBA: Effect of Parameters



- Tradeoff between efficiency and fairness
- The scheme is sensitive to parameters
- Best value of Z = 0.8, R = 0.5*K

UBR + EPD + FBA

			UBR		EPD		SD		FBA	
Conf.	Srcs	Buffers	Eff.	Fairn.	Eff.	Fairn.	Eff.	Fairn.	Eff.	Fairn.
LAN	5	1000	0.21	0.68	0.49	0.57	0.75	0.99	0.88	0.98
LAN	5	2000	0.32	0.90	0.68	0.98	0.85	0.96	0.84	0.98
LAN	5	3000	0.47	0.97	0.72	0.84	0.90	0.99	0.92	0.97
LAN	15	1000	0.22	0.31	0.55	0.56	0.76	0.76	0.91	0.97
LAN	15	2000	0.49	0.59	0.81	0.87	0.82	0.98	0.85	0.96
LAN	15	3000	0.47	0.80	0.91	0.78	0.94	0.94	0.95	0.93
WAN	5	12000	0.86	0.75	0.90	0.94	0.90	0.95	0.95	0.94
WAN	5	24000	0.90	0.83	0.91	0.99	0.92	0.99	0.92	1
WAN	5	36000	0.91	0.86	0.81	1	0.81	1	0.81	1
WAN	15	12000	0.96	0.67	0.92	0.93	0.94	0.91	0.95	0.97
WAN	15	24000	0.94	0.82	0.91	0.92	0.94	0.97	0.96	0.98
WAN	15	36000	0.92	0.77	0.96	0.91	0.96	0.89	0.95	0.97

- FBA improves both efficiency and fairness
- Effect of FBA is similar to that of SD. SD is simpler.

Drop Policies: Results

- Low efficiency and fairness for TCP over UBR
- Need switch buffers = Σ(TCP maximum window sizes) for zero TCP loss
- EPD improves efficiency but not fairness
- Selective drop improves fairness
- Fair Buffer Allocation improves both efficiency and fairness, but is sensitive to parameters
- TCP synchronization affects performance

6. Problem in TCP Implementations

- Linear Increase in Segments:
  CWND/MSS = CWND/MSS + MSS/CWND
- In Bytes: CWND = CWND + MSS*MSS/CWND
- All computations are done in integer
- If CWND is large, MSS*MSS/CWND is zero and CWND does not change. CWND stays at 512*512 or 256 kB.

Solutions

- **Solution 1**: Increment CWND after N acks (N > 1) CWND = CWND + N*MSS*MSS/CWND
- **Solution 2**: Use larger MSS on Satellite links such that MSS*MSS > CWND. MSS $\geq$ Path MTU.
- **Solution 3**: Use floating point
- **Recommendation**: Use solution 1. It works for all MSSs.
- **To do**: Does this change TCP dynamics and adversely affect performance.
- Result: Solution 1 works. TCP dynamics is not affected.

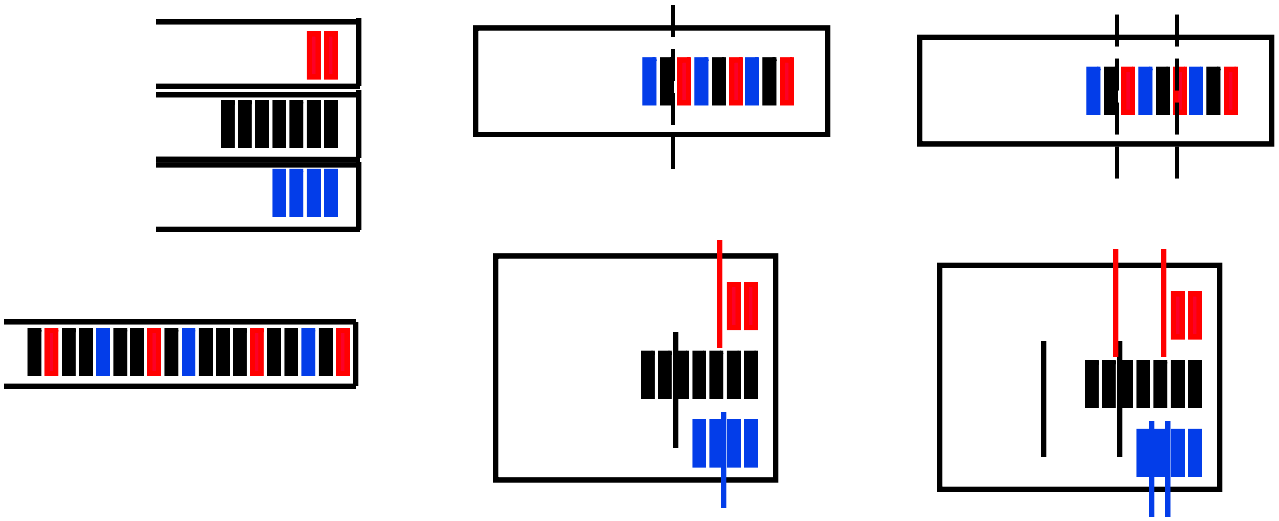
7. Optimize SACK TCP

- SACK helps only if retransmitted packets are not lost.
- Currently TCP retransmits immediately after 3 duplicate acks (Fast retransmit), and then waits RTT/2 for congestion to subside.
- Network may still be congested
  $\Rightarrow$ Retransmitted packets lost.
- Proposed Solution: Delay retransmit by RTT/2, I.e., wait RTT/2 first, and then retransmit.
- New Result: Delayed retransmit does not help.

4a. Guaranteed Frame Rate (GFR)

- UBR with minimum cell rate (MCR) ⇒ UBR+
- Frame based service
  - Complete frames are accepted or discarded in the switch
  - Traffic shaping is frame based.
    All cells of the frame have CLP =0 or CLP =1
  - All frames below MCR are given CLP =0 service.
    All frames above MCR are given best effort (CLP =1) service.
- Allocation of excess (over MCR) is arbitrary

4a. GFR Options

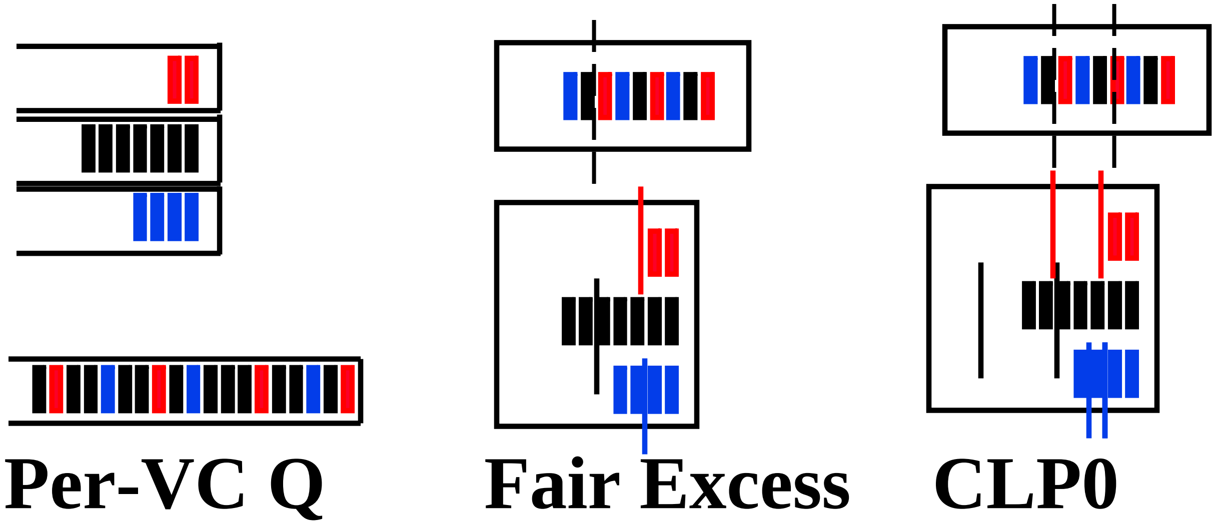


Queuing	Per-VC	FIFO
Buffer Management	Per-VC Thresholds	Global Threshold
Tag-sensitive Buffer Mgmt	2 Thresholds	1 Threshold

Options (Cont)

- FIFO queuing versus per-VC queuing
  - Per-VC queuing is too expensive.
  - FIFO queuing should work by setting thresholds based on bandwidth allocations.
- Buffer management policies
  - Per-VC accounting policies need to be studied
- Network tagging and end-system tagging
  - End system tagging can prioritize certain cells or cell streams.
  - Network tagging used for policing -- must be requested by the end system.

GFR: Results



- Per-VC queuing and scheduling is sufficient for per-VC MCR.
- FBA and proper scheduling is sufficient for fair allocation of excess bandwidth
- One global threshold is sufficient for CLP0+1 guarantees
  Two thresholds are necessary for CLP0 guarantees

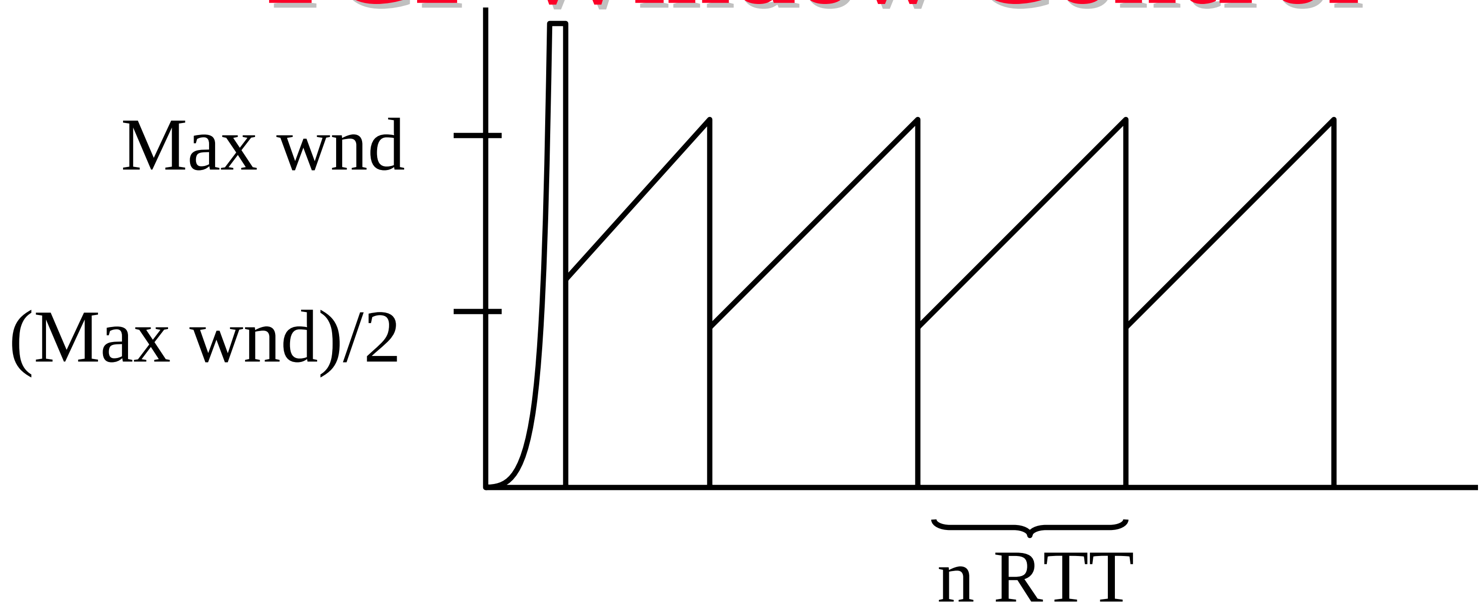
Issues

- All FIFO queuing cases were studied with high target network load, i.e., most of the network bandwidth was allocated as GFR.
- Need to study cases with lower percentage of network capacity allocated to GFR VCs.

Further Study: Goals

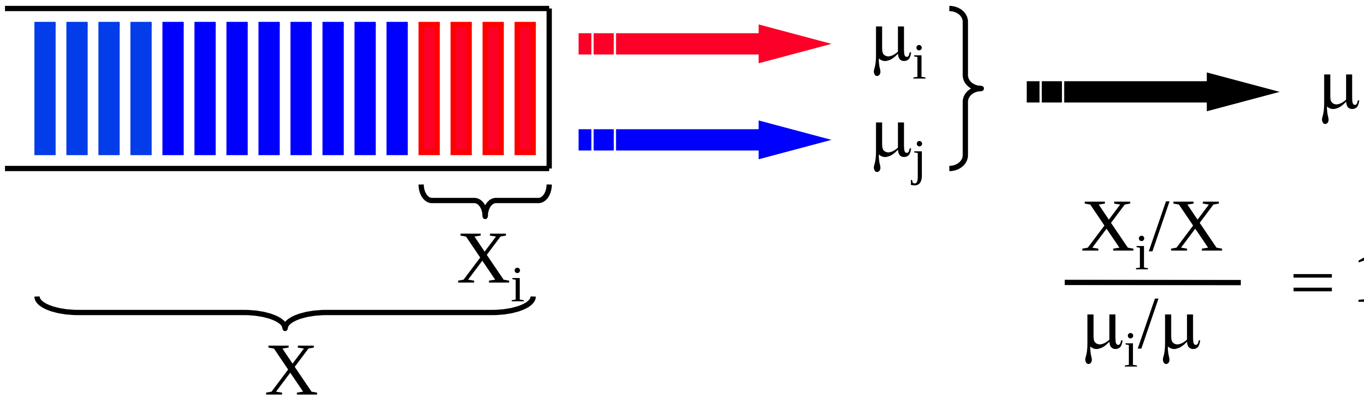
- Provide minimum rate guarantees with FIFO buffer for TCP/IP traffic.
- Guarantees in the form of TCP throughput and not cell rate (MCR).
- How much network capacity can be allocated before guarantees can no longer be met?
- Study rate allocations for VCs with aggregate TCP flows.

TCP Window Control



- For TCP window based flow control (in linear phase)
  - Throughput = (Avg wnd) / (Round trip time)
- With Selective Ack (SACK), window decreases by 1/2 during packet loss, and then increases linearly.
  - Avg wnd = $[\Sigma_{i=1,\ldots,n} (\text{max wnd}/2 + \text{mss}*i)] /n$

FIFO Buffer Management



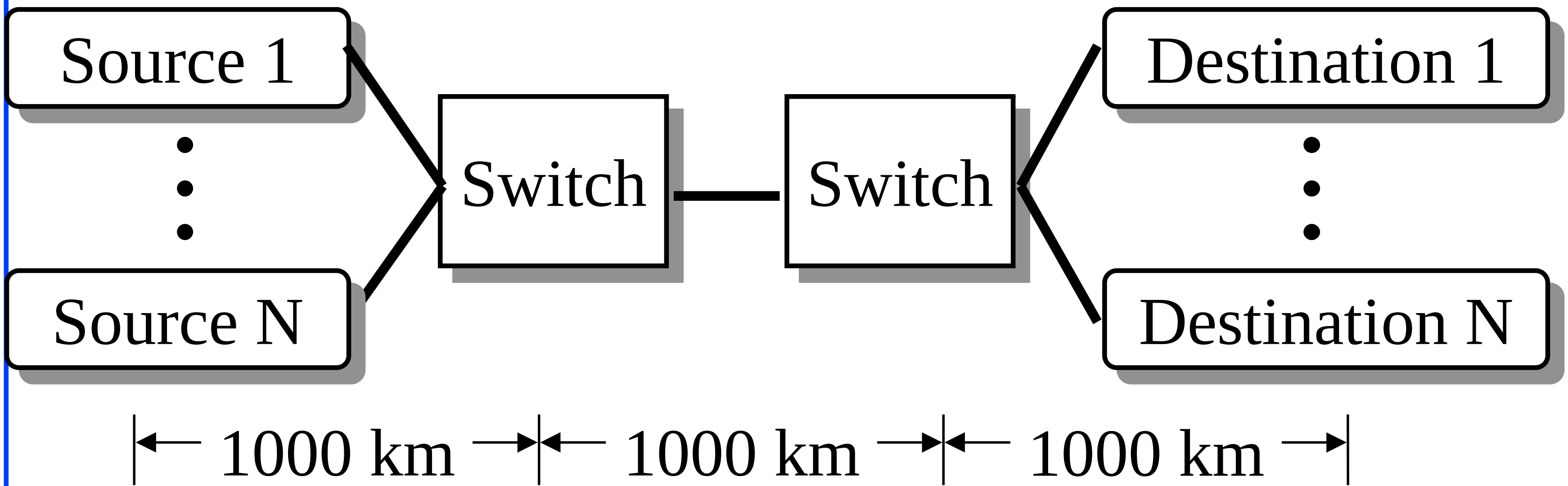
$$\frac{X_i/X}{\mu_i/\mu} = 1$$

- Fraction of buffer occupancy (X_i/X) determines the fraction of output rate (μ_i/μ) for VCi.
- Maintaining average per-VC buffer occupancy enables control of per-VC output rates.
- Set a threshold (R_i) for each VC.
- When X_i exceeds R_i, then control the VC's buffer occupancy.

Buffer Management for TCP

- TCP responds to packet loss by reducing CWND by one-half.
  - When *i*th flow's buffer occupancy exceeds R_i, drop a single packet.
  - Allow buffer occupancy to decrease below R_i, and then repeat above step if necessary.
- K = Total buffer capacity.
- Target utilization = $\Sigma R_i / K$.
- Guaranteed TCP throughput = Capacity $*$ R_i/K
- Expected throughput, $\mu_i = \mu * R_i / \Sigma R_i$. ($\mu = \Sigma \mu_i$)

Simulation Configuration



- SACK TCP.
- 15 TCP sources (N = 15).
- Buffer Size = K = 48000 cells.
- 5 thresholds ($R_1,\ldots,R_5$).

Configuration (contd.)

Sources	Expt 1	Expt 2	Expt 3	Expt 4	Expected Throughput
1-3 (R_1)	305	458	611	764	2.8 Mbps
4-6 (R_2)	611	917	1223	1528	5.6 Mbps
7-9 (R_3)	917	1375	1834	2293	8.4 Mbps
10-24 (R_4)	1223	1834	2446	3057	11.2 Mbps
13-15 (R_5)	1528	2293	3057	3822	14.0 Mbps
$\Sigma R_i/K$	29%	43%	57%	71%	

- Threshold $R_{ij} \propto \lfloor K{*}MCR_i/PCR \rfloor$
- Total throughput μ = 126 Mbps. MSS =1024B.
- Expected throughput = $\mu * R_i / \Sigma R_i$

Simulation Results

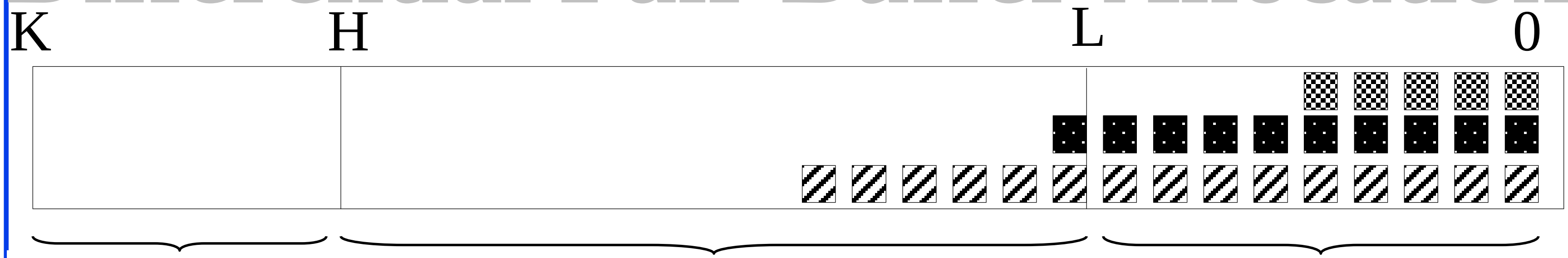
TCP Number	Throughput ratio (observed / expected)			
1-3	1.0	1.03	1.02	1.08
4-6	0.98	1.01	1.03	1.04
7-9	0.98	1.00	1.00	1.02
10-12	0.98	0.99	0.98	0.88
13-15	1.02	0.98	0.97	1.01

- All ratios close to 1.
  Variations increases with utilization.
- All sources experience similar queuing delays

TCP Window Control

- TCP throughput can be controlled by controlling window.
- FIFO buffer $\Rightarrow$ Relative throughput per connection is proportional to fraction of buffer occupancy.
- Controlling TCP buffer occupancy
  $\Rightarrow$ May control throughput.
- High buffer utilization $\Rightarrow$Harder to control throughput.
- Formula does not hold for very low buffer utilization
  Very small TCP windows
  $\Rightarrow$ SACK TCP times out if half the window is lost

Differential Fair Buffer Allocation



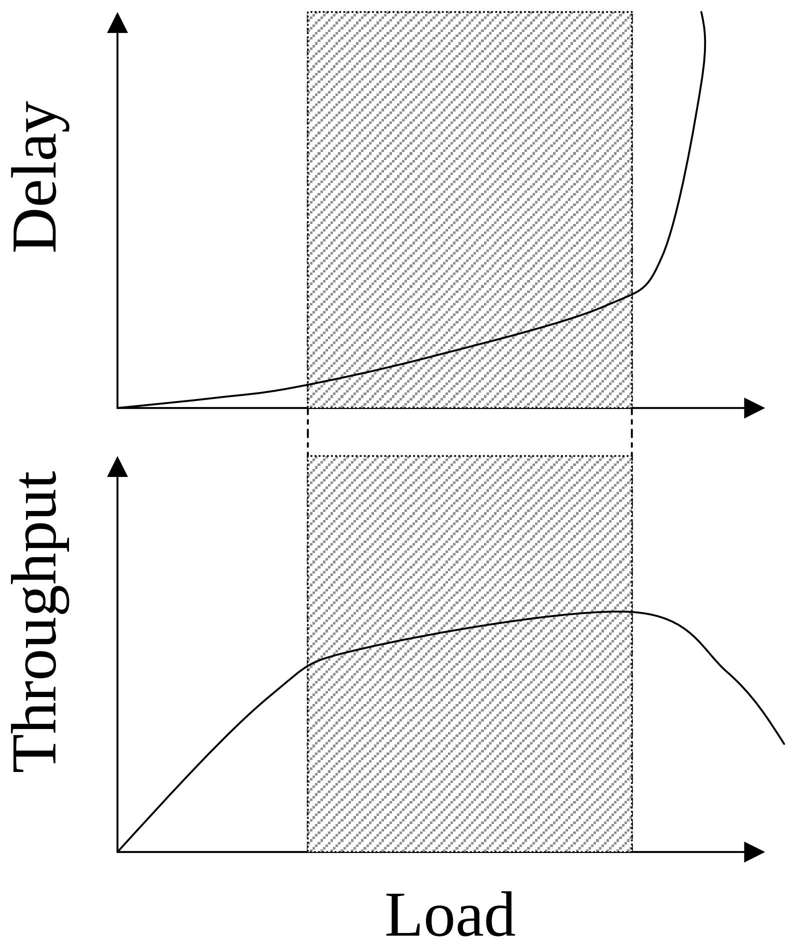
$X > H$
$\Rightarrow$ EPD

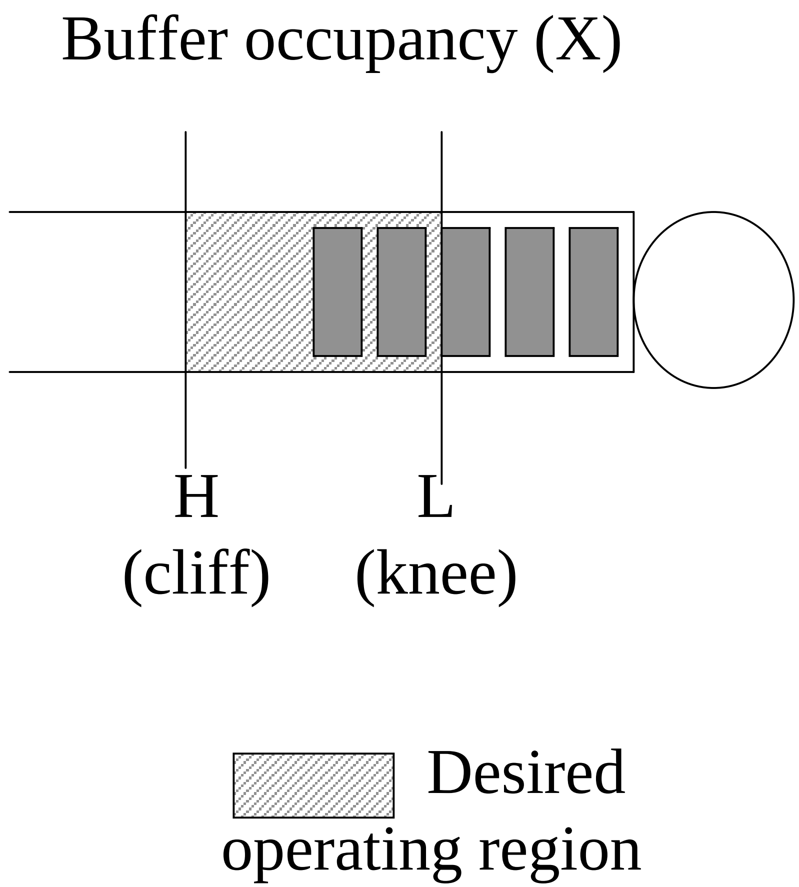
$X > L \Rightarrow$ Drop all CLP1.
$X > L$ and $X_i > X*W_i/W \Rightarrow$
Probabilistic Loss of CLP0

$X \leq L$
$\Rightarrow$ No Loss

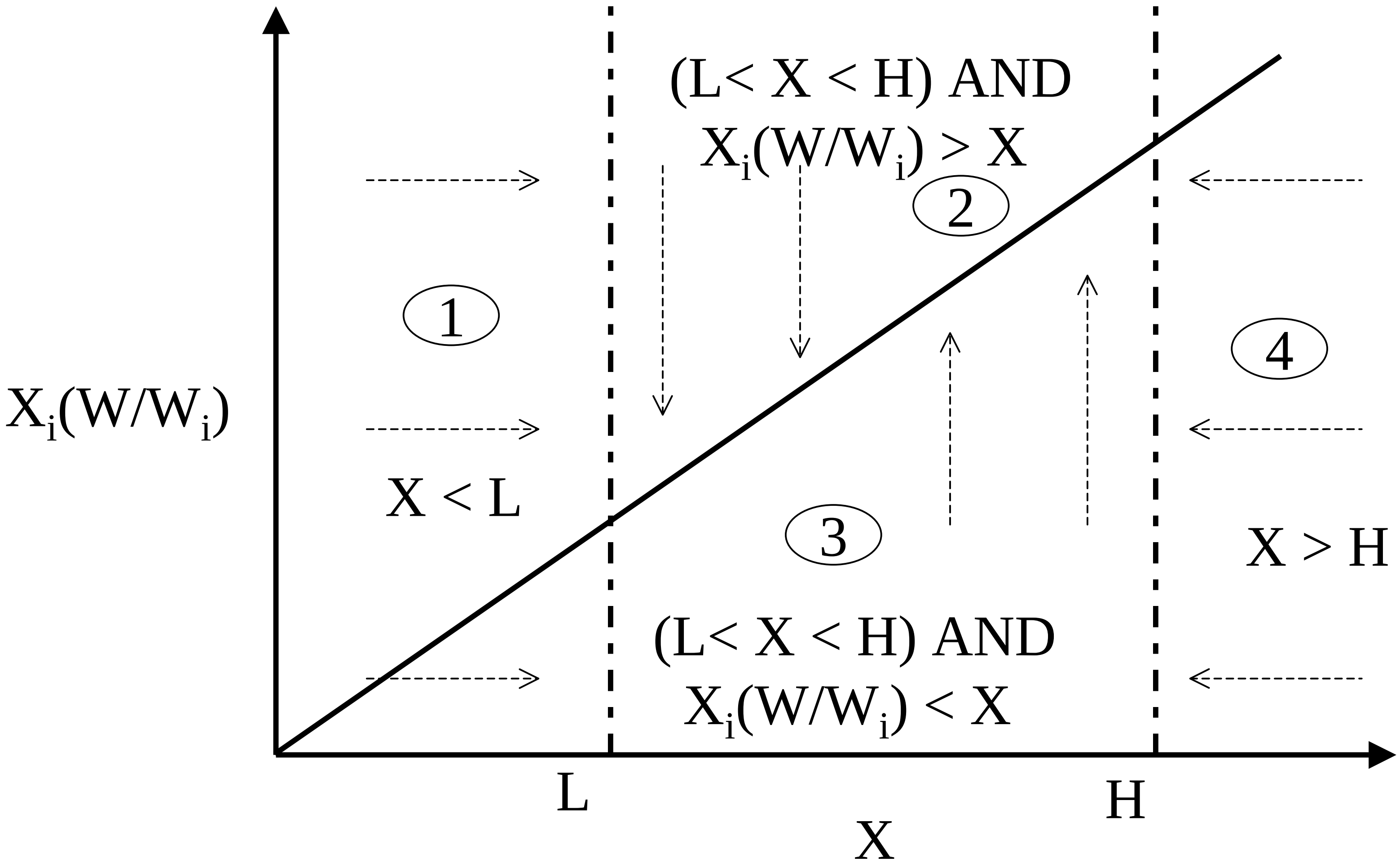
- W_i = Weight of VC_i = MCR_i/(GFR Capacity)
- $W = \Sigma W_i$
- L = Low Threshold. H = High Threshold
- X_i = Per-VC buffer occupancy. ($X = \Sigma X_i$)
- Z_i = Parameter ($0 \leq Z \leq 1$)

DFBA Operating Region





DFBA (contd.)



DFBA (contd.)

Region	Condition	Action
1	Underload	Improve efficiency
2	Mild congestion, more than fair share	Drop low priority packets, bring down to fair share
3	Mild congestion, less than fair share	Drop low priority packets, bring up to fair share
4	Severe congestion	Reduce load

DFBA Algorithm

- When first cell of frame arrives:
- IF (X < L) THEN
  - Accept frame
- ELSE IF (X > H) THEN
  - Drop frame
- ELSE IF ((L < X < H) AND ($X_i > X \times W_i/W$)) THEN
  - Drop CLP1 frame
  - Drop CLP0 frame with

$$P\{Drop\} = Z_i \left(\alpha \times \frac{X_i - X \times W_i/W}{X(1 - W_i/W)} + (1-\alpha) \times \frac{X - L}{H - L} \right)$$

Drop Probability

- Fairness Component
  (VC_i's fair share = $X \times W_i/W$)

$$\frac{X_i - X \times W_i/W}{X \times (1 - W_i/W)}$$

  Increases linearly as X_i increases from $X \times W_i/W$ to X

- Efficiency Component

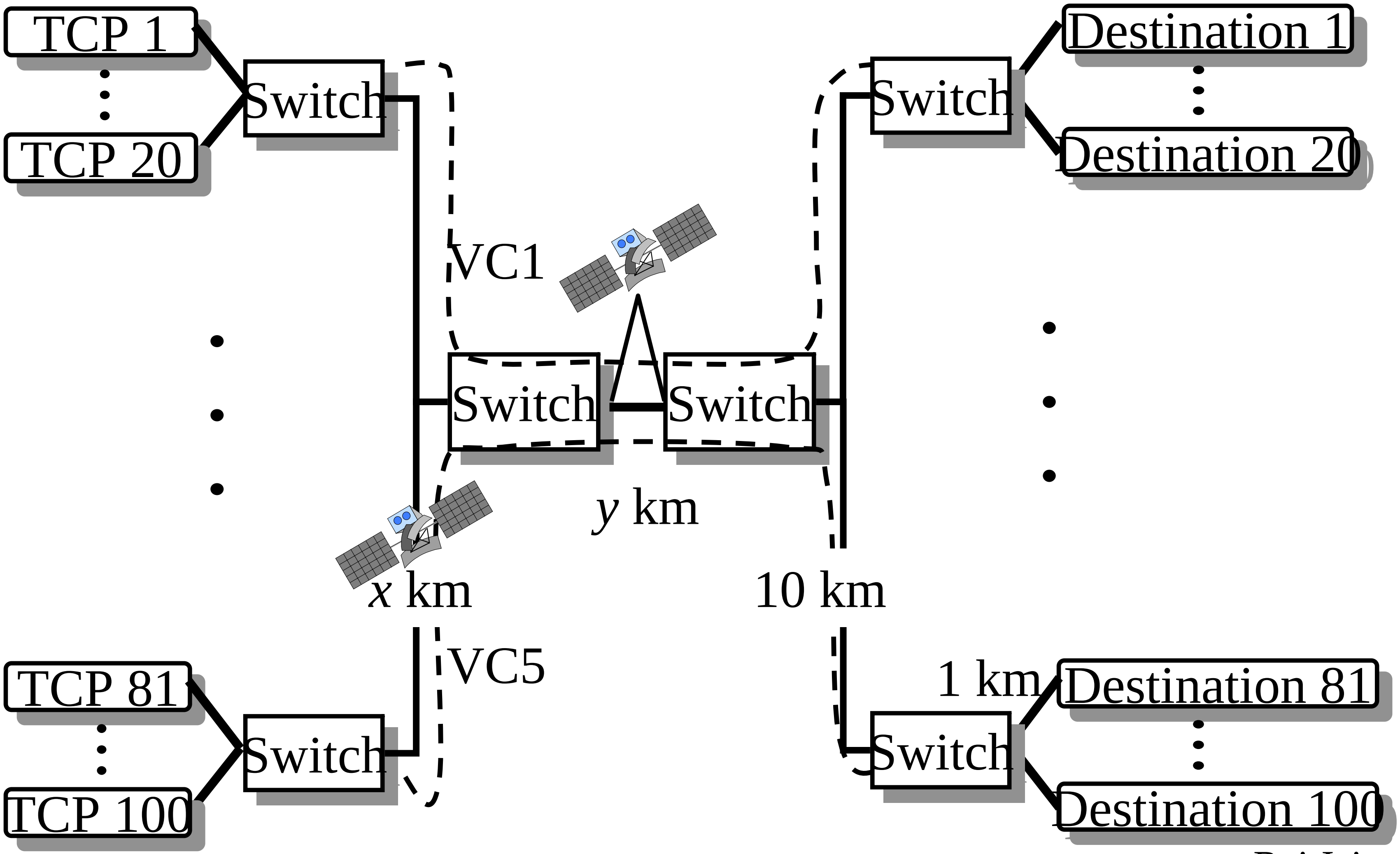
$$\frac{X - L}{H - L}$$

  Increases linearly as X increases from L to H

Drop Probability (contd.)

- Z_i allows scaling of total probability function
  - Higher drop probability results in lower TCP windows
  - TCP window size W ∝ 1/√P{Drop} for random packet loss [Mathis]
    TCP data rate $D \propto \frac{MSS}{RTT \times \sqrt{P(drop)}}$
  - To maintain high TCP data rate for large RTT:
    - Small P(Drop)
    - Large MSS
- Choose small Z_i for satellite VCs.
- Choose small Z_i for VCs with larger MCRs.

DFBA Simulation Configuration



DFBA Simulation Configuration

- SACK TCP, 50 and 100 TCP sources
- 5 VCs through backbone link.
- Local switches merge TCP sources.
- x = Access hop = 50 μs (Campus), or 250 ms GEO
- y = Backbone hop = 5 ms (WAN or LEO) or 250 ms (GEO)
- GFR capacity = 353.207 kcells/sec (≈155.52 Mbps)
- $\alpha = 0.5$

Simulation Configuration (contd)

- 50 TCPs with 5 VCs (50% MCR allocation)
  - MCR_i = 12, 24, 36, 48, 60 kcells/sec, i=1, 2, 3, 4, 5
  - W_i = 0.034, 0.068, 0.102, 0.136, 0.170
  - $\Sigma (MCR_i / \text{GFR capacity}) = \Sigma W_i = W \approx 0.5$

Simulation Configuration (contd)

- 50 and 100 TCPs with 5 VCs (85% MCR allocation)
  - MCR_i = 20, 40, 60, 80, 100 kcells/sec, i=1, 2, 3, 4, 5
  - W_i = 0.0566, 0.1132, 0.1698, 0.2264, 0.283
  - $\Sigma\ (MCR_i\ /\text{GFR capacity}) = \Sigma\ W_i = W \approx 0.85$

Simulation Results

MCR	Achieved Throughput	Excess	Excess / MCR
4.61	11.86	7.25	1.57
9.22	18.63	9.42	1.02
13.82	24.80	10.98	0.79
18.43	32.99	14.56	0.79
23.04	38.60	15.56	0.68
69.12	126.88	57.77	

- 50 TCPs with 5VCs (50% MCR allocation)
- Switch buffer size = 25 kcells
- Z_i=1, for all i
- MCR guaranteed. Lower MCRs get higher excess.

Effect of MCR Allocation

MCR	Achieved Throughput	Excess	Excess/MCR
7.68	12.52	4.84	0.63
15.36	18.29	2.93	0.19
23.04	25.57	2.53	0.11
30.72	31.78	1.06	0.03
38.40	38.72	0.32	0.01
115.2	126.88	11.68	

- 50 TCPs with 5 VCs (85% MCR allocation)
- Switch buffer size = 25 kcells
- Z_i=1, for all I
- MCR guaranteed. Lower MCRs get higher excess

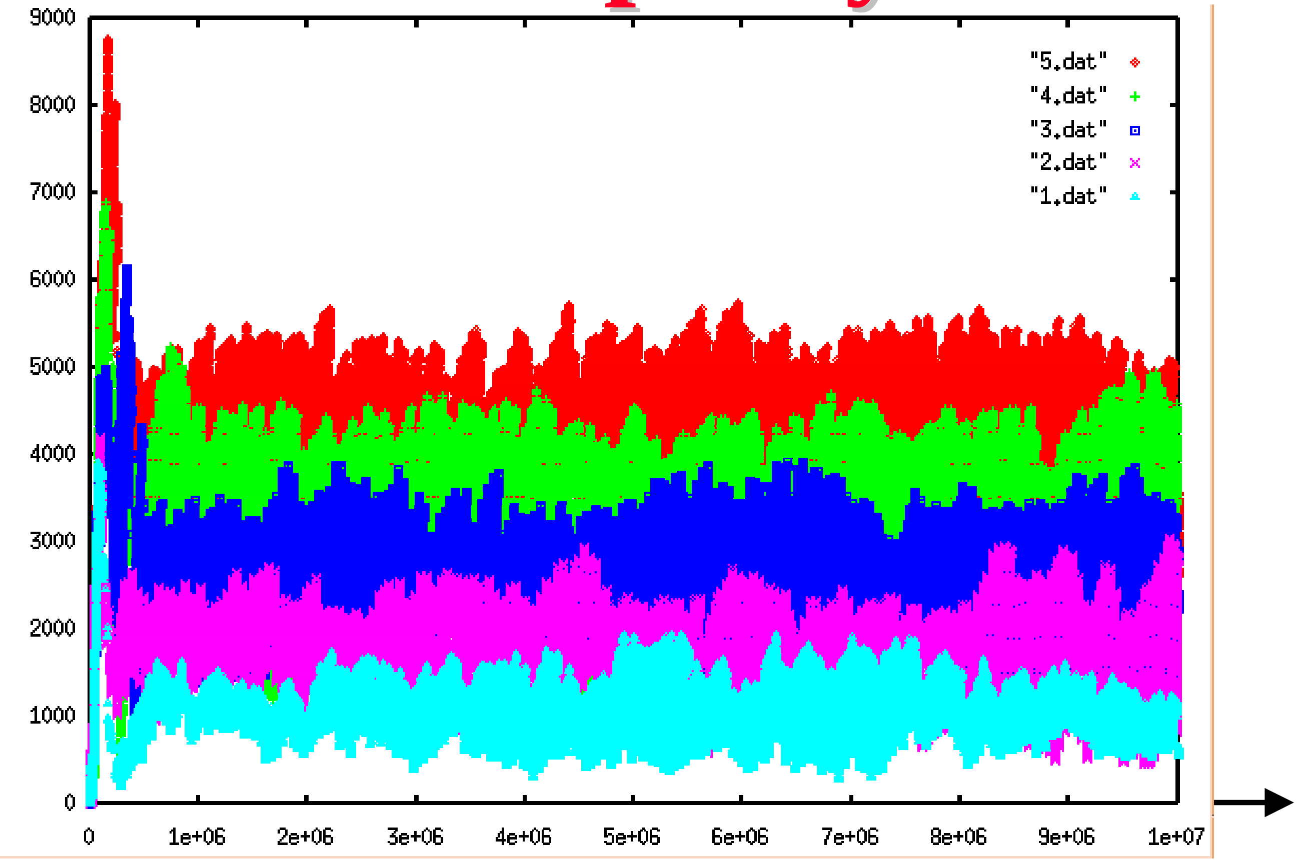
Effect of Number of TCPs

MCR	Achieved Throughput	Excess	Excess/MCR
7.68	11.29	3.61	0.47
15.36	18.19	2.83	0.18
23.04	26.00	2.96	0.13
30.72	32.35	1.63	0.05
38.40	39.09	0.69	0.02
115.2	126.92	11.72	

- 100 TCPs with 5 VCs (85 % MCR allocation)
- Switch buffer size = 25 kcells
- Z_i=1, for all i
- Results are independent of the number of sources

Buffer Occupancy

Cells



Time

- 100 TCPs with 5 VCs (85 % MCR allocation)
- Switch buffer size = 25 kcells
- Queues are approximately proportional to MCRs

Effect of Buffer Size

MCR	Achieved Throughput	Excess	Excess/MCR
7.68	11.79	4.11	0.54
15.36	18.55	3.19	0.21
23.04	25.13	2.09	0.09
30.72	32.23	1.51	0.05
38.40	38.97	0.57	0.01
115.2	126.67	11.47	

- 100 TCPs with 5 VCs (85 % MCR allocation)
- Switch buffer size = 6 kcells (Small)
- Z_i=1, for all I
- MCR guaranteed. Lower MCRs get higher excess.

Buffer Size (Cont)

MCR	Achieved Throughput	Excess	Excess/MCR
7.68	10.02	2.34	0.30
15.36	19.31	3.95	0.26
23.04	25.78	2.74	0.12
30.72	32.96	2.24	0.07
38.40	38.56	0.16	0.00
115.2	126.63	11.43	

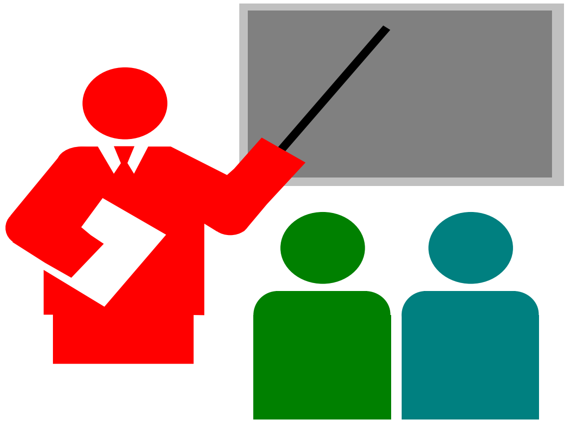
- 100 TCPs with 5 VCs (85 % MCR allocation)
- Switch buffer size = 3 kcells (Small)
- Z_i=1, for all I
- MCR guaranteed. Lower MCRs get higher excess.

Effect of Z_i

$Z_i = 1 - W_i/W$		$Z_i = (1 - W_i/W)^2$	
Excess	Excess/MCR	Excess	Excess/MCR
3.84	0.50	0.53	0.07
2.90	0.19	2.97	0.19
2.27	0.10	2.77	0.12
2.56	0.08	2.39	0.08
0.02	0.02	3.14	0.08

- 100 TCPs with 5 VCs (85 % MCR allocation)
- Switch buffer size = 6 kcells
- Small Z_i for large MCR enables MCR proportional sharing of excess capacity

Summary



- Task 2: Design switch drop policies:
  - Selective drop and Fair Buffer Allocation improve fairness and efficienciy
  - FBA is more sensitive to parameters than SD
- Task 6: Changes to TCP congestion control:
  - Increment CWND after N acks works OK

Summary (Cont)

- Task 7: Optimizing SACK TCP:
  - Delayed retransmit has no effect.
- Task 4a: Guaranteed Frame Rate:
  - SACK TCP throughput may be controlled with FIFO queuing under certain circumstances:
    - TCP, SACK (?)
    - Σ MCRs < GFR Capacity
    - Same RTT (?), Same frame size (?)
    - No other non-TCP or higher priority traffic (?)
  - New Buffer Management Policy: DFBA

References

- All our contributions and papers are available on-line at http://www.cis.ohio-state.edu/~jain/
- See Recent Hot Papers for tutorials.
- Tasks 1 and 2: Analyze and design switch and end-system policies. UBR drop policies.
  Rohit Goyal, et al, "Improving the Performance of TCP over the ATM-UBR service", To appear in Computer Communications, http://www.cis.ohio-state.edu/~jain/papers/cc.htm

References (Cont)

- Task 3: Buffer requirements for various delay-bandwidth products
  - Rohit Goyal, et al, "Analysis and Simulation of Delay and Buffer Requirements of Satellite-ATM Networks for TCP/IP Traffic," Submitted to IEEE Journal of Selected Areas in Communications, March 1998, http://www.cis.ohio-state.edu/~jain/papers/jsac98.htm

References (Cont)

- Task 4: UBR with GR and GFR
  - Rohit Goyal, et al, "Design Issues for providing Minimum Rate Guarantees to the ATM Unspecified Bit Rate Service", Proceedings of ATM'98, May 1998, http://www.cis.ohio-state.edu/~jain/papers/atm98.htm
  - Rohit Goyal, et al, "Providing Rate Guarantees to TCP over the ATM GFR Service," Submitted to LCN'98, http://www.cis.ohio-state.edu/~jain/papers/lcn98.htm

Thank You!

This research was sponsored by
NASA Lewis Research Center.