The War Between Mice and Elephants

Liang Guo and Ibrahim Matta

Boston University

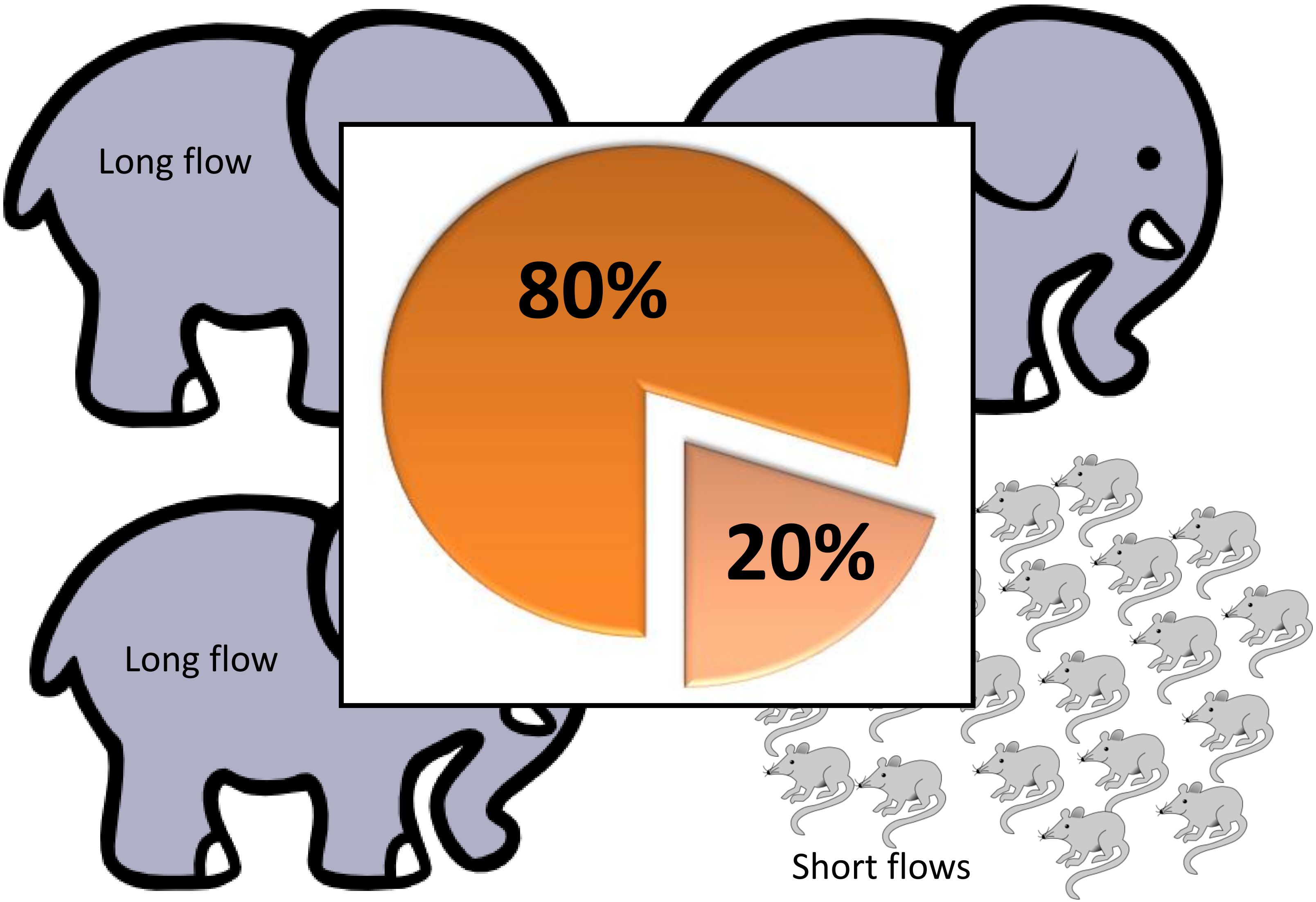
ICNP 2001

Presented by
Thangam Seenivasan

Outline

- Introduction
- Analysis and Motivation
- Architecture
- Simulation Results
- Discussion
- Conclusion

TCP flows in the Internet



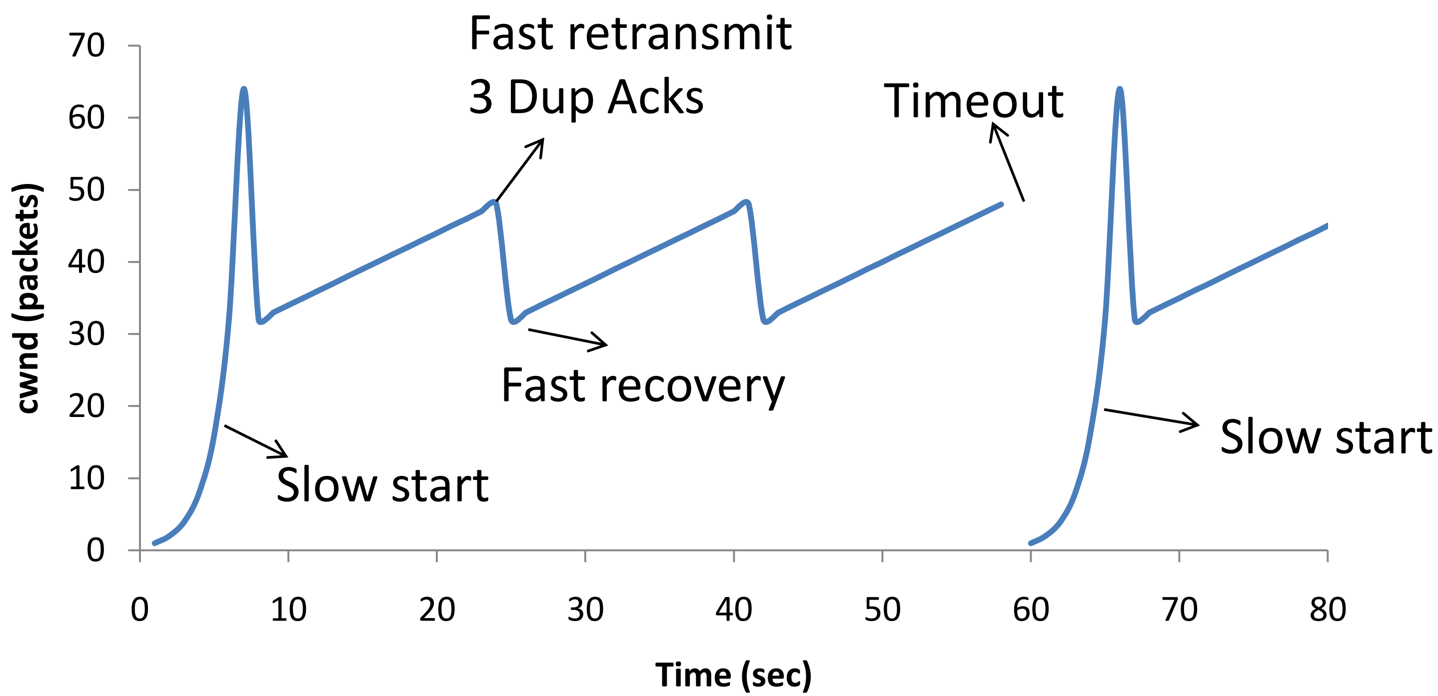
Is Internet fair?

- In a fair network
  - Short connections expect relatively fast service compared to long connections

- Sometimes this is not the case with Internet

Why?

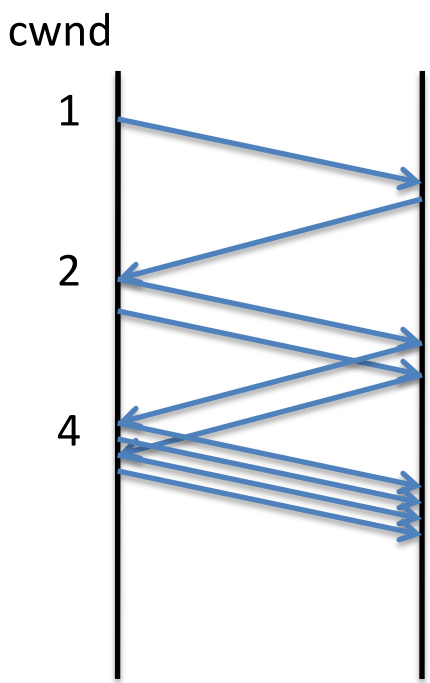
TCP



70
60
50
40
30
20
10
0
cwnd (packets)
0
10
20
30
40
50
60
70
80
Time (sec)
Fast retransmit
3 Dup Acks
Timeout
Fast recovery
Slow start
Slow start

Short TCP flows

1. Most short flows finish before slow start finish

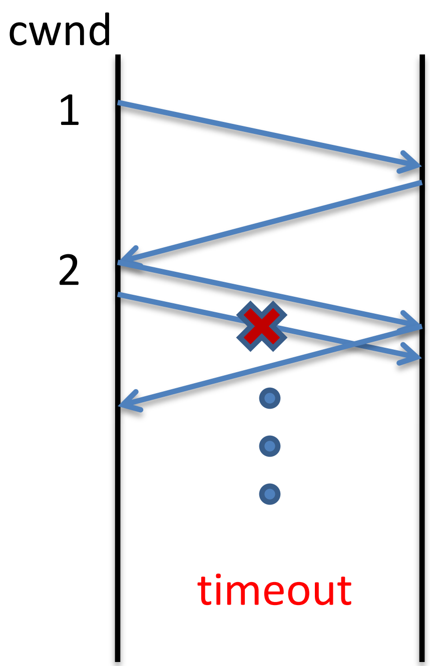


Slow start

- Transmission rate increases slowly
- Does not get the fair share of the bandwidth

Short TCP flows

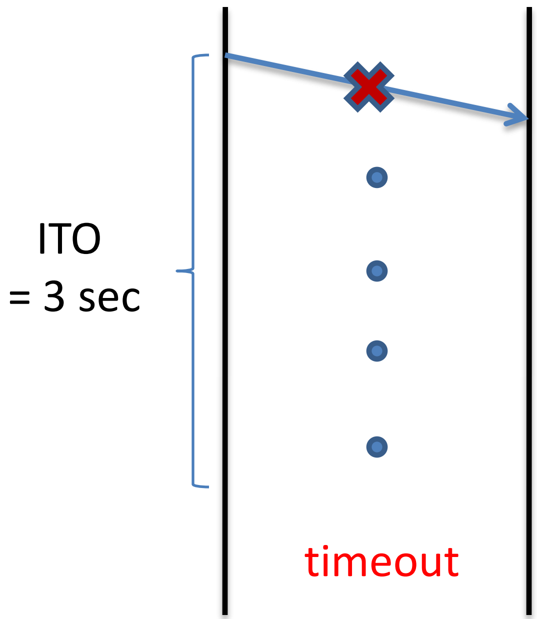
2. Short flows have small congestion window



- Fast retransmit needs 3 dup ACKs
- Small cwnd, not enough packets to activate dup Acks
- So timeout happens
- Timeout severely degrades the performance of TCP

Short TCP flows

3. Conservative Initial Timeout (ITO)



- No sampling data available
- Conservative timeout for (SYN, SYN-ACK) and 1st data packet
- Disastrous effect on short connection performance if these packets lost

Existing and proposed solution

Slow start	Small cwnd & Packet loss	ITO & 1st packet loss
		Reduce ITO
Use large initial window value		
		Get RTT from previous records or neighbors
	Reduce the loss probability these packets	

Preferential treatment to short flows

- Differentiated Services Architecture
  - Classify flows into short and long flows
  - Isolate packets from short flows
  - Reduce the loss probability of these packets

With the help of

- Active Queue Management
  - RED In and Out (RIO)
    - RED with two flow classes (short and long flows)

RIO-PS

RED In and Out with preferential treatment to short flows

Outline

- Introduction
- Analysis and Motivation
- Architecture
- Simulation Results
- Discussion
- Conclusion

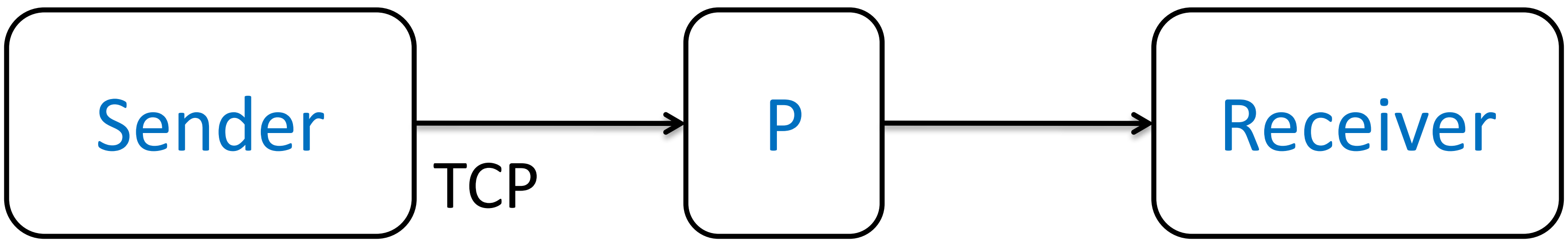
Sensitivity of TCP flows to loss rate

RTO = 4 x RTT

ITO = 3 sec

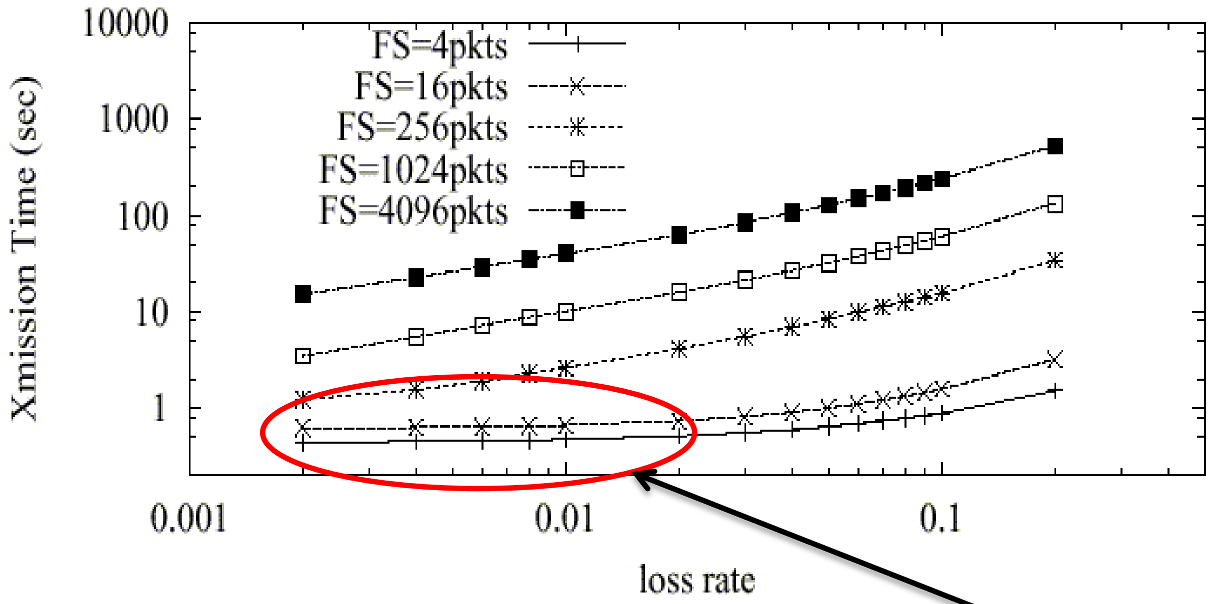
Drops packet with certain probability

RTT = 0.1 sec



4 pkts	Short flows	0.001
16 pkts		⋮
256 pkts		0.01
1024 pkts	Long flows	⋮
4096 pkts		0.1

Average transmission time



(a) Average Transmission Time

No Loss

For short flows,

Xmission time increases drastically after certain loss rate

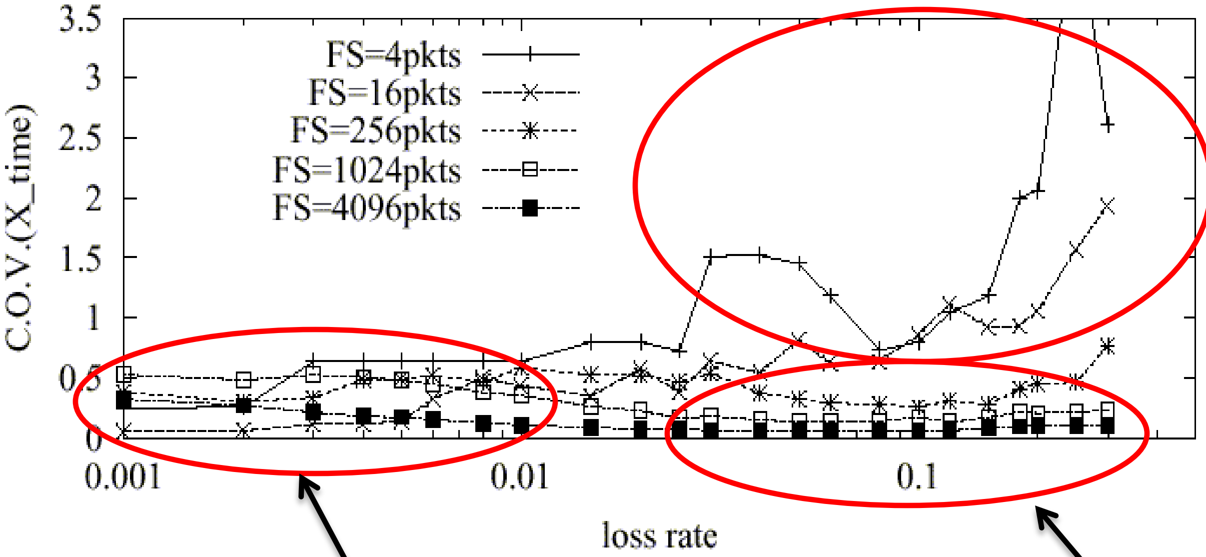
Variance of transmission times

Variation occurs across experiments because

1. When loss rate is high, TCP enters exponential back-off phase
   - ➢ Causes Significantly high variability in transmission time of each individual packet in a flow
2. When loss rate is low, depending on when the loss happens
   - ➢ Slow start phase – aggressive retransmission
   - ➢ Congestion avoidance phase – less aggressive

Variance of transmission times

COV = Standard deviation/mean



(b) Coefficient of Variation

Variability in short flows
Due to 1.
Law of large numbers

Variability in long flows
Due to 2.
Loss in slow start or congestion avoidance

Less variability in long flows
Loss in both slow start and congestion avoidance

Conclusion and Motivation

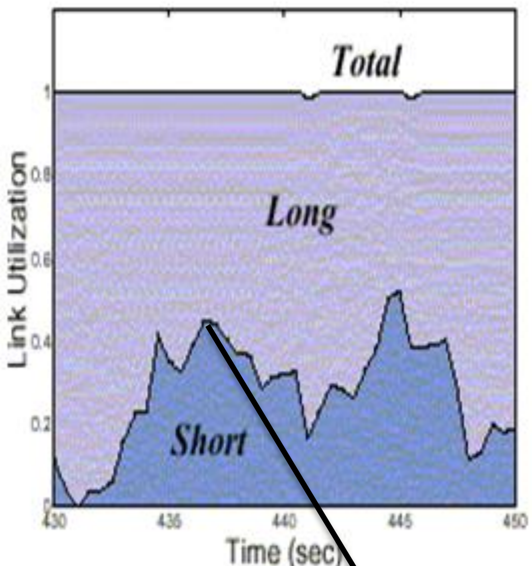
- Short flows are more sensitive to increase in loss probability
- Variability of transmission time is closely related to fairness
- Important to give preferential treatment to short flows
  - Reduce the loss probability for short flows

Preferential treatment to short flows

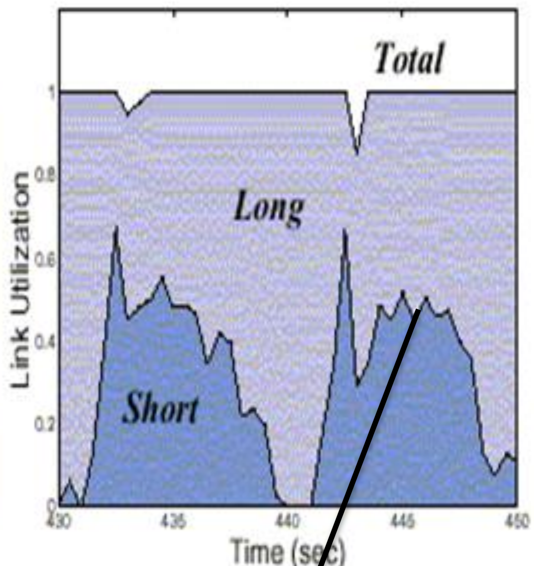
- Simulation – ns simulator
  - 10 long (10000-packet) TCP-Newreno
  - 10 short (100-packet) TCP-Newreno
  - Competing over a 1.25Mbps link
- Vary queue management policy
  - Drop tail
  - RED
  - RIO-PS
    - Reduce the loss probability of short flows

Link Utilization

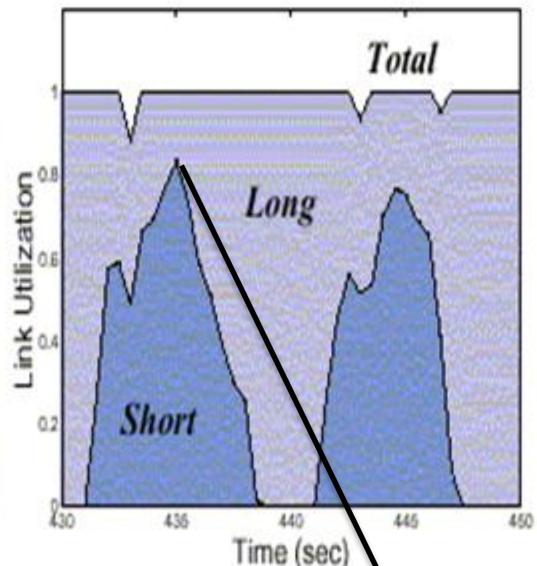
Drop tail



RED



RIO-PS



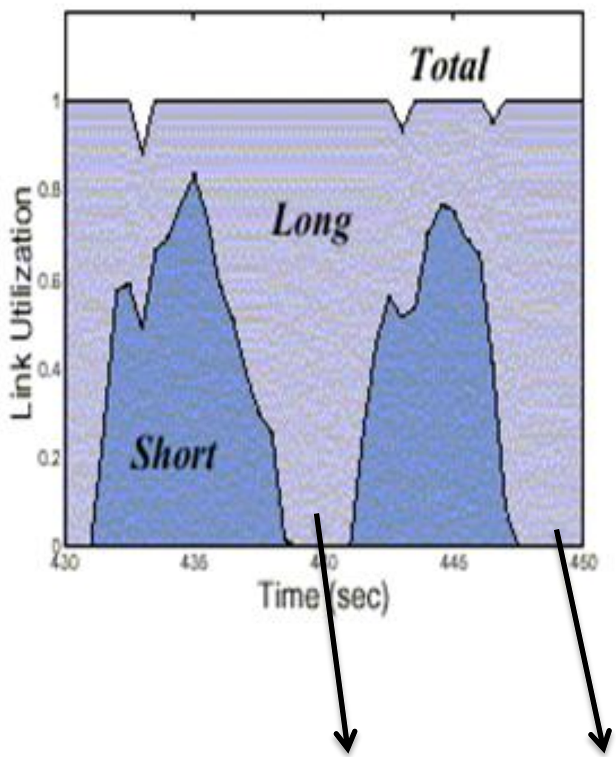
Fails to give fair share to short flows

Favors flows with larger windows

More than fair share to short flows

Almost fair treatment to all flows

Link Utilization - RIO-PS



- Short flows temporarily steal more bandwidth from long flows
- In the long run, their early completion returns an equal amount of resources to long flows

- It might enhance the transmission of long flows
  - Less disturbed by short flow

Network Goodput

More loaded network

RIO-PS has higher goodput

Link B/W	Flows	DropTail	RED	RIO-PS
1.25Mbps	All	153479	154269	154486
	Short	40973	49897	49945
	Long	112506	104372	104541
1.5Mbps	All	185650	184315	183154
	Short	43854	49990	49990
	Long	141796	134325	133164

Less loaded network

DropTail performs slightly better

DropTail drops packets only when queue is full unlike other schemes

Conclusion

- Preferential treatment to short flows
  - Faster response to short flows
  - Improves the overall goodput

Outline

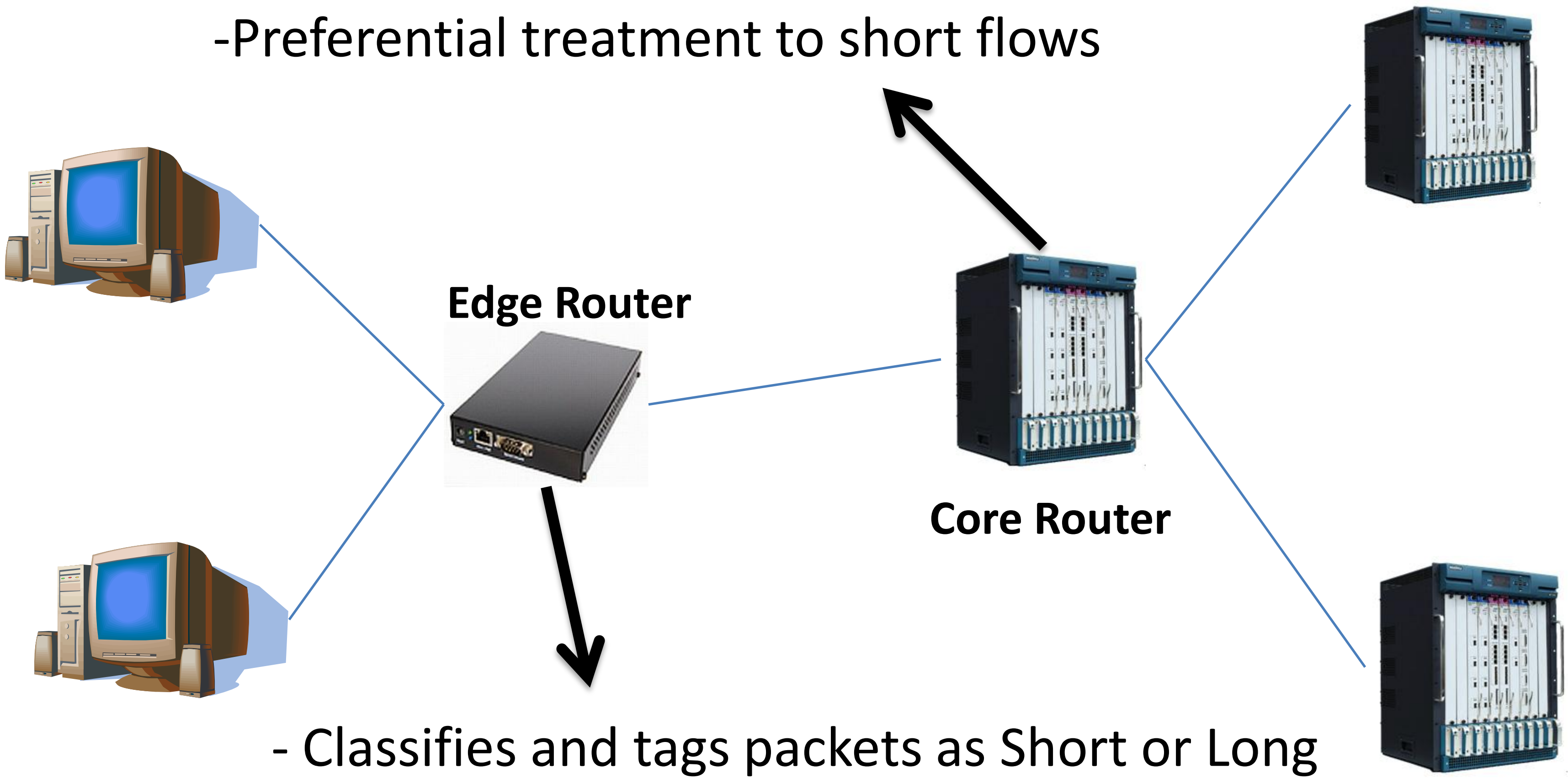
- Introduction
- Analysis and Motivation
- Architecture
- Simulation Results
- Discussion
- Conclusion

Diffserv Architecture

RIO-PS

-Use RED In and Out

-Preferential treatment to short flows



- Classifies and tags packets as Short or Long
- Maintain per flow packet count

Edge Router – Packet classification

Threshold based approach

- Maintains a counter for every flow
  - Counts the number of packet per flow
- Maintain threshold L_t
  - When counter exceeds L_t – tag as long flow
  - Else tag as short flow
- Flow table is updated periodically – Every T_u
  - If no packets from a flow in T_u time units, remove entry

Edge Router – Packet classification

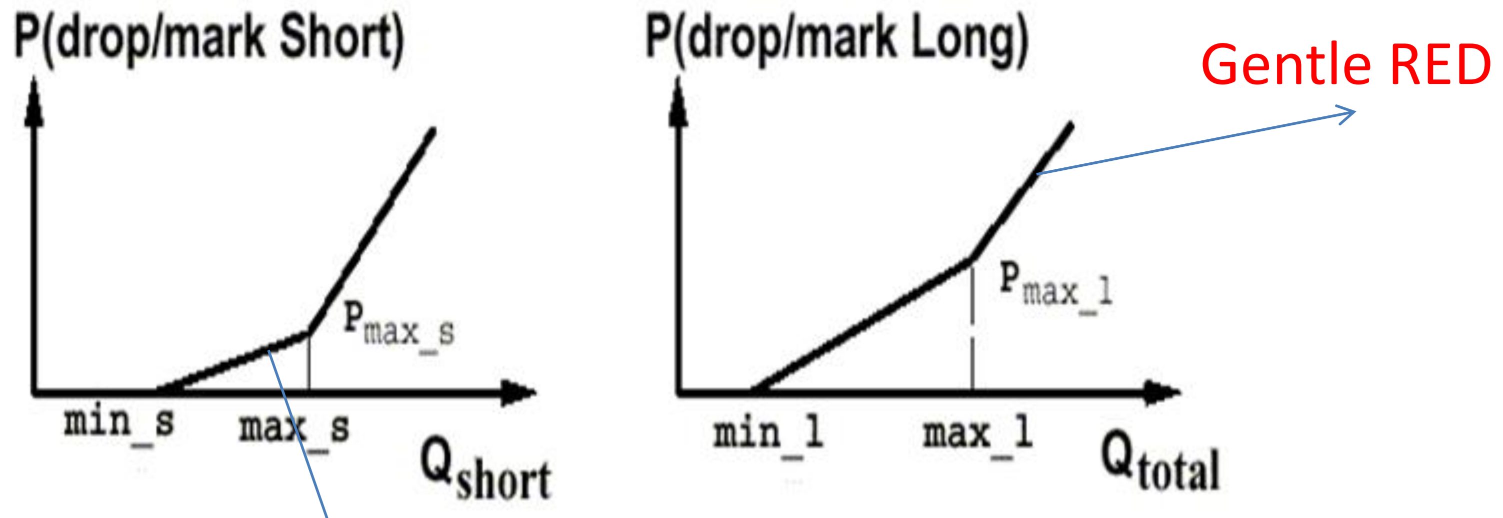
- Threshold L_t adjusted dynamically
  - Balance the number of active short and long flows
- Short-to-Long-Ratio (*SLR*)
  - Configurable parameter
- Every T_c adjust L_t to achieve the target SLR

Core Router – RIO-PS

- RIO - RED with In (Short) and Out (Long)
- Preferential treatment to short flows
  - Short flows
    - Packet dropping probability computed based on the average backlog of short packets only (Q_{short})
  - Long flows
    - Packet dropping probability computed based on the total average queue size (Q_{total})

RIO-PS

Two separate sets of RED parameters for each flow class



Less Packet dropping probability for short flows

Features of RIO-PS

- Single FIFO queue is used for all packets
  - Packet reordering will not happen
- Inherits all properties of RED
  - Protection of bursty flows
  - Fairness within each class of traffic
  - Detection of incipient congestion

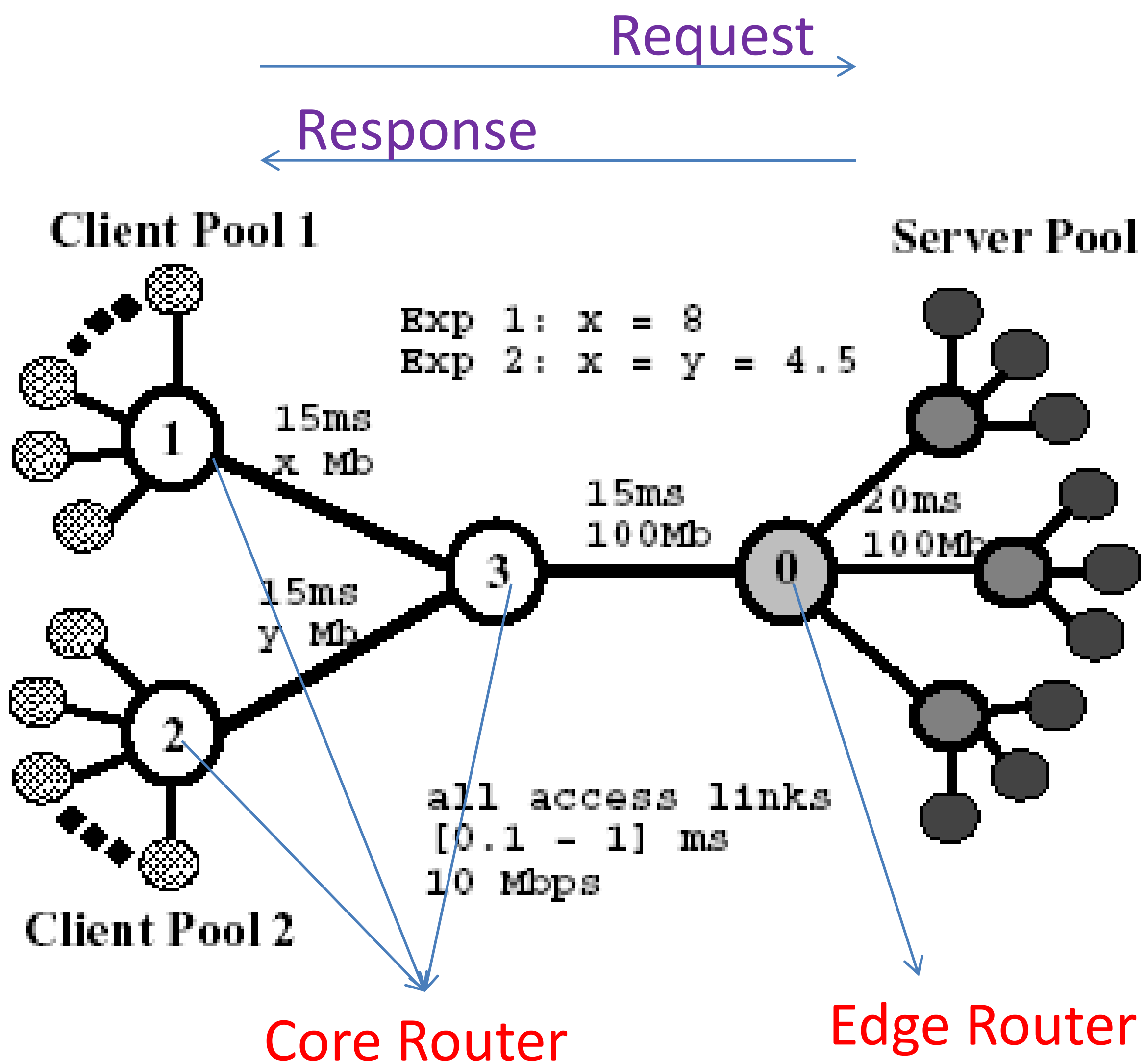
Outline

- Introduction
- Analysis and Motivation
- Architecture
- Simulation Results
- Discussion
- Conclusion

Simulations setup

- ns-2 simulations
- Web traffic model
  - HTTP 1.0
  - Exponential inter-page arrival (mean 9.5 sec)
  - Exponential inter-object arrival (mean 0.05 sec)
  - Uniform distribution of objects per page (min 2 max 7)
  - Object size; bounded Pareto distribution (min = 4 bytes, max = 200 KB, shape = 1.2)
  - Each object retrieved using a TCP connection

Simulation topology



Network configuration

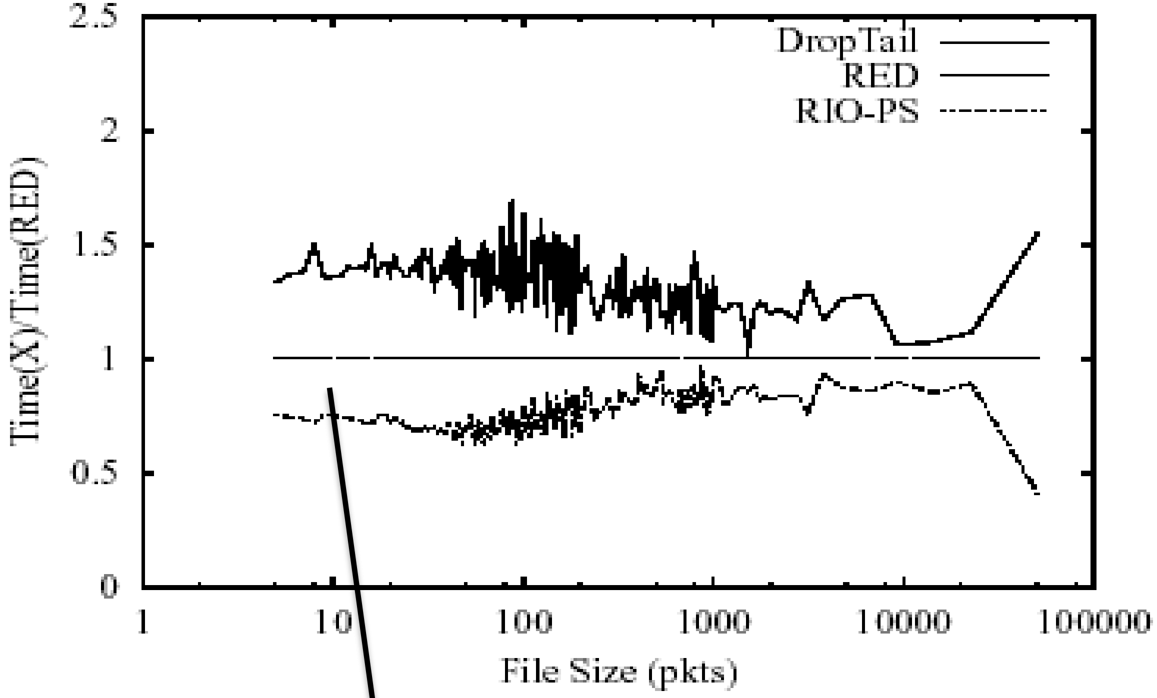
Description	Value
Packet Size	500 bytes
Maximum Window	128 packets
TCP version	Newreno
TCP timeout Granularity	0.1 seconds
Initial Retransmission Timer	3.0 seconds
B/W delay product (BDP)	≈ 200 pkts (Exp1) ≈ 120 pkts (Exp2)
Bottleneck Buffer Size (B)	DropTail: 1.5× BDP RED/RIO-PS: 2.5×BDP
Q. Parameters	$(min_{th}, max_{th}, P_{max}, w_q)$
RED	(0.15B, 0.5B, 1/10, 1/512)
RIO-PS short	(0.15B, 0.35B, 1/20, 1/512)
RIO-PS long	(0.15B, 0.5B, 1/10, 1/512)
RED & RIO-PS	ecn_on, wait_on, gentle_on
Edge Router	$SLR = 3, T_u = 1\ sec, T_c = 10\ sec$
Foreground Traffic	
(Src, Dest)	(Server Pool, Client Pool)
Long Connection Size	1000 packets
Short Connection Size	10 packets

Simulations details

- The load is carefully tuned to be close to the bottleneck link capacity
- RIO parameters
  - Short TCP flows are guaranteed around 75% of the total bandwidth in times of congestion
- Experiments run 4000 seconds with a 2000 second warm-up period

Average response time relative to RED

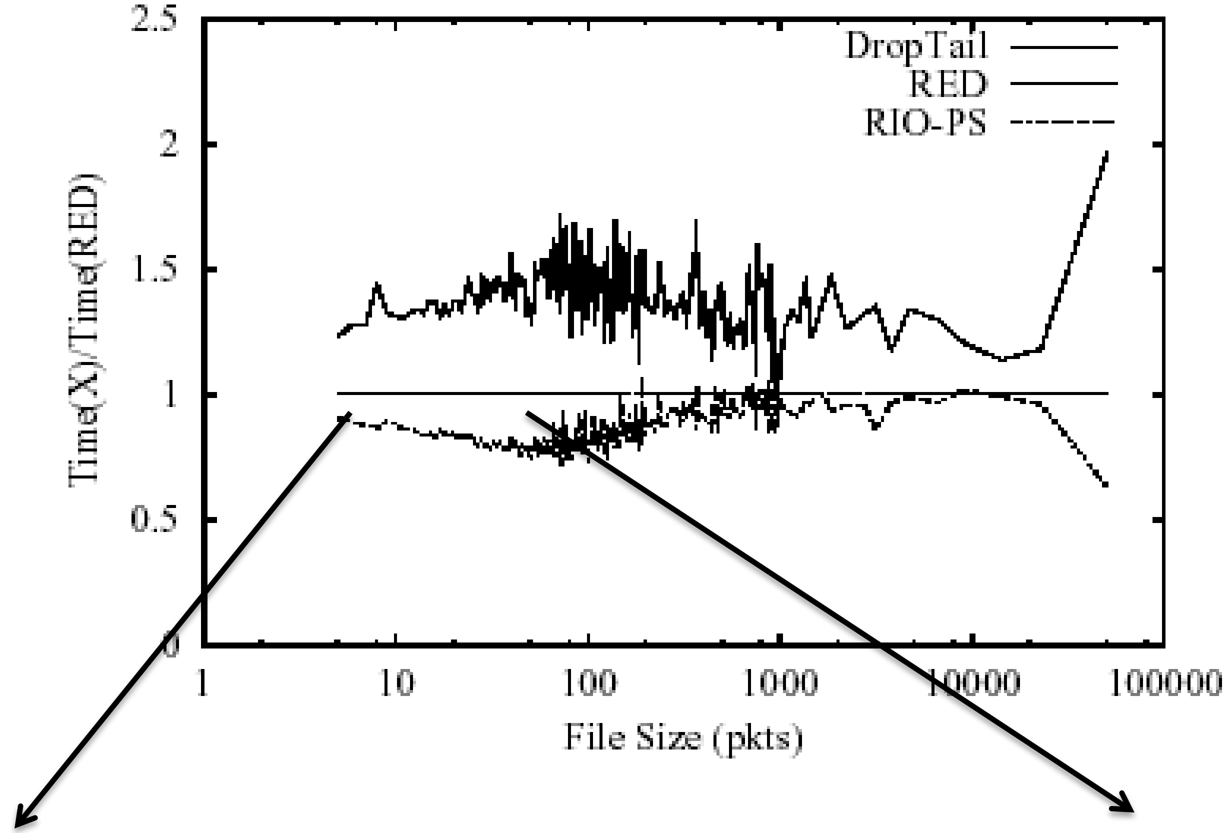
ITO = 3 sec



Average response time reduced by 25-30%
for short and medium sized flows

Average response time relative to RED

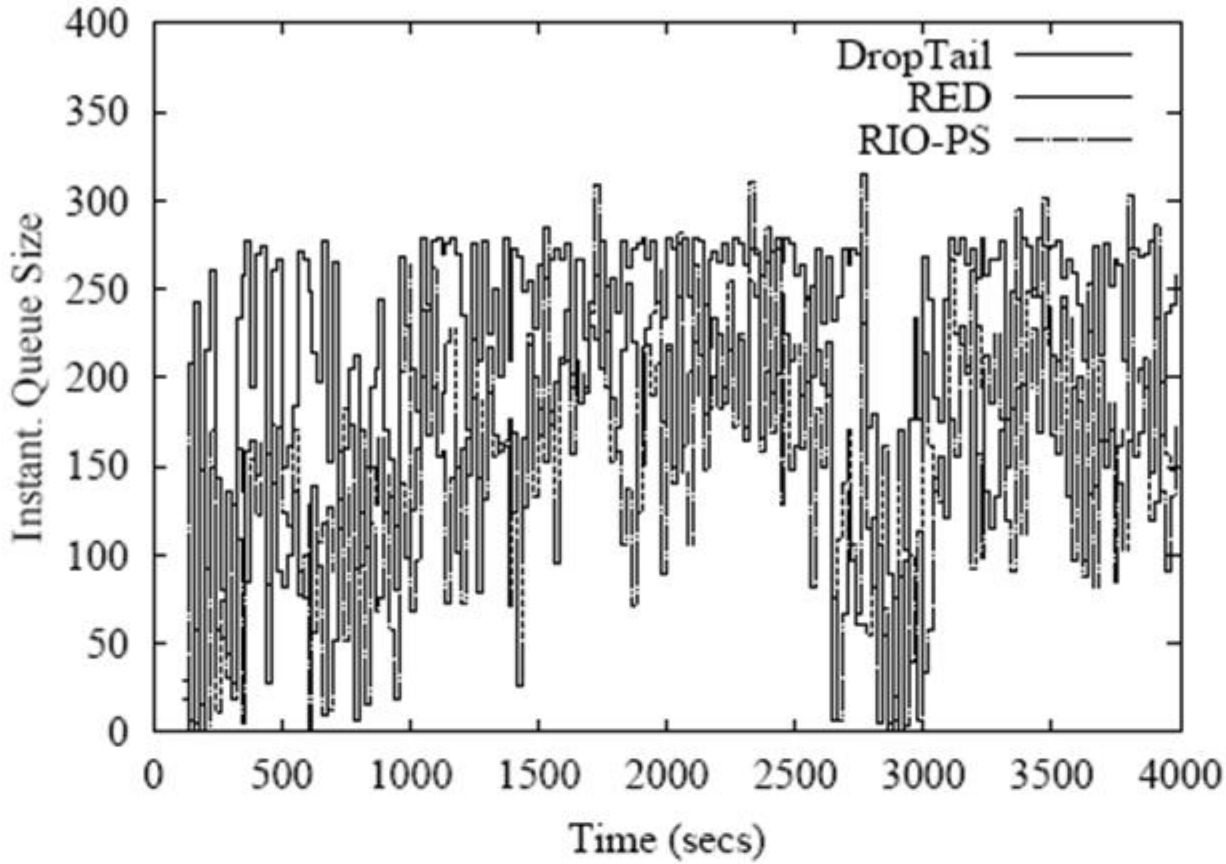
ITO = 1 sec



Average response time reduced by 10-15% for short flows

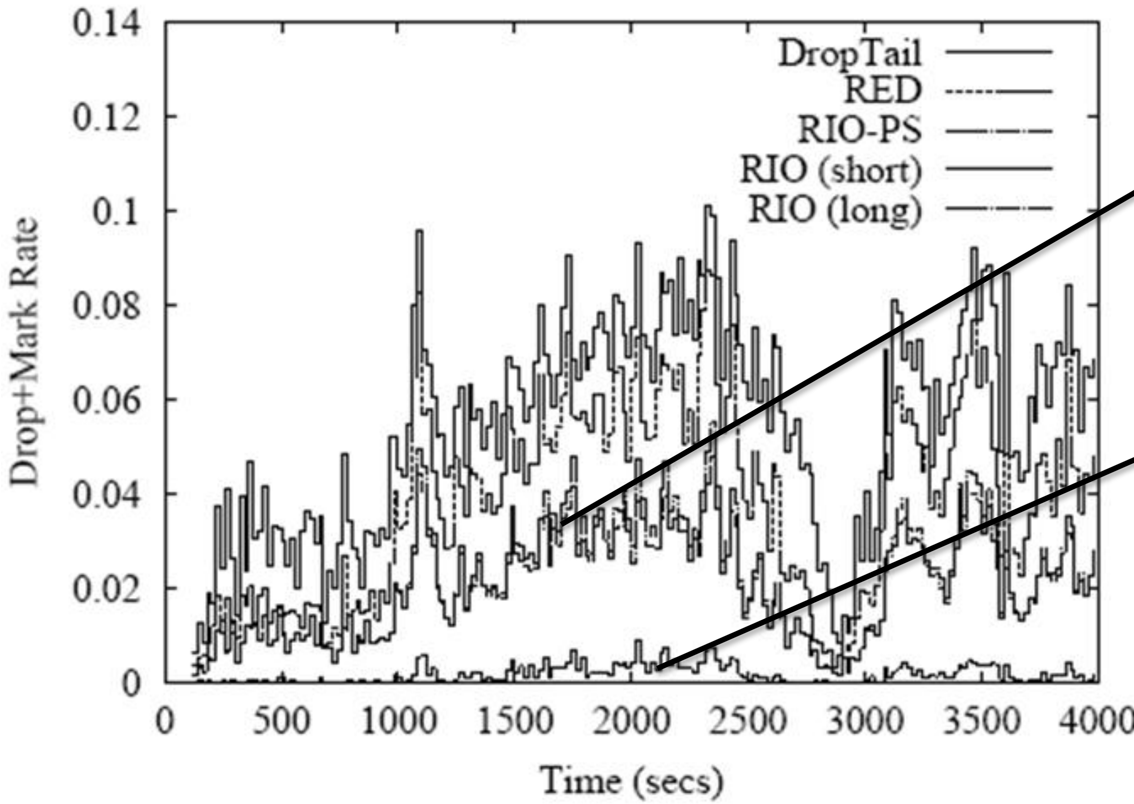
Average response time reduced by 15-25% for medium sized flows

Instantaneous Queue Size



Load in the bottleneck link has high variability over time due to the heavy-tailedness of the file size distribution

Instantaneous Drop/Mark rate



RIO-PS reduces the overall drop/mark probability

Comes from the fact that short flows rarely experience loss

Also, Short TCP flows are not responsible for controlling congestion because of the time scale at which they operate.

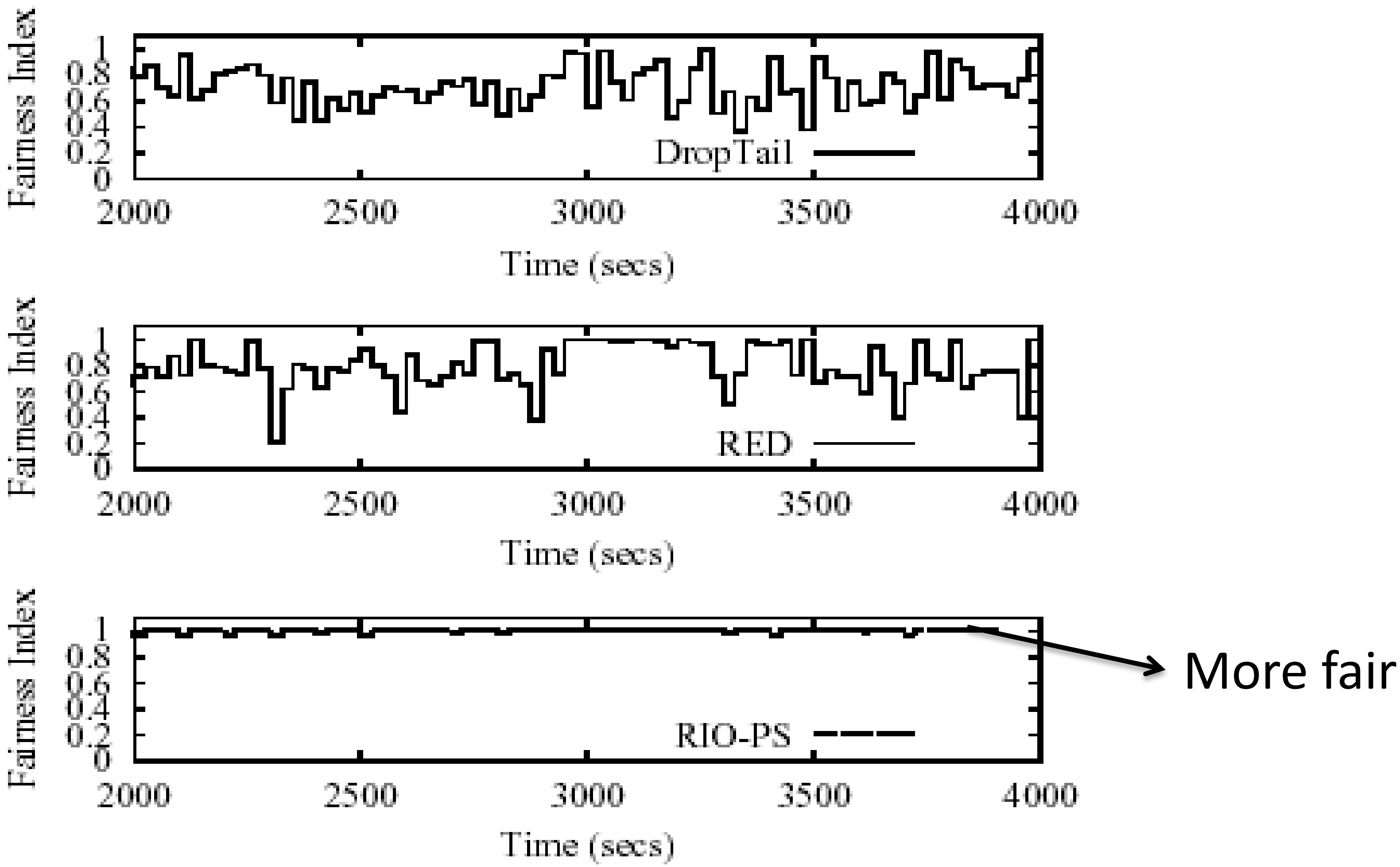
Preferential treatment to short flows does not hurt the network

Study of Foreground Traffic

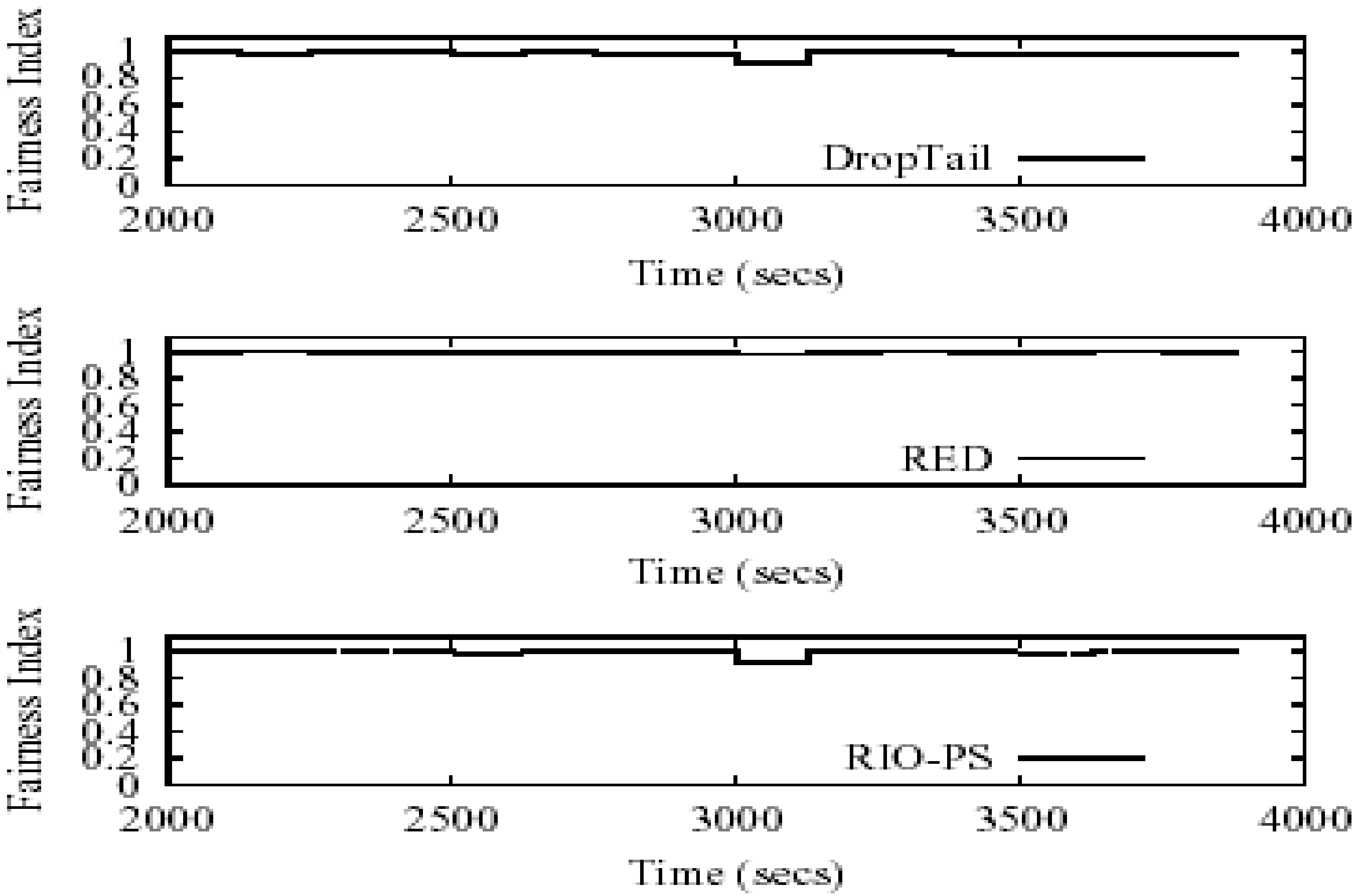
- Periodically inject 10 short flows (every 25 seconds) and 10 long flows (every 125 seconds) as foreground TCP connections and record the response time for i_{th} connection
- Fairness index
  - For any give set of response times (x_1, .., x_n), the fairness index is

$$\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n \sum_{i=1}^{n} x_i^2}$$

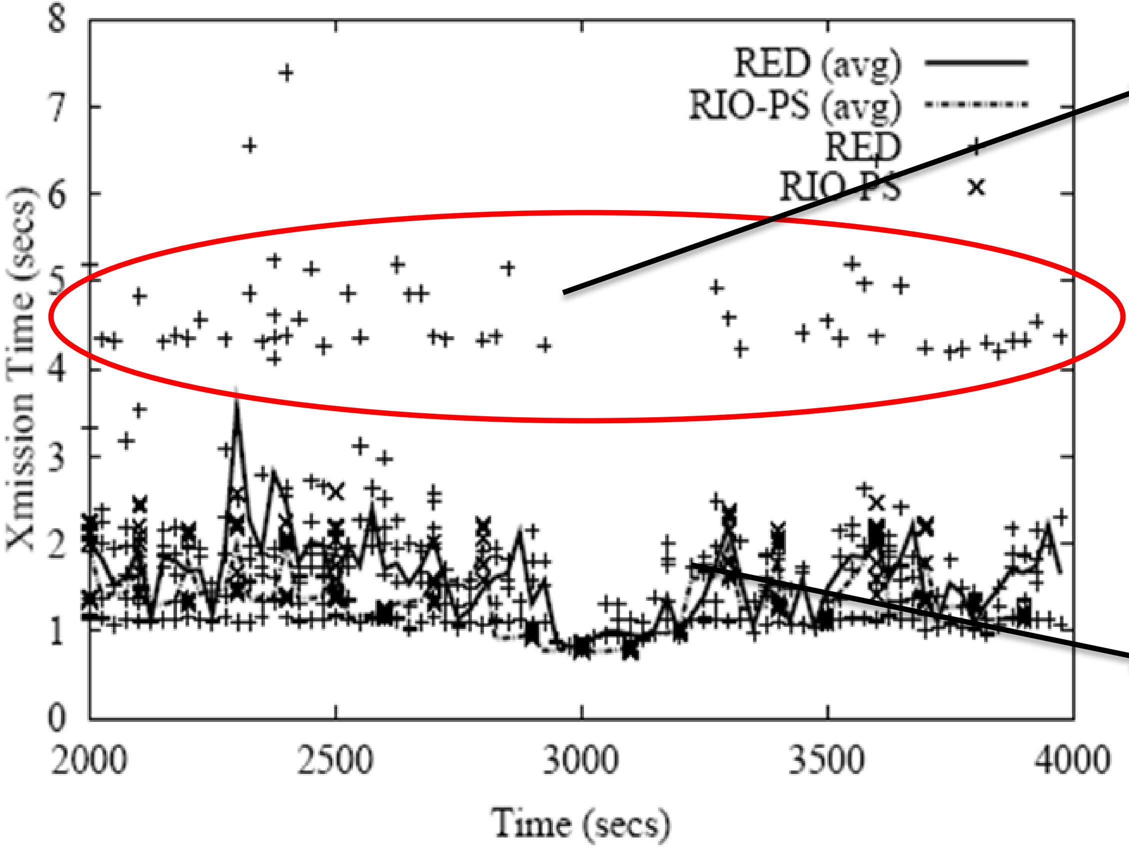
Fairness Index – Short Connections



Fairness Index – Long Connections



Transmission time – short connections

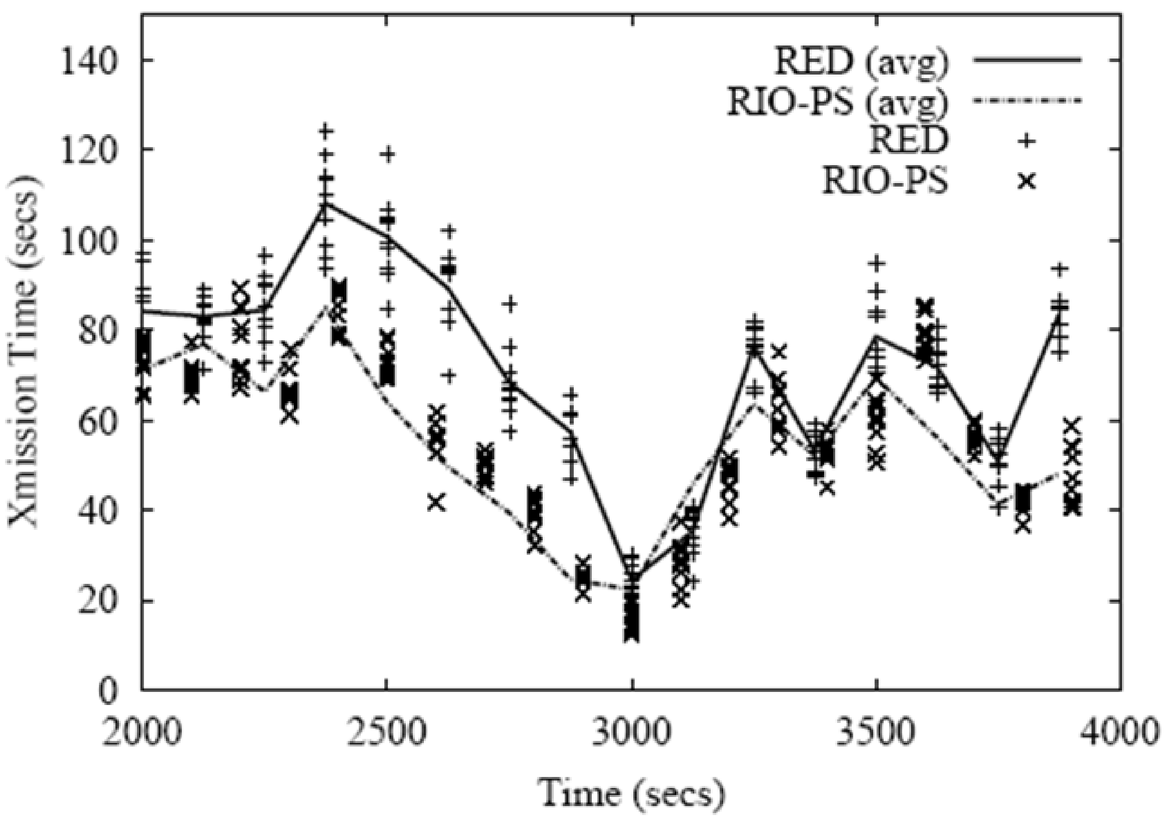


-Even with RED queues, many short flows experience loss

-Some lost first packet and hence timeout (3 sec)

RIO-PS much less drops

Transmission time – long connections



RIO-PS does not hurt long flow performance

Goodput

Scheme	DropTail	RED	RIO-PS
Exp1 (ITO=3sec)	4207841	4264890	4255711
Exp1 (ITO=1sec)	4234309	4254291	4244158
Exp2 (ITO=3sec)	4718311	4730029	4723774

RIO-PS does not hurt overall goodput

Slightly improves over DropTail

Outline

- Introduction
- Analysis and Motivation
- Architecture
- Simulation Results
- Discussion
- Conclusion

Discussion

- Simulation Model
  - Dumbbell and Dancehall (one-way traffic) model
  - All TCP connections have same propagation delay
  - Complicated topologies may impact the performance
- Queue Policy
  - RIO does not provide class based guarantee
  - PI controlled RIO queue or proportional Diffserv gives better control over classified traffic

Discussion

- Deployment Issues
  - Edge routers need to maintain per flow state information.
  - Edge router state maintenance and classification does not have a significant impact on the end to end performance.
  - Incrementally deployable
    - RIO-PS implemented only at bottleneck links
    - Advanced edge devices may be placed in front of busy web server cluster

Discussion

- Flow Classification
  - Threshold based flow classification
  - First few packets of long TCP flow treated same as short flows
  - This mistake enhances performance
    - First few packets of the long flow are similar to short flow and vulnerable to packet losses
    - Makes the system fair to all TCP connections.

Discussion

- Controller Design
  - Edge load control is a topic of further research
  - Preliminary results indicate performance is not sensitive to SLR
  - SLR depends on T_c and T_u
  - Smaller values of T_c and T_u may increase overhead
- Malicious users
  - Users can break their long transmission into small pieces to get fast service
  - This is less likely due to the overhead of fragmentation and reassembly

Outline

- Introduction
- Analysis and Motivation
- Architecture
- Simulation Results
- Discussion
- Conclusion

Conclusion

- TCP major traffic in the Internet
- Proposed Scheme is a Diffserv like architecture
  - Edge routers classifies TCP flow as long or short
  - Core routers implements RIO-PS
- Advantages
  - Short flow performance improved in terms of fairness and response time.
  - Long flow performance is also improved or minimally affected since short flows are rapidly served.
  - System overall goodput is improved
  - Flexible Architecture, can be tuned largely at edge routers

Acknowledgements

- Thanks to Professor. Bob Kinicki, Matt Hartling and Sumit Kumbhar.
- Some figures in this presentation are taken from their class presentation and modified.