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

80% Long flows

20% Short flows
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

- Slow start
- Fast retransmit
- 3 Dup Acks
- Fast recovery
- Timeout
- Slow start

Time (sec) vs cwnd (packets) graph.
Short TCP flows

1. Most short flows finish before slow start finish

- 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 Ack
- 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 1\textsuperscript{st} data packet
- Disastrous effect on short connection performance if these packets lost

ITO = 3 sec
Existing and proposed solution

<table>
<thead>
<tr>
<th>Slow start</th>
<th>Small cwnd &amp; Packet loss</th>
<th>ITO &amp; 1st packet loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use large initial window value</td>
<td></td>
<td>Reduce ITO</td>
</tr>
<tr>
<td></td>
<td>Get RTT from previous records or neighbors</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reduce the loss probability these packets</td>
</tr>
</tbody>
</table>
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

Sender

TCP

P

Receiver

4 pkts  0.001
16 pkts  0.01
256 pkts
1024 pkts
4096 pkts

Short flows

Long flows
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

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
- Fails to give fair share to short flows
- Favors flows with larger windows

RED
- Almost fair treatment to all flows

RIO-PS
- More than fair share to short 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

Less loaded network

DropTail performs slightly better
DropTail drops packets only when queue is full unlike other schemes

More loaded network

RIO-PS has higher goodput

<table>
<thead>
<tr>
<th>Link B/W</th>
<th>Flows</th>
<th>DropTail</th>
<th>RED</th>
<th>RIO-PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25Mbps</td>
<td>All</td>
<td>153479</td>
<td>154269</td>
<td>154486</td>
</tr>
<tr>
<td></td>
<td>Short</td>
<td>40973</td>
<td>49897</td>
<td>49945</td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>112506</td>
<td>104372</td>
<td>104541</td>
</tr>
<tr>
<td>1.5Mbps</td>
<td>All</td>
<td>185650</td>
<td>184315</td>
<td>183154</td>
</tr>
<tr>
<td></td>
<td>Short</td>
<td>45854</td>
<td>49990</td>
<td>49990</td>
</tr>
<tr>
<td></td>
<td>Long</td>
<td>141796</td>
<td>134325</td>
<td>133164</td>
</tr>
</tbody>
</table>
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

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

**Core Router**
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 \( (Q_{\text{short}}) \)
  - Long flows
    - Packet dropping probability computed based on the total average queue size \( (Q_{\text{total}}) \)
RIO-PS

Two separate sets of RED parameters for each flow class

Less Packet dropping probability for short flows

Gentle RED
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

Request

Response

Core Router

Edge Router

Client Pool 1

Server Pool

Client Pool 2

Exp 1: x = 8
Exp 2: x = y = 4.5

15ms
x Mb

15ms
y Mb

15ms
100Mb

20ms
100Mb

all access links
[0.1 - 1] ms
10 Mbps
Network configuration

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Packet Size</td>
<td>500 bytes</td>
</tr>
<tr>
<td>Maximum Window</td>
<td>128 packets</td>
</tr>
<tr>
<td>TCP version</td>
<td>Newreno</td>
</tr>
<tr>
<td>TCP timeout Granularity</td>
<td>0.1 seconds</td>
</tr>
<tr>
<td>Initial Retransmission Timer</td>
<td>3.0 seconds</td>
</tr>
<tr>
<td>B/W delay product (BDP)</td>
<td>( \approx 200 ) pkts (Exp1)</td>
</tr>
<tr>
<td></td>
<td>( \approx 120 ) pkts (Exp2)</td>
</tr>
<tr>
<td>Bottleneck</td>
<td>DropTail: (1.5 \times BDP)</td>
</tr>
<tr>
<td>Buffer Size (B)</td>
<td>RED/RIO-PS: (2.5 \times BDP)</td>
</tr>
</tbody>
</table>

### Q. Parameters

\[
(m_{\text{ini}}, m_{\text{max}}, P_{\text{max}}, w_q)
\]

- **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**
  \[
  \text{ecn\_on, wait\_on, gentle\_on}
  \]
- **Edge Router**
  \[
  SLR = 3, T_u = 1 \text{ sec}, T_c = 10 \text{ sec}
  \]

### Foreground Traffic

<table>
<thead>
<tr>
<th>(Src, Dest)</th>
<th>(Server Pool, Client Pool)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long Connection Size</td>
<td>1000 packets</td>
</tr>
<tr>
<td>Short Connection Size</td>
<td>10 packets</td>
</tr>
</tbody>
</table>
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
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 \((x_1, \ldots, x_n)\), the fairness index is

\[
\text{Fairness index} = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n \sum_{i=1}^{n} x_i^2}
\]
Fairness Index – Short Connections

More fair
Fairness Index – Long Connections

![Graphs showing Fairness Index over time for DropTail, RED, and RIO-PS.]
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

<table>
<thead>
<tr>
<th>Scheme</th>
<th>DropTail</th>
<th>RED</th>
<th>RIO-PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp1 (ITO=3sec)</td>
<td>4207841</td>
<td>4264890</td>
<td>4255711</td>
</tr>
<tr>
<td>Exp1 (ITO=1sec)</td>
<td>4234309</td>
<td>4254291</td>
<td>4244158</td>
</tr>
<tr>
<td>Exp2 (ITO=3sec)</td>
<td>4718311</td>
<td>4730029</td>
<td>4723774</td>
</tr>
</tbody>
</table>

- **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.