97-0831: GFR -- Providing Rate Guarantees with FIFO Buffers to TCP Traffic

Rohit Goyal, Raj Jain, Sonia Fahmy, Bobby Vandalore, Shivkumar Kalyanaraman

Sastri Kota, Lockheed Martin Telecommunications
Pradeep Samudra, Samsung Telecom America, Inc.

Contact: jain@cis.ohio-state.edu
http://www.cis.ohio-state.edu/~jain/

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

The Ohio State University

Raj Jain
Overview

- Guaranteed frame rate
- Goals of this study
- Controlling TCP windows
- Differential Fair Buffer Allocation
- Simulation results
Guaranteed Frame Rate (GFR)

- GFR guarantees:
  - Low loss ratio to conforming frames
  - Best effort to all frames
  - Fair share of unused capacity
    (Not well defined. May be removed.)
- User specifies an MCR and a maximum frame size
- Conforming Frames = Frames which are untagged by the end system and pass the GCRA like policing mechanism.
Motivation

- GFR VCs could be used by routers separated by an ATM cloud.
- Users could also set up GFR VCs for traffic that could benefit from rate guarantees.
- Higher layers would expect some guarantees at that level.
- Higher layer traffic management may interact with GFR traffic management and achieve unfair throughput.
- A good GFR implementation should “work with” most common traffic types.
GFR Implementation Issues

- FIFO queuing versus per-VC queuing
  - Per-VC queuing is too expensive.
  - FIFO queuing should work by setting thresholds based on bandwidth allocations.

- 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. [??]

- Buffer management policies
  - Per-VC accounting policies need to be studied
Summary of Past Results

- In the July meeting it was shown
  - Difficult to guarantee TCP throughput with FIFO queuing.
  - Can do so with per-VC queuing.
- 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.
Goals

- Provide minimum rate guarantees with FIFO buffer for TCP/IP traffic.
- Guarantees in the form of TCP throughput.
- 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 = \( \frac{(\text{Avg wnd})}{(\text{Round trip time})} \)
- With Selective Ack (SACK), window decreases by 1/2 during packet loss, and then increases linearly.
  - Avg wnd = \( \left[ \sum_{i=1,\ldots,n} (\text{max wnd}/2 + \text{mss}\times i) \right]/n \)
FIFO Buffer Management

- Fraction of buffer occupancy \((X_i/X)\) determines the fraction of output rate \((\mu_i/\mu)\) 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.

\[ \frac{X_i/X}{\mu_i/\mu} = 1 \]
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.
- $K = \text{Total buffer capacity}$.
- Target utilization = $\sum R_i / K$.
- Guaranteed TCP throughput = $\text{Capacity} \times R_i / K$
- Expected throughput, $\mu_i = \mu \times R_i / \sum R_i$. ($\mu = \sum \mu_i$)
Simulation Configuration

- SACK TCP.
- 15 TCP sources \((N = 15)\).
- Buffer Size \(= K = 48000\) cells.
- 5 thresholds \((R_1, \ldots, R_5)\).
Threshold $R_{ij} \propto \left[ K \times \frac{MCR_i}{PCR} \right]$

Total throughput $\mu = 126$ Mbps. MSS = 1024B.

Expected throughput $= \mu \times \frac{R_i}{\sum R_i}$

<table>
<thead>
<tr>
<th>Sources</th>
<th>Expt 1</th>
<th>Expt 2</th>
<th>Expt 3</th>
<th>Expt 4</th>
<th>Expected Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3 (R₁)</td>
<td>305</td>
<td>458</td>
<td>611</td>
<td>764</td>
<td>2.8 Mbps</td>
</tr>
<tr>
<td>4-6 (R₂)</td>
<td>611</td>
<td>917</td>
<td>1223</td>
<td>1528</td>
<td>5.6 Mbps</td>
</tr>
<tr>
<td>7-9 (R₃)</td>
<td>917</td>
<td>1375</td>
<td>1834</td>
<td>2293</td>
<td>8.4 Mbps</td>
</tr>
<tr>
<td>10-24 (R₄)</td>
<td>1223</td>
<td>1834</td>
<td>2446</td>
<td>3057</td>
<td>11.2 Mbps</td>
</tr>
<tr>
<td>13-15 (R₅)</td>
<td>1528</td>
<td>2293</td>
<td>3057</td>
<td>3822</td>
<td>14.0 Mbps</td>
</tr>
<tr>
<td>$\sum R_i/K$</td>
<td>29%</td>
<td>43%</td>
<td>57%</td>
<td>71%</td>
<td></td>
</tr>
</tbody>
</table>
Simulation Results

<table>
<thead>
<tr>
<th>TCP Number</th>
<th>Throughput ratio (observed / expected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>1.0 1.03 1.02 1.08</td>
</tr>
<tr>
<td>4-6</td>
<td>0.98 1.01 1.03 1.04</td>
</tr>
<tr>
<td>7-9</td>
<td>0.98 1.00 1.00 1.02</td>
</tr>
<tr>
<td>10-12</td>
<td>0.98 0.99 0.98 0.88</td>
</tr>
<tr>
<td>13-15</td>
<td>1.02 0.98 0.97 1.01</td>
</tr>
</tbody>
</table>

- 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 ⇒ Relative throughput per connection is proportional to fraction of buffer occupancy.
- Controlling TCP buffer occupancy ⇒ May control throughput.
- High buffer utilization ⇒ Harder to control throughput.
- Formula does not hold for very low buffer utilization
  Very small TCP windows ⇒ SACK TCP times out if half the window is lost
Differential Fair Buffer Allocation

<table>
<thead>
<tr>
<th>K</th>
<th>R</th>
<th>R₁</th>
<th>R₂</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>WᵢRᵢ</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>R₃</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **X > R** ⇒ EPD
  - Drop all tagged
- **Xᵢ > Rᵢ** ⇒ Probabilistic Loss,
- **Xᵢ ≤ Rᵢ** ⇒ No Loss

- **Wᵢ** = Weight of VCi.
- **Rᵢ** = per-VC threshold (Rᵢ depends on Wᵢ).
- **Xᵢ** = per-VC buffer occupancy. (X = Σ Xᵢ)
- **Z > 1. Z*Rᵢ** = per-VC high threshold.

The Ohio State University
Raj Jain
Differential Fair Buffer Allocation

When first cell of frame arrives:

- IF \( X_i < R_i \) THEN
  - Accept frame

- ELSE IF \( X > R \) OR \( X_i > Z*R_i \) THEN
  - Drop frame

- ELSE IF \( X < R \) THEN
  - Drop cell and frame with
    \[ P\{\text{drop}\} = W_i \times \frac{X_i - R_i}{R_i \times (Z-1)} \]
Drop Probability

- Increases as $X_i$ increases above $R_i$
  - Indicates higher levels of congestion.
- Proportional to $W_i$
  - With larger window, more packets can be dropped without timing out.
- $X_i > Z*R_i \implies$ EPD is performed.
DFBA Simulation Configuration

TCP 1
TCP 3

Switch

Switch

VC1

1000 km

Switch

Switch

TCP 12
TCP 15

VC5

10 km

1 km

Destination 1

Destination 3

Switch

Switch

Destination 12

Destination 15

Raj Jain

The Ohio State University
DFBA Simulation Configuration

- SACK TCP, 15 TCP sources.
- 5 VCs through backbone link. 3 TCP’s per VC.
- Local switches merge TCP sources.

<table>
<thead>
<tr>
<th>VC Number</th>
<th>Thresholds for backbone switch</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>152  305  611</td>
</tr>
<tr>
<td>2</td>
<td>305  611  1223</td>
</tr>
<tr>
<td>3</td>
<td>458  917  1834</td>
</tr>
<tr>
<td>4</td>
<td>611  1223  2446</td>
</tr>
<tr>
<td>5</td>
<td>764  1528  3057</td>
</tr>
</tbody>
</table>
Simulation Results

<table>
<thead>
<tr>
<th>VC Number</th>
<th>Throughput Ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.04 1.01 <strong>1.16</strong></td>
</tr>
<tr>
<td>2</td>
<td>1.05 1.02 1.06</td>
</tr>
<tr>
<td>3</td>
<td>0.97 1.03 1.05</td>
</tr>
<tr>
<td>4</td>
<td>0.93 1.00 <strong>1.13</strong></td>
</tr>
<tr>
<td>5</td>
<td>1.03 0.99 <strong>0.80</strong></td>
</tr>
</tbody>
</table>

- Achieved throughput per-VC proportional to fraction of threshold allocated to the VC.
- Higher variation with increase in buffer allocation.
Summary

- SACK TCP throughput may be controlled with FIFO queuing under certain circumstances:
  - TCP, SACK (?)
  - \( \sum \) MCRs < Uncommitted bandwidth
  - Same RTT (?), Same frame size (?)
  - No other non-TCP or higher priority traffic (?)

The Ohio State University

Raj Jain
Future Work

- Other TCP versions.
- Effect to non-adaptive (UDP) traffic
- Effect of RTT
- Effect of tagging
- Effect of frame sizes
- Parameter study
- Buffer threshold setting formula?
- How much buffer can be utilized?