Spatial Reuse Protocol Fairness (SRP-fa) and Performance Evaluation

Donghui Xie
Cisco Systems
March 14, 2001
SRP-fa Agenda

- Fairness as An Objective
- SRP Overview
- SRP Fairness Algorithm
- SRP-fa Simulation Evaluation
- Summary
- Appendix
Fairness as An Objective

- Equal opportunity access to ring bandwidth for all stations, no single station should be starved from ring bandwidth.
- Simplify and support distributed dynamic ring bandwidth management.
  - Efficient ring bandwidth allocation and utilization
- Support ring station plug and play by eliminating explicit node ring bandwidth fairness or unfairness configuration, otherwise, it may involve reconfiguring all the nodes on the ring.
- Support great and complex QoS features in higher layer traffic management by providing consistent and deterministic ring access rate.
SRP Fairness Algorithm

- A distributed algorithm
  - each node executes a local copy of SRP-fa
- Periodically propagate and use bandwidth usage information to ensure global fairness
- Control low priority packets ring insertion rate and forwarding rate
- Ensure rapid fairness convergence and adaptation
- Guarantee packet delivery once it is on the ring (no packet loss on the ring)

**Reference:**
SRP-fa Fairness Control

• High Priority Host Packets Are Not SRP-fa Rate Controlled

• SRP Transmit Order
  – High priority transit packets
  – Low priority transit packets if Low Priority TB is full
  – High priority host packets
  – Low priority transit packets if LP TB exceeds low threshold
  – Low priority host packets
  – Low priority transit packets

• Low Priority Host Packets Throttled When
  – My_usage > Allow_usage
  – My_usage > Max_allow
  – LP TB is not empty
    and My_usage > Fwd_rate
SRP-fa Simulation Evaluation

- Simulation One:
  VoIP and TCP applications performance over DPT-OC12 ring

- Simulation Two:
  Unevenly distributed TCP traffic performance over DPT-OC12 ring
• DPT-OC12 ring with 34 nodes
• Link propagation delay 200us (40km), total aggregation link latency 3ms.
• 12 nodes aggregation, routing node ip forwarding speed is 5.32Mpps
• Http, Ftp, UDP and VoIP traffic aggregate to destinations attached to DPT/SRP node_0
• 500 simultaneous callers in each call group
• SRP Configuration:
  → HP transmit buffer 5.6Kbytes
  → HP transit buffer 5.6Kbytes
  → LP transit buffer 512Kbytes
  → LP transmit buffer 512Kbytes
  → LP Tb low threshold 128Kbytes
  → LP Tb high threshold 500Kbytes
  → Max_allow 32000
Simulation Runs

- Referenced VoIP traffics are from CalleeGroup1 (55Mbps) and CallerGroup2 (49Mbps).
- There are four simulation runs
  - Link utilization 70%: (5 node aggregation)
    - VoIP from CalleeGroup1 and CallerGroup2, total 104Mbps
    - Http traffic from WebServer and WebServer2, total 84Mbps
    - Ftp traffic from FtpClient1, 9 and 11, total 168Mbps
    - UDP traffic from UDP_Gen3, total 80Mbps
  - Link utilization 86%: (6 node aggregation)
    - VoIP same as first run
    - Http traffic from WebServer, WebServer2 and 3, total 124Mbps
    - Ftp traffic from FtpClient1, 5, 9 and 11, total 224Mbps
    - UDP traffic from UDP_Gen3, total 80Mbps
  - Link utilization > 100%: (11 node aggregation)
    - VoIP same as first run
    - Http traffic from WebServer, WebServer2, 3 and 4, total 160Mbps
    - Ftp traffic from FtpClient1, 3, 5, 7, 9, 11, 13 and 15, total 304Mbps
    - UDP traffic from UDP_Gen1, 2 and 3, total 250Mbps
  - Link utilization >100% (12 node aggregation)
    - 50Mbps more VoIP traffic from CallerGroup4 to CalleeGroup2, total 150Mbps
    - Http, Ftp and UDP traffics are the same as the third run

---

[Diagram showing total traffic in bits/sec for different runs]
TCP Configuration and Sampled Ftp Traffic Source Profile

- **TCP Configuration**
  - TCP Tahoe with fast retransmission
  - No fast recovery
  - No window scaling
  - Buffer size: 65535 bytes

- **FTP Traffic Configuration**
  - 140 simultaneous users
  - Exponential ftp request inter-arrival, mean 2sec
  - Exponential file size, mean 100kbytes
  - Overall average 56Mbps

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Segment Size (bytes)</td>
<td>Auto-Assigned</td>
</tr>
<tr>
<td>Receive Buffer (bytes)</td>
<td>65536</td>
</tr>
<tr>
<td>Receive Buffer Usage Threshold (of RCV BUFF)</td>
<td>0.0</td>
</tr>
<tr>
<td>Delayed ACK Mechanism</td>
<td>Segment/Clock Based</td>
</tr>
<tr>
<td>Maximum ACK Delay (sec)</td>
<td>0.200</td>
</tr>
<tr>
<td>Slow-Start Initial Count (MSS)</td>
<td>1</td>
</tr>
<tr>
<td>Fast Retransmit</td>
<td>Enabled</td>
</tr>
<tr>
<td>Fast Recovery</td>
<td>Disabled</td>
</tr>
<tr>
<td>Window Scaling</td>
<td>Disabled</td>
</tr>
<tr>
<td>Selective ACK (SACK)</td>
<td>Disabled</td>
</tr>
<tr>
<td>Nagle’s SWS Avoidance</td>
<td>Disabled</td>
</tr>
<tr>
<td>Karn’s Algorithm</td>
<td>Enabled</td>
</tr>
<tr>
<td>Retransmission Thresholds</td>
<td>Attempts Based</td>
</tr>
<tr>
<td>Initial RTO (sec)</td>
<td>1.0</td>
</tr>
<tr>
<td>Minimum RTO (sec)</td>
<td>0.5</td>
</tr>
<tr>
<td>Maximum RTO (sec)</td>
<td>64.0</td>
</tr>
<tr>
<td>RTT Gain</td>
<td>0.125</td>
</tr>
<tr>
<td>Deviation Gain</td>
<td>0.25</td>
</tr>
<tr>
<td>RTT Deviation Coefficient</td>
<td>4.0</td>
</tr>
<tr>
<td>Timer Granularity (sec)</td>
<td>0.5</td>
</tr>
<tr>
<td>Persistence Timeout (sec)</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Sampled VoIP and HTTP Traffic Source Profile

- VoIP traffic profile from CalleeGroup1 and CallerGroup2
- 500 simultaneous callers in each LAN, exponential talk duration (7min), erlang interarrival process (scale 1, shape 6)
- Voice talk spurt exponential (0.352 sec)/silence (0.65 sec)
- Voice encoding: G.711
- 1 voice frame per packet

- Http1.1 traffic profile from WebServer and WebServer2
- 140 simultaneous users in each LAN
- Exponential page interarrival process
- Object number per page: exponential with mean 5
- Object size: exponential with mean 60k bytes
VoIP Traffic on the Ring between the Runs

VoIP traffic sourced on the ring at Node_1

VoIP traffic sourced on the ring at Node_9
- Cumulative Distribution Function (CDF) for voice packet delay
- Largest delay variation is 300us
- As more high priority traffic aggregates on the ring, its delay gets smaller
CallerGroup2 VoIP Performance
VoIP Packet End-to-End Delay

- CDF of voice packet delay
- Largest delay variation is 180us
- Voice packet transit delays at most one low priority packet size
CallerGroup4 VoIP Performance

- CDF of voice packet delay
- Largest delay variation is 230us
Low Priority Traffic Performance

- Low priority traffic conforms to fair transmission rate
- Severe delay increase for excessive low priority traffic
Simulation Two
Unevenly Distributed TCP/FTP Traffics

- Link propagation delay 200us (40km)
- 15 nodes aggregation, (total 34 nodes), routing node ip forwarding speed is 320kpps
- FTP clients traffic aggregation to a common FTP server at node_0
- There are 40~160 simultaneous ftp sources in each 1000Base_X LAN
- SRP Configuration:
  - LP transit buffer 512Kbytes
  - LP transmit buffer 512Kbytes
  - LP Tb low threshold 128Kbytes
  - LP Tb high threshold 500Kbytes
  - Max_allow 32000
Simulation Runs

- Unevenly distributed TCP/FTP traffic aggregates along outer ring with the same traffic source profile as FTP in previous simulation.

- There are three simulation runs:
  - First Run: link utilization 84%,
    total traffic 521.6Mbps
  - Second Run: link utilization 95%,
    total traffic 587.2Mbps
  - Third Run: link utilization: > 100%,
    total traffic > 622Mbps

- Fair rate for the large sources is about 47Mbps
Traffic Aggregation on the Ring

- When overloaded, ring bandwidth is 100% utilized.
Fair and consistent TCP delay performance as more traffic aggregates and the ring is oversubscribed.

Guaranteed TCP delay performance for conforming traffics.
TCP Performance for Non-Conforming Traffic

- Severe performance degradation for the excessive rate in non-conforming TCP traffic when ring is oversubscribed.
- Stable and good delay performance for >90% of the TCP traffic
When the ring is not oversubscribed
- Fair and consistent TCP delay performance for all nodes.
- Fair ring bandwidth access for all nodes.
As more traffic aggregates and the ring is not oversubscribed
  – Fair and proportional TCP delay increase for all nodes
  – The largest traffic gets the largest delay increase
  – Smoothed ring access rate for TCP applications
• As more traffic aggregates and the ring is oversubscribed
  – Fair and proportional TCP delay increase for conforming traffics
  – Very large and severe TCP delay increase for non-conforming traffics
  – Large TCP source nodes are throttled to fair ring access rate
Summary

- SRP-fa is scalable to large and high bandwidth rings for metro, regional and wide area networks.
- SRP-fa provides excellent support for TCP applications by ensuring
  - fair and stable ring access rate
  - stable and consistent end-to-end delay performance for all conforming tcp traffics.
  - only the non-conforming tcp traffic suffers significant performance degradation.
- For high priority traffic, regardless of low priority traffic, SRP-fa guarantees
  - its bandwidth requirement and ring access rate
  - a predictable packet end-to-end delay and jitter performance
Appendix

- Appendix 1  SRP Overview
- Appendix 2a. SRP-fa Rate Counters
- Appendix 2b. SRP-fa Feedback Usage Generation
Appendix 1 SRP Overview

• Spatial Reuse Protocol (SRP) is the new media access control protocol for bi-directional dual counter rotating ring
  – media independent
  – utilize both rings to transport data and control packets
  – support Intelligent Protection Switching (IPS) for ring protection and restoration
  – support plug and play operation

• Enable spatial reuse by destination stripping
  – allow multiple nodes transmitting simultaneously
  – bandwidth consumed only on traversed ring segment
  – Unicast packets travels along ring spans between the src and dest nodes only

• SRP fairness algorithm (SRP-fa) controls access to the ring and enforce fairness

• Scalable to large number of nodes on the ring
Appendix 2a
SRP-fa Rate Counters

- **Transmit Rate Counter**: My_usage
  - Incremented when transmitting low priority transmit packets
    \[
    \text{My usage} = \text{My usage} + \text{Packet Len}
    \]
  - Decremented by a fixed fraction at decay interval
    \[
    \text{My usage} = \text{My usage} - \min\left(\frac{\text{allow usage}}{\text{AGECOEFF}}, \frac{\text{my usage}}{\text{AGECOEFF}}\right)
    \]

- **Threshold Counter**: Allow_usage and Max_allow
  - Allow_usage set to feedback usage from downstream neighbours
  - Allow_usage can decay upwards to Max_allow if Null usage is received
    \[
    \text{allow usage} += \frac{(\text{MAX LRATE} - \text{allow usage})}{(\text{LP ALLOW})}
    \]
  - Max_allow is statically pre-configured.

- **Transit Rate Counter**: Fwd_rate
  - Incremented when transmitting low priority transit packets
    \[
    \text{Fwd rate} = \text{Fwd rate} + \text{Packet Len}
    \]
  - Decremented by a fixed fraction at decay interval
    \[
    \text{fwd rate} = \text{fwd rate} - \frac{\text{fwd rate}}{\text{AGECOEFF}}
    \]
Appendix 2b
SRP-fa Feedback Usage Generation

- LP TB congestion status
  
  congested = (lo_tb_depth > TB_LO_THRESHOLD/2)

- If congested, signal the smallest usage to throttle upstream transmit
  
  if (lp_my_usage < rcvd_usage)
    upstream_usage = lp_my_usage;
  else
    upstream_usage = rcvd_usage;

- If not congested but some downstream node is congested which is caused by upstream node, pass on received usage to throttle upstream
  
  if ((rcvd_usage != NULL) && (lp_fwd_rate > allow_usage)
    upstream_usage = rcvd_usage;

- Otherwise, signal null usage to upstream nodes
  
  upstream_usage = NULL
  if (upstream_usage > MAX_LRATE)
    upstream_usage = NULL