High Speed Networks Need Proactive Congestion Control

Stanford University

Lavanya Jose, Lisa Yan, Stephen Ibanez, Isaac Keslassy, George Varghese, Sachin Katti, Mohammad Alizadeh, and Nick McKeown

Stanford University, ~Technion, ♠Microsoft Research, ⋆Massachusetts Institute of Technology

Fraction of Total Flows in Bing Workload

- Small (1-10KB): 14%
- Medium (10KB-1MB): 50%
- Large (1MB-10MB): 36%

Typical Flow Completion Times

<table>
<thead>
<tr>
<th>Throughput</th>
<th>Median</th>
<th>Tail (99%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 Gbps</td>
<td>10-40 RTTs</td>
<td>7-8 RTTs</td>
</tr>
<tr>
<td>40 Gbps</td>
<td>17-20 RTTs</td>
<td>10 RTTs</td>
</tr>
<tr>
<td>100 Gbps</td>
<td>7-8 RTTs</td>
<td>10 RTTs</td>
</tr>
</tbody>
</table>

Reactive techniques find rates too slowly

Why reactive techniques are slow

- No use of explicit information, can only react and move in right direction

Can we use explicit info (Link Capacities, Traffic Matrix) to schedule flows faster?

Centralized

Fastpass: A centralized, packet scheduling scheme

Distributed

PERC: A distributed, rate allocation scheme

Arbiter schedules every packet’s departure time and path

Links communicate with flows to compute appropriate rates

Can we use explicit info (Link Capacities, Traffic Matrix) to schedule flows faster?

Centralized

Fastpass: A centralized, packet scheduling scheme

Distributed

PERC: A distributed, rate allocation scheme

Arbiter schedules every packet’s departure time and path

Links communicate with flows to compute appropriate rates

Looking Forward

Using Programmable Data Planes

Implementation of distributed proactive congestion control schemes using programmable data planes

New Applications

New applications will be enabled by high-bandwidth, low-latency data-centers:
- Large-scale machine learning applications
- Massive virtual worlds
- Disaggregated flash in data centers

Making PERC practical

Goals

1. Robust to:
   - control packet losses
   - errors in measuring rates, calculations, control messages
2. Simple computation in switches
3. Minimal state/state access per packet

References


