Hybrid network traffic engineering system (HNTES)

Zhenzhen Yan, Zhengyang Liu, Chris Tracy, Malathi Veeraraghavan
University of Virginia and ESnet
Jan 12-13, 2012
mvee@virginia.edu, ctracy@es.net

Project web site: http://www.ece.virginia.edu/mv/research/DOE09/index.html

Thanks to the US DOE ASCR program office and NSF for UVA grants DE-SC002350, DE-SC0007341, OCI-1127340 and ESnet grant DE-AC02-05CH11231
Problem statement

• A hybrid network supports both IP-routed and circuit services on:
  - Separate networks as in ESnet4, or
  - An integrated network as in ESnet5

• A hybrid network traffic engineering system (HNTES) is one that moves science data flows to circuits

• Problem statement: Design HNTES
Two reasons for using circuits

1. Offer scientists rate-guaranteed connectivity
2. Isolate science flows from general-purpose flows

<table>
<thead>
<tr>
<th>Circuit scope</th>
<th>Reason</th>
<th>Rate-guaranteed connections</th>
<th>Science flow isolation</th>
</tr>
</thead>
<tbody>
<tr>
<td>End-to-end (inter-domain)</td>
<td>✔</td>
<td>✖</td>
<td>✗</td>
</tr>
<tr>
<td>Per provider (intra-domain)</td>
<td>✗</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

Request to sites:
• Any information on trouble tickets created by science flows would be appreciated
What type of flows should be isolated?

- Dimensions
  - size (bytes): elephant and mice
  - rate: cheetah and snail
  - duration: tortoise and dragonfly
  - burstiness: porcupine and stingray

alpha flows

- number of bytes in any T-sec interval $\geq H$ bytes
  - if $H = 1$ GB and $T = 60$ sec
  - throughput exceeds 133 Mbps
- alpha flows responsible for burstiness
- alpha flows are caused by transfers of large files over fast links
  - Let’s look at GridFTP usage statistics

GridFTP log analysis

• Two goals:
  - Determine durations of high-throughput GridFTP transfers
    • to use dynamic circuits, since current IDC circuit setup delay is ~1 min, need transfer durations to be say 10 mins
  - Characterize variance in throughput
    • identify causes
GridFTP data analysis findings

- GridFTP transfers from NERSC dtn servers that > 100 MB in one month (Sept. 2010)
- Total number of transfers: 124236
- GridFTP usage statistics

Thanks to Brent Draney, Jing Tie and Ian Foster for the GridFTP data
Top quartile highest-throughput transfers
NERSC (100MB dataset)

<table>
<thead>
<tr>
<th>Throughput (Mb/s)</th>
<th>Min</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>444.5</td>
<td>483.0</td>
<td>596.3</td>
<td>698.8</td>
<td>791.9</td>
<td>4315</td>
</tr>
</tbody>
</table>

- Total number: 31059 transfers
- 50% of this set had duration < 1.51 sec
- 75% had duration < 1.8 sec
- 95% had duration < 3.36 sec
- 99.3% had duration < 1 min
- 169 (0.0054%) transfers had duration > 2 mins
- Only 1 transfer had duration > 10 mins

Need to look for multi-transfer sessions
Throughput variance

There were 145 file transfers of size 32 GB to same client
  - Same round-trip time (RTT), bottleneck link rate and packet loss rate
  - IQR (Inter-quartile range) measure of variance is 695 Mbps
  - Need to find an explanation for this variance
Potential causes of throughput variance

• Path characteristics:
  - RTT, bottleneck link rate, packet loss rate
  - Usage stats do not record remote IP address
  - Can extract from NetFlow data for alpha flows

• Number of stripes

• Number of parallel TCP streams

• Time-of-day dependence

• Concurrent GridFTP transfers

• Network link utilization (SNMP data)

• CPU usage, I/O usage on servers at the two ends
Time-of-day dependence (NERSC 32 GB: same path)

- Two sets of transfers: 2 AM and 8 AM
- Higher throughput levels on some 2 AM transfers
- But variance even among same time-of-day flows
Dep. on concurrent transfers:
Predicted throughput

- Find number of concurrent transfers from GridFTP logs for $i^{th}$ 32 GB GridFTP transfer: NERSC end only
- Determine predicted throughput
- $d_{ij}$: duration of $j^{th}$ interval of $i^{th}$ transfer
- $n_{ij}$: number of concurrent transfers in $j^{th}$ interval of $i^{th}$ transfer

$$\tilde{T}_i = T_{max} \sum_{j=1}^{j_{max}} \frac{1}{n_{ij}} \times \frac{d_{ij}}{D_i}$$
Dependence on concurrent transfers (NERSC 32 GB transfers)

Correlation seen for some transfers
But overall correlation low (0.03)
expl: Other apps besides GridFTP
Correlation with SNMP data

- SNMP raw byte counts: 30 sec polling
- Assume GridFTP bytes uniformly distributed over duration
- Ordered GridFTP transfers by throughput
- Conclusion: GridFTP bytes dominate and are not affected by other transfers – consistent with alpha behavior

<table>
<thead>
<tr>
<th>1st Qu.</th>
<th>if1</th>
<th>if2</th>
<th>if3</th>
<th>if4</th>
<th>if5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.677</td>
<td>0.604</td>
<td>0.719</td>
<td>0.750</td>
<td>0.749</td>
</tr>
<tr>
<td>2nd Qu.</td>
<td>0.419</td>
<td>0.147</td>
<td>0.138</td>
<td>0.327</td>
<td>0.294</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>0.538</td>
<td>0.592</td>
<td>0.543</td>
<td>0.415</td>
<td>0.371</td>
</tr>
<tr>
<td>4th Qu.</td>
<td>0.782</td>
<td>0.872</td>
<td>0.797</td>
<td>0.789</td>
<td>0.790</td>
</tr>
<tr>
<td>All</td>
<td>0.902</td>
<td>0.922</td>
<td>0.919</td>
<td>0.918</td>
<td>0.918</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1st Qu.</th>
<th>if1</th>
<th>if2</th>
<th>if3</th>
<th>if4</th>
<th>if5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.254</td>
<td>0.188</td>
<td>0.429</td>
<td>0.505</td>
<td>0.486</td>
</tr>
<tr>
<td>2nd Qu.</td>
<td>0.269</td>
<td>-0.067</td>
<td>-0.110</td>
<td>0.089</td>
<td>0.071</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>0.059</td>
<td>0.157</td>
<td>0.110</td>
<td>0.015</td>
<td>-0.039</td>
</tr>
<tr>
<td>4th Qu.</td>
<td>0.196</td>
<td>0.328</td>
<td>0.239</td>
<td>0.287</td>
<td>0.276</td>
</tr>
<tr>
<td>All</td>
<td>0.351</td>
<td>0.365</td>
<td>0.443</td>
<td>0.524</td>
<td>0.527</td>
</tr>
</tbody>
</table>

Thanks to Jon Dugan for the SNMP data
Request from sites

- Permission to view GridFTP usage statistics
- Performance monitoring of DTN servers
  - File system usage
  - CPU usage
- MRTG data from site internal links
- Trouble ticket information
Back to HNTES: Role
Usage within domains for science flow isolation

- Ingress routers would be configured by HNTES to move science flows to MPLS LSPs
Three tasks executed by HNTES

1. alpha flow identification
   - Offline flow analysis
   - Online flow analysis
   - End-host assisted

2. Circuit Provisioning
   - Rate-unlimited MPLS LSPs initiated offline
   - Rate-unlimited MPLS LSPs initiated online
   - Rate-specified MPLS LSPs initiated online

3. Policy Based Route (PBR) configuration at ingress/egress routers
   - Set offline
   - Set online

online: upon flow arrival
offline: periodic process (e.g., every hour or every day)
Questions for HNTES design

• Online or offline?
• PBRs: 5-tuple identifiers or just src/dst addresses?
• /24 or /32?
• How should PBR table entries be aged out?
NetFlow data analysis

• NetFlow data over 7 months (May-Nov 2011) collected at ESnet site PE router
• Three steps
  - UVA wrote R analysis and anonymization programs
  - ESnet executed on NetFlow data
  - Joint analysis of results
Flow identification algorithm

• alpha flows: high rate flows
  - NetFlow reports: subset where bytes sent in 1 minute > H bytes (1 GB)
  - Raw IP flows: 5 tuple based aggregation of reports on a daily basis
  - Prefix flows: /32 and /24 src/dst IP
  - Super-prefix flows: (ingress, egress) router based aggregation of prefix flows

• 7-month data set
  - 22041 raw IP flows, 125 (/24) prefix flows, and 1548 (/32) prefix flows
Flow aggregation from NetFlow

- **NetFlow report set**
- **Raw IP flow set**
- **Prefix flow set**

- Length represents #bytes count
- The leftmost color represents src and dst IP/subnet
- The second to the leftmost color represents src, dst port and prot

\[ \alpha \text{-interval (} t_1 \text{)} = 1 \text{ min} \]
\[ \text{aggregation interval (} t_2 \text{)} = 1 \text{ day} \]
Online vs. offline

- 89.84% α-flows are less than 2 min, virtual circuit setup delay is 1 min
- 0.99% of the flows are longer than 10 minutes, but same ID for long and short flows (how then to predict)

Histogram of α-flows with duration < 4.5 mins (0-95\textsuperscript{th} percentile)
Raw IP flow vs. prefix flow

- Port numbers are ephemeral for most high-speed file transfer applications, such as GridFTP
  - Answer to Q: Use prefix flow IDs
- Hypothesis:
  - Computing systems that run the high-speed file transfer applications don’t change their IP addresses and/or subnet IDs often
  - Flows with previously unseen prefix flow identifiers will appear but such occurrences will be relatively rare
When new collaborations start or new data transfer nodes are brought online, new prefix flows will occur.
Effectiveness of offline design

TABLE II: Number of days during which the percentages of α-bytes, and number of raw IP flows, that are not redirected to MPLS LSPs exceed different thresholds. The total number of days is 214.

<table>
<thead>
<tr>
<th>Aging parameter</th>
<th>Three measures</th>
<th>( \geq 100% )</th>
<th>( \geq 75% )</th>
<th>( \geq 50% )</th>
<th>( \geq 25% )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>/24</td>
<td>/32</td>
<td>/24</td>
<td>/32</td>
<td>/24</td>
</tr>
<tr>
<td>7 days</td>
<td>α-bytes</td>
<td>2</td>
<td>15</td>
<td>20</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Number of raw IP flows</td>
<td>2</td>
<td>15</td>
<td>17</td>
<td>76</td>
</tr>
<tr>
<td>14 days</td>
<td>α-bytes</td>
<td>2</td>
<td>8</td>
<td>12</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>Number of raw IP flows</td>
<td>2</td>
<td>8</td>
<td>10</td>
<td>51</td>
</tr>
<tr>
<td>30 days</td>
<td>α-bytes</td>
<td>1</td>
<td>5</td>
<td>8</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Number of raw IP flows</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>35</td>
</tr>
<tr>
<td>∞ days</td>
<td>α-bytes</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Number of raw IP flows</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>23</td>
</tr>
</tbody>
</table>

- 94.4% of the days, at least 50% of the alpha bytes would have been redirected.
- For 89.7% of the days, 75% of the alpha bytes would have redirected (aging parameter = never; prefix identifier is /24)
Effect of aging parameter on PBR table size

- For operational reasons, and forwarding latency, this table should be kept small.

![Graph showing the effect of aging parameter on PBR table size.](image-url)
92% of the alpha bytes received over the 7-month period would have been redirected (aging parameter = never; prefix identifier is /24)
Key points for HNTES 2.0 design

• From current analysis:
  - Offline design appears to be feasible
    • IP addresses of sources that generate alpha flows relatively stable
    • Most alpha bytes would have been redirected in the analyzed data set
  - /24 seems better option than /32

• Aging parameter:
  - 30 days: tradeoff PBR size with effectiveness
Future NetFlow data analyses

- other routers’ NetFlow data
- redirected beta flow bytes experience competition with alpha flows (/24)
- utilization of MPLS LSPs
- multiple simultaneous alpha flows on same LSPs
- match with known data doors
Discussion

• To determine cause of throughput variance
  - Feedback?
  - Need your support to obtain data
• Would trouble ticket log mining be useful to help answer “why isolate science flows”?
• Automatic flow identification and redirection appears feasible
  - How do you feel about this?