Analyzing Peer-To-Peer Traffic Across Large Networks

From: IEEE/ACM TRANSACTIONS ON NETWORKING,
VOL. 12, NO. 2, APRIL 2004

Presented by 江文德
Outline

- Introduction
- Methodology
- Characterization Metrics
- Overview of P2P Traffic and System Dynamics
- Traffic Characterization
- Conclusions and Future Work
Introduction

- Two Categories of P2P Traffic
  - signaling
  - data transfer
- Previous Projects Almost Focused on P2P Signaling Traffic
  - They gather traffic by setting up P2P crawler.
Introduction (Cont.)

- Research Questions for P2P System Design and Traffic Engineering
  - How is the P2P traffic distributed across the Internet?
  - What are the characteristics of the application-level P2P network connectivity?
  - How dynamic are the P2P system, both temporally and spatially?
Methodology

- Popular P2P Applications
  - FastTrack, Gnutella, and DirectConnect

- Measurement Approach
  - Offline analysis of flow-level data gathered from multiple routers across a large ISP’s backbone
  - IP address, network prefix, autonomous system (AS)
  - Cisco’s NetFlow service
Methodology (Cont.)

- Advantages
  - It doesn’t require knowledge about the P2P protocol.
  - It approach is non-intrusive and all the traffic data can be collected.
  - It gathers information on both the signaling traffic and the data download traffic.
  - It is conducive to determining the impact of P2P traffic on certain regions of the network.
Methodology (Cont.)

- Limitations
  - We are not able to obtain application-level details.
  - We may not capture the complete flow of traffic.
Characterization Metrics

- Host Distribution
- Traffic Volume
- Host Connectivity
Characterization Metrics (Cont.)

- Traffic Pattern Over Time

Fig. 1. The binning of netflow records.

\[
\text{Volume}(b_l) = \sum_{i=1}^{n} \text{Volume}(s_{il}).
\]
Characterization Metrics (Cont.)

- **Connection Duration and On-Time**
  - $f_i$ and $f_j$ are concurrent iff
    \[
    \text{StartTime}(f_j) \leq \text{FinishTime}(f_i) + \delta
    \]
    where $i < j$, $f_i$ and $f_j$ are two flows

- **Mean Bandwidth Usage**
  \[
  \text{Bandwidth}_R(h) = \frac{\text{Volume}_R(h)}{\text{OnTime}_R(h)}
  \]
  where $h$ is a given host
Overview of P2P Traffic and System Dynamics

- Summary Statistics for the P2P Data Set

<table>
<thead>
<tr>
<th>Date</th>
<th>Protocol</th>
<th>Number of records</th>
<th>Total number of unique IPs</th>
<th>Number of unique IPs per day</th>
<th>Total traffic volume (GBytes/day)</th>
<th>Traffic volume per IP (MBytes/day)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/10/2001 - 9/15/2001</td>
<td>Gnutella</td>
<td>37,853,281</td>
<td>718,464</td>
<td>197,445</td>
<td>211</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>FastTrack</td>
<td>110,533,024</td>
<td>3,403,900</td>
<td>998,669</td>
<td>773</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>DirectConnect</td>
<td>595,606</td>
<td>22,852</td>
<td>6,244</td>
<td>48</td>
<td>15.4</td>
</tr>
<tr>
<td>10/9/2001 - 10/13/2001</td>
<td>Gnutella</td>
<td>49,649,348</td>
<td>823,532</td>
<td>247,114</td>
<td>272</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>FastTrack</td>
<td>184,113,038</td>
<td>4,450,149</td>
<td>1,485,370</td>
<td>1,153</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>DirectConnect</td>
<td>566,740</td>
<td>23,211</td>
<td>7,193</td>
<td>56</td>
<td>15.6</td>
</tr>
<tr>
<td>12/10/2001 - 12/16/2001</td>
<td>Gnutella</td>
<td>69,578,723</td>
<td>887,520</td>
<td>236,954</td>
<td>242</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>FastTrack</td>
<td>340,690,074</td>
<td>5,924,072</td>
<td>1,934,460</td>
<td>1,776</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>DirectConnect</td>
<td>701,712</td>
<td>29,925</td>
<td>7,213</td>
<td>71</td>
<td>19.6</td>
</tr>
</tbody>
</table>
Overview of P2P Traffic and System Dynamics (Cont.)

- Host Distribution

![Graphs showing host distribution for FastTrack, Gnutella, DirectConnect, IP, Prefix, and AS.]
Overview of P2P Traffic and System Dynamics (Cont.)

- Traffic Volume Distribution

![Graphs showing traffic volume distribution for different P2P networks: FastTrack, Gnutella, DirectConnect.](image-url)
Overview of P2P Traffic and System Dynamics (Cont.)

- Host Connectivity
Overview of P2P Traffic and System Dynamics (Cont.)

- Traffic Pattern Over Time

![Traffic volume (MBytes) over time](image)

![Number of unique IPs, Prefixes, and ASes over time](image)
Overview of P2P Traffic and System Dynamics (Cont.)

- Host Connection Duration and On-Time
Overview of P2P Traffic and System Dynamics (Cont.)

- Host Connection Duration and On-Time

![Graphs showing distribution of IP connections, on-time, and number of connections per day.](image-url)
Overview of P2P Traffic and System Dynamics (Cont.)

- Mean Bandwidth Usage for Hosts

![Graphs showing cumulative distribution of IPs](image-url)
Zipf’s Law
- The rank-frequency plot is the plot of the occurrence frequency $f_r$ of each object versus its rank $r$, in log-log scales.
- The rank frequency version of Zipf’s law states that $f_r \propto 1/r$.
- The generalized Zipf distribution is defined as $f_r \propto 1/r^\theta$, where the slope $-\theta$ is log-log scale can be different than -1.
Traffic Characterization

(a) Unique number of IPs contacted

(b) Unique number of IPs contacted

(c) Traffic volume (MB/yes)

(d) On-time (Minutes)
Traffic Characterization (Cont.)

- Relationships Between Measures

\[ \text{(a) \# On Time - Upstream (sec)} \]
\[ \text{(b) \# Unique IP - Upstream) \]
\[ \text{(c) \# Bw - Upstream (KBytes/s)} \]
Traffic Characterization (Cont.)

- Relationships Between Measures
Traffic Characterization (Cont.)

- P2P Traffic Versus Web Traffic
Conclusions

- Less than 10% of the IPs contribute around 90% of the total traffic volume.
- The P2P traffic distributions for traffic volume, connectivity, on-time, and average bandwidth usage are extremely skewed; but they don’t fit well with power law distribution.
Conclusions (Cont.)

- All three P2P systems exhibit a high level of system dynamics.
- The fraction of P2P traffic contributed by each network prefix remains relatively unchanged more stable than the distribution of either Web traffic or overall traffic over the time period of one month.
Future Work

- Determine accurate estimates of the peak bandwidth usage
- Develop accurate models for the distributions of the traffic metrics