Measurement and Analysis of Traffic in a Hybrid Satellite-Terrestrial Network

Qing (Kenny) Shao and Ljiljana Trajkovic
{qshao, ljilja}@cs.sfu.ca

Communication Networks Laboratory
http://www.ensc.sfu.ca/cnl
School of Engineering Science
Simon Fraser University, Vancouver, Canada
Road map

- Introduction and motivation
- Traffic:
  - collection
  - analysis
  - prediction
- Conclusions
- References
Network traffic measurements

- Focus of networking research during:
  - mid to late 1980’s
  - early 1990’s
- Motivation for traffic measurements:
  - understand traffic characteristics in deployed networks
  - develop traffic models
  - evaluate performance of protocols and applications
  - perform trace driven simulations
Traffic traces

- Most available traffic traces are from the wired networks within research communities:
  - Bellcore, LBNL, Auckland University
- Few traces were collected from wireless or satellite commercial networks
- Various factors affect Internet traffic patterns:
  - Web, Proxy, Napster, MP3, Web mail
- Used to evaluate the AutoRegressive Integrated Moving-Average (ARIMA) model for predicting uploaded and downloaded traffic
DirecPC system

- Satellite one-way broadcast system manufactured by Hughes Network Systems
- DirecPC systems are deployed worldwide
- ChinaSat uses DirecPC system to provide Internet access to over 200 Internet cafés across provinces
- DirecPC utilizes two special techniques to improve network performance:
  - IP spoofing
  - TCP splitting
DirecPC system

- **IP spoofing:**
  - customer’s requests are not directly sent to the website
  - they are rerouted to the satellite network operation center (NOC)
  - NOC resends the request to the website
  - website sends to the NOC data to be downloaded

- **TCP splitting:**
  - terrestrial links use standard TCP
  - to improve throughput, space links with long delay use modified TCP versions with enlarged TCP window size
Traffic collection

Red: uploaded traffic
Green: downloaded traffic
tcpdump trace format

- timestamp src > dst: flags data-seqno ack window urgent options
  - 19:12:45.660701 61.159.59.162.12800 > 192.168.1.169.62246: udp 52
    6541284:6541321(37) ack 1479344110 win 8192 (DF)
    win 8192
  - 19:12:45.676255 61.152.249.71.55901 > 192.168.1.242.40770: P
    2627573783:2627573791(8) ack 5795719 win 63343 (DF)
    2775973525:2775973533(8) ack 11622145 win 64102 (DF)
  - 19:12:45.689095 192.168.1.169.63644 > 202.103.69.103.3010: P
    1969195:1969259(64) ack 2995916216 win 8192 (DF)
  - 19:12:45.692475 202.101.165.134.80 > 192.168.2.3.45585: . ack 3153903 win 6432
  - 19:12:45.699193 207.46.104.20.80 > 192.168.1.239.4912: R
    2405276149:2405276149(0) win 0

- **Red**: uploaded traffic
- **Green**: downloaded traffic
Analysis of weekly billing records

Weekly traffic volume measured in packets (left) and bytes (right)
Traffic data was collected from 09-12-2002 to 15-12-2002
Analysis of daily billing records

- Average traffic volume over a single day measured in packets (left) and bytes (right)
- Traffic data was collected from 9-12-2002 to 15-12-2002
Protocols and applications

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Packets</th>
<th>Packets (%)</th>
<th>Bytes</th>
<th>Bytes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP</td>
<td>36,737,165</td>
<td>84.32</td>
<td>11,231,147,530</td>
<td>94.49</td>
</tr>
<tr>
<td>UDP</td>
<td>6,202,673</td>
<td>14.24</td>
<td>601,157,016</td>
<td>5.06</td>
</tr>
<tr>
<td>ICMP</td>
<td>630,528</td>
<td>1.45</td>
<td>53,128,377</td>
<td>0.45</td>
</tr>
<tr>
<td>Total</td>
<td>43,570,366</td>
<td>~100</td>
<td>11,885,432,923</td>
<td>~100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Applications</th>
<th>Connections</th>
<th>Connections (%)</th>
<th>Bytes</th>
<th>Bytes (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW</td>
<td>304,243</td>
<td>90.06</td>
<td>10,203,267,005</td>
<td>75.79</td>
</tr>
<tr>
<td>FTP-data</td>
<td>636</td>
<td>0.19</td>
<td>1,440,393,008</td>
<td>10.7</td>
</tr>
<tr>
<td>IRC</td>
<td>2,324</td>
<td>0.69</td>
<td>945,965</td>
<td>0.008</td>
</tr>
<tr>
<td>SMTP</td>
<td>562</td>
<td>0.17</td>
<td>2,326,373</td>
<td>0.01</td>
</tr>
<tr>
<td>POP-3</td>
<td>115</td>
<td>0.03</td>
<td>2,326,373</td>
<td>0.02</td>
</tr>
<tr>
<td>Telnet</td>
<td>70</td>
<td>0.02</td>
<td>280,286</td>
<td>0.002</td>
</tr>
<tr>
<td>Other</td>
<td>651</td>
<td>8.84</td>
<td>238,099,412</td>
<td>13.47</td>
</tr>
<tr>
<td>Total</td>
<td>308,601</td>
<td>100</td>
<td>11,885,432,923</td>
<td>100</td>
</tr>
</tbody>
</table>

- Traffic data was collected from 21-12-2002 22:08 to 23-12-2002 3:28
TCP connection level: Web traffic

- Zipf-like distribution: \( f_r \sim 1/r^\beta \)
  
  the number of requests (frequency) is inversely proportional to its rank among the requests

- DGX (discrete lognormal):
  
  \[
  p(x = k) = \frac{A(\mu, \sigma)}{k} \exp\left[-\frac{(\ln k - \mu)^2}{2\sigma^2}\right]
  \]

  \[
  A(\mu, \sigma) = \sum_{k=1}^{\infty} \frac{1}{k} \left[-\frac{(\ln k - \mu)^2}{2\sigma^2}\right]^{-1}
  \]

- DGX distribution fits better than the Zipf-like distribution
TCP connection level: Web traffic

- Traffic is non-uniformly distributed among the Internet hosts
- Ten busiest websites account for 60.23 % of the entire traffic load:
  - all registered under the Asia Pacific Network Information Centre
  - the most popular site: a Chinese search engine website
- Language, geographical, and commercial factors (popular sites) greatly affect the traffic distribution
- Important for designing content delivery networks and caching proxies
TCP packet size

Traffic data was collected from 21-12-2002 22:08 to 23-12-2002 3:28

- Packet size distribution is bimodal:
  - 50 % of packets are less than 200 bytes
  - 30 % of packets are greater than 1,400 bytes
- Most bytes are transferred in large packets
Self-similarity

- Self-similarity implies a “fractal-like” behavior: data on various time scales have similar patterns.
- A wide-sense stationary process $X(n)$ is called (exactly second order) **self-similar** if:
  - $r^{(m)}(k) = r(k), \ k \geq 0, \ m = 1, 2, \ldots, n$
- Implications:
  - no natural length of bursts
  - bursts exist across many time scales
  - traffic does not become “smoother” when aggregated (unlike Poisson traffic)
Estimation of self-similarity
Self-similar processes

- Properties:
  - slow decaying variance
  - long-range dependence
  - Hurst parameter

- Processes with only short-range dependence (Poisson): \( H = 0.5 \)

- Self-similar processes: \( 0.5 < H < 1.0 \)

- As the traffic volume increases, the traffic becomes more bursty, more self-similar, and the Hurst parameter increases
Estimation of self-similarity

\[ H = \text{slope} \]

\[ H = 1 + \frac{\text{slope}}{2} \]

\[ H = \frac{1 - \text{slope}}{2} \]

\[ H = \frac{1 + \text{slope}}{2} \]
Estimation of self-similarity

Traffic data was collected on 09-12-2002
Modeling self-similar processes

- Self-similar process can be generated by aggregating multiple ON/OFF sources
- The ON/OFF periods are heavy-tailed distributed with infinite variance
- Web and ftp file sizes are heavy-tailed
- A probability distribution $X$ is heavy-tailed if:

$$P[X > x] \sim cx^{-\alpha}, \ 0 < \alpha < 2, \ x \to \infty$$

TCP connection model

- We consider two parameters of a TCP connection:
  - connection inter-arrival times
  - number of downloaded bytes per connection
- Four probability distributions:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Probability density</th>
<th>Cumulative probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$f(x) = \frac{1}{\rho} e^{-x/\rho}$</td>
<td>$F(x) = 1 - e^{-x/\rho}$</td>
</tr>
<tr>
<td>Weibull</td>
<td>$f(x) = \frac{1}{a} \left( -\frac{x}{a} \right)^{c-1} e^{-\left(x/a\right)^c}$</td>
<td>$F(x) = 1 - e^{-\left(x/a\right)^c}$</td>
</tr>
<tr>
<td>Pareto</td>
<td>$f(x) = \frac{ak^a}{(x)^{k+1}}$</td>
<td>$F(x) = 1 - \left( \frac{k}{x} \right)^a$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No closed form</td>
</tr>
</tbody>
</table>
TCP connection model

- **Best fit:**
  - Lognormal: downloaded bytes per TCP connection
  - Weibull: inter-arrival times of TCP connections
Traffic prediction

- “Time series analysis - forecasting and control”
  - G. E. P. Box and G. M. Jenkins (1976)
- **AutoRegressive Integrated Moving-Average (ARIMA):**

  \[ X(t) = \phi_1 X(t-1) + \cdots + \phi_p X(t-p) + e(t) + \theta_1 e(t-1) + \cdots + \theta_q e(t-q) \]

  \[ (p, d, q) \times (P, D, Q)_s \]

- past values
  - **AutoRegressive (AR) structure**
- past random fluctuant effect
  - **Moving Average (MA) process**
One week ahead prediction

- We applied Box-Jenkins method to six weeks of billing records
- Derived parameters:
  - $d=0$, $D=1$, $s=168$, $p=1$, $q=0$, $P=0$, $Q=1$
  - collected records fit the model $(1,0,0) \times (0,1,1)_1^{168}$
- Normalized mean squared error (nmse) is used to measure the performance of the predictor:

$$nmse = \frac{1}{\sigma^2 N} \sum_{k=1}^{N} (x(k) - \bar{x}(k))^2$$
Predictability evaluation

Predicting downloaded traffic is more difficult than predicting uploaded traffic.

<table>
<thead>
<tr>
<th>Traffic type</th>
<th>Uploaded traffic</th>
<th>Downloaded traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>nmse</td>
<td>0.3653</td>
<td>0.5988</td>
</tr>
</tbody>
</table>
Conclusions

- Analysis of collected traffic data:
  - Web applications and TCP protocol dominate the collected traffic
  - packet size distribution is bimodal: most bytes are transferred in big packets
  - few Web servers account for majority of data traffic
  - the frequency-rank relation of client connections matches the discrete lognormal distribution
  - various estimators of the Hurst parameter produced inconsistent results
  - more accurate estimation was achieved with the wavelet estimator
Conclusions

- **TCP modeling:**
  - **Weibull:** inter-arrival times of TCP connections
  - **Lognormal:** downloaded bytes per TCP connection

- Traffic prediction using the ARIMA model:
  - performs better for predicting the **uploaded** traffic
  - not suitable for predicting **downloaded** traffic
References


References


