QoS & Transient Simulations of Web Traffic:

Using Quantiles to Characterize User-Perceived Latency in Simulations with Heavy-Tailed Input

Dagstuhl, October 30st, 2002
by Ulrich Fiedler, TIK, ETH-Zurich
Motivation

• Simulations of web traffic are deployed to investigate numerous problems
  • Important performance metrics
    – Server throughput
    – User-perceived latency of downloads
      -> user-centered QoS provisioning
  • Self-similarity ⇒ negative impact on performance
    (Barford, Crovella 1998)
  • Self-similarity ⇐ input: heavy-tailed object size distribution
    – Simulations remain transient during reasonable times
      » Average object size, average latency do not converge
Problem

• Take end-user‘s perspective in client server scenario
• User-perceived latency is sum of latencies of network, server/cache, client
• Latency quantiles (or percentiles)
  • have a natural interpretation
  • do not depend on moments of the distribution
• Are latency quantiles suitable statistics for performance evaluation?
  • Do latency quantiles converge in reasonable times?
Outline

• Web workload modeling
  • Heavy-tailed distribution to model self-similarity, implications of heavy-tailed distributions
• Convergence of simulation input
  • Object size quantiles
• Convergence of simulation output
  • Latency quantiles
• Discussion
Web Workload Modeling I

• **Def.: heavy-tailed distribution**

\[ 1 - F(x) \sim x^{-\alpha} \quad x \to \infty \]

  – Line in log-log representation
  – Infinite variance for shape parameter \(1 < \alpha < 2\)
  – Simplest class of representants: Pareto distributions

\[ F(x) = 1 - \left( \frac{k}{x} \right)^\alpha \quad x \in [k, \infty[ \]
Web Workload Modeling II

- Heavy-tails in object size or think time distribution cause self similarity on the network level
  - On/off model (Willinger 1995) (Likhanov 1995)
  - Effects caused by object sizes dominate effects caused by think time (Park, Kim, Crovella 1996)

- Sampling from heavy-tailed object size distribution, which has infinite variance, ...
  - Average object size in sample does not converge in reasonable times (Central Limit Theorem does not apply any more)
    ⇒ transient simulations (Crovella, Lipsky 2000)
      » Also with a reasonable bound to the object size distribution!
Object Size Quantiles

- Presumably, the $p$-th latency quantile in output can only converge, if the corresponding $p$-th object size quantile (OSQ) has converged.

1. Derive the distribution of sample‘s $p$-th quantile $\xi_q$ around quantile $x_q$ of the distribution which was used for generation of the sample.

2. Derive the asymptotic distribution of sample‘s quantile
   - Normal distribution! (Rao 1973)
   - $\rightarrow$ convergence in reasonable times
### Stabilization of OSQ to 1%

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Heavy-tailed #objects</th>
<th>Exponential #objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>98%</td>
<td>$1.4 \cdot 10^6$</td>
<td>$1.2 \cdot 10^5$</td>
</tr>
<tr>
<td>99%</td>
<td>$2.8 \cdot 10^6$</td>
<td>$1.7 \cdot 10^5$</td>
</tr>
<tr>
<td>99.9%</td>
<td>$2.7 \cdot 10^7$</td>
<td>$8.0 \cdot 10^5$</td>
</tr>
<tr>
<td>99.99%</td>
<td>$2.7 \cdot 10^8$</td>
<td>$4.5 \cdot 10^6$</td>
</tr>
<tr>
<td>Average</td>
<td>$3 \cdot 10^{12}$</td>
<td>800</td>
</tr>
</tbody>
</table>
Latency Quantiles

1. Object size quantiles do converge

2. Exploit theory of robustness for latency quantiles
   - If correlation of a observed random variable is „not too strong“ -> quantiles converge to normality at rate sqrt n (Hampel 1986)

→ Test latency quantiles for convergence to normality
   - Reliable method: normal probability plots (Q-Q plots)
   - Check linearity with linear regression
   - Additionally check consistency (sqrt n rate)
Client Server Scenario

Queue Length: 52KB

Bottleneck Link
Bandwidth: Variable
Delay: 10ms

Access Links
Bandwidth: 10Mb/s
Delay: 0.1ms

Client Side
50 Web Clients

Server Side
5 Web Servers

Access Links
Bandwidth: 10Mb/s
Delay: 0.1ms
Normal Plots

[Graphs showing normal plots with quantiles]
Linearity of N.P. & Consistency

Linearity of Normal Plot (8400Kb/s)

Consistency

Ulrich Fiedler TIK, ETH
Using Quantiles to Characterize User-Perceived Latency
Discussion

• Latency quantiles, e.g. transfer time quantiles, converge if utilization is not too high and the network is not too heterogenous
  • Practical application in performance evaluation of „limited scenarios“
    » Corporate networks, web server, ...

• High utilization
  • Possibly observations of latencies are long range dependent
    ⇒ Quantiles may not converge not to normal, but to \( \alpha \)-stable
    » Exploit Q-Q plots to test for this converge
    » Problems: 1. Need to estimate \( \alpha \) from correlated observations, 2. Likely too slow for practical use
Thanks

• Comments and questions welcome

e-mail: fiedler@tik.ee.ethz.ch
### Source Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Distribution</th>
<th>Average</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of Index Obj.</td>
<td>Pareto vs. Exponential</td>
<td>12000B</td>
<td>1.2</td>
</tr>
<tr>
<td># Embed. Objects</td>
<td>Constant</td>
<td>Zero</td>
<td></td>
</tr>
<tr>
<td>Think Time</td>
<td>Pareto</td>
<td>10 sec</td>
<td>2.0</td>
</tr>
</tbody>
</table>

→ Mean offered load for 50 clients: ~ 480 Kb/s
Characterization of Output

File Size vs. Response Time:
640Kb/s (left) vs. 6400Kb/s (right)