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 $\Rightarrow$ negative impact on performance (Barford, Crovella 1998)
  - Self-similarity $\Leftarrow$ 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

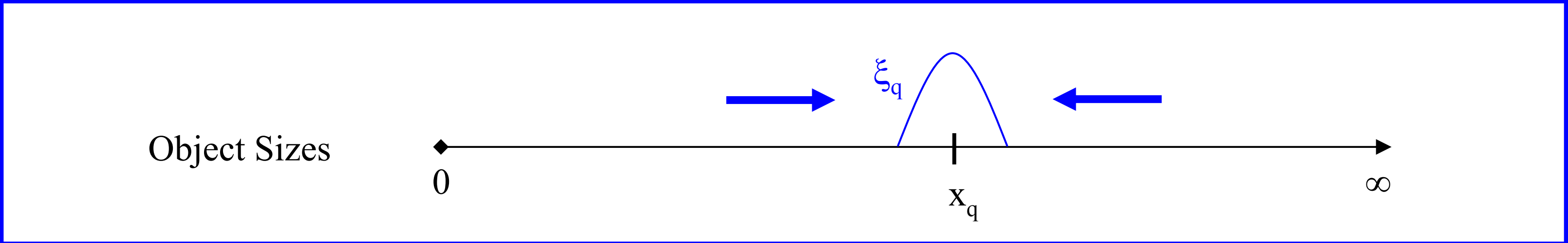
$$F(x) = 1 - \left(\frac{k}{x}\right)^{\alpha} \quad x \in [k, \infty[$$

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) $\Rightarrow$ 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 correponding p-th object size quantile (OSQ) has converged
  1. Derive the distribution of sample‘s p-th quantile ξ_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)
       -> convergence in reasonable times



Stabilization of OSQ to 1%

	Heavy-tailed	Exponential
Quantile	#objects	#objects
98%	$1.4 \cdot 10^{6}$	$1.2 \cdot 10^{5}$
99%	$2.8 \cdot 10^{6}$	$1.7 \cdot 10^{5}$
99.9%	$2.7 \cdot 10^{7}$	$8.0 \cdot 10^{5}$
99.99%	$2.7 \cdot 10^{8}$	$4.5 \cdot 10^{6}$
Average	$3 \cdot 10^{12}$	800

Latency Quantiles

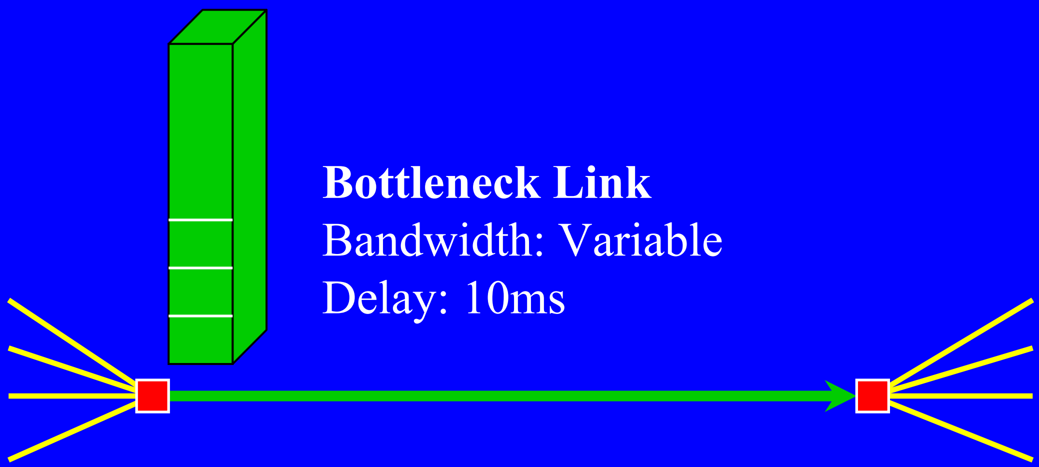
1. Object size quantiles do converge
2. Exploit theory of robustness for latency quantiles
   - If correlation of a observed random variable is „not too strong“ -> quantiles converge to normality at rate sqrt n (Hampel 1986)

→ 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

Server Side

5 Web Servers

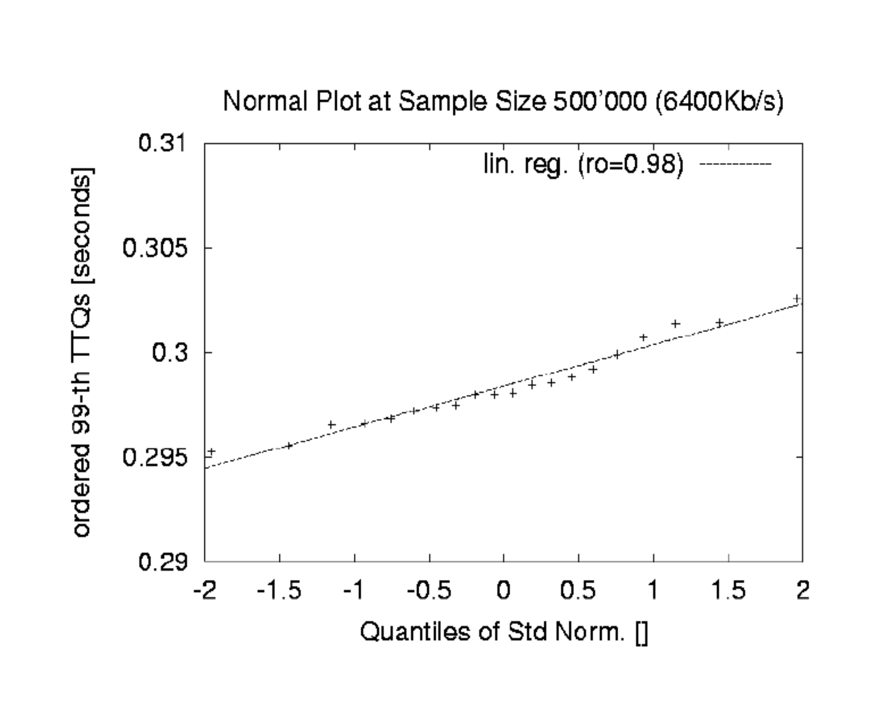
Client Side

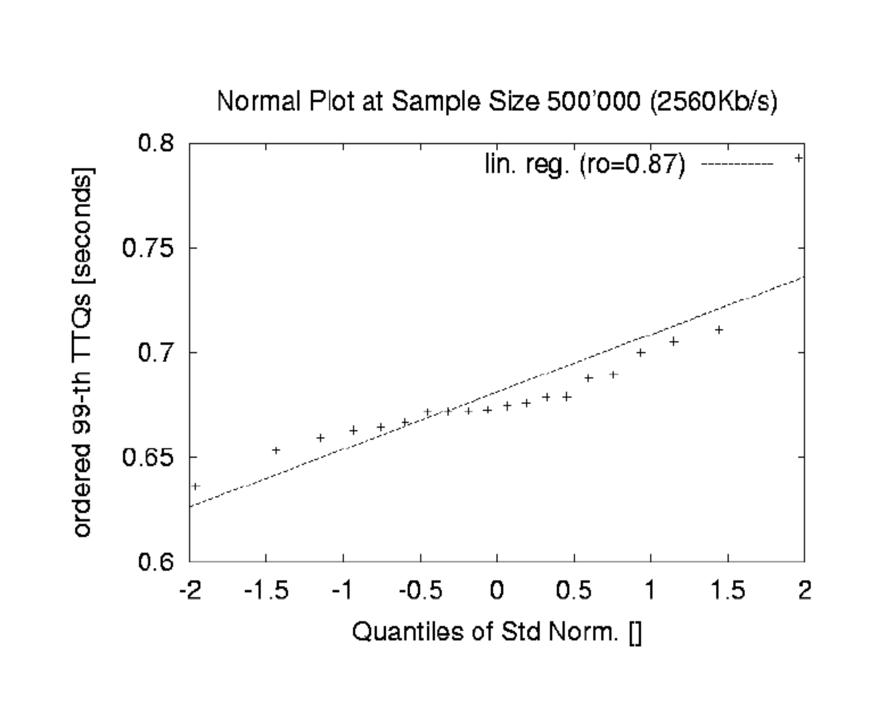
50 Web Clients

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

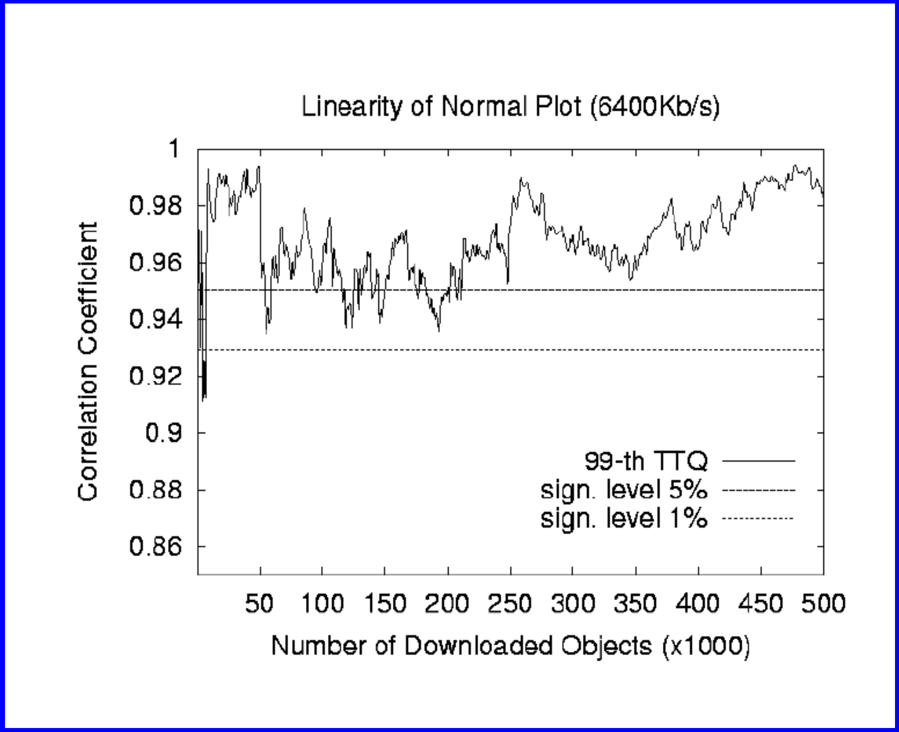
Access Links
Bandwidth: 10Mb/s
Delay: 0.1ms

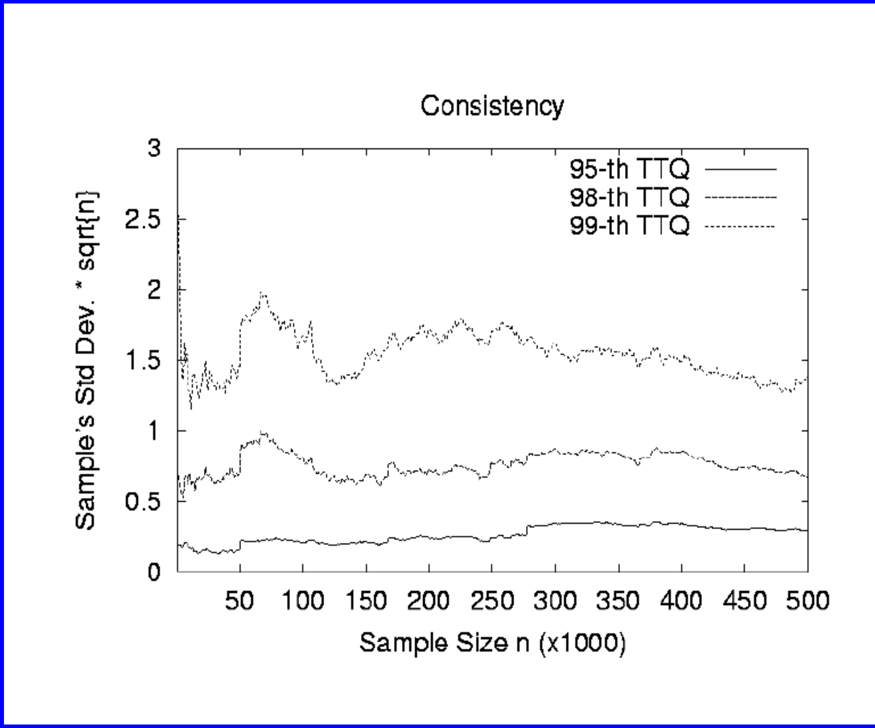
Normal Plots





Linearity of N.P. & Consistency





Discussion

- Latency quantiles, e.g. transfer time quantiles, converge if utilization is not too high and the network is not too heterogenious
    - Practical application in performance evaluation of „limited scenarios“
        - » Corporate networks, web server, ...
- High utilization
    - Possibly observations of latencies are long range dependent $\Rightarrow$ Quantiles may not converge not to normal, but to α-stable
        - » Exploit Q-Q plots to test for this converge
        - » Problems: 1. Need to estimate α from correlated observations, 2. Likely too slow for practical use

Thanks

- Comments and questions welcome

e-mail: fiedler@tik.ee.ethz.ch

Source Model

Parameter	Distribution	Average	Shape
Size of Index Obj.	Pareto vs. Exponential	12000B 12000B	1.2
# Embed. Objects	Constant	Zero	
Think Time	Pareto	10 sec	2.0

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

Characterization of Output

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