Mathematical modeling in support of Service Level Agreements

Prepared For:
IMA/MCIM Industrial Problems Seminar

Telcordia Contact:
Eric van den Berg
(973) 829 4711
February 22, 2002
Outline

- Introduction
- Mathematical work on the structure of Service Level Agreements
- Statistical fluctuations in SLA metrics
- Determining SLA thresholds using Extreme Value Theory
- Conclusions
- Future Work
Introduction

- What is a Service Level Agreement?
- Why are SLAs needed or wanted in packet networks?
- Why are SLAs interesting for mathematicians?
Service Level Agreements: definition

- A **service level agreement** is a **contract** between a customer and a service provider, guaranteeing a specific quality level for a given service.

- **Important ingredients:**
  - **fee** to be paid by customer for guaranteed service
  - **metric** definition for monitoring of service quality
  - **penalty** to be paid by service provider if service level metric exceeds threshold

- **Common features:**
  - availability guarantees
  - maintenance guarantees,
  - customer care guarantees
Service Level Agreements in packet networks

- **Virtual Private Networks** allow (business) customers to economically connect separate locations via a pseudo ‘private’ network, instead of leasing dedicated lines.

- Performance should be like that of a ‘private’ network.
Service Level Agreements in e-commerce

- Web site hosting in **server farms**

  - Bounds on response time critical to avoid missed or delayed transactions
Mathematical work on structure of SLAs

- E. Bouillet et al. “The Structure and Management of Service Level Agreements in Networks”
  - SLA on bandwidth; flow admission control, route selection
  - Web hosting: SLA on response time; solve network flow resource allocation optimization problem
  - SLA on throughput of TCP flows; focus on QoS assurance regions for differentiated services, via stochastic models at different time-scales
An approach to SLAs for IP networks with Differentiated Services

- Consider 3 service classes:
  - EF: voice call class
  - AF1: premium data class (uses TCP)
  - AF2: best effort data class (also uses TCP)

- Model Differentiated services via strict queuing priorities: EF has priority over AF1, which in turn has priority over AF2.

- Quality of Service SLA guarantees:
  - EF: Packet loss probability $\leq 10^{-6}$, Connection blocking prob. $\leq 10^{-3}$
  - AF1: Mean throughput of a connection: $\geq 2000$ kb/sec
    - measured over periods of 1 min, with probability 0.99
  - AF2: Mean throughput of a connection: $\geq 128$ kb/sec
    - measured over periods of 10 min., with probability 0.95
An approach to SLAs for IP networks with Differentiated Services (cont.)

- Combined large collection of (fixed point, traffic, queuing) models at different time-scales into coherent approach to SLA definitions and measurement
- Service Differentiation is difficult. Only a very small region where
  - all QoS assurances are met
  - different QoS assurances are distinguishable
- Points of interest
  - Showed an approach to defining SLAs on throughput of TCP traffic
  - Recognizes the importance of appropriate measurement interval length on metric variability
Maximizing Service Level Agreement profits

- E-commerce setting: Consider a farm of M web servers, hosting N web-sites.
- Assume requests can be divided into K classes: e.g. navigation, shopping card addition, check-out, credit card verification.
- Suppose servers employ weighted processor sharing: class k on server i gets an assigned share $\phi_{ik} C_i$ of server capacity, and requires mean service time $\mu_{i,k}$.
- SLA constraint on response time $T_k$:
  $$P(T_k > z_k) \leq \alpha_k$$
- For each successfully guaranteed request, the provider gets an amount $P_k^+$, and for each violated constraint pays an amount $P_k^-$. 
Maximizing Service Level Agreement profits (cont.)

- Suppose requests for website $j$ of class $k$ arrive according to a Poisson process with rate $\lambda^{(j)}_k$.
- Decision variables:
  - $\lambda^{(j)}_{i,k}$, rate of requests for website $j$ of class $k$ assigned to server $i$.
  - $\phi_{i,k}$, fraction of server capacity of class $k$ at server $i$.

- Optimal values determined by solving a network flow resource allocation problem for the optimal $\lambda^{(j)}_{i,k}$ while keeping $\phi_{i,k}$ fixed, followed by solving for $\phi_{i,k}$ while keeping $\lambda^{(j)}_{i,k}$ fixed, and iterating.
- Profits turn to losses when ratio $P^-_k / P^+_k$ and request load are too large.
SLAs as insurance policies

- Service Level Agreement inherently a form of insurance
- To ‘maximize profits’ from SLAs, it is most important to manage the risk of SLA violations:
  - Violations based on monitored service metric
  - Focus on monitoring of service quality
    - According to mutually accepted metrics
    - By impartial third party?

Service Provider

Independent Monitor
Statistical fluctuations in SLA metrics

- Example SLA metric: end-to-end packet delay measured via injected probe packets

- Analytical approach:
  - Determine effect of long range dependence in packet delays on variance of sample mean probe packet delay

- Statistical approach:
  - study real probe packet data, sampled according to IETF IPPM framework
  - Joint work with Joerg Rothenbuehler, Cornell University
Measured one-way probe packet delays

Estimating Extreme Packet Delays

• Delay measurements can fluctuate strongly, even though average delay is small

• Extreme delays can influence customer experience more than average delay, e.g. in e-commerce setting

• Possible SLA constraint reflecting aversion of extreme delay values:
  – 99% of the agreed-upon measured probe packet delays will be below the value x
  – the agreed-upon measured probe packet delays will not exceed value y in more than one hour per month
Determining SLA thresholds using Extreme Value Theory

• What level $y$ can be guaranteed so that “the agreed-upon metric of performance will not exceed value $y$ in more than one hour per month” is satisfied with confidence level e.g. 95%?

• Would like method that depends on few assumptions, and works also for highly variable measurements
Analysis of Block Maxima

Block 1      Block 2      Block 3     Block 4      Block 5
Generalized Extreme Value distribution (GEV)

- Assumption: block maxima asymptotically follow a Generalized Extreme Value distribution:

\[ H_\xi(x) = \begin{cases} 
\exp \left[-(1 + \xi x)^{-1/\xi} \right] & \text{if } \xi \neq 0 \\
\exp \left[-\exp(-x) \right] & \text{if } \xi = 0 
\end{cases} \]

where \((1 + \xi x) > 0\)
Fisher-Tippett Theorem

Let \( (X_n) \) be a sequence of I.I.D. random variables. Let

\[
M_n = \max(X_1, \ldots, X_n)
\]

If there exist constants \( c_n \) and \( d_n \) such that

\[
\frac{M_n - d_n}{c_n} \overset{d}{\rightarrow} H
\]

as \( n \to \infty \) for some non-degenerate distribution \( H \)

Then

\[
H = H_{\xi}
\]
Assessing the Quality of the Fit

1-Hour Maxima

2-Hour Maxima
Estimation of Extreme Delay Levels

Hourly Maxima

Delay exceeded in one of…
Time-Series Plot Colorado-Harvard

Monday 12:00am - Friday 8 pm
GEV-Fit Results for different Block sizes

- **Block size = 7200 : 108 Blocks**
  
<table>
<thead>
<tr>
<th>xi</th>
<th>sigma</th>
<th>mu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation</td>
<td>-0.3375603</td>
<td>59.75591</td>
</tr>
<tr>
<td>Std. Error</td>
<td>0.0734163</td>
<td>4.69351</td>
</tr>
</tbody>
</table>

- **Block size = 14400 : 54 Blocks**
  
<table>
<thead>
<tr>
<th>xi</th>
<th>sigma</th>
<th>mu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation</td>
<td>-0.4346847</td>
<td>51.7389</td>
</tr>
<tr>
<td>Std. Error</td>
<td>0.1256784</td>
<td>6.2025</td>
</tr>
</tbody>
</table>
## High Level Estimation

### Level exceeded during 1 of 50 hours

<table>
<thead>
<tr>
<th>Block size</th>
<th>Lower Estimate</th>
<th>Upper Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1h</td>
<td>314.8669</td>
<td>326.7487</td>
</tr>
<tr>
<td>1.5h</td>
<td>318.4539</td>
<td>327.3824</td>
</tr>
<tr>
<td>2h</td>
<td>315.7877</td>
<td>324.6415</td>
</tr>
</tbody>
</table>

### Level exceeded during 1 of 100 hours

<table>
<thead>
<tr>
<th>Block size</th>
<th>Lower Estimate</th>
<th>Upper Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1h</td>
<td>322.4220</td>
<td>336.7065</td>
</tr>
<tr>
<td>1.5h</td>
<td>325.7779</td>
<td>335.4343</td>
</tr>
<tr>
<td>2h</td>
<td>324.3893</td>
<td>332.4487</td>
</tr>
</tbody>
</table>
How valid are the predictions?

Obs. = Number of maxima to exceed the corresponding level
Est. = Estimated number of maxima to exceed the corresponding level
L95/U95 = Lower and Upper bound of 95% two-sided CI.

<table>
<thead>
<tr>
<th>Level</th>
<th>Obs.</th>
<th>L95</th>
<th>Est.</th>
<th>U95</th>
<th>U99 (one sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>11</td>
<td>7</td>
<td>10.80</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>20</td>
<td>8</td>
<td>3</td>
<td>5.40</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>30</td>
<td>6</td>
<td>1</td>
<td>3.60</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>40</td>
<td>6</td>
<td>1</td>
<td>2.70</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0</td>
<td>2.16</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>1.08</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>200</td>
<td>0</td>
<td>0</td>
<td>0.54</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Conclusions

• Mathematical modeling to support Service Level Agreements is challenging:
  – requires combining different models for network behavior at widely different time-scales
  – network design process involves much freedom in parameter choice for providers: optimization requires several passes

• SLAs function as insurance for customers:
  – Monitoring based on accepted metric of service quality
  – To obtain achievable SLAs, it is important to model extreme behavior of service metrics

• Extreme Value Theory can help model extremal behavior of metrics
Future work

- Service Level Agreements:
  - Applications of insurance and financial mathematics to structuring SLAs

- Applications of Extreme Value Theory:
  - Design network (capacity) to match given GEV distribution
  - Multivariate extreme value theory to model ‘hot spots’ in network
Thank you!