Service-Level Agreement Durability for Web Service Response Time

Hiranya Jayathilaka
Prof. Chandra Krintz
Prof. Rich Wolski
Computer Science Dept., UC Santa Barbara

IEEE CloudCom 2015
Growth in Web APIs Since 2005

Number of API Today: 14,000+
Source: http://www.programmableweb.com/api-research
Web APIs as IT Resources

APIs do not provide strong guarantees

Client Application

APIS impact user experience

Web Application

Web Application

Web Application
SLAs for Cloud-hosted APIs

• Modern cloud platforms only provide *availability* SLAs for individual APIs
• Cloud platforms do not provide SLAs on deployed user applications and APIs.
• We designed and implemented Cerebro to address these limitations
  – *Response Time Service-Level Agreements for Cloud-hosted Web Applications [SOCC ’15]*
SLA Durability

• Cloud platforms are highly dynamic

• **SLA validity period:** the time until a predicted SLA can no longer be considered correct

• Can we detect when a predicted SLA has become invalid?

• Can we assess the durability of response time SLAs predicted for cloud-hosted web APIs?
Cerebro Architecture

Cloud SDK Monitor

PaaS Cloud

Fetch cloud monitoring data

Static Analyzer

Cloud SDK code

API responds under Q ms, p% of the time

SLA Predictor
(Time series aggregation & QBETS)
Statistical Model

• Suppose at time $t$ Cerebro predicts value $Q$ as the $p^{th}$ percentile of some APIs response time.

• The probability of API’s response time being greater than $Q$:
  ▪ $(1 - 0.01p)$

• Probability of observing $n$ consecutive readings greater than $Q$:
  ▪ $(1 - 0.01p)^n$
A Concrete Example

• Suppose Cerebro predicts that some API responds under 100ms, 95% of the time.
  – Probability of API response time exceeding 100ms is \((1 - 0.01 \times 95) = 0.05\)
  – Probability of observing 3 consecutive such readings is \(0.05^3 = 0.000125\)

• This value 3 is conservative with regard to autocorrelation
  – E.g. To get the same small value 0.000125 with 0.5 autocorrelation, we need to observe 5 events
Detecting SLA Invalidation

• Each time Cerebro makes a prediction, it computes the current autocorrelation in the time series

• Autocorrelation can be used to lookup a table, and determine $C_w$; the number of consecutive readings greater than $Q$, that constitute a rare event

• We consider the SLA to have become invalid if this rare event occurs
SLA Acquisition and Monitoring

• API consumers acquire an initial SLA as part of the API subscription process
  – Cerebro calculates both $Q$ and $C_w$, and records them for future reference

• Cerebro continuously monitors the response time of deployed APIs

• If it observes more than $C_w$ response time measurements greater than $Q$, it considers the prediction to have become invalid
Google App Engine Experiment

- We applied the above statistical model to a set of web APIs deployed in GAE.
- Are the predicted SLAs valid? [SOCC ‘15]
- If so, for how long are they valid?
- What would an individual user experience?
  - SLA validity period
  - Number of renewals due to invalidations
Step 1: Data Gathering

• We deployed a set of APIs in Google App Engine, and monitored their response time over 3 months.
  – Used a set of open source applications

• We also measured and recorded the response time of individual cloud SDK calls made by these APIs.
  – Using Cerebro’s Cloud SDK Monitor
Step 2: SLA Prediction

• We used Cerebro to make response time SLA predictions for the test web APIs.
• Cerebro analyzed the cloud SDK performance data gathered over 3 months, and made 95\textsuperscript{th} percentile predictions for the test web APIs.
  – One prediction per minute, thus forming time series of SLA predictions
  – Each prediction is accompanied by a $C_w$ value
Step 3: Simulation

- We used the predicted SLAs, and the actual API response times measured during the 3 month period in a series of simulations.

<table>
<thead>
<tr>
<th>Predicted SLAs:</th>
<th>40</th>
<th>42</th>
<th>41</th>
<th>41</th>
<th>42</th>
<th>42</th>
<th>42</th>
<th>41</th>
<th>41</th>
<th>41</th>
<th>42</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_w$:</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Actual values:</td>
<td>35</td>
<td>38</td>
<td>39</td>
<td>41</td>
<td>41</td>
<td>42</td>
<td>41</td>
<td>41</td>
<td>41</td>
<td>41</td>
<td>42</td>
</tr>
</tbody>
</table>

User 1 SLA Renewal

125,000+ Users

User 1 SLA Validity Period

User 1 SLA Invalidation
### SLA Validity Periods (In Hours)

<table>
<thead>
<tr>
<th>API</th>
<th>5th Percentile</th>
<th>Mean</th>
<th>95th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>StudentInfo#getStudent</td>
<td>12.97</td>
<td>631.24</td>
<td>1911.19</td>
</tr>
<tr>
<td>StudentInfo#deleteStudent</td>
<td>7.65</td>
<td>472.07</td>
<td>2031.59</td>
</tr>
<tr>
<td>ServerHealth#info</td>
<td>12.96</td>
<td>630.01</td>
<td>1911.19</td>
</tr>
<tr>
<td>Rooms#getRoomByName</td>
<td>8.48</td>
<td>345.13</td>
<td>1096.53</td>
</tr>
<tr>
<td>Rooms#getRoomsInCity</td>
<td>20.56</td>
<td>296.44</td>
<td>1143.45</td>
</tr>
<tr>
<td>Stocks#buy</td>
<td>8.46</td>
<td>411.75</td>
<td>815.5</td>
</tr>
</tbody>
</table>
SLA Renewals Per User

![SLA Renewals Per User Chart]

- **StudentInfo#addStudent**
- **StudentInfo#getStudent**
- **StudentInfo#deleteStudent**
- **ServerHealth#info**
- **Rooms#getRoomsInCity**
- **Rooms#getRoomByName**
- **Stocks#buy**

**Renegotiation Events Experienced by the API Consumer**

**Probability**

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

0 1 2 3 4 5 6
Conclusions

• Web APIs impact the performance of the applications that depend on them.
• Cerebro provides a way to automatically predict response-time SLAs for APIs.
• We present a statistical model that can detect when a predicted SLA has become invalid.
• We extend Cerebro with a simple SLA acquisition and renewal model.
• We show that Cerebro predicted SLAs are highly durable, and the API consumers do not have to renew them too often.