Capacity Planning with Stress Testing

Zhao Cheng, MOGU Inc.

forrest.zhaoc@gmail.com
@forrest_zhao
About Me

Zhao Cheng
Tech Lead, SRE

- Joined MOGU in January 2015
- Senior SWE in HUAWEI 7 Years
- Columnist & Speaker of InfoQ China
- Tencent Cloud Valuable Professional
About MOGU Inc.
Shopping Carnivals

Single’s Day  11.11

12.12

6.18

Valentine’s Day 2.14
The Key Steps of Capacity Planning

- Estimation of Business Model
- Analyzation of Critical Path
- Simulation of Traffic Load
User and System Behaviors

Users
Snaping Up Products

System
Creating Orders
The Most Key Process of Estimation

Estimation of Peak TPS of Order Creation at 0.
Derivation of Peak Order Creation Amount

**Known**
- GMV
- UV
- Conversion Rates
- Per Cust Trans
- Historical Data
- ..........

**Unknown**
- QPS
- TPS
- Sessions
- ..........

POCA = \[ \frac{\text{Objective GMV}}{\text{Historical GMV}} \times \frac{\text{Per Cust Trans}}{\text{Historical Per Cust Trans}} \times \text{Historical POCA} \]
Peak TPS = POCA × Called Times of Order APIs
Estimation of Business Model

Inverted Pyramid

Detail QPS = TPS of Order × (1+conversion rate)

Known: TPS of Order Creation
1K+ Distributed Java Apps and Services
Strong and Weak Dependencies

- Application Level
- Service Level
- API Level
Critical Transaction Path

- Home
- Shop
- Category
- Detail
- Cart
- Comment
- Price
- Coupon
- Payment
- Logistics
- Buy

Critical (Green)
Normal (Blue)
Stress Testing Process

• Single Machine in Each App Cluster
• Single Critical Path
• Full Critical Path
Single Machine in Each App Cluster
Single Machine in Each App Cluster

Requests → Load Balance

10% → RPC Service
10% → RPC Service
60% → RPC Service
Machines of Each Cluster

= 

Estimation of QPS / Single Machine QPS Capacity
SLO Management of Capacity

<table>
<thead>
<tr>
<th>SLOs</th>
<th>Request Per Minute</th>
<th>Peak TPS</th>
<th>Overall Availability</th>
<th>90th Percentile(ms)</th>
<th>95th Percentile(ms)</th>
<th>Error(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018-11-28</td>
<td>1369</td>
<td>40</td>
<td>96.01%</td>
<td>30.21</td>
<td>48.35</td>
<td>0.07</td>
</tr>
<tr>
<td>2018-11-27</td>
<td>926</td>
<td>55</td>
<td>99.99%</td>
<td>32.27</td>
<td>44.41</td>
<td>0.02</td>
</tr>
<tr>
<td>2018-11-27</td>
<td>936</td>
<td>51</td>
<td>99.10%</td>
<td>54.24</td>
<td>73.89</td>
<td>0.12</td>
</tr>
<tr>
<td>2018-11-28</td>
<td>1130</td>
<td>36</td>
<td>99.98%</td>
<td>35.54</td>
<td>49.2</td>
<td>0.03</td>
</tr>
<tr>
<td>2018-11-28</td>
<td>1213</td>
<td>44</td>
<td>97.12%</td>
<td>61.21</td>
<td>81.48</td>
<td>2.87</td>
</tr>
</tbody>
</table>
Single and Full Critical Path

Scenario

- 10% Task
- 30% Task
- .......
- 120% Task

120% Task → Script1 → Test Data1
30% Task → Script2 → Test Data2
 ....... → Script3 → Test Data3
10% Task → Script4 → Test Data4
Automatic Generation of Script and Data

Stress Testing Script

Stress Testing Data
Stress Testing Architecture
# Stress Testing Report

## STATISTICS

<table>
<thead>
<tr>
<th>Executions</th>
<th>Total</th>
<th>OK</th>
<th>KO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>33420</td>
<td>33310</td>
<td>110</td>
</tr>
</tbody>
</table>

| Mean req/s | 106.022 | 105.673 | 0.349 |

<table>
<thead>
<tr>
<th>Response Time (ms)</th>
<th>Total</th>
<th>OK</th>
<th>KO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>5</td>
<td>28</td>
<td>5</td>
</tr>
<tr>
<td>50th percentile</td>
<td>50</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>75th percentile</td>
<td>59</td>
<td>60</td>
<td>16</td>
</tr>
<tr>
<td>95th percentile</td>
<td>93</td>
<td>94</td>
<td>37</td>
</tr>
<tr>
<td>99th percentile</td>
<td>131</td>
<td>131</td>
<td>64</td>
</tr>
<tr>
<td>Max</td>
<td>659</td>
<td>659</td>
<td>92</td>
</tr>
<tr>
<td>Mean</td>
<td>55</td>
<td>55</td>
<td>14</td>
</tr>
<tr>
<td>Std Deviation</td>
<td>20</td>
<td>19</td>
<td>12</td>
</tr>
</tbody>
</table>
Summary

• Estimation of Business Model
• Analyzation of Critical Path
• Simulation of Traffic Load
Thanks