Capacity Planning & Performance for Web-based Services

“What the hell? We could use an idiot.”

by Bill Greenbaum, CSM

for NYCSPIN, May 11, 2010
4 Questions of Capacity Planning

1. What it is?
2. Why is it important to me?
3. Who does it?
4. How to do it?
What is Capacity?

• **Capacity**: how much traffic can your system handle before crashing?
  – **Throughput**: count of requests per period (TPM)
  – **Concurrency**: how many simultaneous requests
What is Capacity?

• **Capacity**: how much traffic can your system handle before crashing?
  – *Throughput*: count of requests per period (TPM)
  – *Concurrency*: how many simultaneous requests

Contrasts with:

• **Performance** = average response time (sec/req)
• **Redundancy** = capacity reserved for business continuity
What is Capacity Planning?

- Capacity Planning is good management.
  - Proactive
  - A form of risk management
  - Ensures competitiveness & profitability

...contrasts with

- Firefighting = fixing issues that already have customer impact
Capacity contrast with Performance

- **Capacity** = throughput (TPM)
- **Performance** = average response time (sec/req)
<table>
<thead>
<tr>
<th>Market</th>
<th>Dec 2008</th>
<th>Dec 2009</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worldwide</td>
<td>89.7</td>
<td>131.4</td>
<td>46%</td>
</tr>
<tr>
<td>Google sites</td>
<td>55.6</td>
<td>87.8</td>
<td>58%</td>
</tr>
<tr>
<td>Yahoo! sites</td>
<td>8.4</td>
<td>9.4</td>
<td>13%</td>
</tr>
<tr>
<td>Baidu.com Inc</td>
<td>8.0</td>
<td>8.5</td>
<td>7%</td>
</tr>
<tr>
<td>Microsoft sites</td>
<td>2.4</td>
<td>4.1</td>
<td>70%</td>
</tr>
<tr>
<td>eBay</td>
<td>1.3</td>
<td>2.1</td>
<td>58%</td>
</tr>
<tr>
<td>NHN Corp</td>
<td>1.9</td>
<td>2.1</td>
<td>9%</td>
</tr>
<tr>
<td>Yandex</td>
<td>1.0</td>
<td>1.9</td>
<td>91%</td>
</tr>
<tr>
<td>Facebook</td>
<td>1.1</td>
<td>1.6</td>
<td>54%</td>
</tr>
<tr>
<td>Ask Network</td>
<td>1.0</td>
<td>1.5</td>
<td>43%</td>
</tr>
<tr>
<td>Alibaba.com Corp</td>
<td>1.1</td>
<td>1.1</td>
<td>-1%</td>
</tr>
</tbody>
</table>

Source: comScore
Who Does Capacity Planning?

- **Capacity Planner** guides others thru the process, does toughest analysis, crafts proposal.
- **Technical Project Managers** are the glue
- **Business Owners** provide critical projections
- **Application Developers**
- **System Developers**
- **Architecture/Hardware Resource team**
- **Database Admins**
- **Disk Storage Admins**
- **Network Admins**
- **QA** validates & does load tests
How (a): Capacity Planning Theory

- Amdahl Scalability Model, 1967
- Gunther Universal Scalability Model, 2000

\[ C(p) = \text{Capacity of } p \text{ Processors} \]

\[ C(p) = \text{Capacity of } p \text{ Processors} \]

\[ \text{Amdahl} \]

\[ \text{Gunther} \]

\[ p = \# \text{ of processors} \]
How (b): Practice

- Obtain Business projections (VoC)
- Model business projections into peak throughput and concurrency
- Load test / Stress test.
- Follow the data to find bottlenecks
- Monitor traffic and look for trends
- Work with lead-time for procurement
How (c): “Follow the data”

- Weak: “+11% YOY, -2% MOM”
- Moderate: Regression lines
- Strong: Control charts include variation in data
Capacity Planning: Example
Further reading

• *The Art of Capacity Planning*, by John Allspaw, 2008, (O’Reilly Media)


• *Building Scalable Web Sites*, by Cal Henderson, 2006 (O’Reilly Media)

• *Guerrilla Capacity Planning*, by Neil J. Gunther, 2007 (Springer)
Back to you, Olly!