Tradeoffs in Designing Web Clusters

Daniel A. Menascé
Department of Computer Science
George Mason University

Typical Questions

• Should I use a large number of low-capacity inexpensive servers or a small number of high-capacity costly ones?
• How many servers of a given type are required to provide a certain performance level at a given cost?
• How many servers are needed to build a Web site with a given reliability?
Comparison Criteria

- Equal Average Response Time
- Equal Cluster Capacity: \( nX = mkX \).
  - \( m = n / k \).
- Equal Cost: \( n \ C(X) = m \ C(kX) \).
  - \( m = n \ C(x) / C(kX) \)
- Equal Reliability.
Reliability Considerations

\[ R_A = 1 - (1 - r_A)^n \]
\[ R_B = 1 - (1 - r_B)^m \]
\[ R_A = R_B \quad \Rightarrow \quad m = \frac{n \log(1 - r_A)}{\log(1 - r_B)} \]

Basic Performance Model (M/G/1)

\[ T = S + \frac{U \times S(1 + C^2)}{2(1 - U)} \]

- T: average response of a Web request
- S: average request service time
- C: coefficient of variation of the service time
- U: server utilization, equal to \( \lambda_w S \)
The utilization has to be < 1 …

\[ U_A = \frac{\lambda}{n} \times \frac{1}{X} < 1 \quad \Rightarrow \lambda < nX \]

\[ U_B = \frac{\lambda}{m} \times \frac{1}{kX} < 1 \quad \Rightarrow \lambda < mkX \]

Response Time Equations

\[ T_A = \frac{1}{X} + \frac{\lambda}{n} \times \left( \frac{1}{X} \right)^2 \left( 1 + C^2 \right) \]

\[ T_B = \frac{1}{kX} + \frac{\lambda}{m} \times \left( \frac{1}{kX} \right)^2 \left( 1 + C^2 \right) \]
Average Number of Requests

• From Little’s Law:

\[ N_A = \lambda T_A \]

\[ N_B = \lambda T_B \]

Equal Response Time Case

\[ T_A = T_B \quad \Rightarrow \quad m = \frac{1}{kX} \left[ \lambda + \frac{1}{2(k-1)} + \frac{k}{\lambda(1+C^2) + \frac{k}{nX - \lambda}} \right] \]

\[ \lim_{\lambda \to nX} m = n/k \]
• When cluster A has 400 servers and $\lambda = 4,810 \text{ req/sec}$, $c$ cluster B needs 160 servers to obtain the same avg. response time of 1.03 sec.

• For a sufficiently large value of $\lambda$, $m$ increases linearly with $\lambda$ at the rate of $\frac{(k - 1)}{(k^2 X)}$. 
Equal Capacity Case

- \( m = n/k \)
- Thus, \( T_B = T_A / k \) \( N_B = N_A / k \)

- Cluster B average response time is always \( k \) times less than that of cluster A.
- Cluster B can handle \( 1/k \) of the requests that cluster A can handle.
Equal Cost Case

- If the cost is proportional to the capacity, then $m$ has the same value as in the equal capacity case.
- Sublinear cost function: there is some economy of scale (i.e., the cost per unit of capacity decreases with the capacity).
- Superlinear cost function: the server cost per unit of capacity increases with the capacity.

Equal Cost Case: sub-linear cost function

$$C(x) = \alpha \sqrt{x} \Rightarrow m = m/\sqrt{k}$$

$$U_A = U_B \sqrt{k}$$
Equal Cost Case: sublinear cost function

Utilization of Servers in Cluster A

Avg. Response Time (sec)

Cluster A

Cluster B

n = 400
m = 179

Equal Cost Case: superlinear cost function

\[ C(x) = \alpha x^2 \Rightarrow m = m/k^2 \]

\[ U_B = kU_A \]
Equal Cost Case: superlinear cost function

Equal Reliability Case
Comparison: $\lambda = 4,800$ req/sec; $TA= 1.025$ sec; $UA= 0.6; n = 400$

<table>
<thead>
<tr>
<th>$T_B$ (sec)</th>
<th>$m$</th>
<th>$U_B$</th>
<th>Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.025</td>
<td>54</td>
<td>89%</td>
<td>Equal response time.</td>
</tr>
<tr>
<td>0.205</td>
<td>80</td>
<td>60%</td>
<td>Equal total capacity</td>
</tr>
<tr>
<td>0.058</td>
<td>179</td>
<td>27%</td>
<td>Equal cost. $C(x) = vx$</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>&gt; 100%</td>
<td>Equal cost. $C(x) = x^2$</td>
</tr>
<tr>
<td>0.083</td>
<td>133</td>
<td>36%</td>
<td>Equal cluster reliability.</td>
</tr>
</tbody>
</table>