Two-Level Iterative Queuing Modeling of Software Contention

Daniel A. Menascé
Dept. of Computer Science
George Mason University
www.cs.gmu.edu/faculty/menasce.html


Motivation

Software View

Hardware View
% Software Contention Time vs. Multithreading Level

% Software Contention Time vs Non-CS/CS Ratio

© 2002 D. Menascé. All Rights Reserved.
If we just consider the SQN we are ignoring time spent at the software resources due to contention for hardware resources.

The HQN model must consider that some processes are not using hardware resources because they are blocked for software resources.
SQN-HQN Scheme

\[ \sum_j s_j \cdot D_{sh}^{i,j,i} \]

\[ \sum_j D_j^f \]

Input Service Demands

<table>
<thead>
<tr>
<th>Hardware Devices</th>
<th>Software Modules</th>
<th>Hardware Demands</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NCS</td>
<td>CS 1</td>
</tr>
<tr>
<td>CPU</td>
<td>0.2000</td>
<td>0.0600</td>
</tr>
<tr>
<td>Disk 1</td>
<td>0.0560</td>
<td>0.0576</td>
</tr>
<tr>
<td>Disk 2</td>
<td>0.0360</td>
<td>0.0000</td>
</tr>
<tr>
<td>Disk 3</td>
<td>0.0360</td>
<td>0.0000</td>
</tr>
<tr>
<td>Software Demands</td>
<td>0.3280</td>
<td>0.1176</td>
</tr>
</tbody>
</table>
Service Demands

<table>
<thead>
<tr>
<th>Hardware Devices (j)</th>
<th>NCS</th>
<th>CS 1</th>
<th>CS 2</th>
<th>Hardware Demands</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>0.2000</td>
<td>0.0600</td>
<td>0.0808</td>
<td>0.3408</td>
</tr>
<tr>
<td>Disk 1</td>
<td>0.0560</td>
<td>0.0576</td>
<td>0.0000</td>
<td>0.1136</td>
</tr>
<tr>
<td>Disk 2</td>
<td>0.0360</td>
<td>0.0000</td>
<td>0.1212</td>
<td>0.1572</td>
</tr>
<tr>
<td>Disk 3</td>
<td>0.0360</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0360</td>
</tr>
<tr>
<td>Software Demands</td>
<td>0.3280</td>
<td>0.1176</td>
<td>0.2020</td>
<td></td>
</tr>
</tbody>
</table>

Software Modules

\[ D_{cs1}^{s} \] (total service time of a software module)

\[ D_{cs1,disk1}^{sh} \] (total service time of a software module at a physical device)

\[ D_{disk1}^{h} \] (total service time of the application at a given device)

SQN-HQN Scheme

\[ D_{i,j}^{sh} \] Adjust demands

\[ R_{i}(N_{h}^{i}) \] (avg. number of processes blocked due to software contention.

\[ \sum_{j} \] Hardware QN

\[ N_{h} \] Hardware QN

\[ B \] Adjust demands

\[ D_{j}^{s} \] Software QN

© 2002 D. Menascé. All Rights Reserved.
SQN-HQN Scheme

\[ \sum_j s_j D_{sh ij} \]

Software QN

Hardware QN

Adjusted demands

\[ R_i(N^h) \]

Avg. number of processes blocked due to software contention.

Adjusted population for HQN

Adjusted demands for SQN.
Basic Idea

• Iteration between solving the SQN and the HQN.
• Number B of processes blocked due to software contention computed through the SQN.
• Population at HQN is reduced by B.
• Service demands at SQN are adjusted to account for physical contention.

SQN-HQN: Initialization

\[
\sum_j s_{ij} D_{sh}^{ij}
\]

Software QN

No adjustment at initialization

Adjust demands

\[
R_i'(N^h_i)
\]

Hardware QN

\[
D_i^h
\]

\[
D_j^s
\]

\[
B
\]

\[
N^h
\]

\[
N
\]
SQN-HQN: Solve SQN – No Hardware Contention

\[ \sum_j s_j D_{ij}^s \]

Software QN

\[ D_j^f \]

Adjust demands

\[ R_i^f(N^h) \]

Hardware QN

\[ N^h = N - B \]

SQN-HQN: Solve HQN

\[ \sum_j D_{ij}^h \]

Software QN

\[ D_j^s \]

Adjust demands

\[ R_i^s(N^h) \]

Hardware QN

\[ N^h = N - B \]
SQN-HQN: Adjust demands for SQN

\[ \sum_j s_j D_{sh} \]

<table>
<thead>
<tr>
<th>Adjust demands</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B )</td>
</tr>
</tbody>
</table>

\[ N^h = N - B \]

SQN-HQN: Solve SQN Again

\[ \sum_j s_j D_{sh} \]

<table>
<thead>
<tr>
<th>Adjust demands</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B )</td>
</tr>
</tbody>
</table>

\[ N^h = N - B \]
**SQN-HQN: Solve HQN**

\[ \sum_{j} s_{j} D_{sh_{j,i}} \]

\[ D_{sh_{j,i}} \]

\[ D_{sh}^{i} \]

\[ B \]

\[ \text{Software QN} \]

\[ \text{Hardware QN} \]

\[ N \]

\[ R_{i}^{'}(N^{h}) \]

\[ N^{h} = N - B \]

**SQN-HQN: Adjust demands for SQN**

\[ \sum_{j} s_{j} D_{sh_{j,i}} \]

\[ D_{sh_{j,i}} \]

\[ D_{sh}^{i} \]

\[ B \]

\[ \text{Software QN} \]

\[ \text{Hardware QN} \]

\[ N \]

\[ R_{i}^{'}(N^{h}) \]

\[ N^{h} = N - B \]

© 2002 D. Menascé. All Rights Reserved.
and so on …

Convergence is checked on absolute relative error on the number of blocked processes in the SQN.

Adjustment of SQN Demands

- Single class case:

\[ D_j^s \leftarrow \sum_i \frac{D_{j,i}^{sh}}{D_i^h} \times R_i^r(\bar{N}_i^h) \]

- Multiple class case:

\[ D_{j;r}^s \leftarrow \sum_i \frac{D_{j;i,r}^{sh}}{D_{i;r}^h} \times R_{i;r}^r(\bar{N}_{i;r}^h) \]
Example

Software QN

Hardware QN

Comparison with other approaches

<table>
<thead>
<tr>
<th>N</th>
<th>SQN-HQN</th>
<th>GB</th>
<th>SQN</th>
<th>HQN</th>
<th>ASM</th>
<th>ASPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.544</td>
<td>1.54</td>
<td>0.27</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2.088</td>
<td>2.11</td>
<td>1.06</td>
<td>4.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.317</td>
<td>2.37</td>
<td>2.22</td>
<td>5.89</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2.428</td>
<td>2.49</td>
<td>2.49</td>
<td>5.99</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.487</td>
<td>2.56</td>
<td>2.86</td>
<td>5.87</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.521</td>
<td>2.60</td>
<td>3.05</td>
<td>5.78</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2.541</td>
<td>2.62</td>
<td>3.00</td>
<td>5.75</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2.555</td>
<td>2.63</td>
<td>2.86</td>
<td>5.77</td>
<td>1.1</td>
<td></td>
</tr>
</tbody>
</table>

GB: global balance equations

SQN is consistently pessimistic.
ASPA is much more complex to implement.
Modeling Non-Software Resources

Client think time
ncs
network
Software QN
B is the avg. no. of processes in the software resource waiting lines.

Client think time
CPU
network
Hardware QN

Open QN at the Software Level

\( \lambda \)
ncs
Software QN

Hardware QN
SQN-HQN Scheme: Open SQN

\[
\sum_j s_j D_{sh}^{ij} \rightarrow \text{Software QN}
\]

\[
D_j^f \rightarrow \text{Adjust demands} \rightarrow N,B
\]

\[
R_i^h (N^h) \rightarrow \text{Hardware QN}
\]

\[
N^h = N - B
\]

Results of Iterations for Open SQN Case

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Nh</th>
<th>Resp. Time</th>
<th>Ns</th>
<th>B</th>
<th>Adjusted SQN Demands</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NCS</td>
</tr>
<tr>
<td>0</td>
<td>0.3280</td>
<td>0.1176</td>
<td>0.2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.295</td>
<td>0.821</td>
<td>1.641</td>
<td>0.346</td>
<td>0.3662</td>
</tr>
<tr>
<td>2</td>
<td>1.440</td>
<td>0.946</td>
<td>1.893</td>
<td>0.453</td>
<td>0.3858</td>
</tr>
<tr>
<td>3</td>
<td>1.513</td>
<td>1.014</td>
<td>2.028</td>
<td>0.515</td>
<td>0.3958</td>
</tr>
<tr>
<td>4</td>
<td>1.550</td>
<td>1.050</td>
<td>2.100</td>
<td>0.549</td>
<td>0.4010</td>
</tr>
<tr>
<td>5</td>
<td>1.570</td>
<td>1.069</td>
<td>2.137</td>
<td>0.568</td>
<td>0.4037</td>
</tr>
<tr>
<td>6</td>
<td>1.580</td>
<td>1.079</td>
<td>2.157</td>
<td>0.577</td>
<td>0.4051</td>
</tr>
<tr>
<td>7</td>
<td>1.585</td>
<td>1.084</td>
<td>2.167</td>
<td>0.582</td>
<td>0.4059</td>
</tr>
<tr>
<td>8</td>
<td>1.588</td>
<td>1.086</td>
<td>2.173</td>
<td>0.585</td>
<td>0.4062</td>
</tr>
<tr>
<td>9</td>
<td>1.589</td>
<td>1.088</td>
<td>2.175</td>
<td>0.587</td>
<td>0.4064</td>
</tr>
<tr>
<td>10</td>
<td>1.590</td>
<td>1.088</td>
<td>2.177</td>
<td>0.587</td>
<td></td>
</tr>
</tbody>
</table>
Related Work

• [Thomasan, 1983]: two-level model (QN+MC). Does not generalize to multiple classes.
• [Jacobson&Lazowska, 1983]: transform network to one with population constraints and solve through approximation.
• [Kahkipuro, 2000] Performance modeling framework for CORBA-based systems. QNs with simultaneous resource possession.

Concluding Remarks

• Simple approach.
• Open, closed, and multiclass QNs can be used at the SQN.
• SQNs can include non-software resources that are not mapped to hardware resources.
• HQNs are closed and can be multiclass.
• Any technique can be used to solve the SQN and HQN. This includes any known approximation to multiple-server devices, priorities, simultaneous resource possession, etc.