SQF: A Slowdown Queueing Fairness Measure

Benjamin Avi-Itzhak, Rutgers University, USA
Eli Brosh, Columbia university, USA
Hanoch Levy, Tel-Aviv University, Israel

Performance 2007
Fairness in Queueing Models

- **Fairness is**
  - Inherent/crucial part of queues
  - Important for many applications

- **Analysis of job fairness**: was little, now growing

- **Our Work**: design an effective scaled fairness measure based on the slowdown principle
Fairness: Size and Seniority

• Inherent physical factors of queueing system

\begin{itemize}
  \item Size \hspace{2cm} Seniority \hspace{2cm} Resources
  \item (Service Requirement) \hspace{2cm} (Arrival Time)
\end{itemize}

• Fairness (and scheduling) deal with these two fundamental physical factors.
The Size vs. Seniority Dilemma

- Mr. Short vs. Mrs. Long

- Is it more fair to serve Short ahead of Long? By how much?
How Fairness is Measured? Prior Work

● **Order based fairness**
  - Avi-Itzhak & Levy (96,04):
  - For G/D/1 system: fairness = variance of the waiting time

● **RAQFM: deviations from equal resource sharing**
  - Raz, Levy & Avi-Itzhak (04,05)
  - Reacts both to size and seniority

● **Expected slowdown fairness criterion**
  - Wierman & Harchol-Balter (03,05):
  - For a given job size mean slowdown is compared with PS
  - Classification of a large variety of policies: always fair/unfair, sometimes fair
## Comparison of Approaches & Motivation

<table>
<thead>
<tr>
<th></th>
<th>Order Based Fairness</th>
<th>RAQFM</th>
<th>Slowdown Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reacts to Seniority</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Reacts to size</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Full Scale</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Yields to Analysis</td>
<td>Non-preemptive Single server ✓ ✓</td>
<td>Complex ✓</td>
<td>Wide range of policies Single server ✓ ✓ ✓</td>
</tr>
</tbody>
</table>
Requirements From a Fairness Measure

- Based on a “sound” intuitive basis
  - *The constant slowdown principle*
- Fit widely accepted intuition in simple cases
  - *Reacts to both seniority and size*
- Yields to analysis
  - *Closed form formulas for a variety of systems*

SQF: A Slowdown Queueing Fairness Measure
The slowdown/proportionality principle
- The sojourn time of a job should be proportional to its size

Absolute fairness if for all jobs: \( \frac{T_i}{x_i} = c \iff T_i = c \times x_i \)
- \( c \): a slowdown constant
Individual and Scenario Unfairness

- **Individual discrimination:**
  \[ D_i = cx_i - T_i \]

- How to compute the constant?
  - Overall system discrimination must equal zero

\[
\sum_{i=1}^{N} D_i = \sum_{i=1}^{N} T_i - \sum_{i=1}^{N} cx_i = 0 \Rightarrow c = \bar{T} / \bar{x}
\]

- **Scenario unfairness:**

  \[
  SQF = \frac{1}{N} \sum_{i=1}^{N} (T_i - cx_i)^2
  \]
System Unfairness

- A stationary system with job size distribution \( b(x) \)

\[
SQF = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} (T_i - cx_i)^2 = \int E[T(x) - cx]^2 b(x) dx
\]

\[
= E[T^2] + \frac{b_2}{b_1} E^2[T] - \frac{2E[T]}{b_1} \int xE[T(x)]b(x) dx
\]

- \( c = \frac{E[T]}{b_1} \): efficiency of an absolute fair system
SQF Properties: Sensitivity to Size and Seniority

- **Theorem:** [Seniority] when two jobs have identical sizes, it is more fair to complete the service of the more senior job earlier.
  (Proof sketch: compare scenarios)

- **Theorem:** [Size] if the arrival times are identical, it is more fair complete the service the shorter job earlier
  (Proof sketch: compare scenarios)

**SQF reacts well to seniority and size**
SQF Analysis: Non-Preempt Polices

- **Example A:** G/G/m with *non-preemptive* scheduling policies
  \[ SQF = E^2[W] (\gamma^2_w + \gamma^2_s) \]
  - *Theorem:* preferential service to senior jobs increases fairness
  - Fairness ranking: (most fair) FCFS < ROS < LCFS (most unfair)

**SQF agrees with the intuition**
SQF Analysis: Preemptive Policies

- **Example B:** M/G/1 Symmetric Queues
  - Processor Sharing, preemptive LCFS, preemptive ROS

<table>
<thead>
<tr>
<th>SQF</th>
<th>Slowdown criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unfairness $= \text{Var}[T]$</td>
<td>Sym Q are always fair</td>
</tr>
<tr>
<td></td>
<td>But $\text{Var}^{\text{PS}}[T] &lt; \text{Var}^{\text{P-LCFS}}[T]$</td>
</tr>
</tbody>
</table>

SQF is more sensitive than the criterion to seniority and agrees with the intuition
SQF Analysis: Multi Server

- **Example C:** Multi-server multi-queue systems
- Monotonic in #servers: SQF(M/M/m+1) < SQF(M/M/m)
  - Under SQF: M/M/m is more fair than M/M/1

- **Intuition:** shorts may complete service before longs even if they arrive later
Measure evaluation

- Evaluate SQF behavior via numerical/simulations for a variety of policies
- Reaction to size and seniority
SQF Evaluation: M/D/1

- Only seniority matters (equal job sizes)
SQF Evaluation: M/G/1, c.v=10

- Large size variability (size dominates seniority)

Non-size based: severe size violation

SRPT, PS: prefer shorts
## SQF: Comparison to other approaches

<table>
<thead>
<tr>
<th></th>
<th>SQF</th>
<th>Order based fairness</th>
<th>RAQFM</th>
<th>Slowdown criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Seniority</strong></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td><strong>Size</strong></td>
<td>√</td>
<td></td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td><strong>Full Scale</strong></td>
<td>√</td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td><strong>Analysis</strong></td>
<td>Non-preemptive &amp; SQ multi server</td>
<td>Non-preemptive Single server</td>
<td>Complex</td>
<td>Wide range Single server</td>
</tr>
<tr>
<td></td>
<td>√ √</td>
<td>√ √</td>
<td>√</td>
<td>√ √ √</td>
</tr>
</tbody>
</table>
Summary

- SQF is a fairness measure that …
  - Bridges the gap between the ‘natural’ waiting time measure and the expected slowdown criterion
  - Reacts to both seniority and size
  - Applicable to complex systems

Thanks you!

Questions?