Trade-offs in Different Modeling Approaches for Performance Prediction of Software Systems

Samuel Kounev, Fabian Brosig, Philipp Meier, Steffen Becker, Anne Koziolek, Heiko Koziolek and Piotr Rygielski

Slides available at http://descartes.tools

Software Engineering 2016, Vienna, February 24, 2016
Main References


Research Context

Software Performance Engineering

Performance = timing behavior + resource usage

Software Performance Engineering (SPE)

“the entire collection of software engineering activities and related analyses used throughout the software development cycle, which are directed to meeting performance requirements.”
Performance-Relevant Concerns Spanning the Software Lifecycle

Response time of service S?
Utilization of server N?
Most suitable architecture?
Performance anti-patterns?
What if changes?
Performance regressions?
SLAs satisfied?

Response time of service S?
...
Performance Prediction

Performance-annotated Software Architecture Model

Transformation to Stochastic Performance Model

Predicted Performance Metrics
Stochastic Performance Models

Statistical Regression

Markov Model

Queueing Network (QN)

Queueing Petri net

Analytical Bounds Analysis

Layered Queueing Network (LQN)
Model Solution

Analytical Analysis

\[
R \geq \max \left[ N \times \max \{D_i\}, \frac{K}{\sum_{i=1}^{K} D_i} \right] \quad X_0 \leq \min \left[ \frac{1}{\max \{D_i\}}, \frac{N}{\sum_{i=1}^{K} D_i} \right]
\]

\[
\frac{N}{\max \{D_i\}[K+N-1]} \leq X_0 \leq \frac{N}{\text{avg}(D_i)[K+N-1]}
\]

Analysis Results

Simulative Analysis

Analysis Results
Modeling Challenges

Which modeling language?

Choice of Modeling Approach

Which modeling granularity?

Analytical solution? Simulation?

Trade-Offs: Accuracy vs. Overhead

Choice of Model Solution Approach

Performance Analyst

Context ➤ Modeling Approaches ➤ Trade-Offs ➤ Outlook
Considered Prediction Approaches

Performance-annotated Software Architecture Model (PCM) (Sec. 2)

Transformation: PCM2SimuCom (Sec. 3.1)

- SimuCom Model

  Model Solution: SimuCom

Transformation: PCM2QPN (Sec. 3.3)

- Queueing Petri Net

  Model Solution: SimQPN, SimQPN-MV

Transformation: PCM2LQQN (Sec. 3.5)

- Layered Queueing Network

  Model Solution: LQNS
Case Studies

Media Store

- Application Server
  - WebGUI
  - MediaStore
- MySQL Database Server
  - Digital Watermarking
  - Database Cache
- Media Database

SPECjEnterprise2010

- Oracle WebLogic Server
  - Delegate Work Order Session
  - Work Order Session
- Oracle Database Server
  - JDBC Driver
  - JDBC Connection Pool

Process Control System

- Server 1:
  - C4
  - C5
  - C6
  - C7
  - C8
  - C9
  - C10
  - C11
  - C3
  - C2
- Server 2:
  - C12
  - C13
  - C14
- Server 3:
  - C17
  - C18
  - C19
  - C20
  - C21
  - C22
  - C23
  - C24
  - C25
  - C26
  - C27

Business Reporting System

- Server 1:
  - Tomcat Webserver
- Server 2:
  - Scheduler
- Server 3:
  - Database
  - Core Graphics Engine
  - Core Online Engine
  - Cache

- 1 request/sec
- 15 requests/sec
- 150 requests/sec
- 48 requests/sec
- 48 request/sec
- 48 request/sec
- 3 CPU cores
## Semantic Gaps

<table>
<thead>
<tr>
<th>Modeling feature</th>
<th>SimuCom</th>
<th>SimQPN</th>
<th>SimQPN-MV</th>
<th>LQNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loops</td>
<td>X</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
</tr>
<tr>
<td>Forks with synchronization barrier</td>
<td>X</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
</tr>
<tr>
<td>Parameter dependencies</td>
<td>X</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
</tr>
<tr>
<td>Response time distributions</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Flexible parameter characterizations</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>Blocking behavior</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>(X)</td>
</tr>
</tbody>
</table>

X support (X) partial support - no support

**TABLE 1**

Semantic gaps.
# Quantitative Evaluation: Accuracy

<table>
<thead>
<tr>
<th>Media Store</th>
<th>SimuCom (reference)</th>
<th>SimQPN</th>
<th>SimQPN (relDiff)</th>
<th>SimQPN-MV</th>
<th>SimQPN-MV (relDiff)</th>
<th>LQNS</th>
<th>LQNS (relDiff)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT(Scenario1)</td>
<td>1.332</td>
<td>1.331</td>
<td>0.0%</td>
<td>1.324</td>
<td>-0.6%</td>
<td>1.288</td>
<td>-3.3%</td>
</tr>
<tr>
<td>TP(Scenario1)</td>
<td>0.751</td>
<td>0.751</td>
<td>0.0%</td>
<td>0.755</td>
<td>0.6%</td>
<td>0.753</td>
<td>0.3%</td>
</tr>
<tr>
<td>U(AppServer_CPU)</td>
<td>0.341</td>
<td>0.345</td>
<td>1.2%</td>
<td>0.340</td>
<td>-0.3%</td>
<td>0.342</td>
<td>0.2%</td>
</tr>
<tr>
<td>U(DBServer_CPU)</td>
<td>1.4%</td>
<td>1.4%</td>
<td>1.6%</td>
<td>1.4%</td>
<td>2.6%</td>
<td>1.4%</td>
<td>2.2%</td>
</tr>
<tr>
<td>U(DBServer_HDD)</td>
<td>64.4%</td>
<td>64.0%</td>
<td>-0.6%</td>
<td>64.5%</td>
<td>0.2%</td>
<td>64.4%</td>
<td>-0.1%</td>
</tr>
<tr>
<td>SPECjEnterprise2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT(Scenario1)</td>
<td>1.060</td>
<td>1.058</td>
<td>-0.2%</td>
<td>1.065</td>
<td>0.4%</td>
<td>1.065</td>
<td>0.4%</td>
</tr>
<tr>
<td>TP(Scenario1)</td>
<td>24.567</td>
<td>24.970</td>
<td>1.6%</td>
<td>24.792</td>
<td>0.9%</td>
<td>25.000</td>
<td>1.8%</td>
</tr>
<tr>
<td>U(Oracle_CPU)</td>
<td>29.1%</td>
<td>29.6%</td>
<td>1.7%</td>
<td>29.5%</td>
<td>1.3%</td>
<td>29.6%</td>
<td>1.8%</td>
</tr>
<tr>
<td>U(WLS_CPU)</td>
<td>51.2%</td>
<td>52.0%</td>
<td>1.5%</td>
<td>51.9%</td>
<td>1.3%</td>
<td>52.2%</td>
<td>2.0%</td>
</tr>
<tr>
<td>Process Control System</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT(Scenario1)</td>
<td>0.00787</td>
<td>0.00764</td>
<td>-2.9%</td>
<td>0.00763</td>
<td>-3.0%</td>
<td>0.00790</td>
<td>0.4%</td>
</tr>
<tr>
<td>RT(Scenario2)</td>
<td>0.06690</td>
<td>0.06636</td>
<td>-0.8%</td>
<td>0.06903</td>
<td>3.2%</td>
<td>0.06710</td>
<td>0.3%</td>
</tr>
<tr>
<td>TP(Scenario1)</td>
<td>149.258</td>
<td>149.254</td>
<td>0.0%</td>
<td>149.254</td>
<td>0.0%</td>
<td>149.254</td>
<td>0.0%</td>
</tr>
<tr>
<td>TP(Scenario2)</td>
<td>15.001</td>
<td>15.000</td>
<td>0.0%</td>
<td>15.000</td>
<td>0.0%</td>
<td>14.999</td>
<td>0.0%</td>
</tr>
<tr>
<td>U(Server1_CPU)</td>
<td>6.5%</td>
<td>6.4%</td>
<td>-2.2%</td>
<td>6.4%</td>
<td>-2.0%</td>
<td>6.5%</td>
<td>-0.5%</td>
</tr>
<tr>
<td>U(Server1_HDD)</td>
<td>45.0%</td>
<td>45.0%</td>
<td>0.0%</td>
<td>45.0%</td>
<td>0.0%</td>
<td>45.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>U(Server2_CPU)</td>
<td>1.1%</td>
<td>1.1%</td>
<td>-4.0%</td>
<td>1.1%</td>
<td>-4.0%</td>
<td>1.1%</td>
<td>-4.0%</td>
</tr>
<tr>
<td>U(Server3_CPU)</td>
<td>55.0%</td>
<td>54.1%</td>
<td>-1.6%</td>
<td>54.0%</td>
<td>-1.8%</td>
<td>55.1%</td>
<td>0.2%</td>
</tr>
<tr>
<td>Business Reporting System</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT(Scenario1 - cons1)</td>
<td>5.223</td>
<td>5.918</td>
<td>13.3%</td>
<td>5.921</td>
<td>13.4%</td>
<td>7.391</td>
<td>41.5%</td>
</tr>
<tr>
<td>TP(Scenario1 - cons1)</td>
<td>1.000</td>
<td>1.000</td>
<td>0.0%</td>
<td>1.000</td>
<td>0.0%</td>
<td>1.000</td>
<td>0.0%</td>
</tr>
<tr>
<td>U(Server1_CPU) Scen. 1</td>
<td>45.0%</td>
<td>44.8%</td>
<td>-0.5%</td>
<td>44.9%</td>
<td>-0.2%</td>
<td>45.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>U(Server2_CPU) Scen. 1</td>
<td>68.8%</td>
<td>68.6%</td>
<td>-0.3%</td>
<td>68.8%</td>
<td>-0.1%</td>
<td>68.9%</td>
<td>0.1%</td>
</tr>
<tr>
<td>U(Server3_CPU) Scen. 1</td>
<td>74.3%</td>
<td>74.8%</td>
<td>0.7%</td>
<td>74.5%</td>
<td>0.3%</td>
<td>74.3%</td>
<td>0.0%</td>
</tr>
<tr>
<td>U(Server4_CPU) Scen. 1</td>
<td>23.5%</td>
<td>23.5%</td>
<td>-0.2%</td>
<td>23.5%</td>
<td>0.1%</td>
<td>23.5%</td>
<td>0.2%</td>
</tr>
<tr>
<td>RT(Scenario2 - Exp(1))</td>
<td>7.302</td>
<td>7.256</td>
<td>-0.6%</td>
<td>7.500</td>
<td>2.7%</td>
<td>7.391</td>
<td>1.2%</td>
</tr>
<tr>
<td>TP(Scenario2 - Exp(1))</td>
<td>1.002</td>
<td>1.000</td>
<td>-0.2%</td>
<td>0.996</td>
<td>-0.6%</td>
<td>1.000</td>
<td>-0.2%</td>
</tr>
</tbody>
</table>

**KEY:** RT = Response Time (sec), TP = Throughput (requests/sec), U = Utilization (%), relDiff = relative difference
## Quantitative Evaluation: Overhead

<table>
<thead>
<tr>
<th></th>
<th>SimuCom (reference)</th>
<th>SimQPN (relDiff)</th>
<th>SimQPN-MV (relDiff)</th>
<th>LQNS (relDiff)</th>
<th>LQNS (relDiff)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Media Store</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logical Simulation Time</td>
<td>25000</td>
<td>25000</td>
<td>25000</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Number of Measurements</td>
<td>18773</td>
<td>18775</td>
<td>18775</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Total Execution Time</td>
<td>31.4</td>
<td>6.4</td>
<td>3.6</td>
<td>1.3</td>
<td>-96.0</td>
</tr>
<tr>
<td>Wall-clock Sim./Ana. Time</td>
<td>25.6</td>
<td>1.7</td>
<td>0.9</td>
<td>0.2</td>
<td>-99.1</td>
</tr>
<tr>
<td><strong>SPECjEnterprise2010</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logical Simulation Time</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Number of Measurements</td>
<td>12257</td>
<td>12485</td>
<td>12485</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Total Execution Time</td>
<td>72.6</td>
<td>7.9</td>
<td>5.6</td>
<td>1.8</td>
<td>-97.6</td>
</tr>
<tr>
<td>Wall-clock Sim./Ana. Time</td>
<td>55.6</td>
<td>3.8</td>
<td>2.1</td>
<td>0.5</td>
<td>-99.2</td>
</tr>
<tr>
<td><strong>Process Control System</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logical Simulation Time</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Number of Measurements</td>
<td>74627</td>
<td>74627</td>
<td>74627</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Total Execution Time</td>
<td>96.7</td>
<td>11.9</td>
<td>7.2</td>
<td>3.9</td>
<td>-95.9</td>
</tr>
<tr>
<td>Wall-clock Sim./Ana. Time</td>
<td>65.4</td>
<td>1.6</td>
<td>2.6</td>
<td>2.0</td>
<td>-96.9</td>
</tr>
<tr>
<td><strong>Business Reporting System</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logical Simulation Time</td>
<td>9994</td>
<td>10000</td>
<td>10000</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Number of Measurements</td>
<td>10016</td>
<td>9997</td>
<td>9997</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Total Execution Time</td>
<td>591.5</td>
<td>74.5</td>
<td>34.1</td>
<td>4.0</td>
<td>-99.3</td>
</tr>
<tr>
<td>Wall-clock Sim./Ana. Time</td>
<td>587.6</td>
<td>35.4</td>
<td>26.6</td>
<td>1.8</td>
<td>-99.7</td>
</tr>
</tbody>
</table>

All times in seconds

---

13 S. Kounev

Context  ➔  Modeling Approaches  ➔  Trade-Offs  ➔  Outlook
## Quantitative Evaluation: Impact of Load

<table>
<thead>
<tr>
<th></th>
<th>SimuCom</th>
<th>SimQPN</th>
<th>SimQPN-MV</th>
<th>LQNS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Media Store (low load)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Execution Time</td>
<td>21.5</td>
<td>5.7</td>
<td>3.4</td>
<td>1.3</td>
</tr>
<tr>
<td>Wall-clock Sim./Ana. Time</td>
<td>16.0</td>
<td>1.1</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td><strong>Media Store (medium load)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Execution Time</td>
<td>31.4</td>
<td>6.4</td>
<td>3.6</td>
<td>1.3</td>
</tr>
<tr>
<td>Wall-clock Sim./Ana. Time</td>
<td>25.6</td>
<td>1.7</td>
<td>0.9</td>
<td>0.2</td>
</tr>
<tr>
<td><strong>Media Store (high load)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Execution Time</td>
<td>40.7</td>
<td>7.3</td>
<td>4.0</td>
<td>1.4</td>
</tr>
<tr>
<td>Wall-clock Sim./Ana. Time</td>
<td>34.1</td>
<td>2.4</td>
<td>1.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

*All times in seconds*
Decision Tree (simplified)

- **Prediction overhead is an issue?**
  - no → Use SimuCom
  - yes
    - **Response time distributions required?**
      - no → Use LQNS
      - yes
        - **Maximum prediction accuracy required?**
          - no → Use SimQPN-MV
          - yes
            - **Solution time must be minimized?**
              - yes → Use SimQPN
              - no → Use SimuCom

- **Accurate representation of low-level platform overheads is important?**
  - yes → Run ProtoCom as a final validation step
  - no
Decision Tree

Source model contains loops?
  yes
  no
  Approximated loop iteration numbers lead to significant inaccuracies?
    yes
    no
    Use SimuCom

Source model contains forks with synchronization barrier?
  yes
  no
  Approximated fork-join behavior leads to significant inaccuracies?
    yes
    no
    Use SimQPN

Source model contains parameter dependencies?
  yes
  no
  Approximated parameter propagation leads to significant inaccuracies?
    yes
    no
    Use SimQPN-MV

Source model contains parameter dependencies?
  yes
  no
  Approximated fork-join behavior leads to significant inaccuracies?
    yes
    no
    Use LQNS

Response time distributions or response time percentiles required?
  yes
  no
  Source model contains flexible parameter characterizations?
    yes
    no
    Approximation with exponential distribution leads to significant inaccuracies?
      yes
      no
      Use SimQPN-MV

Source model contains blocking behavior?
  yes
  no
  Blocking behavior is captured by LQN task multiplicity?
    yes
    no
    Use LQNS

Approximation with exponential distribution leads to significant inaccuracies?
  yes
  no
  Use SimQPN-MV

Use SimQPN

Approximated fork-join behavior leads to significant inaccuracies?
  yes
  no
  Use SimQPN

Use LQNS

Decision Trees are used in decision-making processes to provide a visual representation of the decision-making process. They help in identifying the best course of action by mapping out all possible outcomes and the conditions under which they occur. In this context, the decision tree is used to determine the appropriate modeling approach based on the presence of loops, synchronization barriers, parameter dependencies, and the need for response time distributions or percentiles.
DECLARE Project
Declarative Performance Engineering

DFG Priority Programme 1593

Project PIs
Dr.-Ing. André van Hoorn (Prof.-Vertr.), University of Stuttgart
Prof. Dr.-Ing. Samuel Kounev, University of Würzburg

Members
Dr.-Ing. Dušan Okanović, University of Stuttgart
Dipl.-Inform. Jürgen Walter, University of Würzburg

Associated Partners
Capgemini Deutschland GmbH, Stuttgart, Germany
Software Performance Engineering

Established Methods, Techniques, and Tools

Asking “What?“, Automating the “How?“

Declarative Performance Engineering (DPE)

Declarative Performance Engineering

Performance Concerns

What would be the response times of services X, Y, and Z if the workload intensity doubles over the next week? Rough estimation is sufficient.

Performance Analyst

The resulting response times are...

Software Performance Engineering

Results

Service X: 4ms
Service Y: 5ms
Service Z: 9ms
Resource A: 20%
Resource B: 50%

Analysis Model and Solution Approach

Performance-annotated Software Architecture Model

Asking “What?“, Automating the “How?“

Declarative Performance Engineering (DPE)

References


Questions?

skounev@acm.org

http://descartes.tools