LibReDE – Library for Resource Demand Estimation

SPEC DevOps Performance Working Group
Simon Spinner
25.07.2014
What are resource demands?

Example SEFF in PCM:

A **resource demand** is the time a unit of work (e.g., request or transaction) spends obtaining service from a resource (e.g., CPU or hard disk) in a system.
### Direct Measurement

Requires specialized infrastructure to monitor low-level statistics.

Examples:
- Brunnert et al. [4]
- Magpie [1]

### Estimation

Use of statistical techniques on high-level monitoring statistics.

Examples:
- Linear regression [5-8]
- Kalman filtering [9-11]
- Nonlinear optimization [12-14]
- Maximum likelihood estimation [7] [15]
- Gibbs sampling [16]
- Independent Component Analysis [17]
Why should I use estimation techniques?

- Limitations of monitoring and instrumentation tools
  - CPU time accounting not possible for individual requests
  - CPU time accounting imprecise
  - Fine-granular control flow not available

- Heterogeneous environments
  - Requests is processed in different software stacks
  - Unaccounted work in system or background threads

- Virtualized environments
  - CPU accounting in guests may be wrong
Example

- Least squares regression based on Utilization Law

- Known measurable
  - $U_i$ average utilization in measurement period $i$
  - $X_{i,c}$ average throughput of workload class $c$ in measurement period $i$

- Resource demand $D_c$ of workload class $c$

- Utilization Law for $C$ workload classes:
  $$U_i = X_{i,1} \cdot D_1 + \cdots + X_{i,c} \cdot D_c$$
LibReDE

- Library for Resource Demand Estimation

- Ready-to-use Java implementations of
  - Least-squares regression
  - Kalman filter (2 variants)
  - Non-linear optimization (2 variants)
  - Service Demand Law
  - Response time approximation

References

Estimation process

1. Define estimation problem
2. Select estimation approaches
3. Load monitoring data
4. Check pre-conditions
5. Run estimation approach(es)
6. Cross-Validation
7. Compare estimation accuracy
8. Output results
LibReDE usage

- Offline analysis (Java or Matlab)
  - Measurement traces
  - Standalone application
  - Estimated Demands

- Online analysis (Java)
  - Monitoring tools
  - Custom application
  - LibReDE

Offline analysis:
- .csv files
- Matlab array

Online analysis:
- Custom application
- LibReDE
Example

code:

```c
repository = librede_init({'WC0', 'WC1', 'WC2'}, {'CPU0'});

librede_load_data(repository, 'utilization', 'CPU0', ts, util, 60);
librede_load_data(repository, 'response_time', {'WC0', 'WC1', 'WC2'}, ts, rt, 60);
librede_load_data(repository, 'throughput', {'WC0', 'WC1', 'WC2'}, ts, tput, 60);

[approaches, estimates, relErrUtil, relErrResp] =
    librede_run(repository, ts(1), ts(end), 60, 30);
```

![Bar chart showing mean relative response time error for different approaches and workloads.](chart.png)

Approach
Mean relative response time error

<table>
<thead>
<tr>
<th>Approach</th>
<th>WC0</th>
<th>WC1</th>
<th>WC2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ServiceDemandLaw</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoliaRegression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZhangKalmanFilter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MenasceOptimization</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Work-In-Progress

- GUI to define estimation problem
  - Resources and services
  - Sources of measurement data
  - Configuration of estimation approaches
  - Configuration of validation

- Automatic parameterization of performance models
  - Bridges to DML, QPME, PCM
  - Use performance models for validation

- Additional estimation approaches [7], [15-16]

- Integration with Kieker?
Case studies (1/3): SPECjEnterprise2010

- Extraction of PCM models (all domains)
- Monitoring
  - WebLogic Diagnostics Framework (WLDF) → Response times
  - Operating system → CPU utilization
- Resource demand estimation
  - Response time approximation
  - Service Demand Law

References

Case studies (2/3): Multi-tenant applications

- Admission control of requests based on estimated resource demands
  - Performance isolation
  - QoS differentiation
- Multi-tenant TPC-W in SAP HANA Cloud
- Includes evaluation of resource demand estimators for high number of workload classes

References

Case studies (3/3): Zimbra Server

Number of vCPUs

VM1  VM2  VM3

Hypervisor

End-to-end response time, throughput
Application statistics
CPU utilization, scheduling delays
Hypervisor statistics

vCPU controller

Resource Demands

LibReDE

Application SLO

References

LibReDE

- License: Eclipse Public License (EPL)

- Planned submission to SPEC tools repository (in August)
  - Binaries (Windows/Linux)
  - User guide/tutorial

- Source code available on Bitbucket:
  - https://bitbucket.org/librede/librede
References (1/2)

References