COMPARATIVE ANALYSIS OF FRAMEWORKS FOR THE PERFORMANCE EVALUATION OF MULTI-TIER CLOUD APPLICATIONS

Godofredo R. Garay\textsuperscript{1}, Andrei Tchernykh\textsuperscript{2}, A. Yu. Drozdov\textsuperscript{3}

\textsuperscript{1}University of Camaguey, Camaguey, Cuba, godofredo.garay@gmail.com
\textsuperscript{2}CICESE Research Center, Ensenada, México, chernykh@cicese.mx
\textsuperscript{3}Moscow Institute of Physics and Technology, Moscow, Russia, alexander.y.drozdov@gmail.com

Cloud computing, Education, Research, Development

Moscow, December, 2015
Content

- Objective
- Motivation
- Analytical frameworks review
- Modular Performance Analysis (MPA) with Real-Time Calculus (RTC)
  - RTC Fundamentals
  - RTC model calibration
- Comparison of analytical approaches
- Conclusion
Objective

• We address the performance evaluation of multi-tier clouds applications
• We compare a Real-Time Calculus-based framework with two classical analytical approaches such as queuing theoretic approaches and control theoretic approaches
• We focus on the capabilities of these alternatives for estimating the key Quality of Service parameter - the application response-time
Motivation

Imaginary example of a client session on a basic multi-tier application architecture (note that in virtualized cloud platforms, each software server, i.e., Apache, Tomcat, and MySQL, is run inside of a virtual machine).
Motivation

Focus of attention: Predicting Web-application response-time in cloud computing platform, e.g., does maximum request-to-response latency of a client data access request will not exceed application deadline (with 95% confidence interval)?
Analytical Frameworks Review

- Queuing models
- Control theory models
- Modular Performance Analysis with RTC
Modular Performance Analysis with RTC

• Deterministic analysis (Thiele et. al)
  – RTC belongs to the class of so-called deterministic queuing theories
  – RTC is deterministic in the sense that hard upper and lower bounds of the performance metrics (such as latency) can be always found

• Stochastic analysis (Garay, 2013)
  – Soft real-time guarantees, i.e., guarantees on delays and backlogs that are valid up to a certain level of confidence

RTC Fundamentals

• Arrival and Service Functions
• Arrival and Service Curves
• Worst-case analysis:
  – Maximum Backlog
  – Maximum delay
Arrival and Service Functions

• An event stream can be described by an **arrival function** $R$, where $R(t)$ denotes the number of events that have arrived in the interval $[0, t)$

• A computing or communication resource can be described by a **service function** $C$, where $C(t)$ denotes the number of events that could have been served in the interval $[0, t)$
Arrival and Service Curves

The upper and lower arrival curves, $\alpha^u(\Delta)$, $\alpha^l(\Delta) \in \mathbb{R}_{\geq 0}$ of an arrival function $R(t)$ satisfy the following inequality:

$$\alpha^l(t - s) \leq R(t) - R(s) \leq \alpha^u(t - s), \forall s, t : 0 \leq s \leq t$$
Arrival and Service Curves

The upper and lower service curves,

\[ \beta^u(\Delta), \beta^l(\Delta) \in \mathbb{R}^{\geq 0} \]

of a service function \( C(t) \) satisfy

\[ \beta^l(t - s) \leq C(t) - C(s) \leq \beta^u(t - s) \quad \forall \ s, t : 0 \leq s \leq t \]
Both, $\alpha_f^u$ arrival curve and $\beta_r^l$ service curve are bounding-functions and can be defined using a piecewise linear approximation.
Deriving the $\alpha_f^u$ and $\beta_r^l$ bounding-functions of the processing resource $r$.

RTC model parameters and our metric of interest ($D_{max}$).

Modeling the resource $r$ and obtaining its maximum request-response delay time ($D_{max}$) by using RTC.

$$\text{delay} \leq \sup_{t \geq 0} \{ \inf \{ \tau \geq 0 : \alpha_f^u(t) \leq \beta_r^l(t + \tau) \} \}$$
Modular Performance Analysis with RTC

\[ \beta^l = (((\beta^l_1 \otimes \beta^l_2) \otimes \beta^l_3) \otimes \ldots) \otimes \beta^l_n \]
RTT model calibration

Family of service curves corresponding to a system component with non-deterministic behavior (left part)

Procedure for obtaining its resultant bounding-curve (right part)

RTC model calibration

Deriving the parameters for constructing the $\beta_{ri}^l$ lower service curve of a concrete system component with non-deterministic behavior (e.g., a web, application or database server) from simulations or real traces may give the case where the following assumption holds

$$\exists i, \Delta : \beta_{ri}^l(\Delta) < \beta_{\{ri, reality\}}^l(\Delta)$$

where $i \in (1, 2, 3, \ldots)$, $\beta_{ri}^l$ is a resultant lower service curve derived from a set of lower service curves and $\beta_{\{ri, reality\}}^l(\Delta)$ is an unknown lower bounding-curve of the SUT for the stochastic component being considered.

For this reason, in (Garay, 2013), statistical methods are used in order to demonstrate that the values of the $L$ and $R$ parameters of $\beta_{ri}^l$ have an adequate level of predictability, and, hence, results are valid up to certain level of confidence.

### Comparison of analytical approaches

<table>
<thead>
<tr>
<th>Modeling capabilities</th>
<th>MPA-RTC</th>
<th>Queuing Theory</th>
<th>Control Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-tier cloud Web application</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Hard/Soft response time guarantees</td>
<td>Both</td>
<td>No</td>
<td>Soft guarantees</td>
</tr>
<tr>
<td>Workload models</td>
<td>Real and/or synthetic</td>
<td>Synthetic</td>
<td>Real or synthetic</td>
</tr>
<tr>
<td>Task processing models</td>
<td>Real and/or synthetic</td>
<td>Synthetic</td>
<td>Real or synthetic</td>
</tr>
<tr>
<td>VM provisioning</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>VMs performance interference effect</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Autonomic resource management</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Server consolidation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Horizontal/Vertical scaling</td>
<td>Both</td>
<td>Both</td>
<td>Both</td>
</tr>
</tbody>
</table>

In our paper, references to analytical studies based on queueing theory (QT) and control theory (CT) are given and a discussion on the modeling capabilities of each approach is presented.
Workload models

- Real workload traces
- Naive synthetic workload models (e.g., probability distributions)
- Realistic synthetic workload models
- Combinations of the previous alternatives
Modeling provisioning response delay

Provisioning response delay

Queuing
Provisioning Decision
VM Instantiation
VM deployment

Non-processing time interval
VMs performance interference effect

Contention
Access requests
PM resources
(CPU, memory, I/O, etc)
SLA Violations
Application

m > n
VM1
VM2 VMn
VM1
VM2 VMm
(a)
(b)
RTC-based autonomic resource management

Data Center Infrastructure (Target system)

Control actions (e.g., VM migrations)

HTTP requests model parameters \((p, r, M, b)\)

Software servers model parameters \((L, R)\)

RTC-based Predictive Controller

SLA
Monitoring Interval
Management Strategies

Measured values
VMs deployment scenarios

- **PM1** (Fastest machine)
  - VM1
  - VM2
  - VM3

- **PM2**
  - [OFF]

- **PM3** (Slowest machine)
  - [OFF]

---

Power consumption:

(a) VM1 VM2 VM3

(b) VM1 VM2 VM3

(c) VM1 VM2 VM3
Horizontal scaling

(a) VM1 \downarrow \quad VM2 \quad VM3

(b) VM1 \quad VM2 \quad VM3

VM migration

New server replica (Horizontal scaling)

Workload
Demand

Application
SLA Violations

Workload
Demand

Application
SLA Violations
Conclusion

• We discuss different approaches for modeling cloud-based systems
• We conclude that RTC is suitable framework for estimating statistical response time guarantees
• We consider that contemporary issues in cloud computing research could be analyzed by using MPA-RTC
Conclusion

• We discuss different approaches for modeling cloud-based systems
• We conclude that RTC is suitable framework for estimating statistical response time guarantees
• We consider that contemporary issues in cloud computing research could be analyzed by using MPA-RTC