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

Godofredo R. Garay[1], Andrei Tchernykh[2], A. Yu. Drozdov[3]

[1]University of Camaguey, Camaguey, Cuba, godofredo.garay@gmail.com

[2]CICESE Research Center, Ensenada, México, chernykh@cicese.mx

[3]Moscow Institute of Physics and Technology, Moscow, Russia, alexander.y.drozdov@gmail.com

Cloud computing, Education, Research, Development

Moscow, December, 2015

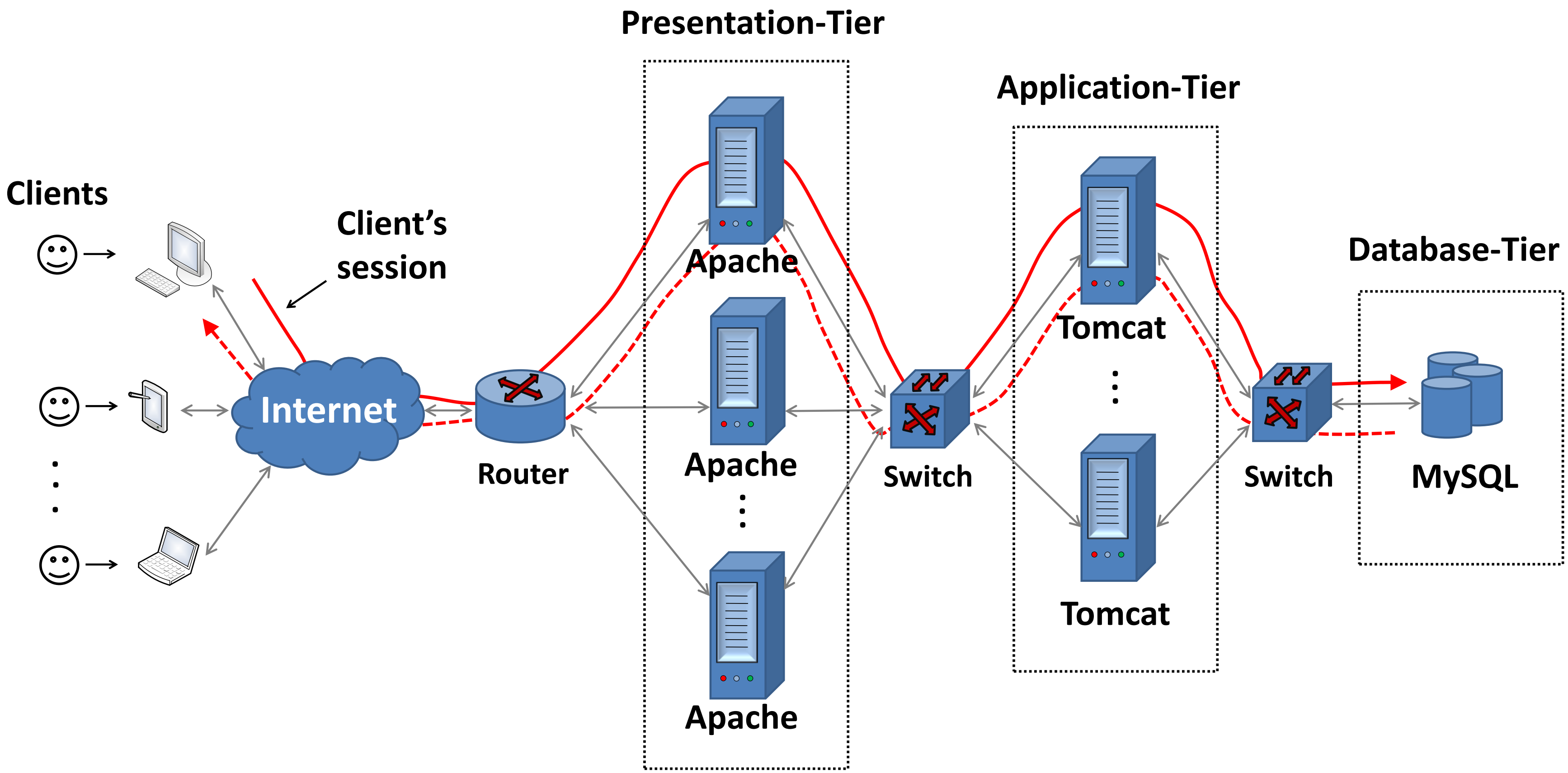
Content

- 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

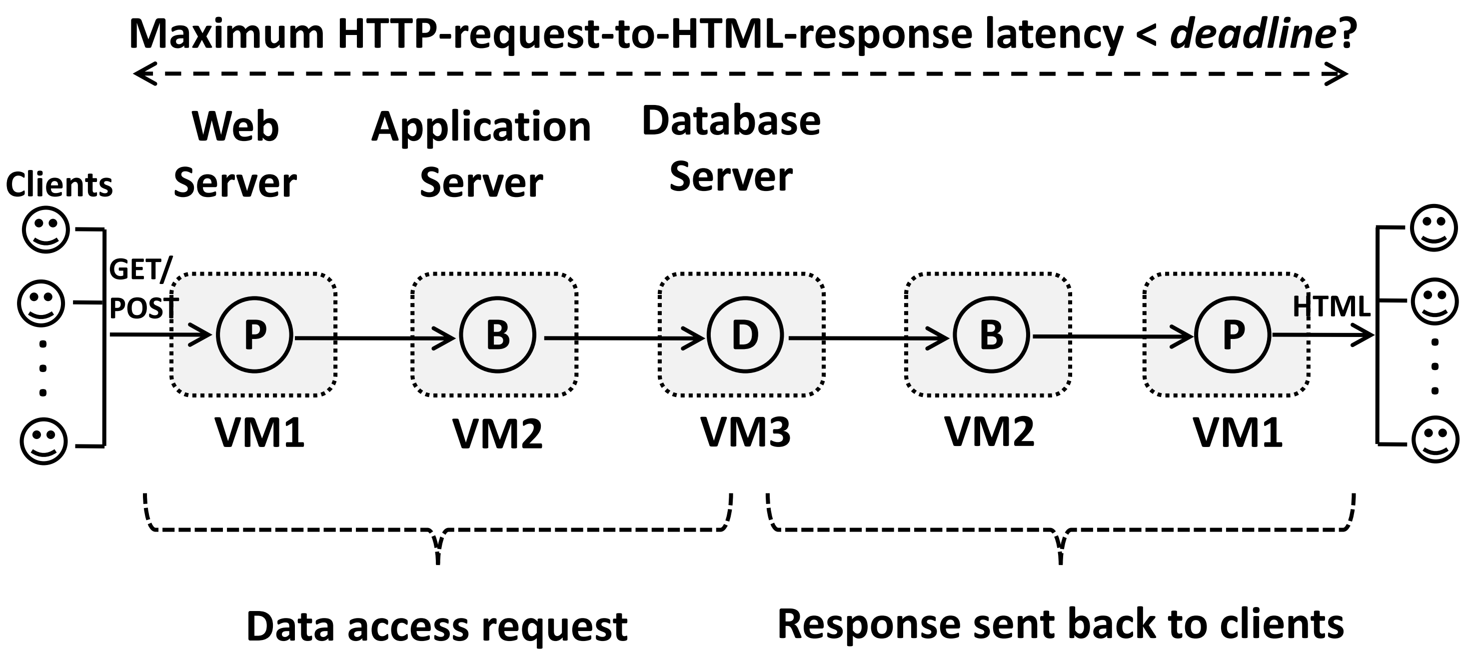
- We address the performance evaluation of multi-tier clouds applications
- We compare a Real-Time Calculus-based framework with two classical analytical approaches such as queuing theoretic approaches and control theoretic approaches
- We focus on the capabilities of these alternatives for estimating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

G. R. Garay, J. Ortega, A. F. Díaz, L. Corrales, and V. Alarcón-Aquino, "System performance evaluation by combining RTC and VHDL simulation: A case study on NICs," *Journal of Systems Architecture,* vol. 59, pp. 1277-1298, 2013.

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

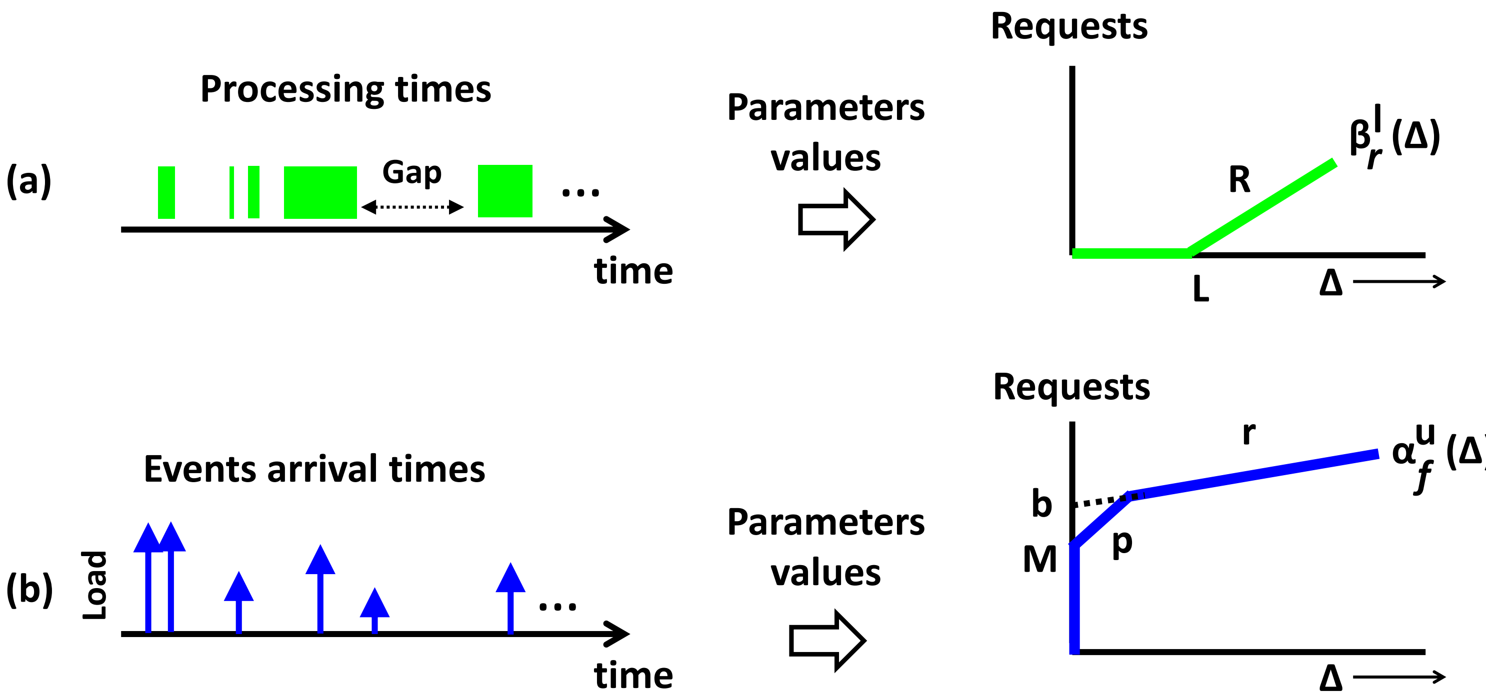
The **upper and lower service curves**,

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

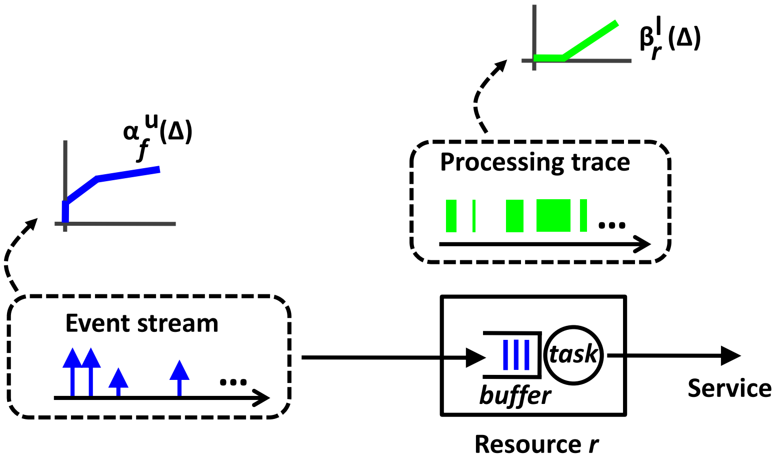
of a service function C(t) satisfy

$$\beta^l(t-s) \leq C(t) - C(s) \leq \beta^u(t-s) \ \forall s, t : 0 \leq s \leq t$$

Modular Performance Analysis with RTC

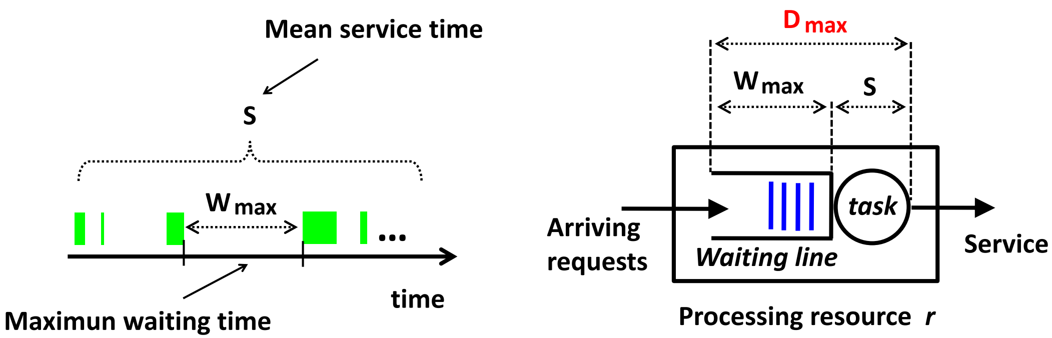


Both, α_f^u arrival curve and β_r^l service curve are **bounding-functions** and can be defined using a piecewise linear approximation



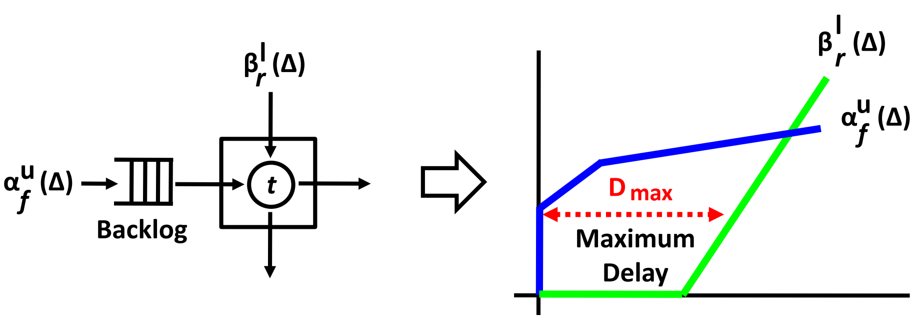
(a)

Deriving the α_f^u and β_r^l bounding-functions of the processing resource r.



(b)

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

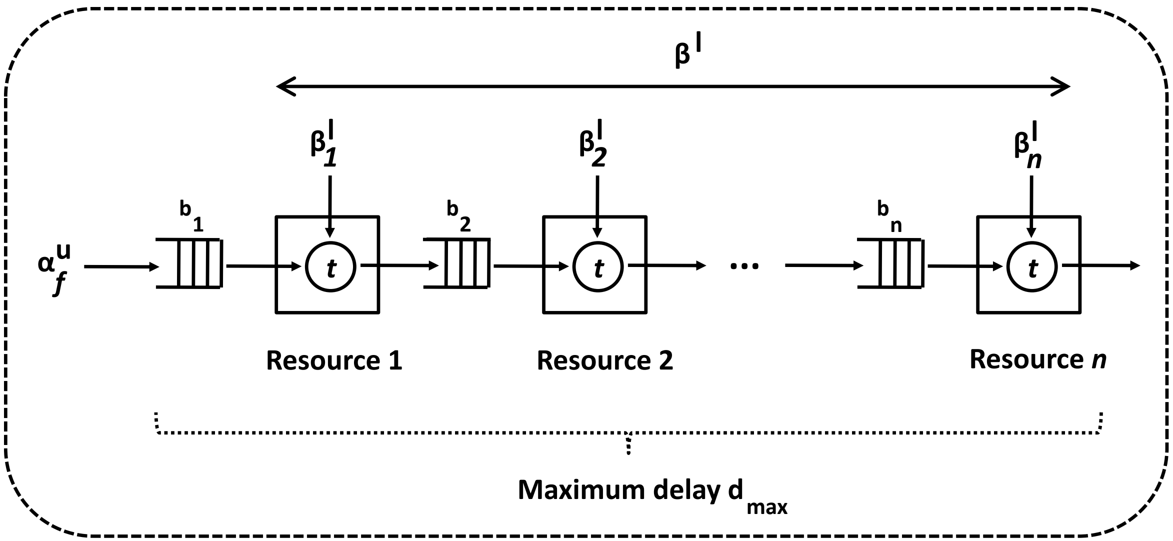


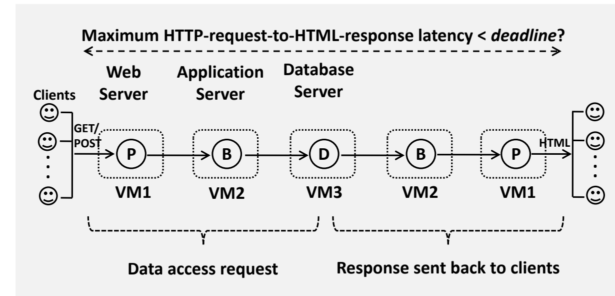
(c)

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

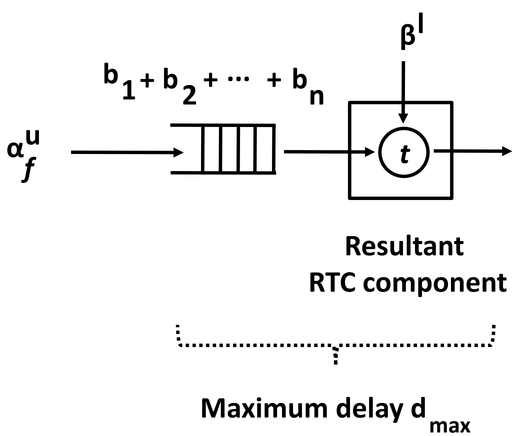
$$delay \leq sup_{t\geq 0}\{\inf\{\tau \geq 0 : \alpha_f^u(t) \leq \beta_r^l(t+\tau)\}\}$$

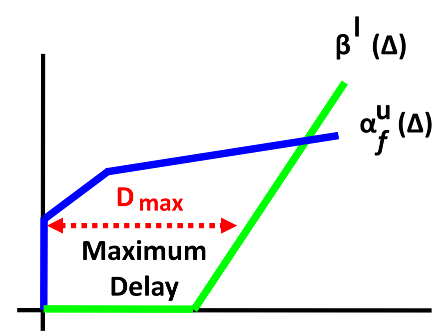
Modular Performance Analysis with RTC



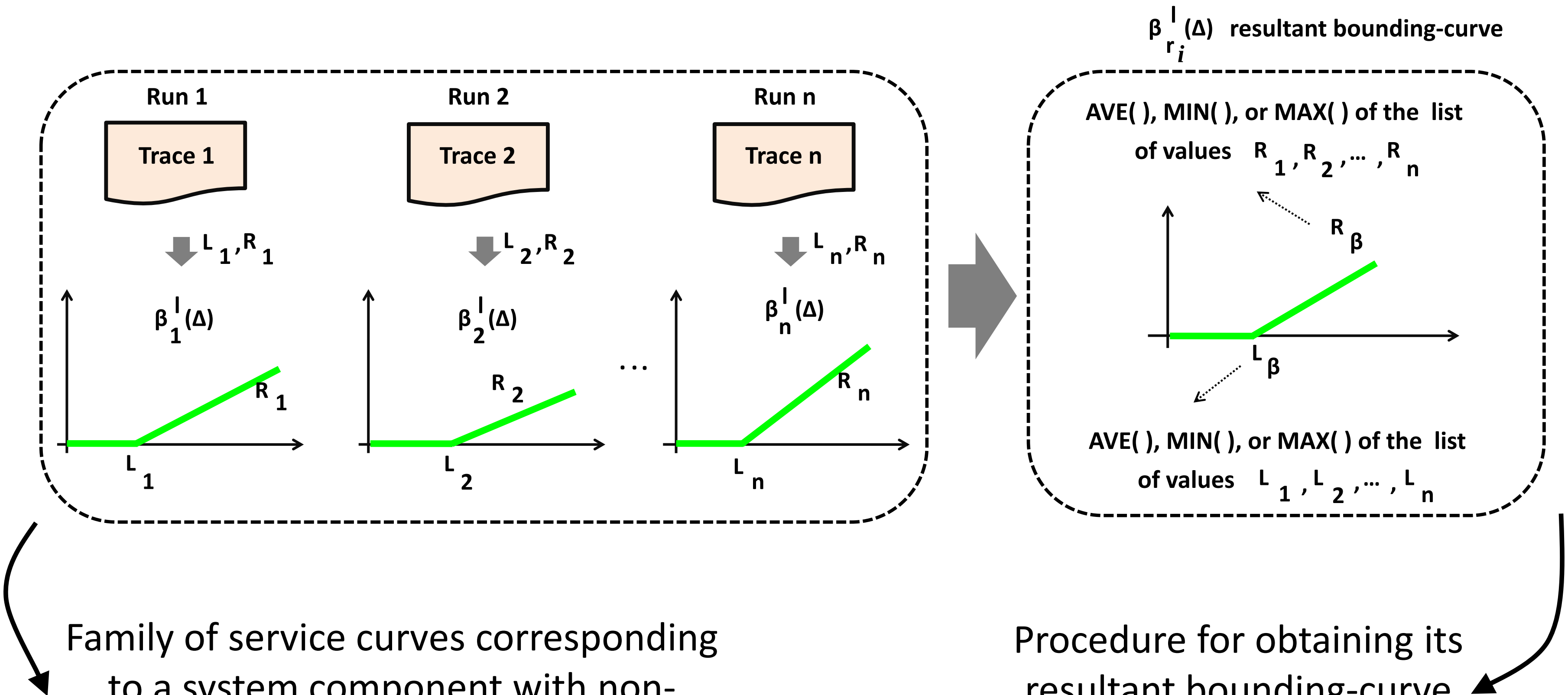


$$\beta^l = (((\beta_1^l \otimes \beta_2^l) \otimes \beta_3^l) \otimes \cdots) \otimes \beta_n^l$$





RTC 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)

G. R. Garay, J. Ortega, A. F. Díaz, L. Corrales, and V. Alarcón-Aquino, "System performance evaluation by combining RTC and VHDL simulation: A case study on NICs," *Journal of Systems Architecture,* vol. 59, pp. 1277-1298, 2013.

RTC model calibration

Deriving the parameters for constructing the $\beta_{r_i}^{l}$ 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

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

where $i \in (1, 2, 3, \dots)$, $\beta_{r_i}^{l}$ is a resultant lower service curve derived from a set of lower service curves and , $\beta_{\{r_i,reality\}}^{l}(\Delta)$ 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 $\beta_{r_i}^{l}$ have an adequate level of predictability, and, hence, results are valid up to certain level of confidence.

G. R. Garay, J. Ortega, A. F. Díaz, L. Corrales, and V. Alarcón-Aquino, "System performance evaluation by combining RTC and VHDL simulation: A case study on NICs," *Journal of Systems Architecture,* vol. 59, pp. 1277-1298, 2013.

Comparison of analytical approaches

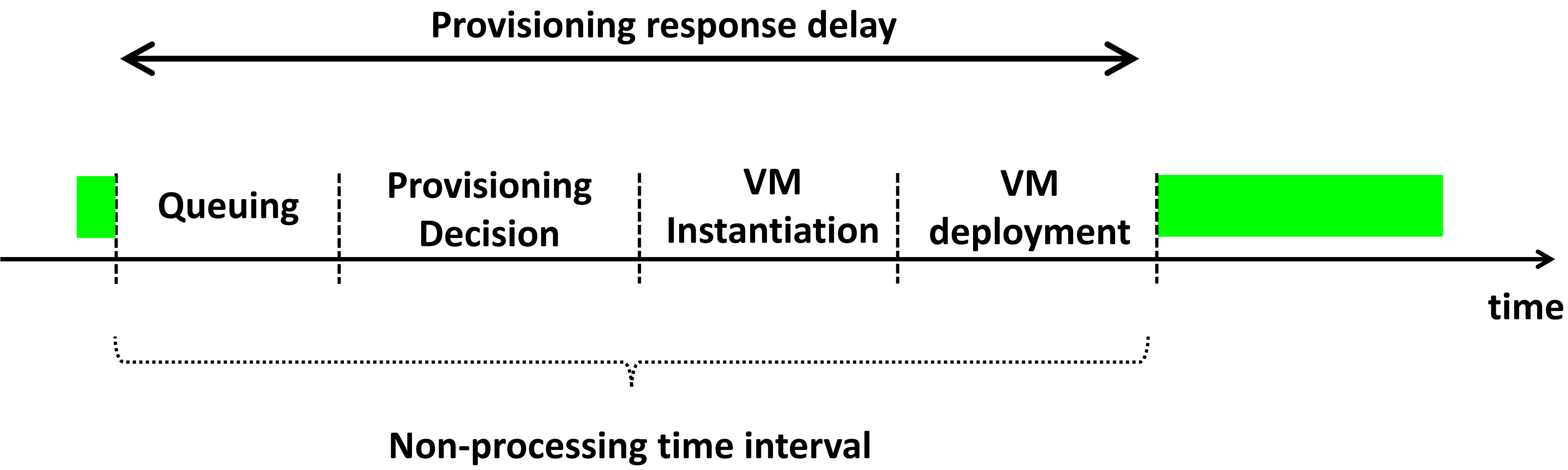
Modeling capabilities	MPA-RTC	Queuing Theory	Control Theory
Multi-tier cloud Web application	Yes	Yes	Yes
Hard/Soft response time guarantees	Both	No	Soft guarantees
Workload models	Real and/or synthetic	Synthetic	Real or synthetic
Task processing models	Real and/or synthetic	Synthetic	Real or synthetic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

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

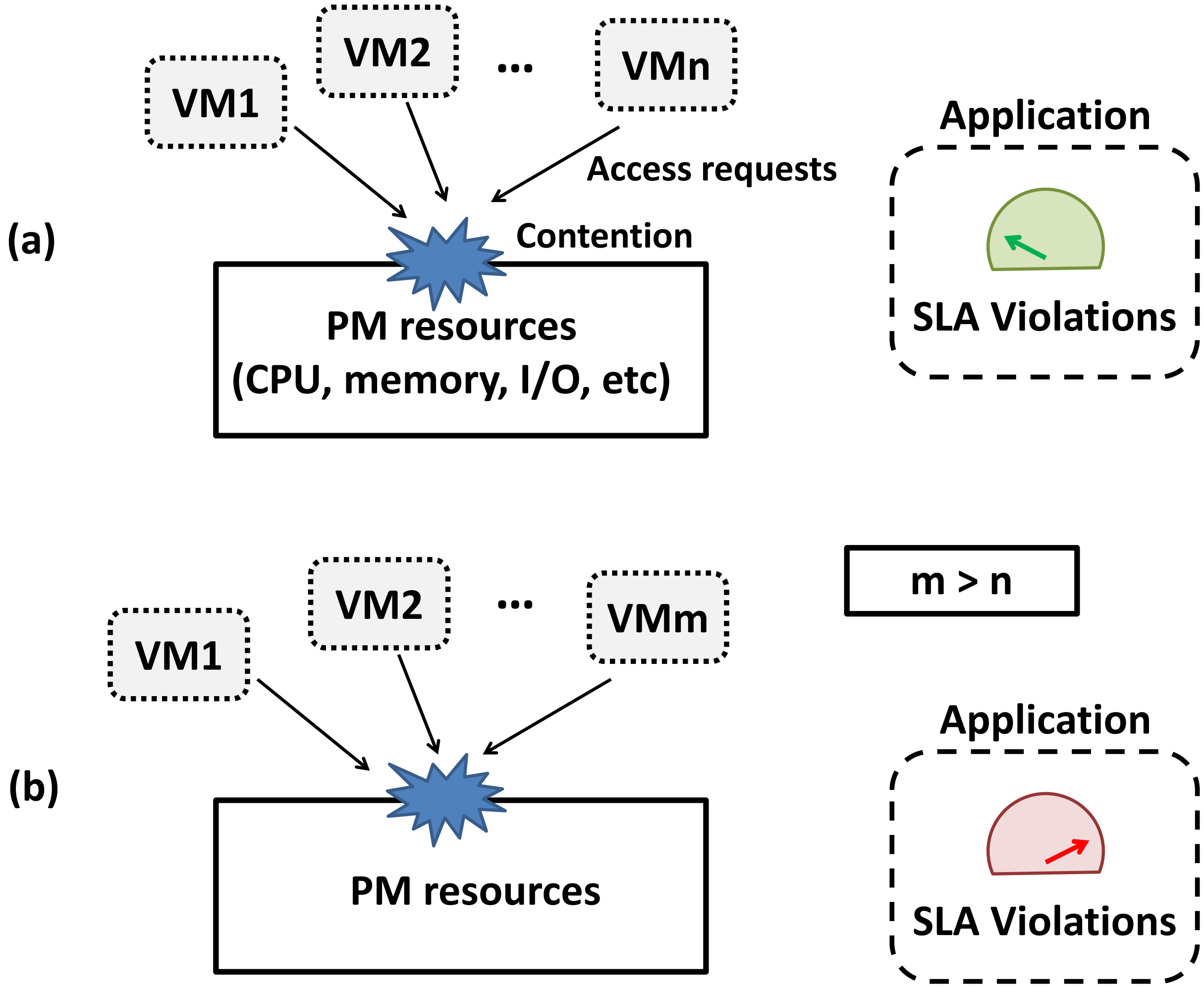
Workload models

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

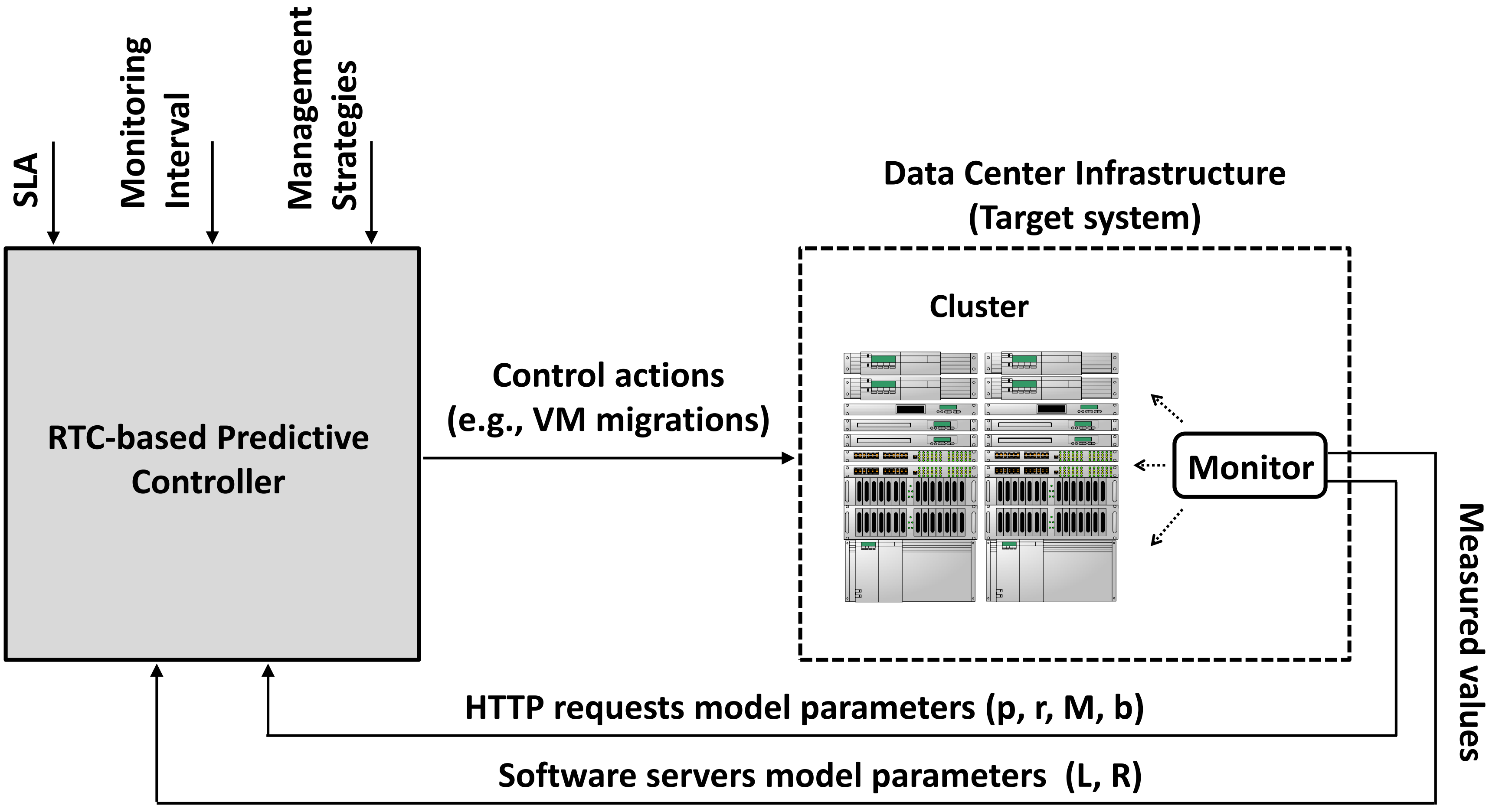
Modeling provisioning response delay



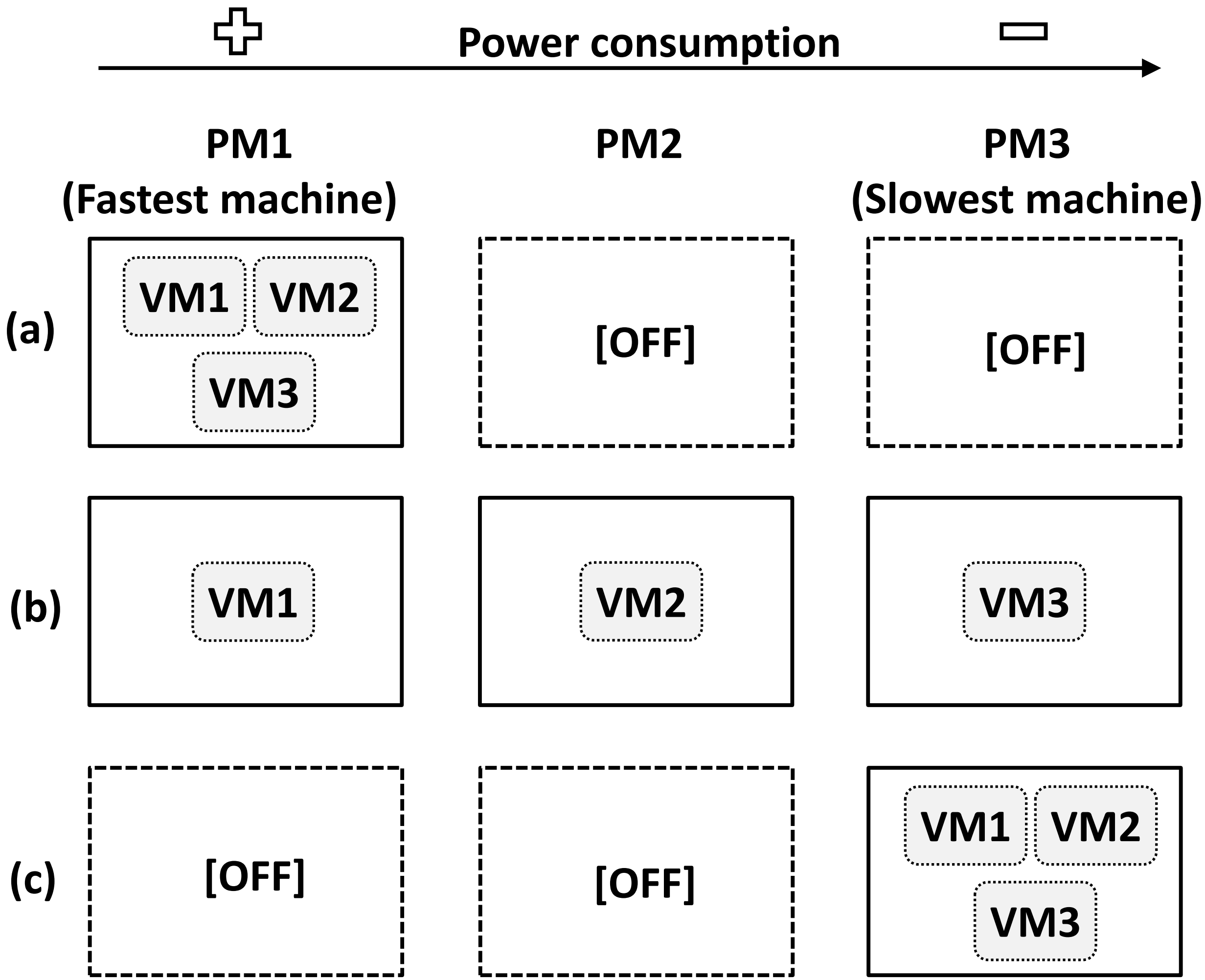
VMs performance interference effect



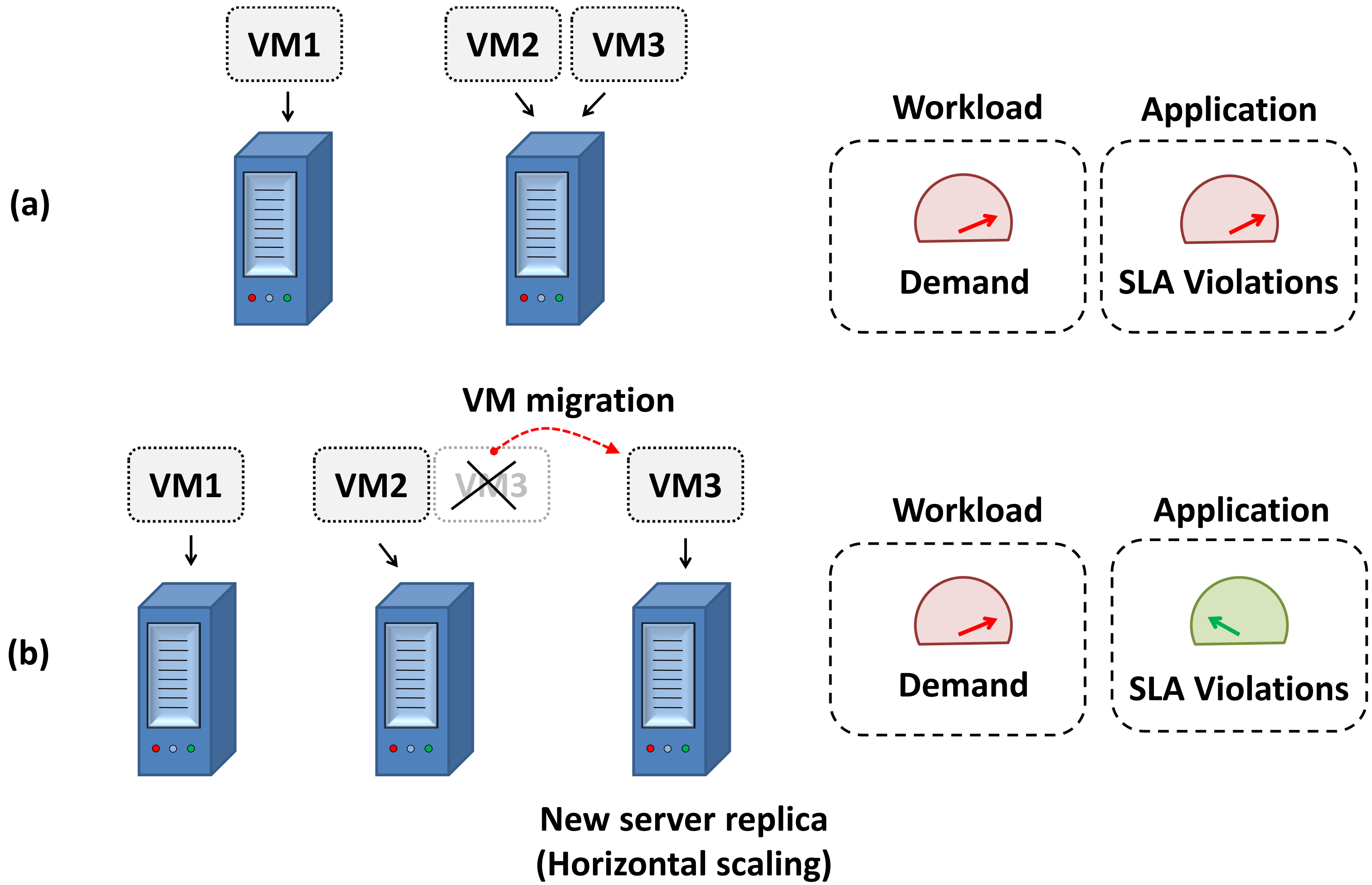
RTC-based autonomic resource management



VMs deployment scenarios



Horizontal scaling



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