Identification of availability and performance bottlenecks in cloud computing systems

Rubens de Souza Matos Júnior

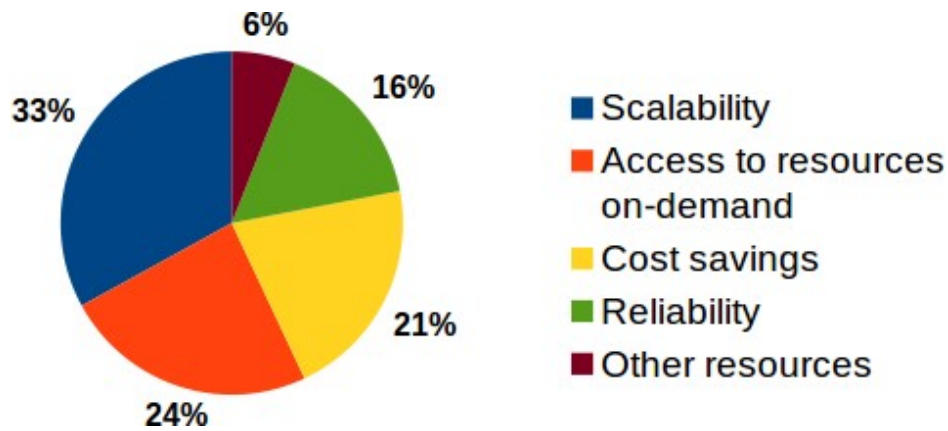
Orientador: Prof. Paulo Maciel
Co-orientador: Prof. Kishor S. Trivedi

Outline

- Motivation
- Background
- Related works
- Proposed methodology
- Case study: Scalable composite web service on private cloud
- Final remarks

Motivation

- Among the major reasons mentioned for **adoption of cloud computing** are:
  - Scalability
  - Access to resources on-demand
  - Cost savings
  - Reliability
- **Problems** in large cloud providers show the importance of proper availability and performance **planning** for cloud infrastructures and their hosted services.



Source: rightscale.com


Home / Reviews / Software / Security / **Amazon Cloud Outage Hits Netflix, Foursquare**

Amazon Cloud Outage Hits Netflix, Foursquare

BY CHLOE ALBANESIUS AUGUST 9, 2011 11:14AM EST 6 COMMENTS

In the same week that a lightning strike in Dublin knocked out service for some European users of A Microsoft's cloud services, Amazon also suffered a stateside cloud outage that affected popular serv Foursquare, Reddit, and Netflix.

Motivation

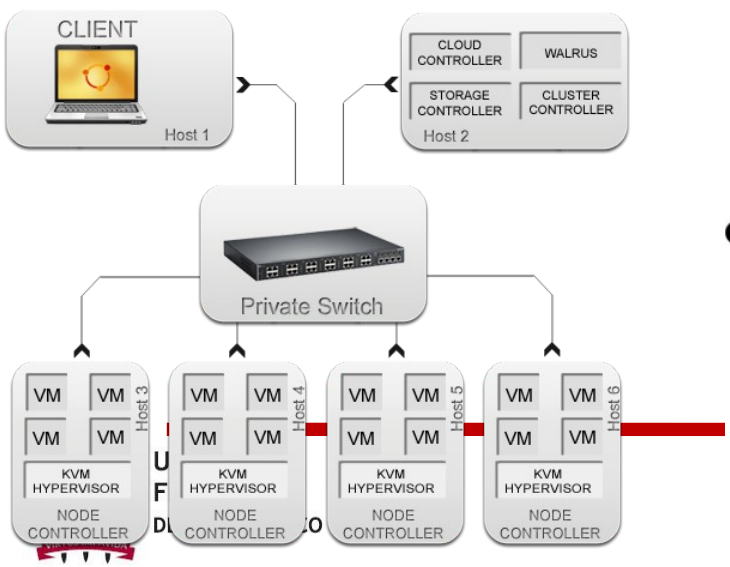
- How to evaluate the performance and availability of cloud computing systems, and detect bottlenecks to **propose improvements**?

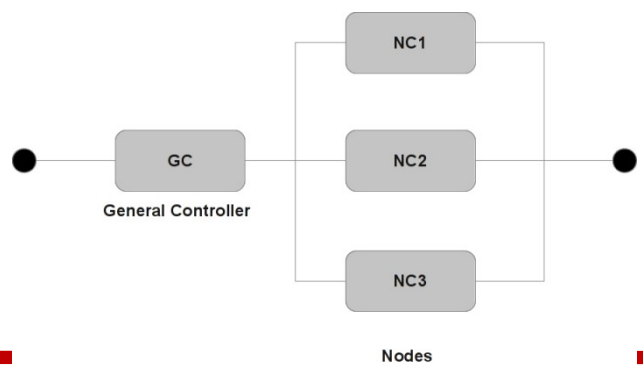
- Cloud computing systems have great complexity, even the private ones and those with small and medium size.
  - Many hardware and software components
  - Interdependence between components

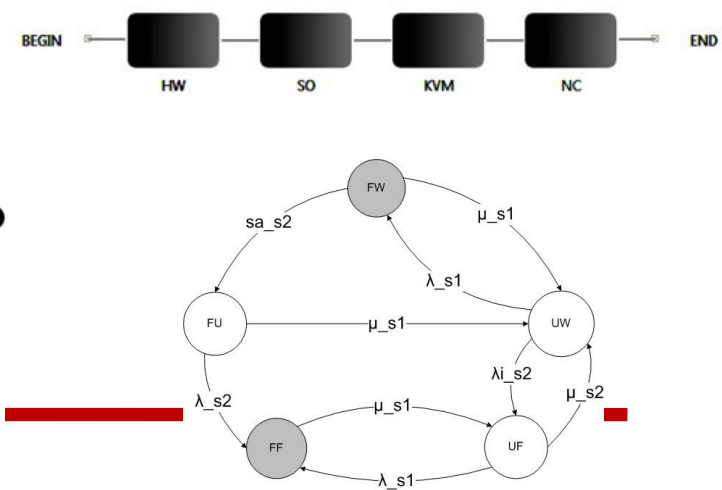
- How to identify what will bring the biggest **gain in quality of service** provided?
  - More powerful and reliable hardware ?
  - More advanced architecture ?
  - A software that provides flexibility, autonomy, resilience ?

Proposed solution

- **Hierarchical models** ease the description of those systems
- They avoid largeness and stiffness issues
- **Sensitivity analysis** techniques are important for bottleneck detection
- It is necessary **adapting** some S.A. techniques to deal with hierarchical models: **compose indices** from distinct models.







Background: Sensitivity analysis

- Parametric sensitivity analysis aims at identifying the **factors** for which the **smallest variation** implies the **highest impact** in model's output measure.
  - Variation of one parameter at a time
  - Partial derivatives
  - Factorial experimental design
  - Correlation analysis
  - Regression analysis
  - Percentage difference

$$S_\theta(Y) = \frac{\partial Y}{\partial \theta},$$

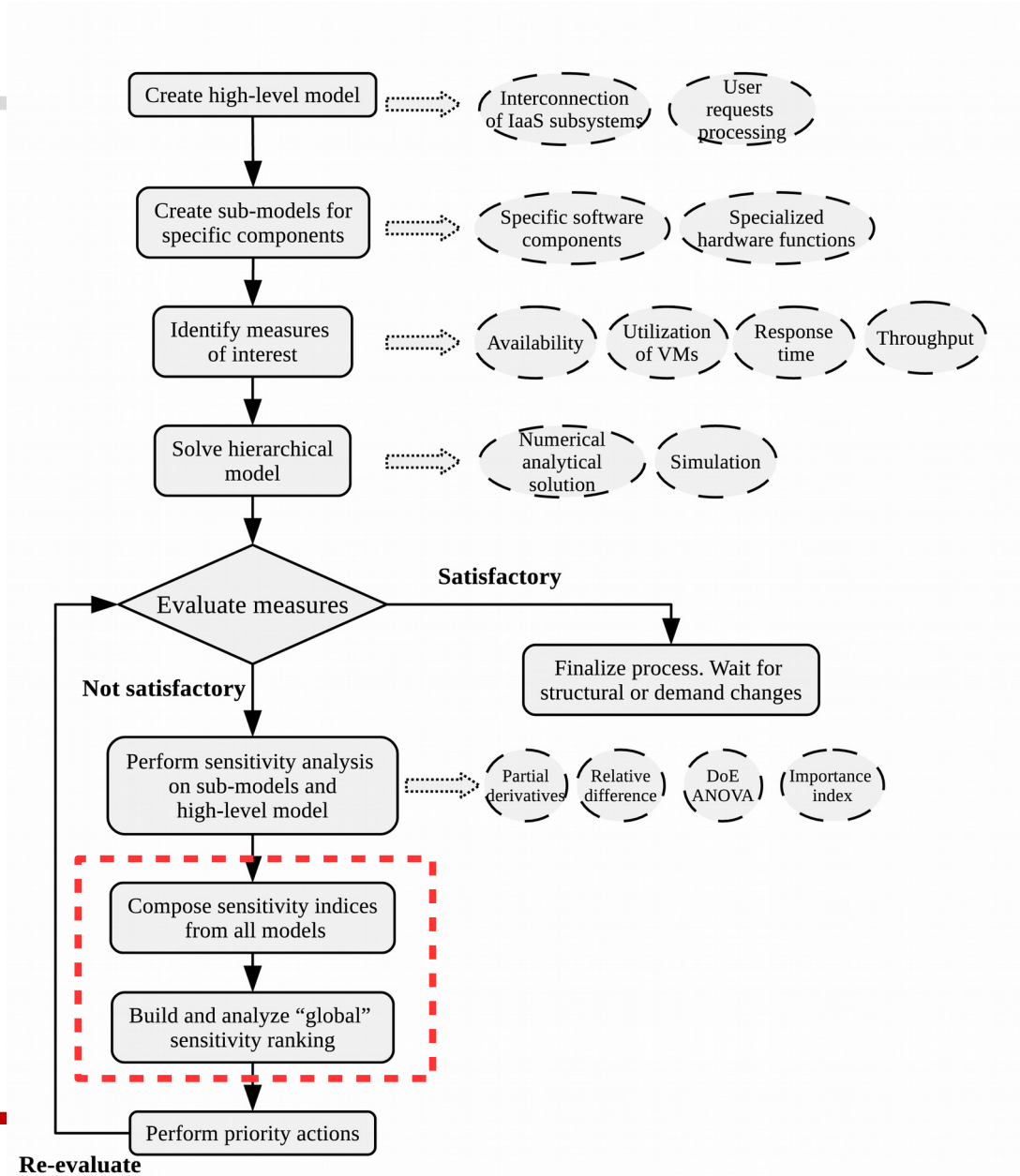
$$SS_\theta(Y) = \frac{\lambda}{Y}\frac{\partial Y}{\partial \theta}.$$

$$S_\theta(Y) = \frac{max\{Y(\theta)\} - min\{Y(\theta)\}}{max\{Y(\theta)\}}$$

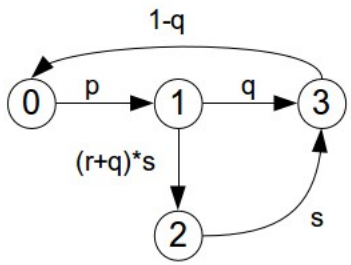
Related works

	Analytical / Simulation Models	Sensitivity indices	Cloud computing perform. and depend.	Optimization
(Sato; Trivedi, 2007)	Single model	Yes	No	No
(Yin et al., 2007)	Single model	Yes	No	No
(Chaisiri; Lee; Niyato, 2013)	No	No	Yes	Yes
(Ou; Dugan, 2003)	Hierarchical non-heterogeneous	Yes	No	No
(Chuob; Pokharel; Park, 2011)	Hierarchical non-heterogeneous	No	Yes	No
(Longo et al., 2011)	Hierarchical non-heterogeneous	No	Yes	No
(Ghosh et al, 2010)	Hierarchical non-heterogeneous	No	Yes	No
(Dantas et al., 2012a,b)	Hierachical heterogeneous	No	Yes	No
(Wei; Lin; Kong, 2011)	Hierarchical heterogeneous	No	Yes	No
My thesis	Hierarchical heterogeneous	Yes	Yes	Yes

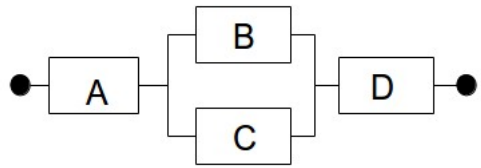
Proposed methodology



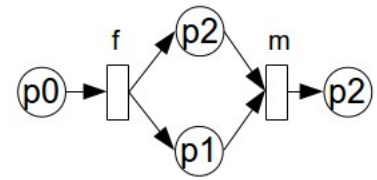
Composition of sensitivity indices



Parameter	S(Y)
q	0.9
s	0.8
r	0.4
p	0.2



Parameter	RI
A	0.7
D	0.6
C	0.4
B	0.4



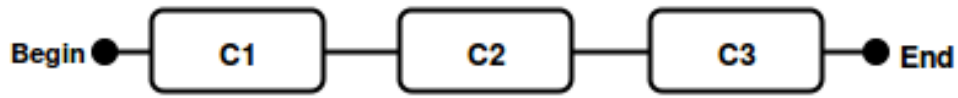
Parameter	S(R)
f	0.7
m	0.2

Composition of sensitivity indices

Parameter	S(X)
q	0.8
A	0.8
s	0.7
...	...
m	0.3
p	0.1

Unified rankingof parameters

Proposed composition techniques: RBD + Other models



Structural equation:

$A_{System} = A_{C1} \times A_{C2} \times A_{C3}$

$A_{C1} = f(p_1, p_2, p_3)$

$A_{C2} = g(p_4, p_5)$

$A_{C3} = h(p_6, p_7, p_8, p_9)$

Structural equation with sub-models functions:

$A_{System} = f(p_1, p_2, p_3) \times g(p_4, p_5) \times h(p_6, p_7, p_8, p_9)$

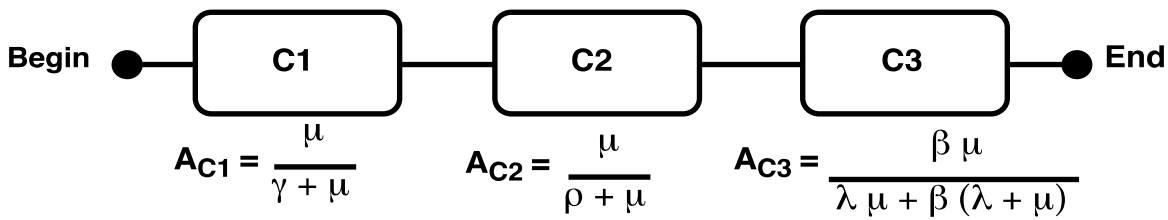
Derivative structural equations:

$\partial A_{System} / \partial p_1 = (\partial A_{C1} / \partial p_1) \times A_{C2} \times A_{C3}$

$\partial A_{System} / \partial p_4 = A_{C1} \times (\partial(A_{C2}) / \partial p_4) \times A_{C3}$

$\partial A_{System} / \partial p_6 = A_{C1} \times A_{C2} \times (\partial(A_{C3}) / \partial p_6)$

Proposed composition techniques: RBD + CTMCs



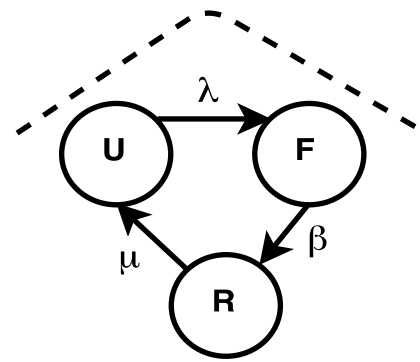
$$A_{C1} = \frac{\mu}{\gamma + \mu} \qquad A_{C2} = \frac{\mu}{\rho + \mu} \qquad A_{C3} = \frac{\beta\,\mu}{\lambda\,\mu + \beta\,(\lambda + \mu)}$$

Structural equation:

$$A = A_{C1} \times A_{C2} \times A_{C3}$$

Derivative structural equation:

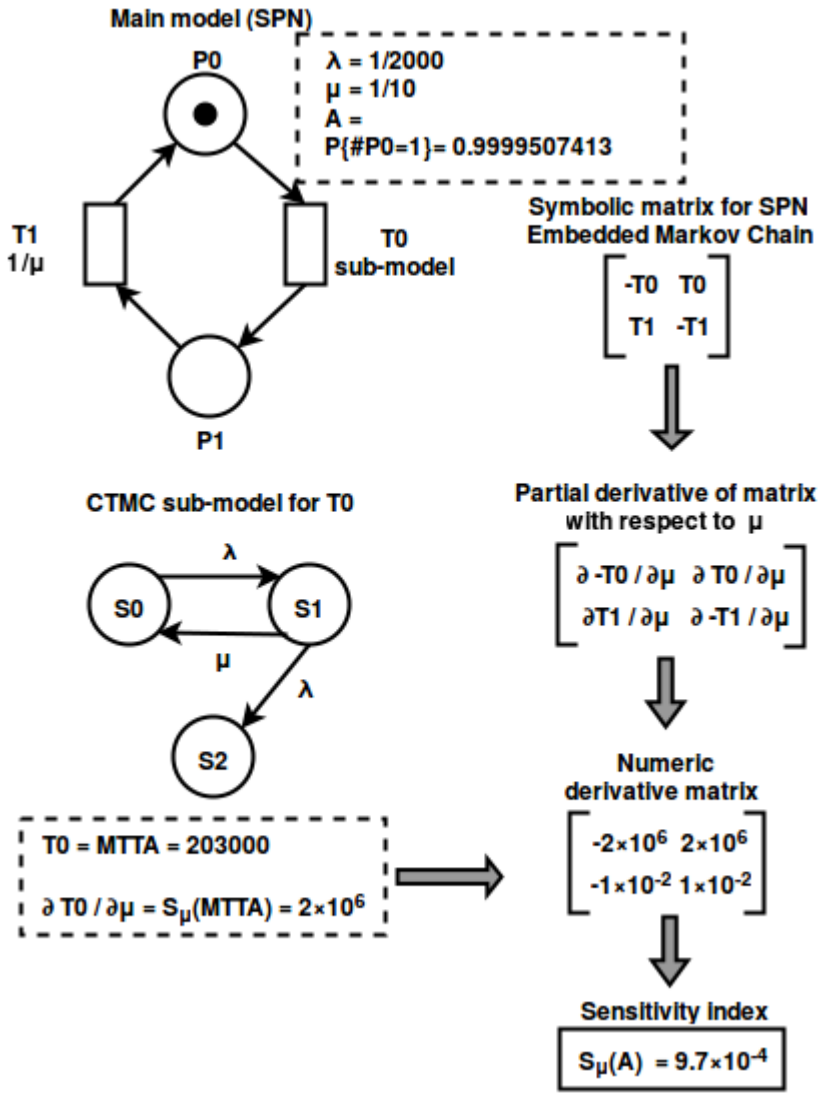
$$S_{\mu}(A) = S_{\mu}(A_{C1}) \times A_{C2} \times A_{C3} + A_{C1} \times S_{\mu}(A_{C2}) \times A_{C3} + A_{C1} \times A_{C2} \times S_{\mu}(A_{C3})$$



$$S_{\mu}(A_{C1}) = \frac{\gamma}{(\gamma + \mu)^2} \qquad S_{\mu}(A_{C2}) = \frac{\rho}{(\rho + \mu)^2} \qquad S_{\mu}(A_{C3}) = \frac{\beta^2\,\lambda}{(\lambda\,\mu + \beta\,(\lambda + \mu))^2}$$

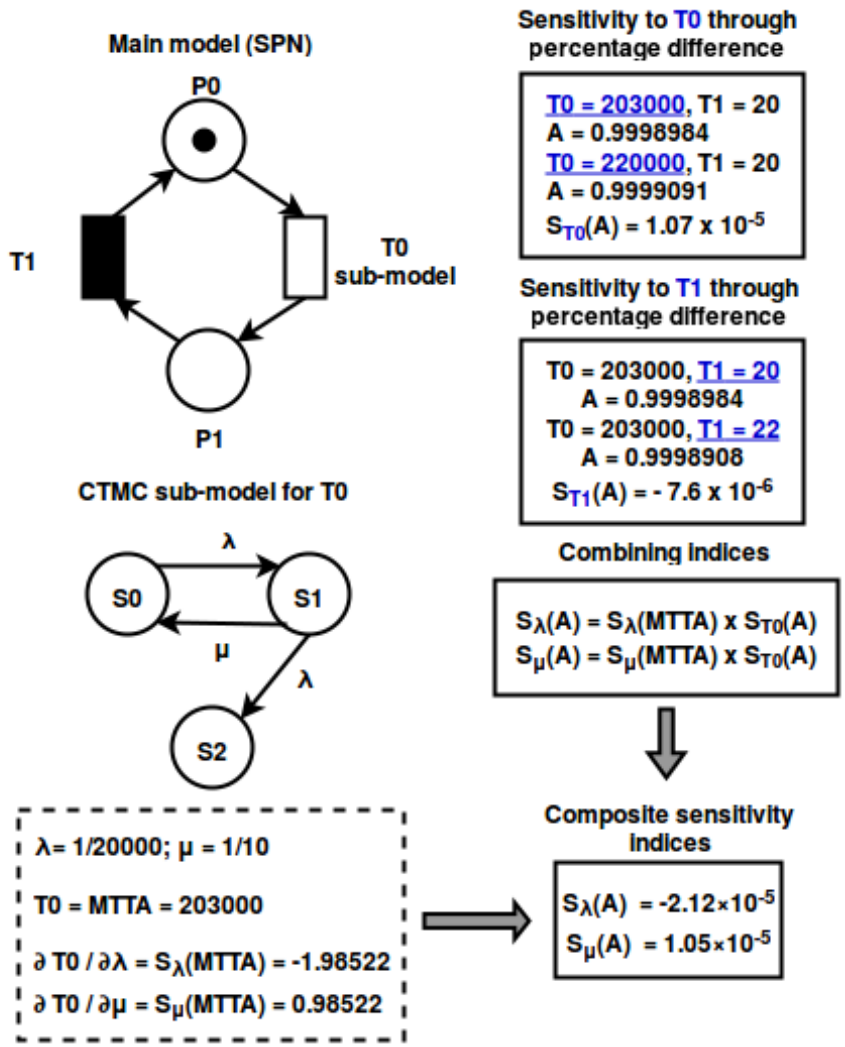
In some cases, CTMC sub-models can be solved through closed-form equations. Their partial derivatives will provide the sensitivity indices.

Proposed composition techniques: SPN + CTMCs



When the SPN can be solved through **numerical analysis**, the sensitivity indices from CTMC sub-models are included directly on derivative of underlying rate matrix.

Proposed composition techniques: SPN (simulation) + CTMCs



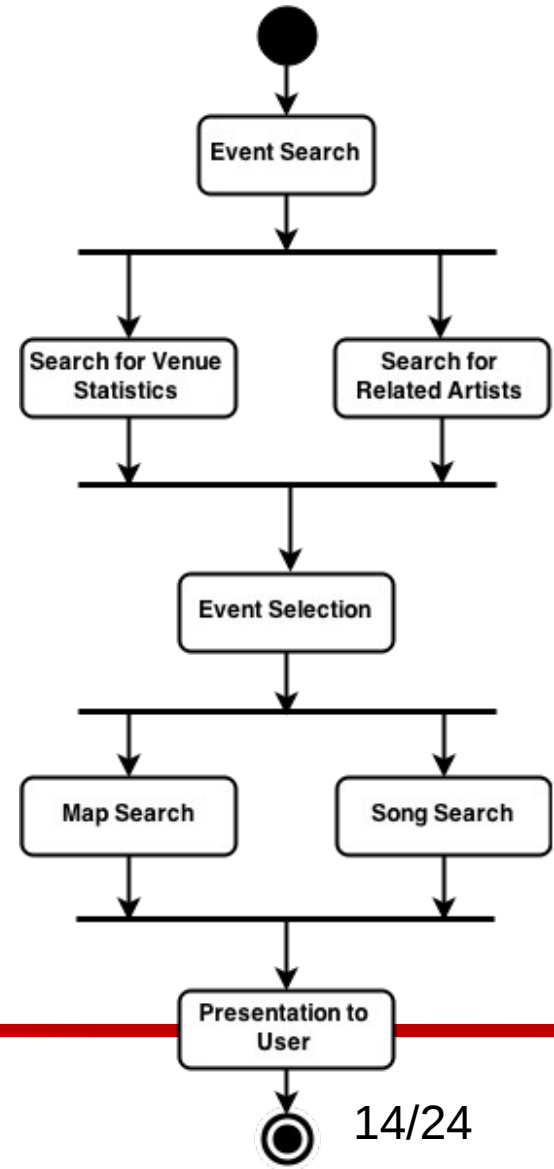
When the SPN is can only be solved through **simulation**, the indices from CTMC sub-models are multiplied by indices of corresponding SPN transitions. Therefore, we follow the **chain rule**:

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx},$$

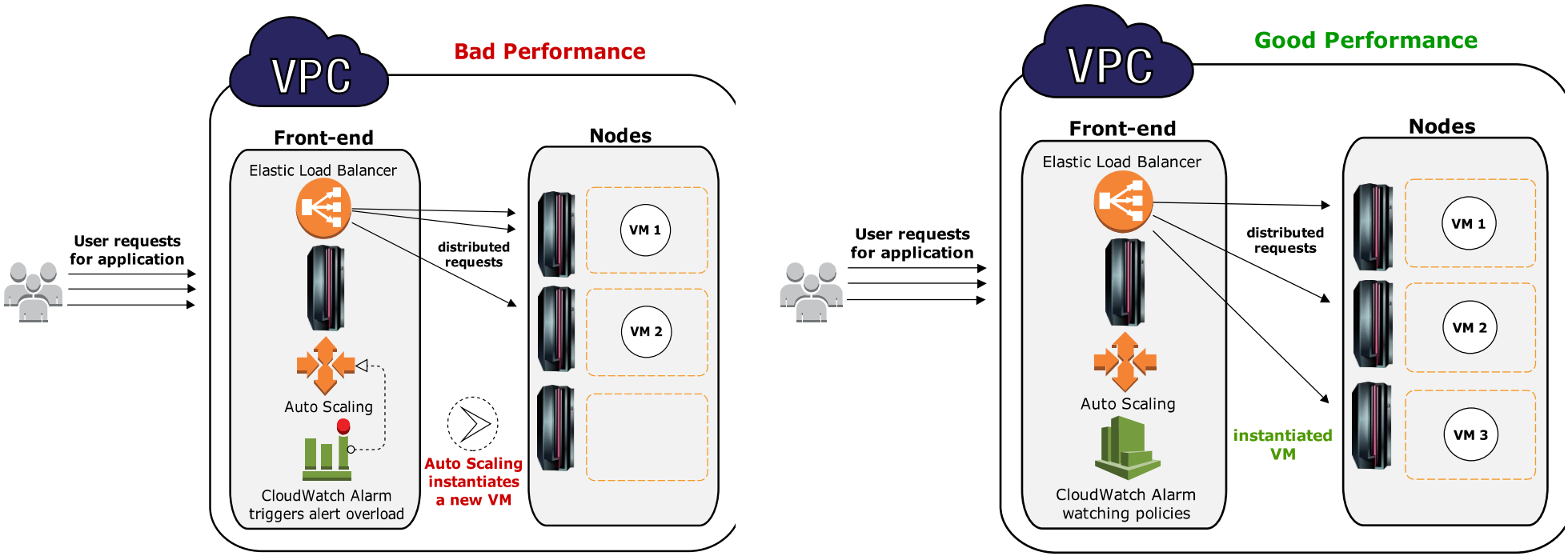
z is the measure from SPN model
x is a parameter from CTMC sub-model
y is a transition from SPN model which has the delay as a function of the parameter x

Case study: Composite web services on private cloud with autoscaling

- Composite **web services** for musical events recommendation
- This mashup runs on a private cloud, with elasticity resources: **automatic creation** and **termination of VMs** according to the workload



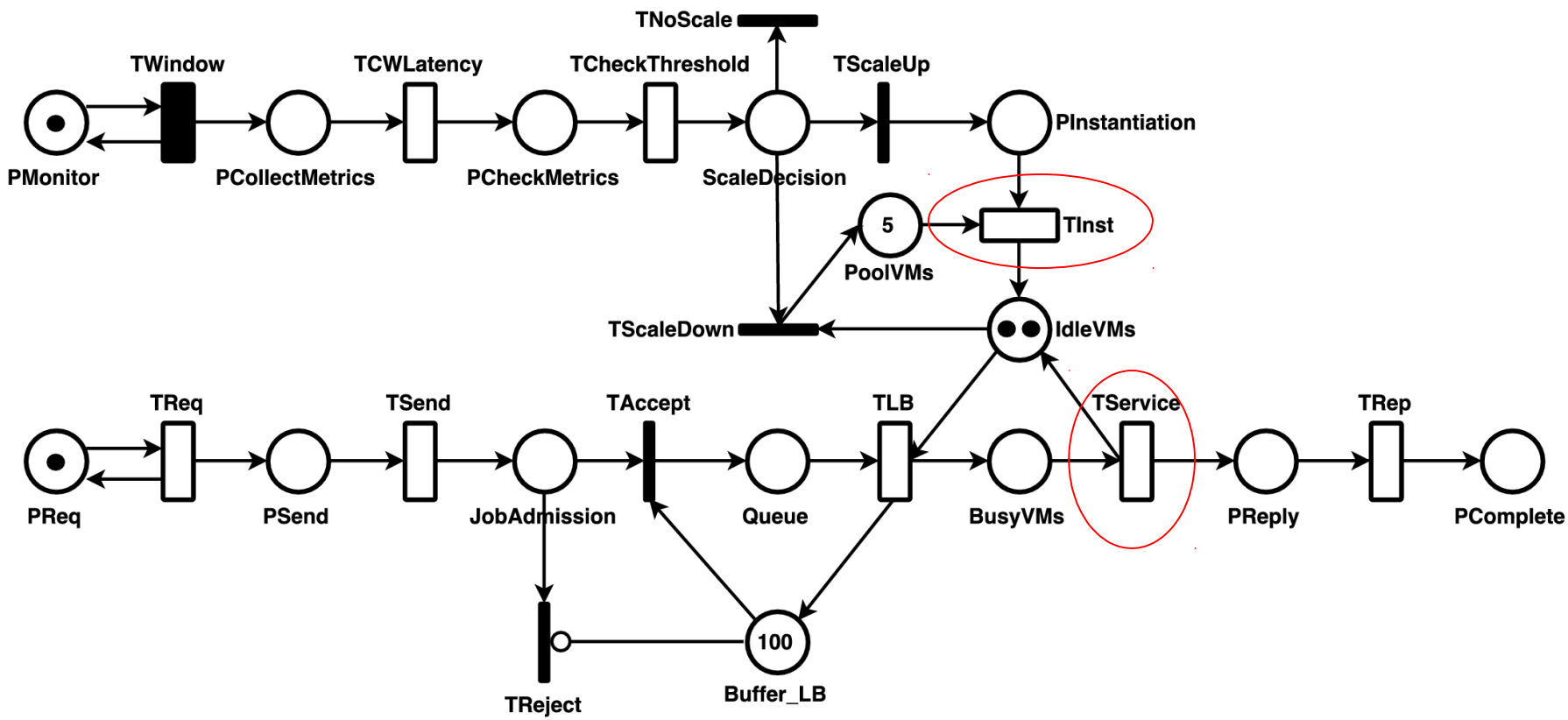
Composite web services on private cloud with autoscaling



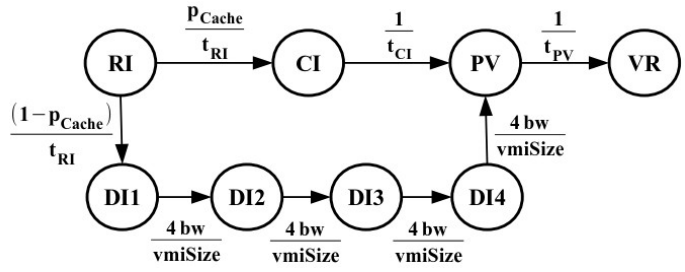
Composite web services on private cloud with autoscaling

- 3 models: 1 SPN + 2 CTMCs:
  - Workload / **autoscaling**
  - VM **instantiation**
  - Web service **execution**

Composite web services on private cloud with autoscaling (Step 1)

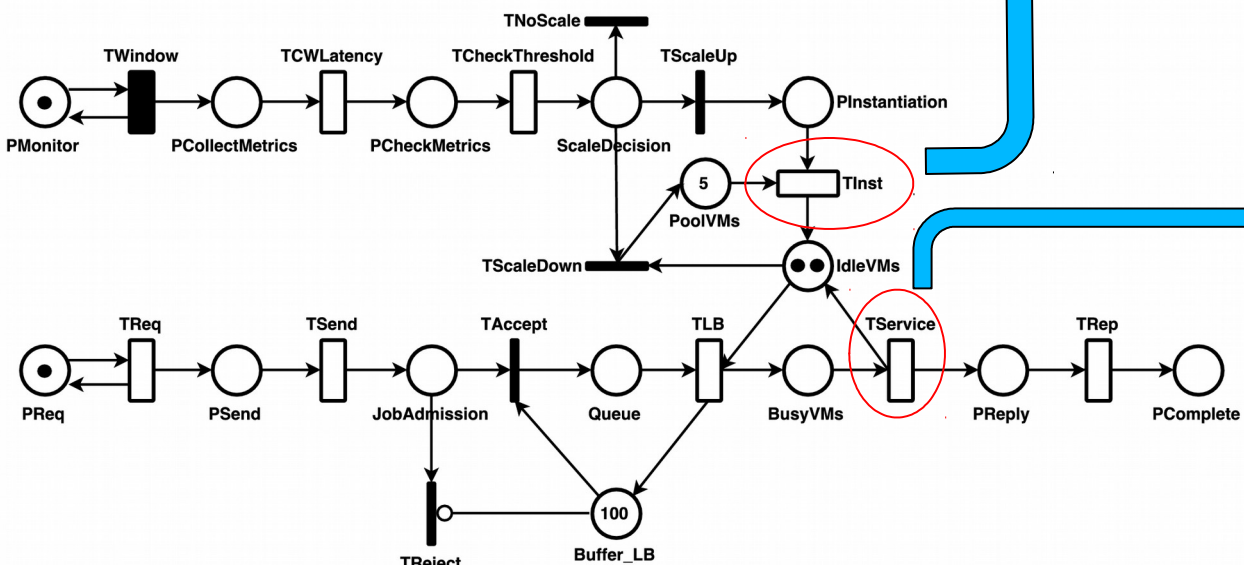


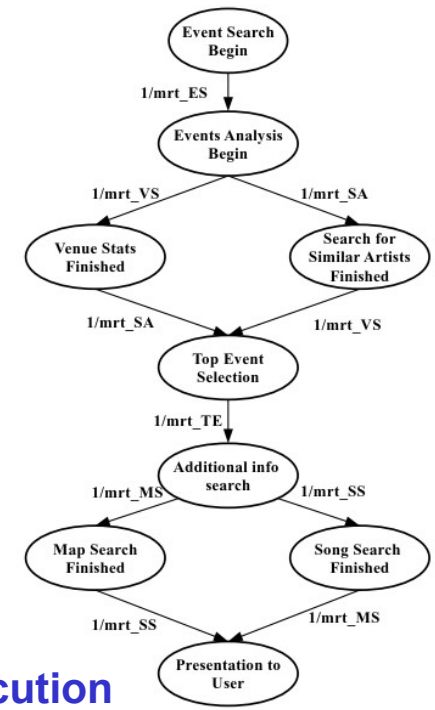
Composite web services on private cloud with autoscaling (Step 2)



VM instantiation

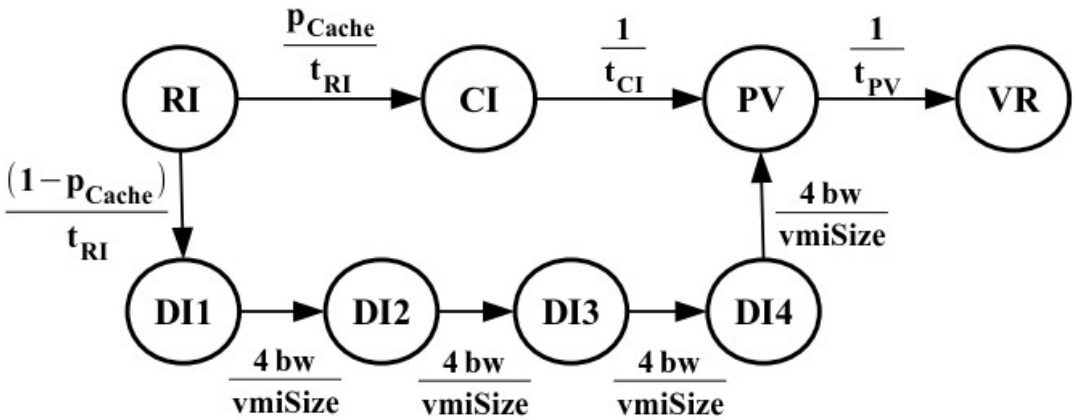
System representation

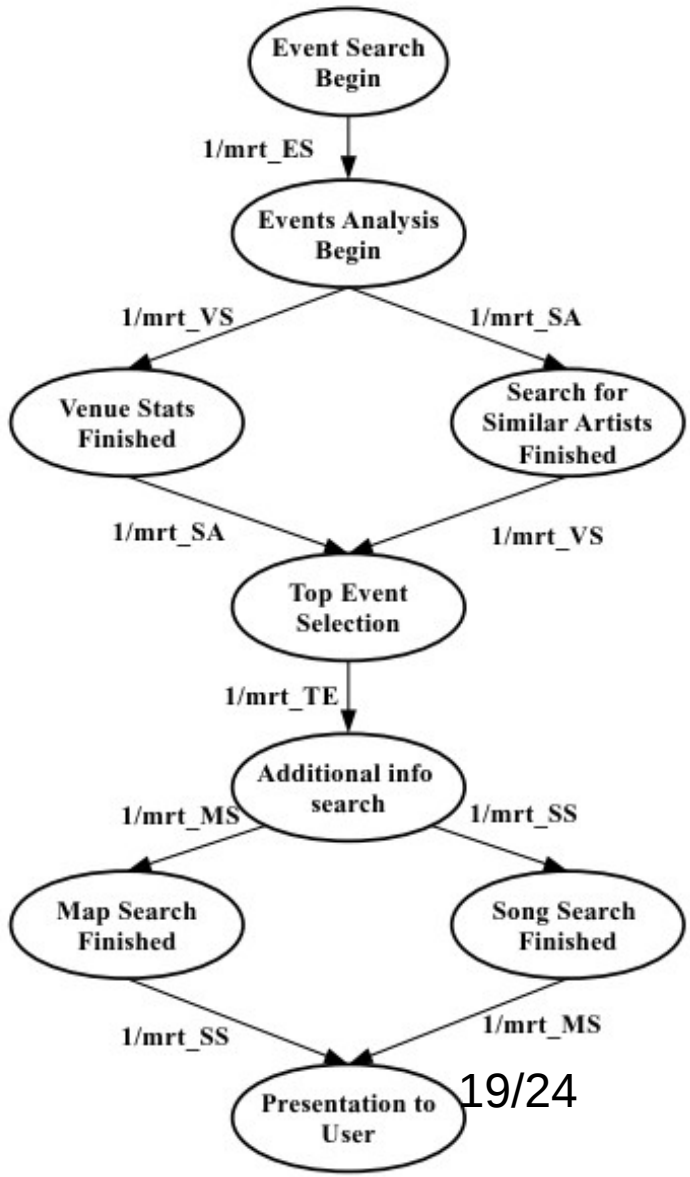




Web service execution

Composite web services on private cloud with autoscaling (Step 2)





Composite web services on private cloud with autoscaling (Step 3, 4, and 5)

Performance measures

Measure	Expression	Value
Utilization of VMs (%)	E{#BusyVMs})/(E{#IdleVMs}+E{#BusyVMs})	38.3 %
Average number of busy VMs	E{#BusyVMs}	1.716
Average number of idle VMs	E{#IdleVMs}	2.773
LB queue size (#of requests)	E{#Queue}	0.432
Mean response time - Rsp - (s)	NRequests/(P{#PReply>0}×(1/TReply))	9.029 s

This is the metric of most interest for the user and it is not is a satisfactory level

Composite web services on private cloud with autoscaling (Steps 6 and 7)

Sensitivity ranking for the main model

Parameter	S(Rsp)
TService	0.45763
TLB	0.13788
TRep	0.11303
TSend	0.11466
TReq	-0.05808
TWindow	0.00617
TCWLatency	0.00489
TInst	0.00256
TCheckThreshold	0.00176

$S_{pCache}(Rsp) = S_{TInst}(Rsp) \times SS_{pCache}(TInst)$

$S_{mrt\ ES}(Rsp) = S_{TService}(Rsp) \times SS_{mrt\ ES}(TService)$

Sensitivity ranking for the VM instantiation submodel

Parameter	SS(TInst)
pCache	-4.52843
vmiSize	0.52363
bw	-0.52363
t_PV	0.28465
t_CI	0.18421
t_RI	0.00752

Sensitivity ranking for the mashup sub-model

Parameter	SS(TService)
mrt_ES	0.33906
mrt_SA	0.32711
mrt_SS	0.26727
mrt_TS	0.03284
mrt_MS	0.02274
mrt_VS	0.01096

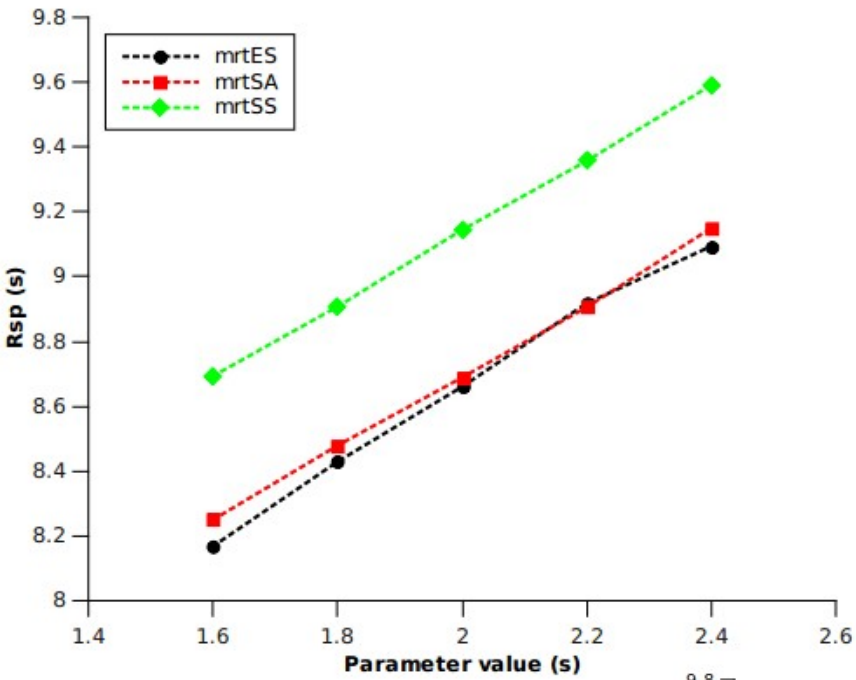
Composite web services on private cloud with autoscaling (Step 8)

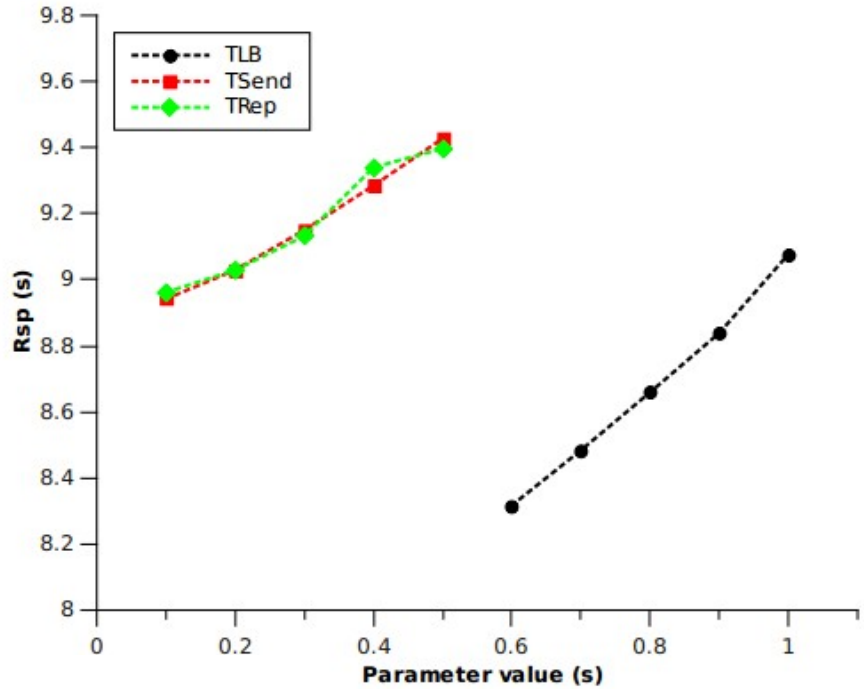
Unified sensitivity ranking for the general model and submodels

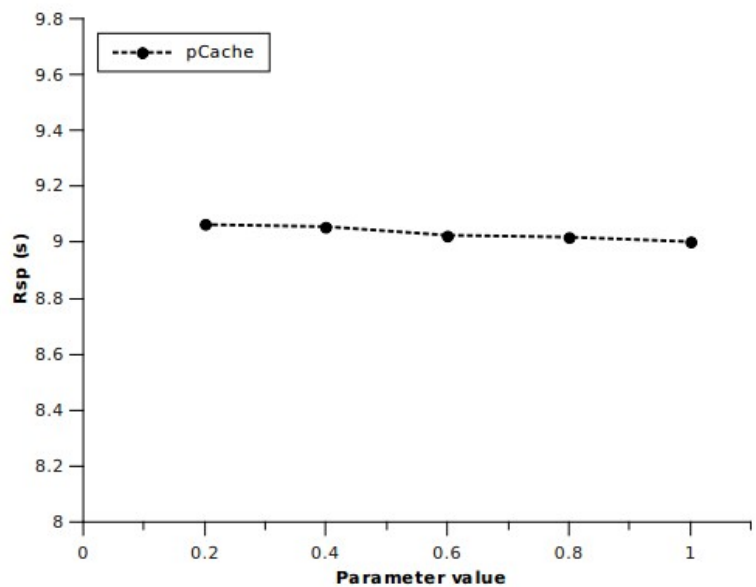
Parameter	S(Rsp)
mrt_ES	0.15517
mrt_SA	0.14969
TLB	0.13977
mrt_SS	0.12231
TSend	0.11466
TRep	0.11303
TReq	-0.05808
mrt_TS	0.01503
pCache	-0.01162
mrt_MS	0.01041
TWindow	0.00617
mrt_VS	0.00502
TCWLatency	0.00489
TCheckThreshold	0.00176
vmiSize	0.00134
bw	-0.00134
t_PV	0.00073
t_CI	0.00047
t_RI	0.00002

- Response time of following web services:
  - **Event Search**
  - **Similar Artists**
  - **Song Search**
- Execution time of **Load Balancer**
- **Network latency** for sending request and receiving reply
- The most important parameter of VM instantiation process (**pCache**) is only **intermediate** when the concern is the total response time of the application

Composite web services on private cloud with autoscaling







Final remarks

- The results **achieved** in this doctoral research were submitted to scientific **journals**. Some papers were already accepted and published, others are under peer-review process.
- The implementation of some features in **Mercury** is in final stage.