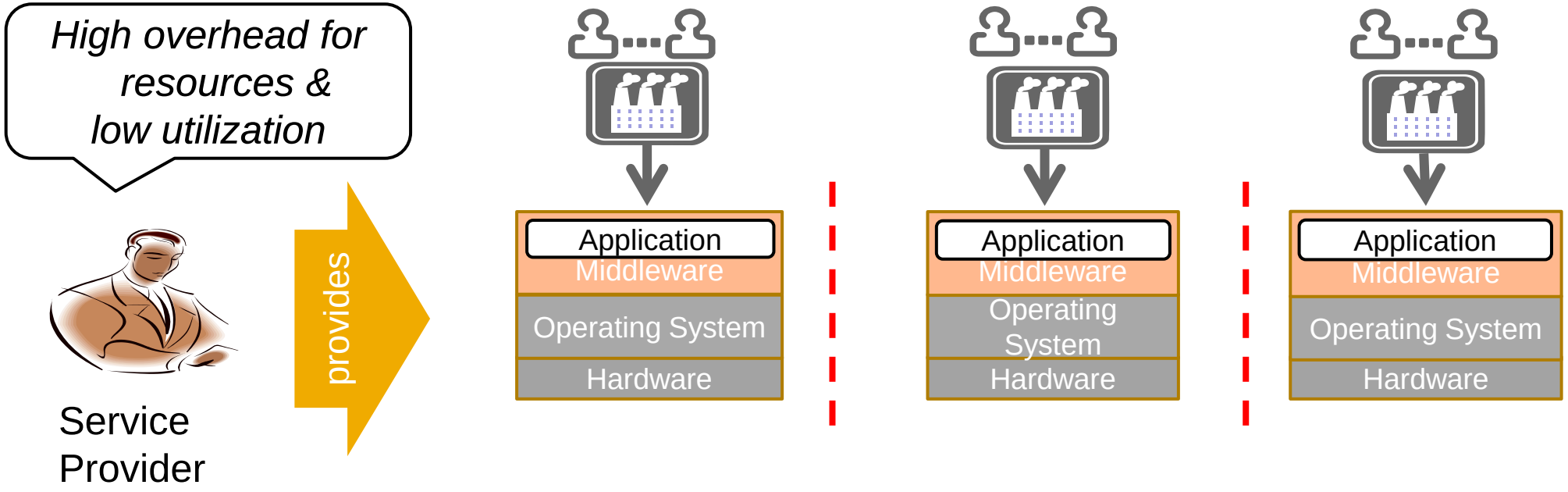
Optimization Method for Request Admission Control to Guarantee Performance Isolation

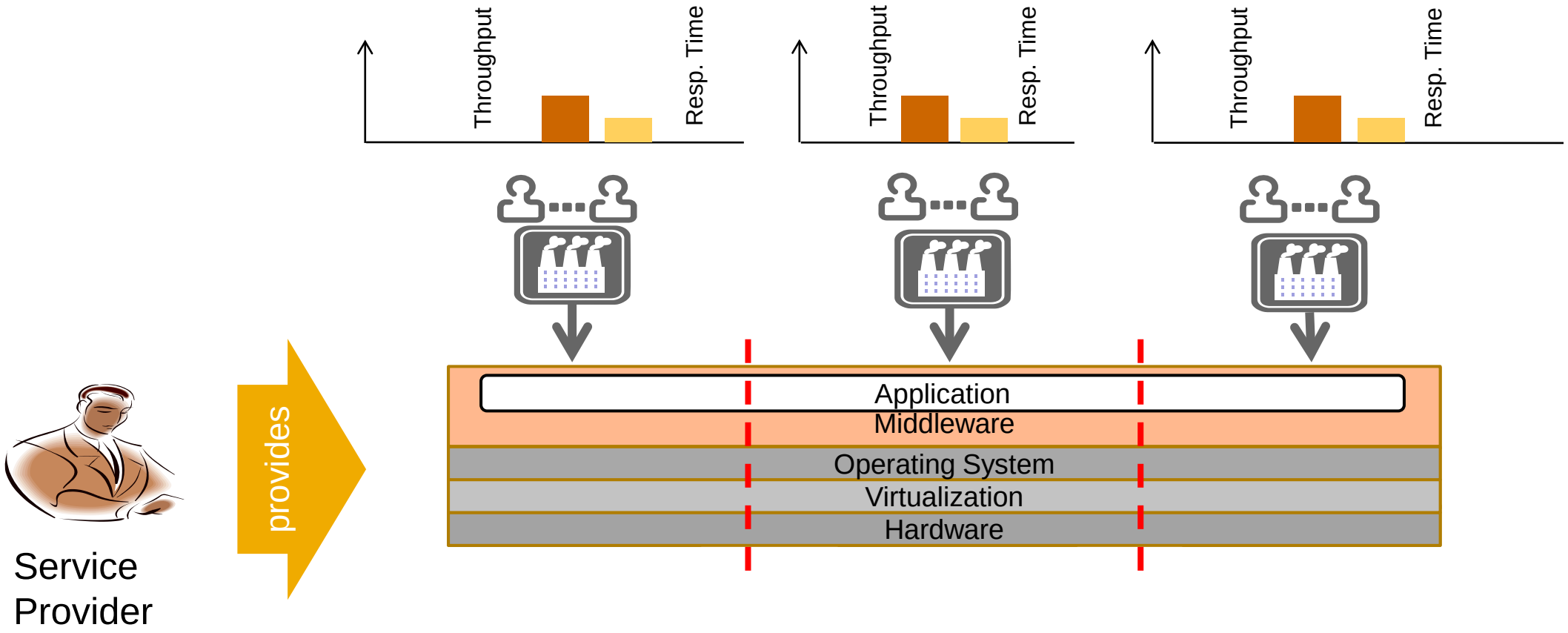
Rouven Krebs, Philipp Schneider, Nikolas Herbst

March 22, 2014 – HotTopiCS 2014

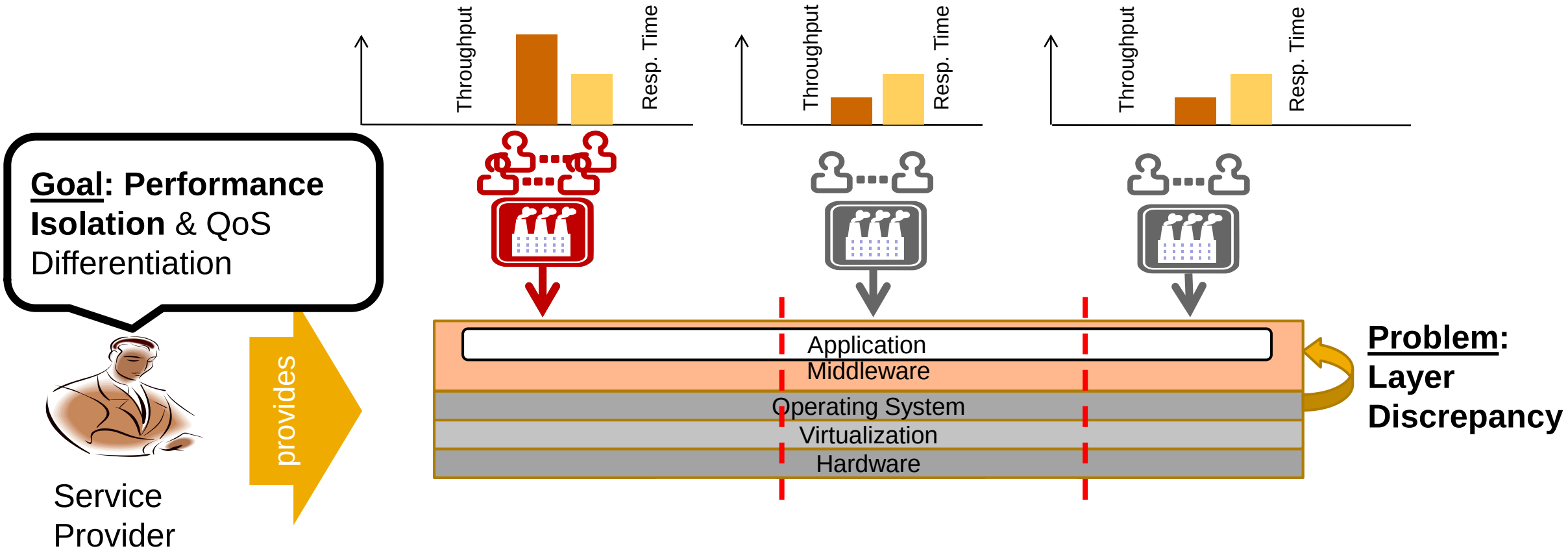
Motivation: Isolation



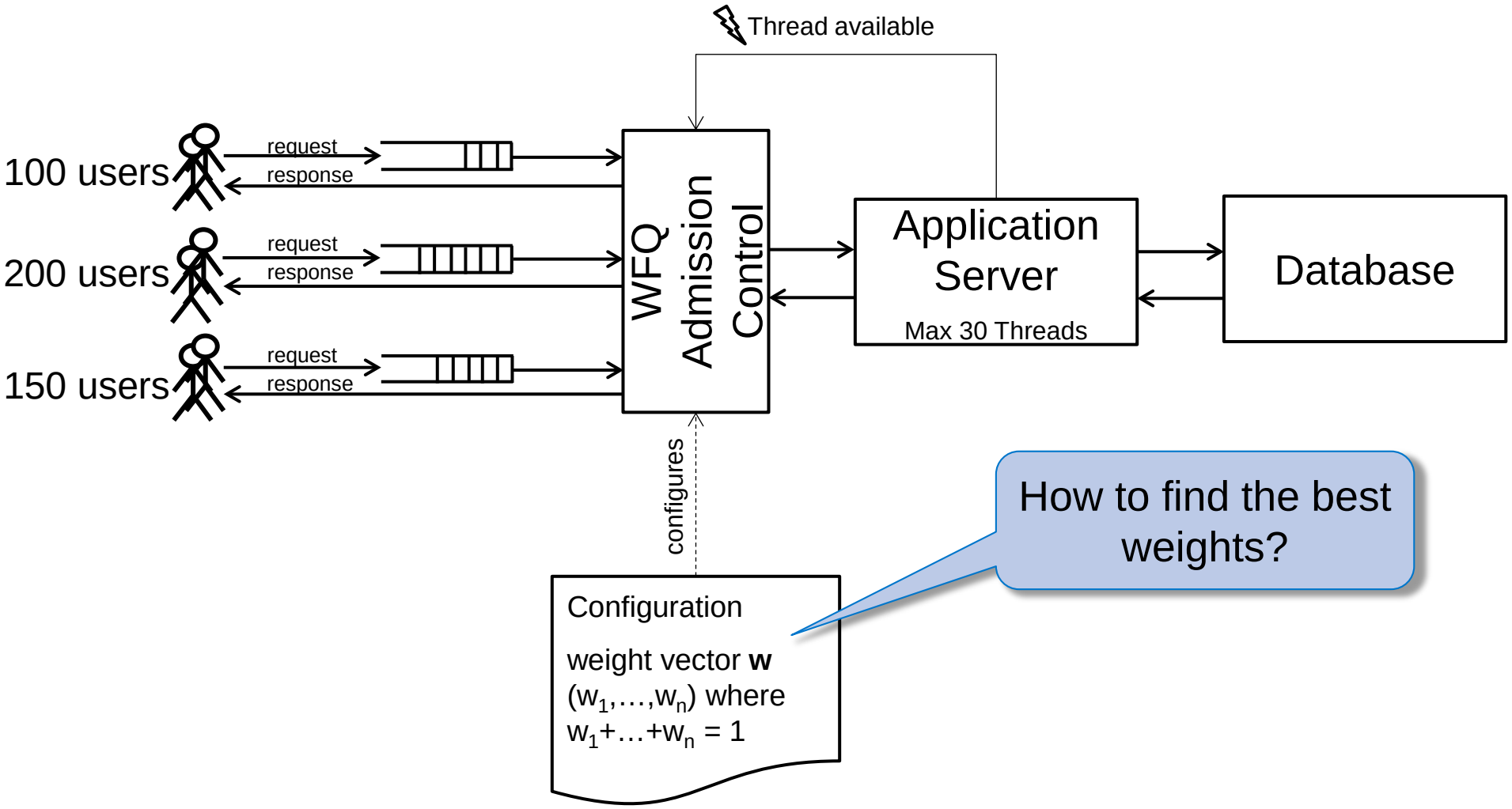
Motivation: Isolation



Motivation: Isolation

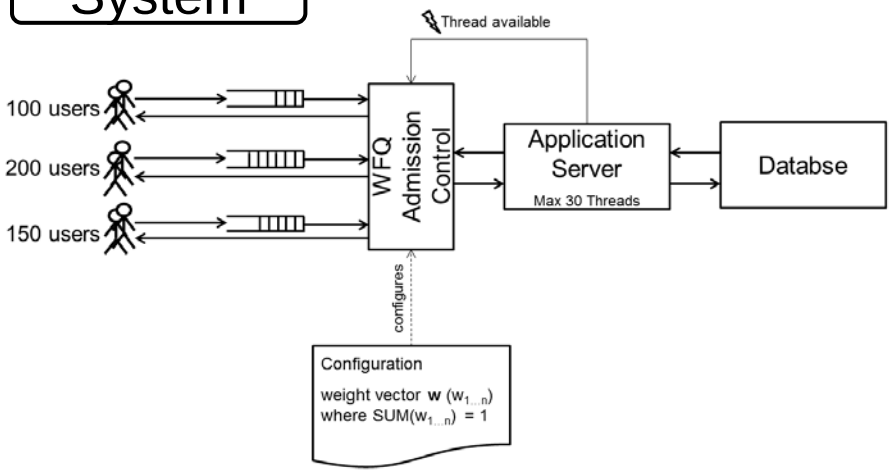


System Overview, Problem and Contribution



General Approach

System



Collect Information

Relevant Information

Information used to derive a model

- Thinktime
- #User
- Throughput
- Service Time per Request

Derive System Function

System Function

Analytical Model to calculate the response time R_i for tenant i dependent on w_i

- Convex
- Differentiable (2 times)

Used to optimze

Optimization

$$\min! \quad f(\mathbf{w}) := \sum_{j=1}^{n} f_j(w_j),$$

subject to

$$\mathbf{w} \in M \quad \Longleftrightarrow \quad \begin{cases} \sum_{j=1}^{n} w_j = 1, \\ w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \end{cases}$$

Computes weights

General Approach

System

Collect Information

Relevant Information

Information used to derive a model

"A system is performance isolated, if for tenants working within their quotas the performance is not affected even if other tenants exceed their quotas" [3]

Computes weights

Optimization

$$\min! \quad f(\mathbf{w}) := \sum_{j=1}^{n} f_j(w_j),$$

subject to

$$\mathbf{w} \in M \quad \Longleftrightarrow \quad \begin{cases} \sum_{j=1}^{n} w_j = 1, \\ w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \end{cases}$$

Used to optimze

System Function

Analytical Model to calculate the response time R_i for tenant i dependent on w_i

- Convex
- Differentiable (2 times)

Optimization

"performance is not affected"

- If the response time > guarantee the weight has to be increased

Easily optimizeable

- Convex
- Minimum within the plane $w_1+\ldots+w_n = 1$

Optimization

$$\text{min!} \quad f(\mathbf{w}) := \sum_{j=1}^{n} \boxed{f_j(w_j),}$$

subject to

$$\mathbf{w} \in M \quad \Longleftrightarrow \quad \begin{cases} \sum_{j=1}^{n} w_j = 1, \\ w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \end{cases}$$

"within their quotas"

- If the quota is exceeded the performance is not longer that important

Configurable

- Importance of guarantee
- Importance of quota

Optimization – Violation Term

"performance is not affected"

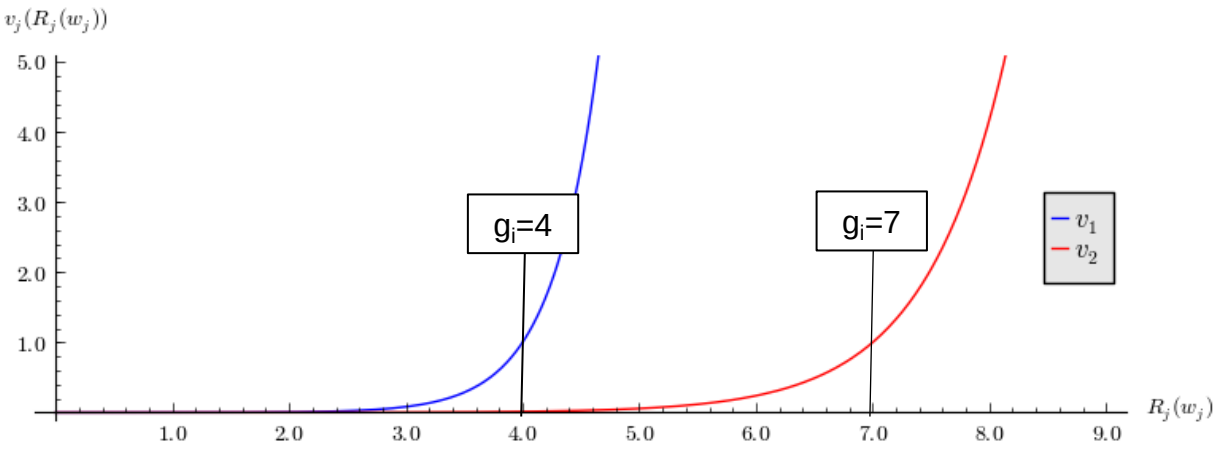
- If the response time > guarantee the weight has to be increased

Tuning Parameter

System function

Guarantee

$$v_i(w_i) := \exp\left(c_v \frac{R_i(w_i) - g_i}{g_i}\right)$$



Optimization

"performance is not affected"

$$v_i(w_i) := \exp\left(c_v \frac{R_i(w_i) - g_i}{g_i}\right)$$

Easily optimizeable

- Convex
- Minimum within the plane $w_1+\ldots+w_n = 1$

Optimization

$$\text{min!} \quad f(\mathbf{w}) := \sum_{j=1}^{n} \boxed{f_j(w_j),}$$

subject to

$$\mathbf{w} \in M \quad \Longleftrightarrow \quad \begin{cases} \sum_{j=1}^{n} w_j = 1, \\ w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \end{cases}$$

"within their quotas"

- If the quota is exceeded the performance is not longer that important

Configurable

- Importance of guarantee
- Importance of quota

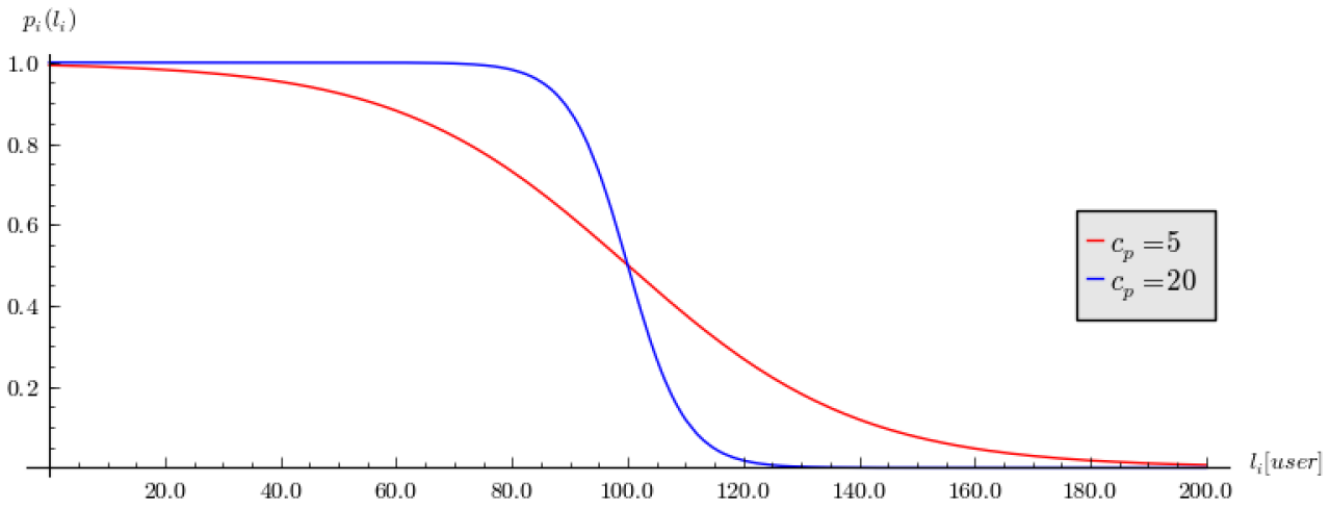
Optimization – Penalty Term

Tuning Parameter

Observed load

Quota

$$p_i(l_i) := (1 + \exp\left(c_p \frac{l_i - q_i}{q_i}\right))^{-1}$$



"within their quotas"

- If the quota is exceeded the performance is not longer that important

Optimization

"performance is not affected"

$$v_i(w_i) := \exp\left(c_v \frac{R_i(w_i) - g_i}{g_i}\right)$$

$$f_i(w_i) := h_i \cdot p_i \cdot v_i(w_i)$$

"within their quotas"

$$p_i(l_i) := (1 + \exp\left(c_p \frac{l_i - q_i}{q_i}\right))^{-1}$$

Optimization

$$\text{min!} \quad f(\mathbf{w}) := \sum_{j=1}^{n} \boxed{f_j(w_j),}$$

subject to

$$\mathbf{w} \in M \iff \begin{cases} \sum_{j=1}^{n} w_j = 1, \\ w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \end{cases}$$

Easily optimizeable

- Convex
- Minimum within the plane $w_1 + \ldots + w_n = 1$

Configurable

- Importance of guarantee
- Importance of quota

Optimization

"performance is not affected"

$$v_i(w_i) := \exp\left(c_v \frac{R_i(w_i) - g_i}{g_i}\right)$$

Heaviness

$$h_i := \frac{l_i \cdot n}{\sum_{j=1}^{n} l_j}$$

Small tenants are preferred by the optimization. h_i compensates this.

$$f_i(w_i) := h_i \cdot p_i \cdot v_i(w_i)$$

h_i and p_i are independent from w_i.

"within their quotas"

$$p_i(l_i) := (1 + \exp\left(c_p \frac{l_i - q_i}{q_i}\right))^{-1}$$

Substituted

$$\exp\left(\alpha_i R_i(w_i) - \beta_i\right)$$

ly optimizeable

Convex

- Minimum within the plane $w_1 + \ldots + w_n = 1$

Configurable

- Importance of guarantee
- Importance of quota

Optimization - Optimzeable

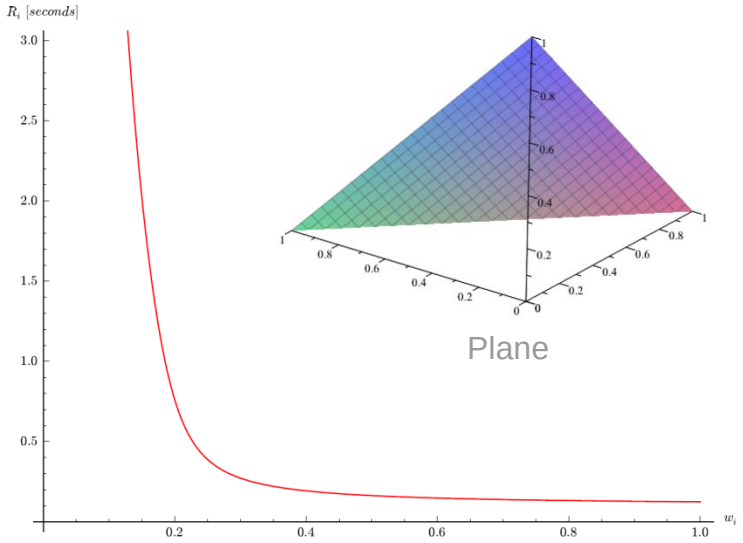
$f_i(w_i) = \exp(\alpha_i R_i(w_i) - \beta_i)$.

If system function $R_i(w_i)$ is convex $\alpha_i R_i(w_i) - \beta_i$ is convex, as well. Since for each convex, function $\varphi : \mathrm{R} \rightarrow \mathrm{R}$ the composition $\exp \circ \varphi$ is convex.

Easily optimizeable

- Convex
- Minimum in the interior of the plane $w_1 + \ldots + w_n = 1$ and for all $w_i : 0 \leq w_i \leq 1$

Suppose the minimum lies on the boundary of the constraint plane. As each weight vector on the boundary has a component which is zero, it holds that there exists an $i \in \{1, \ldots, n\}$, such that $w_i = 0$.
Since $\lim_{w \rightarrow 0} f_j(w) = \infty$ this is a contradiction.



Plane

System Function

Optimization

"performance is not affected"

$$v_i(w_i) := \exp\left(c_v \frac{R_i(w_i) - g_i}{g_i}\right)$$

$$f_i(w_i) := h_i \cdot p_i \cdot v_i(w_i)$$

"within their quotas"

$$p_i(l_i) := (1 + \exp\left(c_p \frac{l_i - q_i}{q_i}\right))^{-1}$$

Optimization

$$\text{min!} \quad f(\mathbf{w}) := \sum_{j=1}^{n} \boxed{f_j(w_j),}$$

subject to

$$\mathbf{w} \in M \iff \begin{cases} \sum_{j=1}^{n} w_j = 1, \\ w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \end{cases}$$

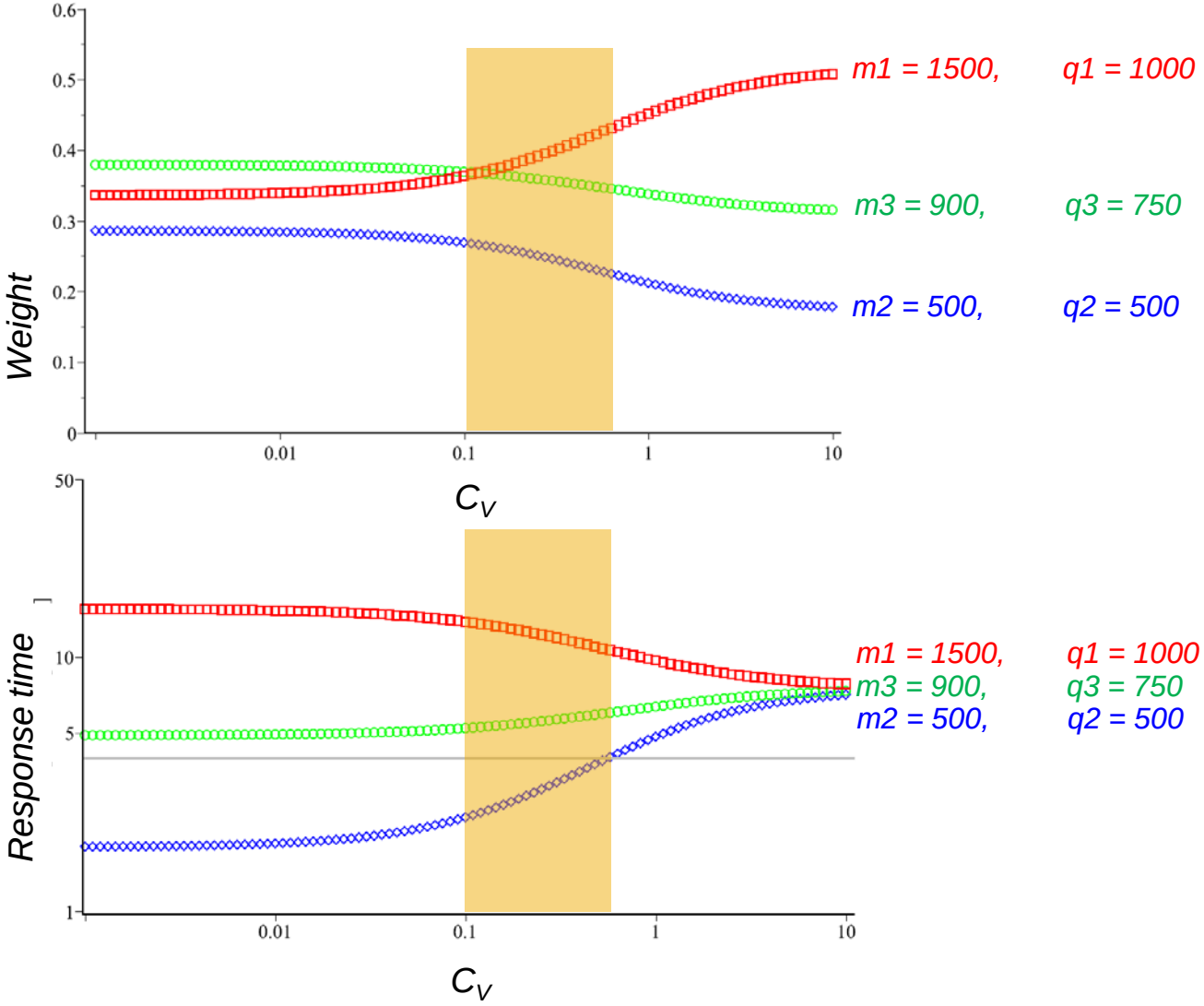
Easily optimizeable

- Convex
- Minimum within the plane $w_1 + \ldots + w_n = 1$

Configurable

- Importance of guarantee
- Importance of quota

Optimization - Configurability (Violation Term)



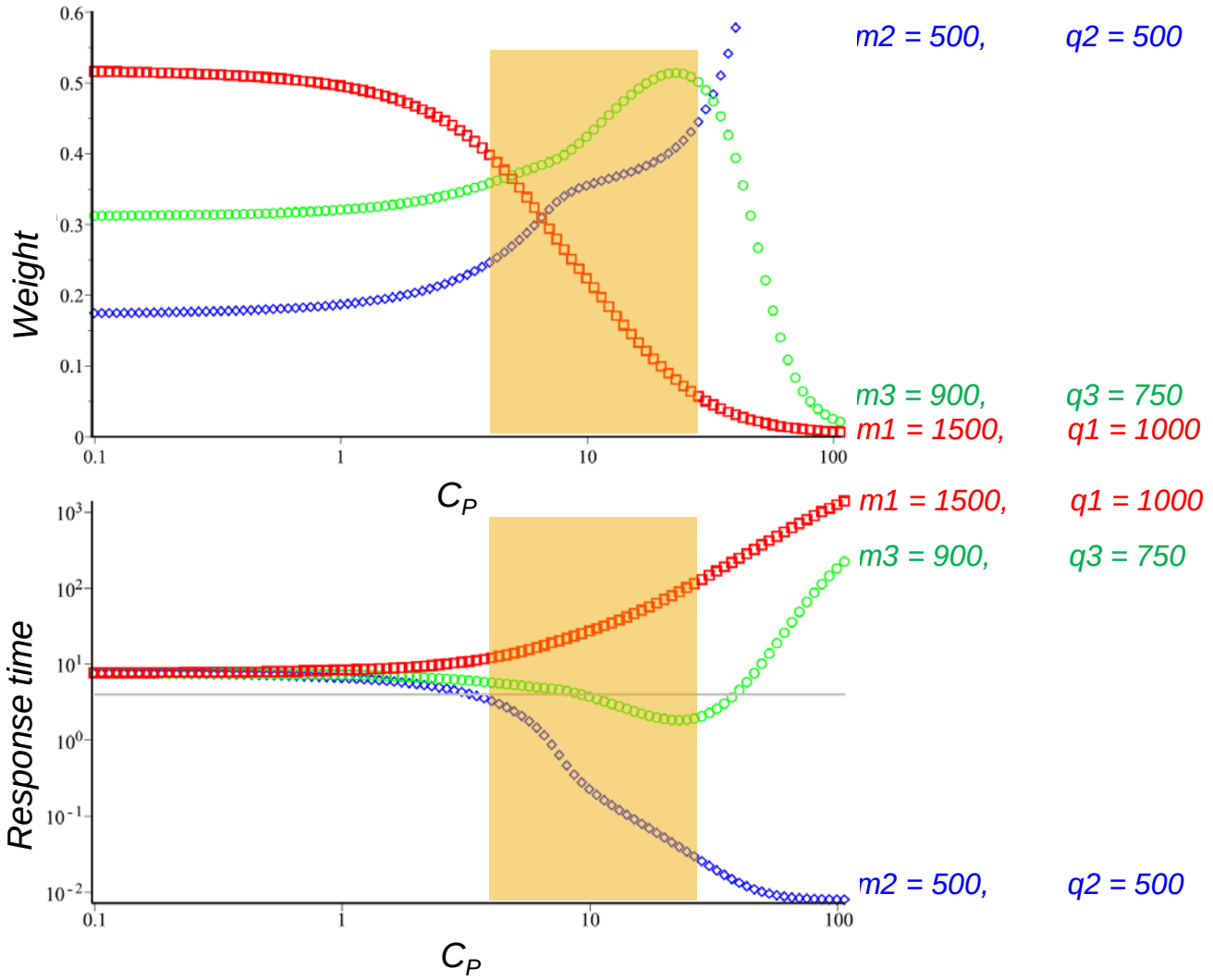
In case of a **low** violation term configuration parameter the quota violation is the **only reference**.

In case of a **high** value, the quota violation is not important any more. Consequently the response time converges. Thus there is **no performance isolation**.

Configurable

- Importance of guarantee
- Importance of quota

Optimization - Configurability (Penalty Term)



In case of a **low** penalty term configuration parameter the guarantee violation is the **only important reference**.

In case of a **high** value, the guarantee violation becomes irrelevant. Consequently the response time depends **only on the quota violation**.

Configurable

- Importance of guarantee
- Importance of quota

Optimization

"performance is not affected"

$$v_i(w_i) := \exp\left(c_v \frac{R_i(w_i) - g_i}{g_i}\right)$$

$$f_i(w_i) := h_i \cdot p_i \cdot v_i(w_i)$$

"within their quotas"

$$p_i(l_i) := (1 + \exp\left(c_p \frac{l_i - q_i}{q_i}\right))^{-1}$$

Optimization

$$\text{min!} \quad f(\mathbf{w}) := \sum_{j=1}^{n} f_j(w_j),$$

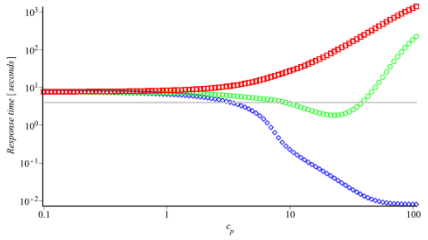
subject to

$$\mathbf{w} \in M \iff \begin{cases} \sum_{j=1}^{n} w_j = 1, \\ w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \end{cases}$$

Easily optimizeable

- Convex
- Minimum within the plane $w_1 + \ldots + w_n = 1$

Configurable



Validation

Optimization

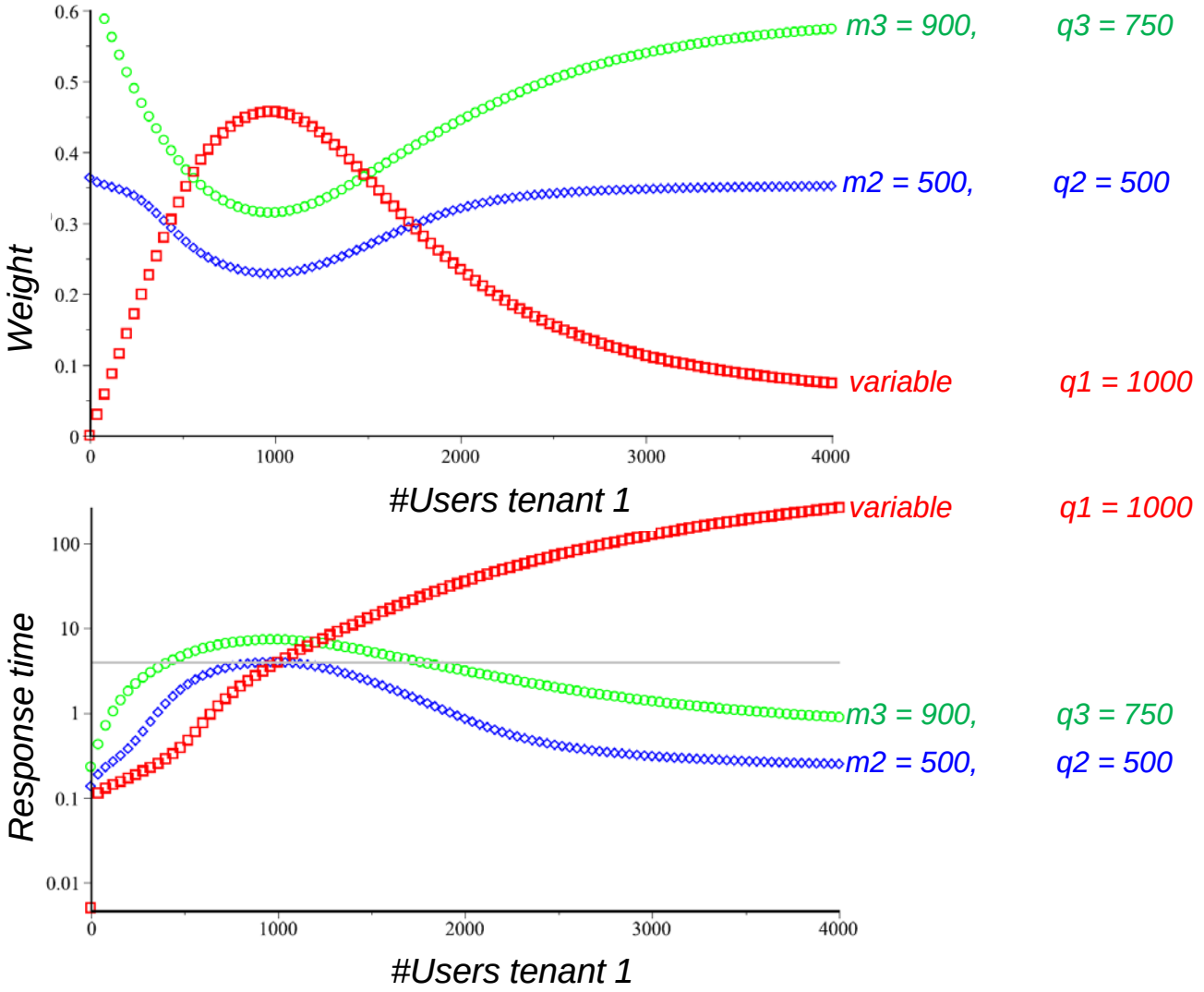
$$\min! \quad f(\mathbf{w}) := \sum_{j=1}^{n} f_j(w_j),$$

subject to

$$\mathbf{w} \in M \quad \Longleftrightarrow \quad \begin{cases} \sum_{j=1}^{n} w_j = 1, \\ w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \end{cases}$$

Validation

- Simulation
- Numerical optimizer
- Random samples of R_i compared with real system running TPC-W, error below 20%.



Validation

"performance is not affected"

$$v_i(w_i) := \exp\left(c_v \frac{R_i(w_i) - g_i}{g_i}\right)$$

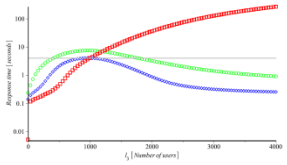
$$f_i(w_i) := h_i \cdot p_i \cdot v_i(w_i)$$

"within their quotas"

$$p_i(l_i) := (1 + \exp\left(c_p \frac{l_i - q_i}{q_i}\right))^{-1}$$

Optimization

$$\text{min!} \quad f(\mathbf{w}) := \sum_{j=1}^{n} f_j(w_j),$$



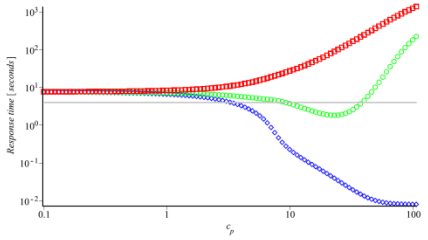
subject to

$$\mathbf{w} \in M \iff \begin{cases} \sum_{j=1}^{n} w_j = 1, \\ w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \end{cases}$$

Easily optimizeable

- Convex
- Minimum within the plane $w_1 + \ldots + w_n = 1$

Configurable



Related Work

Zhang [5] and Fehling [2] are representatives of approaches where isolation is tried to be achieved by placing tenants.

Control theory in general [1] dynamically optimize the system which is prone to need some time and comes along with the risk of oscillating.

The SPIN environment [4] interprets the ratio of the mean arrival rate and mean service rate to detect performance anomalies/overload of the system. The aggressive tenant request flow is adopted.

In our former work [3], we introduced static request admission mechanisms like Round Robin or Black Lists. However, static approach are not suitable with regards to efficiency.

References

[1] J. Hellerstein, Y. Diao, S. Parekh, and D. Tilbury.
Feedback Control of Computing Systems. Wiley, 2004.

[2] C. Fehling, F. Leymann, and R. Mietzner.
A framework for optimized distribution of tenants in cloud applications.
In *CLOUD 2010 IEEE 3rd International Conference on Cloud Computing*, 2010.

[3] R. Krebs, C. Momm, and S. Kounev.
Metrics and Techniques for Quantifying Performance Isolation in Cloud Environments.
Elsevier Science of Computer Programming Journal (SciCo), 2013.

[4] X. Li, T. Liu, Y. Li, and Y. Chen.
SPIN: Service performance isolation infrastructure in multi-tenancy environment.
Service-Oriented Computing, SICSOC 2008, pages 649–663, 2008.

[5] Y. Zhang, Z. Wang, B. Gao, C. Guo, W. Sun, and X. Li.
An effective heuristic for on-line tenant placement problem in SaaS.
Web Services, IEEE International Conference on, 0:425–432, 2010.

Recap

Problem

- Performance Isolation in multi-tenant applications is a challenge
- Request admission control mechanisms are one approach
- How to determine the weights for the admission control

Approach and Contribution

- Derive a violation function describing the difference between each tenants response time guarantee and the observed one
- Derive a penalty function to “remove” a tenant from the optimization if he exceeds the quota
- Integrate both into one function

Benefit

- Opportunity to provide better guarantees in multi-tenant applications

Questions?

"performance is not affected"

$$v_i(w_i) := \exp\left(c_v \frac{R_i(w_i) - g_i}{g_i}\right)$$

$$f_i(w_i) := h_i \cdot p_i \cdot v_i(w_i)$$

"within their quotas"

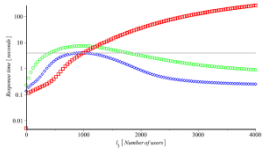
$$p_i(l_i) := (1 + \exp\left(c_p \frac{l_i - q_i}{q_i}\right))^{-1}$$

Optimization

$$\text{min!} \quad f(\mathbf{w}) := \sum_{j=1}^{n} \boxed{f_j(w_j)},$$

subject to

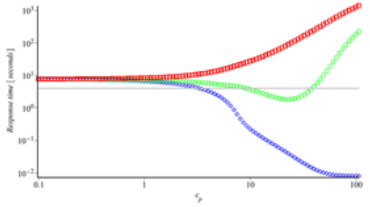
$$\mathbf{w} \in M \iff \begin{cases} \sum_{j=1}^{n} w_j = 1, \\ w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \end{cases}$$



Easily optimizeable

- Convex
- Minimum within the plane SUM($w_{1...n}$) = 1

Configurable



Rouven.Krebs@sap.com Philipp.Schneider01@sap.com herbst@fzi.de

cloudscale-project.eu

descartes.ipd.kit.edu