Optimization Method for Request Admission Control to Guarantee Performance Isolation
Rouven Krebs, Philipp Schneider, Nikolas Herbst
March 22, 2014 – HotTopiCS 2014
Motivation: Isolation

High overhead for resources & low utilization

Service Provider provides

- Application
  - Middleware
  - Operating System
  - Hardware

- Application
  - Middleware
  - Operating System
  - Hardware

- Application
  - Middleware
  - Operating System
  - Hardware
Motivation: Isolation

Throughput Resp. Time

Service Provider provides

Application
Middleware
Operating System
Virtualization
Hardware
Motivation: Isolation

Goal: Performance Isolation & QoS Differentiation

Problem: Layer Discrepancy
System Overview, Problem and Contribution

Application Server
Max 30 Threads

Configuration
weight vector \( \mathbf{w} = (w_1, \ldots, w_n) \) where
\[ w_1 + \ldots + w_n = 1 \]

How to find the best weights?
General Approach

### System

- **100 users**
- **200 users**
- **150 users**

**System Function**

- Derive System Function
  - Analytical Model to calculate the response time $R_i$ for tenant $i$ dependent on $w_i$
  - Convex
  - Differentiable (2 times)

**Relevant Information**

- Information used to derive a model
  - Thinktime
  - #User
  - Throughput
  - Service Time per Request

**General Approach**

- **Collect Information**
- **Information used to derive a model**
- **Derive System Function**
- **System Function**

**Optimization**

- Computes weights

- **Optimization**

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

subject to

$$
w \in M \iff \left\{ \begin{array}{l}
\sum_{j=1}^{n} w_j = 1, \\
w_j \geq 0, \quad \forall j \in \{1, \ldots, n\}
\end{array} \right. $$

© 2014 SAP AG or an SAP affiliate company. All rights reserved.
General Approach

**System**

- 100 users
- 200 users
- 150 users

**VIFO Admission Control**

**Application Server**

**Max 32 Threads**

**Database**

**Collect Information**

**Relevant Information**

Information used to derive a model

**System Function**

Derive System Function

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

- Convex
- Differentiable (2 times)

“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]

**Optimization**

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

subject to

$$w \in M \iff \sum_{j=1}^{n} w_j = 1,$$

$$w_j \geq 0, \ \forall j \in \{1, \ldots, n\}$$
Optimization

“performance is not affected”
- If the response time > guarantee the weight has to be increased

Easily optimizeable
- Convex
- Minimum within the plane \( w_1 + \ldots + w_n = 1 \)

“within their quotas”
- If the quota is exceeded the performance is not longer that important

Configurable
- Importance of guarantee
- Importance of quota

\[
\begin{align*}
\text{min} & \quad f(w) := \sum_{j=1}^{n} f_j(w_j), \\
\text{subject to} & \quad \sum_{j=1}^{n} w_j = 1, \\
& \quad w_j \geq 0, \quad \forall j \in \{1, \ldots, n\}
\end{align*}
\]
"performance is not affected"

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

\[ v_i(w_i) := \exp \left( c_v \frac{R_i(w_i) - g_i}{g_i} \right) \]
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 \( w_1 + \ldots + w_n = 1 \)

Configurable
- Importance of guarantee
- Importance of quota

“within their quotas”
- If the quota is exceeded the performance is not longer that important

Optimization
\[
\begin{align*}
\min f(w) &= \sum_{j=1}^{n} f_j(w_j), \\
\text{subject to} & \\
w \in M & \iff \begin{cases} 
\sum_{j=1}^{n} w_j = 1, \\
w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \end{cases}
\end{align*}
\]
Optimization – Penalty Term

If the quota is exceeded, the performance is not longer that important.

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

\[
\begin{align*}
\text{min} & \quad f(w) := \sum_{j=1}^{n} f_j(w_j), \\
\text{subject to} & \quad w \in M \iff \left\{ \begin{array}{l}
\sum_{j=1}^{n} w_j = 1, \\
    w_j \geq 0, \quad \forall j \in \{1, \ldots, n\}
\end{array} \right. 
\end{align*}
\]

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”

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

\[ f_i(w_i) := h_i \cdot p_i \cdot v_i(w_i) \]

“within their quotas”

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

Heaviness

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

Substituted

\[ \exp \left( \alpha_i R_i(w_i) - \beta_i \right) \]

Small tenants are preferred by the optimization. \( h_i \) compensates this.

\( h_i \) and \( p_i \) are independent from \( w_i \).

Easily optimizeable

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

Configurable

- Importance of guarantee
- Importance of quota
$f_i(w_i) = \exp(\alpha_i R_i(w_i) - \beta_i)$.

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 $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ the composition $\exp \circ \varphi$ is convex.

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 $\lim_{w \to 0} f_i(w) = \infty$ this is a contradiction.

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

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

Easily optimizeable

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

Configurable

- Importance of guarantee
- Importance of quota

Optimization

\[
\begin{align*}
\min \quad & f(w) := \sum_{j=1}^{n} f_j(w_j), \\
\text{subject to} \quad & w \in M \iff \left\{ \begin{array}{l}
\sum_{j=1}^{n} w_j = 1, \\
0 \leq w_j, \quad \forall j \in \{1, \ldots, n\}
\end{array} \right.
\end{align*}
\]
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"

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

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

subject to

\[ w \in M \iff \{ \sum_{j=1}^{n} w_j = 1, \quad w_j \geq 0, \quad \forall j \in \{1, \ldots, n\} \} \]

Easily optimizeable

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

Configurable
Validation

Optimization

$$\min_{\mathbf{w}} \ f(\mathbf{w}) := \sum_{j=1}^{n} f_j(w_j),$$
subject to

$$\mathbf{w} \in M \iff \left\{ \sum_{j=1}^{n} w_j = 1, \ w_j \geq 0, \ \forall j \in \{1, \ldots, n\} \right\}$$

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”

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

Easily optimizeable
- Convex
- Minimum within the plane \( w_1 + \ldots + w_n = 1 \)

Configurable

“performance is not affected”

\[ f(w) := \sum_{j=1}^{n} f_j(w_j), \]

subject to

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


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

\[ f_i(w_i) := h_i \cdot p_i \cdot v_i(w_i) \]

"within their quotas"

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

\[ \min \quad f(w) := \sum_{j=1}^{n} f_j(w_j), \]

subject to

\[ w \in M \iff \begin{cases} \sum_{j=1}^{n} w_j = 1, \\ w_j \geq 0, & \forall j \in \{1, \ldots, n\} \end{cases} \]

Easily optimizeable
- Convex
- Minimum within the plane \( \sum w_i = 1 \)

Configurable