# ACRA: A Unified Admission Control and Resource Allocation Framework for Virtualized Environments

Dimitrios Dechouniotis[1], Nikolaos Leontiou[1], Nikolaos Athanasopoulos[2], George Bitsoris[1] and Spyros Denazis[1]
[1]Electrical and Computer Engineering Department,
University of Patras, Rion Patras, Greece.
{ddexouni, nleontiou, bitsoris, sdena}@ece.upatras.gr
[2]Electrical Engineering Department,
Eindhoven University of Technology, The Neteherlands.
n.athanasopoulos@tue.nl

***Abstract*—Exploiting the benefits of virtualization, web services are consolidated in large data centers. Managing the performance of such complex systems is a critical problem. Providers must offer applications with high quality of service (QoS) and performance and simultaneously achieve optimal utilization of their infrastructure. Meeting their Service Level Objectives (SLOs), such as response time in a dynamic environment (dense load, variable capacity), while minimizing the energy consumption of the data center is an open research problem. Most of the proposed approaches use either admission control or resource allocation techniques to solve it. We present a unified framework, which models the system's dynamic behavior with a group of state-space models, scales between different desired operation points and uses a set-theoretic control technique to solve admission control and resource allocation problems as a common decision problem with stability and robustness guarantees for the system under study.**

## I. Introduction

Modern virtualization technology [1] is the key factor for the consolidation of many distributed web services into large data centers. Cloud providers offer an execution platform with the essential means (i.e hardware infrastructure and software tools) that applications can use on demand. They must achieve the desired QoS of an application and minimize the use of resources and energy consumption. The workload of Internet services varies by orders of magnitude [2] while the requests have different service demands and time constraints. Additionally the underlying servers have constraints on the available computational resources. Thus, the dynamic behavior of web services is complex and nonlinear.

The proposed solutions in the literature can be categorized according to the chosen model, the control method and whether the admission control and resource allocation problems are solved jointly or not. When the modeling approach is based on queueing theory ( [3], [4], [5], [6], [7]), it relies on the assumption that the system is in steady state, which is contradictive to its dynamic nature. On the other hand, the studies adopting linear state-space models ( [8], [9], [10], [11]) cannot efficiently describe the system's complexity. From the aspect of the control methods used, approaches [8], [9], [10], [11] do not take into account system's constraints. The methods proposed in [8], [9], [5], [11] do not guarantee system stability and robustness. Moreover, the previous studies do not consider feasibility of the solution, meaning that it is not examined if there is a control value guaranteeing convergence to the desired operating point. Authors of [9], [5] tackle only the resource allocation problem, possibly leading to an overloading effect when the admitted workload is too large or the system's resources reach their upper limit. On the other hand, studies [8] and [11] proposed admission control techniques with static resource allocation that must reject a large portion of requests in order to avoid overloading. In order to overcome these shortcomings of the above system modeling and control, it is essential to design an autonomic control framework, which should include the following parts: a system identification tool, which models the system's dynamic behavior under any workload conditions and a dual controller, which considers admission control and resource allocation as a common decision problem, scales among different reference points and does not violate the system's constraints.

This paper presents the ACRA (Admission Control and Resource Allocation) framework, which is a time efficient optimization modeling and control methodology that combines resource allocation and admission control. The objective is to maximize the admittance of customers to the provided service, while satisfying customers' QoS requirements. Although ACRA does not consider power management as basic target, it includes the minimization of the necessary computing resources. Instead of a single linear state space model, we use a group of linear state-space models with additive uncertainties in order to cover the variation of workload and quantify the system's nonlinearities. For the design of the dual controller, we use a novel set-theoretic control theory which provides stability and robustness guarantees. Finally, ACRA can adapt between several reference points, because it is guaranteed to drive the system in the neighborhood of a feasible equilibrium point. The rest of the paper is structured as follows: Section II contains an analytical description of the system identification and control scheme. In section III the implementation and evaluation of the approach is presented. Conclusions and future work are outlined in section IV.

## II. Problem Description

### A. System Modeling and Identification

We use a group of linear state space models that covers the range of incoming workload. We assume there are $m$ applications, $n$ servers and each application has one Virtual Machine (VM) on every server. For each application, the incoming request rate is between a minimum ($L_{min_i}$) and a

978-3-901882-48-7 

maximum value ($L_{max_i}$). We suppose that $L_{min_i}$ corresponds to a small positive value and $L_{max_i}$ corresponds to the maximum served request rate when all VMs of the application get their maximum CPU capacity. Next, we divide each interval $[L_{min_i}, L_{max_i}]$ for every application $i = 1, \ldots, m$ in $p$ smaller equidistant parts, e.g. spanning a range of 50 requests each. Thus, combining all applications, a number of $N = p^m$ regions are defined in the "request rate" space. For each region, we assign a linear time invariant (LTI) model affected by additive disturbances denoted by $M_q,\ q = 1, \ldots, N$, which describes the system's dynamics, when the request rates are inside region $i$, i.e.

$$M_q:\ rt(k+1) = A_q rt(k) + B_q u(k) + \eta(k),\ q = 1, \ldots, N \tag{1}$$

where $rt \in \mathbb{R}^m$ is the vector of response time of all $m$ applications (state vector), $u \in \mathbb{R}^{(n+1)m}$ is the input vector that contains the $nm$ VM capacity inputs $cap_{ji}$ and the $m$ admitted request rate variables $lad_i$. Since we use only LTI models, the effect of nonlinearities and uncertainties is represented by the time-varying additive term $\eta \in \mathbb{R}^m$, which is considered here as unknown but bounded disturbance. We use the Recursive Least Square (RLS) algorithm [12] to identify the nominal disturbance-free LTI models,

$$M_q^*:\ rt(k+1) = A_q rt(k) + B_q u(k),\ q = 1, \ldots, N \tag{2}$$

The bounds of vector $\eta(k)$ of each model $M_q$ are determined by the minimal and maximum error of the RLS algorithm during the identification of the corresponding nominal model $M_q^*$.

### B. System's Constraints

There are physical constraints of the state and input variables. The average response time on a specific time interval varies from zero until the value when the system is saturated or the request has expired due to network constraints (TCP). Thus, the state variables are bounded to the constraint set $\mathbb{X}$ of the form

$$\mathbb{X} = \{rt \in \mathbb{R}^m :\ 0 \le rt_i \le rt_{max},\ i = 1, \ldots, m\} \tag{3}$$

Input constraints concern the restrictions on the VM capacity and the admitted request rate. Many VM's can be deployed on a physical server and they share its CPU capacity. For each server $j$, the CPU capacity assigned to each VM varies from a minimum value $c_{min}$ until a maximum $c_{max}$. Additionally the sum of VMs' entitlement should not exceed the total capacity of the server $c_{smax_j}$:

$$\begin{array}{ll} c_{min} \le cap_{ji} \le c_{max} & ,i = 1, \ldots, m\ , j = 1, \ldots, n \\ \sum_{i=1}^{m} cap_{ji} \le c_{smax} & ,i = 1, \ldots, m\ , j = 1, \ldots, n \end{array} \tag{4}$$

The incoming load of each application varies from $L_{min_i}$ to $L_{max_i}$. For each local model the admitted request rate is between $l_{min_i}$ and $l_{max_i}$ values, i.e.

$$l_{min_i} \le lad_i \le l_{max_i},\ \ i = 1, \ldots, m. \tag{5}$$

Unifying the input constraints, we can state that $u \in \mathbb{U}$, where

$$\begin{array}{ll} \mathbb{U} = \{u \in \mathbb{R}^{(n+1)m} :\ & c_{min} \le cap_{ji} \le c_{max} \\ & \sum_{i=1}^{m} cap_{ji} \le c_{smax} \\ & lmin_i \le lad_i \le lmax_i, \} \end{array} \tag{6}$$

Modeling uncertainting are bounded by the minimum and maximum values of the modeling error. Thus, the uncertainties $\eta$ are confined in set $\mathbb{N}$, where

$$\mathbb{N} = \{\eta \in \mathbb{R}^m :\ \eta_{min_i} \le \eta_i \le \eta_{max_i},\ i = 1, \ldots, m\}. \tag{7}$$

### C. Determination of the Equilibrium points

Due to workload fluctuations and the time varying availability of the system's resources, the provider should have the elasticity to dynamically arbitrate the incoming workload and the capacity assigned to each VM by determining a feasible equilibrium point that does not violate the SLO. In order to provide this capability, we allow the provider to switch between many equilibrium points, making the determination of feasible equilibrium point a decision problem. Given a nominal model $M_q^*$, a desired response time vector $RT_{ref} \in \mathbb{X} \subseteq \mathbb{R}^m$ and an admitted request rate is decided by a diurnal pattern $L_{ref} = [lad_{iref}]^T,\ i = 1, \ldots, m$. A suitable reference input vector $U_{ref} = [CAP_{ref}\ L_{ref}]^T, U_{ref} \in \mathbb{U} \subseteq \mathbb{R}^{(n+1)m}$ can be determined such that $RT_{ref} = ART_{ref} + BU_{ref}$. In order to optimally choose a feasible equilibrium point the following linear program is solved:

$$\min_{cap_{ji}, lad_i, d_i} \{W_c \sum_{j=1}^{n}\sum_{i=1}^{m} cap_{ji} + W_d \sum_{i=1}^{m} d_i\} \tag{8a}$$

*subject to*

$$RT_{ref} = A\,RT_{ref} + BU_{ref} \tag{8b}$$

$$RT_{ref} \in \mathbb{X} \subset \mathbb{R}^m \tag{8c}$$

$$u \in \mathbb{U} \subset \mathbb{R}^{(n+1)m} \tag{8d}$$

$$-d_i \le lad_i - L_{ref_i} \le d_i \tag{8e}$$

$$d_i \ge 0, \tag{8f}$$

where $W_c$ and $W_d$ are the weights of the cost function, which allows a trade-off between minimizing the VMs' capacity and achieving the desired admitted rate. We define variables $d_i$ to relax the definition of equilibrium point. Because of (8e),(8f), if the solution of the problem $U_{ref}^*$ of (8b) does not correspond to $L_{ref}$, it will correspond to the admitted request rate variable $L_{ref}^* = [lad_i]^T,\ i = 1, \ldots, m$, which is closest to $L_{ref}$. The above formulation assures that even if $L_{ref}$ cannot be achieved, it will find the closest admitted rate to $L_{ref}$.

### D. ACRA Control Framework

Admission control and resource allocation are two decision problems that should be considered jointly. Along this direction, we present a novel controller that stabilizes the system near to the desired equilibrium point. Most of the proposed solutions try to stabilize the system on a specific reference point. This is impossible because the system's dynamics involves nonlinearities in the state space description as well as time-varying additive terms. Thus, it might not be possible to find

a control law ensuring asymptotic stability with respect to an equilibrium point. Instead set-theoretic approaches deal with the stability analysis and control design problem, identifying and characterizing subsets of the state space containing the desired equilibrium state with special properties: robust positively invariant sets, disturbance invariant sets, or ultimately bounded sets For more information, the interested reader can refer to [13], [14], [15] and [16].

Our goal is to design for each model $M_q,\ q = 1, \ldots, N$, corresponding to a given workload profile, an affine state-feedback control law and compute a domain $\Delta \subseteq \mathbb{R}^m$ of the state-space such that all trajectories starting from $\Delta$ are transferred to a target set $R$, which contains the equilibrium point, in a finite time and remain in it. In contrast to classical control problem formulations, we want to confine all trajectories in a target set $R$ and not to drive it to the equilibrium point since the system is affected by additive disturbances. The resulting closed-loop system is said to be ultimately bounded in $R$ from $\Delta$ [15], [16]. In order to apply these methods, initially we apply a coordinates transformation to obtain a transformed nominal linear model with the equilibrium point to the origin:

$$\begin{array}{c} z(k) = rt(k) - RT_{ref} \\ v(k) = u(k) - U_{ref} \\ \overline{M_q}:\ z(k+1) = Az(k) + Bv(k),\ q = 1, \ldots, N \end{array} \tag{9}$$

The state and input constraints are transformed accordingly:

$$z \in \overline{\mathbb{X}} = \{z \in \mathbb{R}^m :\ rt_{min} - RT_{ref} \leq z \leq rt_{max} - RT_{ref}\} \tag{10}$$

$$v \in \overline{\mathbb{U}} = \{u \in \mathbb{R}^{(n+1)m} : u_{min} - U_{ref} \leq v \leq u_{max} - U_{ref}, \sum_{i=1}^{m} vcap_{ji} \leq c_{smax} - \sum_{i=1}^{m} cap_{jiref}\} \tag{11}$$

where $vcap_{ji} = cap_{ji} - cap_{jiref}$ transformed VMs'capacity inputs. We design a state feedback controller of the following form,

$$\begin{array}{ll} u(k) = K(rt(k) - RT_{ref}) + U_{ref} & \\ or \qquad v(k) = Kz(k) & \end{array} \tag{12}$$

Secondly, we must compute the target set $R$, which must possess the robust invariance property [13], meaning that if the closed-loop system trajectory is in $R$, it remains in it for all future instances. In our case, we obtain a candidate target set $S_1$ by applying the Jordan decomposition [17] of the closed-loop system $A + B\overline{K}$, where $\overline{K}$ is a gain matrix that places its eigenvalues inside the unit rhombus [18] in the eigenvalue space,

$$S_1 = \{z \in \mathbb{R}^m : G_1 z \leq w_1\} \tag{13}$$

where $G_1 \in \mathbb{R}^{2m \times m}$, $G_1 = [V^T - V^T]^T$, $w_1 \in \mathbb{R}^{2m}$ : $w_i = 1,\ i = 1, \ldots, 2m$. Matrix V is the transformation matrix transforming $A + B\overline{K}$ to its real Jordan form. The state and input constraints set $\overline{\mathbb{X}}$ and $\overline{\mathbb{U}}$ are formulated below using the standard half-space represantation of polytopes by $S_z$ and $S_v$,

$$\begin{array}{l} S_z = \{z \in \mathbb{R}^m : G_z z \leq w_z\} \\ S_v = \{v \in \mathbb{R}^{(n+1)m} : G_v v \leq w_v\} \\ where\ G_z \in \mathbb{R}^{2m \times m},\ w_z \in \mathbb{R}^{2m \times 1}, \\ G_v \in \mathbb{R}^{2n(m+1) \times mn},\ w_v \in \mathbb{R}^{2n(m+1) \times 1} \end{array} \tag{14}$$

The bounded uncertainties set $S_\eta$ is,

$$\begin{array}{l} S_\eta = \{\eta \in \mathbb{R}^m : G_\eta \eta \leq w_\eta\} \\ where\ G_\eta \in \mathbb{R}^{2m \times m} :\ G_\eta = \left[\begin{array}{c} I_{m \times m} \\ -I_{m \times m} \end{array}\right], \\ w_\eta \in \mathbb{R}^{2m} \end{array} \tag{15}$$

Thirdly, our goal is to compute a control law of the form (12) and a set $\Delta$ such that the closed-loop is ultimate bounded in $R$. We choose $\Delta$ to have the form,

$$\Delta = \{z \in \mathbb{R}^m : G_1 z \leq a w_1\} \tag{16}$$

Using the generalized Farkas' lemma [19], we can compute a control law $v = Kz$ and a set $\Delta$, which is the largest set fitted in $S_z, S_v$. This can be described by the optimization problem,

$$\begin{array}{rll} \max\limits_{K,H_1,H_2,H_3,\epsilon,a} \{a\} & or & \min\limits_{K,H_1,H_2,H_3,\epsilon,b} \{b\} \qquad (17a) \\ subject\ to & & subject\ to \\ G_1(A + BK) = H_1 G_1 & & G_1(A + BK) = H_1 G_1 \qquad (17b) \\ aH_1 w_2 + d \leq \epsilon a w_2 & & H_1 w_2 + bd \leq \epsilon w_2 \qquad (17c) \\ H_2 G_1 = G_z & & H_2 G_1 = G_z \qquad (17d) \\ aH_2 w_2 \leq w_z & & H_2 w_2 \leq b w_z \qquad (17e) \\ H_3 G_1 = G_v K & & H_3 G_1 = G_v K \qquad (17f) \\ aH_3 w_2 \leq w_v & & H_3 w_2 \leq b w_v \qquad (17g) \\ 0 \leq \epsilon \leq \epsilon^* & & 0 \leq \epsilon \leq \epsilon^* \qquad (17h) \end{array}$$

$$where\ d = \left[\begin{array}{c} \max\limits_i \{V\eta\}_i \\ \max\limits_i \{-V\eta\}_i \end{array}\right]$$

where $H_1,\ H_2,\ H_3$ are non-negative matrices, K is the feedback gain matrix, and the scalar $\epsilon$ is positive between zero and a subunitary value $\epsilon^*$ which is a metric of convergence speed. The optimization problem on the left is non-linear because of the first inequality, thus we define $b = \frac{1}{a}$ and reformulate the problem as linear on the right side. Relations (17b), (17c) guarantee the positive invariance and contractiveness. Relations (17d), (17e) guarantee $\Delta \subseteq \overline{\mathbb{X}}$ and relations (17f), (17g) guarantee $\Delta \subseteq S_{v(z)}$, where $S_{v(z)} = \{z \in \mathbb{R}^m : G_v K z \leq w_v\}$.

## III. Evaluation

We demonstrate an experiment, which shows the method's performance and compares it with two well-known approaches. We assume two applications ($m = 2$), which are hosted on the same CPU core in two separated VMs ($n = 1$). The incoming workload for both applications varies between $[175\ ,\ 325]$ req/s. Applying the system identification method described in section II, we divide the "request rate" space into hypercubes (squares for two-dimensions space) with a step of 50 requests and generate a local linear model for each square, such as in (1). The measurements and control signals are updated every 20 sec. Initially, we assume that the admitted request rate is high and the computing resources for this type of requests (e.g. browsing) are low. Thus the target value corresponds to the reference mean response time of 1 sec and the desired admitted request rate is 300 req/s for both

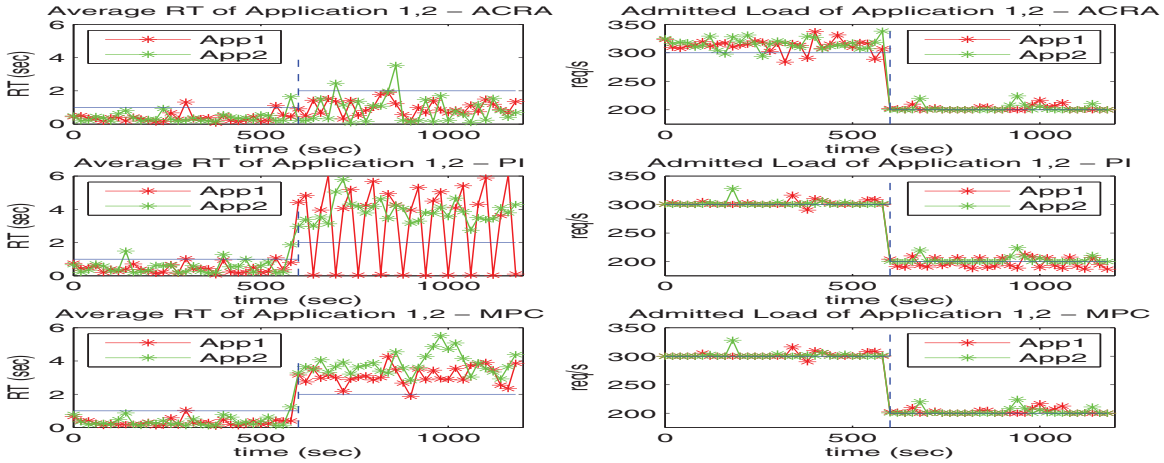


Fig. 1: Response Time and Admitted Request Rate.

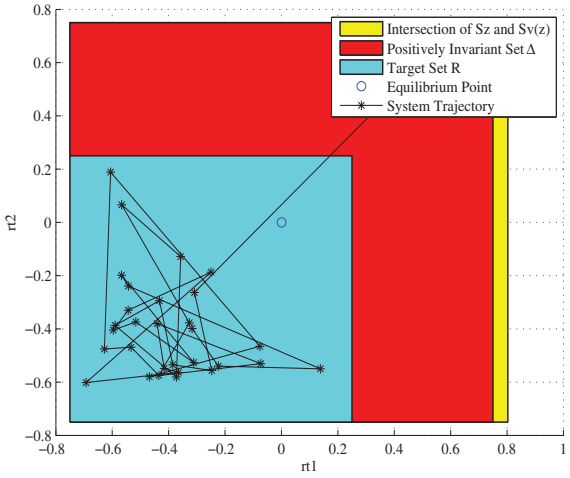


Fig. 2: System Trajectory and Sets

services. After a period the SLO target value changes, since the dominant type of transactions needs more CPU resources and has lower incoming request rate than before. The target of mean response time is 2 sec and desired admitted request rate is 200 req/s now. The modeling uncertainties varies between $[-0.4\ ,\ 0.4]$ for both candidate equilibrium points. We compare ACRA with the controllers of [9] and [11]. The first controller is a PI controller, while the second controller is obtained by applying a model predictive control (MPC) scheme [20]. We select the predictive horizon of the MPC controller to be $H = 5$ and the trade-off parameter of the approach $a = 0.6$. The cost function of the controller is $J = \sum_{k=K}^{K+H-1}(a||z(k)||_2^2 + (1-a)||v(k)||_2^2)$.

Fig. 1 shows the mean response time and the admitted workload for each control scheme. Although all controllers regulate the system near the first equilibrium point, the ACRA controller admits more requests due to its capability to refine the equilibrium point. For the second equilibrium point, the difference between ACRA and the rest approaches is obvious. PI and MPC controllers fail to drive the system near the equilibrium point for aforementioned reasons.

Fig. 2 is a graphical representation of the trajectories in the state space of the transformed model in (9) when the target value of the average response time for both applications is 1 sec. In order to succeed this reference response time, we choose as equilibrium point (blue circle) a value smaller than the target (0.75 sec), which is inside the target set $R$. The yellow set shows the feasibility set produced from the intersection of the state and input constraints set $S_z \cap S_{v(z)}$, the red set depicts the positively invariant set $\Delta$ and the cyan set is the target set $R$. As we observe, starting from the positively invariant set $\Delta$, the trajectory of the closed-loop system is confined in the target set $R$ in finite time.

## IV. Conclusions

ACRA framework models the system's dynamics using a group of linear state space models, which cover all the range of workload conditions, and using a set-theoretic technique it designs a state feedback controller that successfully leads and stabilize the system in the region of the equilibrium point, providing stability guarantee and robustness against system disturbances and nonlinearities. Future research is intended towards the reduction of the number of local models using an optimization criterion, the usage of a workload predictor and application of set-theoretic methods towards the computation of the minimal and maximal robust positively invariant set.

## References

[1] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, "Xen and the art of virtualization," in *Proc. of the ACM Symposium on Operating System Principles (SOSP)*, 2003, pp. 164 – 177.

[2] A. Williams, M. Arlitt, C. Williamson, and K. Barker, *Web Workload Characterization: Ten Years Later*. Springer, 2005.

[3] R. Urgaonkar, U. Kozat, K. Igarashi, and M. Neely, "Dynamic resource allocation and power management in virtualized data centers," in *Proc. of the IEEE Network Operations and Management Symposium (NOMS)*, 2010, pp. 479 – 486.
[4] J. Almeida, V. Almeida, D. Ardagna, I. Cuhna, and C. Francalanci, "Joint admission control and resource allocation in virtualized servers," *Elsevier Journal of Parallel and Distributed Computing*, vol. 70, no. 4, pp. 344–362, 2010.
[5] D. Kusic, J. Kephart, J. Hanson, N. Kandasamy, and G. Jiang, "Power and performance management of virtualized computing enviroments via lookahead control," *Springer Journal on Cluster Computing*, vol. 12, no. 1, pp. 1–15, 2009.
[6] D. Ardagna, C. Ghezzi, B. Panicucci, and M. Trubian, "Service provisioning on the cloud: Distributed algorithms for joint capacity allocation and admission control," in *Proc. of the 3rd European Conference ServiceWave*, 2010, pp. 1 – 12.
[7] D. Ardagna, B. Panicucci, M. Trubian, and L. Zhang, "Energy-aware autonomic reource allocation in multitier virtualized enviroments," *IEEE Transactions on Services Computing*, vol. 5, no. 1, pp. 2–19, 2012.
[8] N. Leontiou, D. Dechouniotis, and S. Denazis, "Adaptive admission control of distributed cloud services," in *Proc. of the IEEE Conference Network and Service Management (CNSM)*, 2010, pp. 318 – 321.
[9] X. Wang and Y. Wang, "Coordinating power and performance management for virtualized server clusters," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 2, pp. 245 – 259, 2011.
[10] P. Giani, M. Tanelli, and M. Lovera, "Controller design and close-loop stability analysis for admission control in web service systems," in *Proc. of the 18th IFAC World Congress*, 2011, pp. 6709 – 6714.
[11] C. Poussot-Vassal, M. Tanelli, and M. Lovera, *A Control-Theoretic Approach for the Combined Management of tje Quality-of-Service and Energy is Service Centers*. Springer, 2010.
[12] P. E. Wellstead, E., and M. B. Zarrop, *Self-Tuning Systems: Control and Signal Processing*. John Wiley & Sons, Inc., 1991.
[13] F. Blanchini and S. Miani, *Set theoretic methods in Control*. Birkhauser, 2008.
[14] F. Blanchini, "Set invariance in control," *Automatica*, vol. 35, no. 11, pp. 1747 – 1767, 1999.
[15] N. Athanasopoulos, G. Bitsoris, and M. Vassilaki, "Ultimate boundedness and robust stabilization of bilinear discrete-time systems," in *Proc. of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, 2011, pp. 4622–4627.
[16] F. Blanchini, "Ultimate boundedness control for uncertain discretetime systems via set-induced lyapunov functions," in *Proc. of the 30th IEEE Conference on Decision and Control(CDC)*, 1991, pp. 1755 – 1760 vol.2.
[17] K. Ogata, *State space analysis of control systems*. Prentice-Hall, 1967.
[18] G. Bitsoris, "Positively invariants polyhedral sets of discrete-time linear systems," *International Journal of Control*, vol. 47, no. 6, pp. 1713 – 1726, 1988.
[19] G. Bitsoris and M. Vassilaki, "Constrained regulation of linear systems," *Automatica*, vol. 31, no. 2, pp. 223 – 227, 1995.
[20] J. Maciejowski, *Predictive Control with Constraints*. Prentice-Hall, 2001.