Dynamic Trade-off Analysis of QoS and Energy Saving in Admission Control for Web Service Systems

C. Poussot Vassal M. Tanelli M. Lovera

VALUETOOLS '09: Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools, Pisa, Italy, 2009

Summarized by: Sean (Shahin) M. Ansari

Introduction

* Increasing Complexity
* Energy Management
* QoS for Variable Workload
* Control Theoretic Techniques
* Dynamic Voltage Scaling
* Admission Control
* Linear Parametrically Varying model vs. Queuing Theory

Problem Statement & Notation

* Average request arrival rate λ_k
* Average requests service time S_k
* Average Server response time T_k
* Average Web Service throughput X_k
* Effective CPU frequency f_k
* Probability of Admission $\mathscr{P}_k$
* Queuing Time ξ_k

Problem Statement

$$s_{f,k} = s_k / f_k$$

$$X_k = \mathscr{P}_k \lambda_k$$

$$T_k = s_{f,k} + \xi_k$$

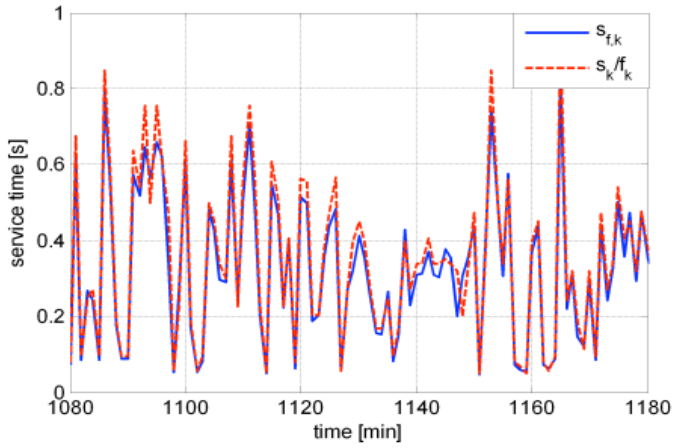


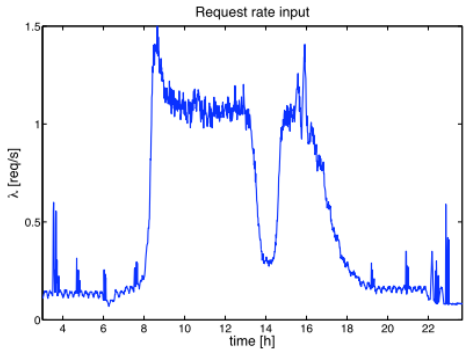
Figure 1: Experimental verification of $s_{f,k} = s_k/f_k$ on a Web server with DVS functionalities.

LPV State Space Models

* LPV vs. LTI – Linear Time Invariant
* Theta = $$\begin{aligned} x_{k+1} &= A(p_k)x_k + B(p_k)u_k \\ y_k &= C(p_k)x_k + D(p_k)u_k, \end{aligned}$$
* Identification – Minimize Cost $$V_N(\theta) := \sum_{k=1}^{N} ||y_k - \hat{y}_k(\theta)||_2^2 = E_N^T(\theta)E_N(\theta)$$

$$E_N^T(\theta) = \left[\left(y_1 - \hat{y}_1(\theta)\right)^T \quad \cdots \quad \left(y_N - \hat{y}_N(\theta)\right)^T \right]$$

* Request Rate applied to LPV:



Identification

* Admission Control Identification $\sigma[s_k] = 4E[s_k]$
* Admission Probability via Server Utilization results

$$\rho_{ac,k} = \lambda_k s_k \mathscr{P}_k)$$

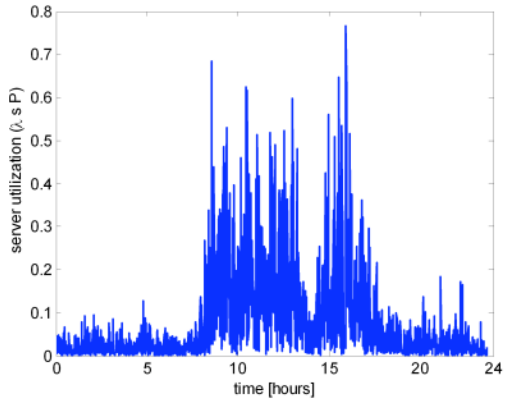


Figure 3: Time history of the server utilization $\rho_{ac,k} = \lambda_k s_k \mathscr{P}_k$ used in the identification experiments for admission control.

* Variability Test with $\sigma[s_k] = qE[s_k]$ and $q = \{2,6\}$

LPV model validation

* Metrics for Quantitative evaluation of identification and validation:
* Percentage Variance Accounted for: $$VAF = 100\left(1 - \frac{Var[y_k - y_{sim,k}]}{Var[y_k]}\right)$$
* Percentage Average Error: $$e_{avg} = 100\left|\frac{E_t[y_k - y_{sim,k}]}{E_t[y_k]}\right|$$
* Results:

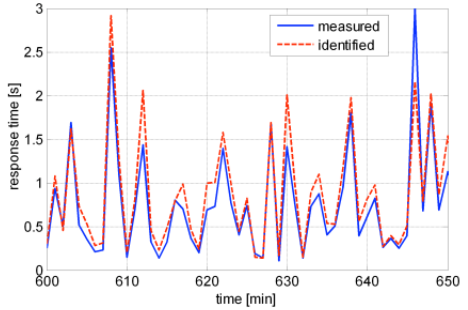


Figure 4: Detail of the measured (solid line) and simulated (dashed line) response time obtained with an LPV-IA model for the admission control dynamics on validation data.

Valid. Performance $\Delta t = 1$ min	$q = 2$	$q = 6$
VAF on 24h	78.38%	74.96%
VAF light load	92.63%	83.01%
VAF heavy load	73.79%	63.57%
e_{avg} on 24h	3.35%	6.60%
e_{avg} light load	0.42%	2.85%
e_{avg} heavy load	5.48%	10.74%

Table 1: Performance of the identified models for admission control with $\Delta t = 1$ min on validation data.

Model Predictive Control

* Guaranteed service time T_{ref} + Maximum users served
* Or guaranteed service time T_{ref} + Minimum power consumption
* Optimal performance calculated via optimization algorithm inspired via LPV-MPC
* MPC – Control in terms of optimization, Cost function, and relevant input, state, and output variables.
* Iterative calculation of optimization problem

Optimal Performance Analysis

* Control Variables: admission probability and effective service time
* Assumptions: request rate, and response time known for near future
* LPV system is a model of web server
* State variables X_k is known

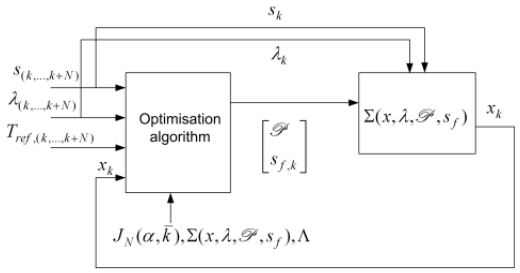


Figure 5: Optimal performance computation scheme.

Performance Calculations

* Cost Function Definitions

$$J_N(\alpha,\bar{k}) = \alpha J_{QoS}(\bar{k}) + (1-\alpha) J_{ES}(\bar{k}) =$$
$$= \alpha \sum_{k=\bar{k}}^{\bar{k}+N-1} \left| \frac{\overline{\mathscr{P}} - \mathscr{P}_k}{\overline{\mathscr{P}} - \underline{\mathscr{P}}} \right| + (1-\alpha) \sum_{k=\bar{k}}^{\bar{k}+N-1} \left| \frac{\bar{s}_f - s_{f,k}}{\bar{s}_f - \underline{s}_f} \right|, \quad (7)$$

* Dynamic Equality Constraints

$$\begin{aligned} x_{k+1} &= Ax_k + \left(B_0 + B_1 s_{f,k} + B_2 s_{f,k}\lambda_k\right)\mathscr{P}_k \\ \xi_k &= Cx_k + \left(D_0 + D_1 s_{f,k} + D_2 s_{f,k}\lambda_k\right)\mathscr{P}_k \\ T_k &= \xi_k + s_{f,k}, \end{aligned}$$

Performance Calculations

* Input and Performance Inequality Constraints

$$\Lambda : \begin{cases} 0 \leq \xi_k \\ \underline{\mathscr{P}} \leq \mathscr{P}_k \leq \overline{\mathscr{P}} \\ \underline{s}_f \leq s_{f,k} \leq \overline{s}_f \\ -\Delta \leq T_k - T_{ref} \leq \Delta \end{cases}$$

* LPV-MPC Optimization Problem

$$J_N^*(\alpha, \bar{k}) = \min\ J_N(\alpha, \bar{k})$$

$$\text{subject to} \begin{cases} \begin{bmatrix} x_{k+1} \\ \xi_k \\ T_k \end{bmatrix} = (8) \\ \Lambda = (9) \end{cases} \quad k \in [\bar{k}, \bar{k}+N-1].$$

Simulation Results

* Alpha = 0
* P_{es} <> max , min
* Alpha = .25 is equilibrium

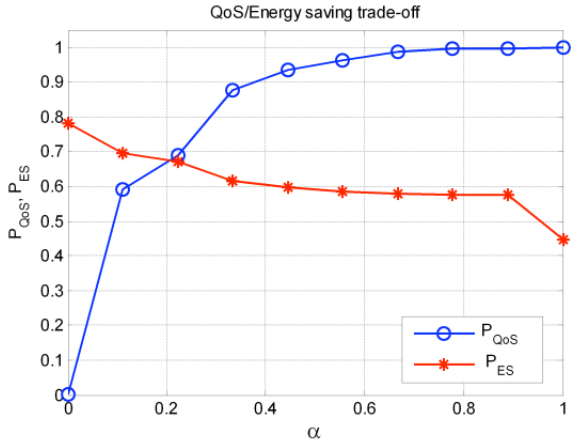


Figure 6: Plot of the performance measures P_{Qos} and P_{ES} as functions of α for $N = 20$.

Final Results