Multi-step ahead response time prediction for single server queuing systems

Amani, Payam; Kihl, Maria; Robertsson, Anders

Published in:
[Host publication title missing]

2011

Multi-step ahead response time prediction for single server queuing systems

Payam Amani, Maria Kihl
Department of Electrical and Information Technology
Lund University
Lund, Sweden
Email: Payam.Amani@eit.lth.se, Maria.Kihl@eit.lth.se

Anders Robertsson
Department of Automatic Control
Lund University
Lund, Sweden
Email: Anders.Robertsson@control.lth.se

***Abstract*—Multi-step ahead response time prediction of CPU constrained computing systems is vital for admission control, overload protection and optimization of resource allocation in these systems. CPU constrained computing systems such as web servers can be modeled as single server queuing systems. These systems are stochastic and nonlinear. Thus, a well-designed nonlinear prediction scheme would be able to represent the dynamics of such a system much better than a linear scheme. A nonlinear autoregressive neural network with exogenous inputs based multi-step ahead response time predictor has been developed. The proposed estimator has many promising characteristics that make it a viable candidate for being implemented in admission control products for computing systems. It has a simple structure, is nonlinear, supports multi-step ahead prediction, and works very well under time variant and non-stationary scenarios such as single server queuing systems under time varying mean arrival rate. Performance of the proposed predictor is evaluated through simulation. Simulations show that the proposed predictor is able to predict the response times of single server queuing systems in multi-step ahead with very good precision represented by very small mean absolute and mean squared prediction errors.**

I. Introduction

Computing systems enable the Telecom operators to provide their customers with a vast variety of services which are aimed to meet their demands and desires. An operator usually uses a network of several such computing systems to facilitate providing the end users with an ever growing variety of services. Optimization of resource allocation in computing systems has attracted much interest in recent years as it directly relates to the performance of these systems.

Node elements (NEs) as building blocks of this network of computing systems, have a requirement for secure, reliable and real-time activation, modification and deactivation of both new and current customers or services. These tasks should be performed fast and in an automated manner. A resource access conflict exists among performing those tasks and providing current customers with their requested services in the network. This fact raises the necessity for a new enterprise provisioning system in the network which is hereby named as management system (MAS). The MAS is equipped with an admission control mechanism which enables it to avoid the resource access conflict by delaying sending of requests to a highly loaded NE and protecting it from becoming overloaded [1]–[3]. This control mechanism usually includes a feed forward controller as it should predict the resource access conflict well before it happens and take action to avoid it. Therefore there is a requirement for a multi-step ahead state predictor for the NEs which precisely represent the dynamics of the NE in its whole operation range. NEs are desired to be loaded as much as possible close to their capacity meanwhile protected from becoming overloaded. One of the main performance measures of computing systems is the response time of the requests sent to them. Businesses and their customers like to minimize the system's response times while maximizing system utilization. Doing this, users will have a positive experience during delivery of services which would lead to increased customer retention and revenue.

It has been shown in [4] that CPU constrained computing systems such as web servers with dynamic content can be modeled as single server queuing systems. These are nonlinear stochastic systems. A nonlinear model is much more capable of representing the dynamics of a single server queuing system compared to a linear model [5]. Many attempts to develop analytical estimators or predictors for single server queuing systems have been presented in the literature. Clarke, in his pioneer work published in 1957, presented maximum likelihood estimator (MLE)'s of arrival and service rates [6]. Basawa et al. in [7] have presented a maximum likelihood estimator for single server queues from waiting time data. In [8], Zheng and Seila have investigated some popular performance measures like waiting time and queue length under frequentist setup and showed their undesirable characteristics like nonexistence of expected value of the estimator and infinite mean-squared error of the estimator. Further, they proposed a set-up to fix that property. For the first time, McGrath et al. in [9], [10] have applied the concept of Bayesian statistical inference to the $M/M/1$ queuing system. Their work has been considerably extended in [11], [12].

The above mentioned analytical approaches to the estimation of single server queuing systems have some unfavorable characteristics for overload protection admission control schemes. Firstly, all of the above mentioned estimation methods can only be applied to steady state and stationary scenarios. Secondly, mean service and arrival rates are assumed to be constant and time invariant. However, in the real world, there are many cases where we are interested in a state estimator that can be applied to a CPU constrained computing system with at least one time varying parameter. Time varying mean arrival rate can be a good example of these parameters. Finally it should be noted that none of the above mentioned methods support multi-step ahead prediction.

The requirement for a nonlinear multi-step ahead response time predictor that can work under stationary and steady state scenarios as well as time varying and non-stationary scenarios led us to a black box approach to identification of single server queuing systems. By means of a nonlinear autoregressive network with exogenous inputs (NARX) neural network we have designed a predictor that covers all the above mentioned characteristics that the other methods lack and also is able to predict the response times of single server systems with very good precision represented by very small mean absolute and mean squared prediction errors.

This paper is structured as follows. System configuration containing the use case scenario, the NARX neural network and the predictor is investigated in section II. Section III is dedicated to specifications of simulation environment and scenarios. Simulation results are summarized in section IV. Finally section V concludes the paper.

II. System Configuration

This section covers three sub-sections. In subsection (A) the pilot system for which a nonlinear multi-step ahead predictor is developed is introduced. Sub-section (B) is dedicated to the introduction of NARX recurrent neural networks. The proposed NARX multi-step ahead response time predictor is presented in sub-section (C).

A. Management System and Node Elements

Communication and computer networks are the media used by Telecom operators to inspire their subscribers with an ever growing variety of services. These networks are usually interconnected and operators provide services to their customers via several of them. The MAS is responsible for real-time, secure and reliable activation, modification and deactivation of subscribers and services in an automated manner. Such management system interacts with many NEs in the network and is usually implemented in distributed server clusters. Figure 1 depicts a generic distributed service management system.

The interactions between the MAS and the NEs should not lead the high loaded NEs to become overloaded. This brings the necessity of an admission control mechanism for overload protection of the NEs into picture. As a real network usually consists of various parts which are provided by several vendors with their own protocol implementations, the admission control mechanism should be implemented in the MAS. Also, it should be based on measurements which can

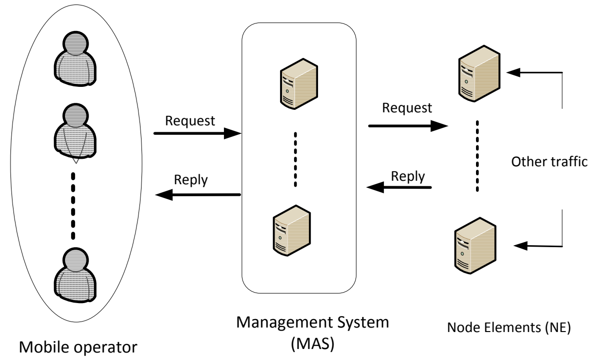


Fig. 1: A generic distributed service management system.

be provided without a need for changing the current protocols and operating systems. A close review of the described setup led us to figure out three measurements that have the above mentioned characteristics. These include inter-arrival times, inter-departure times and response times of the services sent to the node elements from the service management system. These measurements can be retrieved from the time tagged logs of requests traveling in the network. The response time of a service is defined as the duration of time that it spends from the moment that it leaves the MAS for the NE to the time that it leaves the NE. A high response time corresponds to a highly loaded NE and a low response time to a lightly loaded one. Thus response time can be used as an indicator for the NEs' internal state.

The control mechanism for overload protection of the NEs is in the form of admission control so that NEs' traffic from the subscribers is given higher priority compared to the traffic sent from the MAS to the NEs. In case that the NE is heavily loaded and tending to become overloaded, the MAS will back off sending more requests to the NE allowing it to process some of them and reduce its load. As the control action should take place well before an overload occurs in the network, the admission control scheme will consists of not only a feedback loop but also a feed forward part. The requirement for a feed forward controller raises the need for a multi-step ahead predictor for the NEs. In this paper we focus on interaction of one Management Server with one NE. Configuration of the computing system which is to be investigated in this paper is illustrated in Figure 2.

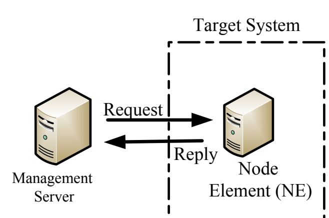


Fig. 2: The pilot computing system configuration.

NEs can be modeled as single server queuing systems. In this paper the problem of nonlinear multi-step ahead prediction of response times of single server queuing systems is investigated. Figure 3 illustrates a single server queuing system in which the distribution of the inter-arrival times and service times are general. The mean arrival rate and the mean service rates of the queuing system are denoted by λ and μ respectively.

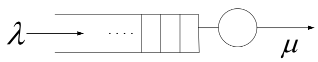


Fig. 3: A single server queue with mean arrival rate λ and mean service rate μ.

B. NARX Neural Network

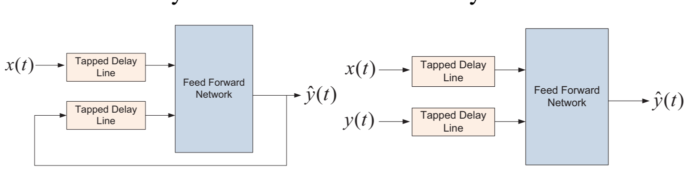
Recurrent neural networks have been widely used for modeling of nonlinear dynamical systems [13], [14]. Among various types of the recurrent neural networks such as distributed time delay neural networks (TDNN) [13], layer recurrent networks [13] and NARX [13], the latest is of great interest in input output modeling of nonlinear dynamical systems and time series prediction [15]–[19].

NARX is a dynamical recurrent neural network based on the linear ARX model. The next value of the dependent output signal $y(t)$ is regressed over the latest n_x values of the independent input signal and n_y values of the dependent output signal. n_x and n_y respectively represent the dynamical order of the inputs and outputs of the NARX. A mathematical description of the NARX model is summarized in (1) in which f is a nonlinear function.

$$y(t) = f(y(t-1), y(t-2), \dots, y(t-n_y), x(t-1), \\ , x(t-2), \dots, x(t-n_x)) \tag{1}$$

A NARX neural network can be implemented in two set-ups namely parallel and series-parallel architectures. These are depicted in Figure 4. In this paper we have used the series-parallel architecture in which during the training period the actual values of the output are fed back to the neural network. This will improve the training precision.

This network consists of three main layers namely input layer, hidden layer and output layer. The input layer consists of the current and previous inputs and previous outputs. These are fed into the hidden layer. The hidden layer consists of one or several neurons resulting in a nonlinear mapping of affine weighted combination of the values from input layer. The output layer consists of an affine combination of the values from hidden layer. In this network the dynamical order of



(a) Parallel (b) Series-Parallel

Fig. 4: NARX (a): Parallel and (b): Series-Parallel Architectures.

inputs and outputs and number of neurons in each layer are pre-determined. Several methods for determination of these values are presented in [13]. A suitable training algorithm and performance measure also should be chosen. Finally, the type of the nonlinear map needs to be defined.

Some pre and post processing on the input and target values should be performed in order to have a valid training [13]. These processes include mapping of the input and target data to values in range of $[-1, 1]$, normalization of the inputs and targets to have zero mean and unity variance and removal of constant inputs and outputs and processing of unknown inputs. As these measurements are very noisy, after normalization we filter both input and target values with a designed Butterworth low pass filter. The bandwidth of the filter is chosen so it suppresses noise as much as possible while not affecting the characteristics of in band part of input and output data sets.

C. NARX Multi-Step Ahead Response Time Predictor Set-up

Our application requires the prediction of response times of the single server queuing system in some time steps into the future before the actual output measurements become available. A black box identification approach was chosen to predict the response time of a service sent to a NE by the Management Server from three measured time values namely inter-arrival, inter-departure and response times of the services. The input vector consists of current inter-arrival times and inter-departure times as two inputs. Output of the neural predictor is the predicted response time. Measured response times are required for training and evaluation of the NARX multi-step ahead response time predictor and are fed back to the input layer of the proposed predictor. Measured data is divided into training, evaluation and test data sets. Prediction horizon m is defined as the shift between corresponding inputs and output values so that current input is used for prediction of output in m time steps in the future. The proposed multi-step ahead response time predictor set-up is illustrated in Figure 5.

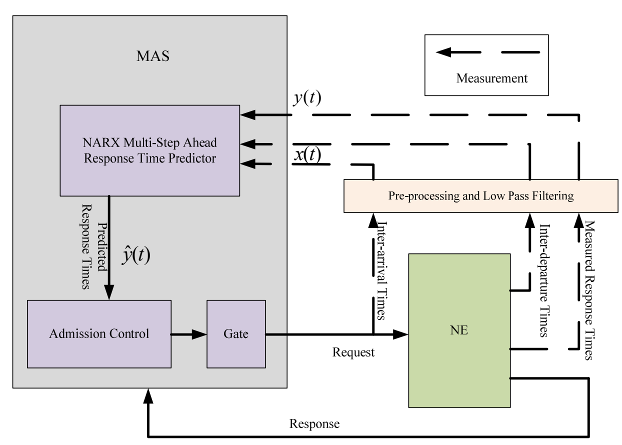


Fig. 5: Multi-Step ahead Response Time Predictor Set-up.

The overload protection admission controller uses a gate for controlling the flow of requests to the NE. The flow of requests from the MAS to the NE cannot get negative values as negative requests do not exist. Also the gate cannot send more requests than the available requests in the MAS. This imposes an input nonlinearity to the NE. We already know that NEs are nonlinear and stochastic computing systems. Thus, a NARX based predictor as a nonlinear predictor has a much better opportunity to grasp the dynamics of the response times of the NE compared to linear predictors [5].

The off-line training process is described as follows. The NE is simulated under high load conditions. The acquired data is then divided to training, validation and test data sets. The NARX multi-step ahead response time predictor is trained using the train data and training is validated and its performance is tested using validation and test data sets. This trained predictor is used for all static and dynamic load conditions. Performance of the predictor is investigated in the following section.

III. Simulation environment and Scenarios

The simulation environment and the simulation scenarios are defined in the sub-sections (A) and (B) respectively.

A. Simulation environment

The NE and the Management Server are implemented in the MATLAB Simulink tool for event based simulations called SimEvents. Considering the pilot system configuration, the MAS is simulated by a request generator and the NE is simulated by a single server queuing system. The NARX multi-step ahead response time predictor is designed using MATLAB Neural Network Toolbox. The low pass filter is designed using the Filter Design Toolbox of MATLAB. All the mentioned tools are included in MATLAB R2010b.

B. Simulation Scenarios

In order to test the performance of the multi-step ahead NARX response time predictor we apply it to the $M/M/1$ queuing system [20]. Also we consider 4 main test scenarios plus an extra scenario which deals with smoothly changing mean arrival rate.

1) Response Time Predictor's Parameters: The NARX multi-step ahead response time predictor is configured as follows. The two dimensional input vector $x(t)$ consists of inter-arrival and inter departure times. The one dimensional output $y(t)$ represents the predicted response times.

The measured response times are fed back to the input layer for training. The tapped delay line in the hidden layer consists of two delays. Three neurons are considered in the hidden layer. The hidden layer neuron's activation function is considered as tangential sigmoid function $tansig$. The output layer consists of one neuron with activation function chosen as linear function $purelin$. This is summarized in Figure 6.

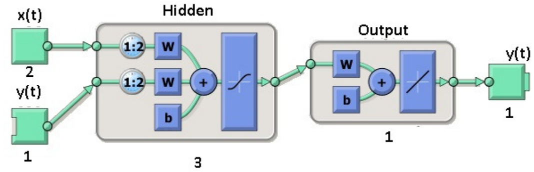


Fig. 6: Neural predictor configuration and parameters.

Several criteria such as sampling time, control structure and constraints affect the choice of the prediction horizon m. As in this paper we only focus on response time prediction of a single server system regardless of the control structure and other criteria that affect the choice of the prediction horizon, a suitable value for m cannot be decided here. However as we know that a multi-step ahead prediction is required we chose $m = 4$ to show that the NARX response time predictor is able to predict the response times of the single server queuing systems in several time steps into the future.

The Levenberg-Marquardt algorithm [13] is chosen as the training algorithm and the performance metric is set to mean squared error (MSE). The predictor is trained with the data from the high load scenario then tested over high load, low load and two varying load conditions.

2) Single Server Queuing System Parameters: Our test set includes a $M/M/1$ queuing system with the following parameters. Mean service rate is set to 1 for all the scenarios. Simulation time is set to 20000. All time values are in simulated seconds. The static scenarios are designed for evaluation of performance of the NARX response time predictor in steady state under low ($\rho = 0.30$) and high ($\rho = 0.95$) load conditions. The dynamical scenarios are meant to evaluate performance of the NARX response time predictor with arrival rate changing from high to low load or vice versa with a step function at time 5000. The extra simulation scenario covers time variant mean arrival rate as a saturated ramp function starting from low($\rho = 0.30$) load at start time and saturating at high ($\rho = 0.95$) load.

Static Scenarios:

- S1: Mean arrival rate is set to 0.95 to simulate a high ($\rho = 0.95$) load scenario. The NARX multi-step ahead response time predictor is then trained, validated and tested using train, validation and test data sets.

- S2: Mean arrival rate is set to 0.3 to simulate a low ($\rho = 0.3$) load condition. Performance of the NARX multi-step ahead response time predictor is tested using the acquired data.

Dynamic Scenarios:

- S3: Mean arrival rate is a step, denoted by $step1$, from 0.3 to 0.95 with the step time set to 5000. Performance of the NARX multi-step ahead response time predictor is tested using the acquired data.

- S4: Mean arrival rate is a step from 0.95 to 0.3 with the same step time as before shown as $step2$. Performance of the NARX multi-step ahead response time predictor is tested using the acquired data.

IV. Simulation Results

Performance of the proposed predictor applied to the defined M/M/1 queuing system is summarized in Table I. In this section MAE stands for mean absolute error and MSE stands for mean squared error. It should be considered that the data used in these simulations is normalized to its maximum value. This is the reason why the maximum value of the response times is equal to one.

TABLE I: Performance (MAE and MSE) of NARX m step ahead response time predictor for $M/M/1$ queuing system in S1 to S4 scenarios with prediction horizon m set to 4.

Performance of the Proposed Predictor in Training Phase			
Scenario	Server load ρ	Performance Measure	Value
S1	$\rho = 0.95$	MSE	$5.1915e-10$
		MAE	0.00174
Performance of the Proposed Predictor in Testing Phase			
Scenario	Server load ρ	Performance Measure	Value
S1	$\rho = 0.95$	MSE	$1.3728e-9$
		MAE	0.002
S2	$\rho = 0.30$	MSE	$4.1312e-9$
		MAE	0.0163
S3	$\rho = step1$	MSE	$2.585e-9$
		MAE	0.002
S4	$\rho = step2$	MSE	$1.8481e-7$
		MAE	0.0037

As it can be seen from the results in TABLE I, the multi-step ahead NARX response time predictor is well trained and shows a promising performance under S1 and S2 considering both MSE and MAE. This shows that the proposed response time predictor is able to accurately predict the response time of the described $M/M/1$ system in 4 steps ahead under static and steady state load conditions.

Performance of the response time predictor under dynamic load conditions especially its performance in transient load conditions is of interest in the following tests. Under S3 both MSE and MAE indicate very good performance of the predictor. Figure 7 depicts measured response time vs. estimated response time under S3. As it can be seen in the upper side of Figures 7,8,9, the measured and predicted response time values are so close that it is really hard to distinguish between them. Thus an additional figure depicting the difference between the measured and predicted response time values or simply prediction error has been added to the lower part of Figures 7,8,9.

Under S4 we observed that the prediction performance degrades compared to S3 but still holds a very small value for MSE and MAE. This performance degradation is most likely caused by the nonlinearity of this queuing system. Figure 8 illustrates the transient behavior of the proposed NARX multi-step ahead response time predictor under S4.

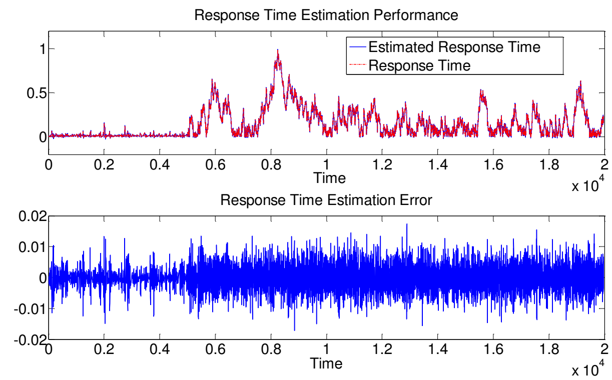


Fig. 7: NARX m step-ahead response time prediction of the M/M/1 queuing system with mean arrival rate changing with a step from 0.3 to 0.95 at time 5000. The prediction horizon m is set to 4. (upper) Measured response times vs. estimated response times. (lower) Difference between measured and estimated response times.

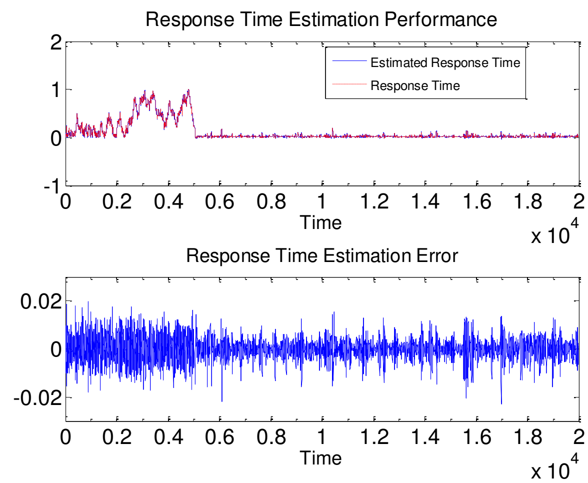


Fig. 8: NARX m step-ahead response time prediction of the M/M/1 queuing system with mean arrival rate changing with a step from 0.95 to 0.3 at time 5000. The prediction horizon m is set to 4. (upper) Measured response times vs. estimated response times. (lower) Difference between measured and estimated response times.

The two last simulations showed that the proposed response time predictor is able to handle transient regimes very well. In the last two simulation scenarios, the time variation in the mean arrival rate was specified as step functions representing sudden change of levels. In the extra simulation scenario performance (MAE and MSE) of the NARX m step-ahead response time predictor applied to a $M/M/1$ queuing system under a smooth slowly varying mean arrival rate is studied. The prediction horizon m is set to 4. The mean arrival rate is changed with a saturated ramp which is depicted in Figure 9. The minimum and maximum values of the mean arrival rate are respectively 0.3 and 0.95. By studying Figure 9 closely we can conclude that the proposed predictor can handle smooth time variant mean arrival rates.

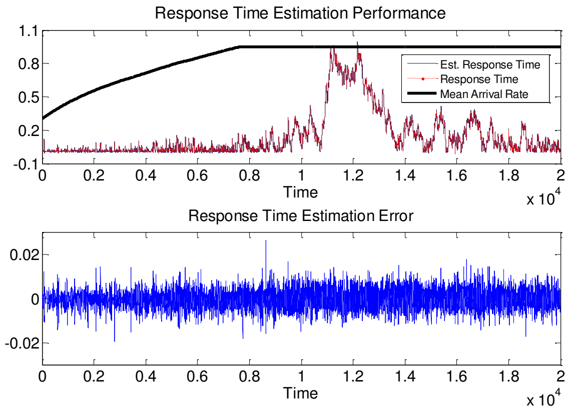


Fig. 9: NARX m step-ahead response time prediction of the M/M/1 queuing system with mean arrival rate as a saturated ramp function.(upper) Measured response times vs. estimated response times and the saturated ramp. (lower) Difference between measured and estimated response times.

Performance of the proposed predictor under some more single server queuing system configurations such as $M/D/1$ and $D/M/1$ for the same sets of scenarios S1-S4 has been investigated via simulations and very small MAE and MSE for the prediction error has been confirmed. Due to the lack of space, we skipped presenting those results.

V. Conclusion

A multi-step ahead NARX response time predictor for single server queuing systems, which represents a CPU constrained computing system, has been proposed and its performance under several test scenarios has been studied. The proposed predictor benefits from several promising characteristics which turns it into a viable candidate for being implemented in admission control products for computing systems. It is nonlinear, it supports multi-step ahead prediction, its structure is simple and its required measurements can be obtained without any requirement on changing communication protocols or operating systems. It has been shown that with being trained in only one high load scenario, it can predict the response times of a single server queuing system in multiple step ahead under high and low load steady state scenarios with a high accuracy. Very good performance of the proposed predictor under time variant and non-stationary scenarios has been confirmed by very small MAE and MSE of the response time prediction. It has also been shown that the proposed predictor is capable of accurate m step ahead response time prediction under time varying mean arrival rate scenarios. For the future work, the proposed predictor will be implemented in our web server lab and its capability of predicting the response times of the node elements will be studied.

Acknowledgment

Payam Amani, Maria Kihl and Anders Robertsson are members of Lund Center for Control of Complex Engineering Systems (LCCC), a Linnaeus Center at Lund University, funded by the Swedish Research Council. Maria Kihl is partly funded by the VINNMER program at the Swedish Governmental Agency for Innovation Systems (VINNOVA). Hereby, the first author appreciates the valuable comments of Dr. Kaan Bur and Mr. Jens Andersson from Department of Electrical and Information Technology of Lund University regarding this paper.

References

[1] M. Kihl, A. Robertsson, M. Andersson, and B. Wittenmark, "Control theoretic analysis of admission control mechanisms for web server systems," *The World Wide Web Journal*, vol. 11, no. 1, 2008.

[2] X. Chen, H. Chen, and P. Mohapatra, "Aces: an efficient admission control scheme for qos-aware web servers," *Computer Communication*, vol. 26, no. 14, 2003.

[3] X. Liu, J. Heo, L. Sha, and X. Zhu, "Adaptive control of multitiered web applications using queuing predictor," in *10th IEEE/IFIP Network Operation Mangement Symposium*, 2006.

[4] J. Cao, M. Andersson, C. Nyberg, and M. Kihl, "Web server performance modeling using an M/G/1/K*PS queue," in *In 10th International Conference on Telecommunications ICT03*, 2003, pp. 1501–1506.

[5] M. Kihl, A. Robertsson, and B. Wittenmark, "Analysis of admission control mechanisms using non-linear control theory," in *IEEE International Symposium on Computer Communications*, 2003.

[6] A. B. Clarke, "Maximum likelihood estimates in a simple queue," *The Annals of Mathematical Statistics*, vol. 28, no. 4, pp. 1036–1040, 1957.

[7] I. V. Basawa, U. N. Bhat, and R. Lund, "Maximum likelihood estimation for single server queues from waiting time data," *Queueing Syst.*, vol. 24, pp. 155–167, 1997.

[8] S. Zheng and A. F. Seila, "Some well-behaved estimators for the M/M/1 queue," *Operations Research Letters*, vol. 26, no. 5, pp. 231 – 235, 2000.

[9] M. F. McGrath, D. Gross, and N. D. Singpurwalla, "A subjective bayesian approach to the theory of queues I modeling," *Queueing Syst. Theory Appl.*, vol. 1, pp. 317–334, June 1987.

[10] M. F. McGrath and N. D. Singpurwalla, "A subjective bayesian approach to the theory of queues II inference and information in M/M/1 queues," *Queueing Syst. Theory Appl.*, vol. 1, pp. 335–353, June 1987.

[11] C. Armero, "Bayesian inference in markovian queues," *Queueing Syst.*, vol. 15, no. 1-4, pp. 419–426, 1994.

[12] A. Choudhury and A. Borthakur, "Bayesian inference and prediction in the single server markovian queue," *Metrika*, vol. 67, no. 3, pp. 371–383, April 2008.

[13] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Prentice Hall, 1998.

[14] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Prentice Hall PTR, 1998.

[15] H. T. Siegelmann, B. G. Horne, and C. L. Giles, "Computational capabilities of recurrent narx neural networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 27, no. 2, pp. 208–215, 1997.

[16] T. Lin, B. G. Horne, P. Tino, and C. L. Giles, "Learning long-term dependencies is not as difficult with NARX recurrent neural networks," *Advances in Neural Information Processing Systems*, vol. 8, pp. 577–602, 1996.

[17] H. Xie, H. Tang, and Y.-H. Liao, "Time series prediction based on NARX neural networks: An advanced approach," in *Machine Learning and Cybernetics, 2009 International Conference on*, vol. 3, 2009, pp. 1275 –1279.

[18] J. M. P. Menezes Jr. and G. A. Barreto, "A new look at nonlinear time series prediction with NARX recurrent neural network," in *Ninth Brazilian Symposium on Neural Networks*, Oct. 2006, pp. 160 –165.

[19] "Multi-step-ahead prediction using dynamic recurrent neural networks," *Neural Networks*, vol. 13, no. 7, pp. 765 – 786, 2000.

[20] L. Kleinrock, *Theory, Volume 1, Queueing Systems*. Wiley-Interscience, 1975.