Delay-Sensitive Services QoS Control in Sensor-Based Mass Applications

S. Marinovic, N. Rozic
Faculty of Electrical Engineering, Mechanical Engineering
and Naval Architecture FESB
University of Split
Split, Croatia
stipe.marinovic@fesb.hr

I. Cubic
Ericsson Nikola Tesla dd
Split, Croatia

Abstract: Wireless Sensor Networks shift from academic realm to industry. The shifting has been started with private, specialized and business-to-business application, but business-to-customers applications and services come in focus now. In business-to-customers (end users, subscribers) applications, Quality of Services is critical issues as well as scalability and reliability. This paper uses robust three-state Markov model of wireless sensor network to obtain probability density function of the service delay for use in modeling of end user sensor based delay-sensitive services from QoS and scalability point of view.

I. Introduction

The computing science and industry have been dealing with reality for decades. Back in time, the once emerging computing science and industry were promised virtual reality. After numbers of takes off, especially successful in movies, virtual reality hype has been reduced to feasible, and less resources demanding, augmented reality. The reduction was initiated by industry impatient to capitalize on new technologies. After high flying and big drop in expectations, it seems that we are, again, going up. The new rise is leveraging by tiny computers capable to constantly monitor physical world and even live creatures. The new reality flavor is not virtual, as not as augmented, it seems more like real reality. In addition to feel real world around, that tiny computers could process gathered measurements and create wireless networks by themselves in order to communicate monitored data. We just summarize concept of wireless sensor network (WSN). The WSN concept, for the truth sake, is not novel at all, but it is definitely mature and ready to go out from scientific laboratories and academy environment.

Constantly big advances in miniaturization of electronic and sensor components, and far slower, but also constant improvement in electric energy sources and batteries take us to the level where sensor network nodes finally looks like in original idea i.e. very small, or better tiny, and smart containers (embedding sensing, processing and radio capabilities) that dramatically extended direct human sensing of nature. Improvements in radio communications, networking, and software also contribute to take WSN concept to the level at which industry recognizes its potential.

In order to offer sensor measurements or sensor-based applications to horizontal, mass market (and end users) industry looks at seamless integration of WSNs with existed networks, primarily with the Internet and mobile networks, as well as quality of service of those new services.

Various transport capacity of data-gathering in wireless sensor networks are analyzed from the organization point of view are discussed in [1]. The quality of service concept [2] as its known in traditional network, defined mainly with network parameters such as end-to-end delay, jitter, and throughput is not sufficient for the new class of services [5]. Other parameters such as reliability and response delay are important for mass-market acceptance and success of sensor-based applications and services. In [6] authors analyze methods for minimizing the maximum delay in WSN by intelligent sink placement. Ref. [7] presents more detailed analysis on Markov model for WSN but without QoS analysis.

In this work we investigate QoS control mechanism for delay-sensitive services in WSN. We define minimal delay that should not be overlapped (delay threshold) based on the delay pdf determined by Markov model of the WSN. Since the service delay highly depends on the traffic load, it is modeled by appropriate forecasting model based on which the application server can in time store and service the requests during the high load conditions.

The rest of this paper is organized as follows: Section II is dedicated to description and analysis of sensor network's Markov model. Section III presents the proposed system for QoS control for delay-sensitive services based on delay distribution and load prediction. In Section IV ARIMA models are shortly discussed suitable for application in forecasting systems are considered. Section V presents numerical results and performance of the proposed QoS control scheme. Section VI concludes the paper with a few remarks on the presented results.

II SENSOR NETWORK MODEL

Wireless sensor networks are composed of a large number of sensing devices, which are equipped with limited computing and radio communication capabilities. [7]

Sensors are used in various fields such as medicine, industries or even at home for measuring or surveillance of various parameters like heart bit rate, body temperature, outdoor temperature, poison detection, smoke detection etc.

In wireless sensor networks sensors are positioned far apart measuring, calculating and sending their data to the main node called Sink. Sink collects data, stores them in data base if necessary, and serves as user interface receiving requests from user and replying to them. Depending on radio communication capabilities and distance to the sink, sensors can communicate with sink either directly or thru nearest neighbor sensors using multihop communications.

Since sensors are running on batteries, to save energy and prolong battery life sensors have two major operational states: active state and sleep state. In active state sensor can generate and process data, or transmit or receive data on the way to the sink.

In sleep state sensor is not active thus it cannot measure, transmit or receive data which results in energy savings. According to [7] sensor once in active state sets time instant in future when it will switch to sleep state. Since sensor can only switch to sleep state if sensor's buffer is empty, if buffer is not empty and time instant in future is reached, the sensor prolongs its active state with restrictions. In prolonged active state sensor can only wait for the neighbor sensors to become available for reception of data and send data to neighbor sensor. Once in sleep state sensor sets time instant in future when switching to active state will occur. Both active state time and sleep state time are geometrically distributed with parameters *p* and *q* respectively.

In this paper Markov chain for sensor with infinite buffer described in [7] and shown in Fig. 1. and DTMM describing the behavior of sensor next hops in Fig. 2. were combined and reduced to simple three state Discrete-Time Markov Model (DTMM) on Fig. 3. shows the reduced model with following states:

T - transmit state in which sensor can generate data, whether sensor is measuring data or receiving it, or send data to Sink or neighbor sensor

W - wait state in which sensor waits for neighbor sensors to become available for reception of data

S - sleep state in which sensor saves energy

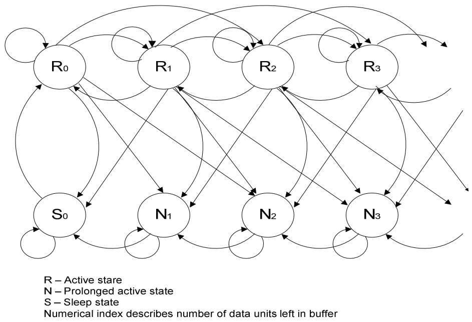
R₀
R₁
R₂
R₃
S₀
N₁
N₂
N₃
R – Active stare
N – Prolonged active state
S – Sleep state
Numerical index describes number of data units left in buffer

Fig. 1 Markov chain describing sensor behaviour [3]

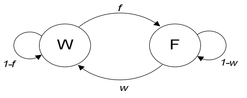

Fig. 2 DTMM model describing the behaviour of the sensor next-hops

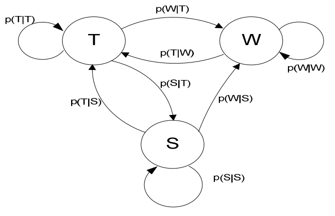

Fig. 3. Reduced discrete-time Markov model

The process of reduction of number of states was following: Transition between all active states with forward capability were summed together and then added with 1 minus sum of all transitions from all states with forward capabilities forming joint self transition probability of transmit state. In the same way self transition probability for wait and sleep states were calculated. Transition probabilities for different states were calculated by summing all probabilities from all forward capable states to all waiting states resulting in transition probabilities from transmit to wait state, and similar for all other transition probabilities. Equation for calculating for transmit state Transition probabilities of DTMM shown in Fig. 3 based on the model shown in Fig. 1 by using probabilities from [7] are:

$$\begin{aligned} p(T \mid T) = {} & p(R^F_{i+1} \mid R^F_i) + p(N^F_i \mid R^F_i) + p(N^F_{i+1} \mid R^F_i) + \\ & + p(R^F_{i+2} \mid R^F_i) + p(N^F_{i+2} \mid R^F_i) + p(R^F_{i-1} \mid R^F_i) + p(N^F_{i-1} \mid R^F_i) + \\ & + p(N^F_{i-1} \mid N^F_i) + 1 - (p(R^W_i \mid R^F_i) + p(N^W_i \mid R^F_i) + p(R^W_{i+1} \mid R^F_i) \\ & + p(N^W_{i+1} \mid R^F_i) + p(R^W_{i+2} \mid R^F_i) + p(N^W_{i+2} \mid R^F_i) + p(R^W_{i-1} \mid R^F_i) + \\ & + p(S^W_0 \mid R^F_i) + p(S^F_0 \mid R^F_i) + p(N^W_{i-1} \mid R^F_i) + p(N^W_i \mid N^F_i) + \\ & + p(N^W_{i-1} \mid N^F_i) + p(S^F_0 \mid N^F_i) + p(S^W_0 \mid N^F_i) + p(N^F_i \mid R^F_i) + \\ & + p(R^F_{i+1} \mid R^F_i) + p(N^F_{i+1} \mid R^F_i) + p(R^F_{i+2} \mid R^F_i) + p(N^F_{i+2} \mid R^F_i) + \\ & + p(R^F_{i-1} \mid R^F_i) + p(N^F_{i-1} \mid R^F_i)). \end{aligned} \tag{1}$$

$$\begin{aligned} p(W \mid T) = {} & p(R^W_i \mid R^F_i) + p(N^W_i \mid R^F_i) + p(R^W_{i+1} \mid R^F_i) + \\ & + p(N^W_{i+1} \mid R^F_i) + p(R^W_{i+1} \mid R^F_i) + p(N^W_{i+2} \mid R^F_i) + p(R^W_{i-1} \mid R^F_i) + \\ & + p(N^W_{i-1} \mid R^F_i) + p(N^W_i \mid N^F_i) + p(N^W_{i-1} \mid N^F_i). \end{aligned} \tag{2}$$

$$p(S \mid T) = p(S^W_o \mid R^F_i) + p(S^F_o \mid R^F_i) + p(S^W_0 \mid N^F_i) + p(S^F_0 \mid R^F_i) \tag{3}$$

Resulting transition probabilities for simplified Markov model are shown in Table 1.

Parameters in Table 1. are defined as follows : [5]

p – parameter of geometric distribution describing sleep state time

q – parameter of geometric distribution describing active state time

g – Bernoulli process parameter describing sensor data generation *α* – probability of data being received from neighboring node in time unit

β - probability if data being send to neighboring node or Sink in time unit

f, w – transition probabilities form Forward to Wait state and vice versa.

So	Sd	P(Sd|So)
T	T	*1-2w+ βpg- βp-β*
T	S	*β (1+p-pg)*
T	W	*2w*
W	W	*1-2f*
W	T	*2f*
W	S	*0*
S	T	*q(f-w+1)*
S	W	*q(1-f+w)*
S	S	*1-2q*

Table 1. Transition probabilities for the reduced model in Fig. 3

Considering sensor model as M/M/1 single server queue we can write the probability that time T_Q an item spends in system is lower or equal to t [11]. This represents cumulative probability distribution

$$\Pr\left[T_Q \le t\right] = 1 - e^{-(1-\rho)t/T_S} \qquad (4)$$

where ρ is utilization of the system, and Ts is service time.

By differencing (4) we get pdf for M/M/1 system with queue

$$p(t) = (1-\rho)\frac{1}{T_S}e^{-(1-\rho)t/T_S} \qquad (5)$$

Since WSN is modeled with three states Markov model as shown on Fig. 3. with inactive (sleep and wait states) system will be available according to state stationary probability which results in increased delay. Considering that WSN is not always available for processing incoming requests according to sum of stationary probability of wait and sleep states, successively from one slot to another WSN will be available with probability:

$$P_A(t) = 1 - (\Pi_S + \Pi_W)^t \qquad (6)$$

where Π_S is stationary probability for sleep state, and Π_W is stationary probability for wait state. Probability density function for data delivery delay for farthest node from the sink is expressed by multiplying (5) and (6) giving:

$$p(t) = \left((1-\rho)\frac{1}{T_S}e^{-(1-\rho)t/T_S}\right)\cdot\left(1-(\Pi_S+\Pi_W)^t\right) \qquad (7)$$

by replacing ρ with λT_S we get delay as function of traffic load:

$$p(t) = \left((1-\lambda T_s)\frac{1}{T_s}e^{-(1-\lambda T_s)t/T_s}\right)\cdot\left(1-(\Pi_s+\Pi_w)^t\right) \qquad (8)$$

Fig 4. shows pdf for data delivery delay report with offset of 6 data units which is minimal delay since data from the farthest node has to travel in average through M=6 nodes till it reaches the sink.

T1, T2 and T3 represents time delays that includes 95% area for traffic loads λ_1,λ_2, and λ_3.

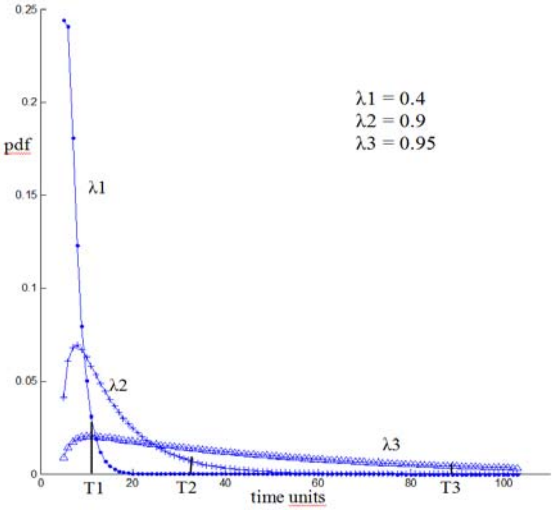

Fig. 4. Data delivery delay pdf for farthest node

III QoS Analysis for Delay-Sensitive Services

A. The proposed system

It is well known that the important strategic issue of the contemporary ICT systems is to provide customer with services available anytime, anywhere and at guaranteed quality. Specifically, in delay-sensitive applications a quality of service (QoS) is dominantly determined by the time instant spent between the customer inquire (by pushing button, launching call, …) of a service and time instant at which the response take place. Sensor wireless network technologies connected to Internet, PSTN, GSM, and other public networks (Fig. 5) are actual and feasible solutions that should satisfy above mentioned strategic QoS issues, but operational aspects are still big job for researchers. One of the most challenged issue in sensor network applications is to guarantee the required QoS for mobile calls that require fast and possibly real-time response enabling users to obtain fresh information at the acceptable low delay.

Subscribers calls impose service requirements (queries) to the wireless sensor networks with an intensity that vary in time, in capacity and in complexity as well. For example, a simple query for data on temperature in the sensor position may vary in frequency during a day, in the required precision causing more bits per response and more processing time as well. Let service priority, capacity and complexity define QoS parameters related with service of a class c.

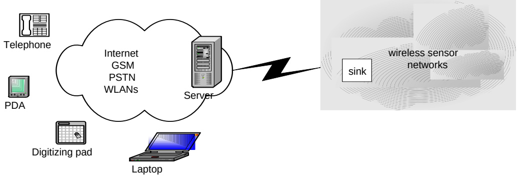

Fig. 5 Architecture of the considered system

As described in Section II, network of sensors operates according to the three-state sensor Markov model which incorporates network functions including communicating, storing and routing. Time consumption of these functions causes delay in response whose PDF depends on transient probabilities of the Markov model, but also on traffic intensity (Fig. 4).

B. The QoS control scheme

Assuming that application server in Fig. 5 is equipped with an infinite storage and very fast server system compared with the sensor network system, the application server can be supported by the traffic load predictive application which should generate traffic load forecasts for enough long horizon. When the forecasting system alert the control system due to ongoing traffic load that can not be satisfied at the guaranteed QoS (D_T) the application server sends request to the sensor network for the data expected to be requested by customers. Application server must send the request in time i.e. enough earlier so that for the predicted traffic load, sensor network can respond and deliver data at a high reliability before a high load starts. This time instant can be determined from the delay PDF as it is marked in Fig. 4 with T_1, T_2 and T_3. This preceding time can be determined from the delay PDF (7) based on the condition

$$\int_{T_i}^{\infty} p(t)\, dt = \text{TRDP} \qquad (9)$$

Fig. 4. shows times T1, T2 and T3 that define request dropping probability of 5% obtained from (9) for $\lambda_1 = 0.4$, $\lambda_2 = 0.9$ and $\lambda_3 = 0.95$.

Let $\hat{\lambda}_c(t + k\Delta t)$ be the class c traffic load forecasted for the horizon $k\Delta t; k = 1, 2, \cdots, K$. The variance of the prediction error $e_c(t + k\Delta t) = \lambda_c(t + k\Delta t) - \hat{\lambda}_c(t + k\Delta t)$ is $\sigma_{e,c}^2(t + k\Delta t)$ and its deviation $\sigma_{e,c}(t + k\Delta t)$ can be used to define the prediction confidence interval. With Normal assumption, 99% confidence limits are given approximately as

$$\hat{\lambda}_c(t + k\Delta t) \pm 3\sigma_{e,c}(t + k\Delta t) \qquad (10)$$

The minimum load to be served in order to satisfy the predicted demand is defined by the upper prediction probability limit, thus

$$\overline{\lambda}_c(t + k\Delta t) = \hat{\lambda}_c(t + k\Delta t) + 3\sigma_{e,c}(t + k\Delta t) \qquad (11)$$

Value $\overline{\lambda}_c$ can be considered as the maximum traffic load that will not be really exceeded. The request dropping probability (RDP) for the class c service is then defined with the delay threshold $D_{T,c}$ which follows from (8) by replacing λ by $\overline{\lambda}_c$. The forecasting system should alert the control system if the forecasted traffic $\overline{\lambda}_c$ exceeds the traffic load λ_A which represents the maximum load at which the delay threshold $D_{T,c}$ does not exceed the targeted request dropping probability (TRDP_C) for the class c service. The QoS control system, when alerted, send a request to the sink and WSN for the data at the instant which precedes the predicted overload instant for $D_{T,c}$ time units. The system continuously monitors actual and forecasted loads λ_c and $\hat{\lambda}_c$ respectively and if $\lambda_C > \lambda_A$ sends requests for data to WSN but at the reduced rate $\lambda_1 = \lambda_A$ by filtering the incoming requests while the rest $\lambda_2 = \lambda - \lambda_A$ are immediately answered based on stored data. In this way, the QoS control system keeps the service delay below the TRDP_C and at the same time minimizes the age of information.
Based on the above analysis we can launce the following algorithm

```
$A_s = 0;$   % normal state
if   $\hat{\lambda}(t + k\Delta t) \geq \lambda_A \rightarrow A_s = 1;$   % allert state
    $t_0 = t,$
    $t = t + \Delta t,$
if $t \geq t_0 + k\Delta t - D_T \rightarrow$ send requests for data to WSN at $\lambda_A$
if $t \geq t_0 + nT_A$   &   $\lambda(t) > \lambda_A \rightarrow$ requests for data to WSN
if $\lambda(t) < \lambda_A$ or $\overline{\lambda}(t + \Delta t) < \lambda_A \rightarrow A_s = 0;$   % normal state
```

The QoS control algorithm

IV ARIMA MODELS FOR LOAD FORECASTING

Many empirical time series such as the number of active channels have no fixed mean value although they exhibit homogeneity in the sense that one part of the series behaves in the same way as the other parts. Autoregressive Integrated Moving Average (ARIMA) models, which describe such homogeneous nonstationary behavior, can be obtained by performing a suitable differentiation of the process to obtain stationary process. In general, the time series also exhibits periodic behavior, and therefore require both nonperiodic and periodic differencing.

The multiplicative ARIMA models incorporate nonperiodic and periodic behavior.

Nonstationary multiplicative models are defined by [6]

$$a(L)A(L^S)\nabla^d \nabla_S^D \lambda_t = b(L)B(L^S)\varepsilon_t \qquad (12)$$

where λ is the considered random variable, $\varepsilon_t \approx N(0, \sigma_\varepsilon^2)$ is noise variable, *d* and *D* are degrees of nonseasonal and seasonal differencing, respectively, *S* is the period, *L* is a backward shift operator and

$$\nabla^d = (1 - L)^d ; \quad \nabla_S^D = (1 - L^S)^D \qquad (13)$$

Coefficients of polynomials *a(L)* and *A(L^S)* are autoregressive aperiodic and periodic parameters, respectively, while *b(L)* and *B(L^S)* are polynomials whose coefficients are moving average aperiodic and periodic parameters, respectively. These polynomials are given by

$$a(L)=\sum_{i=1}^{p} a_i L^i \,;\quad A(L^S)=\sum_{i=1}^{P} a_i L^{iS}\,;$$

$$b(L)=\sum_{i=1}^{q} b_i L^i \,;\quad B(L^S)=\sum_{i=1}^{Q} B_i L^{iS} \text{ with } L^i\cdot\lambda_t=\lambda_{t-i}$$

The resulting model can be denoted as *ARIMA (p,d,q)x(P,D,Q)$_S$*, where *p*, *P* are the numbers of autoregressive parameters and *q*, *Q* are the numbers of moving average parameters, respectively.

Periodic behavior can be expected in campus or in company buildings where people call for some services almost periodically. Typically, teachers and students call for services during hourly breaks etc. Employers in company used to call in the afternoon before leaving the office and so on.

It is also useful to expand the ARIMA model with an intervention model in order to fit the knowledge about future exceptional events that could influence the demand [10][12]. These exceptional events such as emergency situations, sports events, conferences, strikes etc. generate burst-like traffic and could be critical to keep targeted QoS i.e TRDP. The effect of an intervention variable ξ_n on the variable being modeled can also be defined as an ARMA model

$$\delta(L)\lambda_t=\omega(L)\xi_{t-b} \tag{14}$$

where *δ(L)* and *ω(L)* are polynomials whose coefficients are autoregressive and moving average parameters for intervention system respectively.

Based on Eq. (10) we can write the one-step ahead conditional expectation of the considered variable as follows:

$$\langle\lambda_t\rangle = a(L)^{-1}A(L^S)^{-1}b(L)B(L^S)\varepsilon_t \tag{15}$$

where

$$\lambda_t=\nabla^d\nabla_S^D\Lambda_t\,,\ \varepsilon_t{=}0,$$

and variance

$$\sigma^2_{\langle\lambda\rangle}=\sigma^2_{\lambda}\left[\sum_{i=0}^{p}\psi_i^2\right]^{-1} \tag{16}$$

Values of ψ weights follow from $c(L)\psi(L)=1$ with $c(L)=a(L)A(L^S)b(L)^{-1}B(L^S)^{-1}$ and $\psi(L)=\sum_{i=1}^{p}\psi_i L^i$.

Note that ε_t=0 in Eq. (11) does not mean that the right side equals zero since $\varepsilon_{t\text{-}i}$ for i=1,2, …are not equal to zeros.

Despite the model dimensionality, measurement based methods use hidden Markov model, Kalman filter or state-space model, Wiener filter or Bayesian model [6]. Particularly, in real applications time series modules such as exponential smoothing or ARIMA models are useful [12]. Autoregresive and ARIMA models have been proposed to characterize the video traffic.

V Numerical Results

Based on trade of between average data unit delivery delay and average network energy consumption studied in [7] - sleep/active transition rates were chosen to be equal to 0.1. The α and β were also chosen as average values for unconditioned transmission rates computed by model developed in [7] valued as α=0.13 and β=0.5. The *g* was chosen to be 0.005 which represents heavy network load condition. In this work we have considered simulated Poisson-distributed non-stationary request traffic with seasonal behavior. In identification procedure based on analysis of autocorrelation, partial correlation and power spectrum [6] the sequence has been modeled with ARIMA model (8) ARIMA(1,1,0)x(1,1,0)S.

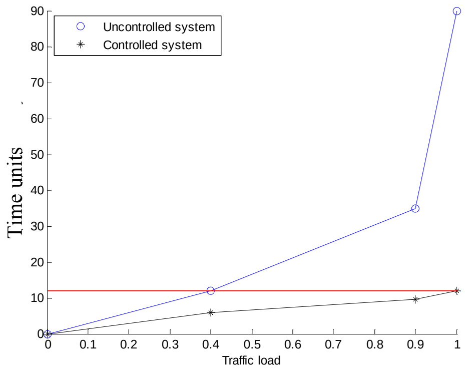

Fig. 7 Request dropping probability versus traffic load

Parameters a_1 and A_1 are estimated from the sample autocorrelation lags via Yule-Walkers equation [9]. This model produces predictions of minimum error variance and error variable presents an independent (white noise) process. The prediction interval was *Δt=60* s and the targeted RDP was set to 0.05 by taking into account 99% upper confidence limit. The forecasted load was $\bar{\lambda}_c=0.63$ and from Fig. 7. followed that the application server should send request for data at *D$_T$=20* time slots in advance.

VI Conclusions

In this paper we have presented a new approach to QoS control for delay-sensitive services based on sensor network Markov model and ARIMA models for traffic load forecasting. The main advantages of the presented approach is in efficiency of QoS control at a minimum cost of information oldness and simplicity of the system implementation at the application server. Simplicity is advantage, because it provides usage of the model in dynamic caching of sensor measurements in network layer as response to service request peaks.

References

[1] Enrique J. Duarte-Melo and Mingyan Liu, "Data-gathering in Wireless Sensor Networks: Organization and Capacity", Computer Networks Volume 43, Issue 4, 15 November 2003, Pages 519-537

[2] Dazhi Chen and Pramod K. Varshney: "QoS Support in Wireless Sensor Networks: A Survery", Proc. of Int. Con. on. *Wireless Networks* (ICWN), Las Vegas, 2004

[3] W. Y. Poe and J. B. Schmitt,"Minimizing the Maximum Delay in Wireless Sensor Networks by Intelligent Sink Placement", Technical Report, Univ. of Kaiserslautern, 2007.

[4] Milan Tafra, Nikola Rožić and Ivica Ćubić, "Database Like Approach to Acquiring Data on Request in Wireless Sensor Networks", WICT SoftCOM 2007.

[5] D. Chen and P.K. Varshney, “QoS Support in Wireless Sensor Networks: A Survey, 2005, pp. 7

[6] Box GEP, Jenkins GM: “Time Series Analysis Forecasting and Control”. Holden-Day 1976.

[7] C.F. Chiasserini and M. Garetto „An Analitical Model for Wireless Sensor Networks with Sleeping Nodes“, IEEE Transactions on Mobile Computing, Vol. 5, No. 12, December 2006.

[8] Peter J. B. King “Computer and Communication Systems Performance Modeling”, Prentice Hall International, 1990.

[9] P.J. Brockwell and R.A Davis, *Time Series: Theory and Methods*, 2nd Ed. New York, Springer Verlag,1991.

[10] G.E.P. Box, G.M. Jenkins, “Time Series Analysis Forecasting and Control” , Holden-Day, 1976.

[11] Stallings, William: “TCP/IP and ATM design principles”, Prentice Hall Inc. 1998

[12] N.Rožić, D.Begušić, G.Kandus: “Application of ARIMA Models for Handoff Control in Multimedia IP Networks”, Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS'03), pp. 787-791, Awaji Island, Japan, December 7-10, 2003.