Dynamic Control and Optimization of Buffer Size for Short Message Transfer in GPRS/UMTS Networks*

Michael M. Markou and Christos G. Panayiotou
Dept. of Electrical and Computer Engineering,
University of Cyprus
Email: {markou,christosp}@ucy.ac.cy

Abstract

In this paper we use the Infinitesimal Perturbation Analysis (IPA) algorithm derived in [5] to obtain derivative estimates of a performance function of interest with respect to a buffer threshold. These estimates are used with stochastic approximation algorithms to determine near optimal buffer thresholds for Short Message Service in wireless networks. Several simulation results are presented that indicate the benefit of the approach.

1. Introduction

The rapid growth of wireless communication networks has put tremendous pressure on the network's resources. As a result, efficient algorithms are required for effective resource management which constitutes a challenging task. This paper investigates one such algorithm [5] as it applies in wireless networks. In Global System for Mobile Communications (GSM) networks, controlling the most successful wireless data service traffic –Short Message Service (SMS) traffic- has become a crucial problem, especially upon the arrival and wide popularity of Multimedia Messaging Services (MMS) [1]. MMS is a new message standard that can be used over General Packet Radio Service (GPRS) / Universal Mobile Telecommunications System (UMTS) SMS network that achieves better utilization of high bandwidth networks, as GPRS and UMTS [2], [4]. As referred in [2], "this service enables the transmission of messages with full content versatility, including text, images, audio and video, between mobile devices and from applications to these devices".

Having in mind the potential provided in these networks and the applications developed based on them, emerges a need to control the traffic generated using the capabilities of the existent infrastructure. This work focuses on the problem of *dynamically* determining the optimal buffer threshold for services in wireless networks, such as message transferring services (SMS, MMS) in GPRS/UMTS networks. Towards this end, we use the approach in [5]. In this approach, we use Infinitesimal Perturbation Analysis (IPA) [3], [6], based on Stochastic Fluid Models (SFM), to obtain sensitivity (derivative) estimates of the performance measure of interest with respect to the control parameter (in this case the buffer threshold). Subsequently, we use the estimates with gradient based stochastic approximation techniques [7] to obtain the optimal buffer threshold. This approach has been applied to a number of models with encouraging results (see [8], [9]).

The major advantages of this approach are: (*i*) the optimization is performed *on line*; as a result, the approach can be used to *continuously* adjust the buffer size as traffic conditions change (no stationarity required). (*ii*) There is no need to know the system's underlying stochastic processes, and (*iii*) implementation of the estimators (in hardware and software) is very simple.

It is worth pointing out that the approach adopted in [5], [8], [9] and in this paper emphasizes on the *control* and *optimization* of a communication system rather than *performance evaluation*. These references derive sensitivity estimates based on a SFM which might not always deliver accurate performance estimates for a "real" discrete-event system. On the other hand, it is not unreasonable to expect that one can identify the solution of an optimization problem based on a model which captures *only* those features of the underlying real system that are needed to lead to the right solution, without the need to estimate the corresponding optimal performance with accuracy. Even if the exact solution cannot be obtained by such lower-resolution models, one can still obtain near-optimal points that exhibit robustness with respect to certain aspects of the model they are based on. Such observations have been made in several contexts (see [5] and references therein).

Motivated by the robustness properties of this approach, we applied the IPA algorithm derived in [5] to the model we consider in this paper. Even though the system dynamics of the two models are quite different (see Section 3 for details), our simulation results show that for small and moderate user mobility, the approach performs satisfactorily; for the performance measures considered in this paper, it identifies optimal or near optimal solutions.

The remaining of this paper is organized as follows. In Section 2 we describe the queuing model that describes a single serving GPRS support node (SGSN). In Section 3 we present the control algorithm and in Section 4 show extensive simulation results. Finally we conclude with Section 5 where we also present plans for future research.

* This work has been partially supported by the Cyprus Research Promotion Foundation under the program for Young Researchers, with contract number ΕΝΤΑΞΗ/0603/11.

2. System Model

This paper is based on the system model presented in [2] which is shown in Figure 1. In this context, $\alpha_n(t)$ is the rate of new mobile-terminated message arrivals. $\alpha_f(t)$ is the rate of forwarded messages that arrive from all neighboring cells while mobility rate $\eta(t;\theta)$ is the rate that queued messages are forwarded to neighboring cells due to user mobility. Furthermore, expiration rate $\sigma(t;\theta)$ is the rate of timeout of queued messages and $\beta(t)$ is the message transmission rate. Based on the above processes, the following processes are derived from the system's dynamic behavior: $x(t;\theta)$ is the buffer content, $\delta(t;\theta)$ is the system's outflow rate and, $\gamma(t;\theta)$ is the rate that packets are dropped when $x(t;\theta) \geq \theta$, where θ is the control parameter, which in this paper is the buffer threshold.

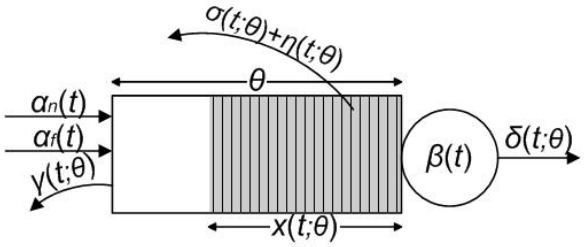

Figure 1. The queueing model of an SGSN.

All above processes are assumed time varying and unlike the work in [2], no Poisson or exponential assumptions are made. Furthermore, as in [2], we consider a heterogeneous GPRS/UMTS SMS network, in which the users move along the service area of SGSN i to the service area of SGSN j according to a routing probability r_{ij}, $i,j = 1,2,\ldots,M$, where M is the number of SGSN in the system.

3. Dynamic Control of the Buffer Size

We assume a packet level admission control policy, where an arriving packet at time t is accepted if the queue length $x(t;\theta) < \theta$, where θ is a controllable threshold, otherwise it is rejected. In this paper, we consider an approach for controlling θ such that a performance measure of interest is minimized. Specifically, we are interested in minimizing

$$\min_{\theta} \bar{J}_T, \ \bar{J}_T = \bar{Q}_T(\theta) + RC \cdot \bar{L}_T(\theta) \quad (1),$$

where $\bar{Q}_T(\theta) = E[Q_T(\theta)]$, $\bar{L}_T(\theta) = E[L_T(\theta)]$ and RC is the rejection cost of each packet. The sample functions Q_T and L_T reflect the cost due to packet workload (or equivalently the cost due to packet delays through Little's law) and packet losses (due to buffer overflow or packet expiration) respectively and are defined below.

$$L_T(\theta) = \sum_{k=1}^{K} \int_{B_k} [\gamma(t;\theta) + \sigma(t;\theta)] dt \quad (2)$$

$$Q_T(\theta) = \sum_{k=1}^{K} \int_{B_k} x(t;\theta) dt \quad (3)$$

where T is the length of the observation interval and B_k denotes the k^{th} busy period (in increasing order), $k = 1,2,\ldots,K$ where K is the generally random number of busy periods observed in the interval $[0, T]$. The overall cost $\bar{J}_T(\theta)$ represents a tradeoff between providing low delay and low packet loss probability. (For small θ the buffer overflow probability is high, however, the average workload or packet delay is small. As the buffer size increases, packet loss probability drops but workload increases. Service providers, depending on their priorities, can define the rejection cost appropriately and as a result, there is a buffer size that minimizes the overall cost $\bar{J}_T(\theta)$). Alternatively, one can formulate the problem as a *constrained* optimization problem, e.g., minimize the workload such that the packet overflow probability is below some threshold.

As already mentioned, our aim is to continuously monitor $\bar{J}_T(\theta)$ and resize the threshold θ in order to minimize the above cost. This objective is achieved by estimating the sensitivities $dL_T(\theta)/d\theta$ and $dQ_T(\theta)/d\theta$ through the Infinitesimal Perturbation Analysis (IPA) algorithm that has been developed in [5] and is presented below. We point out that this algorithm was derived based on the SFM equivalent of the system presented in Figure 1 when $\eta(t;\theta) = \sigma(t;\theta) = 0$.

IPA Algorithm:

- Initialize a counter L=0 and a cumulative timer W=0.
- Initialize τ=0.
- If an overflow event is observed at time t and τ=0:
  - Set τ=t.
- If a busy period ends at time t and τ>0:
  - Set L=L+1 and W=W+(t-τ)
  - Reset τ=0
  - If t=T, and τ>0
    - Set L=L+1 and W=W+(t-τ)

The final values of L and W, divided by the length of the observation interval T, provide the IPA derivative estimates $dL_T(\theta)/d\theta$ and $dQ_T(\theta)/d\theta$ respectively, by which we calculate the estimator of $dJ_T(\theta)/d\theta$, in other words

$$\frac{dJ_T(\theta)}{d\theta} = \frac{W}{T} + RC \cdot \frac{L}{T}.$$

Subsequently these estimates are used to drive the stochastic approximation iterations.

$$\theta_{n+1} = \theta_n - s_n \cdot dJ_T(\theta)/d\theta \quad (4)$$

where s_n is an appropriate step size sequence. The implementation of (5) is shown below.

Stochastic Approximation Algorithm:

- Set an initial, arbitrary value θ_0 for the buffer size.
- For every step n:
  - Calculate the estimator of $dJ_T/d\theta$,
  - If $dJ_T/d\theta = 0$
    - Set $\theta_{n+1} = \theta_n - 1$
  - Else Set $\theta_{n+1} = \theta_n - s \cdot dJ_T/d\theta$

where for simplicity we assume a constant step size s. Furthermore, notice that if θ_0 is set to a very high value, then it is possible that no packets are lost in the interval $[0,T]$ and thus the estimator of $dJ_T/d\theta$ will evaluate to 0 (see the IPA Algorithm). However, this does not necessarily

mean an optimal buffer size (due to excessive workload), therefore, when $dJ_T/d\theta$ evaluates to 0, we reduce θ by 1.

Next we present several simulation results that indicate that even though the IPA algorithm above was developed based on a model with slightly different dynamics, it still delivers a near optimal buffer threshold.

4. Simulation Results

An event-driven simulation is used to obtain numerical results. The first set of simulations considers the case where the system consists of a single SGSN, studying the effects of changing the Call-to-Mobility Ratio (CMR) and the Service-to-Expiration Ratio (SER); CMR is defined as the ratio between the message arrival rate and the portable mobility rate. SER is defined as the ratio between the message transmission rate and the message expiration rate. Simulation results show that for values of CMR greater of 5, the optimization algorithm converges to a near optimal threshold θ within a small number of iterations. The second part of the results presents the effects of changing CMR and SER in a two dimensional GPRS/UMTS network composed of seven SGSNs with equal coverage area, as presented in [2]. For simplicity, for most experiments we used Poisson arrivals and exponential service distributions (except where stated otherwise).

4.1. Single SGSN Simulation Results

The simulation model considers of two packet generators that send packets to queue with mean rates a_n=1 packets/millisecond (p/ms) and a_f=1/10 p/ms. The queue performs admission control to incoming packets and passes the packets to server. It worth to mention that this queue, in contrast with the queue in [5], is not strictly a FIFO queue since a packet that lies in the queue may expire or be moved out of it before is passed to the server for transmission. The service mean rate (transmission time) β=1 p/ms.

This set of results consists of five different simulation cases in which mobility and expiration rates vary as follows:

1. η=1/10 p/ms (CMR=10), σ=1/10 p/ms (SER=10).
2. η=1/20 p/ms (CMR=20), σ=1/20 p/ms (SER=20).
3. η=1/100 p/ms (CMR=100), σ=1/20 p/ms (SER=20).
4. η=1/10 p/ms (CMR=10), σ=1/100 p/ms (SER=100).
5. In this case we substituted the two sources with four ON-OFF sources where the arrivals during each ON period are deterministically distributed, while the duration of the ON and OFF periods are exponential distributed. For each of the four sources, the parameters (mean arrival rate, mean duration of ON period, mean duration of OFF period) are: ($\frac{1}{2}$ p/ms,10ms,10ms), ($\frac{1}{3}$ p/ms,5ms,5ms), ($\frac{1}{4}$ p/ms,5ms,5ms) and ($\frac{1}{10}$ p/ms,1000ms,1000ms).

Figure 2 and Figure 3 show the average waiting time and the packet loss probability functions of buffer size for the above cases 1-4. Figure 2 indicates that as CMR increases, the average waiting time for a packet is increasing. This is easily explained, since a low CMR value (i.e. higher mobility value) indicates that packets are moved out from the queue without being transmitted, thus not considered in the average. Figure 3 indicates that SER has a significant role in packet loss probability, since as SER decreases (i.e. higher expiration rate) the packet loss probability increases. As the mean mobility of the users remains the same, and since a packet expiration is considered as loss, packet loss probability in the case of CMR=10, SER=10 is greater than the case of CMR=10, SER=100. Mobility also affects packet loss probability, since in low CMR (i.e. high mobility) packets leave queue before they expire.

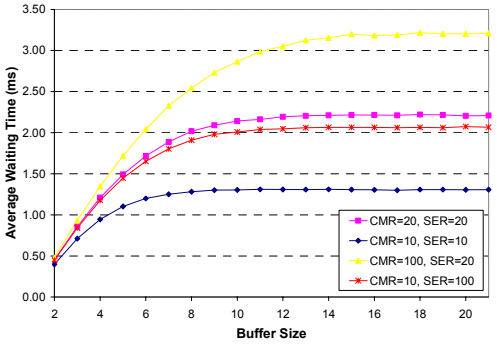

Figure 2. Average waiting time versus buffer size of the single SGSN system for the Cases 1-4.

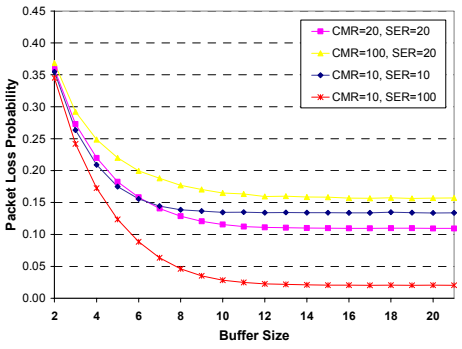

Figure 3. Packet loss probability versus buffer size of the single SGSN system for the four Cases 1-4.

The total cost $J_T(\theta)$, as defined in (1), is shown in Figure 4-Figure 8 for the five simulation cases 1-5 respectively. The "B.F. Simulation" curves correspond to the brute force simulation results. These simulations were executed for a long period of simulation time (30 minutes) in order to extract representative and noise free results for the total cost. Next, we executed the optimization algorithm for each case. We start from an initial buffer size θ=20. We observe the system for a short interval T = 1 second and record the noisy performance estimate of $J_T(\theta)$ and use the derivative estimate in the stochastic approximation algorithm to determine the next buffer threshold.

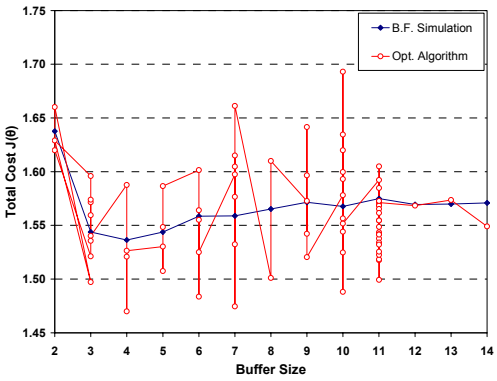

Figure 4. Total cost versus buffer size for Case 1.

From all Figures 5 – 8 we observe that even though the performance estimates are very noisy (due to the short observation interval), the optimization algorithm converges to near optimal buffer sizes within few iterations (few seconds).

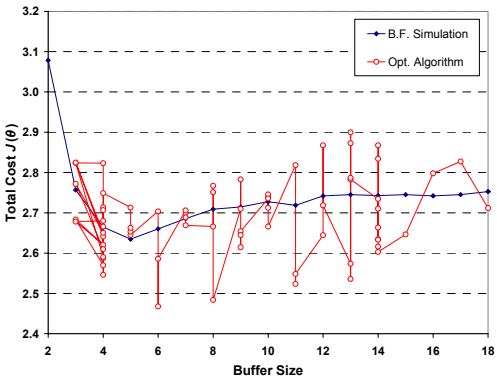

Figure 5. Total cost versus buffer size for Case 2.

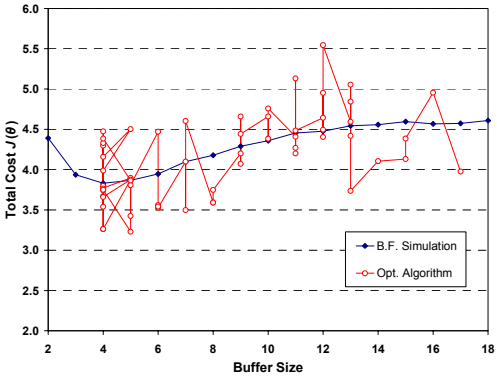

Figure 6. Total cost versus buffer size for case 3.

4.2. GPRS/UMTS Network Simulation Results

The second part of the results present the effects of changing CMR and SER in a two dimensional GPRS/UMTS network composed of seven SGSN's with equal coverage area, as shown in Figure 9.

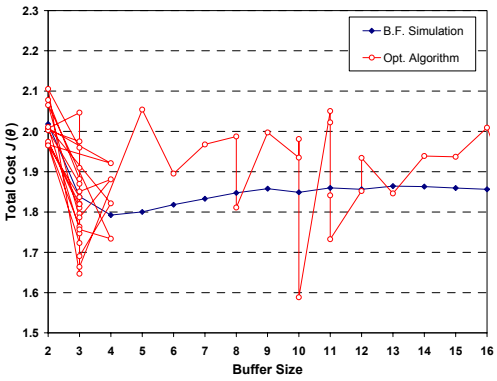

Figure 7. Total cost versus buffer size for Case 4.

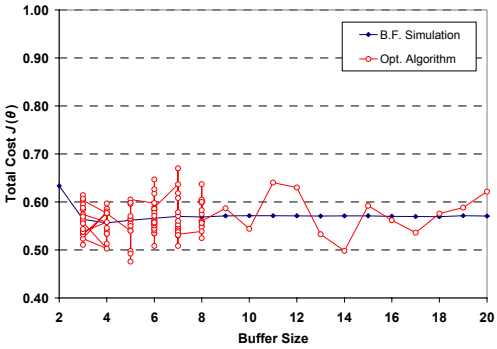

Figure 8. Total cost versus buffer size for Case 5.

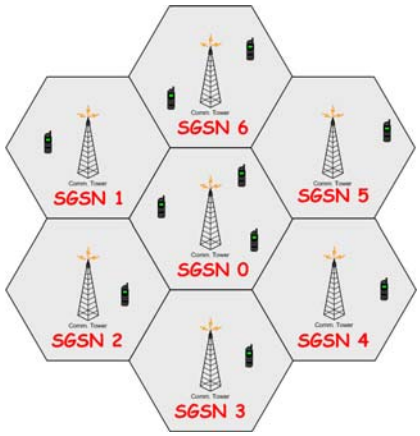

Figure 9. The GPRS/UMTS network model

In each area users move with velocity that is determined by the mobility rate η. Users may move along the service area of SGSN i to the service area of SGSN j according to a routing probability r_{ij}, $i,j = 0, 1,\ldots, 6$, as presented in [2]. The routing matrix $R = [r_{ij}]_{7x7}$,

$$R = \begin{bmatrix} 0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 7/12 & 0 & 1/3 & 0 & 0 & 0 & 1/12 \\ 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 \\ 7/15 & 0 & 1/3 & 0 & 1/5 & 0 & 0 \\ 5/12 & 0 & 0 & 1/3 & 0 & 1/4 & 0 \\ 7/12 & 0 & 0 & 0 & 1/3 & 0 & 1/12 \\ 1/3 & 1/3 & 0 & 0 & 0 & 1/3 & 0 \end{bmatrix}$$

For all simulation experiments we assumed that the mean arrival rate a_n=1 p/ms and service (transmission) rate β=1 p/ms and first consider the case where we set the mean mobility rate η=1 p/ms (CMR=1) and the mean expiration rate σ=1/10 p/ms (SER=10).

Note during these simulation experiments, a_f of an SGSN i is the rate of the total messages that are forwarded from every neighboring SGSN j to SGSN i. Therefore, this a_f depends on the mobility rate that users in the neighboring SGSNs of i have, as well as to the routing probability r_{ji}, where j is a neighbor of i and consequently on the buffer size of the neighboring SGSNs. Performance measures in the following two figures are the *average waiting time (AWT)* that a packet spend in queue before being transmitted and the *loss probability (LP)*. Note that *AWT* is evaluated for packets that are eventually transmitted and not for packets that expire or move out from the cell before they are transmitted. *LP* is defined as the ratio of the packet losses (blocked and expired) over the total number of packets that arrived in queue (before admission control takes place).

Figure 10 and Figure 11 plot *AWT* and *LP* respectively, for the seven SGSNs as functions of buffer size *B* for the first case. We can clearly observe that the performance measures of SGSN 0 are significantly affected by the forwarded traffic, while other SGSNs have about the similar performance.

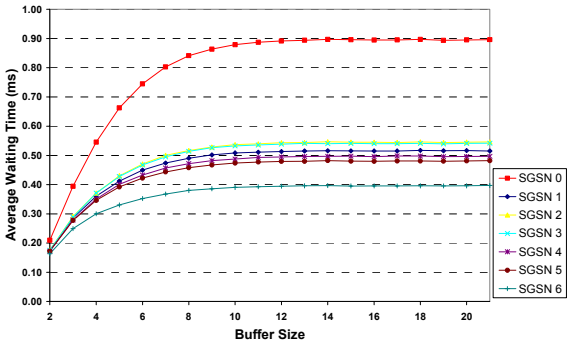

Figure 10. Average Waiting Time for SGSNs when CMR=1 and SER=10.

Next we present the optimization algorithm results for three different cases

- Case I: CMR=5 and SER=10
- Case II: CMR=10 and SER=10,
- Case III: CMR=100 and SER=10.

To simplify the presentations of the results, we illustrate the brute force simulation and optimization algorithm results for the SGSN 0 (hot spot) and SGSN 3 only, since, as we see in the previous figures, SGSN 1-6 have about the same performance.

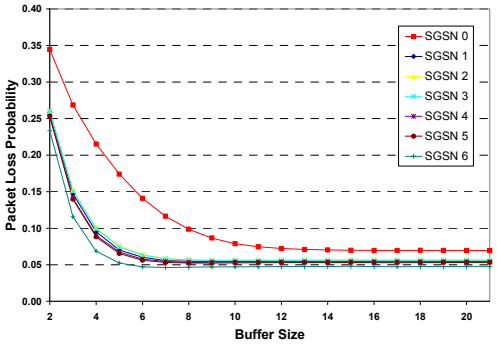

Figure 11. Loss probability for SGSNs when CMR=1 and SER=10.

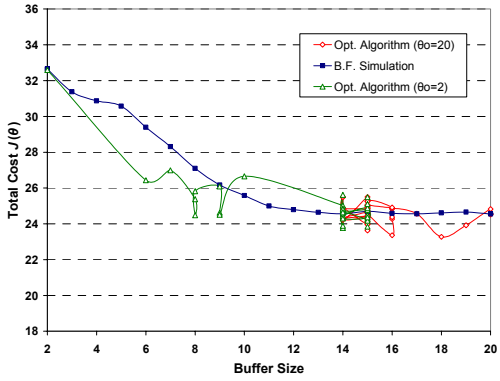

Figure 12. Opt. Algorithm's behavior in Case I for SGSN 0.

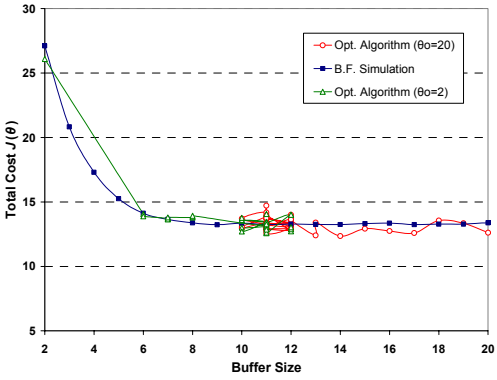

Figure 13. Opt. Algorithm's behavior in Case I for SGSN 3.

In Case I, the optimization algorithm quickly converges to the near optimal buffer size θ^*=15 for SGSN 0 (Figure 12) and θ^*=11 for SGSN 3 (Figure 13). We execute the optimization algorithm for two different initial values of θ, θ=2 and θ=20, observing that for both initial values the algorithms converges to the same values.

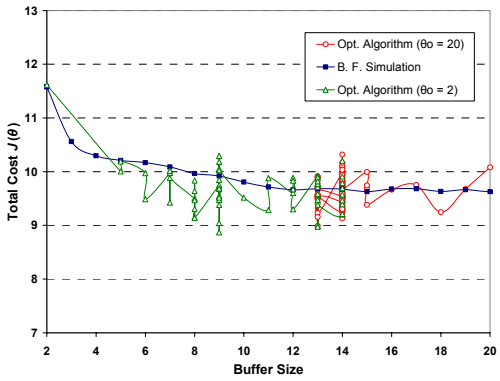

Figure 14. Opt. Algorithm's behavior in Case II for SGSN 0.

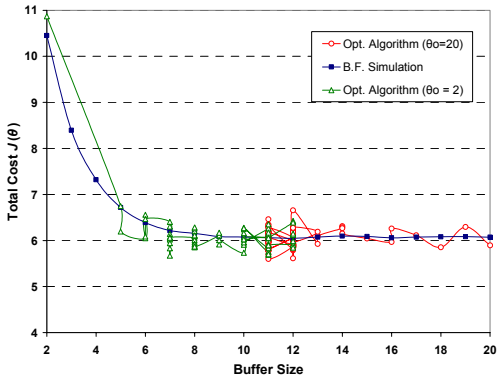

Figure 15. Opt. Algorithm's behavior in Case II for SGSN 3.

In Case II the optimization algorithm converges to the near optimal buffer size θ^*=14 for SGSN 0 (Figure 14) and θ^*=12 for SGSN 3 (Figure 15) . The actual optimal values are 15 and 12 for SGSN 0 and SGSN 3 respectively. Once more, the optimization algorithm is executed for two different initial values of θ, θ=2 and θ=20, with both cases ending to the same estimate for the optimal buffer size.

In case III, where CMR is fairly high, the optimization algorithm again converges to the near optimal buffer size. More precisely, its estimations are θ^*=5 for SGSN 0 (Figure 16) and SGSN 3 (Figure 17), while the actual optimal values are 4 and 5 respectively. These results suggest that for CMR values greater than 5, the optimization algorithm quickly converges to a near optimal buffer size.

5. Conclusions and Future Work

This paper presents a non-linear optimization approach for the on-line control of the buffer size in a SMS wireless network. The simulation results indicate that for moderate user mobility the approach exhibits good convergence properties. A question that we will investigate is how the algorithm can be modified to handle very high user mobility (e.g., highways). Furthermore, the obtained results are based on the cost function defined in (1). We are also investigating other performance measures (e.g., system throughput) as well as ways of providing service differentiation among different types of messages.

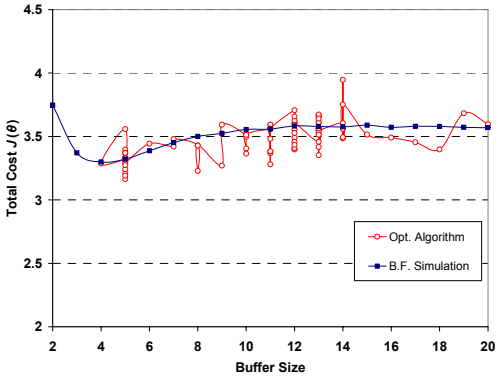

Figure 16. Opt. Algorithm's behavior in Case III for SGSN 0.

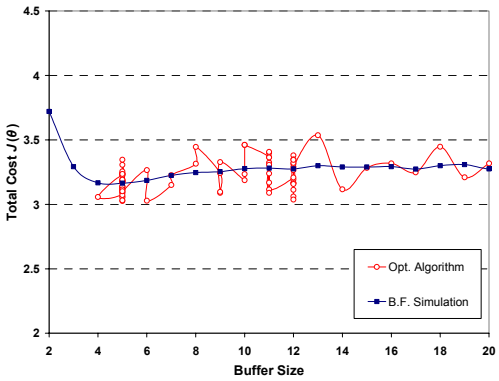

Figure 17. Opt. Algorithm's behavior in Case III for SGSN 3.

6. References

[1] 3GPP TS 23.140, "Multimedia Messaging Service (MMS); functional description"; stage 2, release 5, v5.1.0., 2001.

[2] Y. R. Haung, "Determining the Optimal Buffer Size for Short Message Transfer in a Heterogeneous GPRS/UMTP Network", *IEEE Transactions on Vehicular Technology*, vol. 52, no. 1, 2003.

[3] C.G. Cassandras and S. Lafortune, "Introduction to Discrete Event Systems", Kluwer Academic Publishers, 1999.

[4] 3GPP TS 23.060, "General packet radio service (GPRS); service Description"; stage 2; release 5, v5.0.0., 2002.

[5] C.G. Cassandras, Y. Wardi, B. Melamed, G. Sun, and C.G. Panayiotou, "Perturbation Analysis for On-Line Control and Optimization of Stochastic Fluid Models", IEEE Transactions on Automatic Control, vol. 47, no. 8, pp. 1234 -1248, 2002.

[6] Y.C. Ho and X.R. Cao, "Perturbation Analysis of Discrete Event Dynamic Systems," Kluwer, Boston, MA, 1991.

[7] H.J. Kushner and D.S. Clark, "Stochastic Approximation for Constrained and Unconstrained Systems", Springer-Verlag, Berlin, Germany, 1978.

[8] G. Sun, C.G., Cassandras, Y. Wardi, and C. Panayiotou, "Perturbation Analysis of Stochastic Flow Networks", *Proceedings of IEEE Conference on Decision and Control*, pp. 4831-4838, Dec. 2003.

[9] G. Sun, C.G. Cassandras and C.G. Panayiotou, "Perturbation Analysis of Multiclass Stochastic Fluid Models", *Journal of Discrete Event Dynamic Systems*, 2004. To appear.