International Journal of Multimedia and Ubiquitous Engineering
Vol. 4, No. 4, October, 2009

Delay Calculation of Internet Traffic in Mobile Systems

Seyed Mohammadreza Hashemiannejad
Mobile Communication Company of Iran
M.hashemian@mci.ir

Aziz Mahmoodi
Sadra University
AMahmoodi@sadra.ac.ir

Jahangir Dadkhah Chimeh
Iran Telecommunication Research Center
Dadkhah@itrc.ac.ir

Abstract

An important QoS factor for wireless network planning is delay. We should provide enough resources (bandwidth) for conveying the traffic load. Next generation wireless networks is going to be an all IP network. In this path, voice will be changed to packet. In this paper we introduce a queuing model and a radio resource management algorithm which maximize the system throughput (users), then compute their delays and show this delay is less than the mentioned delay in the 3GPP standards.

1. Introduction

Wireless data services include file, video and voice transfer over the mobile networks. Indeed Internet traffic is a combination of the above traffic services. In order to use radio resources optimally we need clear and precise models for different services. Services are divided to real time and non-real time services. Voice and peer to peer video communications are examples of the real time traffic [1]. Totally traffic models are based on the statistics characteristics of the services. The model presented in [2] like the 3GPP model presented in [3] uses a multi-layer model for describing sources and in the lowest layer, use Weibull and Pareto distributions for ON and OFF packet durations of Internet traffic respectively. The other model presented in [4] and [5] assigns Log-Normal and Pareto distributions to ON and OFF packet durations of Internet traffic respectively.

In third generation WCDMA systems, data applications are expected to finally dominate the overall traffic volume [6]. The traffic generated by data applications is inherently bursty and asymmetric, with higher data rates in the downlink than in the uplink [7].

Koo et al. has introduced a new delay confidence QoS parameter, then analysed a CDMA system capacity supporting voice and delay-tolerant date services based on that parameter [8]. Viterbi in [9] also reviewed the Erlang capacity in a power controlled CDMA system and compared CDMA systems with FDMA and TDMA systems from the capacity point of view.

Because Internet traffic is bursty, radio resource management module can transmit traffic information of other users in the empty sections of the bursts. Activity factor which depends on the traffic type and model, gives a criterion for this burstiness. Therefore advanced cognitive technologies such as spectrum sensing and spectrum mobility which handle packet based transactions, need to be included in radio resource management modules. Besides we treat real time and non-real time services in an all IP network, thus we consider a queue in this module that when there is not any resource for transmitting the packets, they will be inserted into it until transmitted in a suitable time.

Data users as well as the voice users access the services based on the Poisson model. These services also last according to an Exponential model [4], [10]. Thus we can consider their behavior as a M/M/m/K Markov model.

In this paper we consider Internet traffic services and treat voice as data (VoIP). Thus we

pay attention to the traffic load considerations in sections II. In section III we compute the queuing delay, then based on a traffic model and proposed RRM algorithm and that queue model we compute the queuing delay in section IV. Finally we draw the conclusion.

2. Traffic Considerations

Data traffic can be conveyed either through circuit switch systems or packet switch systems. Circuit switch systems have usually constant bit rates, while packet switch systems may have variable bit rates. So far there are some Erlang tables that are only pertinent to voice traffic in circuit switch systems.

Now to handle the data traffic we consider a queuing model which contains m servers. That system includes *K* customers (including the customers in service, K>m). Besides, we assume that the population is infinite (M/M/m/K). We can use this model for data services which are delay-tolerant because when all *m* channels are busy and upon reception of a new call attempt it can be inserted into the queue before it is lost. The birth and death coefficients in this situation are as follow (see also figure 1)

$$\lambda_n = \begin{cases} \lambda & n < K-1 \\ 0 & n \geq K \end{cases} \tag{1}$$

and

$$\mu_n = \begin{cases} n\mu & n \leq m \\ m\mu & m < n \leq K \\ 0 & K < n \end{cases} \tag{2}$$

in which *n* is the number of the subscribers in the queue whom their attempts have been accepted.

We assume P_n is the probability of being in the state *n* (or there exist *n* subscribers in the system). On the other hand it indicates the percentage of the time that the system contains *n* subscribers [11]. We can write [11]

$$P_n = C_n P_0 \tag{3}$$

in which

$$C_n = \begin{cases} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} & n < m \\ \left(\frac{\lambda}{\mu}\right)^n \frac{1}{m!}\frac{1}{m^{n-m}} & m \leq n \leq K \\ 0 & K < n \end{cases} \tag{4}$$

and

$$P_0 = \begin{cases} \dfrac{1}{1+\sum_{n=1}^{m-1}\left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}} & n < m \\ \dfrac{1}{1+\sum_{n=1}^{m-1}\left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}+\frac{1}{m!}\sum_{n=m}^{K}\left(\frac{\lambda}{\mu}\right)^n \frac{1}{m^{n-m}}} & m \leq n \leq K \end{cases} \tag{5}$$

We can write the above equation as ($r = \lambda/m\mu$)

$$P_0 = \begin{cases} \dfrac{1}{1+\sum_{n=1}^{m-1}\left(\dfrac{\lambda}{\mu}\right)^n \dfrac{1}{n!}} & n < m \\ \dfrac{1}{\sum_{n=0}^{m}\left(\dfrac{\lambda}{\mu}\right)^n \dfrac{1}{n!} + \dfrac{1}{m!}\left(\dfrac{\lambda}{\mu}\right)^m r\dfrac{(1-r^{K-m})}{1-r}} & m \le n \le K \end{cases} \quad (6)$$

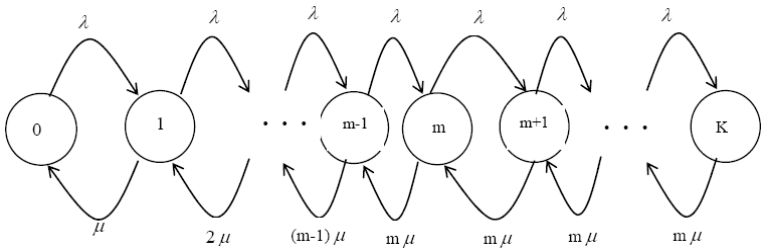


Figure 1. State transition-rate diagram for m servers, finite storage K and infinite population (M/M/m/K)

Now if we assume that there is not any queue in the system so that when all servers are busy the call attempts are lost, the above formula will change to Erlang B formula as

$$P_m = \frac{\left(\lambda/\mu\right)^m / m!}{\sum_{k=0}^{m}\left(\lambda/\mu\right)^k / k!} \quad (7)$$

in which P_m describes the fraction of time that all servers are busy. Calls have a (memory-less) exponential duration distribution with λ the arrival rate of new calls (birth rate) per unit time. h=1/μ Busy Hour Traffic (BHT), is the time duration (in the above unit time) of a call during the busiest period of the operation (we have assumed a call terminates with "rate" μ) We can show (7) by $B = (N, \lambda/\mu)$ and write

$$B(N, A) = \frac{(A)^N / N!}{\sum_{k=0}^{N}(A)^k / k!} \quad (8)$$

where B is probability of blocking, N is the number of trunks (channels) and $A = \lambda h$ total amount of the offered traffic in Erlang.

Because of the similarity in the traffic statistical models of the incoming and outgoing voice and data traffic users, we can use (3) for computing blocking for the three kinds of Telnet, www, Email traffics [4].

3. Queuing Delay

Delay is another important quality of service factor in the mixed traffic systems. According to [12] an end-to-end delay must not exceed 100ms and 200ms for voice and video services respectively. Now if W_q is the average long term waiting time of a subscriber in the queue

then in accordance with the rule of Littel we can write

$$L_q = \lambda W_q \tag{9}$$

in which L_q is the average number of subscribers in the queue in the long term and is equal to

$$\begin{aligned} L_q &= \sum_{n=0}^{\infty} nP_n \\ &= \frac{P_0}{m!}\left(\frac{\lambda}{\mu}\right)^m \cdot \frac{r}{(1-r)^2}\left[1 - r^{K-m+1} - (1-r)(K-m+1)r^{K-m}\right] \end{aligned} \tag{10}$$

with $r = \lambda / m\mu$.

Thus from (9) and (10) we can find W_q which must be greater than the above thresholds.

3. Simulation

A traffic is constituted of a set of sequential packets which are generated due to the user's traffic behavior (traffic model). We plan a radio resource management algorithm for handling www, Telnet, Email and packetized voice. The offered traffic of a user is normally transmitted through a radio link except when the link is busy in which case the offered traffic is inserted to a FIFO storage queue as they come, i.e. according to the packet arrival times they will be inserted to the queue independent of user type. This traffic will be transmitted in a suitable time. We can use preemptive-resume priority in which ongoing service is interrupted by arrival of higher customer, then the service time of the lower priority customer resumes at the point where it was interrupted [14]. We call this algorithm Maximum Bandwidth Usage (MBU) (figure 2).

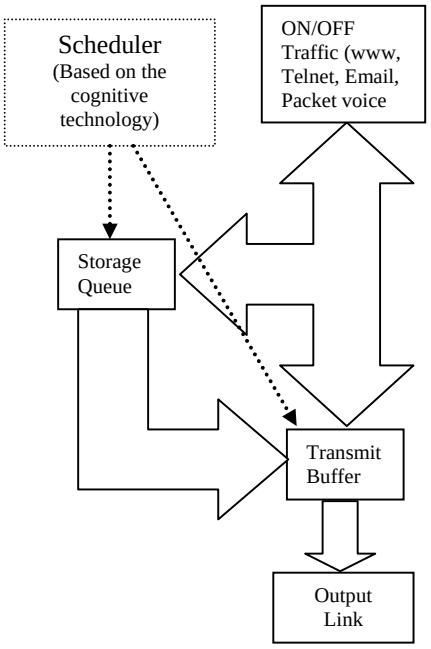


Figure. 2 Call processing layout

We assume system capacity *K(m)* and the number of channels *m* obey the equation

$$K(m)=m+ceil(\alpha *m) \quad (11)$$

which indicates that the larger is the number of channels the larger is the system capacity. This is because a network with larger number of channels must handle more traffic volume. $\alpha = 0.2$ *and* $\lambda = 40$ in this simulation.

Now we consider a system with 1,..,7, 12 and 20 serving channels and a queue with *K(m)-m* length (*K(m)* as in (11)). In simulation we consider various amounts of traffics (or users) and compute traffic delay in queue (figure 3), then according to Table 1 we found figure 4 for Email and Web browsing users each with the rate 384kbps.

Table 1. Activity Factors for data rate 384kbps [13]

	ON duration(s)	OFF duration(s)	Activity Factor
Telnet	0.217	112.79	0.0019
www	1.25	486.84	0.00256
E-mail	0.3	90.8	0.0033

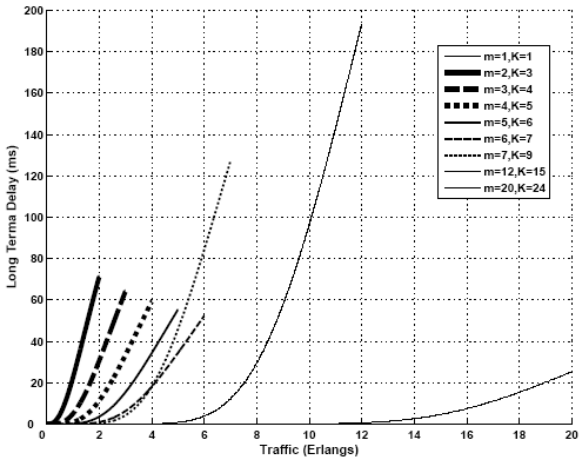
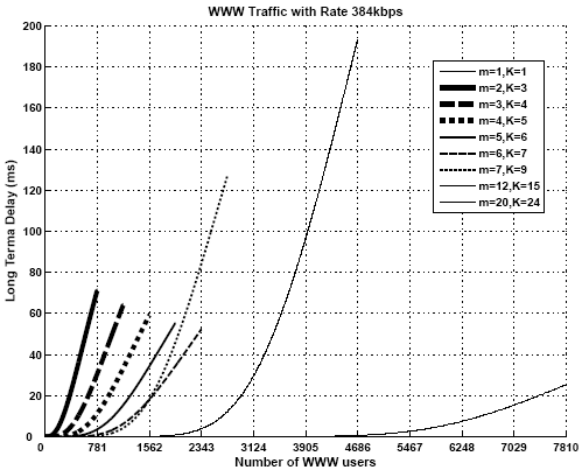
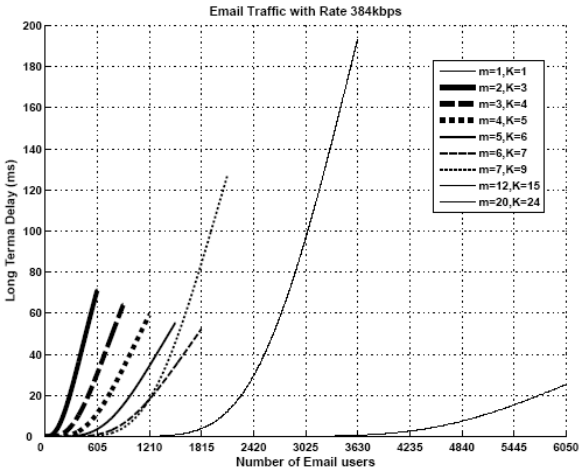


Figure 3. Long term delay versus traffic



(a)



(b)

Figure 4. Long term delay versus the number of the traffic users

As we know maximum carried traffic of m channels is m Erlangs. Thus each curve in figure 4 is plotted up to the amount that the carried traffic is valid. Figure 3 shows the minimum delay according to the proposed RRC algorithm. As we see, when the channel numbers increases the system delay decreases, but when the queue length increases (e.g. m=7, K=9) the delay increases after about 5.2 Elangs. Figures 4 show delay versus maximum WWW and Email users. As indicated in 3GPP standard [12], delay requirement of the traffic is 100ms for conversational and 300ms for streaming services which both are met by this algorithm.

3. Conclusion

We introduced a RRC algorithm and calculated a formula based on a Markov model for evaluating the delay in a wireless system. Then we plotted new curves for the long term delay of the traffic users versus the traffic volume and the number of traffic users and found out that this algorithm and model meets well the delay requirement mentioned in 3GGP standard.

4. Acknowledgement

At the end I feel it is necessary to thank Mobile Communication Company of Iran and Iran Telecommunication Research Center.

References

[1] J. Dadkhah Chimeh, M. Abdoli, M. Hakkak, “A New Radio Resource Management Algorithm for Mixed Traffic Transmission in Mobile System”, The 18th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC2007).

[2] Trung Van Nguyen “Capacity Improve-ment Using Adaptive Sectorization in WCDMA Cellular Systems With Non-Uniform and Packet Mode Traffic” Victoria University, Melbourne, PhD Thesis, March 2005.

[3] 3GPP, “Selection procedures for the choice of radio transmission technologies of the UMTS,” TR 102 112.1998.

[4] David Soldani “QoS Management in UMTS Terrestrial Radio Access FDD Networks” Helsinki University of Technology, Ph.D. Thesis, October 2005.

[5] A. Kuntz and et all., “Analysis of QoS requirement under radio access network planning aspects for GPRS/EDGE and UMTS,” International Conference on Wireless Networks, 2005.

[6] J. Grewal, J. DeDourek, “Provision of QoS in wireless netwoks“, CNSR'04, 2004.

[7] T. Ojanpera, et all., "An Overview of Third-Generation Wireless Personal Communications: A European Perspective", IEEE Personal Communication, pp. 59-65, 1998.

[8] I. Koo, J. R. Yang, K. Kim, ” Erlang Capacity Analysis of CDMA Systems Supporting Voice and Delay-Tolerant Data Services Under the Delay Constraint ”, IEEE Transactions on Vehicular Technology, Vol. 56, No. 4, July 2007.

[9] A. M. Viterbi, A. J. Viterbi, “Erlang Capacity of a Power Controlled CDMA System”, IEEE Journal of Selected Areas in Communications, Vol. 11, No. 6, August 1993.

[10] J.C. Bellamy, *Digital Telephony*, John Wiley, 2000.

[11] L. Kleinrock, *Queueing systems*, Vol I: theory, John Wiley $ Sons, 1975.

[12] 3GGP TS 23 107, Quality of Service (QoS) Concept and Architecture, 2007-7.

[13] J. Dadkhah Chimeh, M. Hakkak, P. Azmi, “Internet Traffic Modeling and Capacity Evaluation in UMTS”, *International Journal of Hybrid Information Technology, Vol. 1, No. 2, April 2008, Page(s): 109-120.*

[14] T. Janevski, Traffic analysis and design of wireless IP networks, Artech House, 2003.

Authors

Seyed Mohammadreza Hashemiannejad was born in Dezful, Iran. He received his BSC from Shahid Chamran (Jondishapur) University in 1995 and his M.S. from Islamic Azad University in Tehran/Iran in 2001 both in electrical engineering. He is now a telecommunication engineer and works in Mobile Communication Company of Iran (MCCI) since 1998. Besides, he has worked in Telecommunication Company of Iran (TCI) three years.

Jahangir Dadkhah Chimeh received his BSC from Sharif University of Technology in 1988 and his M.S. from K.N.T. University of Technology both in Tehran/Iran in 1994. He is now a telecommunication engineering Ph.D. student in Islamic Azad University Science and Research branch, Tehran, Iran. Besides, he is a faculty member of Iran Telecommunication Research Center (ITRC), Tehran. He was project manager of 3G mobile pilot project in 2005 and has some papers and a book in the mobile field.