Performance Estimation of UMTS Release 5 IM-Subsystem Elements

Y. Koucheryavy[1], A. Krendzel[1], S. Lopatin[2], J. Harju[1]

[1] Institute of Communication Engineering, Tampere University of Technology, P.O. Box 553, SF-33101, Tampere, Finland
{yk, krendzel, harju}@cs.tut.fi

[2] LONIIS, St.-Petersburg Research Institute of Telecommunications, 11, Warshawskaya str., 196128, St.Petersburg, Russia
derj@inser.loniis.spb.su

Abstract - **Universal Mobile Telecommunications System (UMTS) is the third generation (3G) mobile communications systems specified by European Telecommunications Standards Institute (ETSI) and the world-wide 3G Partnership Project (3GPP) within the framework defined by the International Telecommunication Union (ITU) and known as International Mobile Telecommunications-2000 (IMT-2000). ETSI and 3GPP are introducing UMTS in phases and annual releases. UMTS is a 3G GSM successor standard using the GSM Phase 2+ enhanced core networks. Currently, there is an evolution from GSM Phase 2+ core network towards UMTS Release 5 multiservice packet network. This process is characterized by creating of Internet Protocol (IP) Multimedia Core Network Subsystem (IM-subsystem), which includes elements necessary to support IM services in UMTS.**

Traffic patterns generated by IM services are different from traditional Poisson models used for circuit switched voice traffic. As a result, the network parameters can be underestimated. Therefore, the UMTS network planning problem arises. For successful solution of the problem it is necessary to estimate a performance of IM-subsystem elements. In this paper we have analyzed the performance of Gateway GPRS Support Node (GGSN) on a base of FBM/D/1/W queueing system (FBM - Fractional Brownian Motion). Submitted results may be applied for estimation of parameters of UMTS Rel'5 IM-subsystem elements.

Keywords: Universal Mobile Telecommunications System, IP-Subsystem, Gateway GPRS Support Node, Serving GPRS Support Node, Fractional Brownian Motion, self-similar traffic

1. Introduction

Currently, there is an ever-growing demand for mobility services and Internet services [1,2] in the telecommunication market. In this respect, UMTS as a part of the IMT-2000 standards family is one of the most important ETSI projects in creating the mass market for high-quality wireless multimedia communications [3]. ETSI and 3 GPP decided to prepare on a yearly basis UMTS specifications named Release' X, $X \subset \{3,4,5,...\}$. UMTS Rel'3 (sometimes called as Rel'99) incorporates enhanced GSM Phase 2+ Core Network (CN). The most important evolutionary step toward UMTS is to introduce a packet switched core network (PS CN) domain [4]. The main function of PS CN domain is to support all services (GPRS, WAP, etc.) provided to both GSM subscribers and UMTS users.

The following phases after Rel'3 specify how voice and multimedia can be supported by IP technology. A new core network subsystem called the IP Multimedia core network (IM CN) subsystem is introduced for the provision of the services [5,6] within UMTS Rel'5 such as, for example, voice telephony, real-time interactive games, videotelephony, instant messaging, emergency calls, multimedia conferencing [7]. The IM-subsystem comprises all PS CN elements for provision of multimedia services.

It is well known [8,9,10] that the use of the classic teletraffic theory for a performance calculation of packet multiservice network elements gives essential faults. In literature [11,12,13] the methods of overcoming such sort of difficulties are considered. We consider the results from [12] concerning self-similar multiservice traffic for the network planning problem solution of Gateway GPRS Support Node (GGSN) probabilistic and time characteristics estimation of UMTS Rel'5 IM-subsystem.

2. UMTS Rel' 5 core network architecture

The architecture of UMTS Rel'5 core network based on configurations [5] is presented on Fig. 1. The shaded area in the picture shows objects included into Intelligent Network/Mobile Management (IN/MM) platform. IM CN subsystem is also denoted.

The main multimedia load created by the network users passes through GGSN. GGSN uses Gn interface to communicate to a set of Serving GPRS Support Nodes (SGSNs) when ones are located inside its network, and Gp interface if the GGSN and SGSNs are located in different networks.

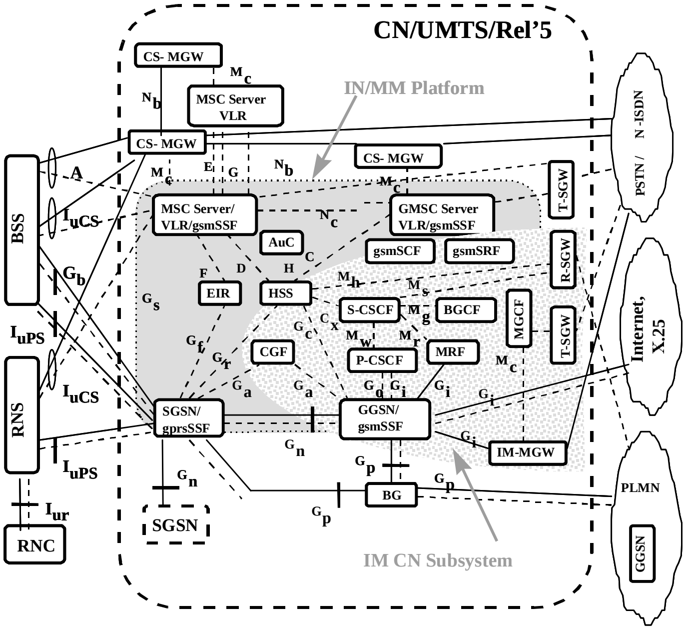

Fig. 1. UMTS Rel'5 core network

The principle of flows aggregation between users and GGSN leads to increasing of the load on the network elements while approaching to GGSN. The network load on GGSN is determined by the intensity of IP packets arrival rate. Currently, a transport technology for IP packets delivery to/from GGSN is not defined uniquely. For example, ATM may be applied as one of the possible cases of such technology.

3. *The load model*

Let us assume that values *s(t)* of a random process with interdependent increments are the total load arrived to the node (server) in the time point t>0. Current values *s(t)* in the time interval [0, t) may be determined by the information units number. If the corresponding process is ordinary then an increment is one information unit. The increments intensity is the rate parameter λ, 1/sec. Realizations of process *s(t)* are non-decreasing step functions with increments taking place in random time points.

Let us consider a random variable $S_T = s(t=T), T \subset [0,t)$ that is a sample of a random process *s(t)*. By definition S_T is a sum of interdependent identically distributed random variables. If $E(S_T) = \lambda T >> 0$ then conditions of the central limiting theorem are fulfilled. Here, $E(\cdot)$ is operator of statistical averaging. Accordingly, S_T may be approximated by a Gaussian random variable [14]. Taking into account abovementioned assumptions S_T may be defined as:

$$S_T = \lambda T + \sqrt{b(T)} \cdot x, \tag{1}$$

where *x = N(0,1)* is a normalized Gaussian random variable with the zero mean and the unit variance, *b(T)* is a variance of S_T.

If $b(t) = \sigma^2 t$ and $t > 0$ then the univariate probabilities distribution of the process *s(t)* coincides with the corresponding distribution of a Brownian motion process or a displaced Wiener process [15,16]. Similarly, the process *s(t)* corresponds to a Poisson process if condition (1) is fulfilled when *x* is a Poisson random variable and $b(t) = \lambda t$.

It is necessary to take into account a self-similarity notion when load modeling in packet data networks. Self-similarity is in different networks, in particular, in local area networks [17], Internet [10], wireless networks [18] and others. It is characterized by stronger dependence of a variance from time then a line dependence.

There are different ways of self-similarity load modeling [11, 12, 17]. With reference to (1) self-similarity may be taken into account as

$$b(T) = (\sigma^2 T)^{2H}, \qquad 0.5 \le H < 1, \tag{2}$$

where H is the Hurst parameter. Expressions (1) and (2) specify a model of a total traffic load arriving to a server input by a time point *t=T*.

4. *The queueing system model*

We assume that *s(t)* arrives to the server input. The server is modeled by queueing system with deterministic rate *C*, 1/sec and the buffer size *(W-1)*, $1 \le W < \infty$. The queueing system is the stable one because there is a stationary probability distribution if $C > \lambda$. In accordance with the Kendall's notation for queues the system is *G/D/1/W* [19]. The corresponding system may be also defined as *FBM/D/1/W* [12] if the expressions (1,2) are fulfilled. Here, the FBM is a normalized fractional Brownian motion, i.e. the corresponding process is a strictly self-similar one.

5. *The task estimation definition*

When stability conditions are fulfilled the average value of the total load arrived to the queueing input by the time point *t=T>0* is less than the queueing system can serve for the same time interval. It should be emphasized that the load is a random process. Therefore, it is possible to appear an event when the buffer will be overflow. The probability of the event is defined by statistical properties of an unserved traffic process that may be written as

$$V(t) = \max[0, S(t) - Ct] \tag{3}$$

The introduction of operator max [0, x] in (3) is caused by nonnegative values of an unserved load. It is similar to the introduction of an adsorbing barrier in the coordinate origin point for the displaced self-similar (fractional) Wiener process.

The estimation problem is to determine values of parameters C and W. It should be done by taking into account the following condition. The probability that the unserved load will be greater than the parameter W must not exceed the preset threshold ε:

$$P[V(t) > W] = \varepsilon, \qquad t > \varepsilon \quad 0 < \varepsilon << 1 \tag{4}$$

6. The estimation task solution

Taking into consideration the approximation of the random process *s(t)* sample by the random Gaussian variable defined by the expressions (1,2) we have the lower bound for the buffer saturation probability

$$P[V(t) > W] \geq \max_{T>0} P[x > \alpha(T)], \tag{5}$$

where $\alpha(T) = [(C-\lambda)T + W]/(\sigma^2 T)^H$.

The expression (5) shows that the probability of events union is not less than the probability of each event. Taking into account that the random variable X is the normalized displaced Gaussian random one, the expression (5) may be transformed as

$$P[V(t) > W] \geq \max_{T>0}\left[\int_{\alpha(T)}^{\infty} \exp(-x^2/2)dx/\sqrt{2\pi}\right] \approx \max_{T>0}\left[0.5\exp(-\alpha^2(T)/2)\right] \tag{6}$$

Let us take into consideration the logarithmic function monotony for the expression that is equivalent (4,6). Then the expression binding the parameters *C, W,* λ and the buffer saturation probability ε is

$$-\ln\varepsilon \approx \min_{T>0}\frac{((C-\lambda)T + W)^2}{(\sigma^2 T)^{2H}} \tag{7}$$

The solution of the equation (7) may be found by the parameter T differentiation and equating of the obtained derivative with zero [12]. It gives the following expression

$$T_m = \frac{WH}{(1-H)(C-\lambda)} = \arg\min_{T>0}\frac{((C-\lambda)T + W)^2}{(\sigma^2 T)^{2H}} \tag{8}$$

Substituting T_m in (7) and transforming the expression we finally get

$$C/\lambda = 1 + n\left(\frac{\sqrt{-2\ln\varepsilon}\,W^{H-1}H^H}{(1-H)^{H-1}}\right)^{1/H}, \qquad 0.5 \leq H < 1, \tag{9}$$

where $n = \sigma^2/\lambda$.

Substituting *n, W,* λ and ε values in (9) we get the upper bound (if H=0.8) and lower bound (if H=0.5) of the server service rate C. One of the main parameters of queueing system is the inverse parameter (9) $\rho = \lambda/C$ called the utilization factor. Fig. 2 illustrates dependences of the utilization factor from the magnitude $n = \sigma^2/\lambda$ for various values of the Hurst parameter and the buffer capacity W when $\varepsilon = 10^{-5}$.

Trends of the curves (Fig.2) show that it is very important to take into account the self-similarity influence while assigning server parameters. According to [20] the Hurst parameter of the corresponding Internet traffic may achieve values exceeding 0.8. It should be emphasized that the area of the dependences when n=1 and H=0.5 (as shown in Fig.2) corresponds to the case of the Poisson arrival process.

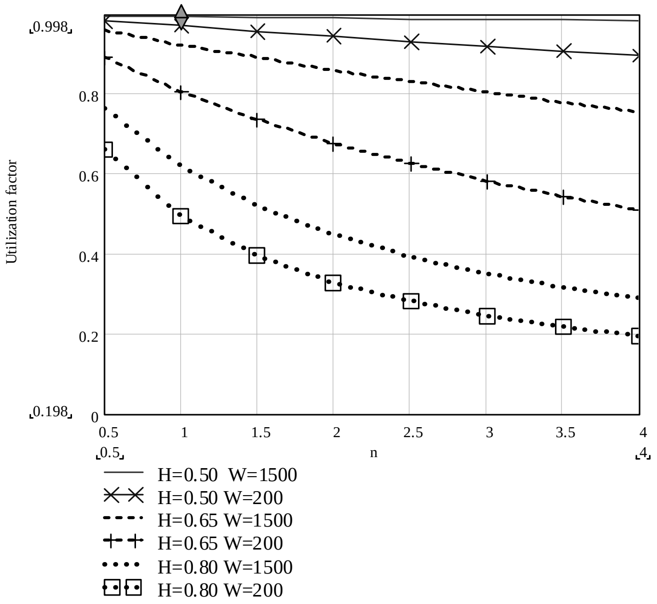

Fig.2. Self-similarity influence on the server utilization factor

It is wise to take the value ε sufficiently small. It enables to have the acceptable probability of massages blocking arriving on the server. In this case the server buffer will be filled partly.

For determination of the upper bound of the average queue length in the server buffer the expression based on results [13,21] may by used:

$$q_{\max} = \frac{(\lambda/C)^{1/2(1-H)}}{(1-\lambda/C)^{H/(1-H)}} \tag{10}$$

The classical result for M/D/1 system may be applied for determination of the lower bound of the average queue length:

$$q_{\min} = \frac{\lambda/C}{1-\lambda/C} - \frac{(\lambda/C)^2}{2(1-\lambda/C)} \tag{11}$$

Little formula [14] enables to determine the upper and the lower bound for the average service time τ using expressions (9-11):

$$\tau_{\max} = q_{\max}/C, \qquad \tau_{\min} = q_{\min}/C, \tag{12}$$

Expressions (9-12) allow estimating bounds of the probabilistic and time characteristics of the single server under the self-similarity load influence.

7. *Case study*

As it may be seen from Fig.1, the GGSN is the node that the most exposed to the self-similarity influence in UMTS. The parameters characterizing the server normal functionality may be estimate by the following way.

Currently, there are no exact regulations on transport network protocols on SCSN-GGSN interface. Therefore, let us assume ATM as underlying technology for IP packets delivery. The rate of information units (ATM cells) arriving on SGSN is multiple to k*2 Mbit/sec.

Let us consider the following example. Let k=10 and in average 30% of the channel throughput is in use during the messages delivery on the RNC – SGSN interface. Then, the value of the ATM cells arriving intensity on the SGSN input is $\lambda \approx 15000\ \mathrm{sec}^{-1}$. If a number of SGSNs connected to the GGSN is 3 then the total value of the ATM cells arriving intensity on GGSN input is $\lambda \approx 45000\ \mathrm{sec}^{-1}$. In accordance with (9-12) the dependences of the GGSN capacity, the upper bounds for average length queue and delay time as functions of *n* $(n = \sigma^2/\lambda)$ are shown in Fig. 3,4,5 respectively ($\varepsilon = 10^{-7}$, W = 100, 500, H = 0,75).

It should be emphasized that the similar approach may be used for SGSN performance calculations.

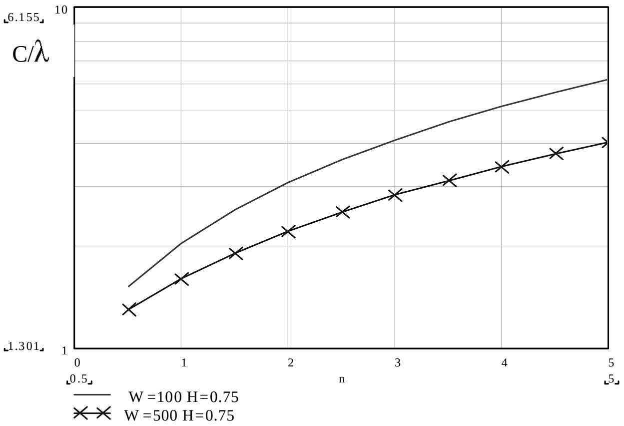

Fig.3. The estimation of the GGSN server capacity

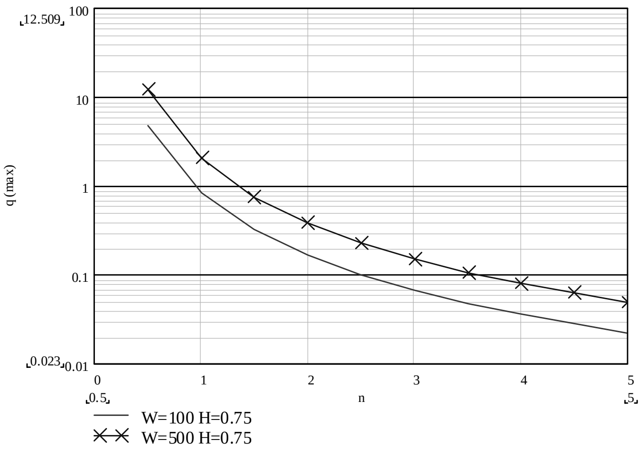

Fig. 4. The upper bound for average queue length in the GGSN buffer

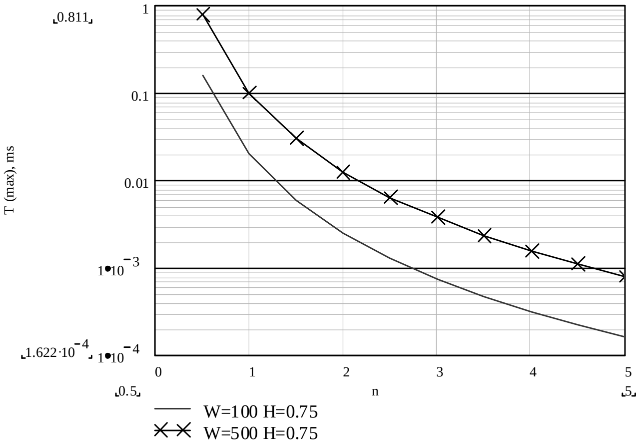

Fig. 5. The upper bound for the average service time in the GGSN

8. *Conclusion*

In this paper we have analyzed the influence of self-similar input on GGSN performance in UMTS Rel'5 IM-subsystem. We have considered the FBM/D/1/W queueing system for GGSN parameters estimation. The submitted approach enables determining the following probabilistic and time characteristics: upper and lower bounds of the GGSN service rate, the average queue length in the server buffer, the average service time of information units. The obtained results point to a need to take into account self-similarity while assigning the GGSN parameters.

As well known, when providing multimedia services based on IP technologies one of the main aspects is Quality of Service guarantee. From this point of view the proposed approach may be used for 3G mobile systems performance estimation, in particular, for UMTS Rel'5 IP multimedia subsystem planning.

ABBREVIATIONS

3G - Third Generation
3GPP - Third Generation Partnership Project
ATM - Asynchronous Transfer Mode
AuC – Authentication Center
BG - Border Gateway
BGSF – Breakout Gateway Control Function
BSS - Base Station System
CGF - Charging Gateway Functionality
CS - Circuit Switched
CSCF – Call Session Control Function
CS-MGW – Circuit Switched – Media Gateway Function
EIR – Equipment Identity Register
ETSI - European Telecommunication Standard Institute
GGSN – Gateway GPRS Support Node
GMSC – Gateway MSC
CN – Core Network
GPRS – General Packet Radio Service
gprsSSF – GPRS Service Switching Function
GSM - Global System for Mobile communications
gsmSCF – GSM Service Control Function
gsmSRF – GSM Specialized Resource Function
gsmSSF – GSM Service Switching Function
HSS – Home Subscriber Server
IM-MGW – IP Multimedia – Media Gateway Function
IM subsystem – IP Multimedia subsystem
IMT-2000 - International Mobile Telecommunications System – 2000
IN - Intelligent Network
IP – Internet Protocol
ISDN – Integrated Services Digital Network
ITU - International Telecommunication Union
MGW – Media Gateway
MGCF – Media Gateway Control Function
MRP – Multimedia Resource Function
MM – Mobile Management
MSC – Mobile Switched Center
N-ISDN – Narrowband ISDN
P-CSCF – Proxy CSCF
PLMN - Public Land Mobile Network
PS - Packet Switched
RNC – Radio Network Controller
RNS – Radio Network System
R-SGW – Roaming SGW
PSTN – Public Switched Telephone Network
S-CSCF – Serving CSCF
SGSN – Serving GPRS Support Node
SGW – Signalling Gateway Function
T-SGW – Transport SGW
VLR - Visit Location Register
UMTS - Universal Mobile Telecommunication System

REFERENCES

1. "Long term potential remains high for 3G mobile data services," *Report # 18 from the UMTS Forum*, February 2002.
2. H. Kaaranen, A. Ahtiainen, L. Laitinen, S. Naghian, V. Niemi, "UMTS networks. Architecture, mobility and services," *John Wiley & Sons*, 2001.
3. "Support of third generation services using UMTS in a converging network environment," *Report # 14 from the UMTS Forum*, February 2002.
4. ETSI TS 123 060 V3.5.0 (2000-10), "Technical specification digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); General Packet Radio Service (GPRS); Service description; Stage 2," 3GPP TS 23.060 version 3.5.0, Release 1999.
5. 3GPP TS 23.002 V5.5.0 (2002-01), "3rd Generation Partnership Project (3GPP); technical specification group services systems aspects; network architecture (Release 5)".
6. "Naming, addressing & identification issues for UMTS," *Report # 12 from the UMTS Forum*, February 2001.
7. M.C. Bale, "Voice and Internet multimedia in UMTS networks," *BT Technology Journal*, vol. 19, №.1, 2001.
8. W. Willinger, D. Wilson, M. Taqqu, "Self-similar traffic modeling for high-speed networks," *ConneXions*, November 1994.
9. W. Stallings, "High speed networks. TCP/IP and ATM design principles," *Upper Saddles River*, NJ: Prentice-Hall, p 576, 1998.
10. J.B. Roberts, "Traffic theory and the Internet," *IEEE Communications Magazine*, pp.94-99, January 2001.
11. R.G. Addie, M. Zukerman, T.D. Neame, "Broadband traffic modeling: simple solutions to hard problems," *IEEE Communications Magazine*, pp. 88-95, August 1998.
12. I. Norros, "On the use of fractional Brownian motion in the theory of connectionless networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, № 6, pp. 953-962, August 1995.
13. I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 16, pp. 387-396, 1994.
14. L. Kleinrock, "Queueing systems. Computer applications," *John Wiley & Sons*, vol.2, 1976.
15. I. Karatsas, S. Shreve, "Brownian motion and stochastic calculus," *New York: Springer-Verlag*, 2nd ed., 1997.
16. A. Papoulis, "Probability, random variables, and stochastic processes," *McGraw-Hill, New York*, 2nd ed., 1984.
17. W. Willinger, M.S. Taqqu, W.E. Leland, D.V. Wilson, "Self-similarity in high-speed packet traffic: analysis and modeling of Ethernet traffic measurements", *Statistical Science*, vol. 10, no. 1, pp. 67-85, 1995.
18. M. Jiang, M. Nicolic, S. Hardy, L. Trajkovic, "Impact of self-similarity on wireless data network performance," *ICC 2001*, *USA*, June 2001.
19. L. Kleinrock, "Queueing systems. Theory", *John Wiley & Sons*, vol. 1, 1975.
20. Y. Koucheryavy, A. Krendzel, "Analysis of Web traffic and users' behavior modeling during busy hour," *6-th EUNICE Open European Summer School, University of Twente, the Netherlands*, pp. 9-12, September 2000.
21. I. Norros, A. Simonian, J. Virtamo, "The Benes method – a unified approach to ATM FIFO queueing," *Proceedings of St. Petersburg International Teletraffic Seminar, St. Petersburg*, pp. 431-449, June 1995.