Analyzing Packet Delay Across A GSM/GPRS Network

Xiaoyan Fang [1] and Dipak Ghosal [2]
[1] Department of Electrical and Computer Engineering
[2] Department of Computer Science
University of California, Davis
Davis, CA 95616, USA
jamesfxy@ece.ucdavis.edu, ghosal@cs.ucdavis.edu
Contact Author: Dipak Ghosal

***Abstract*—**

In this paper we study two major sources of packets delay in the GSM/GPRS wireless network, at Base Station and at GGSN node. Fisrt, for the former one, we present an analytical model to study the performance of channel sharing schemes to support both circuit switched voice and packet data services in a GSM/GPRS network. We study three channel sharing schemes: 1) *fixed sharing* in which cell channels are statically partitioned into two sets one for voice calls and the other for data traffic; 2) *partial sharing* in which n_{data}channels are reserved for data while the remaining $N - n_{data}$channels are shared by voice and data with preemptive priority for voice calls; and 3) *complete sharing* in which all the channels shared by voice and data with preemptive priority to voice calls. We investigate several key issues such as call blocking rate and mean packet delay for different cell loads with the data source modeled by a Markov Modulated Poisson Process (MMPP). We validate the mathematical model through simulations and quantify the impact of the data source model and the voice call load on the mean packet delay for different channel sharing schemes. Secondly, for another souce causing delay, we present the analytical model to quantify the benefit of replicated GGSN with load balancing architecture.Our results show that replicated GGSN architecture with load balancing policy can significantly reduce the packet delay even when the total GGSN capacity remains constant.

I. Introduction

The second generation GSM (Global System for Mobile communications) mobile systems have been introduced into the commercial market for over a decade. The number of global subscribers of GSM has reached to over 860 million, which accounts 78% for the total wireless subscriber in the world by beginning of year 2002 [1]. Before year 2001, most the services provided by the GSM carriers were circuit-switched based built upon a basic data rate service of 9.6 Kbps. With the rapid deployment of IP and IP based services, wireless carriers have now introduced new data services based on packet-switching techniques. GPRS (General Packet Radio Service) with maximum data rate 170 Kbps and EDGE (Enhanced Data for Global Evolution)(384 Kbps) are two well defined and mature technologies that are currently being deployed.

In the design of the first and second generation cellular mobile telephony systems (such as ETACS (Enhanced Total Access Communication System) and GSM in Europe), the technical approach mainly focused on increasing the capacity available for voice services, so as to cope with the explosive growth in the number of subscribers. Today, the need for an increased system capacity is combined with the request for a wider spectrum of telecommunication services, in order to be able to offer data services in addition to plain telephony. This will pave the way to the introduction of wireless multimedia services for mobile users, including voice, data and images. While the performance of cellular telecommunication networks offering mobile telephony services has been investigated [2][3][4] under several different operating conditions, the same cannot be said of networks offering a variety of services in particular voice and data services to mobile users.

First, in this paper, for the packet delay occured at the BTS(base stations), three different sharing schemes have been studied at MAC/RLC layer between the mobile terminals and the base station. Besides, the effect of data source pattern and the voice call load on mean data packet delay is also investigated. We develop an analytical model based on the GSM/GPRS MAC/RLC layer. We provide detailed solution for the analytical model. Based on the analytical model, We have investigated several key QoS issues and their relationship, channel sharing schemes, the offered load and the characteristics of the data traffic. We verify the analytical model using simulation results. Our results quantify the impact of the sharing scheme, data source pattern and the voice call load on the mean packet delay.

Second, in GPRS network, two nodes are added to the standard GSM system: the Gateway GPRS Gateway Support Node (GGSN) and the Serving GPRS Support Node (SGSN). The GGSN acts as a gateway between the Internet (and other GGSNs) and a provider's private GPRS network . The GGSN's main function is to tunnel packets from outside networks (other GPRS networks and the Internet) to the SGSN currently serving the mobile. It does this through an IP based GRPS backbone. In this paper, We propose a load balance scheme to improve the QoS of data traffic at the GGSN nodes. By defining the model, We provide the numerical analysis of the mean packect delay to show the improvement of QoS by this scheme.

In section §II, We give an overview of GSM/GPRS-the network architecture, the protocol layers and the radio interface.

Based on the these, we address the sources of the packect delay in GSM/ GPRS network. In section §III-A We describe the analytical model and the underlying assumptions of the MAC/RLC layer at BTS. The analytical and simulation results are discussed in section §III-A.5. The GGSN nodes with balancing policy is introduced in section §III-B.1 which gives the architecture and routing protocol. The modelling analysis and results are presented in section §III-B.2. In §IV We present the related literature. Finally in section §V We draw the conclusion and outline the future research.

II. GSM/GPRS NETWORK

A. GSM/GPRS Overview

1) System Architecture : GPRS is considered as a service or feature of GSM [12]. Figure 1 illustrates the logical architecture of a GSM network supporting GPRS. GPRS has minor impact on the existing GSM BSS (Base Station System) making it easy to reuse existing component and links without major modifications. This is possible because GPRS uses the same frequency bands and hopping techniques, the same TDMA frame structure, the same radio modulation and burst structure as GSM. A new functional component called packet control unit (PCU) was added to the BSS in the GPRS standard to support the handling of data packets. The PCU (not shown in Figure 1) is placed logically between the BSS and the GPRS NSS (Network Sub-System). Unlike the voice circuit connections however, connections in GPRS have to be established and released between the BSS and the MS (Mobile Station) only when data need to be transported over the air interface. This allows several GPRS users can share the same channel which dramatically increase the bandwidth efficiency.

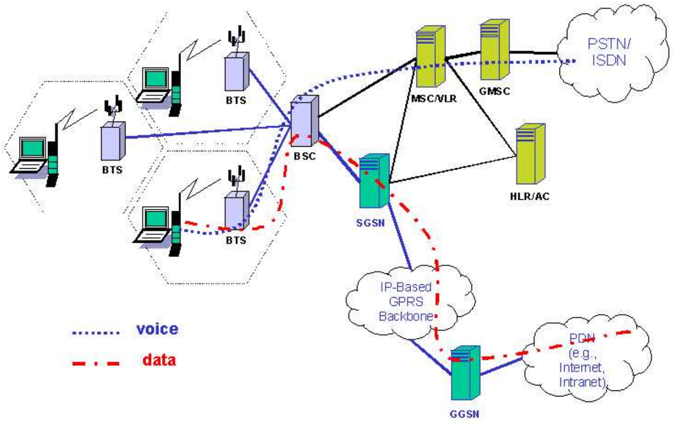

Fig. 1. GSM/GPRS system overview.

The GPRS NSS can be viewed as an overlay network ensuring the link between mobile users and data networks. GPRS introduces a new functional element to the GSM infrastructure as shown in Figure 1: GPRS support node (GSN) which can be either a serving-GSN (SGSN) or a gateway-GSN (GGSN). GGSN provides interworking with external packet-switched networks, publishing subscriber addresses, mapping addresses, routing and tunnelling packets, screening messages, and counting packets.

2) Protocol Architecture : A layered protocol structure is adopted for the transmission and signaling planes in GPRS (Figure 2). The subnetwork dependent convergence protocol (SNDCP) serves as a mapping of the characteristics of the underlying network such as IP. Mobility management functionality is supported by the GPRS mobility management (GMM) and session management (SM) layers. The logical link control (LLC) layer provides a logical link between the MS and the SGSN and manages reliable transmission while at the same time supporting point-to-point and point-to-multipoint addressing. The radio link control (RLC), medium access control (MAC), and GSM RF (radio frequency) layers control the radio link, the allocation of physical channels and radio frequency. LLC PDUs (packet data units) between the MS and the SGSN are relayed at the BSS. The base station system GPRS protocol (BSSGP) layer handles routing and QoS between the BSS and the SGSN. The GPRS tunneling protocol (GTP) is the basis for tunnel signaling and user PDUS between the SGSN and GGSN. Further description can be found in the paper [12]. On the physical layer, GSM uses a combination of FDMA and TDMA for multiple access. Two frequency bands 45 MHz apart have been reserved for GSM operation: $890 - 915$MHz for transmission from the mobile station, i.e., uplink, and $935 - 960$ MHz for transmission from the BTS, i.e., downlink. Each of these bands of 25 MHz width is divided into 124 single carrier channels of 200 kHz width. A certain number of these frequency channels, the so-called cell allocation, is allocated to a BTS, i.e., to a cell [10]. Each of the 200 kHz frequency channels carries eight TDMA channels by dividing each of them into eight time slots. The eight time slots in these TDMA channels form a TDMA frame. Each time slot of a TDMA frame lasts for a duration of 156.25 bit times and, if used, contains a data burst. A burst is a period of RF carrier which is modulated by a data source. It therefore represents the physical content of a timeslot. A timeslot is divided into 156.25 symbol periods. For GMSK modulation a symbol is equivalent to a bit. For 8PSK modulation one symbol corresponds to three bits. The time slot lasts $15/26ms = 576.9\mu s$; so a frame takes 4.613 ms. The recurrence of one particular time slot defines a physical channel. A

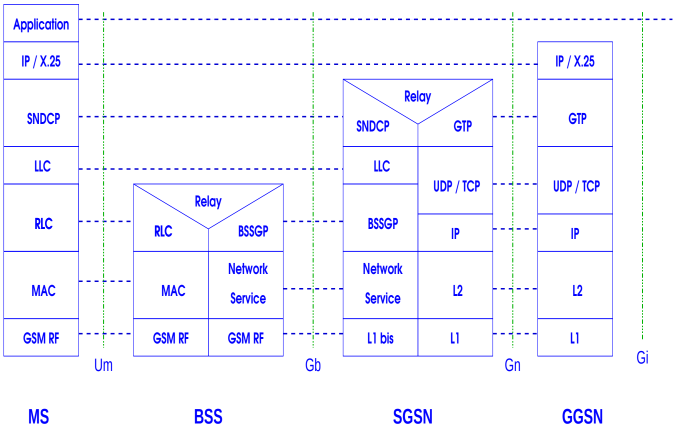

Fig. 2. GPRS protocol architecture.

GSM mobile station uses the same time slots in the uplink as in the downlink. The channel allocation in GPRS is different from the original GSM. GPRS allows a single mobile station to transmit on multiple time slots of the same TDMA frame (multislot operation). This results in a very flexible channel allocation: one to eight time slots per TDMA frame can be allocated for one mobile station. Moreover, uplink and downlink are allocated separately, which efficiently supports asymmetric data traffic (e.g., Web browsing).

3) Channel mapping and allocation algorithms: In conventional GSM, a channel is permanently allocated for a particular user during the entire call period (whether data is transmitted or not). In contrast to this, in GPRS the channels are only allocated when data packets are sent or received, and they are released after the transmission. For bursty traffic this results in a much more efficient usage of the scarce radio resources. With this principle, multiple users can share one physical channel.

Traffic channels (TCHs) are intended to carry either encoded speech or user data in circuit switched mode. All traffic channels are bi-directional Multiple packet data traffic channels can be allocated to the same MS. A PDTCH/F corresponds to the resource allocated to a single MS on one physical channel for user data transmission. Due to the dynamic multiplexing onto the same physical channel of different logical channels, a PDTCH/F using GMSK modulation carries information at an instantaneous bit rate ranging from 0 to 22.8 kbit/s. A PDTCH/F using 8PSK modulation carries information (including stealing symbols) at an instantaneous bit rate ranging from 0 to 69.6 kbit/s [11].

The physical channel dedicated to packet data traffic is called a Packet Data Channel (PDCH). Packet data traffic channels (PDTCH's) are intended to carry user data in packet switched mode. It is a channel allocated for data transfer. In the multislot operation, one MS may use multiple PDTCHs in parallel for individual packet transfer. Different packet data logical channels can occur on the same physical channel (i.e. PDCH)[9].

A cell supporting GPRS may allocate physical channels for GPRS traffic. The PDCHs are taken from the common pool of all channels available in the cell. Thus, the radio resources of a cell are shared by all GPRS and non-GPRS mobile stations located in this cell. The mapping of physical channels to either packet switched (GPRS) or circuit switched (conventional GSM) services can be performed dynamically (capacity on demand principle), depending on the current traffic load, the priority of the service, and the multislot class. According to the current demand, the number of channels allocated for GPRS (i.e., the number of PDCHs) can be changed. Physical channels not currently in use by conventional GSM can be allocated as PDCHs to increase the quality of service for GPRS. When there is a resource demand for services with higher priority, PDCHs can be de-allocated.

The mapping of logical channels onto physical channels has two components: mapping in frequency and mapping in time. The mapping in frequency is based on the TDMA frame number and the frequencies allocated to the BTS and the mobile station. The mapping in time is based on the definition of complex multiframe structures on top of the TDMA frames. Four consecutive TDMA frames form one block. One radio block has 625 bits with duration of 18.452 ms. Based on the definition of the normal burst(NB), the data bits for a radio block is 456 bits. The mean throughput under coding scheme 4(CS-4) of a PDCH equals to $456/(4.613*52) = 22.8$kbps[9]. One PDTCH is mapped onto one physical channel. Up to eight PDTCHs, with different timeslots but with the same frequency parameters, may be allocated to one MS at the same time.

B. *Source of Delay*

First, the data traffic is asymmetric at the wireless channels which means the down link traffic volume is much higher than the one in the uplink. In this paper, we consider the delay caused by the down link channel. Besides, the circuit-switched voice traffic will compete with packet-switched data packets for the source-limited wireless channels at BTS, this will cause to even worse delay of the down link data packets. Second, we can see in figure 1, the incoming packets from the PDN(Packet Data Network) such as Internet to the GPRS network will go through GGSN node. One GGSN is related to several SGSN. GGSN node will tunnel the data packects to relatavie SGSN nodes which maintain the mobility management of mobile stations in different routing area (RA). The aggregated data traffic within one RA is usually different with the one in the other RA, in which case will cause different traffic load on SGSN and GGSN nodes. The GGSN with high traffic load will cause more packet delay than the one with low traffic load. For real-time multimedia application in wirless network, packet delay is major issue of QoS. In the following sections, we will study the delay quantitively by modeling and analysis.

III. Modelling & Discussion

A. *MAC/RLC layer Sharing Algorithm*

In this paper, we study the following three channel sharing policies to address the delay occured at the wireless channels:

- **Fixed Sharing:** In fixed sharing, the N cell channels are statically partitioned into two parts - one is use by the voice calls and the other by the data traffic.
- **Partial Sharing:** In partial sharing, n_{data} channels are reserved for data traffic while the rest of the channels ($N - n_{data}$) are shared by both the voice call and data traffic. Voice call has higher (preemptive) priority over the data packets. Thus, if all the channels are busy, an incoming voice call will preemptively acquire a channel used for data traffic, if the number of channels used by the data traffic is is more than n. If channels are available, a data "call" will acquire a free channel based on first come first serve(FCFS)
- **Complete Sharing:** In complete sharing, all the channels are shared by voice calls and data traffic. Thus partial sharing is same as complete sharing with $n_{data} = 0$.

Those three different policies have different effects on the delay of the packets which is described in the following sections.

1) Model and Analysis: The analytical model developed in this section is based on the following assumptions.

a) We model the downlink of a single cell in cluster of seven cells. We assume the rates at which subscribers move in and out of the cell are the same and hence there is a fixed number of users in the cell. We consider the cell to consists of one TRX (ie. 8 channels).

b) We assume that the SNR and BER ratio are ideal and hence there are no re-transmissions at the MAC/RLC layer.

c) The arrival of voice calls are modelled as Poisson processes with negatively exponential distribution with mean arrival rate λ_v and mean service rate ν. The offered load due to voice call is represented by λ_v/ν Erlang. The mean call duration is assumed to be 180 seconds.

d) The data source is modelled by Markov Modulated Poisson Process (MMPP) with two states - a high state and a low state. The mean duration in the high and low states are $1/r_1$ and $1/r_0$, respectively. In the high state packets are generated with a mean rate of λ_1 pkts/sec and corresponding rate in the low state is λ_0 pkts/sec. The MMPP is specified by the infinitesimal generator matrix $\mathbf{Q}_{MMPP}$ and rate matrix $\mathbf{\Lambda}$ shown in equation (1) and (2), respectively. We introduce two other parameters to describe the MMPP data source, namely, the average arrival rate of data packets,λ_{avg} and the degree of burstiness denoted by B. These are defined in equation(3) and (4).

$$\mathbf{Q}_{MMPP} = \begin{pmatrix} -\mathbf{r}_0 & \mathbf{r}_0 \\ \mathbf{r}_1 & -\mathbf{r}_1 \end{pmatrix} \quad (1)$$

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_0 & \mathbf{0} \\ \mathbf{0} & \lambda_1 \end{pmatrix} \quad (2)$$

$$\lambda_{avg} = \frac{r_1}{r_0 + r_1}\lambda_0 + \frac{r_0}{r_0 + r_1}\lambda_1 \quad (3)$$

$$\mathrm{B} = \frac{\lambda_1}{\lambda_{avg}} \quad (4)$$

e) The service rate for data packet is μ. We assume that data sources are TCP sources with each TCP segment 512 bytes long. We use coding scheme CS-4 with data rate of 22.8 kbps. Thus, the mean data call duration is $512 * 8/22.8ms$. Both the arrival and service rates are negatively exponentially distributed. The mean data traffic load is λ_{avg}/μ Erlang.

f) We assume an infinite buffer for data packets and an Erlang loss model for voice calls.

2) Queueing Model For Partial and Complete Sharing Algorithms : For complete and partial sharing schemes, the system can be described by a three dimension state transition diagram as shown in Figure 3. Each state is represented by a vector (*c*, *i*, *a*) where*c* is the number of data packets in the system (including the ones in service and the ones in the queue buffer), *i* is the number of channels that can be used by data packets, *a* is the MMPP state of the source. Given total N channels, N-i channels are being used by voice calls in partial and complete sharing schemes. In partial sharing $i \geq n_{data}$, n_{data} is the number of reserved data channels. As long as the total Erlang (voice and data) is less than the cell channel capacity, the system will reach to a steady state where each state's probability is expressed as p(c,i,a). To obtain the state probabilities we need to solve

$$\underline{p}\mathbf{Q} = \underline{0} \quad (5)$$

where Q is the generator of the underlying Markov chain of the system. Furthermore,

$$\sum_{i=0}^{\infty} \underline{z}_i \underline{1} = 1 \quad (6)$$

where $\underline{z}_i$ ($i \geq 0$) is a vector of the steady probability of level *i* in the state transition diagram and $\underline{z}_i = (p_{(i,n,0)}, p_{(i,n,1)}, p_{(i,n+1,0)}, p_{(i,n+1,1)}, \cdots, p_{(i,N,0)}, p_{(i,N,1)})$

$$\underline{p} = (\underline{z}_0, \underline{z}_1, \underline{z}_2, \cdots, \underline{z}_i, \underline{z}_{i+1}, \underline{z}_{i+2}, \cdots) \quad (7)$$

$\underline{p}$ is defined in Equation (7).

3) Analysis For Partial and Complete Sharing Schemes: The system can be treated as a CTMC (Continuous Time Markov Chain) process. The CTMC describing this queueing model is a QBD(quasi-birth-death model) process [8]. According to Neut's theory [6] and the algorithm in [7], the steady state probability of CTMC can be solved by exploiting the matrix-geometric properties.

$$\mathbf{Q} = \begin{pmatrix} \mathbf{B}_0 & \mathbf{A}_0 & \mathbf{0} & \cdots \\ \mathbf{B}_1 & \mathbf{A}_1 & \mathbf{A}_0 & \cdots \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_1 & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad (8)$$

The $\mathbf{Q}$ matrix can be expressed as shown above. According to the queueing model in §III-A.2, we can get each element of matrix $\mathbf{B_0}, \mathbf{B_1}, \mathbf{A_0}, \mathbf{A_1}, \mathbf{A_2}$ each of which has dimension of 2(N-n+1) x 2(N-n+1).

$$\underline{z}_i = \underline{z}_{i-1}\mathbf{R} \quad (9)$$

$$\underline{z}_i = \underline{z}_0\mathbf{R}^{i-1}, \quad i = 1, 2, \cdots. \quad (10)$$

From [6], we can get equation (9)(10). Using the global balance equation (5), we get the following results;

$$\begin{aligned} \underline{p}\mathbf{Q} = \underline{0} \quad &\Rightarrow \quad \left[\cdots, \underline{z}_i, \underline{z}_{i+1}, \underline{z}_{i+2}, \cdots\right]\mathbf{Q} = \mathbf{0} \\ &\Rightarrow \quad \underline{z}_i\mathbf{A}_0 + \underline{z}_{i+1}\mathbf{A}_1 + \underline{z}_{i+2}\mathbf{A}_2 = \underline{0} \end{aligned} \quad (11)$$

$$\underline{z}_1\left(\mathbf{R}^2\mathbf{A}_2 + \mathbf{R}^1\mathbf{A}_1 + \mathbf{R}^0\mathbf{A}_0\right) = \underline{0}. \quad (12)$$

Equation (12) can only be true when either $\underline{z}_1 = \underline{0}$, or when the quadratic equation within parentheses equals $\underline{0}$. Since $\underline{z}_1 \neq \underline{0}$, the latter must be the case, and the matrix $\mathbf{R}$ thus follows from the following matrix quadratic equation:

$$\mathbf{R}^2\mathbf{A}_2 + \mathbf{R}^1\mathbf{A}_1 + \mathbf{R}^0\mathbf{A}_0 = \mathbf{0}. \quad (13)$$

From Equation (13) we can derive

$$\mathbf{R} = -(\mathbf{A}_0 + \mathbf{R}^2\mathbf{A}_2)\mathbf{A}_1^{-1} \quad (14)$$

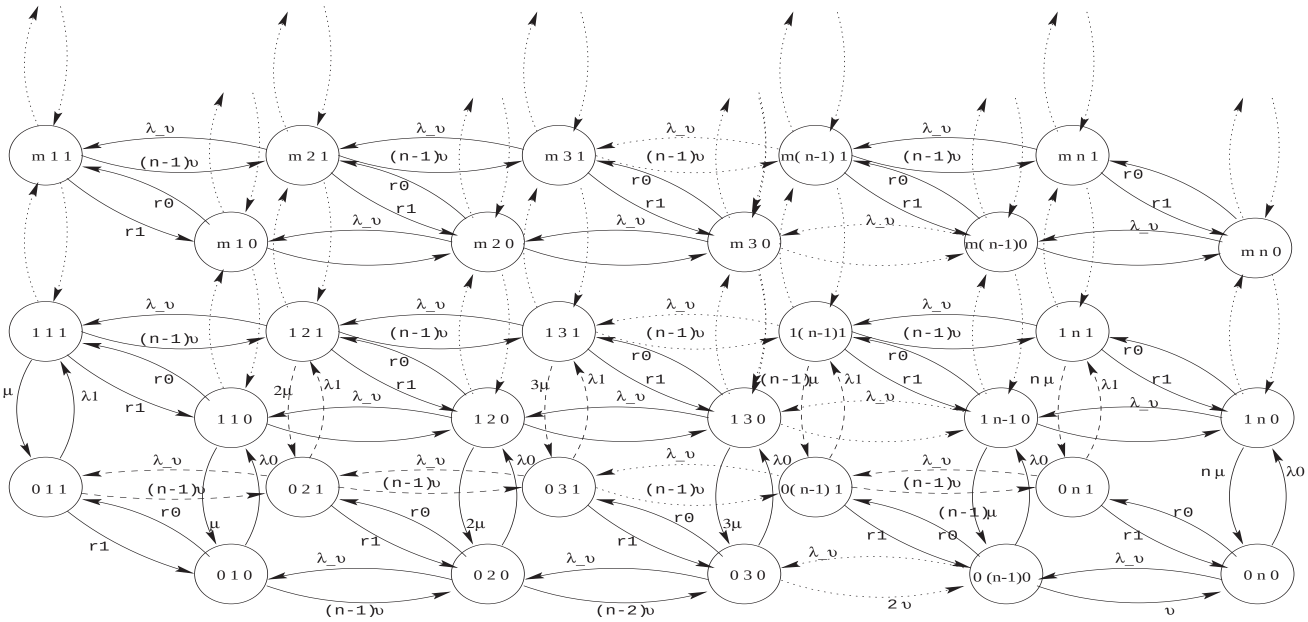

Fig. 3. state transit diagram

Now, taking as a first guess $\mathbf{R}(0) = \mathbf{0}$, we can get the next guess $\mathbf{R}(1) = \mathbf{A}_0\mathbf{A}_1^{-1}$. We obtain successively obtain better approximations of $\mathbf{R}$ as follows:

$$\mathbf{R}(k+1) \quad = \quad -\left(\mathbf{A}_0 + \mathbf{R}^2(k)\,\mathbf{A}_2\right)\mathbf{A}_1^{-1}, \quad k = 1, 2, \cdots \tag{15}$$

The iteration stops when $\|\mathbf{R}(k+1) - \mathbf{R}(k)\| < \epsilon$. In this study we choose $\epsilon = 1E{-}32$. From the global balance equation we can derive

$$(\underline{z}_0, \underline{z}_1)\begin{pmatrix}\mathbf{B}_0\\ \mathbf{B}_1\end{pmatrix} \quad = \quad \underline{0}. \tag{16}$$

Also, since equation (16) is not full rank, the normalization equation (5) has to be used to arrive at a unique solution:

$$\begin{aligned}\sum_{i=0}^{\infty}\underline{z}_i\underline{1} \quad &= \quad \underline{z}_0\underline{1} + \sum_{i=1}^{\infty}\underline{z}_i\underline{1} = \underline{z}_0\underline{1} + \underline{z}_1\left(\sum_{i=0}^{\infty}\mathbf{R}^i\right)\underline{1}\\ &= \quad \underline{z}_0\underline{1} + \underline{z}_1\left(\mathbf{I} - \mathbf{R}\right)^{-1}\underline{1} = 1. \end{aligned}\tag{17}$$

Once the matrix $\mathbf{R}$ and the boundary vector $\underline{z_0}$ and $\underline{z_1}$ are known, we can obtain the average number of data packets in the queue and in service. $E[N]$ denotes the mean number of packets in the system (in queue + in service) and is given by

$$\begin{aligned}E[N] \quad &= \quad \sum_{i=1}^{\infty} i\underline{z}_i\underline{1}\\ &= \quad \underline{z}_1\left(\mathbf{I} - \mathbf{R}\right)^{-2}\underline{1}\end{aligned}\tag{18}$$

The average number of data packets in the queue is $E[N_q] = E[N] - \lambda_{avg}/\mu$. The mean packet delay is $E[W] = E[N_q]/\lambda_{avg}$. We can also get the blocking rate for the voice calls according to Erlang's B formula which is denoted as $B(m, \lambda_v/\nu) = B(m, \rho)$ and is given by:

$$p_m = B(m, \rho) \quad = \quad \frac{\rho^m/m!}{\sum_{j=0}^{m}\rho^j/j!} \tag{19}$$

For simplicity, we normalized the voice and data traffic by the cell channel capacity which is given by $(\lambda_v/\nu + \lambda_{avg}/\mu)/N$.

4) Analysis For Fixed Sharing Algorithm: For fixed sharing algorithm, the total channels in a cell are splitting into two parts; One part is allocated for voice traffic and the other part for data traffic exclusively. $N = n_{voice} + n_{data}$, the voice calls are modeled as a $M|M|n_{voice}|n_{voice}$ queue. Blocking rate for voice calls is still be represented by Equation (19). The data calls now are modeled as a $MMPP|M|n_{data}$ queue. The infinitesimal generator **Q** has the form of a Quasi Birth and Death (QBD) process with complex boundary which is expressed in equation (20).

$$\mathbf{Q} \quad = \quad \left(\begin{array}{cc|ccc}\mathbf{B}_{00} & \mathbf{B}_{01} & \mathbf{0} & \mathbf{0} & \cdots\\ \mathbf{B}_{10} & \mathbf{A}_1 & \mathbf{A}_0 & \mathbf{0} & \cdots\\ \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \cdots\\ \mathbf{0} & \mathbf{0} & \mathbf{A}_2 & \mathbf{A}_1 & \cdots\\ \vdots & \vdots & \vdots & \vdots & \ddots\end{array}\right) \tag{20}$$

$$\underline{z}_i \quad = \quad \underline{z}_{02}\mathbf{R}^i, \quad i \quad = 1, 2, \cdots. \tag{21}$$

$$(\underline{z}_{01}, \underline{z}_{02}, \underline{z}_1)\begin{pmatrix}\mathbf{B}_{01}\\ \mathbf{A}_1\\ \mathbf{A}_2\end{pmatrix} \quad = \quad \underline{0}. \tag{22}$$

$$(\underline{z}_{01}, \underline{z}_{02}, \underline{z}_1) \begin{pmatrix} \mathbf{B}_{01} & \mathbf{B}_{01} \\ \mathbf{B}_{10} & \mathbf{A}_1 + \mathbf{R}\mathbf{A}_2 \end{pmatrix} = (\underline{0}, \underline{0}) . \quad (23)$$

We can solve the above equations to get :

$$\begin{aligned} E\left[N_q\right] &= \sum_{i=0}^{\infty} i \underline{z}_{02} \mathbf{R}^i \underline{1} = \underline{z}_{02} \left(\sum_{i=0}^{\infty} i \mathbf{R}^i \right) \underline{1} \\ &= \underline{z}_{02} \left(\mathbf{R} \left(\mathbf{I} - \mathbf{R} \right)^{-2} \right) \underline{1}. \quad (24) \end{aligned}$$

We can also obtain the mean packet delay $E\left[W\right] = E\left[N_q\right] / \lambda_{avg}$.

5) Results and Discussions: We first study the partial sharing scheme with $n_{data} = 1$. For the voice call, the offered load is fixed at 26.2% of the total cell channel capacity (i.e., 8 channels). With fixed voice load, we increase the data packets arrival rate. As shown in Figure 4, the simulation results almost exactly match the analytical results.

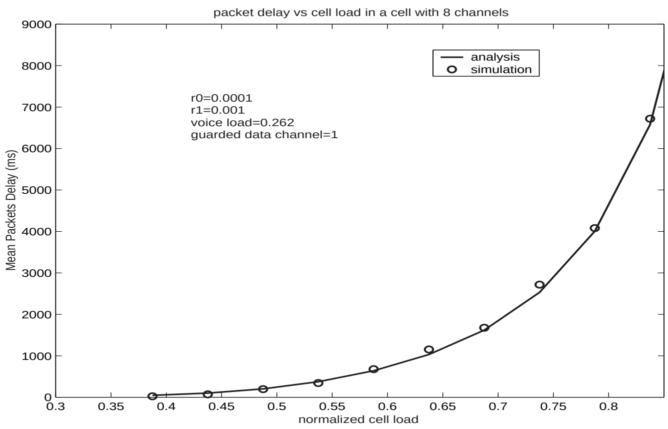

Fig. 4. Mean packet delay as a function of the normalized cell load ($N = 8,\ n = 1,\ \lambda_v/\nu/N = 0.262$).

From Figure 4 we can see that when the offered voice call and data traffic load reaches 65% of the cell channel capacity, the mean packet delay increases rapidly. This is quite critical for real time multimedia application, such as real audio or video which require a small delay. The analytical tool developed in this paper can be useful to determine when to expand the cell capacity.

Second, we investigate the packet mean delay as a function of the burstiness, $\mathbf{B}$, of the data source. In this analysis, we compare he different data source patterns. One is a Poisson process in which $\mathbf{B} = 1$, the other two are MMPPs, each with different $\mathbf{B}$. We fixed the voice calls load while increasing the data stream packets arrival rate. Figure 5 shows that with different burstiness, the packet delays are different. The larger the burstiness, the more delay.

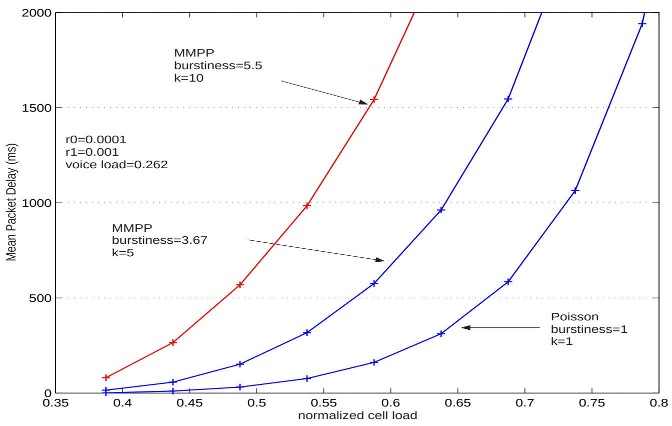

Fig. 5. The Mean packet delay as a function of the burstiness of data packet and arrival rate λ_{avg} $(N = 8,\ n = 1,\ B = 1.0,\ 3.67,\ 5.5 \quad \lambda_v/\nu/N = 0.262)$.

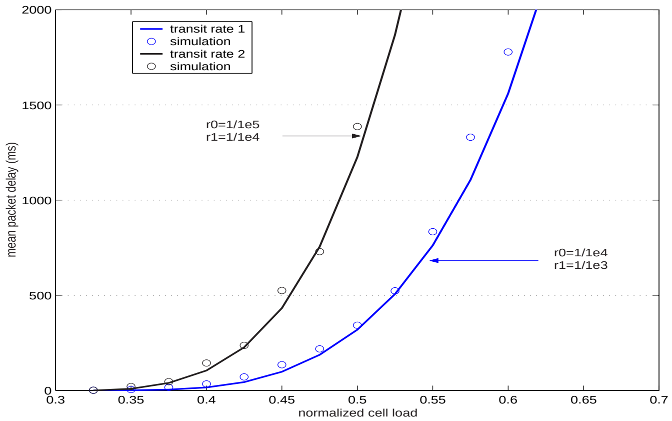

Fig. 6. Mean packet delay as a function of MMPP transition rate and arrival rate $\lambda_{avg}(N = 8,\ n = 1, \quad \lambda_v/\nu\ N = 0.262)$.

We also considered the impact of the transition rates of the MMPP source on the packet delay. We fix the voice call load, keep the same burstiness, $\mathbf{B}$, but with different transition rate for the MMPP data source. The results show that the duration time $(1/r_0, 1/r_1)$ for the MMPP has an effect on the mean packet delay. When the duration time in each state is small, the packets can be more “evenly” distributed in the time domain, thereby reducing the delay. This effect is shown in Figure 6.

In Figure 7, we fix both the load of voice call and data traffic and change the minimum number of reserved channels for data traffic. For ease of comparison, we normalize the delay and blocking rate with the sum of all the corresponding results for different number of reserved channels. The results show that if there are more reserved channels for data traffic, the mean packet delay is reduced. But the tradeoff is obvious; with more reserved channels, the voice call blocking rate is higher. Given a requirement of voice call blocking rate and the average packet delay, it is then possible to determine if a specific partition will satisfy both the requirements. If no partition can satisfy the requirements, then channel capacity must be increased. This provides the motivation for the wireless network provider to consider the algorithms studied in this paper to guarantee QoS

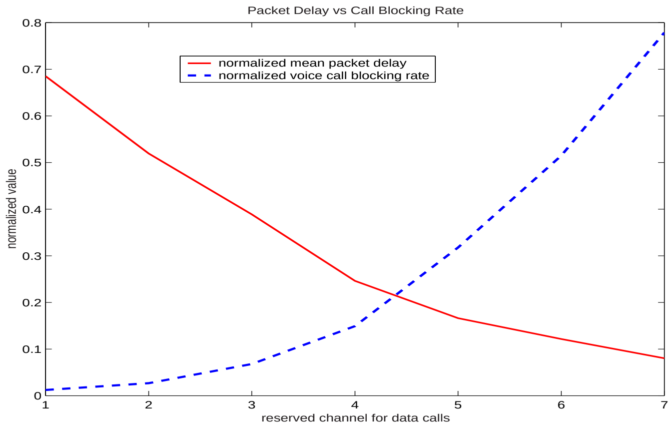

Fig. 7. Effect of minimum number of channels reserved for data on the mean packet delay ($N = 8,\ n = 1, 2, 3 \ldots 7, \quad \lambda_v/\nu/N = 0.3, \quad \lambda_{avg}/\mu/N = 0.4$).

both to the voice call and data traffic.

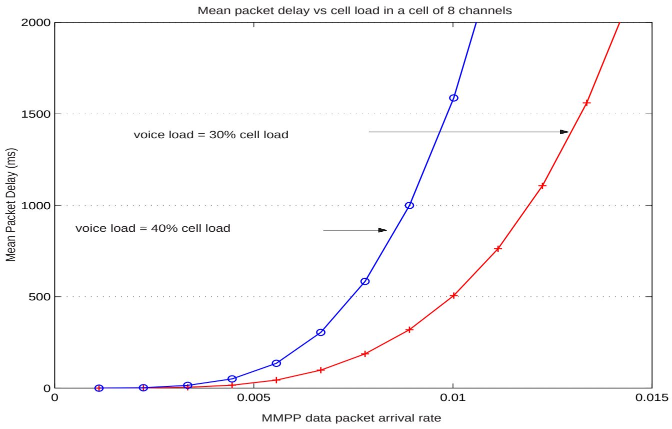

Fig. 8. Effect of voice call load on the mean packet delay(1) ($N = 8,\ n = 1,\ \lambda_v/\nu/N = 0.3,\ 0.4$).

In Figure 8, we investigate the effect of voice call load on the mean delay of data packet. We choose different GSM voice loads in the cell, 30% and 40% respectively, reserved the same number of channels for data while increasing the data packet arrival rate. The results show that voice load has significant effect on the packet delay. This can be explained by that voice call has higher priority over data packet. Even with the same reserved channels for data, more voice calls imply less number of channels available for data traffic. In the partial or complete sharing scheme, the voice call load has serious effect on the QoS of the data traffic.

In Figure 9, at the same cell load of 55%, packet delay is almost double in voice load 40% case than the case when the voice load is 30%.

Finally, we compare the partial sharing scheme with the fixed-sharing and complete sharing schemes. For the fixed sharing scheme we consider the case when total cell channels are split into two equal parts; one part with 4 channels exclusively for voice calls and the other part also with 4 channels exclusively for data traffic. As mentioned, earlier, in complete sharing scheme $n_{data} = 0$. Under the same parameters of the data traffic, voice traffic, we can see that total-sharing policy has the voice block rate $= 0.3\%$, 0.9% for partial-sharing and 14.5% for fixed-sharing. As we can see, even though the fixed-sharing policy has the minimum mean packet delay, the tradeoff is higher voice blocking rate which is not acceptable in real network. While partial-sharing and complete-sharing policy both have a small blocking rate(less than 1%), but the former one has pretty good QoS performance than the late one.

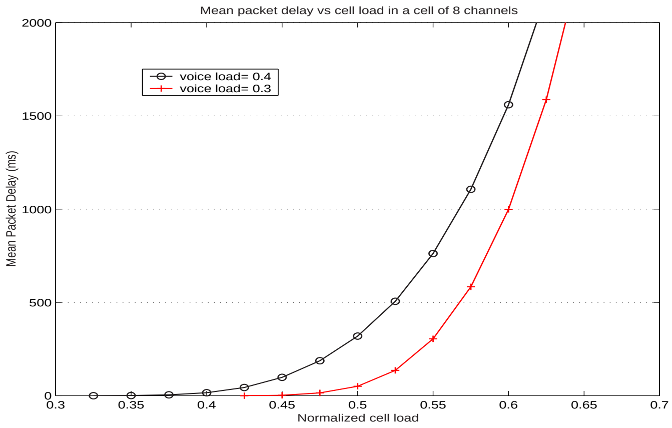

Fig. 9. Effect of voice call load on the mean packet delay(2) ($N = 8,\ n = 1,\ \lambda_v/\nu/N = 0.3,\ 0.4$).

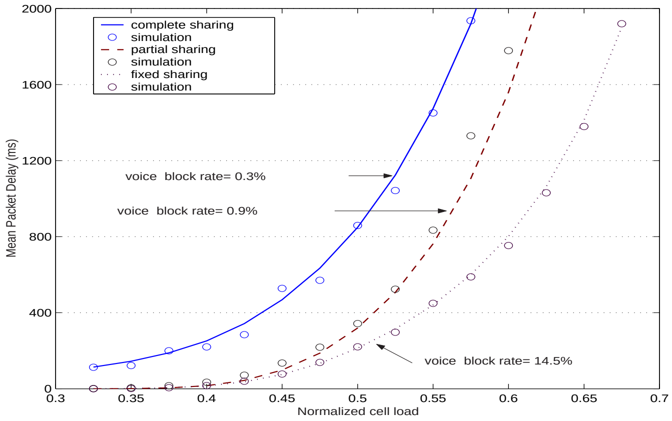

Fig. 10. Comparison of mean packet delay for different channel sharing policies). ($N = 8,\ \lambda_v/\nu/N = 0.3$).

By above analysis, we estimated the data packet delay at MAC/RLC layer of the GSM/GPRS system. The data source pattern (degree of Burstiness, transition rate etc) and the voice call load will both have the effect on the delay of the data packet. If we also consider SNR rartio and channel interferences in the radio air interface, we probably get a worse BER curve related to these issues. For Internet application, such as TCP/IP, the error-prone wireless environment will cause packets to re-transmit to get the data recovered from error. This will further increase the latency of the packets. These delays will significantly impact the end-user's browsing experience. Mobile operators must tackle this problem or face further questions about the usability of the mobile Internet. The combination of high latency and variable latency (jitter) leads to the slowdown or possible failure of Internet applications and unsuccessful delivery of web content.

B. GGSN nodes distribution Algorithm

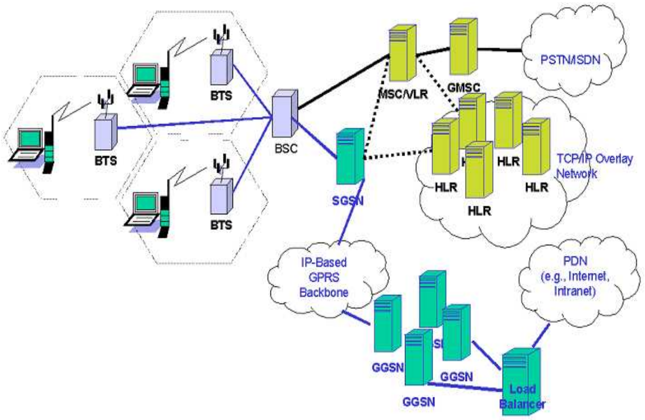

Fig. 11. GSM/GPRS system with load balancing at GGSN nodes.

1) Architecture of GGSN Nodes with balancing scheme : The network configuration is shown in Figure 11. As mentioned before, the GGSN is the ingress point into the GPRS network. All packet-switched traffic to and from mobiles in the provider's network is routed through the GGSN. State information for all users is maintained in the GGSN, as well as context information for all open connections. So while the GGSN cannot be considered an "active agent", it clearly requires more processing power than a normal router, and therefore can become a bottleneck under high load. As the number of GPRS users is expected to increase dramatically, a replicated GGSN architecture can be implemented to transparently provide scalability and fault-tolerance to the GPRS network. There are many issues to be resolved in creating a replicated GGSN architecture. First, the GGSN maintains state information for mobiles in the network. This information will now be replicated, and will need to be kept coherent across all copies. Second, load balancing is critical to the stability and efficiency of the replicated system. Finally, the replicated architecture should be transparent to outside networks (the Internet).

In a replicated GGSN architecture, incoming packets destine for a GPRS mobile will first reach the load balancer. The load balancer will have a hash table of open sessions. It will hash the destination address and port number, and determine the route the packet should take (i.e. which GGSN is currently serving the user). If there is no hash entry, one will be created, and a new GGSN will be assigned to the user, using a load-balancing scheme. This scheme can be simple as simple maintaining a counter of how many open connections each GGSN has an choosing the one with the least number of open connections. More complex methods of load balancing, such as having each GGSN report their load or throughput can also be implemented.

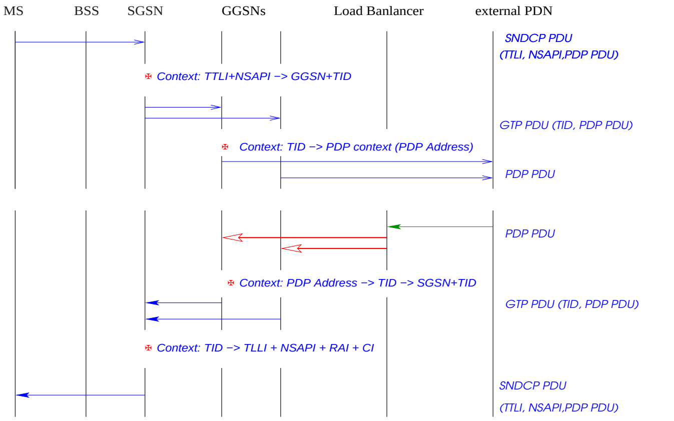
MS
BSS
SGSN
GGSNs
Load Banlancer
external PDN
SNDCP PDU (TTLI, NSAPI,PDP PDU)
Context: TTLI+NSAPI −> GGSN+TID
GTP PDU (TID, PDP PDU)
Context: TID −> PDP context (PDP Address)
PDP PDU
PDP PDU
Context: PDP Address −> TID −> SGSN+TID
GTP PDU (TID, PDP PDU)
Context: TID −> TLLI + NSAPI + RAI + CI
SNDCP PDU
(TTLI, NSAPI,PDP PDU)

Fig. 12. Data traffic routing in GPRS network of replicated GGSNs

Upon a location update, the SGSN will simply multicast the location update message to all the GGSNs. An SGSN can then query any of the GGSNs to determine user state information. The message flow is shown in Figure 12. Note that, since the load balancer takes the place of the "original" GGSN, outside networks such as the Internet will have no knowledge of the internal configuration of the GGSN cluster. GGSNs can be added or removed as the need arises. Only list of GGSNs at the load balancer is affected when a server is added or removed.

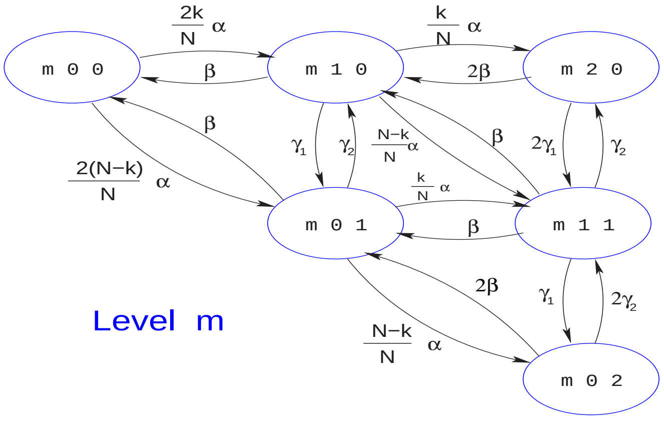

Fig. 13. State transition diagrams within one level

2) Model and Analysis: In this paper, We study the performance of data flow for the downlink that is from the sources to GGSN, from GGSN to mobile user. The overall queueing model is consisting two sources, and four GGSN nodes (N=4). The source can be modeled as a two states MMPP (Markov Modulated Poisson Process). When in ON state, packets are generated according to a Poisson process with rate λ_1 packets/sec, there are no packets generated in OFF state, so the $\lambda_2 = 0$ here in the model, which is IPP (Interrupted Poisson Processes), a special case of MMPP. The source turns on with

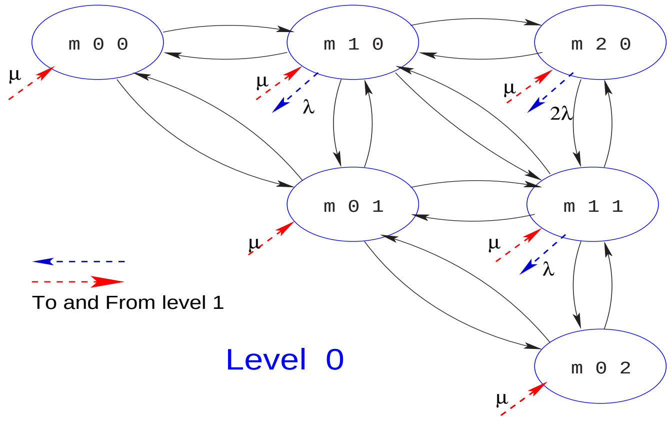

Fig. 14. State diagram of level 0 to level 1

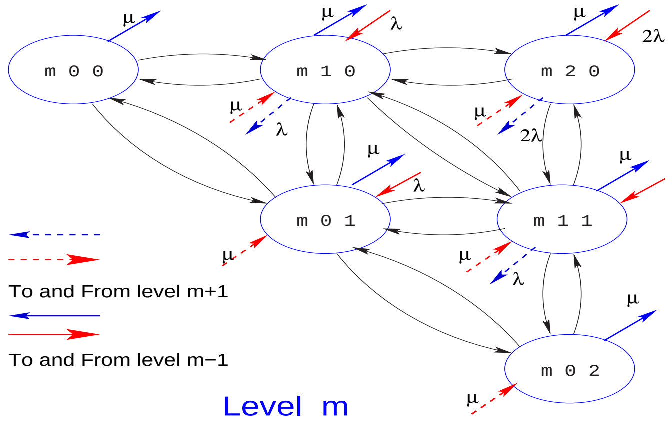

Fig. 15. State diagram of level m to another level (m+1 or m-1)

rate α and turns off with rate β. Each GGSN has the capacity of μ packets/sec with an infinite queue. We study one of the tagged GGSN whose level state diagram is shown in Figure 13, 14 and 15. The state is defined as (m, b, a), m is the packet in the tagged GGSN node, b is the number of On-source transmitting packets to the tagged GGSN node, a is the number of source On-source transmitting packets to the rest of the GGSNs. The transition rates of γ_1 and γ_2 are defined as:

$$\gamma_1 = \frac{N-k}{N}\lambda, \qquad \gamma_2 = \frac{k}{N}\lambda. \tag{25}$$

We represent three different distribution algorithms by the value of k. If $k = 4$, it is the non-distribution policy, all the packets are routed to the tagged GGSN node. If $k = 1$, it is the complete-distribution policy in which all the source packets are evenly distributed among all the GGSN nodes. While $4 > k > 1$, this the partial-distribution policy in which source packets are unevenly distributed among the GGSN nodes.

The infinitesimal generator Q can be drawn from the state diagram. It has the form like in Equation (20). The steady-state vector of Markov chain $\underline{p} = (\underline{z}_0, \underline{z}_1, \underline{z}_2, \cdots, \underline{z}_i, \underline{z}_{i+1}, \underline{z}_{i+2}, \cdots)$ can be found by solving the equations (5) (6). Following the same math analysis from the former section, the mean packet delay can be obtained as:

$$\begin{aligned} E\left[W\right] &= \frac{N}{k\lambda_{avg}}\sum_{i=0}^{\infty} i\underline{z}_{02}\mathbf{R}^i\underline{1} = \frac{N}{k\lambda_{avg}}\underline{z}_{02}\left(\sum_{i=0}^{\infty} i\mathbf{R}^i\right)\underline{1} \\ &= \frac{N}{k\lambda_{avg}}\underline{z}_{02}\left(\mathbf{R}\left(\mathbf{I}-\mathbf{R}\right)^{-2}\right)\underline{1}. \end{aligned} \tag{26}$$

3) Results and Discussions: By adopting three different distribution policies, We compare the mean packet delay by increasing the incoming data sources arrival rate. The analytical result is shown in Figure 16. As we can see the complete-distribution has the best result in terms of mean packet delay while the non-distribution scheme has the worst performance. It verifies that replicated GGSN architecture with appropriate load balancing can improve the QoS in terms of reducing the delay at the GGSN nodes.

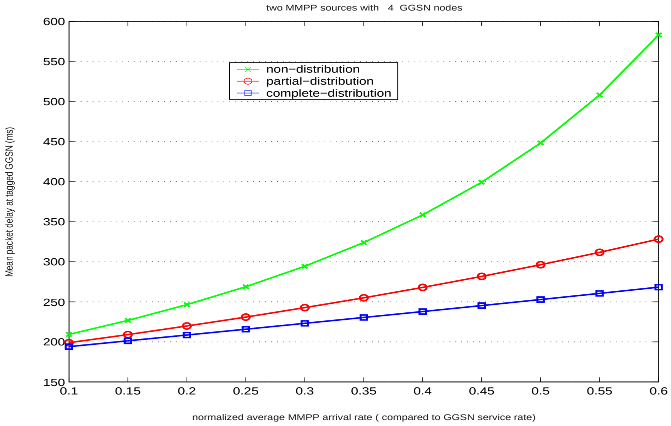

Fig. 16. Mean packet delay with three different load balancing policies

IV. Related Literature

In [3], the authors developed a model for two different types of circuit-switched calls, fresh calls and handoff calls. It uses the static channel allocation algorithm. Only analytical results are presented. In [4], the authors analyze the GPRS network for circuit-switched calls and packet switched GPRS data sessions. The model is based on a five dimensional limited state space which is quite complicated. The author provide the both the analytical and simulation results on some QoS parameters, such as blocking rate and throughput but not the mean data packet delay. In [5], the author studies several partition models based on a loss GPRS system. The results are simulation based and does address the mean packet delay. All these authors do not give an analysis on the GGSN nodes's delay effects which caused by unbalanced data source loads.

In our paper, We use a different analytical technique to model the channel allocation schemes. Using computer simulation We verified our model to be correct. Besides, our primary focus is on the QoS of the MAC/RLC layer and of GSM/GPRS network , specifically on the mean packet delay. We also investigated the effects of data source characteristics and voice calls load effects on the mean packet delay. Besides, the mean packet delay at GGSN nodes has been studied by analysis which is not discussed by the above papers.

V. Conclusions

In this paper, we study the delay sources across the GSM/GPRS network. First, We use a continuous time Markov chain model to analyze the GSM/GPRS MAC/RLC layer function and compare the performance of three different channel allocation algorithms. Using simulation, We verified our analytical model. Our results show that both the data source characteristics such as degree of burstiness and MMPP state transition rate, and the voice call load impact the mean packet delay. The degree of impact is dependent on the channel allocation algorithm. The technique developed in this paper provides a straight-forward way for the wireless carrier to determine when to expand the cell channel capacity or adjust the channel allocation algorithm. Besides, by adopting different load sharing policies, we can improve to a some degree the QoS for the data traffic. Second, We provide a analytical model to study the mean packet delay occurs at the GGSN nodes. By choosing different load balancing policies, the results show the QoS can be improved. Further study is underway to develop new algorithms that can support different QoS to different classes of data stream.

References

[1] Cellular Online, http://www.cellular.co.za/main.htm,

[2] Roger Kalden, Ingo Meirick and Michael Meyer, “ Wireless Internet Access Based on GPRS ”, IEEE Personal Communications , April 2000

[3] M.A. Farahani, M. Guizani “ Markov Modulated Poisson Process Model For Hand-off Calls In Cellular Systems ”, Wireless Communications and Networking Confernce, 2000. WCNC. 2000 IEEE , 2000 Page(s): 1113 -1118 vol.3

[4] Lindemann, C.; Thummler, A. “ Performance analysis of the General Packet Radio Service ”, Distributed Computing Systems, 2001. 21st International Conference on. , 2001 Page(s): 673 -680

[5] M. Ermel, K. Begain, T.Muller, J. Schuler, M. Schweigel “ Analytical Comparision of Different GPRS Introduction Strategies ”, ACM MSWiM 2000 Boston MA USA

[6] Marcel F. Neuts , *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach* , Johns Hopkins University Press, 1981

[7] William J. Stewart , *Introduction to the Numerical Solution of Markov Chains* , Princeton University Press, 1994

[8] Ng Chee Hock , *Queueing Modelling Fundamentals* , John Wiley & Sons , 1996

[9] ETSI TS 100 908 , “ Digital cellular telecommunications system (phase 2+); Multiplexing and multiple access on the radio path ”, V8.10.0 (2001-08)

[10] ETSI TS 100 573 , “ Digital cellular telecommunications system (phase 2+); Physical Layer on the Radio Path (General Description) ”, V4.1.0 (2001-11)

[11] ETSI TS 101 350 , “ Digital cellular telecommunications system (phase 2+); General Packet Radio Service (GPRS); Overall description of the GPRS radio interface; stage 2 ”, V4.3.0 (2002-02)

[12] ETSI TS 101 344 , “ Digital cellular telecommunications system (phase 2+); General Packet Radio Service (GPRS); Service description; stage 2 ”, V7.6.0 (2001-03)