Delay performance analysis and simulation of real-time applications in GPRS networks

Tim Irnich, Dimitri Eges

Communication Networks, Aachen University (RWTH)
Kopernikusstr. 16, D–52074 Aachen, Germany
Phone: +49 241 80 27928, E-mail: {tim|edv}@comnets.rwth-aachen.de

Keywords: General Packet Radio Service, Real-time Applications, Fluid-flow-model

Abstract

Motivated by earlier results [1] we discuss the feasibility of calculating the delay performance of *real-time* applications in packet-based mobile radio networks like the GSM General Packet Radio Service (GPRS) by means of the Fluid-flow modelling (FFM) approach. We determine the required source parameters that represent the mean offered traffic in typical GPRS load scenarios by stochastic simulation. Using these source parameters for definition of a source model that is applicable to Fluid-flow analysis, we determine the mean IP delay for downlink IP GPRS traffic. The FFM analysis' results are compared with results of the GPRS emulation tool GPRSim. Our results show that the FFM's basic assumption of fluid data (i.e., infinitely small packets) and some details of radio resource management and scheduling in GPRS networks lead to deviations between FFM and simulation results, although the traffic source behavior is well captured by the FFM's source model. Based on the observation that under certain conditions the GPRS resource allocation strategy allows the GPRS system to be separated into subsystems, we propose to examine the analytical methodology of multi-queue cyclic service systems for GPRS analysis.

INTRODUCTION

In the context of the evolution towards 3rd Generation (3G) mobile radio networks, packet-switched data services like the *General Packet Radio Service* (GPRS) have recently been introduced into GSM and IS-136 systems worldwide, rising the question of how a mobile radio network integrating circuit- and packet-switched services can be dimensioned properly.

Unfortunately the Erlang theory, which has successfully been applied for dimensioning of circuit-switched commuication systems, does not lead to accurate results for packet-switching systems, because with the change of the switching paradigm the focus in radio network dimensioning is shifted. While dimensioning of circuit switching systems focuses on whole user calls, while *Quality of Service* in packet switching networks has to be defined on packet level.

Although analytical models so far have not been able to significantly contribute in this context in a *quantitative* way, they can be particularly helpful for identifying the crucial influences to traffic performance [2]. Thus, even in case of significant deviations between analytical results and the real system analytical approaches contribute to the process of understanding the complicated interrelations in packet-based mobile radio networks.

We investigate the applicability of *Fluid-flow modelling* (FFM) for dimensioning of packet-switched networks. The FFM is based on a concept developed by L. Kosten [3], and was extended by Anick et al. [4]. For our analysis, we use the notation of Fiedler and Voos [5]. The main result of the FFM approach is the equilibrium buffer size's *Cumulative Distribution Function (CDF)*. Using Little's Law, we derive the mean waiting time of an IP datagram. We do not directly compare the FFM waiting time to the waiting time of data packets in our GPRS simulation system and real GPRS systems, but introduce the *FFM Sojourn Time* as an analytical quantity which is equivalent to the IP datagram delay evaluated in our simulation system.

It is well known that Internet traffic sources can be categorized into elastic and inelastic source types. Elastic traffic sources apply end-to-end error and flow control (usually based on TCP [6]) and thus adapt their offered traffic to the performance perceived in the immediate past. Inelastic source behavior is usually caused by the absence of end-to-end acknowledgements, which is usually the case for real-time application that do not benefit from retransmissions in case of packet loss. Motivated by earlier research [1], where the elastic property of Web browsing traffic was identified to be a dominant reason for deviations between simulation results and Fluid-flow analysis, we focus on traffic types with inherently inelastic behavior. For comparison with simulation results we use the GPRS simulation tool *GPRSim*. Comparison with measurement results from live GPRS networks have shown that this approach leads to very accurate results.

GENERAL PACKET RADIO SERVICE

GPRS has been standardized by the ETSI as part of the GSM *Phase 2+* development to introduce a packet-switched extension to the GSM radio interface, which is essentially a circuit-switched technology. A detailed description of GPRS can be found in [7]; we will limit our description to aspects that are particularly relevant in the context of our work.

Packet switching means that radio resources are used only when users are actually sending or receiving data. Through multiplexing of several logical connections on one or more GSM physical channels, GPRS reaches a flexible use of channel capacity. The basic transmission unit of a GPRS Packet Data Channel (PDCH) is a *radio block* that requires four time slots in four consecutive GSM *Time Division Multiple Access* (TDMA) frames [8]. The length of a TDMA frame is 4.615 ms, and the length of a GPRS Multiframe is 18.46 ms. Every 13th burst is not used for transmission.

Four different *Coding Schemes* (CS) are defined, providing data rates from 9.05 kbit/s to 21.4 kbit/s per PDCH, see Table 1. Since in GPRS the access of all eight slots of a TDMA frame is foreseen, data rates up to 160 kbit/s can be achieved. For a single mobile station its *Multislot Capability* (MSC) defines how many slots within the TDMA frame may be used.

Table 1: GPRS coding schemes (CS)

	CS 1	CS 2	CS 3	CS 4
PDCH data rate [kbit/s]	9.05	13.4	15.6	21.4
MAC block size [bit]	181	268	312	428
RLC block payload [byte]	19	29	35	49

SIMULATION ENVIRONMENT

The (E)GPRS Simulator GPRSim [9] comprises models of Mobile Station (MS), Base Station (BS), Serving GPRS Support Node (SGSN), and Gateway GPRS Support Node (GGSN) (see Fig. 1).

Different from usual approaches to establish a simulator, where abstractions of functions and protocols are used, the GPRSim is based on the detailed implementation of the GSM and (E)GPRS protocols. TCP has been implemented based on the description in [6] including slow start and congestion avoidance algorithms. Radio resource sharing between circuit-switched GSM services (e.g., voice telephony or GSM *Circuit Switched Data* CSD) is modelled as well.

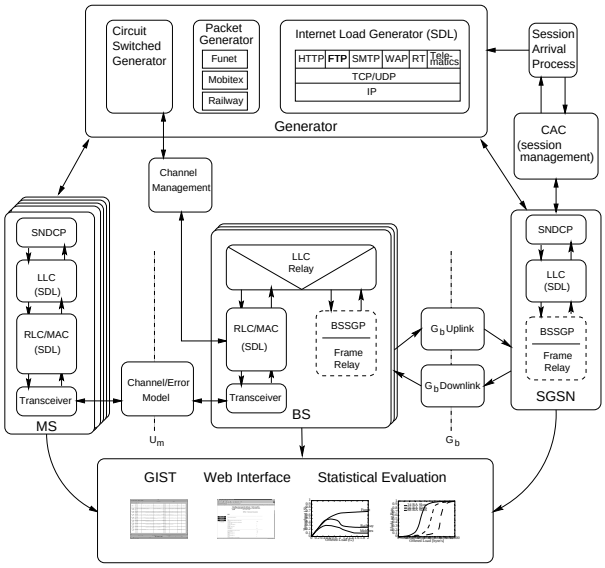


Figure 1: Overview GPRSim structure

Video Streaming Traffic Model

The Video Streaming traffic model used within this work is based on three different video sequence traces. Each video represents a particular group of videos with different intensities of motion.

- `Video 1`: very low motion intensity, characteristic visual telephony or inactive video-conferencing. Resulting mean traffic offer: 10.9 kbit/s.
- `Video 2`: periods with rather high motion and periods of low motion intensity. Represents many kinds of vivid or active video-conferences. Resulting mean traffic offer: 26.7 kbit/s.
- `Video 3`: permanently high motion intensity of both, the actor and the background, characteristic for sport events or movies. Resulting mean traffic offer: 31.7 kbit/s.

We applied a skip factor of 2, leading to a frame rate of 12.5 frames/s. Thus, the IP packet inter-arrival time is 0.08s. From each trace, a packet size table is generated, which specifies the size of the successive packets.

The duration of video sessions is modeled by a negative exponential distribution with an average value of 60s. Between the sessions there is a negative exponential inter-arrival time of 12s and 60s, respectively. For further details on this traffic model please refer to [10].

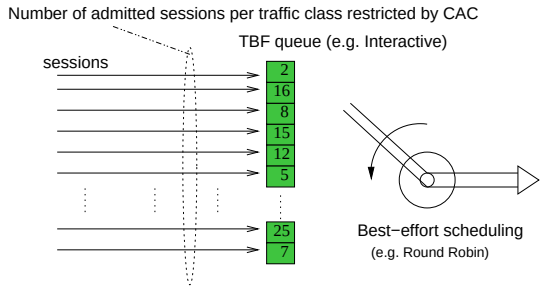


Figure 2: TBF scheduling at MAC layer

Scheduling of downlink traffic at the BSS

The algorithm applied for scheduling of downlink data at the BS is a dominant influence factor to the downlink traffic performance. Scheduling of radio resources at the BSS can be subdivided into two steps: the selection of the next Temporary Block Flow (TBF) and scheduling of the next RLC block of the selected TBF (see Fig.). When multiple TBF's are multiplexed on a given number of PDCH's the TBF scheduler behavior determines how the available capacity is distributed among the concurrent TBF's.

The GPRSim implements a round-robin-like cyclic service scheme. At the time the decision on which TBF is allowed to use a specific PDCH in the next Radio Block Period (RBP), the scheduler starts with the first TBF listed in the queue and checks if it has been allocated to the regarded PDCH. If not, the scheduler moves on to the following TBF. In case the TBF is able to use the regarded PDCH, the related RLC entity is polled for data until the data passd to the scheduler reaches the predefined service quantum or there are no more radio blocks to transmit. Afterwards the following TBF is served. In the scope of this paper the service quantum is set to 10 radio blocks.

In the next step, the RLC entity which has been polled for data by the MAC scheduler checks if there are any data blocks available in the transmit buffer, and selects the next block for transfer.

FLUID-FLOW SOURCE MODELS

In the scope of the FFM, arrival of data is compared to water falling into a reservoir (the network element's buffer memory), which depletes at a constant rate C. The traffic sources alternate between an ON state and an OFF state, the sojourn times in these states are exponentially distributed. While in the ON state, a source transmits data at a constant rate h. A single traffic source is modelled by a *Markov Modulated Rate Process* (MMRP), called *Interrupted Rate Process* (IRP). Multiple equal subscribers are modeled by superposition of multiple IRPs, called NIRP.

Single Traffic Source (IRP)

Each IRP is controlled by a two-state *Markov chain* (MC) with states Λ_0 and Λ_1. In state Λ_0 the source transmits packets at the rate $r_0 = 0$ and in state Λ_1 the transmission rate is $r_1 = h$. The transition rates between the ON state Λ_1 and the OFF state Λ_0 are λ and μ.

Each IRP needs the a set of three parameters to be described completely. The Activity Factor α denotes the fraction of time the source is active. Mean burst length, denoted by EN_B is the mean amount of data generated during an ON period. The arrival rate of this data during the ON period is denoted by h. Thus, λ and μ can be derived to $\mu = \frac{h}{EN_B}$ and $\lambda = \mu \cdot \frac{\alpha}{1-\alpha}$.

The mean transmission rate of the source is $M = \alpha \cdot h$. Thus, the system load is $\rho = \frac{M}{C}$ where C denotes the system capacity. The system load has to be less than one to achieve a stable solution.

Superimposing multiple equal sources (NIRP)

Superimposing N equal IRPs, a so-called N *Interrupted Rate Process* (NIRP) can be defined [4, 5]. A NIRP can be described by a one-dimensional *Markov chain* (MC) (see Figure 3). The state variable of this MC is the number of active IRPs, the total number of states is $N + 1$.

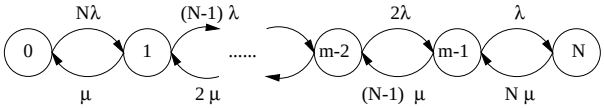


Figure 3: NIRP state transition diagram

For each state Λ_q the data transmission rate is $r_q = q \cdot h$ and the whole NIRP's mean transmission rate is $M = N \cdot \alpha\, h$.

State space subdivision

Based on the arrival rate r_q and the capacity C, the state space of a NIRP can be subdivided into:

- $\Lambda^u = \{\Lambda_q \in \Lambda \text{ with } r_q < C\}$: underload states
- $\Lambda^e = \{\Lambda_q \in \Lambda \text{ with } r_q = C\}$: uniform load states
- $\Lambda^o = \{\Lambda_q \in \Lambda \text{ with } r_q > C\}$: overload states

In an underload state, the buffer content depletes at the rate $C - r_q$, in an overload state the buffer content rises with rate $r_q - C$ and in a uniform load state the buffer content remains constant. It is easy to see that in a NIRP that does not comprise equal or overload states ($N \cdot h < C$) the buffer content is always zero.

FFM SOJOURN TIME

The comparison of Fluid-flow analysis and simulation cannot be based on the waiting time of IP datagrams, because on the one hand the assumption of fluid data does not allow to define the waiting time of a single packet, and on the other hand the cyclic service discipline and the fragmentation of IP packets by the RLC/MAC layer makes the waiting time of an IP packet difficult to define and even more difficult to evaluate. Thus, we derive an analytical quantity that can be directly compared to the IP datagram delay at the GPRS radio interface. Our concept is to add the *minimum* time an IP datagram transmission over the radio interface requires to the mean waiting time in order to establish a lower bound for the total IP delay. We call the resulting quantity the *FFM Sojourn Time*.

The duration of a *Radio Block Period* (RBP) is $t_{\mathrm{RBP}} = 18.46\,\mathrm{ms}$, and the duration of a TDMA frame is $t_{\mathrm{TDMA}} = 4.615\,\mathrm{ms}$. Division of $\overline{N_{\mathrm{IP}}}$ by $N_{\mathrm{RLC,CS}}$, the number of bytes contained in one RLC block (according to the applied coding scheme, see Tab. 1), leads to the mean number of RLC blocks needed for transmission of an IP datagram. The number of RLC blocks per IP datagram additionally has to be divided by the MSC, denoted by N_{msc}, leading to the number of RBPs, R, the transmission of an IP datagram takes:

$$R = \left\lceil \left\lceil \frac{\overline{N_{\mathrm{IP}}}}{N_{\mathrm{RLC,CS}}} \right\rceil \cdot \frac{1}{N_{\mathrm{msc}}} \right\rceil, \qquad (1)$$

$\lceil \cdot \rceil$ denotes rounding to the next greater integer value. Due to the fact that the scheduling at the MS is shifted by two TDMA frame durations, and after every 3 Radio Block periods there is one idle frame, we have to add $(\lfloor R/3 \rfloor + 2)$ TDMA frame durations:

$$t_s = t_w + \left(R \cdot t_{\mathrm{RBP}} + \left(\left\lfloor \frac{R}{3} \right\rfloor + 2 \right) \cdot t_{\mathrm{TDMA}} \right) \qquad (2)$$

For example for CS-2 we have 29 byte payload per RLC block (see Table 1)), this leads to a mean number of 12 RLC blocks per IP datagram for a mean IP datagram size of 340 bytes. Furthermore assuming MSC 4 leads to 3 RBPs required for transmission ($R = 3$). Thus, the lower border for mean transmission duration in this case is 69.225 ms.

RESULTS

First we evaluate the parameters needed for definition of the FFM source models. We discuss the relations between the scenario definition and the resulting IRP parameters and show that the source behavior of the GPRSim's video traffic generator actually is inelastic. Afterwards we discuss the FFM sojourn time delivered by Fluid-flow analysis and compare with simulation results.

IRP parameters

In case of WWW an activity period is defined as the time required for the download of a whole web page, while in case of video-streaming we define an activity period as the whole load generator session, because packets with constant inter-arrival times are generated throughout the whole session.

Fig. 4 shows the behavior of the source parameters under rising system load. The simulation scenario these figures have been generated from is defined by 8 available PDCHs, mobile stations using CS-2 and MSC 4+4, and an error-free radio channel.

The WWW traffic source parameters clearly exhibit the elastic property that is inherent for TCP-based applications. As soon as multiple WWW traffic sources have to share the available capacity at the radio interface, the TCP flow control reduces the data rate h. As the amount of data that is generated by the WWW application model is constant, the activity factor rises proportionally, because it takes longer to transmit the data generated by the source.

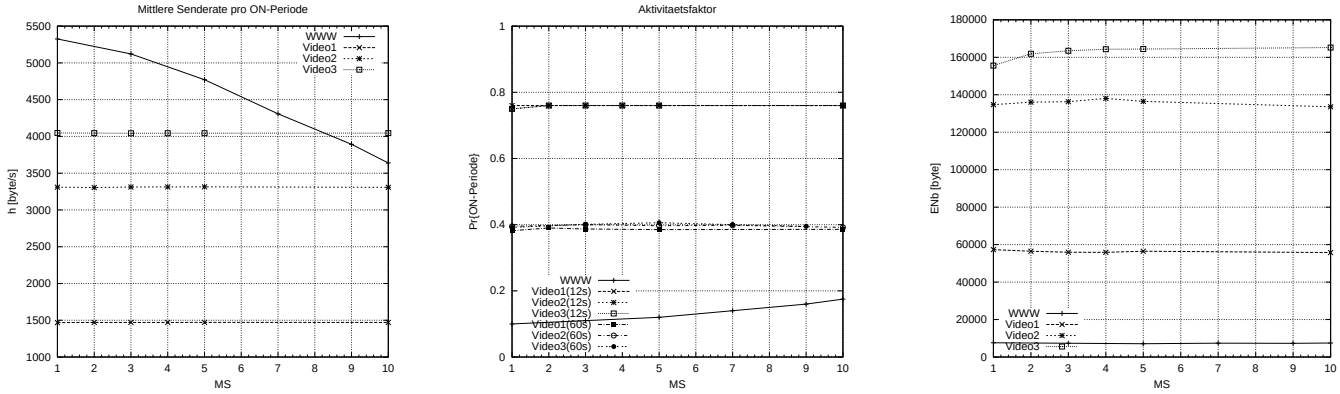
The video source parameters are obviously independent on the available bandwidth per source. Taking a closer look at the results for the source's activity factor in Fig. 4(b), we see that the activity factor is higher for shorter session inter-arrival times. As the accuracy of the FFM results benefits from lower activity factors, we select the results for 60s session inter-arrival time for further analysis. The FFM analysis results presented in the next section are based on the source parameters listed in Tab. 2.

Table 2: IRP parameters, 60s session inter-arrival time

	Video1	Video2	Video3
EN_b [Byte]	56266	139567	168153
h [byte/s]	1465	3311	4040
α	0.38	0.4	0.39

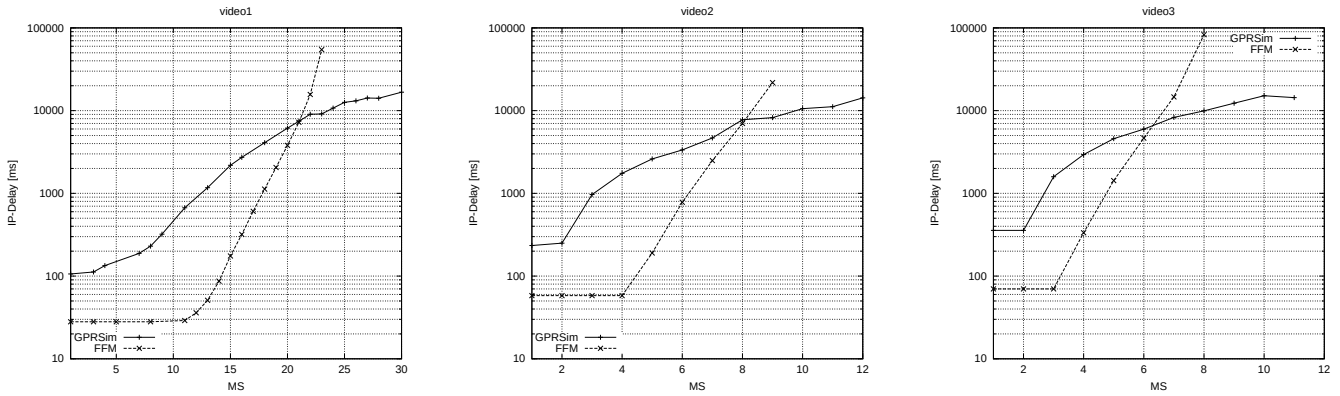
IP Delay Performance Evaluation

In Fig. 5 we compare the behavior of the IP delay performance predicted by FFM analysis with simulation results. The source parameters EN_b, α and h are set according to Tab. 2. Each point in the FFM curves corresponds to the mean IP delay of a NIRP with N sources, while for the



(a) On period arrival rate h (b) Activity factor α (c) Mean burst size EN_B

Figure 4: ON/OFF source parameters from simulation, video-streaming (inelastic) and WWW (elastic)



(a) Video 1 (low bit rate) (b) Video 2 (medium bit rate) (c) Video 3 (high bit rate)

Figure 5: IP delay (simulation) vs. FFM sojourn time (analysis) for all available video types

simulation results each point results from a single simulation with N MS. The scenario parameters for simulation are 8 fixed PDCHs, CS-2, MSC 4+4, service quantum 10, error-free radio channel and 60s session inter-arrival time.

Common to the results for all three video sequences is that the FFM results curves have two regions of different behavior. As long as the corresponding NIRP only consists of an underload range, the waiting time delivered by Fluid-flow analysis is zero. Thus, the IP delay only consists of the IP packet transmission duration (see Eq. (2)) in this range. As soon as the NIRP's state space comprises overload states, the FFM analysis yields an exponential increase (this region is called the *overload range*), until the system becomes unstable (*unstable range*). For video 1 the last stable NIRP is at $N = 23$ MS, for video 2 at $N = 9$ and for video 3 at $N = 8$.

Comparing the FFM values with the simulation results, we note that in both regions the simulation shows not only quantitative deviations, but also a different behavior. In the underload range, where the IP delay should be constant, the simulation shows that the delay is only constant below three MS and increases as soon as three sources are present. Taking into account that the MSC was set to 4+4, this behavior can be explained. As long as only two MS are present each MS receives 4 PDCHs exclusively (see Fig. 6a), because the GPRSim's resource allocation strategy has the goal to minimize the number of active TBFs per PDCH. Once a third MS joins the system, the

first 4 PDCHs have to be shared between MS 1 and 3 (see Fig. 6b), which in general can lead to the situation that IP packets from one MS have to wait until data from the other MS is transferred. Additionally provided that the service quantum applied for the cyclic service discipline (see Sec.) is high enough to ensure exhaustive service for each IP packet, an IP packet that arrives during the service of an IP packet from another MS has to wait until this packet is completely transferred. Please note that this effect is independent on the mean data rate (e.g., 10.9 kbit/s for video 1 compared to $4 \cdot 13.4$ kbit/s available capacity on MAC level for 4 PDCHs and CS-2; see Tab. 1.

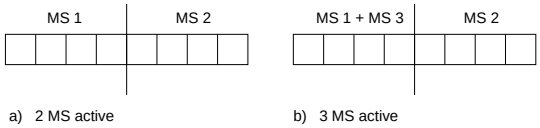


Figure 6: Channel allocation for 2 and 3 MS simultaneously active

Based on these considerations a wider range of constant IP delays (i.e., beyond 3 MS) can be expected for MSC's lower than 4 (e.g., MSC 2+2 or 1+1), which is intended to be subject of future research.

In the overload and unstable range the simulation results show an asymptotic behavior while the FFM results predict exponential growth for this range. The reason for the asymptotic behavior is that every time there is backlogged traffic at session end, these packets are discarded. This has an increasing influence on the delay performance in the simulation system when the system load reaches the stability border and is also the reason why the simulation system is able to continue operation beyond the stability border into the unstable range.

An interesting observation is that for the combination of MSC 4+4 mobiles with the number of available PDCHs being an integer multiple of this figure, e.g., 8, the resulting distribution of concurrent users allows to separate the set of available channels into subsystems (see Fig. 6b). One subsystem is formed by the first four PDCHs, while the other one is formed by the second four PDCHs. Every additional MS that is admitted to the system would be completely allocated to one of these subsystems by the GPRS radio resource management. Inside each of these subsystems the active mobile stations are served according to the cyclic service policy described before. For such systems there are several analytical approaches to performance evaluation known from the literature. Examination of these models is supposed to be a promising field of future research.

CONCLUSION

We have shown that the deviations between simulation and FFM results can be explained by the fact that some general assumptions of FFM do not hold for the GPRSim's combination of cyclic service discipline and the applied channel allocation strategy.

Due to the multislot assignment applied by GPRS's radio resource management, the available system capacity can be divided into subsets. Inside these subsets cyclic service of active TBFs is applied. The service quantum, that was set to 10 in our simulations leads to exhaustive service of single IP packets. Therefore IP packets that arrive during the service duration of an IP packet from a different MS have to wait for completion of the current IP packet's service. In case of two MS sharing the same set of PDCHs, although the capacity provided by these channels in fact is sufficient to serve the data of both MS without delay, waiting times can occur in the simulated system. This situation could be prevented by adapting the number of timeslots allocated to one MS according to the data arrival rate at this MS.

For exhaustive service cyclic service obviously is inherently less efficient and leads to higher delays than the fluid assumption of the FFM, which represents the limiting case of cyclic scheduling with infinitely small service quantum.

References

[1] T. Irnich and P. Stuckmann, "Fluid-flow Modelling of Internet Traffic in GSM/GPRS Networks," in *Proc. Int. Symp. on Perf. Eval. of Computer and Telecomm. Systems (SPECTS)*, pp. 625–632, 2002.

[2] T. Irnich and P. Stuckmann, "Analytical Performance Evaluation of Internet Access over GPRS And its Comparison with Simulation Results," in *Proc. Int. Symp. on Personal, Indoor and Mobile Comm. (PIMRC)*, (Lisbon, Portugal), 2002.

[3] L. Kosten, "Stochasitc theory of a multi entry buffer," in *Delft Progress Report*, vol. 1 of *F*, pp. 10–18, 1974.

[4] D. Anick, D. Mitra, and M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *The Bell System Technical Journal*, vol. 61, pp. 1871–1894, October 1982.

[5] M. Fiedler and H. Voss, *Fluid-flow Modelling of ATM-Multiplexers (in German)*. Herbert Utz Verlag, Munich, Germany, 1997. ISBN 3-89675-251-0.

[6] R. Stevens, *TCP/IP Illustrated*, vol. 1. Addison-Wesley, 1996.

[7] P. Stuckmann, *The GSM Evolution - Mobile Packet Data Services*. John Wiley & Sons, 2002.

[8] B. Walke, *Mobile Radio Networks*. John Wiley & Sons, 2 ed., Nov 2001.

[9] P. Stuckmann, "Simulation environment GPRSim user manual," tech. rep., `http://www.comnets.rwth-aachen.de/~pst`.

[10] C. Hoymann and P. Stuckmann, "On the Feasibility of Video Streaming Applications over GPRS/EGPRS," in *Proc. IEEE Global Telecomm. Conf. (Globecom)*, (Taipei, Taiwan, R.O.C.), November 2002.

Fluid-flow Analysis

In the following we shortly summarize the formulae that were used to obtain our FFM results, for a complete derivation please refer to [5]. We regard the general case of a Fluid-flow multiplexer with buffer size K.

In case N equal ON/OFF sources are attached to the multiplexer, the arrival process is represented by a single NIRP.

Starting with the equilibrium probability of state Λ_q, and the buffer of maximum capacity K being filled with x bytes of data waiting for transfer

$$F_q(x,K) = Pr\{X \leq x \text{ and } \Lambda_q\} \quad ; x \leq K$$

a differential equation system can be set up, leading to an eigenvalue problem. Calculation of the equilibrium buffer size's CDF requires calculation of the eigenvalues z_q, the sum of each eigenvector's components and a set of coefficients to fit the solution to boundary conditions.

The eigenvalues z_q can be derived to:

$$\begin{aligned} z_q &= \frac{1}{2\left(C-qh\right)\left(C-\left(N-q\right)h\right)} \\ &\cdot\Big(NC\left(\lambda+\mu\right)-N^2h\lambda+2\left(N-q\right)qh\left(\lambda-\mu\right) \\ &\quad+\left(2q-N\right)\cdot\Big(C^2\left(\lambda+\mu\right)^2+\left(Nh\lambda\right)^2 \\ &\quad-2NhC\lambda\left(\lambda+\mu\right)+4\left(N-q\right)qh^2\lambda\mu\Big)^{\frac{1}{2}}\Big) \end{aligned}$$

Special cases that have to be treated separately are:

1. $C = qh$:
   The eigenvalue is undetermined.
2. $C = (N-q)h$:
   The eigenvalue is given by:
   $$z_q = \frac{2(N-q)q(\lambda+\mu)^2}{N(2C-Nh)(\lambda+\mu)-(N-2q)^2(\lambda-\mu)}$$

The eigenvectors are calculated using the inverse eigenvalue problem (see [5]). The sum of the eigenvector components can be obtained by evaluating the generating function of the eigenvector components.

$$\begin{aligned} \Phi_q(x,z_q) &= \sum_{i_1=\Gamma_1}^{0}\Bigg(\binom{\Gamma_1}{i_1}\left(-res_{q,1}\left(z_q\right)\right)^{\Gamma_1-i_1} \\ &\quad\cdot\sum_{i_1=\Gamma_1}^{0}\binom{\Gamma_2}{i_2}\left(-res_{q,2}\left(z_q\right)\right)^{\Gamma_2-i_2}x^{i_1+i_2}\Bigg) \end{aligned} \tag{3}$$

and

$$res_{q,1/2}(z)=\tfrac{1}{2\lambda}\Big(\left(\lambda-\mu-zh\right)\pm\sqrt{\left(\mu-\lambda+zh\right)^2+4\mu\lambda}\Big). \tag{4}$$

In the scope of this article the available buffer memory is assumed to be unlimited, because there are no restrictions to the queue length in the simulation system as well. Accordingly, the coefficients that fit the solution to its boundary conditions are

$$a_q(\infty) = -\alpha^N\,\Phi_q(1,z_q)\prod_{\substack{s\in\Lambda^0\\ s\neq q}}\frac{z_s}{z_s-z_q}. \tag{5}$$

The CDF F(x) of the equilibrium buffer size is

$$F(x) = \sum_{\forall q} a_q(\infty)\,\Phi_q(1,z_q)\,e^{z_q x}.$$

Taking into account that the eigenvalue z_0 equals zero, the eigenvector $\vec{\varphi}_0$ equals the vector of the NIRP's steady state probabilities (see [4]), the sum of which is one and regarding the system's boundary conditions, which allow to derive $a_q(\infty)$ for some special cases (see [5] and [1] for a more detailed description), we finally receive for the buffer content's CDF and CCDF (denoted by $G(x)$)

$$F(x) = 1+\sum_{q\in\Lambda^o} a_q(\infty)e^{z_q x} \Rightarrow G(x) = -\sum_{q\in\Lambda^o} a_q(\infty)e^{z_q x}. \tag{6}$$

Integration of $G(x)$ finally delivers the mean equilibrium buffer size:

$$E[x] = \int_0^\infty G(x)dx = \int_0^\infty \left(1-F\left(x\right)\right)dx = \sum_{q\in\Lambda^o}\frac{a_q(\infty)}{z_q} \tag{7}$$

The mean waiting time can be obtained by Little's Law:

$$t_w = \frac{E[x]}{M} = \frac{1}{M}\sum_{q\in\Lambda^o}\frac{a_q(\infty)}{z_q} \tag{8}$$