Quality of Service in GPRS/EDGE Mobile Radio Networks

R. Müllner, C.F. Ball, K. Ivanov, F. Treml and G. Spring
SIEMENS AG, I&C Mobile Networks
Munich, Germany / Vienna, Austria
robert.muellner@siemens.com

Abstract — **Quality of Service (QoS) is the basis for supporting 3G like services in 2G networks as well as for offering seamless service quality in 2G-3G networks, perceived especially by Dual-RAT subscribers. In this paper an advanced QoS strategy comprising 3GPP QoS parameters along with operator's specific weighting factors is used to define the appropriate QoS priority of each service type and user profile. A deterministic up- and downgrading strategy as well as admission control is applied to ensure both a minimum service level defined by the operator for low-priority services and a full bandwidth for delay time sensitive services and premium users. The simulation results show that especially in highly loaded and even overloaded GSM/GPRS/EDGE networks the introduction of QoS provides significant benefits for the end user and offers means to increase the service revenues according to the charging policy adopted by the network operator. The introduction of an appropriate QoS strategy is the prerequisite for an overlay deployment strategy of GSM/EDGE and UMTS. The efficiency of QoS management is crucial for all mobile network technologies and has high impact on fulfilling the demands of mobile subscribers with continuously growing expectations.**

Keywords – Quality of Service; QoS; GPRS; EDGE; radio resource management; scheduler; guaranteed bit rate; target data rate; service sustenance level; performance

I. INTRODUCTION

Applications for packet data services in GSM/GPRS/EDGE mobile networks set specific requirements on throughput, delay and response time. The network is expected to support these applications seamlessly and simultaneously to utilize the available frequency spectrum in a most economic way. In this paper the primary objective is to reveal the benefit of an enhanced QoS strategy in different traffic load scenarios and to monitor the performance of the BSS. The proposed QoS strategy comprises Temporary Block Flow (TBF) ranking according to the QoS priorities as well as deterministic up- and downgrading procedures. The strategy can be easily implemented by modifying existing Radio Resource Management (RRM) and packet data scheduler. This facilitates the system's operation and provides an essential benefit to the network operator: once having introduced QoS management, the operator does not have to care about temporary fluctuations of the traffic load with respect to QoS. With the deterministic up- and downgrading strategy an automatic configuration of the system is provided and only services of higher QoS priority obtain the right of downgrading services of lower QoS priority. Previous work on these topics can be found in [1, 2].

After introduction in Section I, the QoS model is presented in Section II and the simulation assumptions are described in Section III. The simulation results are presented in Section IV. Section V concludes the paper.

II. COMPOUND QUALITY OF SERVICE MODEL

The simulation model comprises 3GPP standard QoS parameters [3] and operator's specific weighting factors. Typical 3GPP QoS parameters are e.g. 'traffic class', 'traffic handling priority', 'allocation/retention priority' and 'guaranteed bit rate' for real-time services. Additionally, in the proposed QoS model operator's specific weighting factors have been introduced, allowing specific priorities for service and/or subscriber groups. Each incoming service request is evaluated according to the QoS parameters assigned to this particular service. Then the overall priority of each service is determined by combining the standard QoS attributes and the operator's specific weighting factors. In case of a new service request the relative priority of this new service with respect to all currently served services in the cell is calculated to determine the appropriate resource allocation. Fig. 1 shows an example for GPRS/EDGE packet data services in a cell with four data calls (TBF1 through TBF4) of different QoS priority. The new service request (TBF5) is ranked in a cell priority list according to its relative QoS priority. In case of resource bottlenecks, a new high priority service may seize parts of the resources assigned to lower priority services, i.e. TBF5 is allowed to "steal" resources from TBF4 and TBF2 but not from TBF3 and TBF1.

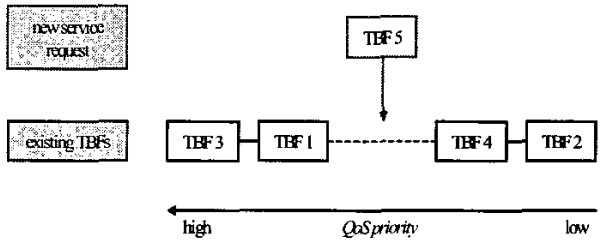

Figure 1. QoS priority of currently served TBFs and a new incoming service request

For real-time services the amount of resources is derived from the standard QoS parameter 'guaranteed bit rate' and is not affected by the QoS priority. For interactive and background

0-7803-8255-2/04/$20.00 ©2004 IEEE.

services as well as for standard and premium subscribers a target throughput (TTP) defined by the operator has been introduced. In order to meet the TTP the following equation is used to define the number of required resources for the k-th packet data call TBF_k:

$$TTP_k = \sum_{i=1}^{max\_num\_ts_k} p_{ik} \cdot CS\_Throughput_k \cdot (1 - BLER_k)$$

with $p_{ik} = 0$, if TBF_k is not allocated on time slot (TS) i, and $num\{p_{ik} \neq 0\} \leq max\_num\_ts_k$. Furthermore p_{ik} is called the share factor of TBF_k on TS i and $max\_num\_ts_k$ is the maximum number of TS allocated for TBF_k due to the multislot class of the mobile. $CS\_Throughput_k$ is the maximum user data rate provided by the selected coding scheme *CS* for TBF_k. $BLER_k$ is the retransmission rate for the selected coding scheme CS for TBF_k depending on radio conditions. The task of the RRM is to optimize the TTP for each service by appropriate time slot allocation and optimum adjustment of the share factors. Furthermore the scheduler multiplexes the services on every time slot according to the assigned share factors. In the following we define the TTP as 'guaranteed bit rate' in case of real-time services and as 'target throughput rate' for non real-time services (WAP, HTTP, e-mail and FTP). In this paper the TTP assumed is 128 kbps for real-time (streaming) services and 32 kbps for non real-time services.

A new service request is admitted by admission control if sufficient resources are available. For packet data real-time services the guaranteed bit rate has to be provided. Less delay sensitive interactive and background services obtain admission if at least a tolerable quality level in the following termed as service sustenance level can be obtained. The service sustenance level is defined as the minimum ratio of assigned throughput and TTP. In these simulations the service sustenance level has been set to 0.1, i.e. at least 10% of the resources necessary to fulfill the TTP requirements have been granted to each service. A new service request is queued for a certain period of time if the necessary resources cannot be provided. For this purpose the QoS model defines a queuing-reject-timer. After expiry of this timer, which has been set to 5 s, a new service request is rejected in this cell, if the required resources cannot be offered. The applied admission control mechanism attempts to allocate and maintain services of high QoS priority at their TTP even in case of high traffic load.

III. Simulation Assumptions

To study the effect of enhanced QoS strategies on the GPRS/EDGE performance, system level simulations have been performed in a typical 4/12 frequency re-use scenario in a 2/2/2 configuration (i.e. three sectors per site with two TRXs per sector). An urban network deployment with slow moving subscribers (3 km/h) has been assumed. A site-to-site distance of 2 km, a propagation index of 3.8, a slow fading standard deviation of 6 dB, a handover margin of 5 dB and a slow fading correlation distance of 20 m have been used.

A mixed voice and GPRS/EDGE data traffic has been considered. The offered voice load has been dimensioned for a voice-only cell at 1% hard blocking (2 signaling channels per cell assumed resulting in 14 traffic time slots) and kept fixed at 7.35 Erlang (worst case) for all simulated scenarios. All 14 traffic channels have been configured as shared channels in a common pool [4]. A mixture of 50% GPRS and 50% EDGE four TS capable MS has been assumed for non real-time services. For streaming services only EDGE terminals are used at a constant data rate of 128 kbps. The performance of the QoS model has been studied at low data load (50% of the user population requests a packet data service in addition to the voice service), medium data load (75%) and high data load (100%). Note that the cell is already fully loaded by voice traffic. Hence the packet data traffic is added on top of the voice traffic such that the performance of the proposed QoS strategy has been proven under these worst case conditions. The type of packet data non-real time service (WAP, HTTP, FTP and e-mail, [5]) is random and equally distributed, while the arrival rate of streaming service requests is 50% of any one of the non-real time services. Streaming services are provided in the RLC non-acknowledged mode while all other data services are provided in the RLC acknowledged mode [6, 7]. TCP/IP effects have not been taken into account.

IV. Simulation Results

The performance of QoS in GPRS/EDGE networks has been studied in different packet data load scenarios (refer to Section III) and the resulting benefit has been evaluated. The benefit from QoS is moderate at low data load. However, even in low data load scenarios, the Target Datarate Factor (TDF) distribution shows a clear separation of the services according to their QoS priority. The TDF is defined as the ratio of assigned to requested resources (TTP). With increasing data traffic in the network the introduction of QoS becomes more and more important in order to keep subscribers satisfied.

Fig. 2 shows the status of the TDF for all allocated TBFs in a particular cell during a period of 15 seconds. The colors indicate the degree to which the TTP of 128 kbps for streaming services and TTP of 32 kbps for WAP, HTTP, e-mail and FTP is met: green means TDF $\in$ [0.9, 1.1[, yellow TDF $\in$ [0.5, 0.9[, red TDF $\in$ [0.1, 0.5[. A TDF equal to 1 implies that the resources necessary to meet the TTP have been fully assigned, while TDF > 1 means that more than the requested resources have been allocated. Note that the amount of required resources depends on the coding scheme selected by dynamic link adaptation [8] and the BLER corresponding to the actual radio conditions. At the beginning of each packet call both an initial coding scheme (MCS-7 for EDGE and CS-3 for GPRS) and a statistical value for the BLER depending on the prevailing radio conditions are used for the initial resource assignment. During the packet call the number of assigned resources is adjusted on each link adaptation step.

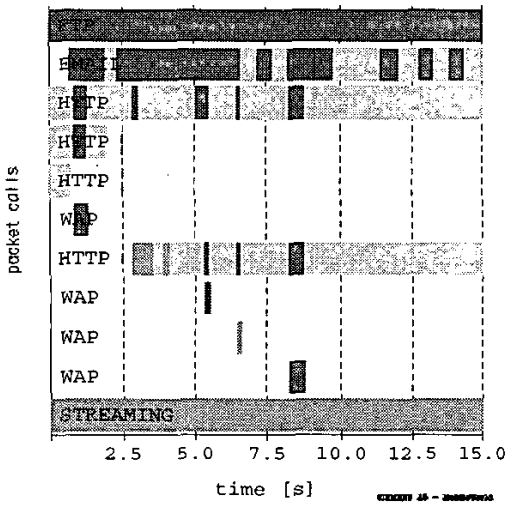

Figure 2. QoS TDF status in a cell vs. time

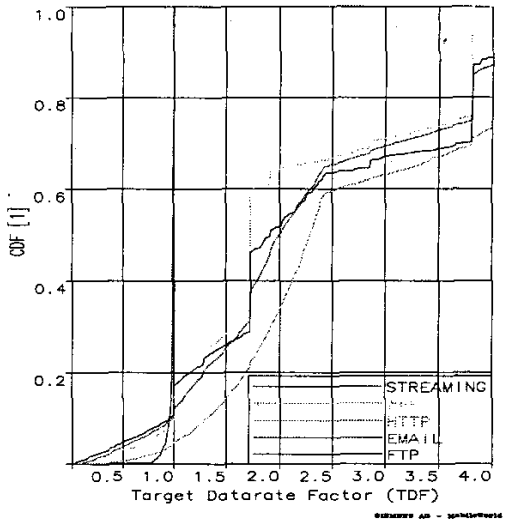

Figure 3. CDF of the mean TDF per session for the low data load scenario

In Fig. 2 at the beginning five TBFs are allocated in the cell. The streaming service is served at the *guaranteed bit rate* of 128 kbps all the time, while the TTP of 32kbps for the interactive HTTP and background (lowest QoS priority) FTP services is provided at TDF ∈ [0.5, 0.9[and TDF ∈ [0.1, 0.5[, respectively. At t = 0.7 s one HTTP service has been released and an e-mail service is allocated and served at the minimum granted QoS level (service sustenance level). At t = 0.8 s a new WAP service is allocated. All TBFs except the streaming service, which is excluded from downgrading, are downgraded to their service sustenance level and the new WAP service is allocated at TDF ∈ [0.1, 0.5[. After release of the WAP service, the previously downgraded HTTP services are upgraded to TDF ∈ [0.5, 0.9[, whereas the TBFs with the lowest QoS priority (FTP and e-mail) remain at TDF ∈ [0.1, 0.5[. The TDF status diagram reveals the desired QoS behavior. The streaming service having highest QoS priority is maintained at its TTP over the complete period. The least delay sensitive service (FTP) is served at the minimum QoS level defined by the network operator and uses additional resources only in case other services do not need them. The target of the QoS based RRM is to distribute the available resources according to the QoS requirements of the respective services.

Fig. 3 displays the CDFs for the mean TDF per session for different service types for the low data load scenario. For streaming services 90% of the sessions have obtained at least 96% of the requested resources with only few sessions experiencing a mean TDF < 1. The main reason for TDF < 1 is related to the MS multislot capability. For maintaining the TTP of 128 kbps with 4 TS MCS-7 has to be used at least. However, if a more robust coding scheme, e.g. MCS-6 is requested by link adaptation the TTP cannot be maintained. If more resources are available than required to meet the TTP for interactive and background services, the TBF is allocated to the maximum number of four time slots, if possible. In this low data load scenario most of the interactive and even background services have obtained more than the required resources for fulfilling the TTP. The 10%-ile TDF for HTTP (WAP) is 1.31 (1.00). Even the less delay sensitive e-mail and FTP services have obtained sufficient bandwidth. For e-mail (FTP) the 10%-ile TDF is 0.99 (0.95).

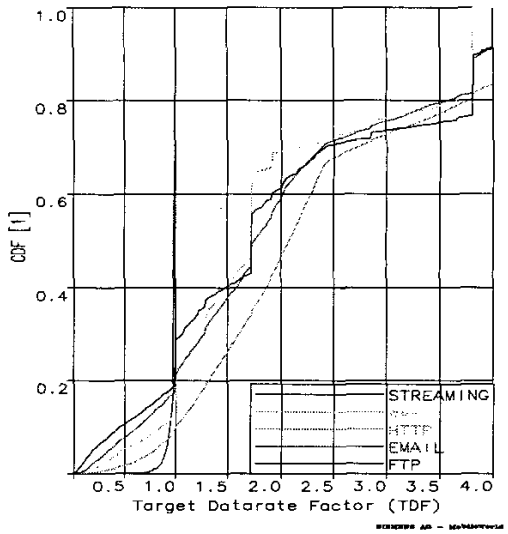

Figure 4. CDF of the mean TDF per session for the medium data load scenario

A significant increase of the percentage of sessions at TDF lower than 1 can be observed in the medium data load scenario displayed in Fig. 4. This diagram shows a clear separation

between more delay sensitive services and less delay sensitive traffic types. For the most delay sensitive streaming service 90% of the sessions have obtained at least 95% of the resources required to maintain the TTP of 128 kbps. The percentage of sessions at TDF < 1 is considerably lower for HTTP and WAP than for e-mail and FTP services. For HTTP the 10%-ile TDF is 1.00, i.e. the necessary resources for fulfilling the TTP of 32 kbps have been assigned to 90% of all HTTP sessions. For WAP the observed 10%-ile TDF is 0.86. The less delay sensitive e-mail and FTP services reveal a 10%-ile TDF of 0.63 and 0.45, respectively.

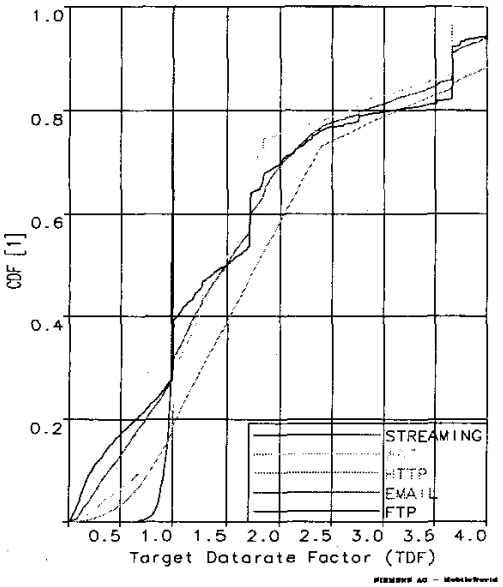

Figure 5. CDF of the mean TDF per session for the high data load scenario

For the high data load scenario the CDFs of the mean TDF per session for different service types are displayed in Fig. 5. For streaming services 90% of the sessions have obtained at least 92% of the requested resources with few sessions experiencing a mean TDF < 1. In overloaded network scenarios another aspect that causes further reduction of the TDF for streaming services can be observed, besides the MS multislot capability limitation. The possibility to utilize additional resources from other services depends on the setting of the service sustenance level. Ongoing packet data services might be downgraded only to that minimum throughput level. If services of lower QoS priority are already served at their sustenance level, no further "bandwidth borrowing" is allowed and the service of higher QoS priority cannot access additional resources.

For HTTP and WAP the number of sessions with mean TDF < 1 is considerably lower (e.g. the 10%-ile TDF is 0.79 and 0.70, respectively) than for the less delay sensitive e-mail and FTP services (the 10%-ile TDF is 0.39 and 0.22, respectively).

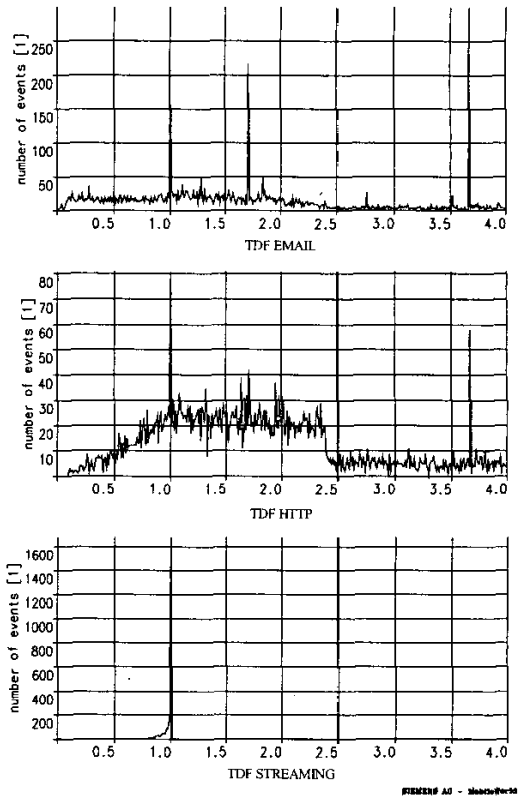

Figure 6. Probability density function for mean TDF per session for the high data load scenario

Fig. 6 displays the probability density function of the mean TDF for three different service types in the high data load scenario. The less delay sensitive e-mail service shows a peak at TDF = 1 and a wide range of nearly uniform distribution below TDF = 1. E-mail services have been served at TDF < 1 for multiple times, which can be observed in the distribution below TDF = 1. Due to statistical fluctuations of the occupied resources, there are also periods during which a high number of resources have been assigned to services of low QoS priority (TDF > 1).

HTTP sessions show a wide range with a considerable number of sessions having a mean TDF between 1 and 2.4. For interactive HTTP the ratio of sessions at mean TDF lower than the TTP is significantly lower than for the less delay sensitive e-mail service. Characteristic peaks can be observed in the histograms for e-mail and HTTP. The peaks at TDF = 1 are resulting from the assignment and maintenance of the requested resources. If there are more resources available than requested, additional resources up to the MS multislot capability have been assigned. The peaks at TDF = 1.70 and TDF = 3.67 are related to the allocation of four time slots to quite a few GPRS or EDGE packet data calls. For the most delay sensitive services (streaming) Fig. 6 exhibits a clear

peak at TDF = 1, i.e. the required resources for achieving the guaranteed bit rate of 128 kbps have been assigned.

In Fig. 7 the CDF of the mean TDF for HTTP is compared with that for an e-mail in the low, medium and high data load scenario, respectively. HTTP shows a significantly lower percentage of sessions at low mean TDF compared to the less delay sensitive e-mail service. For HTTP at low data load the 5% mean TDF is 1.02 implying that 95% of the HTTP sessions achieved a throughput higher than the targeted one while this portion for e-mail sessions is about 90%. At medium data load only 10% of the HTTP sessions get less bandwidth than requested while about 20% of the e-mail service is affected. At high data load 90% of the HTTP sessions get more than 80% of the requested data rate whereas 90% of the e-mail sessions perceive more than 40% of the requested bandwidth.

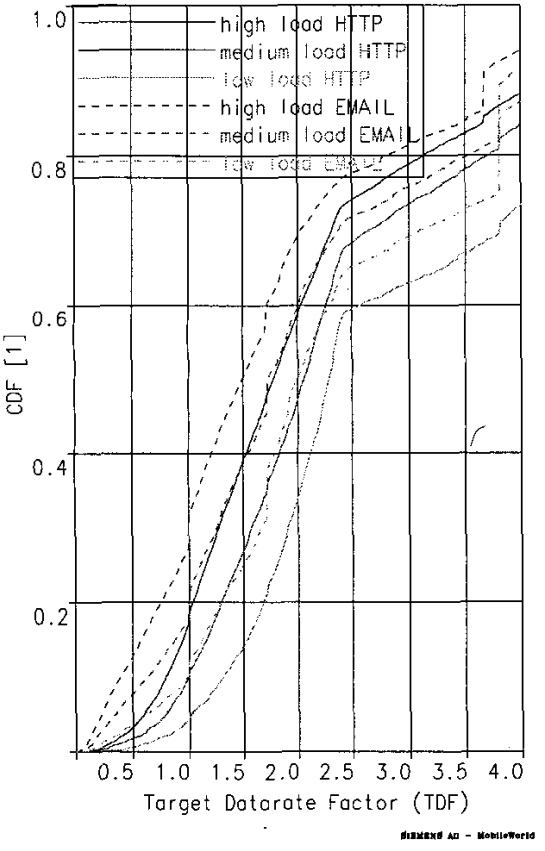

Figure 7. Comparison of the CDFs for the mean TDF in different data load scenarios

V. CONCLUSIONS

The performance of an advanced QoS strategy in GPRS/EDGE networks has been studied for different packet data load scenarios in a fully loaded voice cell. The simulation results demonstrate the maintenance of the requested QoS level for high priority services, whereas a controlled downgrading und upgrading process is applied to services of low priority. Real-time (e.g. streaming) services as the most delay sensitive ones do always obtain the requested bandwidth. Background services are downgraded to a much higher extent than interactive services. Especially in highly loaded GSM/GPRS/EDGE networks the introduction of QoS provides significant benefits for the end user and offers means to increase the service revenues according to the charging policy adopted by the network operator. To cope with the upcoming steadily growing packet data traffic and to avoid unacceptable voice blocking the installation of additional transceivers will then be required.

The introduction of QoS does not provide additional resources but it is able to distribute the resources in a different way, tailored to the expectations of the different user segments and the requirements of the different service types. QoS offers to network operators means to satisfy customers' demands, e.g. in terms of acceptable response time. Finally, QoS enables the introduction of new services, e.g. the delay sensitive streaming services.

VI. REFERENCES

[1] P. Stuckmann, "Quality of Service management in GPRS-based radio access networks", in Telecommunication Systems (19:3), Kluwer Academic Publishers, pp. 515-546, 2002.

[2] D. Fernandez and H. Montes, "An enhanced Quality of Service method for guaranteed bitrate services over shared channels in EGPRS", IEEE VTC Spring, pp. 957-961, 2002.

[3] 3GPP TS 23.107 v5.1.0 (2001-06), "QoS Concept and Architecture (Release 5)".

[4] K.Ivanov, C.F.Ball and F.Treml, "GPRS/EDGE performance on reserved and shared packet data channels", IEEE VTC Fall, Orlando, 2003.

[5] C.F.Ball, K.Ivanov and F.Treml, "Contrasting GPRS and EDGE over TCP/ IP on BCCH and non-BCCH carriers", IEEE VTC Fall, Orlando, 2003.

[6] T. Halonen, J. Romero and J. Melero, "GSM, GPRS and EDGE performance", Wiley & Sons, 2002.

[7] P. Stuckmann, "The GSM Evolution – Mobile Packet Data Services", Wiley & Sons, 2003.

[8] C.F. Ball, K. Ivanov, P. Stöckl, C. Masseroni, S. Parolari, R. Trivisonno, „Link Quality Control Benefits from a Combined Incremental Redundancy and Link Adaptation in EDGE Networks", IEEE VTC Spring, in press, 2004.