ADVANCED QUALITY OF SERVICE STRATEGIES FOR GERAN MOBILE RADIO NETWORKS

R. Müllner[1], C.F. Ball[1], K. Ivanov[1] and H. Winkler[2]

[1] Siemens AG, I&C Mobile Networks, Sankt-Martin-Str. 76, 81541 Munich, Germany, robert.muellner@siemens.com
[2] Siemens AG, PSE, Gudrun Str. 11, 1101 Vienna, Austria, hubert.winkler@siemens.com

Abstract - Customer demand for wireless data services is rapidly increasing. The introduction of GPRS, EDGE and UMTS providing high bit rate radio bearer, however, is not the complete response for satisfying the demands of these new high quality services. An advanced Quality of Service QoS management is necessary to handle the characteristic requirements of both different service types and user expectations. In this paper a new QoS strategy is proposed and analyzed comprising 3GPP QoS parameters along with operator's specific weighting factors to define the appropriate QoS priority of each service type and user profile. Admission control as well as a deterministic up- and downgrading strategy are applied to ensure a minimum grade of service for low-priority applications. Furthermore, delay time sensitive services and premium users are granted a full bandwidth. Simulation results are provided to qualify the behavior of the proposed QoS strategy under different packet data load conditions. Especially in highly loaded and even overloaded GERAN networks the introduction of QoS provides significant benefits for the end user and offers powerful 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.

Keywords - Quality of Service; QoS; GPRS; EDGE; radio resource management; scheduler

I. INTRODUCTION

In GERAN networks packet data applications have specific requirements in terms of e.g. 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. Packet data services vary significantly in their operational requirements. In general applications can be categorized into different groups. Low priority applications tolerate high delay and throughput variations resulting in low bandwidth requirements. In contrast other services have more stringent operational requirements, e.g. a constant bit rate or minimum delay. These are termed high priority applications. Besides the various application types also different user segments might have specific performance requirements and expectations towards the mobile network.

Fig. 1 shows a GERAN network layout featuring multiple Base Transceiver Stations (BTS), each of them covering several cells. The BTS is connected to a Packet Control Unit (PCU) that is typically located in the Base Station Controller (BSC). The PCU is connected via the Serving GPRS Support Node (SGSN) and Gateway GPRS Support Node (GGSN) to the Internet. Previously several attempts aimed at achieving a certain QoS level to the end user. The introduction of reserved and shared packet data channels (PDCH) in the Radio Resource Management (RRM) was an important step. Reserved PDCH are explicitly utilized by packet data traffic. A preemption of packet data by circuit switched traffic, e.g. voice calls, is not allowed. Shared PDCH is used commonly by both circuit and packet data on demand. This simple RRM strategy allows for a sufficient overall QoS for data and voice, however, a specific QoS level required by particular applications or users cannot be guaranteed due to system load [1].

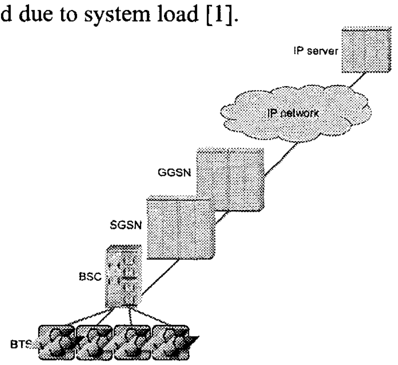

Figure 1. GERAN network layout

Therefore, additionally a limitation of the maximum number of services multiplexed on a single radio channel has been introduced. GSM allows in downlink a multiplexing of up to 16 simultaneous packet data calls per PDCH. A restriction to a maximum of 4-6 multiplexed data calls per PDCH grants an acceptable overall throughput. Again, a dedicated QoS is not provided for a particular application or user. Flow control is another option to introduce QoS aspects in mobile networks. Both a cell specific and a mobile specific implementation have been foreseen in the GSM standard. Flow control aims at avoiding overflow and congestion as well as smoothing the end-to-end data flow (see Fig. 2).

Further measures for QoS improvements are in the scope of network radio planning, configuration and optimization.

0-7803-8523-3/04/$20.00 ©2004 IEEE.

Service dependent channel allocation (SDCA) assigns services on channels that are best suited to meet the QoS requirements. Other examples are load balancing between different transceivers, hierarchical cells and/or frequency bands. The assignment of packet data services on hopping respectively non-hopping channels depending on the selected coding scheme has to be taken into account [2].

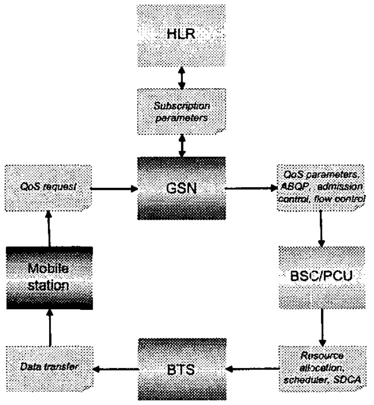
HLR
Subscription parameters
QoS request
GSN
QoS parameters, ABQP, admission control, flow control
Mobile station
BSC/PCU
Data transfer
BTS
Resource allocation, scheduler, SDCA

Figure 2. QoS information flow

Air interface, radio resource management and scheduling algorithm have the highest impact on the end-to-end QoS of the network. Hence full QoS support in future GERAN networks requires substantial modifications in the Radio Resource Management (RRM) as well as in the packet data scheduler. Previous work on these topics can be found in [3, 4]. The following analysis is focusing on QoS aspects of the Radio Access Network (RAN). The primary objective of this study is to reveal the benefit of a novel advanced QoS algorithm at different traffic load. The proposed QoS strategy comprises admission control, ranking of services according to the QoS priorities as well as deterministic up- and downgrading procedures. With the deterministic up- and downgrading strategy an automatic configuration of the system is given and only services of higher QoS priority obtain the right of downgrading services of lower QoS priority. 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. The performance of the proposed enhanced QoS strategy has been investigated by simulations. Detailed results for different load conditions are presented.

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. Quality of Service Model

The QoS model comprises 3GPP standard QoS parameters [5] and operator's specific weighting factors. Typical 3GPP QoS parameters are e.g. "traffic class", "traffic handling priority" and "allocation/retention priority". A particular parameter "guaranteed bit rate" has been introduced for real-time services. Additionally the proposed QoS model has been extended by operator's specific weighting factors. Those allow individual definition of priorities for service types and/or subscriber groups. The overall priority of each service is determined by a combination of the standard QoS attributes and the operator's specific weighting factors. The relative priority of a new service request with respect to all current services in the cell results in an appropriate resource allocation. Fig. 3 shows an example for GPRS/EDGE packet data calls in a cell with 4 ongoing data calls, i.e. temporary block flows (TBFs), having 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 TBF2 and TBF4 but not from TBF3 and TBF1.

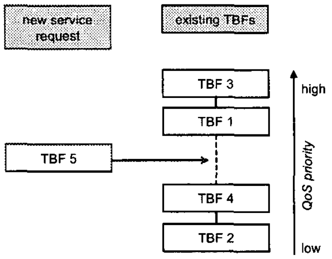

Figure 3. QoS priority of currently served TBFs and ranking of a new incoming service request

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 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 radio resource management (RRM) is to optimize the TTP for each service by appropriate time slot allocation and optimum adjustment of the share factors. The packet data scheduler distributes permission rights to the physical radio transmission resources on the services multiplexed on this time slot according to the 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 and streaming services, and 32 kbps for non real-time services.

Admission control regulates the service access to the RAN in an appropriate way. A new service request is admitted if sufficient resources are available. For packet data real-time services the guaranteed bit rate has to be provided. Interactive and background services are admitted if at least a tolerable quality level in the following termed service sustenance level can be obtained. The service sustenance level is defined as the minimum ratio of assigned data rate and TTP. In the simulation model the service sustenance level is set to 0.1, granting each service at least 10% of the resources required to meet the TTP. An existing TBF is downgraded to its service sustenance level. A new service request is queued for a certain amount of time if the necessary resources cannot be offered. For this purpose the QoS model defines a queuing-reject-timer. After expiry of this timer (5 s), the new service request is rejected. The applied admission control algorithm attempts to allocate services of high QoS priority at their target throughput rate even in case of high cell load.

In case of increasing traffic load, the network first starts downgrading of services from the initially allocated maximum number of resources to the TTP. The downgrading process starts with the service of lowest QoS priority. If further resources are necessary for allocating new service requests or for maintaining the QoS level of services of high QoS priority, a further reduction of the allocated resources below the TTP is performed in two steps. This downgrading procedure starts again with the service of lowest QoS priority. First the TBFs are downgraded from the TTP to the intermediate service sustenance level (corresponding to 50% of the TTP) and then to the minimum serving level (service sustenance level corresponding to 10% of the TTP).

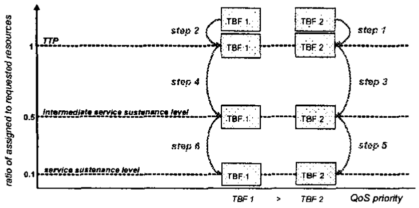

Figure 4: Downgrading sequence for two TBFs of different QoS priority

Fig. 4 shows an example for the downgrading sequence with two TBFs having different QoS priority. TBF 2 has lower priority than TBF 1. Hence the stepwise downgrade via TTP, intermediate and sustenance level is performed in alternating way.

The upgrading procedure is executed in the opposite way, i.e. it is started for the TBF of highest QoS priority. The maintenance of the TTP for each TBF is continuously controlled by the RRM.

III. Simulation Assumptions

The proposed advanced QoS strategy has been studied in a typical 2/2/2 configuration (2 transceivers per sector and three sector sites) with effective frequency reuse of 12. A network deployment with slow moving subscribers (3 km/h) in a mixed voice and GPRS/EDGE data traffic scenario has been assumed. The offered voice load has been dimensioned for a voice-only cell at 1% blocking (2 signaling channels per cell assumed resulting in 14 traffic time slots) and kept fixed for all simulated scenarios. All 14 traffic channels have been configured in a common pool that allows the allocation of both service types, voice services and data services [1]. A mixture of 50% GPRS and 50% EDGE four time slot capable mobile stations 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, e-mail and FTP, [2]) 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 [6], all others in RLC acknowledged mode.

IV. Simulation Results

The performance benefit of QoS in GERAN networks has been evaluated in different packet data load scenarios. The gain from QoS is moderate at low data load. However, even in low data load scenarios a clear separation of services according to their QoS priority can be observed. With increasing data traffic in the network the QoS becomes more and more important in order to keep subscribers satisfied. Fig. 5 shows the status of the Target Datarate Factor (TDF) for each TBF in a particular cell during a period of 30 seconds. The TDF is defined as the ratio of assigned to requested resources (TTP). In Fig. 5 different colors indicate the degree to which the TTP of 128 kbps for streaming services and the TTP of 32 kbps for WAP, HTTP, e-mail and FTP is met: blue color means TDF ≥ 1.1, green means $0.9 \leq TDF < 1.1$, yellow means a TDF in the range of 0.5 to

0.9 and red means a TDF < 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 applied coding scheme by dynamic link adaptation and the BLER corresponding to the actual radio conditions. At the beginning of each packet call an initial coding scheme (MSC-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.

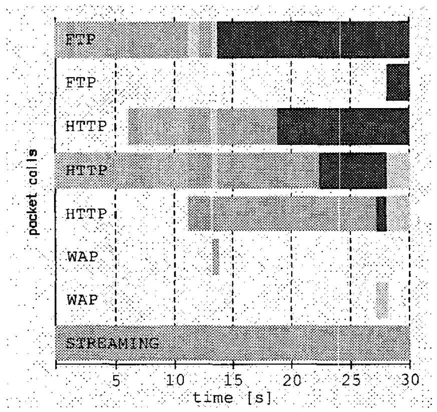

Figure 5. QoS TDF status in a cell vs. time

In Fig. 5 at the beginning three packet data services are allocated in the cell (FTP, HTTP and streaming). In addition seven voice calls are allocated in this cell, those are not depicted in Fig. 5. The streaming service is served at the guaranteed bit rate of 128 kbps (green color), while sufficient resources are initially available to serve the HTTP and FTP service at more than 32 kbps. The latter is indicated by blue color. After 6.1 s a second HTTP data call is allocated and all packet data services are served at their TTP, which is indicated by green color for each TBF. At 8.2 s an additional time slot becomes available for data services due to the release of a voice call. The interactive HTTP services obtain more resources (TDF > 1), while the FTP background service is still kept at its TTP. At t = 11.1 s a new voice call is allocated and the interactive HTTP services are downgraded to their TTP. The background service FTP is also downgraded and is served below its TTP, which is shown by yellow color. At t = 13.2 s another voice call and an interactive WAP service are allocated. The resources assigned to the interactive and background services are downgraded in order to offer the target data rate to the interactive WAP service. This is shown by the transition from green to yellow color. Hence the FTP background service is served below its intermediate service sustenance level (red color). At t = 18.8 s again a voice call is allocated. For maintaining the high priority packet data services at their TTP, the FTP background service and one interactive HTTP service get served at data rates below the intermediate service sustenance level. At t = 22.3 s an additional voice call induces the downgrade of the next HTTP service, since all other packet data services of lower QoS priority are already served at their minimum sustenance level. At t = 28.1 s a new WAP packet call is allocated and served at TDF between 0.5 and 0.9 (yellow color). After the WAP has been released the free capacity is used for upgrading two HTTP sessions.

The TDF status diagram reveals the desired QoS behavior. The streaming service having highest QoS priority is maintained at its TTP over the complete time period. The FTP as least delay sensitive background service was served far below its TTP, whereas the more delay sensitive interactive HTTP services suffer from less downgrades.

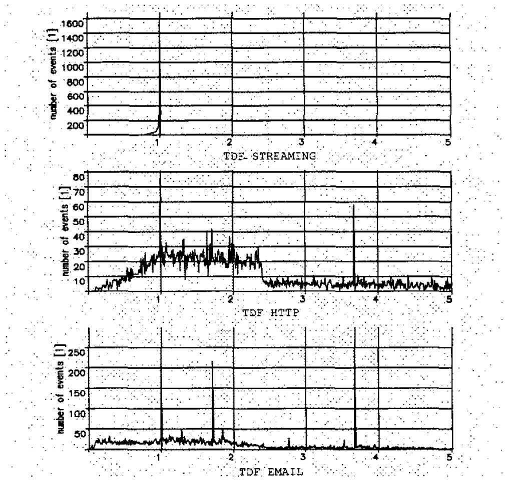

Figure 6. Histogram of the mean Target Datarate Factors TDF per session for the high load scenario

In Fig. 6 and 7 the histogram and the corresponding cumulative density function (CDF) of the mean TDF per session for different service types are presented for the high data load scenario. For streaming services the probability density function exhibits a clear peak at TDF = 1. HTTP sessions show a broad distribution with a significant portion of sessions having a mean TDF between 1 and 2.4. For interactive HTTP the ratio of sessions with mean TDF < 1 is considerably lower than for the less delay sensitive e-mail service. 90% of the streaming sessions have obtained at least 92% of the requested resources with only few sessions experiencing a mean TDF < 1. The main reason for TDF < 1 is related to MS multislot capability. For maintaining the TTP of 128 kbps with 4 TS at least MCS-7 has to be used. However, if a more robust coding scheme, e.g. MCS-6 is requested by link adaptation the TTP cannot be maintained with 4 TS. For HTTP and WAP the number of sessions with mean TDF < 1 is considerably lower (90% of the sessions are characterized by TDF > 0.79 and 0.70, respectively) than for the less delay sensitive e-mail and FTP services (90% of the sessions show TDF > 0.39 and 0.22, respectively).

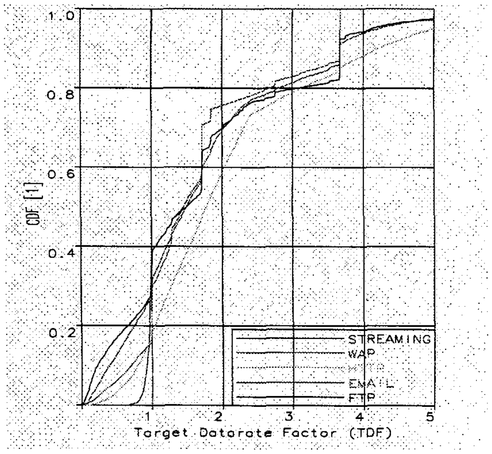

Figure 7. CDF of mean TDF per session (high load scenario)

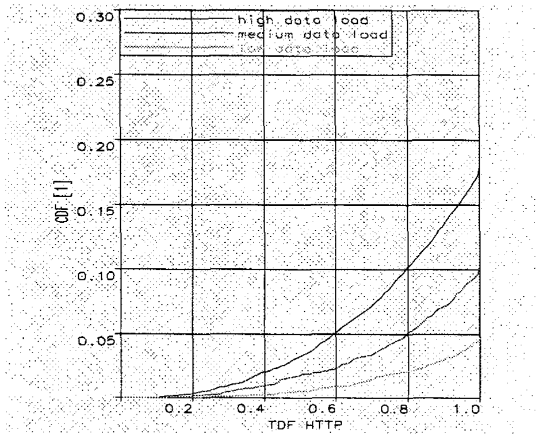

Figure 8. Comparison of the cumulated TDF distribution for HTTP in different load scenarios

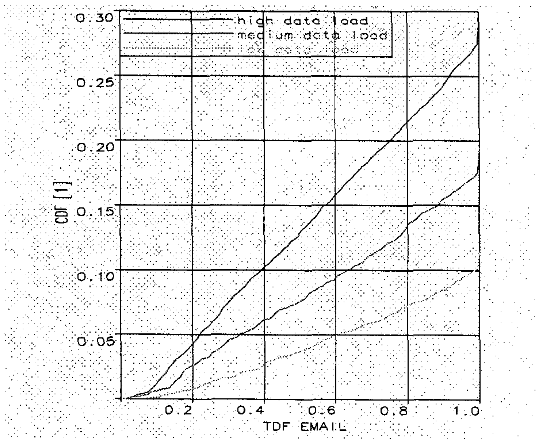

Figure 9. Comparison of the cumulated TDF distribution for e-mail in different load scenarios

In Fig. 8 and 9 the CDFs of the TDF for HTTP and e-mail are compared at varying data load. HTTP shows a significantly lower percentage of sessions at low TDF compared to the less delay sensitive e-mail services. For HTTP at low load roughly 95% of the HTTP sessions achieved a throughput higher than the targeted one (TTP) while this portion for e-mail sessions is about 90%. At medium load only 10% of the sessions get less bandwidth than requested while about 20% of the e-mail service is affected. At high load 90% of the HTTP sessions get more than 80% of the requested resources whereas 90% of the e-mail sessions perceive more than 40% of the requested bandwidth.

V. CONCLUSIONS

The performance of an advanced QoS strategy in GERAN 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 and 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. The downgrading level can be defined by the operator. Especially in highly loaded GERAN 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.

VI. REFERENCES

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

[2] 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.

[3] P. Stuckmann, "Quality of Service management in GPRS-based radio access networks", in Telecommunication Systems (19:3), Kluwer Academic Publishers, April 2002, pp. 515-456.

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

[5] 3GPP TS 23.107 v 5.1.0 (2001-06), "QoS Concept and Architecture (Release 5)".

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