the Iub link carrying the elastic Internet traffic which resides on top of the TCP/IP protocol suite, where the rate of TCP flow adjusts itself to fill the available bandwidth according to the network traffic condition by using the TCP flow control. This dimensioning approach is based on the M/G/R Processor Sharing (M/G/R-PS) queuing model which characterizes TCP traffic at the flow level, where mobile users represent flows generated by downloading Internet objects; the sojourn times represent the object transfer times. The transmission bandwidth of UMTS user connections is limited by their assigned radio access bearer (RAB) type, e.g. 64 kbps or 384 kbps. In order to guarantee a minimum throughput one must have a CAC (Call Admission Control) [6] scheme employed in the network to avoid instability and low "goodput" in case of overload. The admission control for elastic traffic is based on the maximum number of allowed flows. Whenever a per-flow access control is employed a blocking model is used to dimension the network. Thus the original M/G/R-PS method is extended to an M/G/R/N-PS model, which allows a maximum number of N active user connections in the system sharing R servers simultaneously. In addition, since the Iub interface has to fulfill strict delay requirements posed by UMTS specific layers in order to guarantee each radio frame delivered on time according to the air interface, the FP PDU delay is also a very critical dimensioning issue to be specifically considered. Due to this AAL2 was chosen as the adaptation layer which basically is focused on real-time traffic whereas a major portion of UMTS upper layer traffic represents non real-time characteristics.

The remaining parts of this paper are organized in 4 sections. The following section summarizes analytical approaches for UTRAN dimensioning with focus on application performance. In the third section a description of the simulation model is given and the simulation results obtained from this model are presented. In the last section, key conclusions on the dimensioning problems of UMTS and an outlook of future work are given.

## 2. DIMENSIONING METHODS FOR ELASTIC TRAFFIC

Figure 2-1 shows an overview of a UMTS network sketching its main components. Populations of mobile subscribers are connected to the NodeB via the radio interface (Uu-Interface), and then go through the UTRAN and Core Network to the remote Internet servers. The link between the Base Station (NodeB) and the Radio Network Controller (RNC), the Iub interface, is a potential bottleneck within the UTRAN. End users generate service requests to the application servers to download Internet objects such as Web pages, emails over TCP connections, which are established between the end users and the application servers. In this way, each request generates one or more 'TCP flows' over the Iub link.

It is assumed in this analysis that every elastic traffic flow is generated by one file transfer and all users are assigned to the same maximum radio access bit rate, e.g. 128kbps. The file transmission rates are controlled by the TCP feedback algorithm as a network congestion control function. If TCP works ideally (i.e. instantaneous feedback), all elastic traffic flows going over the same link will share the bandwidth resources equally, and thus the system only carrying elastic traffic flows is essentially behaving as an M/G/R-Processor Sharing queue on the Iub link. Each application is assigned a specific radio access bearer with a certain peak data rate. This is modeled by assuming R servers inside the system where each application can (at maximum) utilize the full capacity of a single server. In this paper the focus of the analysis is placed on a single RAB type, i.e. each application receives the same

decrease of the average throughput) due to link congestion. It is a quantitative measure of how link congestion affects the transaction times, taking into account the economy of scale effect. An example of calculating the file transfer delay using formula 2-1 can be seen in Figure 3-1.

## 2.2 Extended M/G/R-PS Model

However, the M/G/R-PS model introduced above assumes ideal capacity sharing among active flows. But in practice the TCP flows are not always able to utilize their fair share of available bandwidth. TCP's effectiveness of capacity sharing is determined by the TCP slow start and congestion avoidance mechanisms which are affected by network conditions such as round-trip times and packet loss probability. Slow start is executed at the beginning of the TCP transmission when the link capabilities are unknown to the senders. During the slow start phase, the TCP congestion window, which represents the number of TCP segments allowed to send to the network when receiving an acknowledgement from the receiver, starts from one segment and then is increased exponentially until a slow start threshold value is reached. Within the slow start phase, the available bandwidth assigned to a connection is not fully utilized. Thereby, the slow start mechanism leads to the increase of the file transfer time by not completely utilizing the available bandwidth at the beginning of each transmission process. Especially in scenarios with low utilization, this influences the achieved throughputs while the number of concurrent flows is not sufficient to utilize the bandwidth left by TCP flows being in slow start phase. To solve this problem, [2] proposes a realistic modeling which considers the impact of slow-start into the M/G/R-PS model.

Given the required delay factor $f_R$ and the maximum rate $r_p$, the amount of data sent up to the time when the sender can start utilizing its bandwidth share is calculated as:

$$x_{slow-start} = \left(2^{n^*} - 1\right) MSS \quad \text{with} \qquad n^* = \left\lceil \log_2 \left( \left\lceil \frac{r_p RTT}{f_R MSS} \right\rceil \right) \right\rceil \tag{2-2}$$

where $n^*$ represents the time step in terms of Round Trip Times (*RTT*) where the available shared bandwidth starts to be fully utilized and *MSS* (Maximum Segment Size) is the maximum allowed TCP packet size (in our scenario *MSS* is set to 1460 bytes). If the total size $x$ of the file is smaller than $x_{slow\text{-}start}$, the sender never reaches the state of fair capacity sharing. Therefore, the approximation of the expected file transfer delay $E\{T(x)\}$ for files of size $x$ (including overhead) can be more accurately described by formula 2-3 (cf. [2]):

$$E\{T(x)\} = \begin{cases} \left\lfloor log_2 \left( \left\lceil \frac{x}{MSS} \right\rceil \right) \right\rfloor RTT + E_{M/G/R}\{T(x - x_{start})\} & x < x_{slow-start} \\ n^* RTT + E_{M/G/R}\{T(x - x_{slow-start})\} & x \geq x_{slow-start} \end{cases} \tag{2-3}$$

As seen from formula 2-3, the computation of the expected time to transfer all data is divided into two parts: the first part gives the sum of all necessary RTTs which are mostly for waiting for acknowledgements as a consequence of the slow start mechanism. The second term considers the time of sending the rest of the data with the available capacity. $x_{start}$ is the amount of data sent within the first part as shown in formula 2-4, which is only used when the file size $x$ is smaller than $x_{slow\text{-}start}$:

### 2.3 Performing the Connection Admission Control

It has been advocated by Roberts et al. [6] that admission control for TCP flows allows guaranteeing a minimum throughput. Let $r_m$ denote the minimum rate which can be used as an admission control threshold. Then the number of admitted flows is limited by: $N = C/r_m$. This behaves as an M/G/R/N-PS model. The results for the M/G/R/N-PS queue follow directly from the M/G/R/∞-PS queue only that the state space corresponding to the number of flows or connections is now restricted to a total of number of $N$. The probability of each state is given in formula 2-7. When the number of connections reaches the maximum value $N$, formula 2-7 gives the blocking probably $p(N)$. When $R = N$, $p(N)$ reduces to Erlang's first formula.

$$p(j) = \begin{cases} \dfrac{(1-r)\dfrac{R!}{j!}(Rr)^{j-R}E_2(R,Rr)}{1-E_2(R,Rr)r^{N-R}r} & (j < R) \\ \dfrac{E_2(R,Rr)r^{j-R}(1-r)}{1-E_2(R,Rr)r^{N-R}r} & (N \geq j \geq R) \end{cases} \tag{2-7}$$

Here $\rho$ is the offered traffic load. With the probability of each queue state, the average number of connections (or mean queue size) can be calculated. By applying Little's law, the average file transfer delay can be obtained by:

$$E\{T\} = \frac{E\{W\}}{l(1-p(N))} \quad \text{with} \quad E\{W\} = \sum_{j=0}^{N} j \cdot p(j) \tag{2-8}$$

## 3. MODEL VALIDATION AND SIMULATION RESULTS

This section validates the M/G/R-PS model for configurations without admission control and the M/G/R/N-PS model for configurations using admission control. The validation is performed by comparing the results of the analytical investigation with simulation results obtained by the UTRAN simulation model developed for this study.

### 3.1 Simulation Model Set-up

The considered traffic type (web, ftp) is assumed to use UMTS QoS "best effort" (cf. [14]) which allows the UTRAN to select a bearer type which is most suitable for a certain UE at a certain time (i.e. higher rates for nearby UEs in unloaded radio cells and lower rates for far-off UEs in loaded cells, etc.). With respect to "best effort" services the properly dimensioned Iub has to achieve two targets. Firstly, the application performance (mainly response time for object transfers) has to fit to the requirements. Secondly, the interaction of the radio link control layer (RLC) and the frame protocol layer as a transport resource needs a timely arrival of the radio frames (i.e. FP PDU), because radio frames arriving (downlink from RNC to NodeB) later than scheduled for the transmission on-air will be discarded resulting in a waste of Iub bandwidth (additional load due to re-transmission of discarded blocks, if RLC operates in acknowledged mode). The percentage of FP PDUs exceeding the allowed time budget is referred to as "delayed FP PDUs" within this paper.

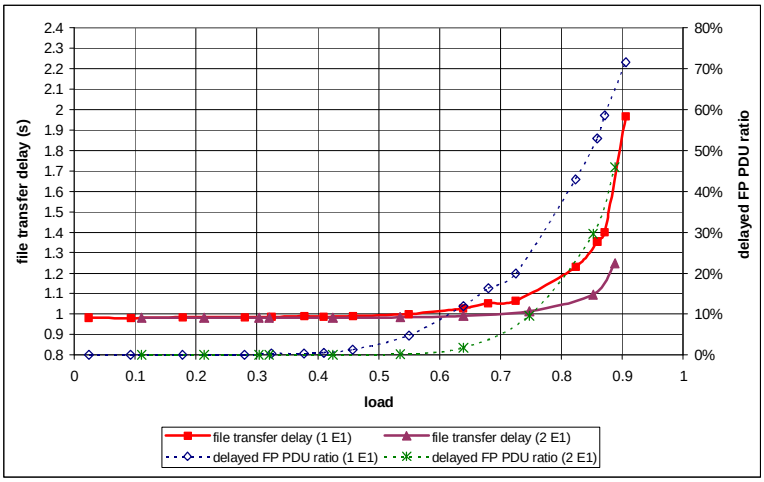


**Figure 3-2: File transfer delay, delayed FP PDU ratio: constant file size, mean =12kB**

Figure 3-2 presents simulation results for the delayed FP PDU ratio and file transfer delays over different loads, for both a single E1 line and 2 E1 lines. The ftp connection performance for load ratios above 0.85 is severely degraded. For dimensioning purposes the file transfer delay in the high-load case (load ratio > 0.8) can be negligible because the requirements for the frame protocol delay become the constraining factor. Not using admission control, in case of both link capacities, when the load ratio is above 0.9 the system is overloaded and the application performance becomes unstable, i.e. result in an extremely high file transfer delay. Figure 3-2 also highlights the achievable multiplexing gain comparing an Iub with a single E1 line with the 2 E1 line case. The available bandwidth ratio delimited by the amount of "delayed" frame protocol PDUs reaching 10% is increased from 0.6 to 0.75 when changing the Iub bandwidth from one 1 E1 to 2 E1. Thus the traffic properly served by the Iub (i.e. expected traffic from the group of radio cells belonging to the corresponding NodeB) could be increased by 150% from 0.6 E1 units to 1.5 E1 (0.75×2 E1) units although the line capacity is "only" doubled. Furthermore, the file transfer delay is also greatly improved by the additional E1 line. At the load ratio of 0.75 the delay obtained from two E1 lines is 7.8% lower than that from one E1 line case. If the mobile network operator aims for a traffic load reaching the maximum acceptable utilization of an E1 link, the acceptable traffic load can be derived from the corresponding results of both application and FP PDU performance. With a maximum expected delay of 1.2 seconds for a file transfer the utilization can reach 80% (1 E1) according to Figure 3-2. With a maximum of 5% of "delayed" FP PDUs only 55% utilization is acceptable, i.e. the FP delay is the limiting QoS constraint and an over-saturation of the application QoS requirement as the ftp delay for 55% utilization is close to the unloaded line results. It has to be noted, that the requirement for FP PDU has to be carefully aligned with implementation and parameter settings of higher layers, e.g. using RLC in acknowledged mode with retransmission is more robust against FP PDU delay/losses than RLC transparent mode.

## 3.3 Scenario with Admission Control

When employing the admission control into the system, the system can avoid the heavy congestion by limiting the number of active user connections simultaneously on the link, and

## 4. CONCLUSION

Within this paper two key constraints (file transfer delay and FP PDU delay) for the dimensioning of the Iub interface in UMTS access network have been discussed. The dimensioning approach based on the M/G/R-PS model has been demonstrated to be able to determine the application performance perceived by a UMTS user. The presented simulation results also prove the necessity of using admission control to protect against instability e.g. overload situations and thus guarantee the quality of service even for "best effort" class. The proposed M/G/R/N-PS model for determining the application performance of applying admission control was validated by comparing with the simulation results. Nevertheless, the importance of the performance on the AAL2 layer is stressed additionally. The related performance curve was obtained from the simulations in this study. The derivation of an analytical approach to predict the AAL2 performance based on a given UMTS traffic profile is subject to further work. Although, the focus of this paper lies only on packet switched user plane traffic, of course the Iub link also has to enable transferring the real-time stream traffic such as voice and video services, in addition certain control data channels exchange data between the NodeB and the RNC. Available dimensioning approaches for mixing stream and elastic traffic giving higher priority for stream traffic have been discussed in [7] and [8]. This will be further investigated. In this paper, the shown simulations emphasize the necessity for a careful network monitoring in evolved UMTS networks.

## REFERENCES


1. Zhong Fan, Marconi Labs Cambridge, UK, "Dimensioning Bandwidth for Elastic Traffic"
2. Anton Riedl; Thomas Bauschert; Maren Perske; Andreas Probst; "Investigation of the M/G/R Processor Sharing Model for Dimensioning of IP Access Networks with Elastic Traffic", Munich University of Technology (TUM), Siemens AG
3. Anton Riedl; Maren Perske; Thomas Bauschert; Andreas Probst, "DIMENSIONING OF IP ACCESS NETWORKS WITH ELASTIC TRAFFIC", Siemens AG; TUM
4. D. P. Heymant; T. V. Lakshman; Arnold L. Neidhardt, "A new method for analysing feedback-based protocols with applications to engineering web traffic over the Internet"
5. R. Vranken; R.D. van der Mei; "Performance of TCP with Multiple Priority Classes"
6. L. Massoulie and J. Roberts, "Arguments in favour of admission control for TCP flows"
7. K. Lindberger, "Balancing quality of service, pricing and utilization in multiservice networks with stream and elastic traffic", in ITC 16, 1999
8. R. Nunez, H. van den Berg and M. Mandjes (1999), Performance Evaluation of Strategies for Integration of Elastic and Stream Traffic, Proceedings ITC 16, Edinburgh.
9. F. Brichet, Admission Control in multiservice networks
10. Anton Riedl; Thomas Bauschert; "A frame work for multi-service IP network planning"
11. Jianhua Cao, Mikael Andersson, Christian Nyberg and Maria Kihl, "Web Server Performance Modeling Using an M/G/1/K*PS Queue", Lund Institute of Technology
12. T. Bonald, "Insensitivity results in statistical bandwidth sharing", France Telecom R&D.
13. 3GPP 25.853, "Delay Budget within the Access Stratum"
14. 3GPP 23.107, "Quality of Service Concept"