Published in IET Software
Received on 16th January 2009
Revised on 3rd July 2009
doi: 10.1049/iet-sen.2009.0006

In Special Issue on Performance Engineering

ISSN 1751-8806

Analytical TCP throughput model for high-speed downlink packet access

L. Bodrog[1] G. Horváth[1] C. Vulkán[2]

[1]Department of Telecommunications, Technical University of Budapest, Magyar tudósok krt. 2, Budapest H-1521, Hungary
[2]Nokia Siemens Networks, Köztelek u. 6, Budapest H-1092, Hungary
E-mail: bodrog@hit.bme.hu

Abstract: This study deals with the throughput analysis of data services over high-speed downlink packet access (HSDPA) systems. The achievable throughput is calculated with an approximate analytical method based on the Padhye model that has two input parameters: the packet loss probability and the transmission control protocol (TCP) round trip time. The proposed solution is to calculate these parameters with an equivalent queuing network model of the HSDPA system that takes into account the possible congestion points in the system and the protocol layers that have dominant impact on the delay and packet drop. The modelling considerations and the analysis method are described in detail. Finally, the model is validated with a performance study of the HSDPA system that is executed with detailed NS2- based simulations too. The proposed method is found to be reasonably accurate requiring less computational effort than the simulations.

1 Introduction

High-speed downlink packet access (HSDPA) is a packet-based downlink service for data users over the universal mobile telecommunications system (UMTS) with data rates ranging up to several megabits per second [1].

In conventional UMTS, Layer 2 protocols of the radio protocol interface, such as radio link control (RLC) and medium access control (MAC) protocol, are terminated in the radio network controller (RNC). Physical layer protocols of the radio interface are implemented in the Node-B that is connected to the RLC via the Iub interface. In acknowledged mode (AM), the RLC is responsible for error-free, in-sequence delivery of the user data. This is achieved by retransmissions based on the automatic repeat request (ARQ) mechanism. RLC retransmissions increase the Layer 2 round trip time (RTT) and may trigger TCP timeouts.

In HSDPA an additional protocol layer located in the Node-B (see Fig. 1), namely MAC-hs, was introduced, which makes Node-B controlled fast adaptation of the modulation and coding scheme, fast scheduling and retransmission handling with the Hybrid ARQ (HARQ) functionality possible. This solution reduced the Layer 2 RTT when retransmissions are required because of erroneous data transfer. Although retransmissions are handled by the Node-B, RLC ARQ was maintained in the system for compatibility reasons with the earlier (Release '99) network solutions, that is, to support the channel switching between HSDPA and Release '99 transport channels, and to provide the capability of soft handover control. The RLC ARQ still handles retransmissions if the maximum allowed number of MAC-hs retransmissions is exceeded or there are packet drops on the transport network. The RLC retransmissions increase the RTT of the data connections using HSDPA service. In addition, as RLC retransmissions might also be triggered by increased transport delays, the RLC contributes to the congestion. These factors and the in-sequence delivery of the user packets by the RLC lead to the fact that, the TCP flow control is incapable of detecting and resolving the congestion situation on the Iub interface. As a result, the TCP notices the congestion only upon Timeout or when finally the RLC discards the packets that have reached the maximum number of retransmissions.

The distribution of the radio protocol architecture between the RNC and the Node-B requires that a flow control algorithm – the HSDPA flow control [2] – is

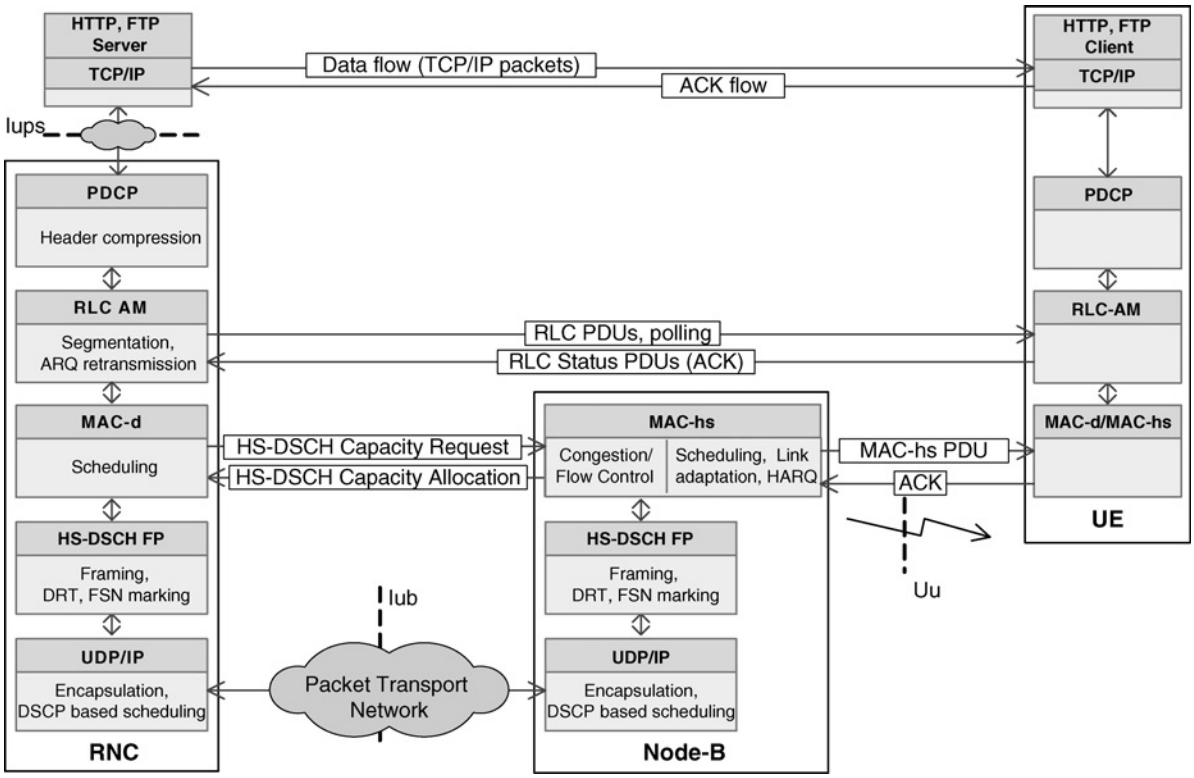


Figure 1 *Overview of the protocols of HSDPA*

implemented. With this algorithm, the Node-B controls the amount of data sent from the RNC in order to keep its buffers at optimal level so that the air interface capacity is not wasted, and at the same time the delay on the Node-B buffer is kept not too high. Typically, the HSDPA flow control monitors the Node-B buffer size and the amount of transferred packet data units (PDUs) over a sampling period without considering the available resources on the Iub transport network shared by real time, non-real time and HSDPA services.

A good indicator of the level of service an HSDPA access network can provide to the mobile users is the achievable TCP throughput. There are numerous papers on the TCP performance in HSDPA. Several papers apply discrete event simulation to study the TCP performance over HSDPA [3, 4]. Some other papers present analytical solutions by modelling the behaviour of the Node-B only. In [5–8] the throughput of TCP traffic is evaluated based on analytical formulas and Monte-Carlo simulation. In [9] a queueing model is presented for the Node-B with compound poisson process traffic. These results cannot be used for transport network dimensioning since they do not model the transport network at all, thus the case when the transport network is the bottleneck is not covered. The analysis of the UMTS terrestrial radio access network (UTRAN) including the transport network is provided in [10] but some random and probabilistic terms are involved in the solution that are difficult to be determined analytically. Therefore simulations have been used to obtain the TCP throughput. In [11] a completely different, elastic approach is used to analyse the throughput in the UTRAN. However, in that paper the focus is on the performance analysis of a specific congestion control algorithm, and it assumes ideal elastic traffic and no buffers in the network, thus it is not suitable for link capacity dimensioning.

In this paper we propose an analytical throughput model for TCP connections in the HSDPA UTRAN. This model includes all the buffers of the network devices and protocol entities the packets have to pass through in the radio access network when transmitted from the RNC to the user equipment. We provide an iterative procedure that approximates the TCP throughput based on the RTT and loss values calculated from the queueing network model. The main purpose of our method is to support transport network dimensioning, thus to investigate how the TCP throughput changes with the link capacity of the transport network.

The rest of the paper is organised as follows. Section 2 gives a short technological overview on the HSDPA UTRAN. It describes how the packets are delivered from the RNC to the user equipments (UE) and introduces a queueing network model of the system. Section 3 summarises the concept of the approximate throughput calculation and describes in detail the queuing network model of the system. Numerical examples and the performance evaluation of the system are provided in Section 4, finally Section 5 concludes the paper.

2 System overview and the equivalent queuing model

The overview of the radio access network configuration in case of HSDPA service is shown in Fig. 1. After header compression in the packet data convergence protocol (PDCP) layer, the incoming data (TCP/IP) packets are segmented and encapsulated by the RLC AM entity. These segments (PDUs) are scheduled by the MAC-d layer according to the HSDPA flow control commands. The RLC entity actively polls the user equipment (UE) that responds with status PDUs indicating the sequence number of lost and received PDUs. Lost PDUs are retransmitted. The master of the HSDPA flow control is the MAC-hs located in the Node-B. It grants resources to the HSDPA connections (MAC-d flows) at the RNC by sending a high-speed dedicated shared channel (HS-DSCH) Capacity allocation message that includes the allocation size, that is, the number of PDUs and their maximum size (HS-DSCH credits, MAC-d PDU length), the interval the data can be sent at (HS-DSCH Interval) and the validity period of the allocation (HS-DSCH Repetition Period). This message is sent either solicited, upon reception of a HS-DSCH CAPACITY REQUEST message from the RNC, or unsolicited. The HS-DSCH frame protocol (FP) assembles a frame out of the scheduled PDUs and transfers it to the ATM adaptation layer type 2 (AAL2), where these frames are segmented to 45 bytes, and encapsulated into common part sublayer (CPS) packets. The size of the CPS-packet header is 3 bytes, thus the maximum size of one packet is 48 bytes. The CPS-packets are eventually assembled into CPS PDUs and sent to the destination via the virtual channel connection (VCC). The CPS-PDU header is 1 byte long, thus at maximum 47 CPS-packet bytes can be fitted into one asynchronous transfer mode (ATM) cell. As queues are intrinsic to the HSDPA system, a natural approach to model the TCP RTT – which is an important parameter with impact on the overall TCP performance – is to create an equivalent queuing model. Accordingly, the potential bottleneck points that dominate the downlink delay have to be identified (in case of mobile services the users are mainly downloading content to their mobiles loading the system mostly in downlink). The developed model focuses on the downlink performance, whereas the uplink delay is modelled with a constant delay. Packet drop (p) can appear at these bottleneck points because of buffer overflow or when the maximum number of retransmissions is reached. There are three such points in the system:

- The buffers of the RLC layer where the RLC PDUs (resulted from the segmentation of the user packets) are stored until a positive acknowledgement arrives or the maximum number of retransmissions is reached and the RLC AM entity discards them. The RLC buffers are scheduled by the MAC-d layer based on the credits received from the Node-B (MAC-hs entity). These credits are calculated in order to maximise the air interface throughput. The congestion situation over the Iub links is not necessarily taken into consideration, thus the RLC layer can easily overload the transport network. In this model, it is assumed that the uplink delay of the HS-DSCH CAPACITY ALLOCATION message is zero.

- The buffers of the AAL2/ATM transport network. As the transport network is a shared and limited resource, congestion may occur leading to increased delay and eventually to packet drops. In this paper the transport network is modelled with one buffer corresponding to the bottleneck link.

- The MAC-hs buffers in the Node-B. There is one buffer per MAC-d flow (HSDPA connection) that stores the MAC-d PDUs waiting for transmission. The amount of the MAC-d PDUs that can be transmitted in a 2-ms long transmission time interval (TTI) depends on the reported channel quality indicator (CQI). In case of transmission failure, the MAC-d PDUs are retransmitted. If the maximum number of retransmissions is reached, the MAC-d PDUs are discarded by the HARQ and the RLC ARQ will handle further retransmissions.

An overview of the queueing network model of the system is shown in Fig. 3. The three components of the queuing model, that is, the RLC buffers, the transport buffer, and the MAC-hs buffers, are located at different protocol layers. Each flow has a dedicated buffer at the RLC layer that stores the PDUs resulting from the segmentation of the TCP packets. The MAC-d schedules these buffers independently based on the credits received from the Node-B. PDUs are discarded in case of buffer overflow or when the maximum number of retransmissions is reached. Each PDU is stored in the buffer until the positive acknowledgement is received or until PDU discard procedure is executed by the RLC.

The transport network is modelled by one buffer representing the bottleneck link. ATM cells are discarded at buffer overflow. At the Node-B, each MAC-d flow has a dedicated buffer. At each 2 ms TTI, the proportional fair (PF) scheduler selects the buffer to be served based on the average throughput of each flow and its instantaneous channel quality. Upon an erroneous transmission over the air interface, the PDUs are retransmitted until the maximum number of transmissions is reached.

3 Concept of the TCP throughput calculation

There are several models available to calculate the TCP throughput. The most popular one is the so-called Padhye

doi: 10.1049/iet-sen.2009.0006

model [12]. This model essentially gives a simple formula that expresses the TCP throughput (B) as a function of the packet loss (p) and round trip time (RTT) (see (1))

In the formula p denotes the packet loss probability, b is the number of packets covered by one acknowledgement ($b = 1$ is assumed in this paper), T_0 is the timeout (we use $T_0 = 1.5$ s), RTT is the round trip time of the packets, W_u is the random variable denoting the unconstrained congestion window size, W_{max} is the maximum value of the constrained Congestion Window size. $E(\cdot)$ is the expectation operator and accordingly $E(W_u)$ is the mean unconstrained window size given by

$$E(W_u) = \frac{2+b}{3b} + \sqrt{\frac{8(1-p)}{3bp} + \left(\frac{2+b}{3b}\right)^2} \qquad (2)$$

$\hat{Q}(w)$ is the probability that a loss in a window of size w is due to timeout, calculated with the formula

$$\hat{Q}(w) = \min\left(1, \frac{(1-(1-p)^3)(1+(1-p)^3(1-(1-p)^{w-3}))}{1-(1-p)^w}\right) \qquad (3)$$

Finally, $f(p)$ is a simplifying notation [12]

$$f(p) = 1 + p + 2p^2 + 4p^3 + 8p^4 + 16p^5 + 32p^6 \qquad (4)$$

Thus, the two unknown parameters of the TCP throughput calculation are the Round Trip Time (RTT) and the packet loss probability (p). Since the major part of the RTT is spent as waiting time in the buffers of the network devices, and the packet loss occurs because of saturated buffers or air interface errors, we model the HSDPA system by a queueing network. To reduce complexity, we decided not to involve the micro-behaviour of the TCP flow control into the model. Instead, we consider the TCP traffic as a flow of packets having a constant intensity.

By assuming a constant rate TCP traffic, the RTT and p are calculated using the queueing network model of the system described in detail in the next sections. Once the RTT and p are known, the TCP traffic intensity corresponding to p and RTT can be calculated with the Padhye model. This value is not necessarily equal to the TCP rate assumed initially. In this case the initially assumed TCP rate is adjusted, and throughput calculation is repeated until the equilibrium is reached. The output of the method will be the TCP rate B^* that – when loaded into the queueing network model – results in a p and RTT with which the Padhye model provides the same TCP throughput, thus $B^* = B(p, \mathrm{RTT})$.

We are aware of that several assumptions of the Padhye model are not met (we have multiple connections, different loss pattern and highly varying RTT) but the complexity of HSDPA transmission made it impossible to come up with a TCP model tailored to HSDPA characteristics. Instead, we follow the tagged traffic approach: a tagged TCP flow is taken, the mean network delay and packet loss probability are calculated, and the mean throughput is obtained with the Padhye model assuming that the tagged one is the only traffic in the system (the same approach has been applied in [13]).

3.1 Overview of the calculation algorithm

Throughout the paper we consider a parameter setting given in Table 1. The selected TCP parameters are typical in the recent TCP implementations over various operating systems. The buffer sizes and the maximum number of retransmissions are typical in case of HSDPA. The P_e and P_s parameters are throughput independent, they depend only on the SIR target, modulation and coding.

As described in Section 3, the TCP throughput over HSDPA is calculated as the load (λ_{TCP}) that carried over the network causes a round trip time RTT and packet loss p such that the Padhye formula (1) results in the same throughput that is, $B(p, \mathrm{RTT}) = \lambda_{\mathrm{TCP}}$.

The equilibrium of the load is calculated by a simple interval bisection method summarised in the Algorithm given in Fig. 2. At the beginning of the algorithm, the lowest possible throughput is initialised to $a_1 = 0$ in line 1. The mean TCP throughput cannot be larger than the average air interface throughput, thus the upper limit of the interval bisection is initialised in line 2 to be $E(S_{\mathrm{Node\text{-}B}})$. In each step, the queueing network shown in Fig. 3 is analysed in line 6, the packet loss and mean RTT is calculated. In lines 8 through 11 the upper and lower bounds of the interval are adjusted depending on the relationship between the actual TCP throughput assumption, λ_{TCP} and the throughput calculated with the Padhye formula $\lambda_{\mathrm{PADHYE}} = B(p, \mathrm{RTT})Kf_{\mathrm{T}}/f_{\mathrm{M}}$.

Parameters p and RTT are results from the analysis of the queueing network (Fig. 3). The users are assumed to be identical, the calculation is performed for one selected (tagged) user. Accordingly, the queueing network seen by the tagged user consists of three nodes: the RLC buffer,

$$B(p, \mathrm{RTT}) = \begin{cases} \dfrac{1-p/p + E(W_u) + \hat{Q}(E(W_u))1/1-p}{\mathrm{RTT}(b/2E(W_u)+1) + \hat{Q}(E(W_u))T_0 f(p)/1-p}, & \text{if } E(W_u) < W_{max} \\[2ex] \dfrac{1-p/p + W_{max} + \hat{Q}(E(W_u))1/1-p}{\mathrm{RTT}(b/8W_{max} + 1-p/pW_{max} + 2) + \hat{Q}(W_{max})T_0 f(p)/1-p}, & \text{otherwise} \end{cases} \qquad (1)$$

Table 1 Parameters contained in `sysparam`

Description	Notation	Value
the number of HSDPA users	K	16
the RLC buffer size (PDUs)		1000
transport node buffer size (ATM cells)	L	2000
Node-B buffer size (PDUs)		1000
maximum number of RLC (re)transmissions	R	6
maximum number of HARQ (re)transmissions	M	3
TCP packets acknowledged by one ACK	b	1
TCP timeout interval	T_0	1.5 s
maximum TCP congestion window size	W_{max}	48 KB
block error rate over the air interface	P_{e}	0.01
prob. of two successive erroneous transmissions	P_{s}	0.001
service distribution at the air interface	$\Pr(\hat{S} = k)$	From trace file
TCP packet size	f_{T}	1500 byte
size of MAC-d and RLC PDUs	f_{M}	336 bit
accuracy parameters	ϵ, ϵ'	1
transport link capacity	C	

the transport (ATM) buffer and the MAC-d buffer at the Node-B. This queueing network does not belong to the class of queueing networks for which an exact solution is known, thus a traffic decomposition-based approximate analysis has been developed [14]. The analysis starts with the first queue. In addition to the performance measures of interest, the output process has to be approximated, too. This approximate departure process is the arrival process at the next queue in the network that can be analysed in the same way. The calculation is repeated until the last queuing node is analysed. As the network has feedback traffic (RLC loss is modelled as if the lost PDUs were re-inserted into the RLC buffer), an iterative solution method has to be used. Initially it is assumed that there is no feedback traffic; the whole network is analysed and the feedback traffic (amount of PDUs that must be retransmitted) is calculated. In the next iteration step this feedback traffic is added to the traffic of the first queue. The iterations are repeated until the difference in the results will not exceed the accuracy parameter. The algorithm is summarised in Fig. 4.

3.2 Model of the RLC buffer

The model of the RLC layer (referred to as `solve rlc` in line 4 in Fig. 4 is based on the observation that the service process of the RLC buffer (thus, the arrival process at the RLC PDUs to the transport network) is controlled by the HSDPA flow control. To achieve efficient air interface resource usage, the Node-B grants credits to each flow based on the reported channel quality and the measured average throughput of the flows. At each HSDPA scheduling interval, the MAC-d scheduler will transmit the amount of PDUs defined by the received credits. In this paper we assume that scheduling interval is 10 ms (that is a typical value), thus PDUs are scheduled at each

The TCP Throughput Calculation Algorithm

INPUT: `sysparam`//the system parameters are listed in Table 1
OUTPUT: γ//the TCP throughput
$a_1 = 0$//the lowest possible throughput value
$a_2 = \mathrm{E}\left(S_{\text{Node-B}}\right)$//the upper bound is the average air interface throughput
set λ_{old} to be larger than $\lambda_{\text{TCP}} = \frac{a_1+a_2}{2}$ plus ϵ.
while $|\lambda_{\text{TCP}} - \lambda_{\text{old}}| > \epsilon$ **do** //the loop of the bisection method
  $\lambda_{\text{old}} = \lambda_{\text{TCP}}$
  $p, RTT =$ `QN Analysis` (λ_{TCP})
  $\lambda_{\text{PADHYE}} = \mathrm{B}\left(p, RTT\right) \cdot K f_T / f_M$ //Apply Padhye model as in (1)
  if $\lambda_{\text{PADHYE}} > \lambda_{\text{TCP}}$ **then**
    $a_1 = \lambda_{\text{TCP}}$
  else
    $a_2 = \lambda_{\text{TCP}}$
  end if
  $\lambda_{\text{TCP}} = \frac{a_1+a_2}{2}$
end while
return $\gamma = \lambda f_M$//harmonize units

Figure 2 *TCP throughout calculation algorithm*

IET Softw., 2009, Vol. 3, Iss. 6, pp. 480–494
doi: 10.1049/iet-sen.2009.0006

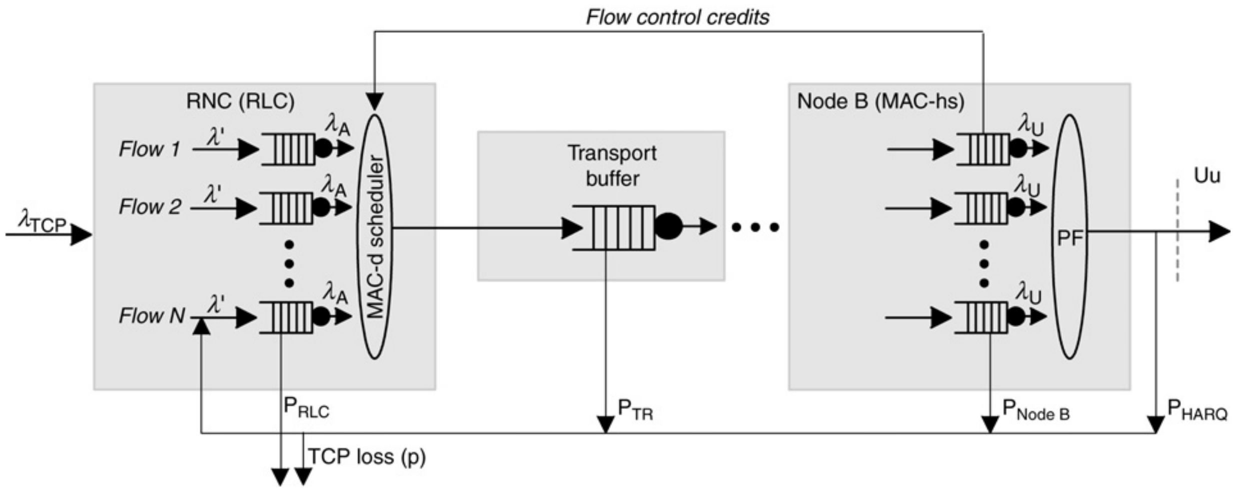


Figure 3 *Overview of the queueing network model of the system*

$\mathrm{TTI}_{\mathrm{RLC}} = 10$ ms. The calculation of the amount of PDUs that can be sent at a given scheduling instance is based on the assumption that the Node-B has perfect knowledge of the air interface conditions, thus the distribution of the number of MAC-d PDUs that can be transmitted over the air-interface at each 2 ms HSDPA TTI is known (see Section 3.4).

Based on this assumption, the number of PDUs the MAC-d scheduler can transfer (given that there are enough PDUs in the buffer) during a 10 ms time slot (denoted by S_{RLC}) is given by

$$S_{\mathrm{RLC}} = \sum_{1}^{5} S_{\text{Node-B}} \tag{5}$$

(Note that here we sum five random variables, thus the distribution of S_{RLC} is the 5 fold convolution of the distribution of $S_{\text{Node-B}}$).

The arrival process at the RLC buffer consists of the traffic at the input of the system (having an intensity of λ_{in}/K) and the PDUs that are retransmitted by the RLC AM entity (this is how RLC losses, denoted by λ_{FB}, are modelled). λ_{FB} is calculated with (38). At this calculation step Poisson traffic with a total arrival rate of λ' is assumed. The Poisson assumption has been used because of its simplicity (it has only a rate parameter) and because it is closed to the basic traffic operations like splitting and superposition.

The `QN Analysis` method

INPUT: λ_{in} //the load generated by the TCP sources
OUTPUT: p, RTT//packet loss and mean round trip time
$\lambda' = \frac{\lambda_{\mathrm{in}}}{K}$//the throughput of the tagged HSDPA user
set λ'_{old} to be larger than λ' plus ϵ'.
while $|\lambda' - \lambda'_{\mathrm{old}}| > \epsilon'$ **do** //loop to find the equilibrium value of λ'
  $(P_{\mathrm{RLC}}, \mathrm{E}(T_{\mathrm{RLC}}), D_{\mathrm{RLC}}) = \texttt{solve rlc}(\lambda')$//see Section 3.2
  $(P_{\mathrm{Tr}}, \mathrm{E}(T_{\mathrm{Tr}}), D_{\mathrm{Tr}}) = \texttt{solve tr}(C, D_{\mathrm{RLC}})$//see Section 3.3
  $(P_{\text{Node-B}}, \mathrm{E}(T_{\text{Node-B}}), \lambda_U) = \texttt{solve node-b}(S, D_{\mathrm{Tr}})$//see Section 3.4
  $p_L \Leftarrow (P_{\mathrm{Tr}}, P_{\text{Node-B}}, P_{\mathrm{HARQ}})$//the loopback probability given by (36)
  $\hat{p} = \frac{\sum_{k=1}^{R}(1-p_L)^{k-1}p_L}{\sum_{k=1}^{R+1}(1-p_L)^{k-1}p_L}$//the probability that the PDU is resent by RLC
  $\lambda' = \frac{\lambda_{\mathrm{in}}}{K} + \hat{p} \cdot p_L \cdot \lambda_A$
  $\lambda'_{\mathrm{old}} = \lambda'$
end while
$(D_u, D_s) \Leftarrow (P_{\mathrm{Tr}}, P_{\text{Node-B}}, P_{\mathrm{HARQ}}, \mathrm{E}(T_{\mathrm{RLC}}), \mathrm{E}(T_{\mathrm{Tr}}), \mathrm{E}(T_{\text{Node-B}}))$//(40)
$RTT = D_{\mathrm{UL}} + \sum_{k=1}^{R} \frac{p_L^{k-1}(1-p_L)}{1-p_L^R}((k-1)D_u + D_s)$//the RTT as in (41)
$p = 1 - \frac{\lambda_U}{\lambda}$//the TCP loss probability given in (42)

Figure 4 *QN Analysis method*

Thus we have

$$\lambda' = \frac{\lambda_{\text{in}}}{K} + \lambda_{\text{FB}} \tag{6}$$

The distribution of the number of packets entering the RLC buffer in a 10 ms interval is calculated as follows

$$\Pr(A_{\text{RLC}} = k) = \frac{(\lambda' \text{TTI}_{\text{RLC}})^k}{k!} \mathrm{e}^{-\lambda' \text{TTI}_{\text{RLC}}}, \quad k = 0, 1, 2, \ldots \tag{7}$$

The distribution is truncated at N such that the probability of the cut-off part of the distribution is reasonably small ($<10^{-6}$).

The queue length evolution embedded at TTI_{RLC} long time slots is then modelled by a discrete time Markov chain (DTMC) according to the following evolution equation

$$X_{n+1} = (\min(X_n + A_{n+1}, L) - S_{n+1})^+ \tag{8}$$

where X_{n+1} is the queue length, A_{n+1} is the number of arrivals and S_{n+1} is the number of packets served in the $n+1$st time slot, L is the buffer length, $(\cdot)^+$ denotes $\max(0, \cdot)$.

The $i\,j$th element of the transition probability matrix ($\boldsymbol{P}$) of the DTMC is given by (L denotes the length of the RLC buffer, N is the support of the truncated Poisson distribution generating the arrivals) (see (9))

In the first case, the queue level is so low that there can be no loss, thus the transition probability equals the probability that there were $j - i$ more packets served than arrived. In the second case, the first (second) term corresponds to arrival sizes without (with) loss, respectively.

The steady-state solution (π) of the DTMC is given by the solution of the linear equation system

$$\pi \boldsymbol{P} = \pi$$
$$\pi \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = 1 \tag{10}$$

Having the steady-state solution, the loss probability at the RLC buffer is calculated as the ratio of the mean number of lost and of the mean number of arrived PDUs during a $\text{TTI}_{\text{RLC}} = 10\,\text{ms}$ time slot

$$P_{\text{RLC}} = \frac{\sum_{i=0}^{L} \pi_i \sum_{k=0}^{N} \max(0, i + k - L) \Pr(A_{\text{RLC}} = k)}{\sum_{i=0}^{L} \pi_i \sum_{k=0}^{N} k \Pr(A_{\text{RLC}} = k)} \tag{11}$$

The system time of the PDUs in the RLC buffer is calculated using Little's theorem

$$E(T_{\text{RLC}}) = \frac{E(X_{\text{RLC}})}{(1 - P_{\text{RLC}}) E(A_{\text{RLC}})} \text{TTI}_{\text{RLC}} + \frac{1}{2} \text{TTI}_{\text{RLC}} \tag{12}$$

where $E(X_{\text{RLC}})$ is the mean queue length. Since this is a discrete time model but arrivals can happen in continuous time, the model does not differentiate between arrivals at the beginning of the scheduling interval and at the end of it, that is, as they would not have different system times. Assuming that the arrival instants are uniformly distributed over the scheduling interval, the system time computed from the embedded DTMC is increased by the half of the interval.

During the analysis of the queueing network, the departure process from the RLC buffers has to be calculated as this is the arrival process at the transport buffer. We assume that the departures are independent and identically distributed (i.i.d.), with the distribution of the number of departing packets in a TTI_{RLC} interval

$$p_{ij} = \begin{cases} \sum_{k=0}^{\infty} \Pr(A_{\text{RLC}} = k) \sum_{\ell=0}^{\infty} \Pr(S_{\text{RLC}} = i + k + \ell), & j = 0,\ i < L - N \\ \sum_{k=0}^{L-i} \Pr(A_{\text{RLC}} = k) \sum_{\ell=0}^{\infty} \Pr(S_{\text{RLC}} = i + k + \ell) & \\ + \sum_{\ell=0}^{\infty} \Pr(S_{\text{RLC}} = L + \ell),\ \sum_{k=L-i+1}^{N} \Pr(A_{\text{RLC}} = k), & j = 0,\ i \geq L - N \\ \sum_{k=0}^{\infty} \Pr(A_{\text{RLC}} = k) \Pr(S_{\text{RLC}} = i - j + k), & j > 0,\ i < L - N \\ \sum_{k=0}^{L-i} \Pr(A_{\text{RLC}} = k) \Pr(S_{\text{RLC}} = i - j + k) & \\ \quad + \Pr(S_{\text{RLC}} = L - j) \sum_{k=L-i+1}^{N} \Pr(A_{\text{RLC}} = k) & j > 0,\ i \geq L - N \end{cases} \tag{9}$$

IET Softw., 2009, Vol. 3, Iss. 6, pp. 480–494
doi: 10.1049/iet-sen.2009.0006

computed by

$$\Pr(D_{\mathrm{RLC}}=k)=\sum_{i=0}^{L}\pi_i\sum_{j=k+1-i}^{\infty}\Pr(A_{\mathrm{RLC}}=j)\Pr(S_{\mathrm{RLC}}=k)$$
$$+\sum_{i=0}^{L}\pi_i\Pr(A_{\mathrm{RLC}}=k-i)\sum_{j=k}^{\infty}\Pr(S_{\mathrm{RLC}}=j) \quad (13)$$

This expression consists of two terms: the first corresponds to the case when there are enough packets in the buffer, the number of departing packets is determined by the number of packets the server can serve whereas in the second term the server could serve more packets than the buffer content.

3.3 Model of the transport buffer

In this paper we consider an AAL2/ATM-based transport network (the transport link buffer model and its solution is referred to in line 5 of Fig. 4). The AAL2 layer is multiplexing the user connections into one constant bit rate (CBR) VCC, with capacity C.

The ATM switch works in continuous time in contrast with the MAC-d and PF schedulers that are working in time slotted manner. In order to avoid mixing the continuous and discrete models, we decided to apply a discrete time model for the transport buffer as well. The RLC buffer is scheduled with $\mathrm{TTI}_{\mathrm{RLC}}=10\,\mathrm{ms}$ transmission interval and the PF scheduler in the Node-B is forwarding PDUs with a $\mathrm{TTI}_{\text{Node-B}}=2\,\mathrm{ms}$. The selected time slot for the transport buffer is the minimum of these two, for example, $\mathrm{TTI}_{\mathrm{Tr}}=2\,\mathrm{ms}$ is used to approximate the transport buffer mainly because this value allows finer resolution in time than a model with 10 ms interval. Another assumption is that in the model the transport buffer stores and transmits RLC PDUs instead of ATM cells. Since the RLC PDUs are the 'data units' in other parts of the network, using the same data unit in the transport buffer simplifies the calculation significantly.

The distribution of the number of arrivals in a time slot is derived from the distribution of the number of departures from the RLC layer (D_{RLC}). The departure process of the RLC corresponds to a 10 ms $\mathrm{TTI}_{\mathrm{RLC}}$, whereas the transport buffer model has a 2 ms $\mathrm{TTI}_{\mathrm{Tr}}$. Thus, as a first step a conversion has to be applied between the MAC-d scheduling interval and the transport time slot, having a departure distribution from the RLC layer in a five-time longer $\mathrm{TTI}_{\mathrm{RLC}}$. Assuming that the i PDUs arrived in a $\mathrm{TTI}_{\mathrm{RLC}}$ are distributed randomly over the 2 ms long $\mathrm{TTI}_{\mathrm{Tr}}$, we can define the following distribution for the number of packets arrived in $\mathrm{TTI}_{\mathrm{Tr}}$

$$\Pr(D_{2\,\mathrm{ms}_{\mathrm{tr}}}=k)=\sum_{i=k}^{\infty}\Pr(D_{\mathrm{RLC}}=i)\binom{i}{k}\left(\frac{1}{5}\right)^{k}\left(1-\frac{1}{5}\right)^{i-k} \quad (14)$$

When calculating the distribution of the number of arrivals to the transport buffer, the whole traffic aggregate has to be considered since each user connection is multiplexed into one VCC. Thus, with denoting the number of HSDPA users by K we have

$$A_{\mathrm{Tr}}=\sum_{1}^{K}D_{2\,\mathrm{ms}_{\mathrm{tr}}} \quad (15)$$

(Note that here we sum K random variables, thus the distribution of A_{Tr} is calculated as the 5- fold convolution of the distribution of $D_{2\,\mathrm{ms}_{\mathrm{tr}}}$.)

The service time of the RLC PDUs in the transport buffer is calculated as

$$D=\frac{\text{RLC packet size with overheads}}{C} \quad (16)$$

The transport overheads are considered with the following formula

$$\text{RLC packet size with overheads}$$
$$=f_{\mathrm{M}}\underbrace{\left(\frac{424}{376}\frac{f_{\mathrm{M}}+24}{f_{\mathrm{M}}}\frac{E(D_{\mathrm{RLC}})f_M+72}{E(D_{\mathrm{RLC}})f_M}\right)}_{\text{overhead}} \quad (17)$$

The size of an ATM cell is 424 bits. The overhead consists of the ATM header (40 bits) plus the 8-bit long CPS PDU start field belonging to the AAL2 protocol; this gives an overhead ratio of 424/376. Additionally, there is a 24-bit long CPS packet header per an RLC PDU ($f_{\mathrm{M}}+24/f_{\mathrm{M}}$) and finally the 72-bit long HS-DSCH FP frame header that carries $E(D_{\mathrm{RLC}})$ RLC packets in an average.

In our system having $\mathrm{TTI}_{\mathrm{Tr}}$ long time slots the number of PDUs served in a time slot is $\mathrm{TTI}_{\mathrm{Tr}}/D$. The problem is that this is a real number – but since this system is discrete, only an integer amount of PDUs can be served in a time slot. To overcome this problem we obtain an average number of $\mathrm{TTI}_{\mathrm{Tr}}/D$ served PDUs by serving either $F=\lfloor\mathrm{TTI}_{\mathrm{Tr}}/D\rfloor$, or $F+1$ PDUs according to a random choice given by the fractional part of $\mathrm{TTI}_{\mathrm{Tr}}/D$ as follows

$$\Pr(S_{\mathrm{Tr}}=F)=1-\left(\frac{\mathrm{TTI}_{\mathrm{Tr}}}{D}-F\right) \quad (18)$$

$$\Pr(S_{\mathrm{Tr}}=F+1)=\frac{\mathrm{TTI}_{\mathrm{Tr}}}{D}-F \quad (19)$$

Of course, all PDUs can be served if the number of PDUs waiting in the buffer is less than F.

The queue length can be modelled by a DTMC similar to the one we applied for the RLC buffer, that is

$$X_{n+1} = (\min(X_n + A_{n+1}, L) - S_{n+1})^+ \tag{20}$$

where X_{n+1} is the queue length, A_{n+1} is the number of arrivals and S_{n+1} is the number of PDUs served in the $n+1$st time slot, L is the buffer length.

Based on the distribution of the number of arrivals and served PDUs, we can create the transition probability matrix of the DTMC such that the $i\,j$th element will be calculated in the same way as in the case of the RLC buffer (see (21))

The computation of the loss probability is similar to the one applied at the RLC modelling

$$P_{\mathrm{Tr}} = \frac{\sum_{i=0}^{L-F} \pi_i \sum_{j=0}^{N} \max(0, i+j-L)\Pr(A_{\mathrm{Tr}} = j)}{\sum_{i=0}^{L-F} \pi_i \sum_{j=0}^{N} j\Pr(A_{\mathrm{Tr}} = j)} \tag{22}$$

The numerator is the expected number of lost PDUs and the denominator is the expected number of PDUs received correctly. Here we also used the truncated and renormalised arrival distribution with support $[0, N]$ and the steady-state solution of the DTMC (π).

The system time of the PDUs in the transport buffer is calculated based on Little's theorem as (see (12))

$$E(T_{\mathrm{Tr}}) = \frac{E(X_{\mathrm{Tr}})}{(1-P_{\mathrm{Tr}})E(A_{\mathrm{Tr}})}\mathrm{TTI}_{\mathrm{Tr}} + \frac{1}{2}\mathrm{TTI}_{\mathrm{Tr}} \tag{23}$$

The calculation of the departure process is similar to that of the same parameter in case of the RLC buffer.

$$\begin{aligned}\Pr(D_{\mathrm{Tr}} = k) = {} & \sum_{i=0}^{L-F} \pi_i \sum_{j=k+1-i}^{\infty} \Pr(A_{\mathrm{Tr}} = j)\Pr(S_{\mathrm{Tr}} = k) \\ & + \sum_{i=0}^{L-F} \pi_i \Pr(A_{\mathrm{Tr}} = k - i) \sum_{j=k}^{\infty} \Pr(S_{\mathrm{Tr}} = j)\end{aligned} \tag{24}$$

3.4 Model of the MAC-hs buffers

In this paper it is assumed that the MAC-hs buffers are scheduled by a proportional fair algorithm. This scheduler makes the scheduling decisions based on the instantaneous channel quality and the average throughput of the users with the scope to achieve high level of resource usage and in the same time to provide high level of fairness to the users. The scheduler selects one user for a transmission at each scheduling instance (at every $\mathrm{TTI}_{\mathrm{Node\text{-}B}} = 2\,\mathrm{ms}$). The reported CQI defines the modulation and coding scheme and thus the number of MAC-d PDUs that can be transmitted during a TTI. Since the channel conditions can change quickly, temporary traffic overload can occur in the Node-B. The arriving PDUs are stored in the MAC-hs buffers (there is a separate buffer for each flow).

The modelling of the HSDPA air interface model is out of the scope of this paper. Instead, the MATLAB-based tool of the Eurane project (see [15]) has been used in order to obtain the distribution of the number of MAC-d PDUs that can be transmitted in a TTI ($P(\hat{S} = k)$). This distribution has been generated by assuming saturated buffers without taking the impact of HARQ into consideration [16].

To obtain the service process of the MAC-hs buffer first the effect of HARQ is included in the model. According to [5, 6] the probability of properly decoding the packet at the user side and thus the probability of the error-free transmission after j trials is

$$P_j = \begin{cases} 1 - P_{\mathrm{e}}, & j = 1 \\ P_{\mathrm{e}}{}^{j-1} P_{\mathrm{s}}{}^{j-2}(1 - P_{\mathrm{e}}P_{\mathrm{s}}), & j > 1 \end{cases} \tag{25}$$

The values of P_{e} and P_{s} are listed in Table 1. A MAC-d PDU can leave the queue either when it has been transmitted successfully or when the transmission fails if the maximum number of retransmissions is reached. Considering that the maximal number of trials is M, the expected number of time slots before a PDU can leave the queue, and thus the expected number of retransmissions before successful transmission or failure, is

$$E(H) = \sum_{j=1}^{M} jP_j + M\left(1 - \sum_{j=1}^{M} P_j\right) \tag{26}$$

Thus the fraction of time slots occupied by retransmissions in which no MAC-d PDU can leave is

$$P_{\mathrm{tl}} = 1 - \frac{1}{E(H)} \tag{27}$$

$$p_{ij} = \begin{cases} \sum_{k=0}^{\infty} \Pr(A_{\mathrm{Tr}} = k) \sum_{\ell=0}^{\infty} \Pr(S_{\mathrm{Tr}} = i + k + \ell), & j = 0,\ i < L - (N - F) \\ \sum_{k=0}^{L-i} \Pr(A_{\mathrm{Tr}} = k) \sum_{\ell=0}^{\infty} \Pr(S_{\mathrm{Tr}} = i + k + \ell) & \\ \quad + \sum_{\ell=0}^{\infty} \Pr(S_{\mathrm{Tr}} = L + \ell) \sum_{k=L-i+1}^{N} \Pr(A_{\mathrm{Tr}} = k), & j = 0,\ i \geq L - (N - F) \\ \sum_{k=0}^{\infty} \Pr(A_{\mathrm{Tr}} = k)\Pr(S_{\mathrm{Tr}} = i - j + k), & j > 0,\ i < L - (N - F) \\ \sum_{k=0}^{L-i} \Pr(A_{\mathrm{Tr}} = k)\Pr(S_{\mathrm{Tr}} = i - j + k) & \\ \quad + \Pr(S_{\mathrm{Tr}} = L - j) \sum_{k=L-i+1}^{N} \Pr(A_{\mathrm{Tr}} = k), & j > 0,\ i \geq L - (N - F) \end{cases} \tag{21}$$

IET Softw., 2009, Vol. 3, Iss. 6, pp. 480–494
doi: 10.1049/iet-sen.2009.0006

Finally, the distribution of the number of MAC-d PDUs that can be transmitted in a TTI taking the retransmissions also into consideration is (we assume that retransmission happen with probability $P_{\rm tl}$)

$$\Pr(S_{\text{Node-B}} = k) = \begin{cases} (1 - P_{\rm tl})\Pr(\hat{S} = k) + P_{\rm tl}, & k = 0 \\ (1 - P_{\rm tl})\Pr(\hat{S} = k), & k \neq 0 \end{cases} \tag{28}$$

The distribution of the number of arrivals at the MAC-hs buffer is calculated by assuming that the packets arriving from the transport network are directed to the buffer of the tagged user according to a random choice with probability $1/K$; resulting in the following binomial distribution

$$\Pr(A_{\text{Node-B}} = k) = \sum_{i=k}^{\infty} \Pr(D_{\rm Tr} = i)\binom{i}{k}\left(\frac{1}{K}\right)^k\left(1 - \frac{1}{K}\right)^{i-k} \tag{29}$$

Contrary to the other two nodes the queue length evolution of the MAC-hs buffer is

$$X_{n+1} = \min\left((X_n - S_{n+1})^+ + A_{n+1}, L\right) \tag{30}$$

This means that only those MAC-d PDUs can be served by the PF scheduler that have arrived before the beginning of TTI. The i jth element of the transition probability matrix is (see (31))

After the computation of the steady-state solution, the loss probability is calculated as the ratio of the lost and arrived PDUs in a TTI as

$$P_{\text{Node-B}} = \frac{\sum_{i=0}^{L} \pi_i \sum_{j=0}^{\infty} \max(0, i + j - L)\Pr(A_{\text{Node-B}} = j)}{\sum_{i=0}^{L} \pi_i \sum_{j=0}^{\infty} j\Pr(A_{\text{Node-B}} = j)} \tag{32}$$

$P_{\text{Node-B}}$ is the loss probability of PDUs because of buffer saturation and tail drop at the Node-B. However, at the Node-B the buffer saturation is not the only event that leads to packet loss. If the air interface quality is bad, and the HARQ mechanism fails, the MAC-hs discards the PDU from the corresponding HARQ register and the retransmission of the PDUs falls back to the RLC layer if the maximal number of retransmissions (M) has been reached. The probability of such events is denoted by $P_{\rm HARQ}$ and computed as

$$P_{\rm HARQ} = 1 - \sum_{j=1}^{M} P_j \tag{33}$$

The system time of the MAC-hs buffer is calculated using Little's theorem as

$$E(T_{\text{Node-B}}) = \frac{E(X_{\text{Node-B}})}{(1 - P_{\text{Node-B}})E(A_{\text{Node-B}})}\text{TTI}_{\text{Node-B}} + \frac{1}{2}\text{TTI}_{\text{Node-B}} \tag{34}$$

where $E(X)$ is the mean queue length, and the addition of the extra time of half-$\text{TTI}_{\text{Node-B}}$ in the second term has the same explanation as in case of the RLC and transport network models.

For the queueing network analysis, the departure intensity of the Node-B buffer is needed. The number of MAC-d PDUs per $\text{TTI}_{\text{Node-B}}$ equals the minimum of the number of packets in the buffer and the number of packets that can be served. This gives

$$\lambda_U = \frac{1}{\text{TTI}_{\text{Node-B}}} \sum_{i=0}^{L} \pi_i \sum_{k=0}^{\infty} \Pr(S_{\text{Node-B}} = k)\min(i, k) \tag{35}$$

3.5 Feedback link

In our queueing model the PDUs lost at different parts of the network are considered as if they were entering the RLC buffer again for repeated transmission. The feedback link in Fig. 2 'collects' these lost packets. In this section we calculate the traffic intensity on the feedback link. This traffic (with Poisson assumption [14]) is added to the traffic entering the network during the analysis of the RLC model.

As a first step the probability of a PDU loss (due to any reason) in the network after leaving the RLC buffer is calculated. This probability is denoted by $p_{\rm L}$ and computed by

$$p_{\rm L} = P_{\rm Tr} + (1 - P_{\rm Tr})P_{\text{Node-B}} + (1 - P_{\rm Tr}) \times (1 - P_{\text{Node-B}})P_{\rm HARQ} \tag{36}$$

It can happen that a retransmitted PDU is lost. After a given number of RLC-level retransmission attempts (R) that equals the maximum number of RLC retransmissions the PDU is discarded and loss is detected by the TCP flow control. In this case this PDU does not enter the RLC buffer again

$$p_{ij} = \begin{cases} \sum_{k=0}^{j-1} \Pr(A_{\text{Node-B}} = k)\Pr(S_{\text{Node-B}} = i - j + k) \\ \quad + \Pr(A_{\text{Node-B}} = j)\sum_{k=i}^{\infty} \Pr(S_{\text{Node-B}} = k), & i < k_m \\ \sum_{k=0}^{\infty} \Pr(A_{\text{Node-B}} = k)\Pr(S_{\text{Node-B}} = i - j + k), & i \geq k_m \end{cases} \tag{31}$$

(as long as the higher layer entity does not resend it). The probability that PDU loss did not reach the maximum number of retransmission attempts thus increasing the load of the RLC buffer is calculated with

$$\hat{p} = \frac{\sum_{k=1}^{R} (1-p_{\mathrm{L}})^{k-1} p_{\mathrm{L}}}{\sum_{k=1}^{R+1} (1-p_{\mathrm{L}})^{k-1} p_{\mathrm{L}}} \tag{37}$$

(we assumed truncated geometrical distribution for the distribution of the number of retransmissions).

With the above considerations the traffic of the feedback link is computed by

$$\lambda_{\mathrm{FB}} = \hat{p} p_L \lambda_A \tag{38}$$

where λ_A denotes the mean departure rate of the RLC buffer.

3.6 TCP-level packet loss and the RTT

In this section we describe the calculation of the TCP-level performance measures based on the buffer-wise performance measures [given by (11), (12), (22), (23), (32), (33) and (34)].

The delay of one packet assuming that it has not been lost in the system is composed by the delay gathered in the RLC buffer $E(T_{\mathrm{RLC}})$ in the transport buffer $E(T_{\mathrm{Tr}})$ and in the buffer of the Node-B $E(T_{\text{Node-B}})$. Thus, if a packet starting in the RNC buffer arrives at the user equipment successfully, its delay is

$$D_{\mathrm{s}} = E(T_{\mathrm{RLC}}) + E(T_{\mathrm{Tr}}) + E(T_{\text{Node-B}}) \tag{39}$$

If it has been lost somewhere, it is put back into the RNC buffer for retransmission, but before it gets lost it spends some time in the network that counts into RTT, denoted by D_{u}. This delay depends on in which queue the loss occurred. With probability $P_{\mathrm{Tr}}/p_{\mathrm{L}}$ the packet got lost in the transport network buffer (given that is has been lost during the transmission trial) and before getting lost it spent $E(T_{\mathrm{RLC}})$ time in the network. If it has been successfully transmitted on the transport network but the loss occurred because of the saturated buffer at the Node-B (that happens with probability $(1-P_{\mathrm{Tr}})P_{\text{Node-B}}/p_{\mathrm{L}}$), it spent $E(T_{\mathrm{RLC}}) + E(T_{\mathrm{Tr}})$ time in the network before getting lost. Finally, with probability $(1-P_{\mathrm{Tr}})$ $(1-P_{\text{Node-B}})P_{\mathrm{HARQ}}/p_{\mathrm{L}}$ it is possible that the packet passed through the transport buffer and the buffer at Node-B, but it got lost on the air interface because of the bad channel conditions. In this case the delay before getting lost is $E(T_{\mathrm{RLC}}) + E(T_{\mathrm{Tr}}) + E(T_{\text{Node-B}})$. Thus, the delay of an unsuccessful transmission trial is as follows

$$\begin{aligned} D_{\mathrm{u}} = [&P_{\mathrm{Tr}}E(T_{\mathrm{RLC}}) + (1-P_{\mathrm{Tr}})P_{\text{Node-B}}(E(T_{\mathrm{RLC}}) + E(T_{\mathrm{Tr}})) \\ &+ (1-P_{\mathrm{Tr}})(1-P_{\text{Node-B}})P_{\mathrm{HARQ}} \\ &\times (E(T_{\mathrm{RLC}}) + E(T_{\mathrm{Tr}}) + E(T_{\text{Node-B}}))]/p_{\mathrm{L}} \end{aligned} \tag{40}$$

If a packet has been transmitted k times before successfully received, the mean RTT can be calculated as the sum of the mean delays of $k-1$ unsuccessful trials and one times the delay of a successful transmission. Using the geometric distribution assumption for the number of retransmission attempts again, we have

$$\mathrm{RTT} = D_{\mathrm{UL}} + \sum_{k=1}^{R} \frac{p_{\mathrm{L}}^{k-1}(1-p_{\mathrm{L}})}{1-p_{\mathrm{L}}^{R}}((k-1)D_{\mathrm{u}} + D_{\mathrm{s}}) \tag{41}$$

where D_{UL} denotes the mean delay in uplink direction, which is considered to be constant as the UTRAN is typically not congested in uplink direction.

The loss at the TCP layer is simply calculated by one minus the ratio of the traffic entering and leaving the system

$$p = 1 - \frac{\lambda_{\mathrm{U}}}{\lambda} \tag{42}$$

4 Numerical results

In this section, we study several performance and efficiency aspects of the HSDPA system with the presented analytical model.

4.1 Application for transport network dimensioning

One possible application of the presented model is the dimensioning of the transport network. The numerical results of this section are compared to simulation results to evaluate the accuracy of the analysis method.

An NS-2 simulation scenario has been created based on a topology consisting of one RNC and one Node-B. It is assumed that there is only one MAC-d flow and one priority queue per HSDPA user. The scheduler is a proportional fair scheduler. The number of HARQ processes is six, the maximum number of MAC-hs retransmissions is three, whereas the maximum number of RLC retransmissions is six. The number of HS-DSCH codes per cell is five; code multiplexing is not implemented. HSDPA users are connected to the Node-B via HS-DSCH in downlink and via DCH in uplink. Each HSDPA UE is of category 5/6. The Iub interface and the user plane of the radio layer protocols (MAC-d, MAC-hs, RLC, PDCP) are implemented in detail. The transport network of the Iub consists of one CBR VCC. HSDPA

IET Softw., 2009, Vol. 3, Iss. 6, pp. 480–494
doi: 10.1049/iet-sen.2009.0006

users are originating file (FTP) downloads from servers located on the Internet. The transport protocol was TCP Reno; the maximum advertised window size was 48 kbytes; the maximum TCP/IP packet size was set to 1500 bytes. The HSDPA UE reports the observed channel quality (CQI) to the Node-B. Based on this, the amount of data to be sent to the UE is defined. The radio channel condition is simulated separately for each UE. Negative – when the signal-to-noise ratio (SNR) is below the required threshold – or positive acknowledgement is generated upon reception of a MAC-hs frame. The CQI estimation error is modelled with a constant delay of 6 ms. Users are modelled with ITU-T Pedestrian B model, velocity 3 km/h and Vehicular B model, velocity 30 km/h, assuming that chase-combining is implemented in the UEs. The initial distance of the users from the Node-B was set to 500 m. The SNR is calculated considering the following: distance loss according to the Okumara–Hata model for urban cell, with a base station antenna height of 30 m, a mobile antenna height of 1.5 m and a carrier frequency of 1950 MHz [17]; multi-path (fast) fading; Rake receiver assuming that channel estimation is ideal and the power levels of all paths are known; shadow (slow) fading (log-normal distribution correlated in time [18]); constant Node-B antenna gain constant (17 dBi); inter-cell interference (−70 dBm) and intra-cell interference (30 dBm).

In this simulation tool there was no possibility to obtain confidence intervals for the results. The only way we could control the accuracy was adjusting the simulation time. First we ran the simulation for a very long time and took the result as a reference. Then we decreased the simulation time successively and found that by simulating 600 s the difference in the results increased to 5%. This setting (600 s) of simulation time has been used in all the simulation runs in this section. To demonstrate the confidence of the simulation results, we started the simulations with five different random seeds. The simulation plots depict the maximum, minimum and average of the results.

In the analytical model two kinds of air interface trace files have been generated with the MATLAB scripts of the Eurane project (see [15]) with parameters defined in Table 2.

Next, the distribution of the number of RLC PDUs that can be transmitted by the Node-B (denoted by $\hat{S}$ in Section 3.4 and also in Table 1) is extracted and the analysis method is executed at several link capacity settings (Fig. 5).

Figs. 5*a* and *b* confirm that the largest error of the approximation is around 10% (it is the largest absolute difference between analysis and simulation results relative to the simulation results). The most important application of analytical throughput computation methods like the one presented in this paper is the transport link dimensioning. During the link dimensioning, the mean throughput (Fig. 5) is calculated and the optimal link capacity is usually selected to be the intersection of the linear asymptotes of the curve before and after the knee point. Above this point the increase of link capacity does not introduce a significant increase in the TCP throughput, whereas below this point the air interface can be underutilised. The optimal link capacity determined by the analysis and simulation is about the same in the pedestrian case (2200 kbps) and the difference in the simulation and analysis results is around 10% in the vehicular case (1500 against 1650 kbps). Another possibility to support finding

Table 2 Parameters of the air interface profile

	Case 1	Case 2
profile	Ped-B	Veh-B
speed	3 km/h	30 km/h
distance	500 m	500 m
trace length	900 s	900 s

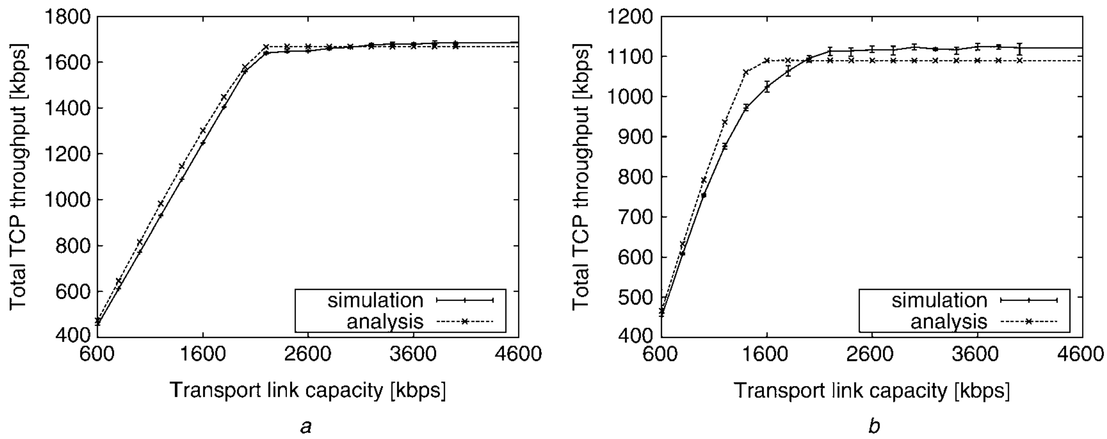


Figure 5 *Comparison of the analysis and simulation results*

a 16 Pedestrian B users
b 16 Vehicular B users

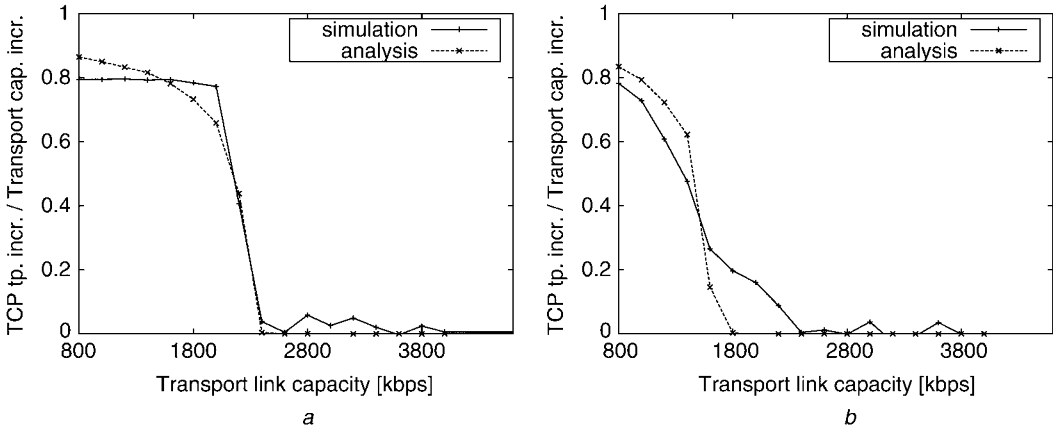


Figure 6 *The ratio of TCP throughput and transport link capacity increments*

a 16 Pedestrian B users
b 16 Vehicular B users

Table 3 Execution times of the analysis and simulation against the capacity of the transport link

Capacity, kbps	Analysis time, s	Simulation time, s
600	57	267
1200	62	419
1800	67	525
2400	66	559
3000	66	554
3600	66	528

the optimal link capacity is to investigate the resulting TCP increment because of the link capacity increment, shown in Fig. 6 (note that it equals the derivative of Fig. 5). Based on these figures the network planner has an overview on how much the HSDPA users gain in download speed with increasing investment into transport network capacity.

(We note that the simulation results are less accurate in the vehicular case since the SNR fluctuation at the air interface is much more pronounced. The results of the multiple simulation runs started with different random seeds were practically equal for the pedestrian case.)

Table 3 depicts the execution times of the analysis method and of the simulation. It confirms that our method can be used for transport link dimensioning with a much lower computational effort compared to simulations.

4.2 Analysis of the RLC retransmissions

In HSDPA the retransmissions are still handled by RLC if the maximum allowed number of MAC-hs retransmissions is exceeded or there are packet drops on the transport network. The RLC retransmissions increase the RTT of the data connections using HSDPA service. This causes that the TCP flow control is not able to detect and resolve the congestion situation on the Iub interface. As a result, the TCP notices the congestion only upon Timeout or when finally the RLC

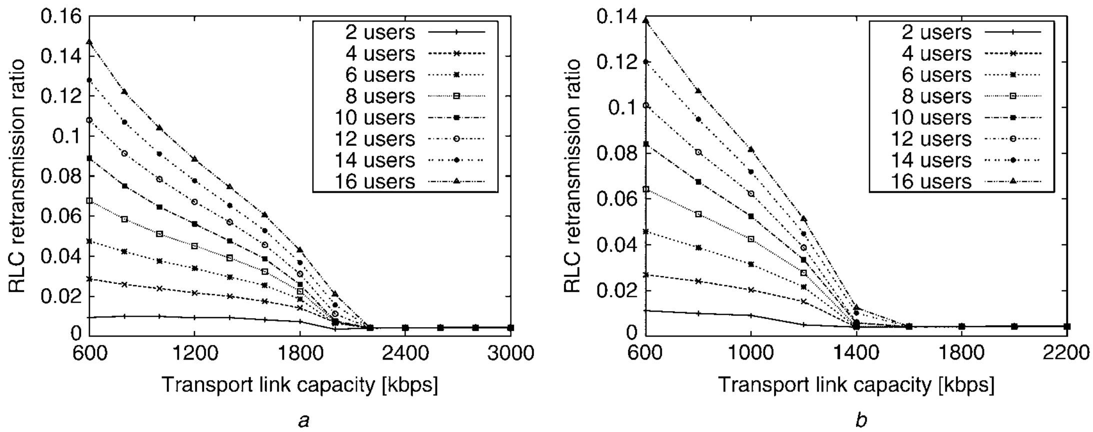
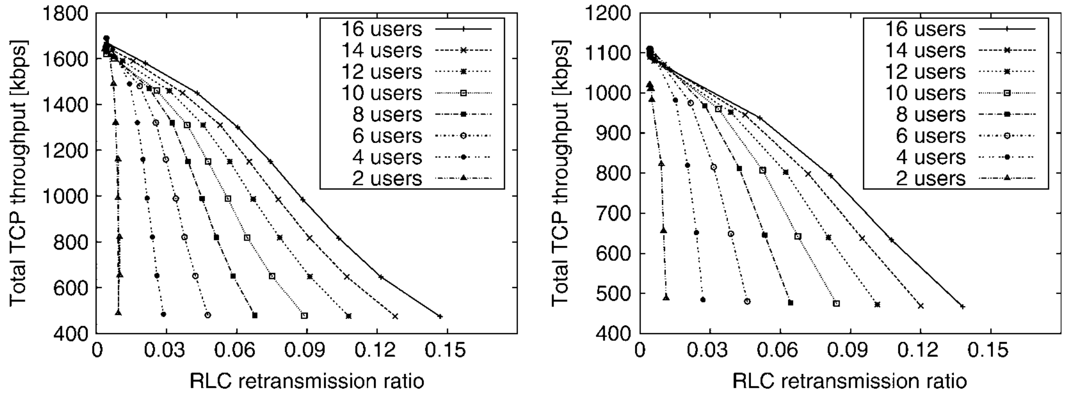


Figure 7 *RLC retransmission ratio as the function of transport link capacity*

a Pedestrian B users
b Vehicular B users

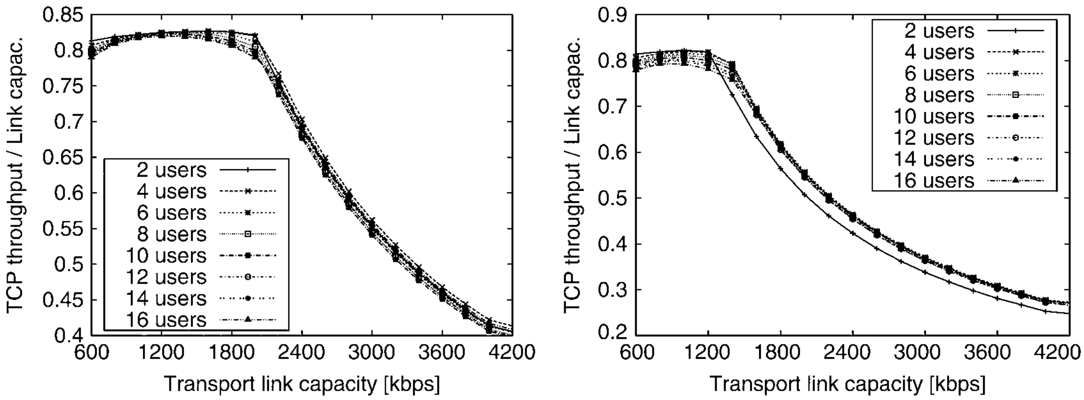
doi: 10.1049/iet-sen.2009.0006



a *b*

Figure 8 *TCP throughput as the function of RLC retransmission ratio*

a Pedestrian B users
b Vehicular B users



a *b*

Figure 9 *Ratio of TCP throughput and link capacity as the function of the link capacity*

a Pedestrian B users
b Vehicular B users

discards the packets that has reached the maximum number of retransmissions.

In this section we investigate how often the RLC ARQ is used during the transmission.

It can be seen in Fig. 7 that in both case of user types the RLC retransmissions occur much more frequently when the transport network capacity is low. In this case the loss at the transport buffer causes an RLC retransmission that actually increases the load of the transport buffer even more. At a given number of RLC retransmissions, the loss is reported at the network layer causing the TCP to decrease its transmission rate. Still, the ratio of RLC retransmissions, appear as a useless extra load that decreases the efficiency of the system. Fig. 7*a* and *b* aid to select an optimal transport link capacity by which the ratio of RLC retransmissions is negligible. These optimal capacities are around 2300 kbps in the case of pedestrian users and it is around 1700 kbps in the case of vehicular users. Note that these values coincide with the ones chosen by the dimensioning approach suggested in the previous section.

In Fig. 8 the TCP throughput is depicted as a function of the RLC retransmissions. As expected, higher TCP throughput corresponds to lower ratio of RLC retransmitted traffic.

4.3 The utilisation of the transport link

An interesting quantity to investigate is the ratio of the TCP throughput and the transport link capacity, and thus the ratio of the link capacity that is used to transmit useful data. The results are depicted in Fig. 9. As long as the transport link is the bottleneck of the system, the transport link utilisation is the largest and constant. Note that the largest utilisation value on the plots is 0.83 because of the protocol overheads dominated by the ATM header and the AAL2 packet header. As it can be seen on both sub-figures in Fig. 9, the plots are not completely horizontal in the lower transport link capacity range, the utilisation decreases slightly as the capacity is getting extremely low. The reason is that the RLC retransmission traffic decreases the TCP performance significantly at such low-link capacities.

As the bottleneck moves towards the air interface, the transport link utilisation decreases – as expected. In this case the breaking point does not coincide with the optimal transport link capacities determined in the previous sections. According to the results, the transport link utilisation is around 75% of the optimal link capacity both in the pedestrian and vehicular cases. Furthermore, the results are practically independent of the number of TCP flows in the system, which can be explained by the TCP behaviour which tries to utilise the available bandwidth.

5 Conclusion

In this paper we described an approximate performance model for the TCP throughput over HSDPA. We identified the relevant congestion points in the system that have dominant impact on the TCP throughput and developed Markov models to calculate the performance measures. An iterative solution method is provided to solve the queueing network model of the system. Numerical examples have been evaluated to demonstrate that our model is reasonably accurate and is able to capture the most important features of the system with much less computational effort compared to simulations. The faster computation times allow the application of the method for the dimensioning of the transport link capacity in HSDPA systems.

6 Acknowledgment

The research work of Levente Bodrog and Gábor Horváth is partially supported by the Hungarian Research Found (OTKA) under grant K61709. The content of this paper has been developed in cooperation with Nokia Siemens Networks.

7 References

[1] HOLMA H., TOSCALA A.: 'HSDPA/HSUPA for UMTS' (Wiley, 2006)

[2] LEGG P.J.: 'Optimised Iub flow control for UMTS HSDPA'. Vehicular Technology Conf., 2005 VTC 2005-Spring 2005 IEEE 61st, 30 May–1 June 2005, 4: Vol. 4, pp. 2389–2393

[3] ALEXIOU A., BOURAS C., IGGLESIS V.: 'Performance evaluation of TCP over UMTS transport channels'. Int. Symp. on Communications Interworking, 2004

[4] HAIDER A., HARRIS R., SIRISENA H.: 'Simulation-based performance analysis of HSDPA for UMTS networks'. Proc. Australian Telecommunication Networks and Applications Conference (ATNAC '06), 2006

[5] BADII J., DJAMAL ZEGHLACHE M.A.: 'Effect of TCP on UMTS-HSDPA system performance and capacity'. In IEEE Global Telecommunications Conf., GLOBECOM '04, 2004, vol. 6, pp. 4104– 4108

[6] DJAMAL ZEGHLACHE M.A.: 'Cross-layer design in HSDPA system to reduce TCP effect', *IEEE J. Selected Areas Commun.*, 2006, **24**, (3), pp. 614–625

[7] ASSAAD M., ZEGHLACHE D.: 'TCP performance over UMTS-HSDPA systems' (Auerbach Publications, Boston, MA, USA, 2006)

[8] ASSAAD M., ZEGHLACHE D.: 'Analytical model of HSDPA throughput under Nakagami fading channel', *IEEE Trans. Veh. Technol.*, 2009, **58**, (2), pp. 610–624

[9] VAN DO T., CHAKKA R., HARRISON P.G.: 'An integrated analytical model for computation and comparison of the throughputs of the UMTS/HSDPA user equipment categories'. MSWiM '07: Proc. 10th ACM Symp. on Modeling, Analysis, and Simulation of Wireless and Mobile Systems, ACM, 2007, pp. 45–51

[10] HAIDER A., HARRIS R.: 'A note on the performance of TCP over HSDPA '. Sixth Int. Conf. on Information, Communications & Signal Processing, 2007, pp. 1–5

[11] WEERAWARDANE T., PERERA R., TIMM-GIEL A., GORG C.: 'A Markovian model for HSDPA TNL congestion control performance analysis'. In IEEE 68th Vehicular Technology Conf., VTC 2008-Fall, 2008, pp. 1–6

[12] PADHYE J., FIROIU V., TOWSLEY D., KUROSE J.: 'Modeling TCP throughput: a simple model and its empirical validation'. Proc. ACM SIGCOMM '98 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communication, SIGCOMM '98 1998, (ACM Press), pp. 303–314

[13] STANKIEWICZ R., JAJSZCZYK A.: 'Modeling of TCP behavior in a DiffServ network supporting assured forwarding PHB', IEEE Int. Conf. on Communications, 2004, vol. 4, pp. 2071–2075

[14] BOLCH G., DE MEER H., GREINER S., TRIVEDI K.S.: 'Queueing networks and Markov chains : modeling and performance evaluation with computer science applications' (Wiley-Interscience, 1998)

[15] Eurane. 'The Eurane project'. 2004, http://www.ti-wmc.nl/eurane/

[16] CS VULKÁN G.H.: 'Throughput analysis of the proportional fair scheduler in HSDPA'. Proc. European Wireless 2008 (EW2008)

[17] HOLMA H., TOSKALA A.: 'WCDMA for UMTS' (Wiley, New York, NY, USA, 2002)

[18] BROUWER F., DE BRUIN I., SILVA J.C., SUOTO N., CERCAS F., CORREIA A.: 'Usage of link-level performance indicators for HSDPA network-level simulations in E-UMTS'. Proc. IEEE ISSSTA '04, 2004

doi: 10.1049/iet-sen.2009.0006