BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
FACULTY OF ELECTRICAL ENGINEERING AND INFORMATICS
DOCTORAL SCHOOL OF INFORMATICS
DEPT. OF TELECOMMUNICATIONS AND MEDIA INFORMATICS

ANALYSIS AND DESIGN OF RADIO ACCESS TRANSPORT NETWORK CONGESTION CONTROL AND DIMENSIONING METHODS

Pál L. Pályi

MSc. in Technical Informatics

Summary of the Ph.D. Dissertation

Supervised by

Dr. Sándor Molnár
High Speed Networks Laboratory,
Dept. of Telecommunications and Media Informatics

Advisors:
Dr. Sándor Rácz* and Dr. József Bíró‡
*Ericsson Research, Budapest; ‡HSN Lab.

Budapest, Hungary
2013

1 Introduction

Modern mobile telecommunications networks and high speed data packet services need the elaboration of several new congestion control methods in order to ensure the appropriate Quality of Service (QoS). During the development of the methods, it is not enough to apply existing techniques, but the characteristics of the network under development and the quality demanded by future users should be taken into account as well. Throughout this adaptation, many challenging assignments are to be solved in the scientific point of view.

The Radio Access Network (RAN) Transport Network (TN) of a 3rd Generation (3G) (Dahlman et al. (2008)) or 3rd Generation Partnership Project (3GPP) Long-Term Evolution (LTE) (3GPP (2008)) mobile telecommunications system comes with new problems to be solved. This thesis considers the Transport Network of the 3G High Speed Downlink Packet Access (HSDPA) (3GPP (2004)) system and I adapt my methods to this system.

1.1 Historical background

With the introduction of HSDPA, the RAN Transport Network may already be a bottleneck for the system. The Transport Network can also be congested because for HSDPA flows no resource reservation is made. In practice, the increased air interface capacity of HSDPA did not always come with similarly increased TN capacity. Network operators often upgrade the base stations (Node Bs) first and delay the upgrade of the Transport Network until there is significant HSDPA traffic. In some cases also the cost of Iub transport links is still high, however, it decreases significantly with the introduction of new mobile backhaul technologies, see Ericsson White Paper (2008). Thus it is a common scenario that the throughput is limited by the capacity available on the Iub TN links and not by the capacity of the air interface. Moreover, the Transport Network using optical fiber cables may also be a bottleneck in the case of a common HSDPA and LTE transport, where more NodeBs' and eNodeBs' traffic are multiplexed together. If an operator that provides HSDPA services introduces the LTE system, the existing 3G/HSDPA architecture may be reused for cost-efficiency reasons, see Ekstrom et al. (2006).

These issues provoke new research challenges. The RAN Transport Network needs special handling in terms of congestion control and dimensioning. The scarce resources of the RAN TN should be carefully dimensioned. Moreover, a system-specific congestion control is needed because the Radio Link Control (RLC) protocol (3GPP (2009)) does not have congestion control functionality and the TCP congestion control can not operate efficiently above the Acknowledged Mode RLC.

2 Research Objectives

The research objectives of the dissertation are twofold. On one hand, it is to design new RAN Transport Network congestion control methods and improvements in order to provide the QoS needed, taking into account new aspects. On the other hand, it is to extend dimensioning methods to model the typical traffic of the RAN Transport Network.

- In the first part the goal was to work out a method which improves the fairness characteristics of a system.
- In the second part it was to design a new RAN Transport Network congestion control.
- The third part deals with the characteristics of peak-rate limited Discriminatory Processor Sharing (DPS) with bandwidth-efficient rate sharing.
- The last part analyzes the flow-level performance of a multi-rate system supporting stream and elastic services.

3 Methodology

In the first part of the thesis, the algorithm design was supported and validated by analytical methods and simulations. The simulation tools were a simple self-developed flow-level Additive Increase Multiplicative Decrease (AIMD) simulator and a complex packet-level HSDPA protocol simulator. The protocol simulator was needed because the proposed method was applied in a high-complexity system that cannot be evaluated by numerical methods.

In the second part of the thesis, numerical method development was validated by mathematical methods.

4 New Results

4.1 Fairness-Optimal Initial Shaping Rate

In the HSDPA (High-Speed Downlink Packet Access) Iub Transport Network (TN) which connects the Radio Network Controller (RNC) with base stations (Node Bs), congestion control is needed [B1], see Figure 1. Because of the TN's often narrow resources, fairness of resource sharing is also an important issue.

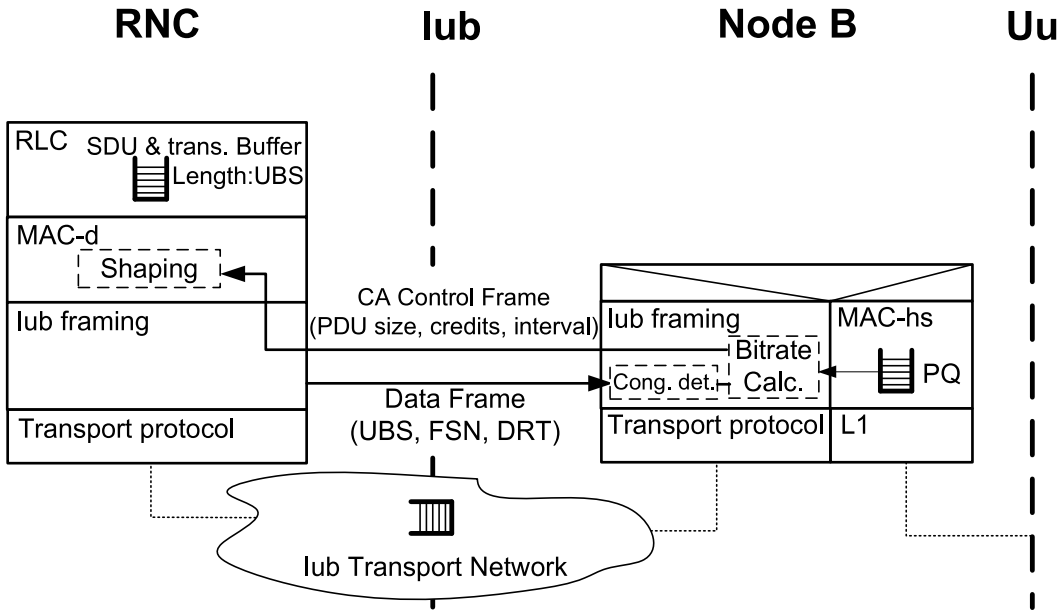

Figure 1: Congestion Control in the Radio Access Network

The Additive Increase Multiplicative Decrease (AIMD) congestion control scheme guarantees convergence to fairness in the long run; all flows converge to an equal share of resources in steady state, where no flows join or leave, see Chiu & Jain (1989). However, incoming flows may decrease the level of fairness, that is why transient fairness should be taken into account as well. AIMD does not define the starting rate of a flow. The scheme deals only with the dynamic behaviour of flows, after congestion.

The proposed method provides fairness-optimal initial rate for incoming HSDPA flows. With this fairness-optimal initial value, fairness can be improved to the greatest possible extent. The proposed method also improves average fairness. The method can be applied in a rate based congestion control where flows share the same bottleneck. A general solution for fairness-optimal initial rate in the case of second-order fairness measures is also given.

4.1.1 General formula for second-order fairness measures and fairness-optimal initial rate solution

Let X_i denote the actual bandwidth allocation of flow i and n the number of flows. Let us define second-order functions of these variables in the following way:

$$F_2(X_1, X_2, \ldots, X_n, n) = F_2\left(\sum_{i=1}^{n} d_i X_i, \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} X_i X_j, n\right),\ d_i, c_{ij} \in \mathbb{R}. \quad (1)$$

Let us assume that F_2 fulfills the following two requirements:
Req-(i): $F_2(Y, X_2, \ldots, X_n, n)$ has only one extreme value in Y.
Req-(ii): $F_2(Y, X_2, \ldots, X_n, n)$ is differentiable in Y.

In Thesis 1, I show that if function F_2 satisfies additional requirements (7)-(11) then it is a second-order fairness measure.

The well-known Jain's fairness index defined by Jain et al. (1984) is given by the following formula:

$$J(X_1, \ldots, X_n) \stackrel{\text{def}}{=} \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n \sum_{i=1}^{n} X_i^2}. \quad (2)$$

Thesis 1 *I have defined the general class of second-order fairness measures – including Jain's fairness index – and I have given the general fairness-optimal initial rate solution for this group of fairness measures*

The fairness-optimal initial shaping rate of the new flow in the case of Jain's fairness index can be determined in the following way:

$$Y^* = \arg\max_{Y} J(Y, X_1, X_2, \ldots, X_n), \quad (3)$$

where

$$J(Y, X_1, X_2, \ldots, X_n) = \frac{\left(\sum_{i=1}^{n} X_i + Y\right)^2}{(n+1)\left(\sum_{i=1}^{n} X_i^2 + Y^2\right)}. \quad (4)$$

$J(Y, X_1, X_2, \ldots, X_n)$ is differentiable with respect to Y and has only one maximum, which occurs at $Y^* = \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i}$. If this value is applied as the initial rate of a new flow, Jain's fairness index is improved to the greatest possible extent at the arrival of the new flow. This fairness gain depends only on n and $J(X_1, \ldots, X_n)$:

$$J(Y^*, X_1, \ldots, X_n) = J(X_1, \ldots, X_n)\frac{n}{n+1} + \frac{1}{n+1}. \quad (5)$$

In Fig. 2, relative fairness increase (from (5)) – gained by a new fairness-optimal flow – against current fairness is plotted; e.g., at n=5 if fairness is 0.4 currently it

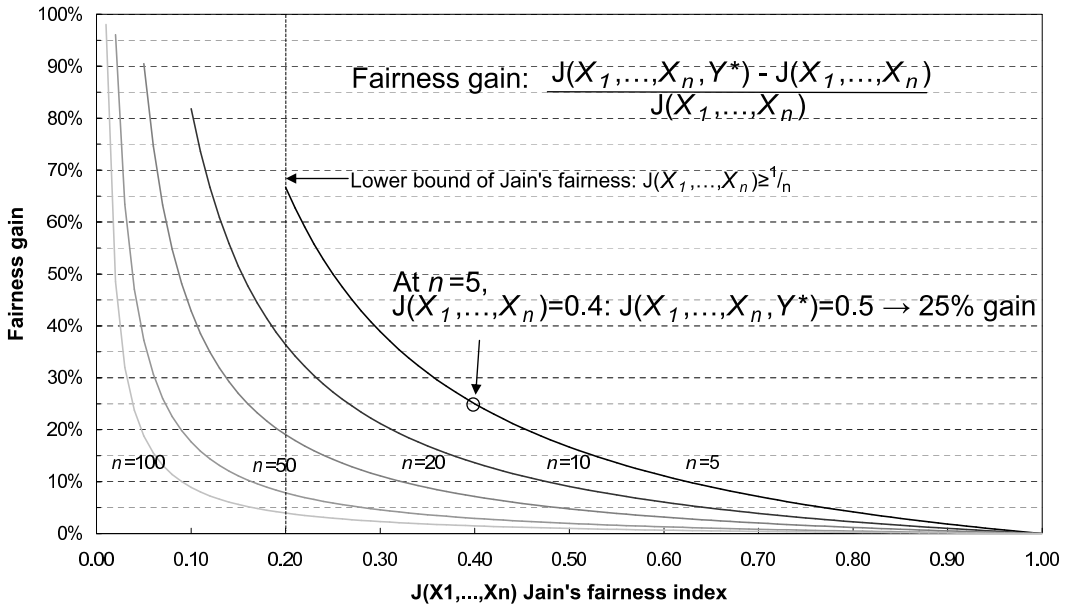

Figure 2: Relative increase of fairness (5) due to a fairness-optimal flow arrival

will change to 0.5 after adding Y^*, which means a 25% gain in fairness. Note that the method also provides gain compared to the trivial solution where we apply the average rate of ongoing flows as initial rate.

Thesis 1.1 *[C2] I have shown that the general form of second-order fairness measures is:*

$$F_2(X_1,\ldots,X_n,n) = \Phi\left(\frac{\frac{\sum_{i=1}^n X_i^2}{n}}{\left(\frac{\sum_{i=1}^n X_i}{n}\right)^2}\right), \tag{6}$$

where $\Phi(\alpha)$ is an arbitrary strictly monotonic function, where $\Phi(1)=0$ and $\Phi(\alpha)>0$ if $\alpha>1$. I have shown that this measure fulfills the following natural requirements for fairness functions, defined by Jain et al. (1984):

$$\begin{aligned}
F_2(X_1,\ldots,X_n) &= 0 \Leftrightarrow X_1=X_2=\cdots=X_n && (7)\\
F_2(X_1,\ldots,X_n) &\geq 0 && (8)\\
F_2(cX_1,\ldots,cX_n) &= F_2(X_1,\ldots,X_n)\ c\in\Re && (9)\\
F_2(X_1,\ldots,X_n) &= F_2(X_1,X_1,\ldots,X_n,X_n) && (10)\\
F_2(X_{i_1},\ldots,X_{i_n}) &= F_2(X_1,\ldots,X_n) && (11)\\
\forall\{i_1,\ldots,i_n\} &\in \prod\{1,\ldots,n\}
\end{aligned}$$

Thesis 1.2 *I have given the general fairness-optimal initial rate solution for the introduced second-order fairness measures, and shown that it is independent of function* Φ*. The general fairness-optimal solution is:*

$$Y^* = \frac{\sum_{i=1}^{n} X_i}{n} + \frac{\frac{\sum_{i=1}^{n} X_i^2}{n} - \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2}{\frac{\sum_{i=1}^{n} X_i}{n}}. \tag{12}$$

I have shown that this solution is in accordance with the one found for Jain's fairness index and $\Phi_{\text{Jain's}}(\alpha) = \frac{1}{\alpha}$.

It means that the fairness-optimal rate is independent of the fairness measure in the considered group.

The simplest $\Phi(\alpha)$ fulfilling the requirements above is $\Phi_1(\alpha) = \alpha - 1$. Equation (12) can be interpreted as follows. If the system is completely fair, i.e., the bandwidth allocation of each flow is the same, the fairness-optimal shaping rate of the new flow is the common rate. If there is some level of unfairness among the bandwidth allocations of the flows, then the fairness-optimal shaping rate of the new flow is the average bandwidth allocation of the ongoing flows increased with the relative variance of the bandwidth allocations of ongoing flows.

In Thesis 1.3, transient and long-term average fairness is considered. On transient fairness I mean the fairness of resource sharing in an arbitrary time instant (see Fig. 3). On long-term average fairness I mean the average of transient fairness values for all time instants. It is a natural requirement from a resource sharing algorithm that long-term average fairness should be independent from the number of parallel flows.

The method can be applied in a rate-based congestion control where flows share the same bottleneck. The common bottleneck is needed because fairness among flows could not be interpreted otherwise.

Thesis 1.3 *[C1][C2][P1] I have proposed a method in which the initial rate of a new HSDPA flow is set to the fairness-optimal value (Eq.* (12)*) with the Capacity Allocation (CA) Control Frame based on the CAs of ongoing flows; the method is described by the pseudo-code of Algorithm 1. I have extensively evaluated the performance of the proposed method (as an extension of an existing rate-based congestion control solution) in terms of fairness and shown that the proposed method improves both the transient fairness characteristics at user arrivals and the long-term average fairness (up to 30%, on average 20% compared to slow start) of the existing solution. I have also shown that the long-term average fairness is independent from the number of users in the case of this method.*

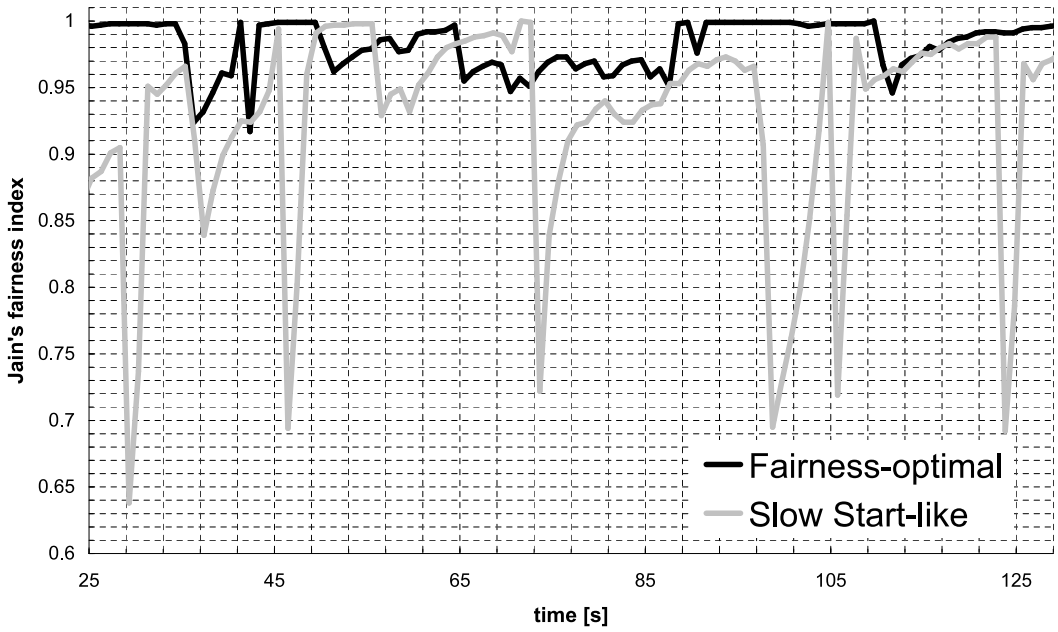

Figure 3: TCP-level fairness with and without the Fairness-Optimal method

The applied simulation tool was developed by a vendor and product development within the system is based on it. I have studied extensive simulation scenarios considered relevant by the vendor from the product point of view. A patent application [P1] has also been filed concerning the proposed method.

```
//Proposed method (in addition to the existing method):
//Variable c[i]: actual shaping rate of flow i
//At new flow arrival
Input: c[] for all ongoing flows
    if c[] not empty // if there are ongoing flows; calc. the init.
        then c[new flow]=AVG(c[])+VAR(c[])/AVG(c[]) //rate of new flow
    else use existing slow-start-like method to determine c[new flow]
Output: Shaping rate of the new flow c[new flow]
```

Algorithm 1: Sketch of the proposed method implemented in Node B

4.2 RLC-based HSDPA transport network congestion control

With the introduction of HSDPA, the transport network of the radio access network may already be a bottleneck for the system. TN can also be congested because for HSDPA no resource reservation is made. Loss or delay increase may occur due to congestion below the Radio Link Control (RLC) layer in the wired part. If the rest of the system is left unchanged, RLC cannot efficiently resolve these congestion situations, because it was not prepared for congestion control, but only for radio link failures. As a consequence, a mechanism is needed to resolve potential TN congestion, because RLC is hiding it from TCP by considering it as radio link failure (see 3GPP (2005), Nádas et al. (2007), and Vulkán & Nagy (2009)) and simply resending data suspected to be lost.

Considering the protocols in question, an interesting duality can be noticed between TCP and RLC. TCP was originally designed to handle loss as an indicator of congestion, i.e. if loss occurs rate reduction is needed. However, loss due to radio link failures (that are typically not because of congestion) should be handled in a different way, otherwise TCP is not efficient over wireless networks. On contrary, the original purpose of the RLC protocol is just the very opposite. RLC was designed to handle loss and increased delay due to radio link failures only, i.e., data units that are suspected to be lost are simply resent. Loss or delay increase due to congestion should be handled in a different way, otherwise RLC does not work efficiently over congested wired links such as the transport network in the case of HSDPA.

Thesis 2 *[C1] [J3] I have proposed and evaluated a possible extension of the Radio Link Control (RLC) protocol with congestion control functionality. I have shown that this method can handle congestion in the HSDPA Transport Network.*

Thesis 2.1 *[C1] [J3] I have proposed a window-based congestion control solution which controls the RLC transmission window and the MAC-d shaper according to the congestion level, with the appropriate transmission window size and shaping rate adjustment, and fits in the 3GPP architecture.*

The proposed scheme is an extension of the RLC protocol with congestion control functionality and it relies on the 3GPP congestion detection and signalling framework at Iub Framing Protocol layer. Therefore, it is a cross-layer solution. Modifications are only needed in the RNC and the solution is still standard compliant. The congestion control follows the AIMD scheme for the RLC transmission window size adjustment. This transmission window determines the average allowed bitrate of the flow. Due to the RLC-specific acknowledgment mechanism, traffic can become bursty. In order to perform traffic smoothing, the MAC-d shaper is also used. The

actual shaping rate, which limits the allowed peak rate of the flow, is in proportion with the average flow bitrate estimation to minimize average throughput limitation.

Thesis 2.2 *[C1][C2] I have adapted the method from Thesis 1 originally designed for rate-based congestion control to the proposed window-based congestion control. I have shown that the fairness-optimal solution for the initial transmission window size of a new flow is* $V^* = \frac{\sum_{i=1}^{n} W_i}{n} + \frac{\frac{\sum_{i=1}^{n} W_i^2}{n} - \left(\frac{\sum_{i=1}^{n} W_i}{n}\right)^2}{\frac{\sum_{i=1}^{n} W_i}{n}}$*, where* W_i *denotes the actual transmission window size of flow* i*.*

Since all the flows share the same bottleneck in the Transport Network, their Transport Network Round Trip Times are approximately equal in congestion. Therefore, actual bitrates are proportional with the corresponding transmission window sizes. As a consequence, the Fairness-optimal method (Thesis 1) can be adapted to window-based congestion control.

Thesis 2.3 *[C1] [J3] I have shown that the proposed method meets the low delay (below 100 ms), loss (∼1%) and high utilization (above 90%) requirements (3GPP (2001)) of a RAN TN Congestion Control by using a detailed protocol simulator that accurately models a WCDMA/HSDPA system. I have proposed an optimal parameter set for the proposed method. I have extensively evaluated the performance of this method (with the proposed parameter set), which is based on Thesis 2.1 and Thesis 2.2, in the 3GPP HSDPA architecture, where the HSDPA congestion control works according to Thesis 2.1 and the initial window size of a new flow is determined based on Thesis 2.2; the method is described by Algorithm 2.*

The same simulation tool was used for the performance evaluation as referred to in Thesis 1. The proposed window-based method provides very high (∼96%) Transport Network utilization compared to alternative solutions. Moreover, this high utilization is performed in a wide range of scenarios. At the same time, the proposed method provides fair resource sharing. Moreover, this fair resource sharing is also performed among flows with different RTTs (e.g. local server vs. overseas server).

```
//Variable tx[i]: actual RLC Tx window size of flow i in PDUs
//Regular update of the Tx window sizes of ongoing flows
for each received CA
    // based on information in the received CA
        if CA has Status Loss or Delay
            then
                decrease tx[i]
                state[i] = Congestion Avoidance
        else
            if state == CongestionAvoidance
                then
                    tx[i]+=(Nr of ACKs in received CA)/tx[i]
            if state == ExponentialStart
                then
                    tx[i]+=(Nr of ACKs in received CA)
end for each received CA

//Updating MAC-d shaping rate
for each ACKed RLC PDU calculate RTT
if Nr of ACKed RLC PDUs  > tx[i]
    then
        estimate AVG(RTT) based on per PDU RTTs
        estimate AVG(RLC bandwidth) based on AVG(RTT)
        set shaping rate to AVG(RLC bandwidth) * F_shaper

//At new flow arrival
Input: tx[] for all ongoing flows
    if tx[] not empty //if there are ongoing flows; calc. init. window
      then tx[new flow]=AVG(tx[])+VAR(tx[])/AVG(tx[]) //of new flow
    else tx[new flow]=W_init then state[new flow] = ExponentialStart
Output: Tx window size of the new flow tx[new flow]
```

Algorithm 2: Sketch of the proposed method implemented in the RNC

4.3 Peak-rate limited DPS with bandwidth-efficient rate sharing

Bandwidth sharing schemes are often modelled by processor sharing models. Discriminatory Processor Sharing (DPS) (Ayesta & Mandjes (2009)) is an important generalization of the (multi-class) egalitarian processor sharing discipline. In DPS, to each traffic class we assign a weight (e.g., service- or priority-class determined by the user's subscription). The weight of class-i is denoted by g_i. The bandwidth share of flows are proportional to these weights. More formally, two requirements can be identified on the capacity shares in DPS:

requirement-A: $\dfrac{c_i}{c_j} = \dfrac{g_i}{g_j}$, and

requirement-B: $\displaystyle\sum_{i=1}^{K} N_i c_i = C,$

where c_i's are the bandwidth shares, K is the number of traffic classes, N_i is the number of class-i users in the system, and C is the server capacity. These requirements are uniquely fulfilled by

$$c_i = \frac{g_i C}{\sum_{j=1}^{K} g_j N_j}, \quad i \in \{1, \ldots, K\}.$$

This is also referred to as the work-conserving property, i.e., either all flows get all the bandwidth they required or the system is serving on its full capacity.

The peak rate limitation means in the model that each traffic class has its own maximal rate that is denoted by b_i for class-i. If there is enough capacity then the flows receive their peak bandwidths. When there is not enough capacity for all ongoing flows to get their peak rates, that is, $\sum_{i=1}^{K} N_i b_i > C$, then some flows or all flows will be "compressed" in the sense of their reduced service rates. This is the elastic "regime" of the model.

According to bandwidth-efficient rate sharing, unused capacity is redistributed among flows in proportion to their weights (requirement-A). For a while, let us assume that the set of compressed ($\mathcal{Z}$:$\{\forall i, c_i < b_i\}$) and uncompressed ($\mathcal{A}$) classes are known in a given state $\underline{N} = (N_i,\ i \in \{1, \ldots, K\})$. In the uncompressed case, $c_i = b_i,\ i \in \mathcal{A}$. Since these flows cannot utilize their bandwidth shares, they leave

$$\sum_{i \in \mathcal{A}} \left(\frac{g_i N_i}{\sum_{j=1}^{K} g_j N_j} C - N_i b_i \right)$$

capacity which is re-distributed among compressed flows. If $j \in \mathcal{Z}$, the original bandwidth share is increased due to the redistribution. The redistribution should be

proportional to the weights g_i, in order to keep a similar requirement to requirement-A. Between two compressed classes, $c_i/c_j = g_i/g_j$, $i,j \in \mathcal{Z}$, and between a compressed and an uncompressed class, $c_i > c_k \frac{g_i}{g_k}, \forall i \in \mathcal{Z}, \forall k \in \mathcal{A}$. (The latter requirement is needed to ensure that $\mathcal{Z}$ is unique for given $\underline{N}$.) This results

$$c_i = \frac{g_i}{\sum_{j=1}^{K} g_j N_j} C + \frac{g_i}{\sum_{k \in \mathcal{Z}} g_k N_k} \sum_{l \in \mathcal{A}} \left(\frac{g_l N_l}{\sum_{j=1}^{K} g_j N_j} C - N_l b_l \right), i \in \mathcal{Z}. \qquad (13)$$

Due to our assumption, constraint $c_i \leq b_i$ is fulfilled for $i \in \mathcal{Z}$.

For implementing a calculation of bandwidth-efficient rate shares based on (13) I first show a simpler form of that equation, and then using this simpler form, I present a method for determining $\mathcal{Z}$ and $\mathcal{A}$.

Thesis 3 *[J2] I have characterized the state space and the bandwidth sharing scheme of the peak-rate limited DPS with bandwidth-efficient rate sharing*

Thesis 3.1 *[J2] I have shown that the service rate of the compressed classes' users formulated in* (13) *can be re-written as*

$$c_i = \frac{g_i}{\sum_{j \in \mathcal{Z}} g_j N_j} \left(C - \sum_{k \in \mathcal{A}} N_k b_k \right), i \in \mathcal{Z}. \qquad (14)$$

The immediate consequence is that c_i, $i \in \mathcal{Z}$ can be considered as the bandwidth allocation of a reduced Discriminatory Processor Sharing system with capacity $(C - \sum_{k \in \mathcal{A}} N_k b_k)$ and traffic classes $\mathcal{Z}$ in state $\underline{N}$.

Thesis 3.2 *[J2] I have shown that the following order of classes:*

$$\frac{g_1}{b_1} \leq \frac{g_2}{b_2} \leq \ldots \leq \frac{g_K}{b_K}\ , \qquad (15)$$

based on the ratios g_i/b_i, is directly related to the compressed and uncompressed classes in such a way that:
If a class with given g/b value is compressed, then all classes with lower g/b are compressed, and if a class with given g/b value is uncompressed, then all classes with higher g/b are uncompressed.

I have also shown that the compression order depends on neither the server capacity nor the number of users.

I have given a method (Algorithm 3) for determining the set of compressed classes $\mathcal{Z}$ and the bandwidth share c_i of each flow.

1. $\mathcal{Z} = \{1, 2, ..., K\}$

2. *while* $\max_{i \in \mathcal{Z}} \left\{ \frac{g_i}{b_i} \right\} \geq \frac{\sum_{j \in \mathcal{Z}} N_j g_j}{C}$ *and* $\mathcal{Z} \neq \emptyset$ *do*

$$i' = \arg\max_{i \in \mathcal{Z}} \left\{ \frac{g_i}{b_i} \right\}$$
$$\mathcal{Z} \leftarrow \mathcal{Z} \backslash \{i'\}$$
$$C \leftarrow C - N_{i'} b_{i'}$$

3. *for* $i = 1$ *to* K *do*

$$\textit{if } i \in \mathcal{Z} \textit{ then } c_i = \frac{g_i}{\sum_{j \in \mathcal{Z}} N_j g_j} C$$
$$\textit{else } c_i = b_i.$$

Algorithm 3: Determining the set of compressed classes

4.4 Flow level performance analysis of a multi-rate system supporting stream and elastic services

The integration of stream and elastic traffic on flow level is analyzed; the average throughput is calculated for elastic traffic classes. Stream flows are characterized by a fixed bandwidth assignment (e.g., Guaranteed Bit Rate, GBR). Elastic flows (e.g., data) can adapt their service requirements and share the capacity left over by stream calls. The sojourn time of elastic calls is affected by the assigned service rate.

An efficient numerical approximative algorithm is proposed that handles many stream and many elastic classes with different peak bandwidth requirements and mean holding times. A two-dimensional macro-state representation is introduced for the micro-state model, in order to calculate the average throughputs. The macro-state model provides a means to describe the original multi-rate stream elastic system with fewer details based on aggregated statistical descriptors. Table 1 summarizes the input parameters that describe the system.

Thesis 4 *[J1] I have characterized the stationary probabilities of macro-states and evaluated the accuracy of the approximation method.*

Thesis 4.1 *[J1] I have shown that the stationary probability of the macro-states can be efficiently calculated using the following three steps:*

- *expressing the probability of all macro-states as a linear combination of the probability of macro-states in* $\mathcal{S}_0 = \{\{k, j\} : \ j = 0 \dots b_m - 1, \ k = 0 \dots T_{st}\}$*; in this step we use the global-balance equations of macro-states* $\{\{k, j\} : \ j = 0 \dots T_{el} - b_m, \ k = 0 \dots T_{st}\}$*; where* m *is the index of the elastic class which has the highest peak rate;*

System descriptors		
C	link capacity	$[BU]$
T_{st}	limit on total stream occupancy	$[BU]$
T_{el}	limit on total elastic occupancy assuming peak rates	$[BU]$

Descriptors of the i-th traffic class		
TYPE	stream or elastic class	-
b_i	peak rate (limitation)	$[BU]$
λ_i	flow arrival rate	$[flows/sec]$
t_i	mean flow holding time; $\mu_i = 1/t_i$	$[sec]$
ρ_i	$\rho_i = \lambda_i/\mu_i$ (and $\rho_i b_i$ is the offered load)	

Table 1: Input parameters

- *determining the probabilities of states in $\mathcal{S}_0$; in this step we use the global-balance equations of states $\{\{k, j\} : \ j = T_{el} - b_m + 1 \dots T_{el}, \ k = 0 \dots T_{st}\}$ and solve a linear system of equations of size $b_m \times (T_{st} + 1)$.*

- *determining the probabilities of states in $\mathcal{S} \setminus \mathcal{S}_0$ using the probabilities of states in $\mathcal{S}_0$.*

The computational complexity of this method is significantly lower, because it is enough to solve $b_m \times (T_{st} + 1)$ linear equations instead of solving $(T_{el} + 1) \times (T_{st} + 1)$.

Thesis 4.2 *[J1] I have shown that the proposed approximation method provides accurate results – it is more exact for larger systems and for less compressed elastic flows – by extensive accuracy evaluation.*

I used the results of $\sim 10\,000$ random scenarios to demonstrate the accuracy of the method. For reference, the original micro-state model was solved numerically. The systems under investigation have four classes, class-1 and class-2 are stream classes and class-3 and class-4 are elastic classes. Figure 4–7 show the accuracy of the method. In the figures, r denotes the average compression of elastic classes (value 1 denotes no compression), and the following error measures are used: the *per-class error* (e_i) in Figure 7, the *average per-class error* (e_A) both in Figure 4 and Figure 6, and the *system throughput error* (e_B) in Figure 5. The per-class error of the throughput is defined as:

$$e_3 = \frac{\widetilde{\theta}_3 - \theta_3}{\theta_3} \qquad e_4 = \frac{\widetilde{\theta}_4 - \theta_4}{\theta_4}.$$

where $\widetilde{\theta}_3$ and $\widetilde{\theta}_4$ denotes the performance measures provided by the method for the two elastic classes. θ_3 and θ_4 denotes the exact values of average throughputs.

In addition, let us define the average per-class error as the average of per-class errors weighted by the carried load of elastic classes:

$$e_A = \sqrt{\frac{(1-B_3)\rho_3 b_3 e_3^2 + (1-B_4)\rho_4 b_4 e_4^2}{(1-B_3)\rho_3 b_3 + (1-B_4)\rho_4 b_4}}.$$

where B_3, B_4 denotes the blocking probabilities.

The system throughput error is defined as:

$$e_B = \frac{\widetilde{\theta}_{avg} - \theta_{avg}}{\theta_{avg}}.$$

where,

$$\theta_{avg} = \frac{(1-B_3)\rho_3 b_3 \theta_3 + (1-B_4)\rho_4 b_4 \theta_4}{(1-B_3)\rho_3 b_3 + (1-B_4)\rho_4 b_4} \qquad \widetilde{\theta}_{avg} = \frac{(1-B_3)\rho_3 b_3 \widetilde{\theta}_3 + (1-B_4)\rho_4 b_4 \widetilde{\theta}_4}{(1-B_3)\rho_3 b_3 + (1-B_4)\rho_4 b_4}$$

Both the values of the average per-class error and the system throughput error are concentrated at zero in Figure 4 and Figure 5, respectively. Moreover, the system throughput error is less than the average per-class error.

Figure 6 shows that the method is more accurate for larger systems. The error decreases for each compression group if we increase the capacity.

Figure 7 shows that the method provides more accurate results for less compressed elastic flows. This figure depicts the per-class error of the second elastic class versus the per-class error of the first elastic class. Scenarios with the least compressed elastic flows have the smallest errors. Values in the figure are concentrated around zero. The higher the level of compression, the bigger the average distance from zero. If elastic flows are not compressed the method has no error.

As a consequence, the proposed method provides a good approximation.

5 Applicability of the Results

Most of the work presented in this PhD thesis was done at the Traffic Analysis and Network Performance Laboratory of Ericsson Research at Ericsson Hungary. I worked as a team member in a larger international research group. This group of people dealt with transport issues of RAN – mainly focused on HSDPA – and related research tasks. Some of the research results presented in this work are filed by Ericsson as patent application [P1]. Thesis 1.3 and Thesis 2.3 are directly applicable in HSDPA systems. Results in Thesis 3 and Thesis 4 are applicable in dimensioning methods modelling the typical traffic of the RAN TN.

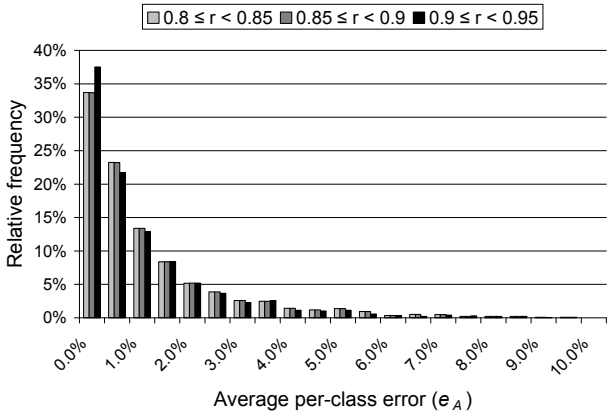

Figure 4: Histogram of the average per-class error, e_A

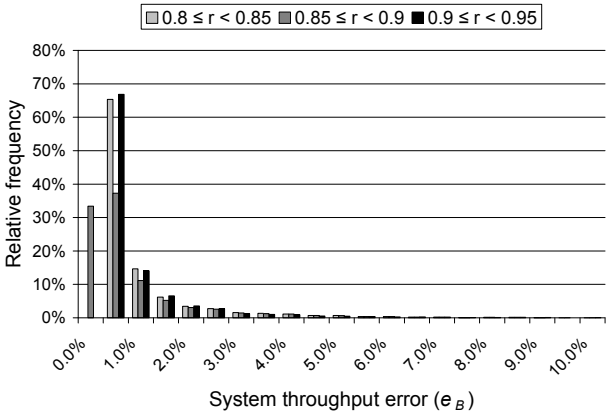

Figure 5: Histogram of the system throughput error, e_B

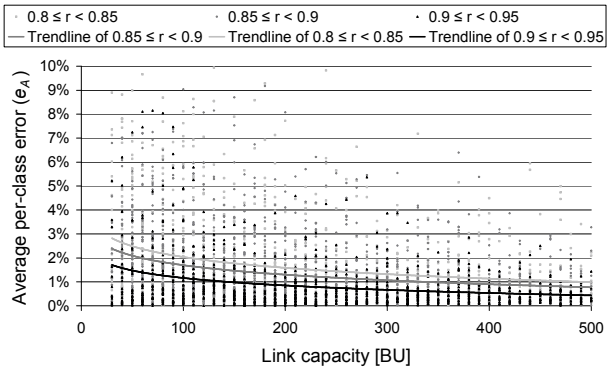

Figure 6: Average per-class error (e_A) vs. capacity

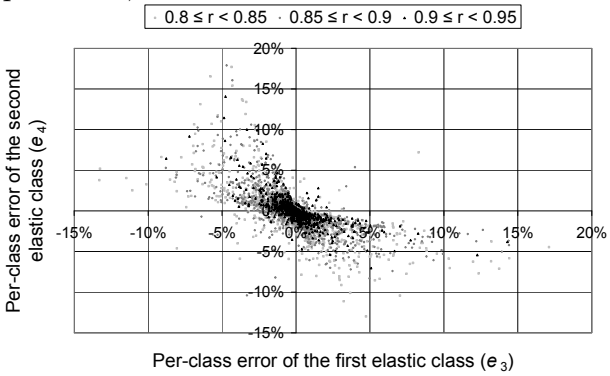

Figure 7: Per-class error of the second vs. the first elastic class, (e_3, e_4)

Acknowledgements

First of all, I would like to say thanks to all of my three advisors for their support. Especially to my industrial advisor, Dr. Sándor Rácz at Ericsson Research Hungary who introduced me as still an MSc student to the very interesting research field of mobile broadband and guided me throughout all my research activities. His exceptional professional knowledge, insight, enthusiasm, encouragement and support have given me an essential factor of motivation to finish (and even start) the present PhD thesis.

Most of the work presented in this dissertation was completed at the Traffic Analysis and Network Performance Laboratory of Ericsson Research Hungary. I wish to say thanks to my colleagues and former colleagues for their support during the work; especially Szilveszter Nádas, Zoltán Nagy, István Ketykó and Balázs Péter Gerő.

References

3GPP (2001). TS 25.853 V4.0.0 (Delay budget within the access stratum).

3GPP (2004). TS 25.308 version 6.3.0 (UTRA High Speed Downlink Packet Access (HSDPA) Overall Description).

3GPP (2005). TR 25.902 version 6.0.0 (UMTS Iub/Iur congestion control).

3GPP (2008). TS 36 series, release 8 LTE (Evolved UTRA) and LTE-Advanced radio technolgy.

3GPP (2009). TS 25.322 V8.5.0 (Radio Link Control (RLC) protocol specification).

Ayesta, U. & Mandjes, M. (2009). Bandwidth-sharing networks under a diffusion scaling. *Annals Operation Research*, *170*(1), 41–58.

Chiu, D. & Jain, R. (1989). Analysis of the increase/decrease algorithms for congestion avoidance in computer networks. *Computer Networks and ISDN*, *17*(1), 1–14.

Dahlman, E., Parkvall, S., Skld, J., & Beming, P. (2008). *3G Evolution – HSPA and LTE for Mobile Broadband*. Academic Press.

Ekstrom, H., Furuskar, A., Karlsson, J., Meyer, M., Parkvall, S., Torsner, J., & Wahlqvist, M. (2006). Technical solutions for the 3G long-term evolution. *Communications Magazine, IEEE*, *44*, 38–45.

Ericsson White Paper (2008). High-speed technologies for mobile backhaul. `http://www.ericsson.com/technology/whitepapers/pdf/High-speed-technology-mobile-backhaul.pdf`. 284 23-3119 Uen Rev B.

Jain, R., Chiu, D., & Hawe, W. (1984). A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. Technical Report TR-301, DEC Research.

Nádas, S., Rácz, S., Nagy, Z., & Molnár, S. (2007). Providing congestion control in the Iub Transport Network for HSDPA. In *Globecom*, (pp. 5293–5297).

Vulkán, C. & Nagy, Z. (2009). Iub/Iur HSDPA Congestion Control. In *ICT-MobileSummit 2009*.

Publications

Book chapters

[B1] Sz. Nádas, S. Rácz and **P. Pályi**. HSPA Transport Network Layer Congestion Control. In *HSDPA/HSUPA Handbook*. CRC Press, Taylor & Francis Group, 2010. pp. 297–330.

Journal papers

[J1] B. Gerő, **P. Pályi** and S. Rácz. Flow level performance analysis of a multi-rate system supporting stream and elastic services. In *International Journal of Communication Systems*, 2012. doi: 10.1002/dac.1383.

[J2] **P. Pályi**, A. Körösi, B. Székely, J. Bíró, S. Rácz. Characterization of Peak-Rate Limited Bandwidth-Efficient Discriminatory Processor Sharing. in *Acta Polytechnica Hungarica*, 2012. pp. 151–164.

[J3] **P. Pályi**, M. Horváth and S. Rácz. RLC-alapú HSDPA szállítóhálózati torlódásvezérlés. In *Híradástechnika*, 2010. pp. 15–20.

Conference papers

[C1] **P. Pályi**, S. Rácz and Sz. Nádas. Window-based HSDPA Transport Network Congestion Control. In *Proc., European Wireless 2010*, Lucca, Italy, April 2010, pp. 123–131.

[C2] **P. Pályi**, S. Rácz and Sz. Nádas. Fairness-Optimal Initial Shaping Rate for HSDPA Transport Network Congestion Control. In *Proc., IEEE ICCS 2008*, Guangzhou, China, November 2008, pp. 1415–1421.

[C3] N. Ukic, M. Zemljic, I. Markota, **P. Pályi**, D. Asztalos. Evaluation of Bridge-Point Model-Driven Development Tool in Distributed Environment. In *Proc., Soft-COM 2011*, Split, Croatia, September 2011.

[C4] **P. Pályi**, S. Molnár. Fairness Study of HSDPA in the Transport Network. In *Proc., TRANSCOM 2007*, Zilina, Slovak Republic, June 2007, pp. 197–200.

[C5] **P. Pályi**. Application of elastic models in HSDPA transport network dimensioning. In *Proc., Poster 2007*, Prague, Czech Republic, May 2007.

[C6] **P. Pályi**. HSDPA fairness analysis in the transport network. In *Proc., HTE-BME 2007 students conference*, Budapest, Hungary, May 2007.

Patents

[P1] **P. Pályi**, S. Rácz, Sz. Nádas and Z. Nagy. Method For Achieving an Optimal Shaping Rate For a New Packet Flow. Patent application, Filing Number: PCT/SE2009/050726, 2009.

[P2] S. Rácz, Sz. Nádas and **P. Pályi**. Delayed Flow Control Action in Transport Network Layer WCDMA Communications. Patent Application, U.S. Application Serial No. 12/730,752, 2010.

[P3] **P. Pályi**, M. Skarve, S. Rácz and Sz. Nádas. Non-congestive loss in HSPA congestion control. Patent Application, Filing Number: 13162136.9, 2013.

Other publications

[O1] **P. Pályi**. RLC-based HSDPA transport network congestion control. Poster presentation in *High Speed Networks Laboratory Workshop 2010*, Balatonkenese, Hungary, May 2010.

[O2] **P. Pályi**, S. Rácz, Sz. Nádas. Fairness-Optimal Initial Shaping Rate for HSDPA Transport Network Congestion Control Presentation in *High Speed Networks Laboratory Workshop 2009*, Balatonkenese, Hungary, May 2009.