Performance Analysis of Bandwidth Allocation Schemes in Multiservice IP Networks using Utility Functions

Veselin Rakocevic[a], John Griffiths[a], Graham Cope[b]

[a]Department. of Electronic Engineering, Queen Mary, University of London, London E1 4NS, UK, Email: {v.rakocevic, j.m.griffiths} @elec.qmul.ac.uk

[b]Fujitsu Telecommunications Europe Ltd, Birmingham, UK, Email: G.Cope@ftel.co.uk

Abstract: The Internet needs to evolve from a single-service data network into the multiservice intelligent network capable of satisfying diverse performance requirements. The way network resources, primarily bandwidth, is shared, is one of the key issues in this evolution. This paper analyses the concept of bandwidth partitioning, in which each link in the network is partitioned into a number of sublinks, each sublink serving a single traffic class. We analyse a new approach to bandwidth partitioning, *Dynamic Bandwidth Partitioning*. In this scheme bandwidth partitioning is related to the average level of end-user satisfaction with the network performance according to a simple linear control algorithm. The scheme is user-oriented and adaptive, designed to maximise the overall end-user utility. Furthermore, this paper observes a number of bandwidth allocation schemes in the mutliservice Internet environment, and compares their performance by comparing the average *connection utility* - the quality of service level which a user of an Internet application derives from the network performance. Our analysis shows that Dynamic Bandwidth Partitioning generates higher overall utility in the multiservice Internet environment than the current best-effort scheme, mainly because it takes account of the quality of service requirements of the real-time Internet traffic.

1. Introduction

Bandwidth allocation in the multiservice communication networks presents a very important problem in the design of the future multi-class Internet. The main motivation for the research in this field lies in the necessity for structural changes in the way the Internet is designed. The current Internet offers a single class of 'best-effort' service, although some traffic prioritisation will be active in the new network router implementations. A lot of work has been done [1][2][3] concerning the issues of bandwidth allocation and fairness in a single-service environment, for the network serving a single type of *elastic* traffic [4]. The objective there was to use all available bandwidth while trying to achieve some fairness in the way the bandwidth is shared between different traffic flows.

However, the Internet is changing. New sophisticated real-time applications (video conferencing, video on demand, distance learning, etc) require better and more reliable network performance. Furthermore, these applications require firm guarantees from the network that certain resources (bandwidth, buffer space) will be reserved for them. Network designers are facing a complicated problem of optimising the network control to satisfy both the issue of *fairness* for the data traffic and the issue of *performance guarantees* for the real-time traffic.

This paper analyses resource partitioning as a means to provide quality of service to large number of end-users, while preserving a high level of network performance. Although network resources are numerous, we focus in this work on the link bandwidth. We propose a bandwidth partitioning scheme which is application-oriented and dynamic, allocating the network bandwidth to maximise the end-users' utility. Bandwidth partitioning makes traffic from different traffic classes independent, and therefore protects high-priority traffic from the effect of sudden burstiness of the low-priority traffic. The concept of bandwidth partitioning has been analysed previously (e.g. [5]), with a common conclusion that it is ineffective, due to low network utilisation. It is considered to perform better then the best-effort scheme only in the environment of very high traffic loads.

In this work we present a novel approach to bandwidth partitioning, showing that an adaptive partitioning scheme can perform better then the best-effort scheme even for moderate traffic loads. Our scheme is user-oriented, designed to generate maximal user-utility on the network. We start from the fact that each user of an Internet application derives a certain utility from the network performance. Each user of an Internet application defines its utility function, which relates the allocated bandwidth level to the user satisfaction, rating that satisfaction on a scale from 0 to 1. The current Internet works on the premise that network capacity is shared proportionately between the users, by using implicit control mechanisms such as TCP. This assumes that the utility function that drives QoS allocation is uniform among the users. However, the utility of a service is flexible according to user's subjective perceptions. Furthermore, we should have in mind that the users are interested primarily in the quality of service, rather than just having the certain amount of network bandwidth available.

A number of works recently use end-users' utility as the maximizing objective for resource allocation schemes. All of these approaches have a common objective of maximising the network performance in terms of the users' utility. Kelly [3] argues that bandwidth should not be shared so as to maximise the link utilisation, but rather to maximise an objective function representing the overall utility of the flows in progress. A similar approach can be found in [6], where a *utility max-min* fairness is introduced. Shenker and Breslau [7] define the appropriate mathematical expression for the non-concave utility function for adaptive applications, and analyse the necessity for the admission control and bandwidth reservation in the case of adaptive real-time applications. Sarkar and Tassiulas [8] also present an algorithm for computation of maximally fair utility allocation, and use a stepwise utility function for video applications. Very interesting work by Rajkumar et al [9][10] presents a framework for utility optimisation in systems that must satisfy application needs along multiple dimensions. Their work presents the way to derive the end-user utility function from multiple utility functions, each relating to a different resource.

The current traffic heterogeneity in the Internet brings numerous different applications with hundreds of different utility functions. It is possible that every single user of an Internet application has a different utility function. The precise definition of all utility functions is therefore a very complicated problem, as discussed in [11]. That is why in this work we made an approximation by defining a finite number of traffic classes, and defining a single utility function per each traffic class. We believe that by using well-defined utility functions we can efficiently evaluate the network performance.

The paper is organized as follows. Section 2 of the paper presents the Dynamic Bandwidth Partitioning scheme, including the algorithm for the dynamic change of the partitioning parameters. Section 3 presents the performance evaluation of the scheme and the comparison of the scheme with other bandwidth allocation schemes. This section presents the simulation model that has been developed, and the mathematical expressions for the used utility functions. Finally, the simulation results are discussed in section 3.4.

2. Dynamic Bandwidth Partitioning

2.1. Model Description

Consider a network with K links, where each link $k = 1,...,K$ has bandwidth B_k. The network is serving numerous user-applications. Each user-application j is allocated a certain amount of bandwidth b_j while active in the network. There is a utility function $u_j(b_j)$ defined for each user-application. If we assume that utility functions are additive, the overall utility generated in the network at any time is $U = \sum_j u_j(b_j)$. The optimal bandwidth allocation on such a model calculates the bandwidth b_j to be allocated to each traffic flow j in order to maximize the overall utility U generated in the network. This presents a multi-constrained optimization problem, which heavily depends on the defined utility functions. Furthermore, we must not observe this problem only from the analytical point of view, since the current Internet requires a bandwidth allocation scheme which is adaptable, easy to implement and compatible with the existing network controls, as well as being optimal. Previous research [10] has proven that allocation of multiple resources in order to maximize the overall utility is NP-hard. Therefore, only an approximate solution is possible.

In this work we analyze possible implementation of a bandwidth partitioning concept into this problem. Although it is more correct to say that each end-user defines its own utility function, in this work we make an approximation that all end-users of a particular application have equal utility function. Consider that all active traffic in the network can be classified into I traffic classes. A bandwidth partitioning scheme decouples the link bandwidth into I independent 'sublinks', where each of the sublinks serves one traffic class. The partitioning is defined with the set of partitioning parameters $\alpha_{ik}, i = 1,...I,\ k = 1,...,K$. Each individual sublink i occupies the bandwidth $B_{ik} = \alpha_{ik} B_k$ on the link k. Sublinks can be treated independently, therefore the performance measures of interest can be easily calculated.

The level of bandwidth allocated to each traffic flow depends heavily on the partitioning parameters on the links that flow traverses. Therefore, the above optimisation problem can be mapped on the problem of finding the vector of parameters $\boldsymbol{\alpha}$ which maximises the overall utility. Previous work [12] has shown that, if we know the utility functions for each of the traffic classes and the traffic intensity, it is possible by using linear programming, to calculate an optimal set of partitioning parameters which maximise the overall utility on the link.

Bandwidth partitioning is usually considered to be efficient only when the traffic load is very heavy. However, in practice bandwidth partitioning can prove to be an interesting resource allocation scheme. An interesting implementation of this model is in the case when network capacity needs to be partitioned for some other purposes, such as the case with Virtual Private Networks [13], where the network capacity is partitioned and leased to a number of users, who pay for part of the network capacity. Complete bandwidth partitioning will never provide highest capacity utilisation, but, observed in the multiservice environment of a network whose goal is to utilise its capacity in terms of maximising the revenue, quality of service and user's satisfaction, bandwidth partitioning can still be seriously considered as a resource allocation scheme.

In order to increase the efficiency and adaptability of the bandwidth partitioning concept, in this paper we present a Dynamic Bandwidth Partitioning scheme. In this scheme, the partitioning parameters α_{ik} are not fixed and pre-defined, but they are variable, and the parameters change with the change in the traffic intensity and the current end users' utility. This paper will show how a simple dynamic partitioning scheme can improve the overall utility of the network.

2.2. Partitioning Algorithm

In order to design an algorithm for dynamic bandwidth partitioning, we observe a simplified network model that consists of a single link, of bandwidth B, which serves the traffic belonging to I different traffic classes. We denote partitioning parameters with α_i, and the bandwidth allocated to each traffic flow from the traffic class i, is $b_i = \alpha_i B/n_i$, where n_i is the number of active traffic flows belonging to traffic class i. Each traffic class i has its utility function, $u_i(b_i)$, defined. Therefore, a traffic class is characterised by its transmission rate b_i and by its utility function $u_i(b_i)$. The objective of a bandwidth allocation scheme is to calculate the bandwidth for each traffic flow that maximises the sum of the utilities over b_i subject to capacity constraints.

The way bandwidth is allocated between active traffic flows determines the experienced utility. The idea for the dynamic change in partitioning parameters in the Dynamic Bandwidth Partitioning scheme is that every time the utility decreases for a certain value, a change in the partitioning parameters happens, in the direction which increases the overall utility.

Let us define the *normalised* utility Ψ:

$$\Psi = \frac{\sum \omega_i n_i u_i(b_i)}{\sum \omega_i n_i} \tag{1}$$

where ω_i are the scaling factors. The scaling factors are introduced to show that the defined traffic classes should not be treated with the same priority. Without the scaling factors, the generated utility will be determined by the number of the active flows from each of the classes. However, we argue that prioritisation is necessary in the multi-class IP environment. It is far more complicated to serve a customer that uses sophisticated video connection, then the one doing a simple file transfer.

If the normalised utility Ψ between two time instances, $t-\tau$ and t, decreases for more then a certain specified amount δ, a change of the partitioning parameters happens. The direction of the change in defined by the change of the utilities for each traffic class, $\Delta u_i(t) = u_i(t) - u_i(t-\tau)$. The parameter with the largest decrease in the utility is increased, while all other parameters are decreased. The linear control algorithm is used to calculate the new value for the partitioning parameters. The value of the partitioning parameter α_i at time t is then,

$$\alpha_i(t) = \alpha_i(t-\tau) + f\left[\Psi(t), \Psi(t-\tau)\right] \tag{2}$$

A linear control algorithm has been successfully used in the TCP congestion control mechanism. TCP uses additive increase, multiplicative decrease rule [15], with transmission rates being increased until the network signals the loss of packets. Then, the transmission rates are decreased by multiplying the current transmission rate with some constant, usually 0.5. We use a similar approach to design the partitioning algorithm. The function $f\left[\Psi(t), \Psi(t-\tau)\right]$

contains the information about the change in the utility. Based on that information, it generates the appropriate change for the partitioning parameters. Let us introduce an indicator for the direction of change, $\theta_{\mathrm{i}}(t) \in \{0,1\}$, where $\theta_i(t)=1$ indicates the increase, and $\theta_i(t)=0$ indicates the decrease. It is clear that only one traffic class will have $\theta_i(t)=1$, since only one of the parameters α_i will be increased. The partitioning algorithm can now be defined as follows:

$$\alpha_i(t)=\alpha_i(t-\tau)+\left[\varepsilon_i^{inc}\theta_i(t)-\varepsilon_i^{dec}(1-\theta_i(t))\right] \tag{3}$$

ε_i^{inc} and ε_i^{dec} are additive parameters, for increase and decrease respectively. Finding the optimal values for these parameters presents a very interesting problem for the future research. Equation (3) clearly shows that our algorithm follows the additive increase, additive decrease control rule. The main idea is to always perform the change that will increase the utility of the active traffic flows. The only constraint is that the partitioning parameters need to be within the interval $\alpha_i \in \{\alpha_{i\min}, 1\}$, where $\alpha_{i\min}$ defines the part of the bandwidth that is reserved for the traffic class i. To understand the need for the reservation of a portion of bandwidth, we have to understand the nature of Internet applications. A dynamic scheme like this provides variable bandwidth levels to all Internet applications. While many real-time applications are able to adapt to the variable bandwidth levels in the network, they still require a certain minimal *guaranteed* bandwidth level to be able to operate. The network must be able to guarantee minimal bandwidth to real-time traffic classes. That is why real-time traffic classes require minimal values for partitioning parameters, $\alpha_{i\min}$. On the other hand, data traffic flows do not require guaranteed minimal bandwidth level to operate. However, it is possible to introduce a minimal partitioning parameter for data traffic class, as well, to prevent it from being completely pushed aside by the higher-priority traffic.

3. Performance Evaluation

3.1. Simulation Model

In order to evaluate the efficiency of the Dynamic Bandwidth Partitioning scheme, a comparison between this scheme and a number of other schemes has been performed. A simulator has been specially built for that purpose. The simulation is done on the level of traffic flows, with traffic flows from all traffic classes arise as a Poisson process, and have the duration/size exponentially distributed. The traffic is differentiated into three major traffic classes, which will be explained in detail in the next section.

Bandwidth allocation schemes are compared on the basis of an evaluation metric which is called *connection utility*. Connection utility is calculated when the traffic flow terminates, by simply calculating the time average utility while the flow was active. For analytical purposes, we can approximate the calculation of this average with an integral in equation (4). The connection utility v_{ji} of a traffic flow j that belongs to the traffic class i can be seen as the approximation of the network performance traffic flow received while in the network.

$$v_{ji} = \frac{1}{T_{dur}} \int_0^{T_{dur}} u_i\left[b_j(t)\right] dt \tag{4}$$

where T_{dur} is the time the traffic flow spent on the link. We observe the mean connection utility for a large number of flows and use it as a main comparison parameter for comparing the Dynamic bandwidth partitioning scheme with other bandwidth allocation schemes.

3.2. Traffic Differentiation and Utility Functions

In order to extend the basic traffic differentiation to real-time and non-real-time traffic, it is very interesting to note that not all real-time Internet applications expect circuit-switched service from the network. Instead, they are designed to adapt [7][16] to the currently available bandwidth. This flexibility of applications such as streaming video and audio is a very important feature which needs to be taken under consideration when designing new resource allocation schemes for IP networks. That is why in our model we are looking at a case of three traffic classes, real-time traffic with strict performance requirements, real-time traffic that is adaptable to the network state, and non-real-time, elastic traffic. Each of the three traffic classes and appropriate utility functions for them will be introduced in this section.

We call the first traffic class *brittle* traffic. Traffic belonging to this class requires strict end-to-end performance guarantees and does not show any adaptive properties. A brittle traffic flow is not allowed to enter the network if there is not enough bandwidth available. While in the network, a brittle traffic flow occupies a constant amount of bandwidth on the link. Analogies for this traffic class can be found in the CBR traffic class of ATM, Guaranteed Service in the Integrated Services architecture, or Expedited Forwarding per-hop behaviour in the Differentiated Services architecture. Typical applications belonging to this traffic class would be video telephony, highly secure data transactions, telemedicine etc. The user of a brittle traffic connection is interested only in high level of quality of service. If the network is not capable to guarantee the required performance for a traffic flow belonging to this traffic class, the end-users' utility will be 0. That is why for brittle traffic class we use a very simple utility function (Fig.1):

$$u_b(b_b) = \begin{cases} 1, \text{ if } b_b \geq b_{b\min} \\ 0, \text{if } b_b < b_{b\min} \end{cases} \tag{5}$$

where b_b is the allocated bandwidth, and $b_{b\min}$ is the minimum required bandwidth.

The second traffic class is the *stream* traffic. Traffic belonging to this class results from audio and video applications and requires the network to provide a *minimum* level of network performance guarantees. If, due to low network utilisation, the network signals that more resources can be available, these adaptive applications will change their sending rate, or their layer coding technique, thus providing much better quality of service to the end-user. A survey of adaptation techniques can be found in [16]. Stored video and audio sequences accessed remotely across the network can be considered as stream traffic. Playout buffers and retransmissions serve well those types of applications.

For the stream traffic, traffic flows require some minimum level of bandwidth $b_{s\,min}$. Traffic belonging to this traffic class is rate-adaptive. It is able to change its rate between the minimum required rate $b_{s\,min}$, and its peak rate $b_{s\,max}$. Nevertheless, admission control for this traffic class is necessary, and therefore the optimal number of active traffic flows on the link should be finite. If n stream traffic flows use a single link of bandwidth B, the optimal number of flows, n_{opt} that maximises the overall utility, $nu_s\left(B/_n\right)$, must not be infinite, but has to be $n_{opt} = B/b_{s\,min}$. The utility function that can approximate such behaviour is (Fig.2):

$$u_s(b_s) = 1 - e^{-a_{s2}\frac{b_s^2}{a_{s1}+b_s}} \tag{6}$$

where b_s is the allocated bandwidth. The expression in (6) comes from the work of Shenker and Breslau [7]. They used utility functions to analyse the problem of admission control in communication networks. The constant a_{s1} is easily calculated after the value for $b_{s\,min}$ is known, and the constant a_{s2} is a scaling constant. In addition, if we observe the function on Fig.2, it is interesting to make a connection between the utility function and the set of applications (playback video and audio) we are modelling by this utility function. The small levels of bandwidth are not very useful, so that at low bandwidths the marginal utility of additional bandwidth is small. At high bandwidths the signal quality is already good enough and the marginal utility of additional bandwidth at high bandwidths is also small. It is only at intermediate levels, that the marginal utility of extra bandwidth is significant. We can see from Fig. 2 that at low bandwidth values the function is convex. The most important feature of this utility function is its non-concavity, which makes it different from the utility function for the elastic traffic.

Elastic traffic forms the third traffic class in the model. Elastic traffic flows are established for the transfer of digital documents (files, pictures), and only have loose response time requirements. In the case of the elastic traffic flows, there is no minimum bandwidth requirement. There is no need for the admission control, and optimal number of active flows is infinite. The utility function that models such requirements should be always concave, but not linear. The function we propose in this paper is (Fig.3):

$$u_e(b_e) = 1 - e^{-\frac{a_e b_e}{b_{e\max}}} \tag{7}$$

where b_e is the allocated bandwidth, and $b_{e\,max}$ denotes the peak rate for the elastic flow (in the case of the best-effort scheme, $b_{e\,max} = B$). The constant a_e is a scaling constant. Following the same logic as above, since the utility function $u_e(b_e)$ is always concave, the optimal number of traffic flows on the link is $n_{opt} = \infty$, i.e. the optimal admission policy is “accept all”, which is a feature of current best-effort communication network. If we look at Fig.3, we can see that the marginal utility of extra bandwidth is larger when the bandwidth is small. In the area of high bandwidth, adding extra bandwidth does not improve utility as much as when bandwidth is small.

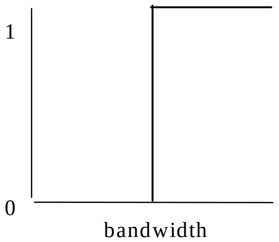

Figure 1 Utility function for the brittle traffic

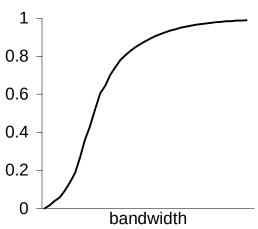

Figure 2 Utility function for the stream traffic

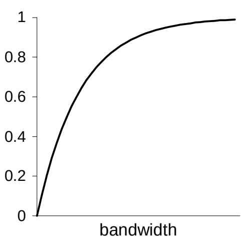

Figure 3 Utility function for the elastic traffic

3.3. Other Bandwidth Allocation Schemes

Three other bandwidth allocation schemes have been used for the performance comparison:
- *Best-effort* scheme (complete sharing) is the scheme without resource reservation and admission control. All traffic flows are accepted to the network and they all receive the equal share of the network capacity. The only limitation is that real-time traffic flows, both brittle and stream, have the maximum rate defined, and therefore are never allocated more bandwidth than their maximum rate.
- *Complete Partitioning* scheme, in which the partitioning parameters are fixed, i.e. the dynamic partitioning algorithm is not deployed. The partitioning parameters in this scheme are arbitrarily chosen to be $\alpha_s = 0.33,\ \alpha_e = 0.33,\ \alpha_b = 0.33$.
- *Trunk reservation*. In this scheme, there is admission control for all traffic classes. An incoming elastic flow is accepted in the network, only if the utility level for stream traffic flows at that moment is greater or equal to some pre-defined parameter η:

If $[u_s(b_s(t)) \geq \boldsymbol{\eta}]$ *then* accept the incoming elastic flow
else reject it

It is important to note that, for bandwidth allocation schemes that deploy admission control procedure, each flow rejection generates a negative utility $v_{ji} = -u_i[b_i(t)]$, where $b_i(t)$ is the bandwidth allocated to the traffic flows belonging to traffic class i at the moment of the rejection of the incoming flow.

3.4. Results and Analysis

Simulation was performed on the connection level. The schemes were compared based on the mean connection utility and the mean file transfer time for the elastic flows. The traffic loads for all three traffic classes in the experiments were equal.

First results on Fig. 4 show the comparison in the average connection utility when the utility functions for all classes, as they are defined in section 3.2, are scaled to 1 (meaning $\omega_i = 1$). That means that all three traffic classes were treated in the same way. We can see from Fig. 4 that in this case trunk reservation and best-effort schemes perform better then the partitioning schemes. This is because the influence of the performance guarantees provision for the brittle traffic class to the overall network performance is proportional to the load of this traffic class. However, as we have discussed before, the differentiation is not only about classification of the traffic into a number of traffic classes, but it is very important to understand the nature of different traffic classes as well. Pricing and prioritisation of services may become important issues here. It is likely that brittle applications, because they require strict network performance guarantees, are going to be priced more than elastic services. Therefore, a possible blocking of a brittle traffic flow, or inability of the network to provide necessary bandwidth requirement has to be valued/penalised more than when it comes to the elastic traffic flows.

Having this in mind, the next step is the introduction of scaled utility functions. Scaling factors ω_i have been allocated the following values: $\omega_b = 5.0,\ \omega_s = 2.0,\ \omega_e = 0.5$. Fig. 5 shows the comparison of the bandwidth allocation schemes after this scaling has been done. The results show that the dynamic bandwidth partitioning scheme generates higher average utility for users then both best-effort scheme and the complete partitioning scheme.

The complete partitioning scheme performs very badly in the area of high loads, due to the fact that high priority brittle and stream traffic cannot use any of the bandwidth reserved for the elastic traffic, even though they need more bandwidth. The best-effort scheme performs badly when it comes to brittle traffic flows. A clear picture of how the bandwidth allocation scheme influences the average per-class utility is given in Fig. 7. We can see that the best-effort scheme performs better than the dynamic partitioning scheme when it comes to elastic and stream traffic flows. The unlimited access that traffic flows have to capacity in the best-effort scheme is obvious there. However, the performance of the best-effort scheme when it comes to brittle traffic is very bad. We can see from the Fig. 7 that only in the areas of low loads is the best-effort scheme able to provide brittle traffic flows with requested capacity. As the load increases, since there is no protection and no capacity guarantees for the brittle flows, their share of the link capacity falls below requested value and corresponding utility becomes 0. If we look on the results for the overall average utility on Fig. 5, we can see the effect of this QoS-unawareness. Dynamic bandwidth partitioning scheme is able to provide requested bandwidth to high-priority traffic flows. The average utility for this traffic class falls below $\omega_b = 5$ only for the very high-load environment, when the percentage of rejected (blocked) traffic flows becomes high.

Fig. 6 shows the comparison of the average file transfer time for the elastic traffic flows. This figure is interesting because it shows that observing just one performance measurement can give us a completely wrong performance evaluation of different bandwidth allocation schemes. Furthermore, these results, in comparison with the results on Fig. 7, give a better view of the meaning of the utility margin.

Fig. 5 clearly shows that the trunk reservation scheme performs much better than both dynamic partitioning and best-effort. This is mainly due to the fact that if the congestion occurs on the link, it is the elastic traffic flows that are being blocked. That does generate a negative utility, but after scaling the effect is smaller than when the real-time flows get blocked. However, even though dynamic bandwidth partitioning does not provide maximal average utility for the users, it still performs better than the best-effort scheme and better than other bandwidth partitioning schemes.

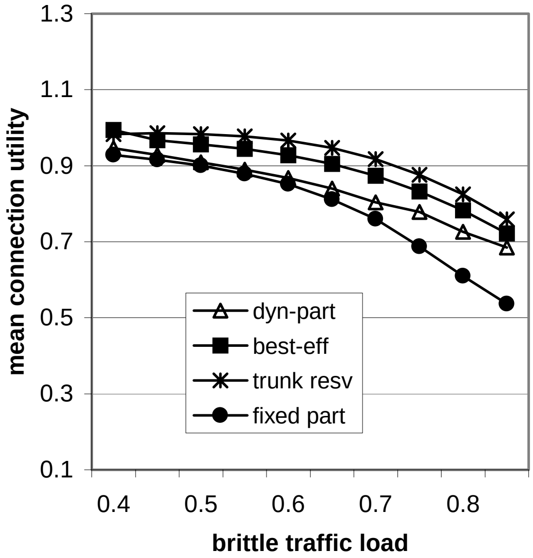

Figure 4 Comparison of the average connection utility

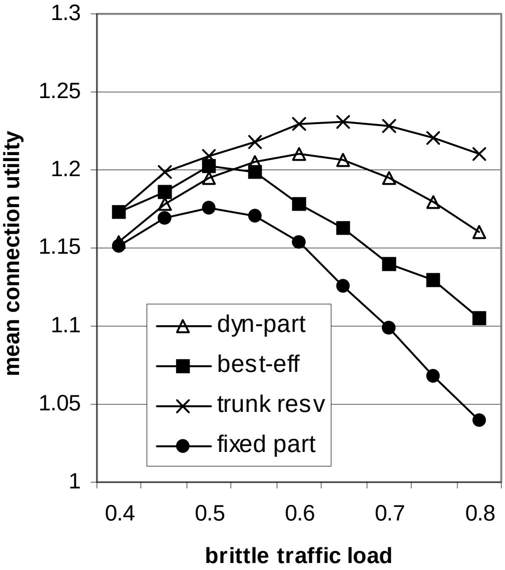

Figure 5 Comparison of the average connection utility, scaled utility functions

Another very important conclusion that can be made from the simulation results presented here is as follows: we can see from the Fig. 5 that the dynamic bandwidth partitioning performs better than the best-effort scheme for traffic loads of 50% and higher. Below that traffic load, the unlimited access to capacity provided by the best-effort scheme is much more efficient. This proves once more the general conclusion about partitioning schemes performing better only in the areas of high loads. Therefore, the network will find it optimal to use Dynamic bandwidth partitioning in the environment of *limited capacity* and *traffic diversity*. Our scheme is unlikely to be the optimal resource allocation scheme for the core IP network, where high-speed optical routers are likely to provide enough bandwidth so that simpler allocation mechanisms can be used. However, in the networks with limited capacity, such as access networks, this type of bandwidth allocation can prove to be very useful. Access links, even after using sophisticated technologies such as ADSL, still present a bottleneck for the network. In the networks with limited resources, we need a resource allocation scheme that can at the same time provide service for as many as possible traffic connections, and give appropriate performance to those connections, maximizing its own utility and revenue at the same time.

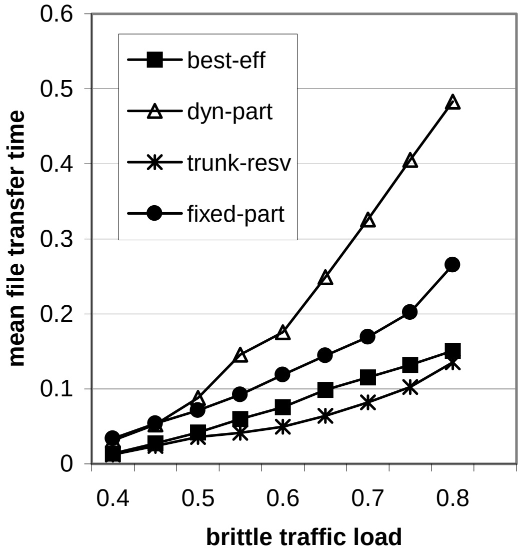

Figure 6 File transfer time for elastic traffic

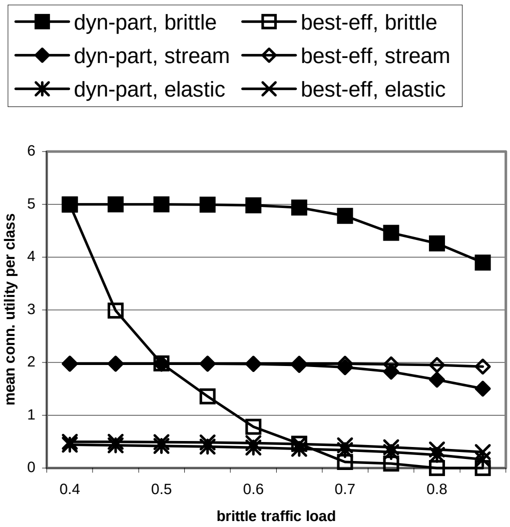

Figure 7 Average connection utility, individual classes

4. Conclusion

This paper presents a new bandwidth allocation scheme for the multiservice IP network, Dynamic bandwidth partitioning. The scheme is adaptive and application-oriented, designed to maximise the overall end-users' utility in the network. In this paper the performance of the Dynamic bandwidth partitioning scheme has been evaluated by measuring the average utility end-users get from the network. The simulation comparison with other bandwidth allocation schemes shows that in the presence of high network load the Dynamic bandwidth partitioning performs better than both the best-effort and the complete bandwidth partitioning schemes.

Bandwidth partitioning became increasingly important in recent days, with a rising interest in Multiprotocol Label Switching and Virtual Private Networks It is not an optimal scheme to be used in the core network, where a large amount of bandwidth is available. However, in the areas of limited capacity (e.g. access networks), where the load on the network is high, Dynamic bandwidth partitioning may prove to be an efficient solution.

The future work will include research on utility-based end-to-end bandwidth allocation, further comparison of the scheme with other bandwidth allocation concepts, and experimenting with different utility functions and increase/decrease parameters in the linear control algorithm.

References

[1] L.Massoulie, J.Roberts, "*Bandwidth Sharing: Objectives and Algorithms*", IEEE INFOCOM 99, New York, USA

[2] S. H. Low, "*Equilibrium Allocation of Variable Resources for Elastic Traffics*", INFOCOM98, San Francisco, USA, 1998

[3] F. Kelly, *"Charging and Rate Control for Elastic Traffic"*, European Trans. on Telecommunications, Vol. 8, pp. 33-37, 1997

[4] S. Shenker, *"Fundamental Design Issues for the Future Internet"*, IEEE Journal on Selected Areas in Telecommunications, Vol.13, No.7, September 1995

[5] R.Nunez-Queija, H. van den Berg, M.Mandjes, "*Performance Evaluation of Strategies for Integration of Elastic and Stream Traffic"*, ITC16, Edinburgh, UK, June 1999

[6] Z. Cao, E. W. Zegura, "*Utility max-Min: An Application-Oriented Bandwidth Allocation Scheme*", IEEE INFOCOM99, New York, USA, 1999

[7] L.Breslau, S.Shenker, "*Best-Effort versus Reservations: A Simple Comparative Analysis*", ACM Computer Communications Review, vol. 28, pp. 131-143, September 1998

[8] S. Sarkar, L. Tassiulas, *"Fair Allocation of Utilities in Multirate Multicast Networks"*, Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, Urbana, Illinois, USA, 1999.

[9] R, Rajkumar, C. Lee, J. Lehoczky, D. Siewiorek, "*A Resource Allocation Model for QoS Management*", IEEE Real-Time Systems Symposium, December 1997

[10] C. Lee, J. Lehoczky, R. Rajkumar, D. Siewiorek, "*On Quality of Service Optimization with Discrete QoS Options"*, IEEE Real-Time Technology and Application Symposium, June 1999

[11] A.Bouch, M.A.Sasse, H.DeMeer, *"Of Packets and People: A User-centred Approach to Quality of Service"*, IWQOS2000, Monterey, CA,USA

[12] K. W. Ross, *"Multiservice Loss Models for Broadband Telecommunication Networks"*, Springer, 1995

[13] V.Rakocevic, J.Griffiths, G.Cope, "*Dynamic Resource Management in Virtual Private Networks*", FITCE Congress 2001, Barcelona, Spain, August 2001

[14] V.Rakocevic, J.Griffiths, G.Cope, "*Dynamic Partitioning of Link Bandwidth in IP/MPLS Networks*", IEEE International Conference on Communications 2001, Helsinki, Finland, June 2001

[15] D. Chiu and R. Jain, *"Analysis of the Increase and Decrease Algorithms for Congestion Avoidance in Computer Networks"*, Journal of Computer Networks and ISDN, vol.17, No.1, June 1989

[16] B.Vandalore, W.Feng, R.Jain, S.Fahmy, *"A Survey of Application Layer Techniques for Adaptive Streaming of Multimedia"*, OSU Technical Report, OSU-CISRC-5/99-TR-14, available through http://www.cis.ohio-state.edu/¬jain/papers.html

[17] P.Kirkby, et.al, "*The use of economic and control theory analogies in the design of policy based Dynamic Resource Control (DRC) network architectures*", ITC16, Edinburgh, UK, 1999