CRATOS
Center for Research on the Applications of Telematics
to Organisations and Society

Charging Schemes for Reservation-based Networks

L. Delgrossi, D. Ferrari
{ldgrossi, dferrari}@pc.unicatt.it

Quaderni del CRATOS

Serie di Telematica

CTR-T98-001

Università Cattolica del Sacro Cuore – Piacenza, Italy

Abstract

The definition of an adequate charging scheme for reservation-based multi-service networks and internetworks is one of the key factors destined to influence the wide use and deployment of these networks in the next decades. This paper presents some of the authors' ideas on this subject and is intended to be a framework on which future works in this area can be based. It comprises an approach to the design of a charging scheme and a first proposal for a charging formula. The approach is based on the fundamental belief that, in the context of reservation-based networks, there should be a direct relationship between user fees and network resources utilisation. The scheme can be applied both to multi-service networks like ATM ones and to such internetworks as the Integrated Services Internet.

1. Introduction

The definition of an adequate charging scheme for reservation-based multiple service networks and internetworks is one of the key factors that are destined to influence the wide use and deployment of these networks in the next decades. So far, however, the attention that the research community has devoted to this problem does not match its importance. Part of the reason is probably due to the charging scheme currently adopted for the Internet, where costs are undertaken by a number of public organisations and private corporations so that, in some cases, the users may have the perception that the service they receive is actually free. Also, the fact that the Internet currently offers all its users the same best-effort type of service does not facilitate the spreading of a culture where charging is based on the actual service received from the network. As the Internet evolves towards an integrated-services internetwork and as other types of multiple-service networks like end-to-end ATM services are considered, it seems reasonable to claim that a charging scheme will soon be needed: users who reserve guaranteed channels to run collaborative applications based on video-conferencing should pay higher fees than users who browse the network and occasionally download information from a web server over a best-effort channel. When to services that are different in their nature and provision costs do not correspond different user fees, it is not possible to prevent users from requesting the most sophisticated service type even when they do not actually need it, i.e., from requesting a guaranteed channel to run a textual network browser.

This paper presents some of the authors' ideas on this subject and is intended to provide a framework on which future works in this area can be based. It comprises an approach to the design of an adequate charging scheme for reservation-based multiple service networks and a first proposal for a charging formula. When writing this paper, we had the following three objectives in our minds: a) to identify a set of properties that are desirable for any charging schemes; b) to define a number of principles to be taken into account when designing the scheme; and c) to propose a method to measure the impact of a single data stream on a network resource in terms of its capacity. In any case we felt that, in the context of reservation-based networks, there should be a direct relationship between fees and network resources utilisation: resources capacity is limited and, unlike best-effort service, each reservation lowers the probability that other users may get access to the network. The scheme proposed in this paper is based on the generic notion of resource and can be applied both to multiple-service networks like ATM ones and such internetworks as the Internet.

The paper is organised as follows: the rest of this introduction presents the current state of the art by recalling a number of previous related works and current research efforts in this area (Section 1.1); it also defines the context of reservation-based networks by introducing a service model (Section 1.2) and a resource and connection model (Section 1.3). Section 2 discusses a number of desirable properties for any network charging schemes (Section 2.1) and a set of design principles (Section 2.2). The proposed charging scheme is presented and discussed in Section 3, where some examples of its application are also given. Conclusions are summarised in Section 4.

1.1 Related Work

Cocchi et al. [1] provide an initial study of pricing in multiple service networks (1991). They compare two different pricing policies (flat *per-byte* fees and *graduated* fees per priority class) and show that users are more satisfied with the latter scheme. User satisfaction is a combination of both pricing and quality of service received. Perhaps the most interesting remark is that the presence of multiple services within a network make usage fees necessary as incentives for proper user behaviour.

Parris, Ferrari and Keshav, [2] [3] concentrate on the effects of pricing on network utilisation and compare a number of pricing schemes, including *per-packet*, *setup*, and *peak-load* pricing (1992); a criterion is given to compare different pricing policies based on the idea that, for the same revenue generated, a scheme is better than another scheme if its blocking probability (due to admission control) is lower. This work is the first to the authors' knowledge in which network resources utilisation and issues related to admission control are taken into account.

More recently, some research efforts related to ATM networks have been started (1996). The ACTS CANCAN project funded by the ACTS programme of the European Community [9] aims at the definition of an appropriate charging policy for end-to-end ATM environments, including IP over ATM traffic. CANCAN's objectives are similar to those of Ca$hman, a project in which a number of partners from the telecommunications industry and the academic world are involved. Final results from these two projects have not been published yet, but it is possible to monitor progresses at site [16] for CANCAN and [17] for Ca$hman.

Information on the works by Varian and others can be retrieved at [18]. In [7], Mackie-Mason and Varian answer many relevant questions on the role of charging schemes and their impact on network utilisation, while in [5] they study the problem of charging for congestionable networks. The bidding scheme proposed in such a context has several nice features but it seems difficult to apply it in practice as it requires a high involvement of the users. Also, it is not targeted to reservation-based networks, where congestion may be avoided by the admission control function.

Klopfenstein provides an annotated bibliography on the subject of network charging and economy [8].

Our work is original with respect to these efforts because it focuses on the notion of resource capacity and it takes into account reservation-based networks. At this stage, we do not have information on comparisons with other approaches. In the near future, we expect to test the efficiency of our scheme through the use of simulations in software and to be able to compare our results with those of other similar projects.

1.2 Service Model

We are interested in the definition of a charging scheme for multiple-service networks or internetworks that are able to carry both real-time and non-real-time traffic. We say that the network can offer its users several *types of service* (ToS). Examples of ToS are traffic with deterministic guarantees, traffic with statistical guarantees and best-effort traffic. For each type of traffic (except best-effort), the user may request channels with a certain *quality of service* (QoS), expressed in terms of a flow specification (*flowspec*) including a series of QoS parameters. We assume that the flowspec comprises at least the following parameters:

- S maximum packet size (bytes)

- R transmission rate (packets/sec)[1]
- D end-to-end delay (sec)
- J delay variation (sec)

Depending on the number of receivers associated with it, a connection may be classified as unicast, multicast or broadcast. For the sake of simplicity, in the rest of this discussion, we will consider unicast connections only, while the other types of connections are left for further study. For the same reason, we shall not discuss traffic with statistical guarantees.

1.3 Resource and Connection Model

A network may be represented as a set of links connecting a set of nodes where each node comprises a number of resources that are responsible for the transmission of the data packets from one node to another. Resources may include the CPU, network adapters and memory buffers. Each resource along the communication path contributes to the delivery of the data by processing incoming data and forwarding it to the next resource until the final destination is reached. We may represent a single resource as an entity that can serve a number of incoming messages in a given time unit with a number of associated queues to store messages that are not processed immediately. **Figure 1** represents a generic resource with its input and output queues. The assignment to a specific queue and the correspondent service time depends on the scheduling policy that is adopted.

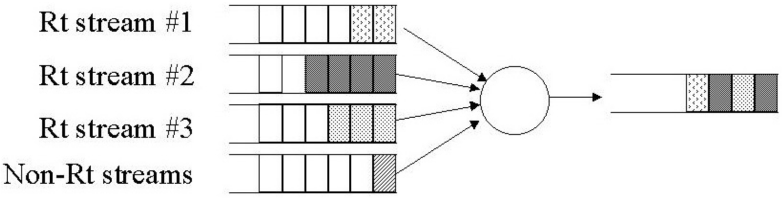

Figure 1: Internal Behaviour of a Generic Resource

The end-to-end path of a real-time connection can be modelled as a cascade of queueing servers (resources) [10], each with a finite queue scheduled, for example, by the earliest due date policy (EDD), where links are characterised by their propagation delay and bandwidth and the first and last queues belong to the source host and destination host respectively (**Figure 2**).

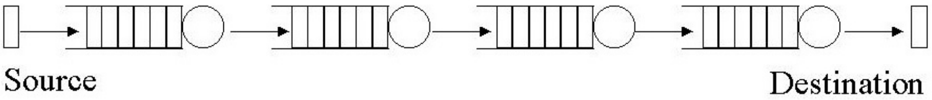

Figure 2: Real-Time Connection Model

2. Properties and Principles for Charging Schemes

In this section, we discuss a number of properties that are felt desirable for any network service charging schemes and we enumerate some principles that we feel should be considered when designing a new charging scheme. The list presented below is not exhaustive and represents solely our point of view in terms of what may

---

[1] As an alternative, transmission rate may be expressed as minimum inter-packet transmission time (sec), as in the Tenet protocol suite [11].

be most desirable from the perspectives both of the network service provider and the user.

2.1 Desirable Properties

We feel that network service providers would be interested in a charging scheme that can offer at least the following properties:

1) *High probability of cost recovery*: the scheme should reflect well the behaviour of the network and ensure that all associated costs can be recovered;
2) *Encouragement* of user behaviour that enhance the network's efficiency;
3) *Competitiveness of prices*: prices should not be higher than those of other networks offering similar services;
4) *Inexpensiveness of implementation*: implementing the scheme in the network should incur reasonable costs;
5) *Inexpensiveness of use*: the accounting and billing operations associated with the scheme should not be too expensive.

Users are likely to be mostly interested in:

6) *Fairness*: two users who establish two identical or very similar connections at the same time with the same type and quality of service should be charged the same price;
7) *Reproducibility / Predictability*: users should be able to roughly estimate in advance how much they are going to pay for the service they are going to receive;
8) *Stability*: prices should be fairly stable over time and not fluctuate in a significant way;
9) *Comprehensibility / Transparency*: all price components and the way each of them contributes to the computation of total charges should be clearly described;
10) *Controllability*: the user should have the means to verify that the charges applied correspond to the service received.

2.2 Principles

As a first principle, user fees should depend on the type of service (ToS) that is offered (e.g., the transportation of real-time packets should cost more than that of non-real-time packets, because the former have priority over the latter in all routers). Also, fees should depend on the quality of service (QoS) associated with each connection, so that, among connections with the same ToS, those that are more demanding in terms of delivery conditions (and therefore tend to stress more the internal resources of the network) result in higher fees.

We feel that usage-sensitive pricing is desirable but not a must; for convenience, it may be replaced by flat-fee pricing. As in all transportation systems, the distance between end-peers should also be taken into account. This imposes a choice when designing a charging scheme, because this distance may be expressed in several ways (e.g., number of hops, Euclidean distance, length of the actual path from the source to the destination) with a number of different implications.

A final remark on charging schemes' properties may help clarifying the charging scheme role and relevance in this context: a well chosen charging scheme leads to prices that are competitive enough to incentivate the use of the network and, at the same time, sufficiently high to regulate access to the network and help reduce congestion.

3. A Reservation-based Charging Scheme

The charging formula that we introduce and discuss in this section can be used to calculate the charge associated with real-time and non-real-time traffic. In the case of real-time traffic, the formula permits the calculation of the charge associated with a single connection. In a multi-service reservation-based network, charge K is given by the following two components:

$$K = K_{reservation} + K_{transport}$$

where:

$$\begin{aligned} K_{reservation} &= 0 && \text{(for non-real-time traffic)} \\ &= T \cdot K_{connection} && \text{(for real-time traffic)} \end{aligned}$$

and

$$\begin{aligned} K_{transport} &= \rho \cdot V && \text{(for non-real-time traffic)} \\ &= (\rho + \sigma) \cdot V && \text{(for real-time traffic)} \end{aligned}$$

In this formula, T is the overall duration of the connection in seconds and $K_{connection}$ is the charge of the resource reservation associated with the connection (we show how to calculate $K_{connection}$ in the rest of this section). Also, ρ and σ are coefficients ($\rho > 0$, $\sigma > 0$) and V is the volume of information transported expressed in bytes. Real-time traffic results in a higher $K_{transport}$ because real-time packets are assigned higher priority within the network nodes.

3.1 Charges associated with a single resource

Before we show how $K_{connection}$ ca be calculated, it is useful to recall the internal behaviour of a generic resource. In reservation-based networks, when a data stream is associated to a resource as a result of the channel setup operations by the setup protocol, a portion of the resource's capacity is reserved, i.e., made only available for the processing of the data belonging to the stream. As data streams are accepted by the admission control algorithm associated with the resource, the portion of free capacity decreases and, eventually, some requests are rejected because the resource does not have sufficient free capacity to meet the data stream's service requirements.

Unfortunately, to provide a precise definition for "resource capacity" is not straightforward. We express "resource capacity" in terms of three resource types:

- *Buffers*, i.e., the amount of memory buffers that can be used to store incoming packets until they are served,
- *Processing power*, i.e., the number of incoming packets that can be processed in a given time unit (this may correspond to computational power in the case of a CPU or to bandwidth for transmission links),
- *Schedulability*, i.e., the capability to assign appropriate execution priorities to incoming packets so that all service time requirements of each single data stream are met.

The capacity of each of these resource types is always limited. When a sufficient amount of memory buffers is not available or when there is not enough bandwidth to process additional incoming packets, the admission control algorithm cannot accept new data streams. However, there also exist situations in which, although sufficient buffers and enough bandwidth are available, the admission control function cannot accept new data streams because, due to factors related to the scheduling policy, it would not be possible to serve packets belonging to the new data stream within their service time constraints. In such cases, we say that the schedulability resource is insufficient. The schedulability element is strictly related to the user requirements in terms of end-to-end transmission delay, and requests for low delay values tend to make it more difficult for the scheduler to find appropriate priority assignments.

The charging scheme that we propose reflects this situation: the charge associated with a single queueing server (see **Figure 1**) depends on the fraction of capacity (expressed in terms of buffers, processing power, and schedulability) that the server utilises to serve a stream's data packets. Local charges are then combined to calculate the overall charge of the connection (Section 3.2).

Let us now calculate the charge associated with a single queueing server. If β is the cost per unit time attributed to the *buffers* component in a given queueing server (usually different from the cost of the space itself), and b is the fraction of the total buffer space reserved for our connection, then the cost to be attributed to the connection is given by $\beta \cdot b \cdot T$; the same may be said about the *processing power* component: the cost to be attributed to fraction c is $\gamma \cdot c \cdot T$.

For *schedulability* the problem is more complicated. A scheduler is never saturated, as it is always possible to add another connection to those already accepted by it provided the new connection is able to tolerate a long enough local delay bound and has a minimum inter-packet interval that is long enough. The addition of a new connection only causes the minimum delay bound that can be offered by the scheduler to increase, and this makes it less and less probable that new connections will be able to go through the server. Thus, it is hard to represent the amount of *schedulability* taken away by a new connection as a function of an available total.

By considering the scheduling discipline used by the queueing server, we may derive indexes that are reasonable measures of the portion of schedulability taken by the establishment of a single connection. For example, if the discipline is FIFO, and the scheduler has a maximum time t_{cycle} within which it must fit all the real-time packets (none of which will therefore incur a local delay greater than $t_{cycle} - t_{max}$, where t_{max} is the longest packet processing time for any packet, including non-real-time ones), we can define:

$$d = t / (t_{cycle} - t_{max})$$

for all connections, having also assumed that all real-time packets require the same processing time t. The index d can be used as a measure of how much capacity is reserved for a single connection in terms of schedulability. If the scheduler is based on the EDD (Earliest Due Date) policy, we note that, to represent the greater difficulties caused by a smaller local delay bound, we could determine the number of connections that will be allowed to have a local delay bound smaller than d_k by a new connection in the queueing server *k*, characterised by (t_k, d_k), t_k being the largest packet processing time for the connection's packets and d_k being the local delay bound assigned by the setup procedure to its packets, if it is accepted (assuming for simplicity that all t_k's are equal to t):

$$\text{Floor}\,[(d_k - t_{max}) / t\,]$$

The reciprocal of this number (eliminating the Floor for simplicity),

$$d = [\, t / (d_k - t_{max})],$$

is a reasonable measure of the impact of admitting connection (t_k, d_k) to the server. It is also a reproducible measure, since it does not depend on anything but the characteristics of connection k and of the server (note that t_{max} can be assumed to characterise the server).

In conclusion, we may define a function of the available buffers, processing power, and schedulability and say that, for the i_{th} queueing server along the communication path, the associated cost can be expressed as:

$$T \cdot (\beta \cdot b_i + \gamma \cdot c_i + \delta \cdot d_i)$$

where β, γ, and δ are positive coefficients, which we assume to be independent of *i*, i.e., the same for all queueing servers, an assumption that is valid for homogeneous networks.

3.2 Charges associated with a reserved channel

Once we have calculated the charge for a single queueing server, we sum them over the connection path to compute the overall connection charge $K_{connection}$:

$$K_{connection} = T \cdot \sum_{i=1}^{N} (\beta \cdot b_i + \gamma \cdot c_i + \delta \cdot d_i)$$

One way to compute this charge is to calculate it as part of the channel setup phase. This can easily be done by such reservation protocols as RCAP [12], ST2+ [14] and RSVP [13]. A *charge* field could be added to the *flowspec* and each router over the connection path could add its local charge to the current value of this field. The final charge should be presented to the user at a time where he/she can still decide whether to accept or refuse the connection. The idea of a charge function computed by the nodes of the network was already present in SRP [15].

At this time, we are not clear whether there should be a correlation (and which) among the b, c, and d components of the sum. It is possible that future simulations show that there are situations in which they do not have the same importance. This can be partially taken into account by assigning appropriate values to the coefficients β, γ, and δ. It is also possible that a more appropriate combination function than $\beta b+\gamma c+\delta d$ is found. In any case, some function that takes into account the internal behaviour of the resources should be adopted.

Calculating the sum over all nodes on the connection path implies the assumption that all nodes contribute in the same way to the delivery of the data. This is not necessarily true, however, we feel that it is complicated to estimate the contributions of single resources,, and that it is likely that this will be regulated by the charges attributed by network service providers to each resource.

Finally, the charging formula presented in this paper depends on the routing algorithm which is adopted within the network. There seem to be both good and bad reasons for relating charges to the path. We leave this type of issue for further study.

3.3 Revenues

The network service provider revenues over a time period Y are given by:

$$R = R_{nrt} + R_{rt}$$

$$R_{nrt} = \Sigma (\rho V)$$

$$R_{rt} = \Sigma [(\rho+\sigma)V] + Y [\beta \Sigma (Ub) + \gamma \Sigma (Uc) + \delta \Sigma (Ud)]$$

where the sums are computed over the connections that were alive during period Y, except the last three in the expression of R_{rt}, which are computed over all queueing servers in the network, and Ub, Uc, Ud are the mean "real-time utilisation" of the resources in each queueing server. The instantaneous real-time utilisation of a resource at time t is the ratio between the amount reserved at that time and the total supply of the resource that can be reserved. The means Ub, Uc, Ud are the utilisations averaged over time period Y.

If we divide the total desired revenue R among the five contributions R_{nrt}, $\Sigma(\rho+\sigma)V$, βSum(Ub), $\gamma\Sigma$ (Uc), and $\delta\Sigma$ (Ud), and if we know (from simulations or measurements) the values of the sums, then we can derive all the coefficients ρ, σ, β, γ, and δ.

3.4 An Example

As an example, let us now consider an application requesting a guaranteed real-time channel over a connection path that traverses an ATM switch. The switch offers a MAXCELL maximum speed per port (cells/sec) and has MAXBUF buffer space (cells). The switch allows the users to specify a contract by means of the following usage parameters:

- Type of traffic (CBR, VBR, or UBR)
- Peak Cell Rate (cells/sec)
- Sustainable Cell Rate (cells/sec)
- Maximum Buffer Size (cells)

The user may for instance request a connection characterised by CBR, 500 cells/s and 200 buffer cells. It is possible that the user application makes use of a different QoS interface, e.g., it may use the flowspec described in Section 1.2. In this case, a translation of the QoS parameters into the correspondent format requested by the network is necessary. This is a common situation when building reserved channels over an integrated-services internetwork. The formula introduced in Section 3.1 may be computed as follows: $\beta\cdot(500/MAXCELL)+\gamma\cdot(200/MAXBUF)+\delta\cdot d$, where the value of d depends on the scheduling policy adopted within the ATM switch.

4. Conclusions

In this work, we propose that reserved data channels are charged based on the portion of resource capacity they make use of. Our discussion showed that capacity may be expressed as a combination of a number of factors, and that one of them depends on the scheduling policy used within the resource. A short example showed how to calculate the formula at the basis of the charging scheme in the case of an ATM switch. In the near future, we intend to build a simulator in software to verify whether the proposed charging scheme has indeed the desirable properties that we have described in this paper. Further study is required in several areas, including that of the relationship between the routing algorithm and the charging scheme.

5. References

[1] R. Cocchi, D. Estrin, S. Shenker, L. Zhang: "*A Study of Priority Pricing in Multiple Service Class Networks*", TR 91-07, 1991.

[2] C. Parris, D. Ferrari: "*A Resource-based Pricing Policy for Real-Time Channels in a Packet-Switching Network*", TR-92-018, International Computer Science Institute, Berkeley, 1992.

[3] C. Parris, S. Keshav, D. Ferrari: "*A Framework for the Study of Pricing in Integrated Networks*", TR-92-016, International Computer Science Institute, Berkeley, 1992.

[4] R. J. Edell, N. McKeown, P. P. Varaiya: "*Billing Users and Pricing for TCP*", IEEE Journal on Selected Areas in Communications, 1995.

[5] J. K. MacKie-Mason, H. R. Varian: "*Pricing Congestible Network Resources*", University of Michigan, 1994.

[6] D. Botvich, Y. Chen, T. Curran, B. Kerswell, J. McGibney, D. Morris*: "On Charging for Internet Services provided over an ATM network*", Teltec Ireland Web Site <URL> http://www.teltec.dcu.ie/cancan/papers/ieee_botvich/paper.htm, 1997.

[7] J. K. MacKie-Mason, H. R. Varian: "*Economic FAQs About the Internet*", Web Site <URL> http://www.virtualschool.edu/mon/Economics/VarianInternetEconomics.html, 1994.

[8] B. C. Klopfenstein: "*Internet Economics: An Annotated Bibliography*", to appear in the Journal of Media Economics, <URL> http://www.bgsu.edu/departments/tcom/annota.html, 1997.

[9] ACTS Project AC014: CANCAN: "*Phase A Field Trial and User Forum Report*", Deliverable 7, October 1996.

[10] D. Ferrari: "*A New Admission Control Method for Real-Time Communication in an Internetwork*", in S. H. Son Ed., Advances in Rea-Time Sytems, Ch. 5, pp: 105-116, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[11] D. Ferrari: "*The Tenet Group*", University of California, Berkeley and International Computer Science Institute, Berkeley, California, 1992.

[12] D.Ferrari, D. Verma*: "A Scheme for Real-Time Channel Establishment in Wide Area Networks*", IEEE Journal on Selected Areas in Communications, Vol. 8, No. 3, 1990.

[13] L. Zhang, S. Deering, D.Estrin, S. Shenker, D. Zappala: *"RSVP: A New Resource Reservation Protocol*", IEEE Network, 1993.

[14] L. Delgrossi, L. Berger*: "Internet Stream Protocol Version 2 (ST2) protocol specification -Version ST2+*", Internet RFC 1819, 1995.

[15] D. P. Anderson, R. G. Herrtwich, L. Delgrossi: "*SRP: A Resource Reservation Protocol for Guaranteed-Performance Communications in the Internet*", ICSI Berkeley Technical Report, January 1991.

[16] <URL> http://www.teltec.com/cancan;

[17] <URL> http://www.isoft.intranet.gr/cashman
[18] <URL> http://www.sims.berkeley.edu