A meta-model for multiobjective routing in MPLS networks

**José Craveirinha · Rita Girão-Silva ·
João Clímaco**

Abstract   MPLS (Multiprotocol Label Switching) enables the utilisation of explicit routes and other advanced routing mechanisms in multiservice packet networks, capable of dealing with multiple and heterogeneous QoS (Quality of Service) parameters. Firstly the paper presents a discussion of conceptual and methodological issues raised by multiobjective routing optimisation models for MPLS networks. The major contribution is the proposal of a multiobjective routing optimisation framework for MPLS networks. The major features of this modelling framework are: the formulation of a three-level hierarchical routing optimisation problem including network and service performance objectives, the inclusion of fairness objectives in the different levels of optimisation and a two-level stochastic representation of the traffic in the network (traffic flow and packet stream levels). A variant of the general model for two classes

J. Craveirinha · R. Girão-Silva · J. Clímaco
Institute of Computers and Systems Engineering of Coimbra (INESC-Coimbra),
R. Antero de Quental, 199, 3000-033 Coimbra, Portugal

J. Craveirinha (✉) · R. Girão-Silva
Department of Electrical Engineering Science and Computer, University of Coimbra,
Pólo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
e-mail: jcrav@deec.uc.pt

R. Girão-Silva
e-mail: rita@deec.uc.pt

J. Clímaco
Faculty of Economics, University of Coimbra, Av. Dias da Silva, 165,
3004-512 Coimbra, Portugal
e-mail: jclimaco@inescc.pt

of traffic flows, QoS traffic and Best Effort traffic, is also presented. Finally a stochastic teletraffic modelling approach, underlying the optimisation model, is fully described.

1 Introduction

1.1 Background and motivation

Modern multiservice network routing functionalities imply the necessity of dealing with multiple, heterogeneous and multifaceted QoS (Quality of Service) requirements. This has led to routing models, the aim of which is the calculation and selection of a sequence of network resources (designated as routes) satisfying certain QoS constraints and the optimisation of route-related metrics. These routes correspond to 'constrained' paths in the capacitated graph representative of the network structure, i.e., in a graph the arcs of which have a limited capacity expressed in terms of bandwidth or number of 'channels' (where each channel corresponds to a traffic carrying capacity relative to each type of traffic flow offered to the network). The term traffic flow will designate in this context, a sequence of node-to-node connection requests, of a certain service class, with certain requirements. Note that some of the typical objectives in routing models have a conflicting nature and are interdependent. For example the objective of maximizing the total revenue associated with all traffic flows carried in the network (for a given routing solution for every node-to-node traffic flow) may conflict with the objective of minimising the blocking probability of some traffic flows.

There are potential advantages in formulating important routing problems in these types of networks as multiple objective optimisation problems. This enables the trade-offs among distinct QoS parameters and relevant network cost function(s) to be pursued in a fully consistent manner. It should be noted that the specification of the objective functions and constraints depends strongly on the type of technological platform used by the network, and the features of the carried traffic flows associated with different service types.

In this work we will address key methodological and modelling issues associated with route calculation and selection in networks using a recent multiservice Internet technology designated as MPLS (Multiprotocol Label Switching). In the modelling of the routing calculation problem in MPLS networks, we need to take into account some traffic engineering-related concepts, now briefly reviewed according to Awduche et al. (1999, 2002), and Rosen et al. (2001).

In such networks traffic flows are composed (at the physical level) of packet streams that are forwarded from node to node, according to specific technical rules. When the packets enter the network, they are grouped in different FECs (Forward Equivalence Classes) according to specific criteria, such as the originating node, the destination node and the grade of service that has to be provided. The concept of 'traffic trunk' as an aggregation of traffic flows of a certain class is important in the present context. The traffic trunks can be characterised by the ingress and egress nodes, the FEC

they are associated with, and a set of parameters/attributes with impact on the traffic engineering schemes, which define some essential requirements of the routing models. The routing mechanism for packets used in the MPLS networks is based on the establishment of the so-called LSPs (Label Switched Paths).[1] At the ingress node, the packets are "labelled", in the sense that a label containing information on the FEC is associated with them. At each intermediate node, the LSRs (Label Switching Routers) forward the packets using a specific label switching technique: the label is an index into a routing table with information on the next hop and the next label to be assigned to the packet.

This enables the establishment of end-to-end "explicit routes" in association with the implementation of advanced QoS routing mechanisms. In particular explicit routes enable source routing mechanisms characterised by the fact that the route followed by each packet stream (of a given connection) is entirely determined by the ingress router. This is an inherent advantage by comparison with the hop-by-hop (i.e., node by node) routing system typical of IP routing.

Having in mind these features and capabilities of MPLS routing a significant number of routing models has been proposed in the literature in recent years. These approaches often differ in key instances of the modelling framework. In particular such differences are concerned with: (i) the scope of the routing optimisation (where we may distinguish network-wide optimisation models and flow-oriented models as discussed in Sect. 2); (ii) the nature of the optimisation model in terms of the objective function(s) and constraints (single/multiobjective, type of QoS-related or other constraints, etc.); (iii) the level of representation of the traffic flows (representation at the level of 'micro-flows' of packet streams carried on a certain LSP or at the level of the traffic trunks—aggregation of traffic flows of the same class placed on the same LSP). These remarkable differences (which in most papers are not discussed by the authors but presented a priori as a "natural" form of formulating the routing model) and our interest in developing new multiobjective routing models capable of capturing several QoS instances and cost functions as well as different levels of traffic representation gave the motivation for the present work.

1.2 Contributions of the paper

In this paper we will begin by analysing and discussing key conceptual and methodological issues associated with routing optimisation models for MPLS networks proposed in the literature. In particular we will present an overview and analysis of multiobjective routing models.

The major contribution of this work is the presentation and discussion of a meta-model for multiobjective routing in MPLS networks which provides a framework for multiobjective routing optimisation in this type of network. A first important feature of this model is the use of hierarchical optimisation typically with three optimisation levels: the first priority objective functions refer to the network level; the second priority

---

[1] If multi-path routing is allowed for the traffic flows, the packets will follow one of the paths associated with the FEC.

objective functions refer to performance metrics for the different types of services supported by the network; the third priority is concerned with performance metrics for the 'micro-flows' of packet streams of the same FEC.

A second feature of the model is the "dual" stochastic representation of traffic flows in the network: 'macro' level, simply designated as *traffic flow* level, and 'micro' level, corresponding to micro-flows of packet streams.

The third feature is the explicit consideration of 'fairness' objectives, at the three levels of optimisation, expressed through the minimisation of the worst performance level associated with the QoS-related metrics defined at each level.

We also propose a variant of this meta-model which considers two major classes of traffic flows: QoS traffic, with guaranteed requisites of certain QoS parameters, and BE (Best Effort) traffic, which is carried on a "best effort" basis, and is in fact second priority traffic.

The paper also presents a complete traffic model which underlies the proposed routing optimisation model, based on the use of the concept of effective bandwidth for macro-flows, on a multidimensional Erlang model for estimating the blocking probability on the links and on a queue with hyper-exponential service times for approximate calculation of average packet delays.

Finally we outline open issues and further work concerning the application and resolution of this type of model.

2 Overview of multiobjective routing models for MPLS

An overview of multiobjective routing models that have been recently proposed in the literature is presented in this section. Models for MPLS networks, and models for other multiservice networks which may be in principle applicable to MPLS, are examined. Different authors use significantly different assumptions as well as different objectives and constraints in the formulation of routing models, leading to distinct approaches to the routing problem. The focus of this overview is on multiobjective routing models, for which the underlying principles, the objective functions and main constraints are outlined. The resolution approaches for the problems are also briefly mentioned.

Erbas (2003), Erbas and Erbas (2003), Erbas and Mathar (2002) formulate a routing problem in MPLS networks, with the traffic represented at the traffic trunk level in a deterministic way, by assuming that a known fixed average bandwidth is offered from origination to destination node, corresponding to a given traffic trunk. The routing model assumes the splitting of the bandwidth required by each traffic trunk by several LSPs. The routing optimisation problem is of the mixed integer type and it is assumed to be solved off-line. The authors consider three different objectives subject to constraints in their formulation of this problem: the minimisation of the expected delay experienced by the packets in the traffic trunks along the chosen paths, the optimisation of the utilisation of resources (which is achieved by optimising the load distribution in the network) and the minimisation of the number of LSPs used. These last two objectives are clearly in conflict: on the one hand, to guarantee a proper load distribution in the network and to prevent the over-utilisation of some links, many paths must be taken into account and the traffic should be split among those paths; on the other hand

the minimisation of the number of LSPs used is also desired. A link load balancing function is defined as a convex, increasing and piece-wise linear function. The purpose of this function is to penalise increasingly the sending of packets over a link as the utilisation of that link becomes higher. In Erbas and Mathar (2002), the routing problem formulation is multiobjective and solutions are found through the minimisation of a single objective function, which is a weighted sum of the three mentioned objectives plus a forth objective: a penalisation cost for the amount of traffic which is not being routed. In the original formulation of the problem, Erbas et al. assume there are different classes of service, namely QoS and BE traffic. However, the problem is solved for each class of service separately. In Erbas and Erbas (2003) the objective of the minimisation of the number of used LSPs is transformed into a constraint, where an upper bound is imposed on the number of used paths. Therefore, the problem is formulated as a bi-objective one and the calculation of the Pareto optimal solutions is made by the method of the Chebyshev lexicographic weighted metric. The same three objectives are at the basis of the formulation of the routing problem in Erbas (2003). However, the minimisation of the delay experienced by the traffic trunks in the chosen path and the optimal utilisation of the resources (achieved by an optimal load distribution in the network) are transformed into constraints. The two objectives are the minimisation of the number of LSPs used and the minimisation of the penalisation cost of the traffic which is not being routed. This problem is solved using a hybrid heuristic method combining an evolutionary method with mathematical programming.

A different approach is followed by Haßlinger and Schnitter (2003) and Schnitter and Haßlinger (2002). Here the routing problem in an MPLS network is bi-objective, of the mixed integer type with constraints, and it is solved off-line. The traffic is again represented at the traffic trunk level in a deterministic way by assuming the average bandwidth offered from origin to destination is known, as in Erbas (2003); Erbas and Erbas (2003) and Erbas and Mathar (2002). Traffic splitting is not allowed in this formulation, because of the delay variations it may cause. The two objectives are the minimisation of the maximum link utilisation (which intends to achieve an optimal load distribution in the network) and the minimisation of the number of hops in a path (which intends to achieve the minimisation of the delay and of the fault probability in the path). This objective is secondary, in the sense that the pursuit of the minimisation of the number of hops in a path should be made without affecting the primary objective of load balancing. The utilisation of the links is defined as the ratio of the used bandwidth of the link and the capacity of that link. The minimisation of the number of hops in a path is assumed to be equivalent to the minimisation of the sum of the aggregated traffic in all the links. To solve this problem, a heuristic based on the cost of the links is used. The cost of a link increases as the load carried increases. The goal of the heuristic is to find the lowest cost paths. The order in which the requests are serviced is important for the quality of the final results, as analysed in Haßlinger and Schnitter (2002).

Mitra and Ramakrishnan (2001) formulate the routing problem in an MPLS network as a bi-criteria optimisation problem. The traffic is represented at the traffic trunk level and traffic splitting is allowed. The problem can be solved on-line for batches of demands and on a semi-on-line basis if reconfigurations of the network for larger batch sizes are needed. The traffic engineering design is deterministic, of the

Multicommodity Flow (MCF)-based type. Different classes of service are considered, namely QoS and BE traffic. The QoS service classes have priority over the BE traffic when accessing the available bandwidth. The two objectives to be maximised are network revenues for both types of traffic, where the revenues are calculated in terms of the carried bandwidth on all the routes between every pair of nodes and for every service. The objectives are prioritised: the first objective is the maximisation of the revenue for the QoS service classes and the second objective is the maximisation of the revenue for the BE class. To solve this problem, the authors propose, in a first step, the resolution of the routing problem for the QoS traffic viewed as the only type of traffic on the network, followed by the resolution of a "combined" problem of maximisation of the revenue for the BE class, subject to the constraint that the revenue for the QoS traffic is optimal. A simplified approach based on a multilayered decomposition of the combined problem is also mentioned by the authors. At the first layer, the routing problem is solved only for the QoS service classes; at the second layer the objective is the minimisation of the bandwidth-hops used by the QoS traffic, while guaranteeing the optimality of the network revenue for the QoS traffic; in the last layer, the routing problem is solved for the BE service class in the residual network that results from the resolution in the previous layers. A simplified form of representing the stochastic nature of the traffic, in the context of the model, is also proposed.

Now we present an overview of some multiobjective routing models formulated for packet or other types of multiservice networks which may be applicable (with some adaptations) to MPLS networks.

Knowles et al. (2000) propose a multiobjective routing model, of the mixed integer type, for packet switched networks with the traffic represented at the traffic trunk level in a deterministic way. No traffic splitting is considered. The problem is solved off-line. The three considered objectives are the minimisation of the deviation between the bandwidth required by the traffic routed in a link and the capacity of that link (corrected by an attribute, designated as Maximum Allocation Multiplier, that measures the desired degree of allocation of the real capacity of each link, taking into account the traffic patterns), the minimisation of the deviation between the utilisation of the links and a pre-defined value of utilisation and the minimisation of the costs of communication, associated with the utilisation of each link. The authors address the resolution of this problem by an evolutionary method.

Widyono (1994) also proposes a multiobjective routing model with constraints, of the mixed integer type, for packet networks. The traffic is represented at the packet level in a stochastic form and traffic splitting is allowed. The problem is solved on-line. The two objectives are the optimisation of the utilisation of the network (related to the efficient management of the network resources) and the minimisation of the number of hops in a path. Constraints on the maximum delay and the maximum variation of delay experienced by the packets, and also on the buffer overflow probability, are imposed. The utilisation of the resources must be well distributed in the network: this is the reason why traffic splitting is allowed. For the resolution of the routing problem, the author uses a heuristic algorithm based on a constrained minimum cost tree (constrained Steiner tree).

Pornavalai et al. (1998) formulate a bi-objective routing problem with constraints in networks supporting multimedia applications. The possibility of traffic splitting is

not considered. The problem is of the mixed integer type. The two objectives are the minimisation of the number of hops in a path (so that the number of resources used is minimum) and the minimisation of the cost of a path (where the cost of a link is defined in terms of the used bandwidth in that link). Four constraints are taken into account: the bandwidth of a path should be above a minimum value, there is an upper bound on the delay and on the variation of the delay the packets can experience in a path, and the number of hops in a path should be below a maximum value. The problem is solved using a heuristic procedure, based on the composition of shortest paths obtained when only one of the objectives is minimised. An exact algorithm to compute, in an efficient manner, the whole Pareto solution set, in this type of bi-objective routing problems, is given in Clímaco et al. (2003).

Martins (2004) and Martins et al. (2003, 2006) propose and describe MODR-S, a multiobjective dynamic routing model for multiservice networks equivalent in the traffic plane to multirate loss traffic networks. The traffic flows are represented as multirate Poisson processes, the calls of which are characterised by their effective bandwidth and mean holding time. A dynamic alternative routing principle is used. The model includes the possibility of existence of multiple types of service, with different QoS requirements. The traffic engineering design uses stochastic traffic models. The optimisation is performed at two hierarchical levels: at the network level, the objectives are the maximisation of the revenue of the total traffic carried in the network for a given routing solution and the minimisation of the maximal average blocking probability for the different services; at the services level, the objectives are the minimisation of the average blocking and the minimisation of the maximal point-to-point blocking for each service. The network level objectives have priority over the services level objectives. The algorithm to choose candidate solutions for the routing optimisation problem is a bi-objective constrained shortest path algorithm (MMRA-S). This algorithm is based on the calculation of k-shortest paths with a maximum number of arcs, where the "length" of a path is given by the weighted average of two metrics: the blocking in the links for each service and the implied cost associated with the acceptance of a call of a certain service in a specific link. This metric is quite powerful because it represents the cost of carrying a call in a link and it includes the effects in the remaining flows of the network resulting from the acceptance of that call. The actual choice of non-dominated solutions is made by using preference thresholds (defined in the objective function space) corresponding to requested and acceptable threshold values for the objective functions associated with each of the metrics, which define priority regions in the objective function space. The model is dynamic of periodic type, that is, the paths are periodically updated, in a reaction to estimated changes in the node to node offered traffic of the different service types. A heuristic is devised to choose which paths in the network should be changed at each route updating period. This heuristic includes service protection mechanisms, to prevent performance degradation in overload situations.

Finally it must be remarked that several authors formulate routing models applicable to MPLS, using a single objective function. Examples of these models can be found in Bourquia et al. (2003); Elwalid et al. (2001); Franzke and Pönitz (2002) (for MPLS networks) and Kochkar et al. (2001); Mitra et al. (1999); Rombaut et al. (2002). For instance, in Mitra et al. (1999) propose a generic routing model for multirate

circuit-switched loss networks. The objective is the maximisation of the long-run average revenue for the network subject to a constraint reflecting an arbitrary admission control scheme.

3 A meta-model for multiobjective routing in MPLS

3.1 Underlying concepts and model features

Following the analysis in the previous section, it is apparent that the proposals of multiobjective routing models for MPLS networks often differ in key instances of the routing framework. This is also true for the single-objective models in this area mentioned in the previous section.

A first instance has to do with the routing optimisation framework where we may distinguish *network-wide* optimisation models and *flow-oriented* models. The former are characterised by the consideration of objective function(s) formulated at network level, that depend explicitly on all traffic flows in the network, such as total traffic carried, total expected revenue, average packet delay or a function which seeks an optimisation of the use of all links in terms of their occupation levels. Examples of this type of models are Mitra et al. (1999); Mitra and Ramakrishnan (2001); Erbas and Mathar (2002); Erbas and Erbas (2003); Erbas (2003) and the model in Elwalid et al. (2003) with application to GMPLS (Generalized MPLS). In flow-oriented optimisation models the objective function(s) are formulated at the level of each node-to-node traffic flow or call (the term "call" is viewed here in its widest sense as a connection request of any service/application type), such as number of arcs or cost of the path, mean packet delay on each packet stream or end-to-end blocking probability. Examples of this type of approach are the numerous QoS routing models based on single-objective constrained shortest path problem formulations (a comprehensive review of these may be seen in Kuipers et al. 2002a,b) and the works on the cognitive packet network (CPN) routing protocol which is based on an adaptive routing control model using neural network reinforcement learning (Gelenbe et al. 2002, 2004).

A second instance that characterises the different models is the nature of the model in terms of the specified objective function(s) and constraints [namely whether the routing optimisation model is single or multiple objective, the technical-economic nature of the objective function(s) and of the QoS-related or other constraints, as well as the way in which they are modelled in mathematical terms] and other key features of the underlying routing system (namely whether the routing system is on-line or off-line, static or dynamic and other functional features).

A third important instance is the representation in the model of the "traffic" to be routed. In this respect two main aspects are at stake. A first aspect is the level of representation or 'granularity'. In principle, this can be done at the level of the traffic trunks or at the level of the packet streams of the same FEC placed on a certain LSP, or combining both levels in some manner. A second aspect is the nature of the representation of the traffic offered. We often find deterministic type representations typically based on the description of the required fixed bandwidth from the originating to the destination node, associated with a traffic trunk, such as in Erbas and Mathar (2002);

Erbas and Erbas (2003) and Erbas (2003). This is common to most modelling appro-
aches in the framework of the classical multicommodity network flow problem for-
mulations. Other approaches involve a stochastic representation of the traffic flows
by recurring to some type of stochastic approximation to describe the arrival in time
of connection demands (input traffic model) and recurring to an appropriate model
from Teletraffic Theory to estimate relevant parameters in the network links (typically
blocking probabilities or average delays). References with this type of approach are in
Mitra et al. (1999); Mitra and Ramakrishnan (2001) and Martins et al. (2003). Need-
less to say the stochastic-based models significantly increase the analytical complexity
and the numerical cost of the mathematical model underlying the routing optimisation
model, but are in essence more realistic.

The purpose of the meta-model for multiobjective routing in MPLS networks
described in the next section is to provide a framework for multiobjective routing
optimisation in this type of network, including the possibility of alternative routing.
The essential features of the model are now discussed.

Firstly it is a network-wide routing optimisation approach of a new type, leading to
a hierarchical multiobjective optimisation model. This has to do with the nature of the
adopted objective functions. In this respect we would like to note that various multiob-
jective approaches previously proposed use objective functions chosen to reflect only
indirectly technical-economic aims. An example is the optimisation of a utilisation
cost for all arcs expressed, through empirical functions, in terms of the loads in the
links (measured by the occupied bandwidth) as in Erbas and Mathar (2002), Erbas
and Erbas (2003), Erbas (2003) and Knowles et al. (2000). The pursued objective
is really to optimise the total traffic carried in the network. That type of objective
function, which we will designate as an 'indirect' or 'refracted' objective function,
has the advantage of being naturally adjusted to multicommodity network flow prob-
lem formulations. Similar considerations might be applied when the minimisation of
the number of arcs per path is used to seek the minimisation of the delay or fault
probability of a path, as in Haßlinger and Schnitter (2003). Nevertheless this type of
approach is just a rough approximation to the 'true' objective function(s) the model
seeks to incorporate, especially taking into account the random nature of traffic pat-
terns, even in stationary or quasi-stationary network working conditions. Instead, our
approach considers an explicit and 'direct' representation of the most relevant tech-
nical-economic objectives in a network-wide routing optimisation, such as the total
expected revenue (expressed in terms of the traffic carried of all service types) or the
average expected delay averaged over all types of packet flow streams. This modelling
approach aspect is in line with the school of thought adopted by Kelly (1988, 1990)
and Mitra et al. (1999), in the context of single-objective routing models.

Secondly we propose a hierarchy of objective functions by considering three lev-
els of optimisation with several objective functions in each level. The first level (first
priority) includes objective functions formulated at network level, considering the
combined effect of all types of traffic flows in the network. The second level refers
to average performance metrics of the traffic flows associated with the different types
of services supported by the network. The third level refers to average performance
metrics of the μ-flows of packet streams, obtained by aggregation of delay-based QoS
metrics.

A third important feature of the model is the explicit consideration, as objective functions, of 'fairness' objectives, at the three levels of optimisation. Such objectives are of min-max type and seek to make the most of the proposed multiobjective formulation. Usually, in previous formulations of routing problems for these networks, such type of aims related to fairness are not considered explicitly in any form or are represented through constraints (upper bounds) on certain performance metrics.

A fourth important feature of the model is the consideration of a bi-level stochastic representation of the traffic flowing in the network. The first or 'macro' level leads to the definition of 'traffic flows' which correspond to a stochastic representation of the MPLS traffic trunks associated with explicit routes. A second 'micro' level representation involves the definition of stochastic streams of packets of the same FEC, which belong to a certain traffic trunk (hence to a given macro traffic flow). The assumptions and basic sub-models associated with this dual stochastic traffic representation are described in Sect. 3.4.

Finally it should be stressed that the model is intended for application both to static or dynamic routing systems by using the necessary adaptations. All these features and the way in which they are interrelated constitute the essence of the substantive model (in the sense defined in Wierzbicki et al. 2000, Chap. 1) developed for the addressed automated routing decision problem.

3.2 Base-model description

Firstly we will describe the proposed model[2] assuming multiple service types, defined by the set $\mathcal{S}$ of traffic flow types assumed to be of the same class, that is, without qualitative differences concerning priority in the access. Here, the concept of *traffic flow* (defined at macro level) $f_s = \left(v_i, v_j, \overline{\gamma}_s, \overline{\eta}(f_s)\right)$ corresponds to a stochastic process, in general a marked point process, that describes the arrivals and basic requirements of μ-flows, originated at the MPLS ingress node v_i and destined to the MPLS egress node v_j, using the same LSP and characterised by the vectors of 'attributes' $\overline{\gamma}_s$ and $\overline{\eta}(f_s)$. The vector $\overline{\gamma}_s$ describes the traffic engineering attributes of flows of service type s. The vector $\overline{\eta}(f_s)$ enables the representation of the mechanism(s) of access control to all links l_k in the network by calls of flow f_s and includes, in general, traffic engineering attributes associated with f_s calls and all the links which may be used by f_s, including priority features. In the teletraffic modelling approach described in detail in Appendices, such attributes include the required effective bandwidth d_s and the mean duration $h(f_s)$ of each μ-flow in f_s. Note that in our model a 'μ-flow' corresponds to a call (the term call is used in its broadest sense) of the traffic flow. The use of the concept of effective bandwidth (Kelly 1996) in this context (MPLS networks using explicit routes) was proposed in Mitra et al. (1999) and used in Martins et al. (2003, 2006). This enables a representation of the network at a macro level, that is in the traffic plane level, by an equivalent multirate loss traffic network, which is used to estimate blocking probabilities for the different f_s.

---

[2] The notation used in this sub-section is listed in Appendix 5.

We assume that we dispose of an approximate teletraffic model that is capable of estimating the blocking probabilities $B(f_s)$ for all flows f_s of all service types, in terms of which one can calculate the average loss (or blocking probability) B_{ms}, for all traffic flows of type s (for a given routing choice for all traffic flows), as well as the maximal average blocking probability among all service types B_{Mm},

$$B_{Mm} = \max_{s \in \mathcal{S}} \{B_{ms}\}. \tag{1}$$

Also the total expected network revenue, W_T, can be calculated in terms of the expected revenues $w(f_s)$ associated with calls of all flows $f_s, \forall s \in \mathcal{S}$ and of the values of carried traffic A_s^c for all service types,

$$W_T = \sum_{s \in \mathcal{S}} W_s = \sum_{s \in \mathcal{S}} A_s^c w_s \tag{2}$$

assuming that $w(f_s) = w_s, \forall f_s \in \mathcal{F}_s$, where $\mathcal{F}_s$ is the set of traffic flows of type s.

Hence the two objective functions of the first optimisation level (global network level) are the total expected network revenue W_T (to be maximised) and B_{Mm}, where this objective function represents the fairness objective of higher priority (to be minimised), defined over all service types. Note that W_T is the objective function usually considered in "classical" single-objective formulations of network-wide routing optimisation and is an obvious direct technical-economic indicator of upper-level.

The second level of optimisation concerns service types and includes $2|\mathcal{S}|$ objective functions to be minimised: the mean blocking probability for flows of type s,

$$B_{ms} = \frac{1}{A_s^o} \sum_{f_s \in \mathcal{F}_s} A_t(f_s) B(f_s) \tag{3}$$

and the maximal loss B_{Ms}, defined over all flows of type s,

$$B_{Ms} = \max_{f_s \in \mathcal{F}_s} \{B(f_s)\}. \tag{4}$$

Again B_{Ms} represents the fairness objective defined for each service type s.

It must be remarked that up to this point the model is analogous to the two-level multiobjective optimisation model of MODR-S, described in Martins et al. (2006).

The third and lower level of routing optimisation has to do with the double representation of the traffic in the network, namely the representation of μ-flows of packet data streams imbedded in the macro traffic flows, the packets of which are forwarded along the same LSP, as mentioned in the previous section. The QoS parameters relevant to these packet streams are naturally average packet delays. We consider as the first objective function the packet total average delay

$$D_T' = \frac{1}{\gamma_T'} \sum_{s \in \mathcal{S}} D_{ms}' \gamma_s' \tag{5}$$

weighted by the relative bandwidths $\frac{\gamma_s'}{\gamma_T'}$ [see (27)–(28)] offered by all the packet streams of type s to the network. The parameter D_{ms}' [see (30)] is the average delay experienced by packets of flows of type s, weighted with bandwidths. Alternatively, one might consider the average packet delay D_T, calculated as in (31), taken over all network packet streams. While D_T defines exactly the mean delay experienced by an arbitrary packet carried in the network, the parameter D_T' reflects the importance of bandwidth load associated with μ-flows for different types of traffic flows. This seems more adequate in a routing optimisation model since it reflects the network loading aspects at the "bit/s" level as well as possible tariff structures based on linear functions of throughput volumes, in bits. This type of average delay measure is analogous to one of the objective functions in the model (Erbas and Erbas 2003).

The second objective function (the delay fairness objective, also to be minimised) at this optimisation level is the maximal average delay experienced by all types of packet streams,

$$D_{Mm} = \max_{s \in \mathcal{S}}\{D_{ms}\} \tag{6}$$

where D_{ms} is the average packet delay for μ-flows in flows of type s [see (29)].

Therefore, we may formulate the *hierarchical multiobjective routing optimisation problem* in the form:

<div align="center">

Problem P-M3-S   (7)

</div>

Network objectives:	$\min_{\overline{R}_t}\{-W_T\}$
	$\min_{\overline{R}_t}\{B_{Mm}\}$
Service objectives:	$\min_{\overline{R}_t}\{B_{ms}\}$
	$\min_{\overline{R}_t}\{B_{Ms}\}$
	$\forall s \in \mathcal{S}$
μ-flow network objectives:	$\min_{\overline{R}_t}\{D_T'\}$
	$\min_{\overline{R}_t}\{D_{Mm}\}$

subject to constraints defined explicitly or implicitly by the underlying teletraffic model and possible additional constraints. The decision variables $\overline{R}_t$ are the sets of feasible route sequences $\{R_t(f_s), \forall f_s \in \mathcal{F}_s, s \in \mathcal{S}\}$ assigned to all traffic flows at the time period t and will be designated as network *routing plans*. The routing plans represent here assignments of MPLS traffic trunks to LSP tunnels (defining explicit routes).

We emphasise that this model should be envisaged as a multiobjective routing optimisation framework dedicated to this type of network, with a significant degree of flexibility. In fact, the proposed model, or better, *meta-model*, i.e., the model underlying concepts and logical relations, may be configured to other definitions or choices of objective functions or to the introduction of possible constraints of a different nature, as long as the basic structure of the meta-model is preserved. Also simplifications or extensions of the model may be considered by eliminating or adding some objective function(s) and/or constraint(s). Such changes or adaptations naturally require corresponding modifications in the underlying teletraffic model or other model-related mathematical structures. In particular, the model can be adapted to static or dynamic

routing formulations and to alternative routing (as in the traffic model in Appendices) or traffic splitting principles.

A significant and useful example of this adaptability is described in the next section by considering the application of the meta-model to a situation, typical in this type of network, where two qualitatively different service classes are considered for the traffic flows: QoS and Best Effort traffic.

3.3 Model for QoS and Best Effort service classes

Let us now consider two classes of services, namely QoS (i.e., services with guaranteed QoS levels relative to some specific QoS parameters) and BE (Best Effort), where the corresponding traffic flows are carried in order to obtain the best possible quality of service but seeking not to jeopardise the QoS of the QoS traffic flows. The traffic flows in each class may be of different types and the service types of each class are grouped in the sets S_Q (for QoS service types) and S_B (for BE service types). Hereafter we will assume this distinction between *service class* and *service type*. Note that in the model, the QoS attributes related to class and type of service are supposed to be included in $\overline{\gamma}_s$ and $\overline{\eta}(f_s)$.

The treatment, in terms of routing, of two (or more) classes of traffic flows is a complex issue and different approaches have been proposed in the literature. An example concerning flow-oriented models is in Medhi and Sukiman (2000), where the authors propose an admission control technique based on the reservation in the links of a certain bandwidth BW_1 for the traffic flows of the QoS service class, while those with lower 'priority' (BE) will only be accepted if the available bandwidth is greater than BW_1; otherwise the acceptance is made according to a certain empirical probability. Also a number of routing models have been proposed for dealing with this issue based on the concept of residual virtual bandwidth. These virtual bandwidths are then associated with arc costs in order to compute paths with minimal costs. Examples of these approaches are in Ma and Steenkiste (1999) and Kochkar et al. (2003).

Concerning network-wide optimisation approaches, Mitra and Ramakrishnan (2001) proposes a bi-objective routing optimisation model where a primary objective function is the weighted sum of the carried bandwidth associated with QoS traffic flows and a secondary objective function of the same type is defined for the BE traffic. A heuristic procedure based on a three-stage decomposition technique is developed for obtaining solutions to the problem, seeking to guarantee the optimality of the revenue associated with QoS traffic.

In the framework of the meta-model described in the previous section we may formulate a three-level multiobjective routing optimisation problem by separating the total expected revenue in two parts: $W_{T|Q}$ for the traffic flows of QoS type and $W_{T|B}$ for the traffic flows of BE type. While $W_{T|Q}$ will be a first priority objective function, together with the maximal blocking probability for all QoS service types, $B_{Mm|Q}$, $W_{T|B}$ will be a second level objective function. This guarantees that the routing of BE traffic, in a quasi-stationary situation, will not be made at the cost of the decrease in revenue or at the cost of an increase in the blocking probability of QoS traffic flows. Assuming that the most relevant performance metrics for BE flows are average delays,

there is no need to include explicitly in the model objective functions corresponding to blocking probabilities for this type of traffic, hence avoiding additional complexity in the model.

These considerations lead to the following formulation of a *three-level hierarchical optimisation problem for two service classes*:

$$\text{Problem P-M3-S2} \tag{8}$$

1st level	QoS - Network objectives: $\min_{\overline{R}_t}\{-W_{T\vert Q}\}$
	$\min_{\overline{R}_t}\{B_{Mm\vert Q}\}$
2nd level	QoS - Service objectives: $\min_{\overline{R}_t}\{B_{ms\vert Q}\}$
	$\min_{\overline{R}_t}\{B_{Ms\vert Q}\}$
	$\forall s \in \mathcal{S}_Q$
	BE - Network objectives: $\min_{\overline{R}_t}\{-W_{T\vert B}\}$
3rd level	Network objectives for the μ-flows: $\min_{\overline{R}_t}\{D'_T\}$
	$\min_{\overline{R}_t}\{D_{Mm}\}$

Note that while QoS and BE traffic flows are treated separately in terms of upper-level objective functions so as to reflect their different 'prioritisation' in the optimisation model, the interactions among all traffic flows are fully represented in the model. This is in fact guaranteed by the traffic modelling approach underlying the optimisation model, since the link traffic model must integrate the contributions of all traffic flows which may use every link. This is another major difference in comparison to other routing models proposed for networks with two service classes.

The definition and calculation of the parameters in the expressions are given in Appendix 5.

3.4 Traffic modelling approach

An important element of the routing optimisation approach is the underlying traffic model that enables the calculation of all the traffic related parameters, under certain simplifying assumptions.

Next we describe the main features of a possible traffic modelling approach that may be used in the context of the meta-model described in previous sections, having in mind the two-level representation of traffic flows.

Taking into account the concept of traffic flow defined at the macro level (see Sect. 3.1) and its representation through a marked point process, it is necessary to obtain a network representation and a stochastic model for the links, enabling the calculation of all the relevant traffic related parameters. In this context, the concept of *effective bandwidth* is used, in association with the definition of the MPLS explicit routes. The effective bandwidth can be viewed as a stochastic measure of the utilisation of network transmission resources capable of representing, in an approximate manner, the variability of the rates of different traffic sources, as well as the effects of statistical multiplexing of different traffic flows in a network. The effective bandwidth enables the traffic behaviour and the QoS features at cell (in ATM networks) and packet level (in IP/MPLS based networks) to be 'encapsulated' in a simplified

Fig. 1 Traffic model for a link
at the traffic flow level

$$\left. \begin{array}{c} \overrightarrow{\mathcal{P}'_{f_{s_1}}(d_{ks_1})} \\ \cdots \\ \overrightarrow{\mathcal{P}''_{f_{s_1}}(d_{ks_1})} \\ \cdots \\ \cdots \\ \cdots \\ \overrightarrow{\mathcal{P}'_{f_{s_m}}(d_{ks_m})} \\ \cdots \\ \overrightarrow{\mathcal{P}''_{f_{s_m}}(d_{ks_m})} \end{array} \right\} \longrightarrow \quad \underbrace{\bigcirc\bigcirc\cdots\bigcirc}_{C_k \text{ servers (channels)}}$$
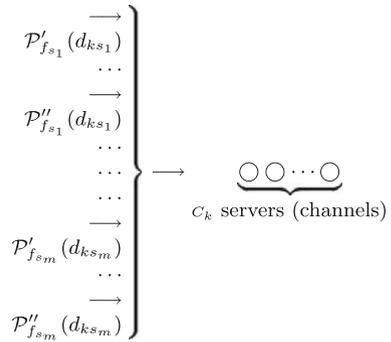
manner. Note that the 'hiding' of packet dynamics details is unavoidable for obtaining tractability in network-wide routing optimisation models. In fact, the use of more exact traffic representations would lead to more complex traffic models with a high numerical burden. A review of the definition, theoretical foundation and application of the effective bandwidth concept can be seen in Kelly (1996). The use of this conceptual tool in routing optimisation models of multiservice networks of various types can be seen for example in Mitra et al. (1999) and Martins et al. (2003). In this manner and for the purpose of calculating blocking probabilities, the network may be represented in the traffic plane by a multiclass loss traffic network, equivalent to a multirate traffic circuit-switched network.

The stochastic model, at the traffic flow level, for the purpose of calculating the blocking probabilities B_{ks} experienced by the flows f_s on a link l_k is a multidimensional Erlang system of type $M_1 + M_2 + \cdots + M_n / M / C_k / 0$ and is represented in Fig. 1. The input traffic is a superposition of n independent multirate marked Poisson processes $\mathcal{P}_{f_s}(d_{ks})$, associated with flows f_s, the routing plan of which uses the link l_k, where d_{ks} is the equivalent required effective bandwidth of f_s in l_k, expressed in the same unit as the equivalent number of servers C_k. The intensity of each process $\mathcal{P}_{f_s}(d_{ks})$ is $I_t(f_s) p_k(f_s)$ where $p_k(f_s)$ is the probability of a call from f_s being offered to l_k.

The probabilities B_{ks} may then be expressed in the form

$$B_{ks} = \mathcal{L}_s \left(\overline{d_k}, \overline{\rho_k}, C_k \right) \tag{9}$$

where $\mathcal{L}_s$ represents the function implicit in the analytical model used to calculate the marginal blocking probabilities, B_{ks}, in terms of $\overline{d_k} = (d_{k1}, \ldots, d_{k|\mathcal{S}|})$ (vector of equivalent effective bandwidths) and $\overline{\rho_k} = (\rho_{k1}, \ldots, \rho_{k|\mathcal{S}|})$ (vector of reduced traffic loads ρ_{ks} offered by flows of type s to l_k).

The Kaufman (or Roberts) algorithm (Kaufman 1981; Roberts 1983) can be used to calculate the functions $\mathcal{L}_s$ for small values of C_k. For larger values of C_k, the use of approximations based on the uniform asymptotic approximation (UAA) (Mitra and Morrison 1994) is recommended. This type of approximation was suggested by Mitra et al. (1999) in the context of off-line single-objective multiservice routing optimisation models and was also used in the multiobjective dynamic alternative routing model (Martins et al. 2006). The use of very efficient and robust approximations in

$$
\left.
\begin{array}{c}
\mathcal{P}^k_{f'_{s_1}}\left(\overrightarrow{h^k(f'_{s_1})}\right) \\
\cdots \\
\mathcal{P}^k_{f''_{s_1}}\left(\overrightarrow{h^k(f''_{s_1})}\right) \\
\cdots \\
\cdots \\
\cdots \\
\mathcal{P}^k_{f'_{s_m}}\left(\overrightarrow{h^k(f'_{s_m})}\right) \\
\cdots \\
\mathcal{P}^k_{f''_{s_m}}\left(\overrightarrow{h^k(f''_{s_m})}\right)
\end{array}
\right\}
\longrightarrow
\boxed{\begin{array}{c}\text{Effective access func-}\\ \text{tion for each } \mathcal{P}^k_{f_s}\end{array}}
\xrightarrow{\text{M}^*}
\underbrace{\square\square\cdots\square}_{\infty\ \text{queue}}\ \underbrace{\bigcirc}_{1\ \text{server}}
$$
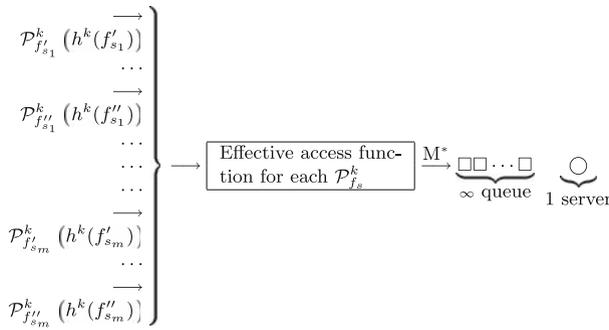
Fig. 2 Traffic model for a link at the packet μ-flow level

this context, such as the one suggested, is absolutely critical in a model of this nature, for tractability reasons.

Concerning the traffic modelling at the packet μ-flow level, firstly we will have to characterise the corresponding point processes offered to each link. For simplifying purposes we will assume an approximation through marked Poisson point processes with intensities $I'_t(f_s)$ (packet/s) expressed in terms of the μ-flow arrival intensities $I_t(f_s)$ during the time period t, and the average number of packets per μ-flow in f_s, given by $\overline{b}(f_s)$ [see (13)]. A *potential traffic offered* to link l_k by packet μ-flow streams from f_s is defined by $\rho^k_t(f_s) = I'_t(f_s)h^k(f_s)$ [Erl], where $h^k(f_s)$ is the mean service time in l_k of a packet from μ-flows in f_s, as calculated in (14). Having in mind the existence of a loss mechanism (at macro level) or possible control access mechanisms described probabilistically by the functions ψ_{ks}, one has to calculate the *reduced offered packet traffic load* on l_k, $\rho^{k*}_t(f_s)$ [Erl], associated with packet streams in f_s [see (40)]. These aspects can be represented through a multidimensional effective access function (the most general type of which is defined in Tralhão et al. 1994) applied to each of the Poisson processes describing the potential traffic offered by the μ-flows. That access function represents a "thinning" mechanism applied to these processes and guarantees the coherence between the traffic models at macro and micro levels.

Therefore we consider as an approximate model for the links at the packet μ-flow level, the system illustrated in Fig. 2.

In the figure, the symbols $\mathcal{P}^k_{f'_{s_i}}$ represent the aforementioned marked Poisson processes with intensities $I'_t(f'_{s_i})$, the "calls" of which are packets with mean service time $h^k(f'_{s_i})$.

Finally, to estimate the average packet delay $D_k(f_s)$ experienced in l_k by packets in μ-flows from f_s we may use the $M/GI/1/\infty$ queue model. As a first approximation to model the service time distribution, consistent with the assumed Poissonian nature of the input processes (satisfying superposition and decomposition properties), we may consider hyper-exponential distributions of order up to $|\mathcal{S}|$, characterised in Appendix 7. The weights π^k_s of these distributions represent the probability of an arbitrary packet offered to l_k being originated from a μ-flow in traffic flow f_s. With this simplification, the mean packet delays $D_k(f_s)$ may be easily calculated by applying

the Pollaczek–Khintchine formula, in terms of the *equivalent total offered traffic* [Erl] on link l_k, given by

$$\rho_t^{k*} = \sum_{s \in \mathcal{S}} \sum_{f_s \in \mathcal{F}_s} \rho_t^{k*}(f_s) \text{ [Erl]} \tag{10}$$

Hence ρ_t^{k*} is the mean value of the (fictitious) Poisson *equivalent total traffic offered* to the link l_k at period t, represented by M* in Fig. 2.

The complete analytical model resulting from the described traffic modelling approach is given in Appendices.

Naturally other more exact (from a purely stochastic analysis point of view) modelling approaches might in principle be used in association with the discussed routing optimisation model. Nevertheless most of such approaches would rapidly become intractable when applied to a routing model of this type, even in a network of small dimension. We think the proposed approximation is a convenient trade-off between exactitude and numerical tractability, while satisfying the basic requirements of the proposed meta-model.

4 Conclusions and further work

A discussion of conceptual and methodological issues raised by routing optimisation models for MPLS networks was put forward. Then a meta-model for multiobjective routing in MPLS networks that provides a framework for multiobjective routing optimisation in this type of network, was presented. The proposed model uses hierarchical optimisation typically with three optimisation levels: the first priority objective functions refer to the network level; the second priority objective functions refer to performance metrics for the different types of services supported by the network; the third priority is concerned with performance metrics for the micro-flows of packet streams of the same FEC. A second feature of the model is the two-level stochastic representation of traffic in the network: a macro level, corresponding to traffic flows that represent a sequence of μ-flows of the same type carried over an explicit route, and a micro level, corresponding to micro-flows of packet streams. A third feature is the explicit consideration of fairness objectives, at the three levels of optimisation, expressed through the minimisation of the worst performance level associated with the QoS-related metrics defined at each level. An application of this meta-model considers two major classes of traffic flows: QoS traffic (with guaranteed requisites of certain QoS parameters) and Best Effort traffic (which is carried on a "best effort" basis and is in fact second priority traffic).

A teletraffic modelling stochastic approach that may be used in the context of this meta-model, for calculating all the traffic related parameters has also been proposed. The theoretical foundations of the proposed approximations as well as the complete analytical model for calculating traffic flow blocking probabilities and average packet delays, in an articulated manner, have also been presented.

We must stress that this meta-model should be envisaged as a multiobjective routing optimisation framework dedicated to this type of network, with a significant degree of

flexibility. In fact, the underlying concepts of the model and logical relations may be adapted to other definitions or choices of objective functions, or to the introduction of possible constraints of different nature, as long as the basic structure of the meta-model is preserved. Also simplifications or extensions of the model may be considered by eliminating or adding some objective function(s) and/or constraint(s). Such changes or adaptations naturally require corresponding modifications in the underlying teletraffic model or other model-related mathematical structures.

Finally the meta-model may be useful as a framework for getting a better (and deeper) insight of particular multicriteria routing models in this area, namely with respect to assumptions and approximations in key instances of each model. In particular it may help in understanding what is 'gained' or 'lost' with a specific model in relation to the broad routing framework provided by the meta-model.

This modelling framework, significantly 'ambitious' in its scope, naturally raises a number of open issues and difficulties that will have to be tackled in the future. Firstly there is the necessity of developing resolution approaches of a heuristic or meta-heuristic type for solving the complex hierarchical multiobjective routing optimisation problem. These resolution procedures have inherent difficulties associated with: the great complexity of the problem, NP-hard in a strong sense; the interdependencies among the objective functions; the treatment of inaccuracy and uncertainty in many of the parameters involved, having in mind the stochastic nature of the traffic representation; last but not least, the representation of the system of preferences necessary for an automatic ordering and selection of candidate solutions.

Also the traffic modelling approach requires the development of adequate models for estimating (from real-time measurements) the effective bandwidth of the various types of traffic flows in MPLS networks, as well as other input parameters to the traffic model.

Finally an experimental test bed, based on a simulation platform, will be necessary for evaluating the performance and the implications/limitations of the developed resolution approaches. The study of those limitations, namely those associated with the required computational cost of a model of this nature, are particularly relevant in dynamic routing formulations in the meta-model framework.

Appendices

5 Notation and calculation of basic parameters

- $\mathcal{F}_s$: set of traffic flows of the service type s ($s \in \mathcal{S}$);
- $R_t(f_s) = \{r^1(f_s), \ldots, r^M(f_s)\}$: routing plan for the flow f_s, in the time interval t;
- $\overline{R}_t = \{R_t(f_s), \ \forall f_s \in \mathcal{F}_s, s \in \mathcal{S}\}$: global routing plan for the network;
- μ-flow: flow which is equivalent to a "call" in the MPLS routing model with explicit routes;
- $b(f_s)$: continuous r.v. representing an approximation to the number of packets in a μ-flow of f_s. The average value of $b(f_s)$ is $\overline{b}(f_s)$;

- $S(f_s)$: continuous r.v. representing an approximation to the size of packets in a μ-flow of f_s. The average value of $S(f_s)$ is $\overline{S}(f_s)$;
- $I_t(f_s)$: traffic intensity for flow f_s (average number of μ-flows arriving during a second) in the time interval t;
- $h(f_s)$: average service time for a μ-flow;
- $A_t(f_s)$: average number of offered μ-flows of f_s, during the average service time of a μ-flow, $h(f_s)$,

$$A_t(f_s) = I_t(f_s) \cdot h(f_s) \text{ [Erl]}; \tag{11}$$

- ψ_{ks}: access control function to the link l_k for calls of the service s (its values are probabilities of access to l_k);
- ρ_{ks} [Erl]: total traffic of type s flows offered to the link l_k;
- d_{ks}: effective bandwidth of the flows $f_s \in \mathcal{F}_s$, of the service type s, in the link l_k, in number of channels [i.e., it corresponds to a bandwidth $d'_{ks} = d_{ks}u_0$ (bit/s)];
- d_s: effective bandwidth of the flows of the service type $s \in \mathcal{S}$;
- C'_k (bit/s): transmission rate or capacity of the link l_k;
- C_k: capacity of the link l_k (bit/s) in terms of basic units of transmission u_0 (bit/s), where $u_0 = 64$ or 16 kbit/s, for instance. The unit of C_k is "channels" or "circuits",

$$C_k = \frac{C'_k}{u_0}; \tag{12}$$

- $\rho_t^k(f_s)$ [Erl]: packet potential traffic offered by the flow f_s to the link l_k in the time interval t;
- $\rho_t^{k*}(f_s)$ [Erl]: packet traffic that is actually offered by the flow f_s to the link l_k in the time interval t;
- ρ_t^{k*} [Erl]: total packet traffic offered by all the flows to the link l_k in the time interval t (reduced offered traffic);
- $I'_t(f_s)$: total intensity of packet traffic associated with f_s in the time interval t,

$$I'_t(f_s) = I_t(f_s) \cdot \overline{b}(f_s) \text{ (packet/s)}; \tag{13}$$

- $h^k(f_s)$: average service time for a packet of the flow f_s in the link l_k,

$$h^k(f_s) = \frac{\overline{S}(f_s)}{C'_k}; \tag{14}$$

- B_{ks}: blocking probability of a call of the service s on the link l_k;
- $L_{r^p(f_s)}$: blocking probability of a call of f_s on the route $r^p(f_s)$;
- $B(f_s)$: point-to-point blocking probability for the flow f_s, considering M disjoint alternative routes,

$$B(f_s) = \prod_{p=1}^{M} L_{r^p(f_s)}; \tag{15}$$

- B_{ms}: mean blocking probability of the flows of type s, calculated as in (3);
- $B_{ms|Q}$: mean blocking probability of the QoS flows of type $s \in \mathcal{S}_Q$,

$$B_{ms|Q} = \frac{1}{A_s^o} \sum_{f_s \in \mathcal{F}_s} A_t(f_s) B(f_s), \quad s \in \mathcal{S}_Q; \tag{16}$$

- B_{Mm}: maximum of the average blocking probabilities experienced by all types of flows, calculated as in (1);
- $B_{Mm|Q}$: maximum of the average blocking probabilities experienced by all types of QoS flows,

$$B_{Mm|Q} = \max_{s \in \mathcal{S}_Q} \{B_{ms}\}; \tag{17}$$

- B_{Ms}: maximum of the point-to-point blocking probabilities for all the flows of type s, calculated as in (4);
- $B_{Ms|Q}$: maximum of the point-to-point blocking probabilities for all the QoS flows of type s,

$$B_{Ms|Q} = \max_{f_s \in \mathcal{F}_s} \{B(f_s)\}, \quad s \in \mathcal{S}_Q; \tag{18}$$

- A_s^o: total traffic offered by the flows of the service type s in the time interval t,

$$A_s^o = \sum_{f_s \in \mathcal{F}_s} A_t(f_s) \text{ [Erl]}; \tag{19}$$

- A_s^c: total traffic carried for all flows of the service type s in the time interval t,

$$A_s^c = \sum_{f_s \in \mathcal{F}_s} A_t(f_s)(1 - B(f_s)) = A_s^o(1 - B_{ms}) \text{ [Erl]}; \tag{20}$$

- $\lambda_{r^P(f_s)}$ [Erl]: marginal carried traffic of calls of the flow f_s in the route $r^P(f_s)$;
- W_s: expected revenue associated with the carried traffic for all flows of the service type s, where $w(f_s) = w_s$, $\forall f_s \in \mathcal{F}_s$,

$$W_s = A_s^c w_s; \tag{21}$$

- W_T: expected revenue (on average) of the network, associated with the total traffic carried for all types of service, calculated as in (2);
- $W_{T|Q}$: total expected revenue for the QoS traffic,

$$W_{T|Q} = \sum_{s \in \mathcal{S}_Q} W_s = \sum_{s \in \mathcal{S}_Q} A_s^c w_s; \tag{22}$$

− $W_{T|B}$: total expected revenue for the BE traffic,

$$W_{T|B} = \sum_{s \in \mathcal{S}_B} W_s = \sum_{s \in \mathcal{S}_B} A_s^c w_s; \tag{23}$$

− $D_k(f_s)$: average delay (average waiting time) in the link l_k, experienced by an arbitrary packet of the flow f_s;
− $D_{r^P(f_s)}$: average waiting time experienced by a packet of the flow f_s, when the corresponding μ-flow uses the route $r^P(f_s)$;
− $D(f_s)$: average waiting time of the packets of the flow f_s in a "successful" call;
− γ_s: total average intensity of packets from the flows f_s of the service type s,

$$\gamma_s = \sum_{f_s \in \mathcal{F}_s} I_t'(f_s) \text{ (packet/s)}; \tag{24}$$

− γ_T: total intensity of packets offered to the network,

$$\gamma_T = \sum_{s \in \mathcal{S}} \gamma_s = \sum_{s \in \mathcal{S}} \sum_{f_s \in \mathcal{F}_s} I_t'(f_s) \text{ (packet/s)}; \tag{25}$$

− $\gamma'(f_s)$: average intensity of bits offered by f_s to the network (i.e., average bandwidth offered),

$$\gamma'(f_s) = I_t'(f_s)\overline{S}(f_s) \text{ (bit/s)}; \tag{26}$$

− γ_s': average bandwidth offered to the network by flows of the type s,

$$\gamma_s' = \sum_{f_s \in \mathcal{F}_s} \gamma'(f_s) \text{ (bit/s)}; \tag{27}$$

− γ_T': total average bandwidth offered to the network by all the flows,

$$\gamma_T' = \sum_{s \in \mathcal{S}} \gamma_s' = \sum_{s \in \mathcal{S}} \sum_{f_s \in \mathcal{F}_s} \gamma'(f_s) \text{ (bit/s)}; \tag{28}$$

− D_{ms}: average delay experienced by packets of flows of the type s,

$$D_{ms} = \frac{1}{\gamma_s} \sum_{f_s \in \mathcal{F}_s} I_t'(f_s) D(f_s); \tag{29}$$

− D_{ms}': average delay experienced by packets of flows of the type s, weighted with bandwidths,

$$D_{ms}' = \frac{1}{\gamma_s'} \sum_{f_s \in \mathcal{F}_s} \gamma'(f_s) D(f_s); \tag{30}$$

− D_T: average delay experienced by packets from an arbitrary flow in the network,

$$D_T = \frac{1}{\gamma_T} \sum_{s \in \mathcal{S}} D_{ms} \gamma_s; \tag{31}$$

− D'_T: average delay experienced by packets from an arbitrary flow in the network, weighted with bandwidths, calculated as in (5);
− $D'_{T|Q}$: average delay experienced by packets from an arbitrary QoS flow in the network, weighted with bandwidths,

$$D'_{T|Q} = \frac{1}{\gamma'_T} \sum_{s \in \mathcal{S}_Q} D'_{ms} \gamma'_s; \tag{32}$$

− D_{Ms}: maximum value of the average delay experienced by flows of type s,

$$D_{Ms} = \max_{f_s \in \mathcal{F}_s} \{D(f_s)\}; \tag{33}$$

− D_{Mm}: maximum value of the average delay experienced by all types of flows, calculated as in (6);
− $D_{Mm|Q}$: maximum value of the average delay experienced by all the QoS flows,

$$D_{Mm|Q} = \max_{s \in \mathcal{S}_Q} \{D_{ms}\}. \tag{34}$$

6 Model of the links for calculating the blocking probabilities

The calculation of the blocking probabilities B_{ks} is made as outlined in Martins (2004); Martins et al. (2006) for a multirate loss traffic network. Consider that the arrivals are according to a Poisson process, the service time follows a negative exponential distribution, there is one alternative path ($M = 2$), the occupations of the links in each path are statistically independent and the two paths $r^1(f_s)$ and $r^2(f_s)$ are disjoint. Therefore, B_{ks} is calculated as in (9).

For the calculation of ρ_{ks}, we begin by calculating

$$L_{r^p(f_s)} = 1 - \prod_{l_j \in r^p(f_s)} \psi_{js}(1 - B_{js})$$

$$= 1 - \prod_{l_j \in r^p(f_s)} (1 - B_{js}) \quad \text{with } p = 1; 2 \tag{35}$$

$$\lambda_{r^1(f_s)} = A_t(f_s) \prod_{l_j \in r^1(f_s)} \psi_{js}(1 - B_{js})$$

$$= A_t(f_s) \prod_{l_j \in r^1(f_s)} (1 - B_{js}) \text{ [Erl]} \tag{36}$$

$$\lambda_{r^2(f_s)} = A_t(f_s)L_{r^1(f_s)} \prod_{l_i \in r^2(f_s)} \psi_{is}(1 - B_{is})$$

$$= A_t(f_s)L_{r^1(f_s)} \prod_{l_i \in r^2(f_s)} (1 - B_{is}) \text{ [Erl]} \tag{37}$$

where we assume that the access is complete, i.e., $\psi_{ks} = 1, \forall l_k \in \mathcal{L}, s \in \mathcal{S}$. Therefore,

$$\rho_{ks} = \sum_{f_s:l_k \in r^1(f_s)} A_t(f_s) \prod_{l_j \in r^1(f_s)\backslash\{l_k\}} (1 - B_{js})$$

$$+ \sum_{f_s:l_k \in r^2(f_s)} A_t(f_s)L_{r^1(f_s)} \prod_{l_i \in r^2(f_s)\backslash\{l_k\}} (1 - B_{is}) \text{ [Erl]}. \tag{38}$$

7 Model of the links for calculating the packet delays

As a first approach, we consider that the links can be modelled by a $M/GI/1/\infty$ queue.

The value of the *potential packet traffic offered* by the flow f_s [Erl] to the transmission link (or server), with a transmission rate C'_k (bit/s), during the time interval t, is given by

$$\rho_t^k(f_s) = I_t(f_s) \cdot b(f_s) \cdot \frac{\overline{S}(f_s)}{C'_k} = I'_t(f_s) \cdot h^k(f_s) \text{ [Erl]} \tag{39}$$

where $I_t(f_s) \cdot b(f_s)$ is the average number of packets offered by the flow f_s, in a time unit. A full availability system is considered.

The actual value of packet traffic offered by f_s to l_k, $\rho_t^{k*}(f_s)$ (*reduced offered packet load*), is a result of the access functions (expressed by ψ_{ks}) and the routing at the μ-flow level.

Consider that there is one alternative path ($M = 2$), the occupations of the links in each path are statistically independent and the two paths $r^1(f_s)$ and $r^2(f_s)$ are disjoint. Therefore,

$$\rho_t^{k*}(f_s) = \begin{cases} \rho_t^k(f_s) \prod_{l_j \in r^1(f_s)} \psi_{js}(1 - B_{js}) & \text{if } l_k \in r^1(f_s) \\ \rho_t^k(f_s)L_{r^1(f_s)} \prod_{l_i \in r^2(f_s)} \psi_{is}(1 - B_{is}) & \text{if } l_k \in r^2(f_s) \end{cases} \text{ [Erl]}. \tag{40}$$

In order to apply the $M/GI/1/\infty$ model, the total packet traffic ρ_t^{k*} that is actually offered to the link l_k has to be calculated [see (10)].

According to the PASTA property of the Poisson processes, the average waiting time of a packet from the flow f_s in l_k will be $D_k(f_s)$, which is the average waiting time in the $M/GI/1/\infty$ system. This model (in spite of being just a first approximation) is more adequate for the calculation of the packet delays, than the $M/M/1/\infty$ model, due to the variability of the values of $\overline{S}(f_s)$ for different flows.

The following simplifications are considered:

$$\overline{b}(f_s) = \overline{b}_s, \quad \forall f_s \in \mathcal{F}_s, s \in \mathcal{S} \tag{41}$$

$$\overline{S}(f_s) = \overline{S}_s, \quad \forall f_s \in \mathcal{F}_s, s \in \mathcal{S} \tag{42}$$

$$h^k(f_s) = \frac{\overline{S}_s}{C_k^t} = h_s^k = \frac{1}{\mu_s^k}, \quad \forall f_s \in \mathcal{F}_s, s \in \mathcal{S}. \tag{43}$$

As an approximation to the model of the service time distribution, we could use the n-hyper-exponential distribution, where $n = |\mathcal{S}|$ (the total number of types of service). In this case,

$$f_k(t) = \sum_{s=1}^{n} \pi_s^k \mu_s^k \exp\left(-\mu_s^k t\right) \tag{44}$$

with the weights

$$\pi_s^k = \frac{\sum_{f_s \in \mathcal{F}_s : l_k \in r^1(f_s) \cup r^2(f_s)} \frac{\rho_t^{k*}(f_s)}{h^k(f_s)}}{\sum_{s \in \mathcal{S}} \sum_{f_s \in \mathcal{F}_s : l_k \in r^1(f_s) \cup r^2(f_s)} \frac{\rho_t^{k*}(f_s)}{h^k(f_s)}}, \quad s = 1, 2, \ldots, n \tag{45}$$

which represent the probability of an arbitrary packet offered to l_k being from the flow f_s of type $s \in \mathcal{S}$. The variance and the mean of this distribution are, respectively,

$$\sigma_k^2 = 2 \sum_{s \in \mathcal{S}} \pi_s^k \left(h_s^k\right)^2 - \left(\sum_{s \in \mathcal{S}} \pi_s^k h_s^k\right)^2 \tag{46}$$

$$m_k = \sum_{s \in \mathcal{S}} \pi_s^k h_s^k = \overline{h}^k. \tag{47}$$

Therefore, the quadratic coefficient of variation is

$$c_k^2 = \frac{\sigma_k^2}{m_k^2} \tag{48}$$

and the Pollaczek–Khintchine formula [Akimaru and Kawashima 1993, Eq. (3.20), p. 44] states that (assuming $\rho_t^{k*} < 1$)

$$D_k(f_s) = D_k = \frac{\rho_t^{k*}}{1 - \rho_t^{k*}} \frac{1 + c_k^2}{2} \overline{h}^k \quad \forall f_s \in \mathcal{F}_s, s \in \mathcal{S}. \tag{49}$$

This model can be easily extended to a more general case where the simplifying assumptions which we considered are not used. For instance, there might be different $h^k(f_s)$ for different flows f_s. In this situation, a n-hyper-exponential distribution, with n given by the number of different $h^k(f_s), \forall f_s \in \mathcal{F}_s, s \in \mathcal{S}$ and $l_k \in r^1(f_s) \cup r^2(f_s)$, should be taken into account. This would bring changes to the Eqs. (45)–(47).

The average waiting time for the packets of f_s when this flow uses the route $r^p(f_s)$ is

$$D_{r^p(f_s)} = \sum_{l_k \in r^p(f_s)} D_k(f_s). \tag{50}$$

The calculation of the average delay experienced by packets of the flow f_s depends on the routing conditions. Therefore, if there are M alternative paths (i.e., all the traffic carried from flow f_s will be routed on one of the M possible paths), then

$$D(f_s) = \sum_{p=1}^{M} \left(\prod_{q=1}^{p-1} L_{r^q(f_s)} \right) \left(1 - L_{r^p(f_s)}\right) D_{r^p(f_s)} \tag{51}$$

with the convention $\prod_{q=1}^{0} L_{r^q(f_s)} = 1$ and $L_{r^p(f_s)}$ calculated as in (35).

The probability of an arbitrary packet (in the set of packets generated by the flow $f_s \in \mathcal{F}_s$) being from f_s is $\pi(f_s) = \frac{I_t'(f_s)}{\gamma_{s,}}$. Therefore, the average delay experienced by packets of flows of the type s is D_{ms} given by (29) and the average delay experienced by packets from an arbitrary flow in the network is D_T given by (31).

If we want the performance measures to take into account the bandwidth factor as in several models (Erbas and Mathar 2002, for instance) then the performance measures associated with the packets delay will be D_{ms}' [see (30)] and D_T' [see (5)].

References

Akimaru H, Kawashima K (1993) Teletraffic: theory and applications, telecommunication networks and computer systems. Springer, Heidelberg

Awduche D, Malcolm J, Agogbua J, O'Dell M, McManus J (1999) Requirements for traffic engineering over MPLS. RFC 2702, Network Working Group

Awduche D, Chiu A, Elwalid A, Widjaja I, Xiao X (2002) Overview and principles of Internet traffic engineering. RFC 3272, Network Working Group

Bourquia N, Ameur WB, Gourdin E, Tolla P (2003) Optimal shortest path routing for Internet networks. In: Ben-Ameur W, Petrowski A (eds) Proceedings of the international network optimization conference (INOC 2003). Institut National des Télécommunications, Evry/Paris, France, pp 119–125

Clímaco JCN, Craveirinha JMF, Pascoal MMB (2003) A bicriterion approach for routing problems in multimedia networks. Networks 41(4):206–220

Elwalid A, Jin C, Low S, Widjaja I (2001) MATE: MPLS adaptive traffic engineering. In: Proceedings of IEEE Infocom 2001, Anchorage, Alaska

Elwalid A, Mitra D, Saniee I, Widjaja I (2003) Routing and protection in GMPLS networks: from shortest paths to optimized designs. J Lightwave Technol 21(11):2828–2838

Erbas SC (2003) Utilizing evolutionary algorithms for multiobjective problems in traffic engineering. In: Ben-Ameur W, Petrowski A (eds) Proceedings of the international networks optimization conference (INOC 2003). Institut National des Télécommunications, Evry/Paris, France, pp 207–212

Erbas SC, Erbas C (2003) A multiobjective off-line routing model for MPLS networks. In: Charzinski J, Lehnert R, Tran-Gia P (eds) Proceedings of the 18th international teletraffic congress (ITC-18). Berlin, Germany. Elsevier, Amsterdam, pp 471–480

Erbas SC, Mathar R (2002) An off-line traffic engineering model for MPLS networks. In: Bakshi B, Stiller B (eds) Proceedings of the 27th annual IEEE conference on local computer networks (27th LCN). Tampa, FL. IEEE Computer Society, pp 166–174

Franzke M, Pönitz A (2002) Global shortest path solutions for the traffic engineering problem. In: Proceedings of NETWORKS 2002—10th international telecommunication network strategy and planning symposium. München, Germany, pp 275–278

Gelenbe E, Lent R, Montuori A, Xu Z (2002) Cognitive packet networks: QoS and performance. In: MAS-COTS symposium. IEEE Computer Society

Gelenbe E, Lent R, Nunez A (2004) Self-aware networks and QoS. Proc IEEE 92:1478–1489

Haßlinger G, Schnitter S (2002) Algorithms for traffic engineering. In: Proceedings of 6th INFORMS telecommunications conference. Boca Raton, FL

Haßlinger G, Schnitter S (2003) Optimized traffic load distribution in mpls networks. In: Telecommunications network design and management. Kluwer, Dordrecht, pp 125–141

Kaufman JS (1981) Blocking in a shared resource environment. IEEE Trans Commun COM-29 10:1474–1481

Kelly FP (1988) Routing in circuit-switched networks: optimization, shadow prices and decentralization. Adv Appl Probab 20(1):112–144

Kelly FP (1990) Routing and capacity allocation in networks with trunk reservation. Math Oper Res 15(4):771–793

Kelly F (1996) Notes on effective bandwidths. In: Kelly FP, Zachary S, Ziedins I (eds) Stochastic networks: theory and applications. Royal Statistical Society Lecture Notes Series, vol 4. Oxford University Press, New York, pp 141–168

Knowles J, Oates M, Corne D (2000) Advanced multi-objective evolutionary algorithms applied to two problems in telecommunications. BT Technol J 18(4):51–65

Kochkar H, Ikenaga T, Oie Y (2001) QoS routing algorithm based on multiclasses traffic load. In: Proceedings of global telecommunications conference—GLOBECOM'01, San Antonio, TX, pp 2193–2198

Kochkar H, Ikenaga T, Hori Y, Oie Y (2003) Multi-class QoS routing with multiple routing tables. In: Proceedings of IEEE PACRIM'03, Victoria, BC, Canada, pp 388–391

Kuipers F, Van Mieghem P, Korkmaz T, Krunz M (2002a) An overview of constraint-based path selection algorithms for QoS routing. IEEE Commun Mag, pp 50–55

Kuipers FA, Korkmaz T, Krunz M, Van Mieghem P (2002b) A review of constraint-based routing algorithms. Technical report. Technical University Delft, Delft

Ma Q, Steenkiste P (1999) Supporting dynamic inter-class resource sharing: a multi-class QoS routing algorithm. In: Proceedings of IEEE Infocom'99. New York, pp 649–660

Martins L (2004) Um Método de Encaminhamento Dinâmico Multiobjectivo em Redes de Telecomunicações - Modelação e Desempenho. Ph.D. thesis, Faculdade de Ciências e Tecnologia da Universidade de Coimbra, Coimbra

Martins L, Craveirinha J, Clímaco J (2003) A new multiobjective dynamic routing method for multiservice networks—modelling and performance. In: Ben-Ameur W, Petrowski A (eds) Proceedings of the International Networks Optimization Conference (INOC 2003). Institut National des Télécommunications. Evry/Paris, France, pp 404–409

Martins L, Craveirinha J, Clímaco J (2006) A new multiobjective dynamic routing method for multiservice networks: modelling and performance. Computat Manag Sci 3(3):225–244

Medhi D, Sukiman I (2000) Multi-service dynamic QoS routing schemes with call admission control: a comparative study. J Netw Syst Manag 8(2):157–190

Mitra D, Morrison JA (1994) Erlang capacity and uniform approximations for shared unbuffered resources. IEEE/ACM Trans Netw 2(6):558–570

Mitra D, Ramakrishnan KG (2001) Techniques for traffic engineering of multiservice, multipriority networks. Bell Labs Tech J 6(1):139–151

Mitra D, Morrison JA, Ramakrishnan KG (1999) Optimization and design of network routing using refined asymptotic approximations. Perform Eval 36–37:267–288

Pornavalai C, Chakraborty G, Shiratori N (1998) Routing with multiple QoS requirements for supporting multimedia applications. Telecommun Syst 9(3-4):357–373

Roberts JW (1983) Teletraffic models for the Telecom 1 integrated services network. In: Proceedings of 10th international teletraffic congress. Montreal, Canada

Rombaut M, Hébuterne G, Priem M (2002) The limited version of the mille feuilles: a routing algorithm for packet networks and circuit switched networks with a constraint number of paths. In: Proceedings of NETWORKS 2002—10th international telecommunication network strategy and planning symposium. München, Germany, pp 365–372

Rosen E, Viswanathan A, Callon R (2001) Multiprotocol label switching architecture. RFC 3031, Network Working Group

Schnitter S, Haßlinger G (2002) Heuristic solutions to the LSP-design for MPLS traffic engineering. In: Proceedings of NETWORKS 2002—10th international telecommunication network strategy and planning symposium. München, Germany, pp 269–273

Tralhão L, Craveirinha J, Paixão J (1994) A study on a stochastic system with multiple MMPP inputs subject to access functions. In: New progress in probability and statistics—4th international symposium on statistics of the base country. VSP International Publishers, pp 415–428

Widyono R (1994) The design and evaluation of routing algorithms for real-time channels. Technical Report TR-94-024. Tenet Group, University of California at Berkeley & International Computer Science Institute

Wierzbicki AP, Makowski M, Wessels J (eds) (2000) Model-based decision support methodology with environmental applications. Kluwer, Dordrecht