

Asymptotic Performance Limits for Routing and Rate Allocation in Rate-Based Multi-Class Networks

Aniruddha Diwan

Indian Institute of Science Bangalore 560012

Joy Kuri

Indian Institute of Science Bangalore 560012

Abstract— In this paper we investigate the performance of Routing and Rate Allocation (RRA) algorithms in rate-based multi-class networks, in particular, a network of routers that support the IntServ Guaranteed Services. We measure the performance of an RRA algorithm in terms of its weighted carried traffic. We derive an upper bound on the weighted carried traffic of any RRA algorithm. The bound can be computed by linear programming. Moreover we show that a Fixed RRA algorithm achieves the bound asymptotically. Therefore the bound can be used as an absolute benchmark for the performance of various RRA algorithms. We illustrate this by comparing the simulated performance of a simple shortest-path RRA (SP-RRA) algorithm with the bound.

I. INTRODUCTION

In this paper, we consider the QoS routing problem in the context of a rate-based multi-class network. In this model, a connection request belongs to one of several pre-specified classes. Each class is characterized by a rate that might depend on the route parameters. Upon arrival, a connection requests establishment of a route between its source and destination such that the required rate is available on this route during the connection holding time. We call this QoS routing problem “Routing and Rate Allocation (RRA)” problem. Note that the rate-based multi-class network model considered in this paper is a fairly general model. In ATM networks, routing a Virtual Circuit (VC) or Virtual Path (VP) involves the selection of a route within the network from the source to the destination that supports the rate requirement of the VC/VP. In the Internet, the IETF IntServ Working Group has defined the Guaranteed Services framework [1] that can provide the connection with certain QoS guarantees by proper rate allocation.

Related Work: Traditionally, the design of routing schemes in VC routed networks has been based on the concept of *Least Loaded Routing* (LLR) [2]. As an alternative to the load-balancing concept, VC *packing* techniques were proposed in [3] for VC routing. [4] proposed the concept of *load profiling* as an alternative to load balancing and packing schemes for VC routing. Load profiling techniques allow the distribution of available bandwidth across a set of candidate routes to match the characteristics of incoming VC requests. In the context of QoS support in the Internet, [5] proposed the *shortest-widest* path scheme as a way of minimizing the call blocking rate. [6] proposed the *widest-shortest* path scheme to extend OSPF for QoS routing. [7] mentions *dynamic alternative* path routing and also presents schemes for var-

ious QoS guarantees in networks with rate-based schedulers.

Our Goal: Though these schemes aim at increasing the network performance, it is not known whether any of the schemes really extracts the best performance from the network. In this paper, we are concerned with precisely this problem. We ask the following question: what is the maximum performance that an RRA algorithm can obtain from the network? In other words, we want to derive a tight upper bound on the performance of any RRA algorithm. Though our work is applicable to any rate-based multi-class network, we consider a network of routers that support the Guaranteed Service framework in this paper.

Motivation: The motivation for our approach in this work comes from similar problems that arise in other types of networks. The first among these is the routing problem in circuit-switched telephone networks. Here we must route connections by selecting a path for each connection such that there is a circuit available to accommodate the call in every link on the path. The routing problem in circuit-switched networks has been examined extensively in [9]. [8] derived an asymptotically tight upper bound on the carried traffic. Using the model in [8], [10] reported an asymptotically tight upper bound on the carried traffic of any routing and wavelength assignment (RWA) algorithm for wavelength routed optical networks and [11] reported an asymptotically tight upper bound on the carried traffic of any fixed channel assignment algorithm for channelized cellular telephone systems. The work reported in [9] is motivated by an objective similar to ours. But the approach is computationally demanding as it involves computing blocking probabilities.

Our Contributions: Motivated by the above works, in this paper we obtain an asymptotically tight upper bound on the performance of RRA algorithms. Using the results in [8], we show that an appropriately scaled optimal solution of a linear program (LP) is an upper bound on the weighted carried traffic that any RRA algorithm can achieve for this network. Moreover, for the traffic models that satisfy the Asymptotic Traffic Property (ATP) [8] (this includes the widely used Poisson call arrival model), we present a Fixed RRA algorithm that achieves the upper bound asymptotically, emphasizing the fact that the upper bound is asymptotically tight. Various heuristic algorithms can be evaluated in absolute terms using the upper bound.

II. THE GUARANTEED (G) SERVICES FRAMEWORK

The G service [1] provides a guaranteed bandwidth, a firm end-to-end delay bound and no queueing loss for packets of a compliant data flow that is admitted. The flow characteristics are specified in terms of its maximum packet size L^{\max} and its token bucket parameters that include b , the token bucket size, ρ , the token accumulation rate, and \hat{p} , the peak rate. Network element l allocates a rate R for the flow and exports parameters C_l and D_l that decide the level of service it can provide to flows that traverse it. C_l and D_l approximate the departure from the "fluid model" of service. Consider a flow traversing a series of N network elements. The end-to-end delay on the packets of the flow is:

$$D^{\text{bound}} = \begin{cases} \frac{(b-L)(\hat{p}-R)}{R(\hat{p}-\rho)} + \frac{L^{\max}}{R} + \sum_{l=1}^N \left[\frac{C_l}{R} + D_l \right], & \hat{p} > R \\ \frac{L^{\max}}{R} + \sum_{l=1}^N \left[\frac{C_l}{R} + D_l \right], & \hat{p} \leq R \end{cases} \quad (1)$$

If D^{reqd} is the worst-case end-to-end delay that is acceptable for the packets of a flow, $D^{\text{bound}} \leq D^{\text{reqd}}$ gives the minimum rate g that has to be allocated at each router on the route of the flow. Let r_l be the capacity of link l on a given path. Then, for example, from the first part of the end-to-end delay bound,

$$g(r_1, \dots, r_N) = \frac{\sum_{i=1}^N C_i + L^{\max} - \frac{(b-L^{\max})\hat{p}}{(\hat{p}-\rho)}}{D^{\text{reqd}} - \sum_{i=1}^N D_i + \frac{(b-L^{\max})}{(\hat{p}-\rho)}}$$

We show the dependence of g on r_1, \dots, r_N because D_i may actually vary with r_i . In the rest of the paper we are going to consider an asymptotic regime in which link capacities grow in proportion. Let $r_{\min} = \min_i r_i$. Then under our asymptotic regime $r_{\min} \rightarrow \infty$ such that $r_1/r_{\min} = \delta_1, \dots, r_N/r_{\min} = \delta_N$ is maintained. We assume that $g(r_1, \dots, r_N)$ decreases as r_{\min} grows large maintaining the proportion and

$$\lim_{r_{\min} \rightarrow \infty} g(r_1, \dots, r_N) = g^\infty(\delta_1, \dots, \delta_N)$$

The motivation is as follows. D_l usually represents the non-preemption delay at node l which reduces as the link capacity at node l increases. Thus the end-to-end delay requirement can be satisfied by allocating less rate as link capacities grow. Note that g^∞ is not the fluid limit rate as it has terms involving C_l . The assumption is true for scheduling policies falling under the G Service. In the rest of the paper we work with this assumption. We do not explicitly show the dependence of g on (r_1, \dots, r_N) and that of g^∞ on $(\delta_1, \dots, \delta_N)$. This is implicit when we use g and g^∞ .

III. NETWORK MODEL AND PROBLEM FORMULATION

We consider a network of routers that support the Guaranteed Services. Let N denote the number of nodes

and L the number of links in the network. Let r_l be the capacity of link l , $1 \leq l \leq L$. Assume that there is a set of P source-destination (SD) pairs. All connection requests are assumed to originate between these pairs. Let M be the number of available paths on which connections can be routed. The set of paths could either be given or can be computed.

Let $A = (a_{kj})$ be the $M \times P$ path-SD pair incidence matrix, i.e.,

$$a_{kj} = \begin{cases} 1 & \text{if path } k \text{ is between SD pair } j, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Let $B = (b_{kl})$ be the $M \times L$ path-edge incidence matrix, i.e.,

$$b_{kl} = \begin{cases} 1 & \text{if link } l \text{ is on path } k, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Further we assume that an arriving connection request belongs to one of the pre-specified traffic classes and each traffic class is identified by its token bucket parameters and end-to-end delay requirement. Let C be the number of traffic classes. We associate a rate g_{ik} with class i and path k , i.e., a class i connection that is to be routed on path k requires g_{ik} amount of bandwidth to be allocated on each link of path k .

On connection arrival, a Routing and Rate Allocation (RRA) algorithm selects a path according to some policy and allocates a rate on that path for the duration of the connection. If the RRA algorithm can not find a suitable path, it blocks the connection and blocked calls never return. For an RRA algorithm, let s_{ij} denote the number of class i connections present on SD pair j at any epoch. Thus $S = (s_{ij})$ is a $C \times P$ matrix that represents the network state at any epoch. Because of the finite link capacities, the set of the network states is also finite. Denote by F_S , the set of all feasible network states. Let n_{ik} be the number of class i connections carried on path k and let $N = (n_{ik})$ be the $C \times M$ matrix. If a network state S is feasible, then there exists an N matrix (which need not be unique) such that,

$$s_{ij} = \sum_{k=1}^M n_{ik} a_{kj} \quad \forall i \in C, \forall j \in P \quad (4)$$

$$\sum_{i=1}^C \sum_{k=1}^M g_{ik} n_{ik} b_{kl} \leq r_l \quad \forall l \in L \quad (5)$$

where $n_{ik} \geq 0$ and integer. This simply says that a network state should not violate the link bandwidth constraints. We say that S is described by N . Note that at any epoch, the network state S due to an RRA algorithm satisfies Eqn. (4) and (5). Denote by F_S^L , the set of all S satisfying the above constraints with integer constraints on s_{ij} and n_{ik} relaxed, i.e., when s_{ij} and n_{ik} are non-negative real numbers. We associate a positive weight w_i with class i (which may be considered as the revenue generated by a connection belonging to class i). Then, if the network is in state S at any epoch, the weighted carried traffic at that epoch is $\sum_{i=1}^C w_i \sum_{j=1}^P s_{ij}$.

The problem that we investigate in our work is to obtain an upper bound on the weighted carried traffic of

RRA algorithms. We also consider the problem of designing an optimal RRA algorithm that achieves this upper bound asymptotically, thereby proving the tightness of the upper bound. In this context, we first investigate the problem with respect to the offline case to gain important insights. We then study the problem with respect to the practically important online case. In this paper, we do not include proofs due to lack of space.

IV. THE OFFLINE CASE

In this case, λ is the total number of connections to be routed simultaneously in the network and $p_{ij}\lambda$ is the number of class i connections to be routed on SD pair j . An RRA algorithm has to decide the admissibility of these offered connections. It may not be possible to admit all the λ connections in the network as that may violate the link capacity constraints. In that case, an RRA algorithm has to accept a subset s_{ij} of the $p_{ij}\lambda$ connections so that the network state S is feasible. An optimal RRA algorithm is one that maximizes the weighted carried traffic $\sum_{i=1}^C w_i \sum_{j=1}^P s_{ij}$. Then, the optimal RRA algorithm is found by solving the following ILP whose value is denoted by $I(\lambda, p)$.

$$I(\lambda, p) = \max \sum_{i=1}^C w_i \sum_{j=1}^P s_{ij}$$

subject to

$$s_{ij} = \sum_{k=1}^M n_{ik} a_{kj} \quad \forall i \in C, \forall j \in P \quad (6)$$

$$\sum_{i=1}^C \sum_{k=1}^M g_{ik} n_{ik} b_{kl} \leq r_l \quad \forall l \in L \quad (7)$$

$$s_{ij} \leq p_{ij}\lambda \quad \forall i \in C, \forall j \in P \quad (8)$$

$$s_{ij}, n_{ik} \geq 0, \text{ integer}$$

Eqns. (6) and (7) ensure that S belongs to the set of feasible network states. That carried traffic should be less than the offered traffic is specified by Eqn. (8). Let $F_S^{I(\lambda, p)}$ denote the set of feasible network states for the above ILP. $I(\lambda, p)$ is an upper bound on the weighted carried traffic of *any* RRA algorithm for the offline case. But solving the ILP is computationally expensive and hence the quantity $I(\lambda, p)$ is hard to compute. We now give an LP formulation and show that the scaled optimal solution of the LP asymptotically achieves $I(\lambda, p)$, where the asymptotic regime that we consider is as follows.

Without loss of generality, we assume that the links are ordered such that $r_1 \leq r_2 \leq \dots \leq r_L$. Let $\frac{\lambda}{r_1} = r$, $\frac{r_1}{r_1} = \delta_1 (= 1)$, $\frac{r_2}{r_1} = \delta_2, \dots, \frac{r_L}{r_1} = \delta_L$. We consider the asymptotic scaling in which the offered load λ and the link capacities r_l grow large *in proportion*, i.e., the ratios $r, \delta_1, \delta_2, \dots, \delta_L$ remain unchanged. Let r_{\min} denote the minimum link bandwidth.

The LP is obtained as follows. We relax the integer constraints on the variables s_{ij} and n_{ik} of the $I(\lambda, p)$, replace g_{ik} by its asymptotic limit g_{ik}^∞ (Section 2) in the LP relaxation and divide throughout by r_{\min} . Then we

obtain the following LP whose value is denoted by $L_0(r, p)$ and whose feasible set of $S = (\alpha_{ij})$ matrices is denoted by $F_S^{L_0(r, p)}$.

$$L_0(r, p) = \max \sum_{i=1}^C w_i \sum_{j=1}^P \alpha_{ij}$$

subject to

$$\alpha_{ij} = \sum_{k=1}^M \beta_{ik} a_{kj} \quad \forall i \in C, \forall j \in P \quad (9)$$

$$\sum_{i=1}^C \sum_{k=1}^M g_{ik}^\infty \beta_{ik} b_{kl} \leq \delta_l \quad \forall l \in L \quad (10)$$

$$\alpha_{ij} \leq p_{ij}r \quad \forall i \in C, \forall j \in P \quad (11)$$

where α_{ij} and β_{ik} are nonnegative. It can be shown that

$$\text{Lemma 4.1: } \frac{1}{r_{\min}} I(\lambda, p) \leq L_0(r, p).$$

Let β'_{ik} and α'_{ij} yield an optimal solution to the $L_0(r, p)$ LP.

Lemma 4.2: Consider the following assignment to the $I(\lambda, p)$ ILP, $n_{ik} = \lfloor r_{\min} \frac{g_{ik}^\infty}{g_{ik}} \beta'_{ik} \rfloor$, $s_{ij} = \sum_{k=1}^M n_{ik} a_{kj}$. Then n_{ik} and s_{ij} yield a feasible solution to the $I(\lambda, p)$ ILP. Let the objective value produced by the above feasible solution be denoted by $I_F(\lambda, p)$.

$$\text{Lemma 4.3: } \lim_{r_{\min} \rightarrow \infty} \frac{1}{r_{\min}} I_F(\lambda, p) = L_0(r, p).$$

We now state the main result of this Section which follows directly from Lemma 4.1 and Lemma 4.3.

$$\text{Theorem 4.1: } \lim_{r_{\min} \rightarrow \infty} \frac{1}{r_{\min}} I(\lambda, p) = L_0(r, p).$$

Thus we have shown that, as the offered load and link capacities grow larger in proportion, the maximum weighted carried traffic $I(\lambda, p)$ is achieved by scaling the optimal solution of $L_0(r, p)$.

The Fixed RRA algorithm that achieves the maximum weighted carried traffic asymptotically is described as follows. Of the offered $p_{ij}\lambda$ class i connections on SD pair j , accept s_{ij} connections in such a way that S is described by N , where s_{ij} and n_{ik} are given by Lemma 4.2. Block the rest of the connections.

V. THE ONLINE CASE

In this case, λ is the offered network traffic in Erlangs and the offered class i traffic on SD pair j is $p_{ij}\lambda$ Erlangs. Once a connection request arrives, an RRA algorithm checks the admissibility of that connection request and makes a decision about the acceptance or rejection of that request. Therefore, at any epoch, the network is in a random state $S = (s_{ij})$ such that,

$$S \in F_S \quad (12)$$

$$E(s_{ij}) \leq p_{ij}\lambda \quad \forall i \in C, j \in P \quad (13)$$

We denote by $E(S)$ the $(E(s_{ij}))$ matrix. Eqn. (12) says that the random state S must belong to the set of feasible network states F_S and Eqn. (13) says that the average number of ongoing class i connections on SD pair j . Consider an $S = (\alpha_{ij})$ matrix satisfying the following,

$$\alpha_{ij} = \sum_{k=1}^M \beta_{ik} a_{kj} \quad \forall i \in C, \forall j \in P \quad (14)$$

$$\sum_{i=1}^C \sum_{k=1}^M g_{ik}^\infty \beta_{ik} b_{kl} \leq \delta_l \quad \forall l \in L \quad (15)$$

where α_{ij} are nonnegative real numbers and let F_S^0 denote the set of all S that satisfy the above. Note that the above constraints are nothing but Eqn. (9) and (10) of the $L_0(r, p)$. Then it is easy to see that $F_S \subseteq F_S^L \subseteq r_{\min} F_S^0$. Note that $E(S)$ is some convex combination of the states in F_S . Therefore $E(S) \in F_S^L$ and hence $E(S) \in r_{\min} F_S^0$. Also $E(s_{ij}) \leq r_{\min} p_{ij} r$ from Eqn. (13). Thus $\frac{1}{r_{\min}} E(S) \in F_S^{L_0(r, p)}$, i.e., $\frac{1}{r_{\min}} E(S)$ is a feasible point of the $L_0(r, p)$ LP (where, now λ is the offered load of online case and p_{ij} its distribution). The weighted carried traffic of an RRA algorithm and in particular the optimal RRA algorithm in the online case is given by $I(\lambda, p) = E(\sum_{i=1}^C w_i \sum_{j=1}^P s_{ij}) = \sum_{i=1}^C w_i \sum_{j=1}^P E(s_{ij})$ by linearity of expectation. With this we get the following Lemma.

Lemma 5.1: $\frac{1}{r_{\min}} I(\lambda, p) \leq L_0(r, p)$

What we have shown is that $r_{\min} L_0(r, p)$ is an upper bound on the maximum weighted carried traffic even in the online case. We now describe a *Fixed RRA* algorithm and show that asymptotically, its weighted carried traffic achieves the bound of Lemma 5.1; thereby proving the tightness of the bound.

Let β'_{ik} and α'_{ij} denote the optimal solution to the $L_0(r, p)$ LP of Section 4. Let n_{ik} and s_{ij} be obtained as follows, $n_{ik} = \lfloor r_{\min} \frac{p_{ik}}{\beta'_{ik}} \beta'_{ik} \rfloor$, $s_{ij} = \sum_{k=1}^M n_{ik} a_{kj}$. Recall from Section 4 that under our asymptotic regime,

$$\lim_{r_{\min} \rightarrow \infty} \frac{1}{r_{\min}} n_{ik} = \beta'_{ik} \text{ and } \lim_{r_{\min} \rightarrow \infty} \frac{1}{r_{\min}} s_{ij} = \alpha'_{ij} \quad (16)$$

Fixed RRA: The Fixed RRA for connection admission is as follows. At time t , let the network state be S^t described by N^t . Suppose a class i connection request arrives on SD pair j at epoch t_+ .

1. If $s_{ij}^t = s_{ij}$, block the connection request.
2. If $s_{ij}^t < s_{ij}$, then accept the connection request on any one of those paths k for SD pair j for which $n_{ik}^t < n_{ik}$.

Essentially, we have divided the bandwidth space on every link in such a way that at most a fixed number s_{ij} of class i connections can be supported simultaneously on SD pair j . Let $I_F(\lambda, p)$ be the weighted carried traffic of the Fixed RRA algorithm. It can be shown that, if the arrival process of connection requests satisfies a certain *Asymptotic Traffic Property (ATP)* [8], then the weighted carried traffic of the Fixed RRA algorithm achieves the bound of Lemma 5.1 asymptotically as r_{\min} grows large. It is shown in [11] that the well-known Poisson arrival process satisfies the ATP. We assume that the per-class per-SD pair offered traffics are independent of one another and the traffic model satisfies the above mentioned ATP. Then,

Theorem 5.1: $\lim_{r_{\min} \rightarrow \infty} \frac{1}{r_{\min}} I_F(\lambda, p) = L_0(r, p)$.

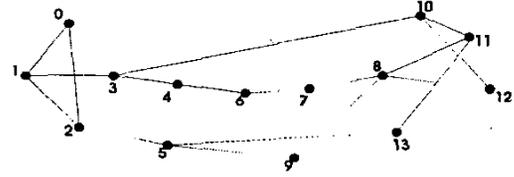


Fig. 1. Network to illustrate the performance of the SP-RRA scheme.

TABLE I
PARAMETERS FOR TRAFFIC MODELS OF TYPE VOICE AND VIDEO.

Class	b (kB)	ρ (Mbps)	D_i^{reqd} (ms)	L^{max} (kB)
1: voice	0.1	0.064	50	0.1
2: video conf	10	0.5	75	1.5
3: st video	100	3	100	1.5

VI. THE SHORTEST PATH RRA SCHEME

In this Section we show how one might utilize the LP upper bound $L_0(r, p)$ of Section 5 by comparing the simulated performance of a simple shortest path (SP) RRA heuristic with the LP upper bound. Heuristic schemes play an important role in real scenarios because the optimal algorithms may be difficult to obtain or may require knowledge of the offered load and offered load may not be known in practical situations. The SP-RRA scheme is as follows. We order the paths between an SD pair in non-decreasing hop-length and consider only the first K paths. If the minimum rate for the first path is available on its links, the connection is admitted on that path. Otherwise, we check the second path in the list and so on. If no path is adequate, the connection request is blocked.

The network under consideration is shown in Fig 1 and the traffic descriptors are as in Table I (and $\hat{p} = \infty$ for all classes). We assume that the well known weighted fair queueing (WFQ) scheduling policy is used on every link. We fix the weight of class i , w_i , equal to ρ_i as it reflects the bandwidth requirement of class i in some sense. We assume a Poisson arrival process and exponential holding time. We consider the case of uniform traffic, $p_{ij} = \frac{1}{CP}$.

In Fig 2-3, we plot the $L_0(r, p)$ bound, the weighted carried traffic for the Fixed RRA scheme and that for the SP-RRA scheme for various scale factors. We consider the case $K = 10$. The axes are normalized by r_{\min} in all plots. It can be observed that at very low loads, the SP-RRA is close to the LP upper bound. Note that the Fixed RRA is optimal only asymptotically, i.e., as the scale factor increases it approaches the LP upper bound. This behaviour can be seen in the plot for scale factor of 20.

We observe from the plots that the performance of the SP-RRA scheme is close to maximum at very low offered loads and it starts degrading later. The performance

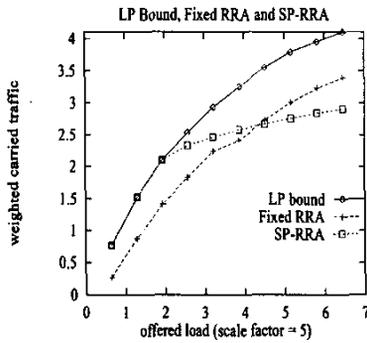


Fig. 2. Weighted carried traffic obtained from the LP upper bound, the Fixed RRA and the SP-RRA in the low blocking region for scale factor = 5.

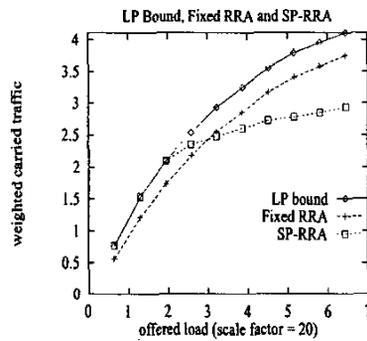


Fig. 3. Weighted carried traffic obtained from the LP upper bound, the Fixed RRA and the SP-RRA in the low blocking region for scale factor = 20.

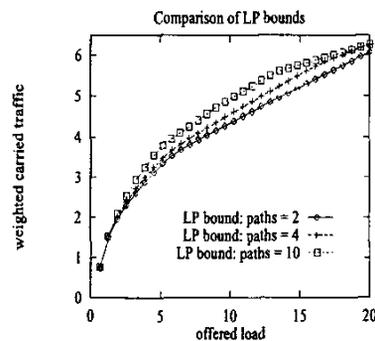


Fig. 4. The LP bound for the cases where at most 2, 4 and 10 paths per SD-pair are allowed.

of Fixed RRA is close to the maximum at higher offered loads, especially at higher scale factors. Thus these schemes perform well over limited regions and none of these guarantees a good performance over the entire region. An important design issue is how many paths one should consider before rejecting a connection request. In Fig 4 we plot the LP bound for the cases where $K = 2, 4$ and 10. An interesting observation is that the performance improvement is marginal for the 10 paths case. This leads to the conclusion that allowing more number of paths per SD pair might not really yield significant performance improvement. That the SP-RRA scheme leaves much room for improvement can be concluded by using the LP upper bound. Note that the LP upper bound makes it possible to draw conclusions about the performance of any heuristic scheme in a similar fashion.

VII. CONCLUSIONS

In this paper we considered the problem of Routing and Rate Allocation (RRA) in rate-based multi-class networks. We derived an asymptotically tight upper bound on the weighted carried traffic of any RRA algorithm in such networks. We showed that the Fixed RRA algorithm achieves the bound asymptotically. The Fixed RRA, though optimal, is not a viable scheme because it requires knowledge of the offered load. Various heuristics are suggested in the literature in order to extract good performance from the network. The LP bound enables us to check the performance of these heuristics in absolute terms.

REFERENCES

- [1] S. Shenker and C. Partridge and R. Guérin, "Specification of guaranteed quality of service", RFC 2212.
- [2] S. Gupta and K. Ross and M. ElZarki, "Routing in virtual path based ATM networks", In Proc. *GLOBECOM*, 1992.
- [3] I. Matta and M. Krunz, "Packing and least-loaded based routing in multi-rate loss networks", In Proc. *IEEE ICC*, 1997.
- [4] I. Matta, A. Bestavros, and M. Krunz, "Load profiling based routing for guaranteed bandwidth flows", *European Transactions on Telecommunications*, 10(2), pp. 165-182, 1999.
- [5] Z. Wang and J. Crowcroft, "QoS routing for supporting multimedia applications", *IEEE JSAC*, 14(7), pp. 1228-1234, 1996.
- [6] R. Guérin and A. Orda and D. Williams, "QoS routing mechanisms and OSPF extensions", *Internet Draft*, 1996.
- [7] Q. Ma and P. Steenkiste, "Quality-of-Service routing for traffic with performance guarantees", In Proc. *IFIP IWQoS*, 1997.
- [8] R. McEliece and K. Sivarajan, "Maximizing marginal revenue in generalized blocking service networks", In Proc. *30th Annu. Allerton Conf. Commun., Contr. Comput.*, 1992.
- [9] D. Mitra and J. Morrison and K. Ramakrishnan, "ATM network design and optimization: a multirate loss network framework", *IEEE Trans. on Networking*, 4(4), pp. 531-543, 1996.
- [10] R. Ramaswami and K. Sivarajan, "Routing and wavelength assignment in all-optical networks", *IEEE/ACM Trans. on Networking*, 3(5), 1995.
- [11] R. McEliece and K. Sivarajan, "Performance limits for channelized cellular telephone systems", *IEEE Trans. on Info. Theory*, 40(1), 1994.