Title
The effect of bandwidth and buffer pricing on resource allocation and QoS

Journal
Computer Networks-The International Journal of Computer and Telecommunications Networking, 46(1)

ISSN
1389-1286

Authors
Jin, Nan
Jordan, Scott

Publication Date
2004-09-01

DOI
10.1016/j.comnet.2004.03.023

Peer reviewed

The effect of bandwidth and buffer pricing on resource allocation and QoS ⋆

Nan Jin and Scott Jordan

University of California, Dept. EECS
Irvine, CA 92697-2625

Abstract

Congestion-based pricing of network resources is a common approach in evolving network architectures that support Quality of Service (QoS). Resource usage and QoS will thus fluctuate in response to changes in price, which must be dynamically controlled through feedback. Such feedback algorithms typically assume that network resources behave as Normal goods, i.e. that an increase in the price of a resource results in a decreased demand for that resource. Here, we investigate the sensitivity of resource allocation and the resulting QoS to resource prices in a reservation-based QoS architecture that provides guaranteed bounds on packet loss and end-to-end delay for real-time applications.

We derive necessary and sufficient conditions for bandwidth and buffer to act as Normal goods, showing that this depends on the shapes of the utility and QoS functions. We then show that the minimum total cost is a decreasing convex function of loss. When the delay constraints are absent or not binding, we prove that if a resource is a Normal good, then an increase in the price of that resource causes the loss on that link to increase, the loss on all other links to decrease, and the total loss to increase. We also give sufficient conditions to establish that an increase in the price for a resource results in a decreased demand for that resource, an increased demand for the other resource at that node, and an increased demand for resources at all other hops. Finally, when the delay constraint is binding, we give sufficient conditions to establish that an increase in the price of bandwidth at one node results in increased loss and delay at that node, and decreased loss and delay at all other nodes.

Key words: Resource allocation, Utility, QoS, Pricing

⋆ This work was supported by NSF and by DARPA.
Email addresses: njin@uci.edu (Nan Jin), sjordan@uci.edu (Scott Jordan).

1 Introduction

Evolving network architectures that support Quality of Service (QoS) rely on mechanisms to dynamically allocate network resources. Many papers in the literature have proposed policies for resource allocation, including a large number of proposals using *pricing*, c.f. [1–14]. While they often differ in detail, many pricing methodologies set prices based upon network congestion. Resource usage and QoS will thus fluctuate in response to changes in price, which must be dynamically controlled through feedback.

In this paper, we investigate the sensitivity of resource allocation and QoS to resource prices in a reservation-based QoS architecture. Such sensitivities are crucial to guide development of price adjustment algorithms. Intuition would lead one to believe that a increase in QoS would be paired with increases in all network resources, i.e. that network resources act as Normal goods. One would also expect that an increase in the price of network resources would result in a decrease in QoS. Our principal finding is that these relationships may not hold; indeed we derive their necessary and sufficient conditions.

As a platform for this work, we consider congestion-based pricing for real-time applications in a network with a reservation-based QoS architecture. By "reservation-based QoS architecture", we include any network architecture that can reserve bandwidth and buffer, whether for a single flow or an aggregate of flows, for the purpose of QoS. Such architectures could potentially include RSVP, MPLS, and ATM. By "real-time applications", we include any applications whose QoS depends on the amount of reserved bandwidth and buffer. In this paper, we consider packet loss probability and maximum end-to-end delay as the QoS measures.

Common network mechanisms to ensure QoS for real-time applications in a reservation-based QoS architecture include flow control, connection admission control (CAC), and dimensioning. Flow control is usually applied on a time-scale less than or equal to a round trip time, CAC on a time scale greater than a round trip time, and dimensioning on a relatively long time scale. Any such QoS architecture would also include many other elements, potentially including traffic characterization, traffic measurement, application-level QoS characterization, scheduling policies, dropping policies, and traffic smoothing.

We focus here on congestion-based pricing that addresses the CAC time-scale and accomplishes resource reservation and admission control. Such pricing policies typically decide what network resources to reserve for each flow (or aggregation of flows) on the basis of how the application values QoS (commonly called utility) and on congestion in the network. Prices can be used to provide a mechanism in which a minimal amount of information is exchanged between applications and the network in a distributed fashion.

Such pricing approaches therefore typically rely on information from both the application and the network. In our model, we abstract what we believe are the key relationships between these entities. On the application side, we presume the existence of two functions, one which describes each flow (or aggregate of flows), and one which describes how the flow measures satisfaction with its QoS. For the flow characterization, we presume the existence of a function that maps the number of sources in an aggregate and the amount of buffer and bandwidth reserved at each hop into the loss probability that these flows experience. Such a function might be derived using the literature on traffic characterization. Some researchers suggest using multi-parameter models to characterize flows, e.g. Markov-modulated Poisson processes or Markov-modulated fluid flows (c.f. [15–18]). Many researchers have suggested using effective bandwidth characterizations (c.f. [19–24]). However, we do not necessarily need to know such a function a-priori, if we can instead measure the loss experienced, such as often assumed in measurement-based admission control policies, (c.f. [25–28]). For the QoS satisfaction, we presume the existence of a function that maps the number of sources in an aggregate and the loss probability into a *utility*. This utility can be either interpreted directly as the amount that these sources would be willing to pay for this level of QoS, or indirectly interpreted simply as a numerical measure of satisfaction. Unfortunately, there is less work in the literature on deriving such functions, as the function would likely be application-specific (c.f. [29,30]).

On the network side, we presume the existence of a reservation-based QoS architecture. In our approach, we assume the network uses some type of route pinning (i.e. virtual circuit) for real-time applications, e.g. using MPLS or ATM. We also assume the existence of scheduling policies that are capable of assigning bandwidth and buffer to aggregates of flows. Such an architecture would likely specify which flows share resources, and thus determine the multiplexing gains in the network. Research on this topic includes effective bandwidth results that describe multiplexing gains by sharing of buffer and bandwidth (c.f. [31,32]). Here we merely presume that such multiplexing gains are incorporated into the flow characterization, whether known or measured.

The mechanism that decides what network resources to reserve for each aggregate should take into account this information from the applications and the network. Congestion-based pricing is often used to signal such information in a distributed fashion with a minimal exchange. There is a significant body of literature on the use of pricing in network operation, in order to accomplish a wide variety of goals. Early versions of congestion-dependent pricing considered charges per packet, c.f. [1–3]. Such approaches often use an auction to determine the optimal price per packet, resulting in prices that vary with demand. However, per-packet pricing can not easily address any flow-based

QoS metrics. The most common version of congestion-based pricing is to set the price per unit bandwidth according to the marginal cost, c.f. [4–6]. Utility and pricing has also been used to allocate bandwidth and describe congestion control dynamics. Kelly (c.f. [33]) proposed an optimization framework in which the objective is to maximize the aggregate source utility over their transmission rates. The centralized problem is decomposed into a separate problem for each user, which indicates the willingness-to-pay, and the network, which allocates bandwidths given each user's willingness-to-pay. Low (c.f. [34]) considers the same optimization problem, but decomposes it such that users choose transmission rates given prices and the network determines the price given the differential between total transmission rate and capacity. In both approaches, the utility function for each user is thought of as determined by the flow control algorithm. Such approaches can capture QoS metrics such as loss, but do not easily capture delay since buffer is not explicitly modelled. Some congestion-based pricing use both bandwidth and buffer as resources, c.f. [7–9]. In addition, some pricing algorithms view the price as the result of a game between users or between the users and the network, c.f. [10–12]. Finally, experiments with pricing in an ISP are described in [13,14]. While these congestion-based pricing approaches differ in detail, they generally assume that resources act as Normal goods.

In this paper, we investigate the ability of such congestion-based pricing schemes to allocate bandwidth and buffer and achieve desired levels of loss and delay, with the goal of maximizing the total utility of all users in the network. Our model was set out in [35], which proposed distributed pricing roles for each user, network node, and an *arbitrager* layer in between the user and the network that sells QoS (loss) to the users, and purchases from the network the least cost bundle of resources (bandwidth and buffer) that achieves the desired QoS. In contrast to most of the previous work on pricing, we consider arbitrary utility functions which can be expressed as a function of QoS, e.g. loss. The QoS is in turn a function of the allocated resources along the route. Finally, we explicitly model these resources as both bandwidth and buffer at each router.

In section II, we briefly review our model, two proposed distributed implementations using pricing, and results concerning optimality. In section III, we derive necessary and sufficient conditions for resources to act as Normal goods. We then show that the minimum total cost is a decreasing convex function of loss. In section IV, we turn to characterizing the sensitivity of the optimal allocation with respect to changes in resource price, e.g. caused by changes in demand of other users. We first consider the case when the delay constraints are absent or not binding. We give sufficient conditions to establish that an increase in the price for a resource results in an increase in the loss at that node, a decrease in the loss at all other nodes, and an increase in the total loss. We also give sufficient conditions to establish that an increase in the price for a resource results in a decreased demand for that resource, an increased demand for the other resource at that node, and an increased demand for resources at all other hops. Finally in section V, we consider the case in which the delay constraint is binding, and give sufficient conditions to establish that an increase in the price of bandwidth at one node results in increased loss and delay at that node, and decreased loss and delay at all other nodes.

2 The pricing framework

2.1 Network and user models

The complete model consists of a network model, which describes what type of service the network offers, and a user model, which describes how the user behaves. We presume the existence of a reservation-based QoS architecture using virtual circuits for real-time applications and using scheduling policies that are capable of assigning bandwidth and buffer to aggregates of flows. Assume the network supports m classes of virtual-circuit real-time traffic. Assume each class can reserve bandwidth on each link it traverses, and buffer at each router it passes through. Specifically, we assume class j can reserve bandwidth BW_{jl} on link l, and buffer BF_{jl} at the router just before link l. If link l has a total (unidirectional) bandwidth BW_l available to real-time traffic, then the bandwidth reservation allocations must obey $\sum_{j=1}^{m} BW_{jl} \leq BW_l$. Assume routers are output-buffered and that the total buffer available to real-time traffic with output link l is BF_l. The corresponding buffer constraint is therefore $\sum_{j=1}^{m} BF_{jl} \leq BF_l$. The maximum delay for class j traffic at link l is therefore $D_{jl} = BF_{jl}/BW_{jl}$. Note that we do not consider the actual queueing delay experienced by the traffic, but rather only a bound on the delay at each router. The delay bound at each router therefore does not depend directly on the traffic characteristics, but only upon the allocated bandwidth and buffer. Each user class can specify a bound on the end-to-end delay, which is interpreted by the network as a bound on the sum of the delay bounds at each router along the path, $\sum_{l=1}^{n_j} D_{jl} \leq D_j$, where n_j is the number of links on the virtual path for class j traffic.

We use the term *user*, or class, to refer to an aggregate of flows with similar traffic characterizations and similar utility functions. Utility is assumed to be a function of loss probability, which in turn depends on the reserved bandwidth and buffer at each node. The user model consists of two functions: a traffic model, which describes the statistics of each real-time flow aggregate, and a QoS model, which describes how each user measures satisfaction with its QoS.

The traffic model is a function that maps the number of sources in an aggregate and the amount of bandwidth and buffer reserved at each hop into the loss probability that these flows experience. Such a function might be derived using the literature on traffic characterization or measured. Specifically, assume there are N_j independent and identically distributed flows within class j. We denote the probability of loss of the multiplexed class on link l, L_{jl}, as a function of the bandwidth allocated to the class, BW_{jl}, the buffer allocated to the class, BF_{jl}, and the number of sources N_j, as $L_{jl} = g_j(BW_{jl}, BF_{jl}, N_j)$. The loss function for class j traffic, g_j, might be given for instance by effective bandwidth results. We presume that this loss function is independent of the link number, which is reasonable if the allocated bandwidth and buffer are sufficient to decouple the effective bandwidths [36]. We also presume that each source within a class experiences the same loss probability. Finally, we presume that the loss function is decreasing, differentiable, and jointly strictly convex in $\{BW_{jl}, BF_{jl}\}$. We assume that the loss on each link is small and independent of other losses within the path, given bandwidth and buffer allocations. Therefore the total end-to-end loss probability for class j is $L_j = \sum_{l=1}^{n_j} L_{jl}$.

The QoS model is a function that maps the number of sources in an aggregate and the loss probability into a *utility*. This utility can be either interpreted directly as the amount that these sources would be willing to pay for this level of QoS, or indirectly interpreted simply as a numerical measure of satisfaction. Specifically, we assume class j derives a satisfaction $U_j(L_j, N_j)$ from supporting N_j sources with a probability of loss of L_j. In this paper, we will assume that $U_j(L_j, N_j)$ is decreasing, differentiable, and strictly concave in L_j. The utility can also be thought of directly as a function of reserved bandwidth and buffer, in which case it can be shown that $U_j(L_j, N_j)$ is strictly concave in $\{BW_{jl}, BF_{jl}\}$. We will also assume that marginal utilities with respect to bandwidth and buffer are each infinite when these resources are zero, so that infinite resource prices imply zero demand. We note that many application's utility curves may not be strictly concave with respect to bandwidth and buffer. Researchers have often argued that the utility function of elastic applications is usually strictly concave everywhere with respect to bandwidth, but the utility function of delay-adaptive audio and video applications is usually S-shaped and convex in a region of small bandwidths, c.f. [37]. Rationale users, however, either operate in the concave region or drop out. We do not consider this case in this paper, but leave the issue to the admission control policy.

2.2 Distributed resource allocation using pricing

We now pose our network resource allocation problem. We assume that the network attempts to maximize total utility of all active users, by choosing bandwidth and buffer allocations on each link and at each router. The corresponding problem is **Problem U**:

$$\max_{BW_{jl},BF_{jl}} \sum_{j=1}^{m} U_j(L_j, N_j)$$
$$s.t. \quad \sum_{j=1}^{m} BW_{jl} \leq BW_l, \ \sum_{j=1}^{m} BF_{jl} \leq BF_l$$
$$\sum_{l=1}^{n_j} D_{jl} \leq D_j, \ BW_{jl} \geq 0, \ BF_{jl} \geq 0$$

We show in [38] that this is a concave program, the solution is unique, and that it can be characterized using shadow costs corresponding to each of the constraints:

Theorem 1. *The resource allocation* $\{BW_{jl}, BF_{jl}\}$ *solves Problem U if and only if there exists a set of nonnegative shadow costs* $\{\alpha_j, \beta_l, \gamma_l\}$ *such that:*

$$\frac{\partial U_j}{\partial BW_{jl}} = \beta_l + \alpha_j \frac{\partial D_{jl}}{\partial BW_{jl}}, \ \frac{\partial U_j}{\partial BF_{jl}} = \gamma_l + \alpha_j \frac{\partial D_{jl}}{\partial BF_{jl}}$$
$$\beta_l(\sum_{j=1}^{m} BW_{jl} - BW_l) = 0, \ \gamma_l(\sum_{j=1}^{m} BF_{jl} - BF_l) = 0$$
$$\alpha_j(\sum_{l=1}^{n_j} D_{jl} - D_j) = 0 \tag{1}$$

We interpret β_l and γ_l as prices for bandwidth and buffer, respectively, on link l. We also interpret $\alpha_j \frac{\partial D_{jl}}{\partial BW_{jl}}$ and $\alpha_j \frac{\partial D_{jl}}{\partial BF_{jl}}$ as surcharges for bandwidth and buffer, respectively, to class j traffic on link l if the delay constraint for that class is binding. This pricing model suggests that the resource allocation can be implemented as a distributed optimization. Suppose the network and users follow the following strategies:

Network Algorithm N1: Update the prices from $\{\alpha_j, \beta_l, \gamma_l\}$ to $\{\alpha'_j, \beta'_l, \gamma'_l\}$ as follows: $\alpha'_j = max(\alpha_j + \Delta\alpha_j, 0)$, $\beta'_l = max(\beta_l + \Delta\beta_l, 0)$, and $\gamma'_l = max(\gamma_l + \Delta\gamma_l, 0)$ where the feedback algorithms are designed so that $sgn(\Delta\alpha_j) = sgn(\sum_{l=1}^{n_j} D_{jl} - D_j)$, $sgn(\Delta\beta_l) = sgn(\sum_{j=1}^{m} BW_{jl} - BW_l)$, and $sgn(\Delta\gamma_l) = sgn(\sum_{j=1}^{m} BF_{jl} - BF_l)$.

User Algorithm U1: In each class j, choose bandwidth and buffer allocations that maximize surplus, where surplus is defined as utility minus cost:

$$\max_{BW_{jl},BF_{jl}} U_j(L_j, N_j) - \sum_{l=1}^{n_j} \alpha_j \frac{BF_{jl}}{BW_{jl}} - \sum_{l=1}^{n_j} \beta_l BW_{jl} - \sum_{l=1}^{n_j} \gamma_l BF_{jl}$$

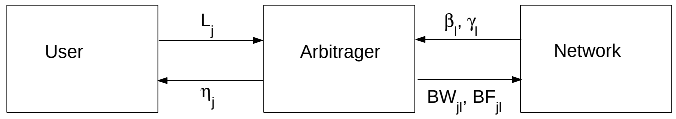

Fig. 1. Communication between user, arbitrager, and network

A few caveats are in order here. We are not suggesting that these algorithms should be implemented in the Internet as written, but only mean to demonstrate what a pricing approach might attempt to accomplish. Any user response would undoubtedly have to be done via a user agent to automate the response, the time scales for each iteration would have to be chosen and the feedback algorithms would have to be designed to guarantee convergence. Also, much additional work would have to be done to make any such approach implementable, including development of signalling protocols, reservation protocols, measurement algorithms, and scheduling policies.

The equilibrium point of this distributed resource allocation process is optimal.
Theorem 2. *The resource allocation* $\{BW_{jl}, BF_{jl}\}$ *solves Problem U if and only if it is an equilibrium for Network Algorithm N1 and the User Algorithms U1 for each class.*

The proof is omitted here.

To end users, how to allocate buffer and bandwidth is meaningless. Instead, they care about the QoS of their applications. So we consider loss and delay as intermediate variables and based on that we consider a second implementation in which an arbitrager layer is introduced in between the user and the network. Each network link periodically updates the prices for buffer and bandwidth on that link, based on the differences between the total demands and supplies for buffer and bandwidth. The arbitrager for class j periodically receives a request for a specified loss level from the class j users and finds the minimum cost allocation of buffer and bandwidth on each link on route j. The arbitrager also calculates a cost per unit loss, η_j, and advertises this cost to the class j users. The class j users periodically calculate the desired loss L_j based on the cost η_j and the utility function for that class. This process is illustrated in figure 1.

To simplify the notation, we denote $U_w^{(jl)} \equiv \partial U_j/\partial BW_{jl}$, $U_f^{(jl)} \equiv \partial U_j/\partial BF_{jl}$, $U_g^{(j)} \equiv \partial U_j/\partial L_j$, $U_{gg}^{(j)} \equiv \partial U_g^{(j)}/\partial L_j$, $g_w^{(jl)} \equiv \partial L_j/\partial BW_{jl}$, $g_f^{(jl)} \equiv \partial L_j/\partial BF_{jl}$, $g_{ww}^{(jl)} \equiv \partial g_w^{(jl)}/\partial BW_{jl}$, $g_{wf}^{(jl)} \equiv \partial g_w^{(jl)}/\partial BF_{jl}$, and $g_{ff}^{(jl)} \equiv \partial g_f^{(jl)}/\partial BF_{jl}$. The marginal benefits to class j for adding more bandwidth and buffer can be represented as $U_w^{(jl)} = U_g^{(j)} g_w^{(jl)}$ and $U_f^{(jl)} = U_g^{(j)} g_f^{(jl)}$.

The network algorithm (N1) remains the same as above. The arbitrager and user algorithms are formalized as follows:

Arbitrager Algorithm A2: In each class j, choose the minimum cost bandwidth and buffer allocations that achieve a target loss L_j:

$$\min_{BW_{jl},BF_{jl}} \sum_{l=1}^{n_j}(\beta_l+\hat{\beta}_{jl})BW_{jl}+\sum_{l=1}^{n_j}(\gamma_l+\hat{\gamma}_{jl})BF_{jl}$$
$$s.t.\ \sum_{l=1}^{n_j} g_j(BW_{jl},BF_{jl}) \le L_j$$

Set a price *per unit loss* for class j as the minimum marginal cost:

$$\eta_j \equiv min(\frac{\beta_l+\hat{\beta}_{jl}}{g_w^{(jl)}},\frac{\gamma_l+\hat{\gamma}_{jl}}{g_f^{(jl)}})$$

User Algorithm U2: In each class j, choose loss that maximizes consumer surplus, where the surplus is defined as utility minus cost,

$$\max_{L_j} U_j(L_j)-\eta_j L_j \tag{2}$$

This distributed process simplifies the task for the users by removing direct consideration of network resources and instead focusing on the QoS parameters of concern. In addition, the arbitrager could be used to guarantee a fixed price for each user connection by acting as an insurance agent who charges a small premium in order to smooth out fluctuations in the resource prices, although we do not consider this here.

In [38], we show that the equilibrium point of this distributed resource allocation process is optimal:
Theorem 3. *The resource allocation $\{BW_{jl}, BF_{jl}\}$ solves Problem U if and only if it is an equilibrium for Network Algorithm, the Arbitrager Algorithm for each class and the User Algorithm for each class.*

Theorem 3 establishes that the approach using separate user, arbitrager, and network algorithms achieves the maximal utility in a distributed fashion. It is a result on equilibrium, but not on dynamics. Freedom is left to design feedback algorithms that achieve convergence on a time-scale of interest.

3 Sensitivity to QoS

Feedback algorithms for congestion-based pricing typically assume that network resources behave as Normal goods, i.e. that an increase in the price of a resource results in a decreased demand for that resource. We examine this assumption by considering the relationship between network resources, QoS, and cost. While the user and arbitrager algorithms operate on each class, the network algorithm sets prices for bandwidth and buffer at each node in response to *total* demand for these resources from all classes. A change in the traffic of one class will cause changes in the prices for bandwidth and buffer at all nodes on its route, and therefore affect the resources that should be allocated to other traffic sharing the route as well as the resulting QoS. It is important to understand the sensitivity of the optimal allocation and resulting QoS to such price changes in order to understand how resources should be allocated. These sensitivities are crucial to the design of feedback algorithms for congestion-based pricing.

We begin by examining the set of minimum cost resource allocations under a range of loss. We next characterize the form of the resulting minimum cost function.

3.1 Minimum cost allocations

We consider here the set of minimum cost allocations of bandwidth and buffer that class j on link l will adopt under a fixed set of prices, over a range of loss probabilities L_{jl}. We wish to characterize the form of the set, when the delay constraint for class j is not binding. Since loss probability L_{jl} is jointly strictly convex in bandwidth and buffer, it follows that buffer is strictly convex in bandwidth given fixed loss. The minimum cost allocation for a specified loss and price ratio is the point on the given loss contour that is tangent to a cost contour (if any). Since the loss contour is strictly convex, this minimum cost allocation is unique. The set of such minimum cost allocations of bandwidth and buffer over different loss probabilities thus forms a unique curve. We call this curve the *expansion path.*

An example of expansion path is shown in figure 2. In this example, we consider the traffic as an aggregate of on/off fluid flows with independent and exponentially distributed on and off times. Unless otherwise stated, we assume 500 sources (N_j) with a mean on-time of 340ms and mean off-time 780ms, corresponding to a duty cycle of $p_j \approx 0.3036$. We measure bandwidth in multiples of 8kbps (the peak rate) and buffer in multiples of 340B (the mean number of bytes per cycle). We calculate loss using an effective bandwidth function derived by Morrison [24].

The fundamental question is whether bandwidth and buffer act as Normal goods, i.e. whether a decrease in packet loss is paired with an increase in both bandwidth and buffer. To characterize the expansion path, we investi-

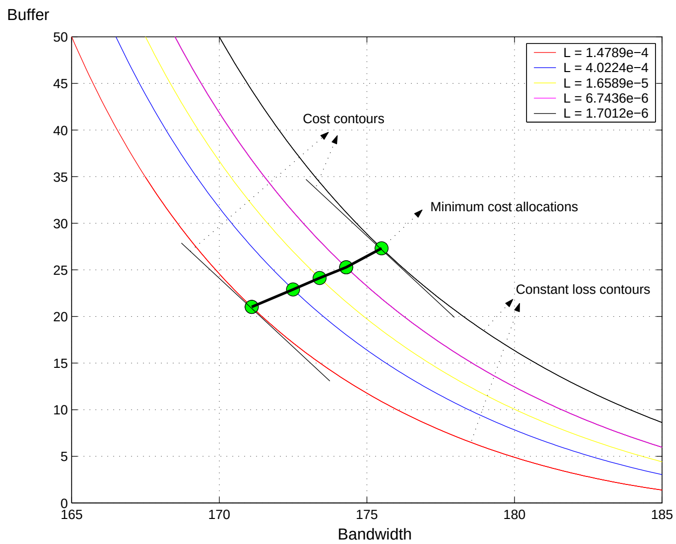

Fig. 2. Minimum cost allocations

gate the sensitivity of bandwidth and buffer on link l along the expansion path to loss on link l, denoted $dBW_{jl}/dL_{jl}|_{ep}$ and $dBF_{jl}/dL_{jl}|_{ep}$ respectively, and the corresponding slope of the expansion path, denoted $dBF_{jl}/dBW_{jl}|_{ep}$. The sign of these sensitivities is of particular interest. Normally one would expect that in order to decrease the loss on a particular link, one would increase both the bandwidth and buffer allocations, namely $dBW_{jl}/dL_{jl}|_{ep} < 0$, $dBF_{jl}/dL_{jl}|_{ep} < 0$, and $dBF_{jl}/dBW_{jl}|_{ep} > 0$. In economics terminology, this means that bandwidth and buffer on link l are *Normal goods*, and that the slope of the expansion path is positive. We find however that these signs hold only under certain conditions.

The following notation will be helpful:

$$x_l = (g_f^{(jl)}, -g_w^{(jl)}),\ \ G_{jl} = \begin{pmatrix} g_{ww}^{(jl)} & g_{wf}^{(jl)} \\ g_{wf}^{(jl)} & g_{ff}^{(jl)} \end{pmatrix}$$

$$h_g^{(jl)} = |G_{jl}| \tag{3}$$

$$k_1^{(jl)} = g_w^{(jl)} g_{ff}^{(jl)} - g_f^{(jl)} g_{wf}^{(jl)} \tag{4}$$

$$k_2^{(jl)} = g_f^{(jl)} g_{ww}^{(jl)} - g_w^{(jl)} g_{wf}^{(jl)} \tag{5}$$

$$\begin{aligned} K^{(jl)} &= x_l G_{jl} x_l^t \\ &= (g_f^{(jl)})^2 g_{ww}^{(jl)} - 2 g_w^{(jl)} g_f^{(jl)} g_{wf}^{(jl)} + (g_w^{(jl)})^2 g_{ff}^{(jl)} \end{aligned} \tag{6}$$

We say link l has *increasing bandwidth along the expansion path*, denoted $IBW^{(jl)}$, if $k_1^{(jl)} < 0$, namely if $g_{wf}^{(jl)} < (g_w^{(jl)}/g_f^{(jl)}) g_{ff}^{(jl)}$. Similarly, we say link l has *increasing buffer along the expansion path*, denoted $IBF^{(jl)}$, if $k_2^{(jl)} < 0$, namely if $g_{wf}^{(jl)} < (g_f^{(jl)}/g_w^{(jl)})g_{ww}^{(jl)}$. Finally, we say link l has *increasing bandwidth and buffer along the expansion path*, denoted $IWF^{(jl)}$, if $k_1^{(jl)} < 0, k_2^{(jl)} < 0$, namely if $g_{wf}^{(jl)} < min[(g_w^{(jl)}/g_f^{(jl)})g_{ff}^{(jl)}, (g_f^{(jl)}/g_w^{(jl)})g_{ww}^{(jl)}]$.

Our first key result is whether these properties hold, and correspondingly whether bandwidth and buffer act as Normal goods, depends on the shape of the utility and loss functions:

Theorem 4.

$$\frac{dBW_{jl}}{dL_{jl}}|_{ep} = \frac{k_1^{(jl)}}{K^{(jl)}}, \ \frac{dBF_{jl}}{dL_{jl}}|_{ep} = \frac{k_2^{(jl)}}{K^{(jl)}} \tag{7}$$

$$\frac{dBF_{jl}}{dBW_{jl}}|_{ep} = \frac{k_2^{(jl)}}{k_1^{(jl)}} \tag{8}$$

Proof. From the optimal shadow costs in equation (1), when the delay constraint for class j is not binding, we obtain:

$$\frac{\gamma_l}{\beta_l} = \frac{g_f^{(jl)}}{g_w^{(jl)}} \tag{9}$$

Consider a small change in loss, ΔL_{jl}, on link l, produced by small changes in bandwidth and buffer, respectively denoted ΔBW_{jl} and ΔBF_{jl}:

$$\Delta L_{jl} = g_w^{(jl)}\Delta BW_{jl} + g_f^{(jl)}\Delta BF_{jl} \tag{10}$$

To be on the expansion path, the ratio $g_f^{(jl)}/g_w^{(jl)}$ at the new loss level must still be equal to the price ratio γ_l/β_l, namely:

$$\begin{aligned}
&\lim_{\Delta L_{jl}\to 0} \frac{g_w^{(jl)} + g_{ww}^{(jl)}\Delta BW_{jl} + g_{wf}^{(jl)}\Delta BF_{jl}}{g_f^{(jl)} + g_{wf}^{(jl)}\Delta BW_{jl} + g_{ff}^{(jl)}\Delta BF_{jl}} = \frac{\beta_l}{\gamma_l} \\
\Rightarrow &\lim_{\Delta L_{jl}\to 0} \frac{g_{ww}^{(jl)}\Delta BW_{jl} + g_{wf}^{(jl)}\Delta BF_{jl}}{g_{wf}^{(jl)}\Delta BW_{jl} + g_{ff}^{(jl)}\Delta BF_{jl}} = \frac{\beta_l}{\gamma_l}
\end{aligned} \tag{11}$$

Solving equations (10) and (11), and substituting (9), we can establish the sensitivity of bandwidth and buffer on link l to loss on link l, along the expansion path, given by equations (7). Combining these two expressions, we can establish the slope of the expansion path, given by (8). □

We can now prove the fundamental result that bandwidth on link l is a Normal good iff $IBW^{(jl)}$ holds, and that buffer on link l is a Normal good iff $IBF^{(jl)}$ holds.

Corollary 1. *$dBW_{jl}/dL_{jl}|_{ep} < 0$ iff $IBW^{(jl)}$, $dBF_{jl}/dL_{jl}|_{ep} < 0$ iff $IBF^{(jl)}$, and $dBF_{jl}/dBW_{jl}|_{ep} > 0$ iff $IWF^{(jl)}$.*

Proof. We start with the denominators of (7), $K^{(jl)}$, which can be expressed as $x_l G_{jl} x_l^t$. By assumption, g^j is decreasing and jointly strictly in $\{BW_{jl},\ BF_{jl}\}$. It follows that the hessian of g_j is positive-definite, and therefore the denominator is positive. The sign of the numerators are determined by the hypotheses. The theorem follows. □

Theorem 4 gives the slope of the expansion path. This information can be used to adjust bandwidth and buffer in the optimal proportion in response to changes in desired levels of loss, as opposed to methods which may either increase only bandwidth (with fixed buffer) or may increase both bandwidth and buffer in a fixed linear proportion. Corollary 1 gives sufficient conditions for the slope of the expansion path to be positive.

In our example, the traffic model we used has the property that $IBW^{(jl)}$ and $IBF^{(jl)}$ hold for $BW_{jl} \in (N_j p_j, \infty)$ and $BF_{jl} \in (0, \infty)$, where $N_j p_j$ is the mean rate of the class j. In other words, bandwidth and buffer of the class j are both Normal goods, and the slope of the expansion path is positive, as pictured in figure 2.

3.2 Minimum cost function

Having established the variation of the optimal resource allocation with changes in loss probability, we turn to characterizing the corresponding minimum cost of these allocations.

First we consider a single link. Denote by $C_{jl}(L_{jl})$ the minimum cost on link l to obtain a loss probability of L_{jl}, namely:

$$\begin{aligned} C_{jl}(L_{jl}) = & \min_{(BW_{jl}, BF_{jl})} \beta_l BW_{jl} + \gamma_l BF_{jl}, \\ & s.t. \quad g_j(BW_{jl}, BF_{jl}) = L_{jl} \end{aligned}$$

Lemma 1. *The minimum link cost $C_{jl}(L_{jl})$ is a decreasing and convex function of loss probability on that link L_{jl}.*

Proof. The marginal cost with respect to loss at the cost minimization point is $\partial C_{jl}/\partial L_{jl} = \beta_l / g_w^{(jl)}$ and therefore the minimum cost is a decreasing function of loss probability. To demonstrate that it is also convex:

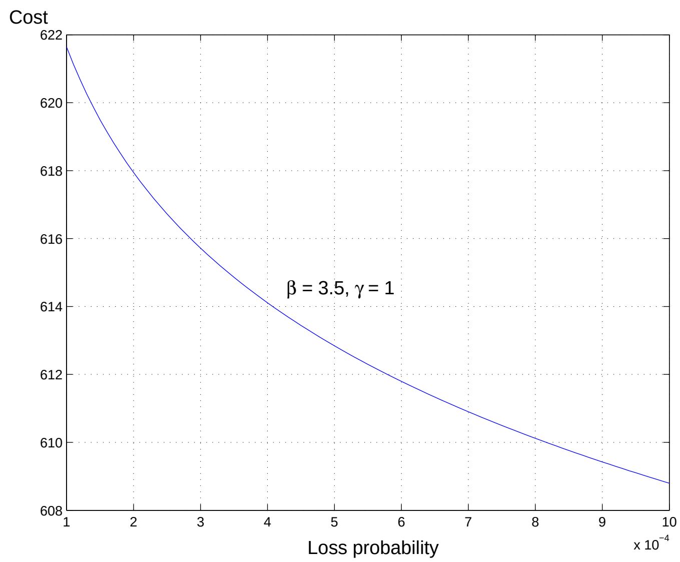

Fig. 3. Minimum cost function with $\beta = 3.5$ and $\gamma = 1$

$$\begin{aligned}\frac{\partial^2 C_{jl}}{\partial L_{jl}^2} &= -\frac{\beta}{(g_w^{(jl)})^2}\frac{dg_w^{(jl)}}{dL_{jl}}|_{ep} \\ &= -\frac{\beta}{(g_w^{(jl)})^2}\left[g_{ww}^{(jl)}\frac{dBW_{jl}}{dL_{jl}}|_{ep} + g_{wf}^{(jl)}\frac{dBF_{jl}}{dL_{jl}}|_{ep}\right]\end{aligned}$$

Substituting equations (7) and simplifying:

$$\frac{\partial^2 C_{jl}}{\partial L_{jl}^2} = -\frac{\beta}{g_w^{(jl)}}\frac{h_g^{(jl)}}{K^{(jl)}}$$

where $K^{(jl)}$ is the denominator in (7). Now $h_g^{(jl)} > 0$ since it is the Hessian of a positive-definite function, $K^{(jl)} > 0$ was shown above, and $-\beta/g_w^{(jl)} > 0$, so it follows that $\partial^2 C_{jl}/\partial L_{jl}^2 > 0$, and hence that the minimum cost is a convex function of loss probability. $\square$

Given the previous example, the cost function with respect to loss when the price for bandwidth is 3.5 and the price for buffer is 1 is illustrated in figure 3. It shows the cost is convex and decreasing in loss probability.

Now we consider the total minimum cost that achieves a loss probability of L_j:

$$\begin{aligned}C_j(L_j) = &\min_{\{(BW_{jl}, BF_{jl})\}} \sum_{l=1}^{n_j} C_{jl}(g_j(BW_{jl}, BF_{jl})) \\ &s.t. \qquad \sum_{l=1}^{n_j} g_j(BW_{jl}, BF_{jl}) = L_j\end{aligned}$$

Note that the minimum cost solution results in the optimal allocation of losses to each link and in the optimal allocation of bandwidth and buffer on each link to achieve the corresponding loss. We can establish that this minimum cost solution has a similar form to the cost of an individual link:

Theorem 5. *The minimum total cost $C_j(L_j)$ is a decreasing and convex function of the total loss probability L_j.*

Proof. The optimal shadow costs in equation (1) can be used to establish that, at equilibrium,

$$-U_g^{(j)} = -\frac{\beta_l}{g_w^{(jl)}} = -\frac{\gamma_l}{g_f^{(jl)}} \quad \forall l$$

which means the marginal cost per unit loss, at the optimal allocation, is the same for each link along the route, namely:

$$\frac{\partial C_j}{\partial L_j} = \frac{\partial C_{jl}}{\partial L_{jl}} \quad \forall l \tag{12}$$

Furthermore, these marginal costs are negative, from Lemma 1, establishing that $C_j(L_j)$ is a decreasing function of the total loss probability L_j.

Differentiating (12) again:

$$\frac{\partial^2 C_j}{\partial L_j^2} = \frac{\partial^2 C_{jl}}{\partial L_{jl}^2}\frac{dL_{jl}}{dL_j} \quad \forall l \tag{13}$$

Also,

$$\sum_{l=1}^{n_j} \frac{dL_{jl}}{dL_j} = 1 \tag{14}$$

Solving (13) for dL_{jl}/dL_j and substituting into (14) gives:

$$\frac{\partial^2 C_j}{\partial L_j^2}\sum_{l=1}^{n_j} \frac{1}{\frac{\partial C_{jl}}{\partial L_{jl}}} = 1$$

Combining the terms in the summation into a common denominator and moving it to the right hand side, we get:

$$\frac{\partial^2 C_j}{\partial L_j^2} = \frac{\prod_{l=1}^{n_j} \frac{\partial^2 C_{jl}}{\partial L_{jl}^2}}{\sum_{l=1}^{n_j}(\prod_{i\neq l} \frac{\partial^2 C_{ji}}{\partial L_{ji}^2})}$$

The numerator consists of the product of positive terms, by Lemma 1, and the denominator similarly consists of the sum of positive terms. Therefore the entire expression is positive, establishing that the minimum total cost $C_j(L_j)$ is a convex function of the total loss probability L_j. □

Congestion-based pricing algorithms typically are based on feedback algorithms in which optimality is achieved at an equilibrium point characterized by equality of demand and supply. Theorem 5 implies that the price per unit loss, η_j, that class j faces has an increasing marginal cost for each marginal decrease in loss probability. Since a class's utility function is assumed to be decreasing and concave in loss, the marginal utility a class gains is decreasing for each marginal decrease in loss probability. The combination of an increasing marginal cost with a decreasing marginal utility (as loss decreases) results in a unique equilibrium at which marginal cost equals marginal utility. This form is crucial for the design of stable feedback algorithms for congestion-based pricing.

4 Sensitivity to price without a delay constraint

We can use the properties of the minimum cost functions to characterize the sensitivity of the optimal allocation with respect to changes in a resource price, e.g. caused by changes in demand of other users. In the first subsection, we give sufficient conditions to establish that an increase in the price for a resource results in an increase in the loss at that node, a decrease in the loss at all other nodes, and an increase in the total loss. In the next subsection, we also give sufficient conditions to establish that an increase in the price for a resource results in a decreased demand for that resource, an increased demand for the other resource at that node, and an increased demand for resources at all other hops.

4.1 Sensitivity of loss to price

We start by examining sensitivities of loss to resource prices. We prove that if a resource is a Normal good, then an increase in the price of that resource causes the loss on that link to increase, the loss on all other links to decrease, and the total loss to increase.

Theorem 6. *If $IBW^{(jl)}$ holds, then $\partial L_{jl}/\partial\beta_l > 0$, $\partial L_{jl'}/\partial\beta_l < 0$ $\forall l' \neq l$, and $\partial L_j/\partial\beta_l > 0$. If $IBF^{(jl)}$ holds, then $\partial L_{jl}/\partial\gamma_l > 0$, $\partial L_{jl'}/\partial\gamma_l < 0$ $\forall l' \neq l$, and $\partial L_j/\partial\gamma_l > 0$.*

Proof. We begin by aggregating all links other than link l into a single virtual link denoted $\bar{l}$. The virtual link satisfies the property that the minimum total cost of a user class on those aggregated links is decreasing and convex in the corresponding total loss, given by Theorem 5. Using this property, it follows that minimizing total cost along route j is equivalent to minimizing the sum of the costs on link l and virtual link $\bar{l}$. The minimum cost to achieve an aggregate loss of $L_{j\bar{l}}$ on virtual link $\bar{l}$ is:

$$C_{j\bar{l}}(L_{j\bar{l}}) = \min_{\{L_{jl'},\ l' \neq l\}} \sum_{l' \neq l} C_{jl'}(L_{jl'}),\ \ s.t. \sum_{l' \neq l} L_{jl'} = L_{j\bar{l}}$$

We can then write the surplus for class j as:

$$f_1(L_{jl}, -L_{j\bar{l}}, \beta_l) \equiv U_j(L_{jl} + L_{j\bar{l}}) - C_{jl}(L_{jl}) - C_{j\bar{l}}(L_{j\bar{l}})$$

The optimal loss on links l and $\bar{l}$ are then given by $\arg\max_{L_{jl}, L_{j\bar{l}}} f_1(L_{jl}, -L_{\overline{j\bar{l}}}, \beta_l)$. The proof proceeds in two steps. First, we show that $f_1(L_{jl}, -L_{j\bar{l}}, \beta_l)$ is parameter monotonic in β_l, which will establish that $\partial L_{jl}/\partial \beta_l > 0$ and $\partial L_{j\bar{l}}/\partial \beta_l < 0$. Second, we show that $\partial L_{j\bar{l}}/\partial \beta_l < 0$ implies that $\partial L_{jl'}/\partial \beta_l < 0\ \forall l' \neq l$.

Define $s = (L_{jl}, -L_{j\bar{l}})$, $S = \mathbb{R}_+ \times \mathbb{R}_-$ as the set of possible values of s, and $\Theta = \mathbb{R}_+$ as the set of possible values of β_l. Then S is a compact sublattice of $\mathbb{R}^2$, and Θ is a sublattice of $\mathbb{R}$. The function $f_1 : S \times \Theta \rightarrow \mathbb{R}$ is a continuous function on S for each fixed β_l.

We wish to first establish that f_1 is supermodular in $\{L_{jl}, -L_{j\bar{l}}, \beta_l\}$. Consider the cross-derivatives of $f_1(L_{jl}, -L_{j\bar{l}}, \beta_l)$. Of these three variables, note that U_j is not a function of β_l, $C_{jl}(L_{jl})$ is not a function of $L_{j\bar{l}}$, and $C_{j\bar{l}}(L_{j\bar{l}})$ is not a function of β_l nor of L_{jl}. Therefore:

$$\frac{\partial^2 f_1}{\partial L_{jl} \partial \beta_l} = -\frac{\partial^2 C_{jl}}{\partial L_{jl} \partial \beta_l}$$

Differentiating C_{jl} first with respect to β_l, we get:

$$\frac{\partial^2 C_{jl}}{\partial L_{jl} \partial \beta_l} = \frac{dBW_{jl}}{dL_{jl}}|_{ep} \tag{15}$$

Furthermore, $dBW_{jl}/dL_{jl}|_{ep} < 0$ iff $IBW^{(jl)}$ by Corollary 1, so $\partial^2 f_1/\partial L_{jl} \partial \beta_l > 0$. Also,

$$\frac{\partial^2 f_1}{\partial L_{jl} \partial(-L_{j\bar{l}})} = \frac{\partial^2 U_j}{\partial L_{jl} \partial(-L_{j\bar{l}})} = -U_{gg}^{(j)} > 0 \tag{16}$$

since U_j is assumed to be strictly concave. Finally,

$$\frac{\partial^2 f_1}{\partial \beta_l \partial(-L_{j\bar{l}})} = 0 \tag{17}$$

Together, (15), (16), and (17) prove that f_1 is supermodular in $\{L_{jl}, -L_{j\bar{l}}, \beta_l\}$ (c.f. [39], Theorem 10.4).

It follows that f_1 satisfies increasing differences in (s, β_l) and is supermodular in s for each fixed β_l (c.f. [39], Theorem 10.3). These conditions are sufficient to conclude that $\arg\max\{f_1(s,\beta_l)|s \in S\}$ admits a greatest element $s^*(\beta_l)$, and that this greatest element is parameter monotonic in β_l (c.f. [39], Theorem 10.7). However since $\arg\max\{f_1(s,\beta_l)|s \in S\}$ contains a single point for each fixed β_l, then it follows that $f_1(L_{jl}, -L_{j\bar{l}}, \beta_l)$ is parameter monotonic in β_l.

In particular, this proves that if $IBW^{(jl)}$ holds, then $\partial L_{jl}/\partial\beta_l > 0$. To show that $\partial L_{jl'}/\partial\beta_l < 0\ \forall l' \neq l$, we must still show that $\partial L_{j\bar{l}}/\partial\beta_l < 0$ implies that $\partial L_{jl'}/\partial\beta_l < 0\ \forall l' \neq l$.

To demonstrate this, consider a link $l' \neq l$. We now aggregate all links in $\bar{l}$ except for link l' into a virtual link $\overline{l'}$. The minimum cost to achieve an aggregate loss of $L_{j\overline{l'}}$ on virtual link $\overline{l'}$ is:

$$C_{j\overline{l'}}(L_{j\overline{l'}}) = \min_{\{L_{jl''}, l'' \neq l, l'\}} \sum_{l'' \neq l, l'} C_{jl''}(L_{jl''})$$

Define:

$$f_2(L_{jl'}, L_{j\bar{l}}) \equiv -C_{jl'}(L_{jl'}) - C_{j\overline{l'}}(L_{j\bar{l}} - L_{jl'})$$

Consider the cross-derivative of $f_2(L_{jl'}, L_{j\bar{l}})$:

$$\frac{\partial^2 f_2}{\partial L_{jl'} \partial L_{j\bar{l}}} = \frac{\partial^2 C_{j\overline{l'}}}{\partial L_{jl'} \partial L_{j\bar{l}}} = \frac{\partial^2 C_{j\overline{l'}}}{\partial L^2_{j\overline{l'}}} > 0$$

using Theorem 5. It follows that $f_2(L_{jl'}, L_{j\bar{l}})$ is supermodular in $(L_{jl'}, L_{j\bar{l}})$, and therefore that it is parameter monotonic in $L_{jl'}$ with respect to $L_{j\bar{l}}$. Combining this with $\partial L_{j\bar{l}}/\partial\beta_l < 0$ proves that $\partial L_{jl'}/\partial\beta_l < 0\ \forall l' \neq l$. A similar approach can be taken to prove that if $IBF^{(jl)}$ holds, then $\partial L_{jl}/\partial\gamma_l > 0$ and $\partial L_{jl'}/\partial\gamma_l < 0\ \forall l' \neq l$.

To determine the sign of the sensitivity of the *total* loss to prices, we rewrite the surplus for class j as:

$$f_3(L_{jl}, L_j, \beta_l) \equiv U_j(L_j) - C_{jl}(L_{jl}) - C_{j\bar{l}}(L_j - L_{jl})$$

A similar approach as we took with f_1 above can be used to prove that if $IBW^{(jl)}$ holds then $\partial L_j/\partial\beta_l > 0$, and that if $IBF^{(jl)}$ holds then $\partial L_j/\partial\gamma_l > 0$. $\square$

Theorem 6 is an important property in the design of the arbitrager feedback algorithms. If a resource is a Normal good, then when a network node raises the price, the arbitrager should respond by buying a combination of resources on that link that produces a higher loss and a combination of resources on other links that produce lower losses. As a consequence, the arbitrager should respond by raising (in absolute value) the price per unit loss, which in turn results in an increase in the requested loss probability in that class.

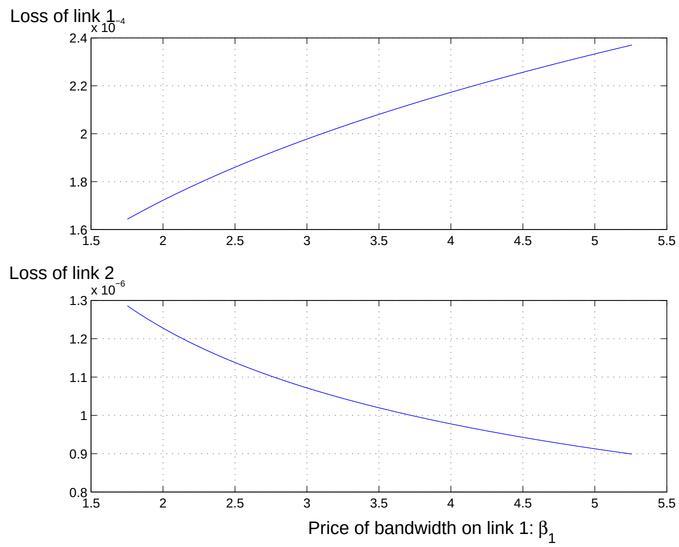

Fig. 4. Sensitivity of loss to price

Figure 4 illustrates the result in theorem 6. We consider an aggregate of data flows that has the same traffic model as given in the previous example and traverses two links in the network. As we discussed before, this traffic model has the property that both bandwidth and buffer of class j are Normal goods. When the price of bandwidth on link 1 increases, we observe that the loss probability on link 1 increases, and that the loss on link 2 decreases.

4.2 Sensitivity of resources to price

The last theorem gives sufficient conditions to establish that an increase in the price for a resource results in an increase in the loss at that node, a decrease in the loss at all other nodes, and an increase in the total loss. It does not, however, address the resulting changes in buffer and bandwidth at each node. We consider these in this subsection.

Let $\mu_j = [\beta_1, \gamma_1, \cdots, \beta_{n_j}, \gamma_{n_j}]^t$ denote the set of resource prices for class j, $x_j = [BW_{j1}, BF_{j1}, \cdots, BW_{jn_j}, BF_{jn_j}]^t$ denote the set of resource allocation for class j, and $U_j(x_j)$ denote the resulting utility. Let $\tilde{x}_j$ denote the resource allocation for class j that maximizes the class's surplus, i.e. $\tilde{x}_j = \arg\max_{x_j} U_j(x_j) - \sum_{l=1}^{n_j} (\beta_l BW_{jl} + \gamma_l BF_{jl})$. So we have:

$$\nabla_{\mu_j} x_j = \begin{pmatrix} \frac{\partial BW_{j1}}{\partial \beta_1} & \frac{\partial BF_{j1}}{\partial \beta_1} & \dots & \frac{\partial BW_{jn_j}}{\partial \beta_1} & \frac{\partial BF_{jn_j}}{\partial \beta_1} \\ \frac{\partial BW_{j1}}{\partial \gamma_1} & \frac{\partial BF_{j1}}{\partial \gamma_1} & \dots & \frac{\partial BW_{jn_j}}{\partial \gamma_1} & \frac{\partial BF_{jn_j}}{\partial \gamma_1} \\ \dots & & & & \\ \frac{\partial BW_{j1}}{\partial \gamma_{n_j}} & \frac{\partial BF_{j1}}{\partial \gamma_{n_j}} & \dots & \frac{\partial BW_{jn_j}}{\partial \gamma_{n_j}} & \frac{\partial BF_{jn_j}}{\partial \gamma_{n_j}} \end{pmatrix} \tag{18}$$

which is the sensitivity matrix that we wish to understand. And we also have

$$\nabla^2 U_j(x_j) = \begin{pmatrix} U_{ww}^{(11)} & U_{fw}^{(11)} & ... & U_{ww}^{(n_j1)} & U_{fw}^{(n_j1)} \\ U_{wf}^{(11)} & U_{ff}^{(11)} & ... & U_{wf}^{(n_j1)} & U_{ff}^{(n_j1)} \\ ... & & & & \\ U_{ww}^{(1n_j)} & U_{wf}^{(1n_j)} & ... & U_{ww}^{(n_jn_j)} & U_{fw}^{(n_jn_j)} \\ U_{wf}^{(1n_j)} & U_{ff}^{(1n_j)} & ... & U_{wf}^{(n_jn_j)} & U_{ff}^{(n_jn_j)} \end{pmatrix}$$

which is the Hessian matrix of utility with respect to network resources.
Lemma 2. *At the optimum $x_j = \tilde{x}_j$, if $\nabla^2 U_j(x_j)$ is positive definite, then $\nabla_{\mu_j} x_j = [\nabla^2 U_j(x_j)]^{-1}$, $\forall j$.*

Proof. When the delay constraint is not binding, Theorem 1 states that:

$$\nabla U_j(x_j)|_{x_j=\tilde{x}_j} = \mu_j \quad \forall j$$

Writing out the chain rule for the derivative of this gradient with respect to μ_j gives:

$$[\nabla^2 U_j(x_j)|_{x_j=\tilde{x}_j}] \cdot [\nabla_{\mu_j} x_j|_{x_j=\tilde{x}_j}] = I \quad \forall j \tag{19}$$

□

Lemma 2 tells us that the gradient of the optimal resource allocations depend on the Hessian of the utility function with respect to bandwidth and buffer. The next theorem gives the explicit expressions for these sensitivities, which are directly derived from lemma 2.

The following mappings will be helpful:

$$U_{ww}^{(ll)} = \frac{\partial^2 U_j}{(\partial BW_{jl})^2} = U_{gg}^{(j)}(g_w^{(jl)})^2 + U_g^{(j)} g_{ww}^{(jl)}$$
$$U_{ff}^{(ll)} = \frac{\partial^2 U_j}{(\partial BF_{jl})^2} = U_{gg}^{(j)}(g_f^{(jl)})^2 + U_g^{(j)} g_{ff}^{(jl)}$$

$$
\begin{aligned}
U_{wf}^{(ll)} &= \frac{\partial^2 U_j}{\partial BW_{jl} \partial BF_{jl}} = U_{gg}^{(j)} g_w^{(jl)} g_f^{(jl)} + U_g^{(j)} g_{wf}^{(jl)} \\
U_{ww}^{(ll')} &= \frac{\partial^2 U_j}{\partial BW_{jl} \partial BW_{jl'}} = U_{gg}^{(j)} g_w^{(jl)} g_w^{(jl')} \\
U_{ff}^{(ll')} &= \frac{\partial^2 U_j}{\partial BF_{jl} \partial BF_{jl'}} = U_{gg}^{(j)} g_f^{(jl)} g_f^{(jl')} \\
U_{wf}^{(ll')} &= \frac{\partial^2 U_j}{\partial BW_{jl} \partial BF_{jl'}} = U_{gg}^{(j)} g_w^{(jl)} g_f^{(jl')}
\end{aligned}
$$

Theorem 7. *When delay constraint is not binding, and $\nabla^2 U_j(x_j)|_{x_j=\tilde{x}_j}$ is positive definite, we have:*

1) The following sensitivities hold for all links l:

$$
\begin{aligned}
\frac{\partial BW_{jl}}{\partial \beta_l} &= \frac{1}{H}\{U_{gg}^{(j)} g_{ff}^{(jl)} \sum_{k\neq l}(K^{(jk)} \prod_{i\neq l,k} h_g^{(ji)}) + U_{ff}^{(jl)} \prod_{i\neq l} h_g^{(ji)}\} \\
\frac{\partial BF_{jl}}{\partial \gamma_l} &= \frac{1}{H}\{U_{gg}^{(j)} g_{ww}^{(jl)} \sum_{k\neq l}(K^{(jk)} \prod_{i\neq l,k} h_g^{(ji)}) + U_{ww}^{(jl)} \prod_{i\neq l} h_g^{(ji)}\}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial BF_{jl}}{\partial \beta_l} = \frac{\partial BW_{jl}}{\partial \gamma_l} = \\
&-\frac{1}{H}\{U_{gg}^{(j)} g_{wf}^{(jl)} \sum_{k\neq l}(K^{(jk)} \prod_{i\neq l,k} h_g^{(ji)}) + U_{wf}^{(jl)} \prod_{i\neq l} h_g^{(ji)}\}
\end{aligned}
$$

where $h_g^{(ji)}$ is given by (3), $K^{(jk)}$ is given by (6),

$$
H = U_g^{(j)} U_{gg}^{(j)} \sum_{k=1}^{n_j}(K^{(jk)} \prod_{i\neq k} h_g^{(ji)}) + (U_g^{(j)})^2 \prod_{i=1}^{n_j} h_g^{(ji)}
$$

and where by convention the summations evaluate to 0 if they have no terms and the products evaluate to 1 if they have no terms.

2) The following sensitivities hold for all pairs of links l and $l' \neq l$:

$$\frac{\partial BW_{jl'}}{\partial \beta_l} = -\frac{U_{gg}^{(j)}}{H} k_1^{(jl')} k_1^{(jl)} \prod_{i \neq l, l'} h_g^{(ji)}$$

$$\frac{\partial BF_{jl'}}{\partial \gamma_l} = -\frac{U_{gg}^{(j)}}{H} k_2^{(jl')} k_2^{(jl)} \prod_{i \neq l, l'} h_g^{(ji)}$$

$$\frac{\partial BW_{jl'}}{\partial \gamma_l} = -\frac{U_{gg}^{(j)}}{H} k_1^{(jl')} k_2^{(jl)} \prod_{i \neq l, l'} h_g^{(ji)}$$

$$\frac{\partial BF_{jl'}}{\partial \beta_l} = -\frac{U_{gg}^{(j)}}{H} k_2^{(jl')} k_1^{(jl)} \prod_{i \neq l, l'} h_g^{(ji)}$$

where $k_1^{(jl)}$ and $k_2^{(jl)}$ are given by (4) and (5).

Proof. We only need to show that equation (19) is satisfied if all the sensitivities in the sensitivity matrix (18) are given by theorem 7. Then according to the uniqueness of the inverse matrix, we know the theorem is proved.

Consider the product of the sensitivity matrix and the Hessian of utility, and assume all the sensitivities in the sensitivity matrix are given by theorem 7. First let's inspect the diagonal elements of the product matrix, e.g. the s^{th} diagonal element, which is given by the product of row s of the sensitivity matrix and column s of the Hessian of utility. When s is an odd element:

$$\begin{aligned}
&\left(\tfrac{\partial BW_{j1}}{\partial \beta_s} \; \dots \; \tfrac{\partial BF_{jn_j}}{\partial \beta_s} \right) \left(U_{ww}^{(s1)} \; \dots \; U_{wf}^{(sn)} \right)^t \\
&= \frac{\partial BW_{js}}{\partial \beta_s} U_{ww}^{(ss)} + \frac{\partial BF_{js}}{\partial \beta_s} U_{wf}^{(ss)} + \sum_{m \neq s} \frac{\partial BW_{jm}}{\partial \beta_s} U_{ww}^{(sm)} \\
&\quad + \sum_{m \neq s} \frac{\partial BF_{jm}}{\partial \beta_s} U_{wf}^{(sm)}
\end{aligned} \tag{20}$$

Substitute the sensitivities given by theorem 7 and the second derivatives of utility with respect to bandwidth and buffer into equation (20). By further simplifying it, equation (20) yields a result of 1.

Similarly, if s is even:

$$
\begin{aligned}
&\left(\tfrac{\partial BW_{j1}}{\partial \gamma_s}\ \dots\ \tfrac{\partial BF_{jn_j}}{\partial \gamma_s}\right)\left(U_{wf}^{(s1)}\ \dots\ U_{ff}^{(sn)}\right)^t \\
=&\frac{\partial BW_{js}}{\partial \gamma_s}U_{wf}^{(ss)}+\frac{\partial BF_{js}}{\partial \gamma_s}U_{ff}^{(ss)}+\sum_{m\neq s}\frac{\partial BW_{jm}}{\partial \gamma_s}U_{wf}^{(sm)} \\
&+\sum_{m\neq s}\frac{\partial BF_{jm}}{\partial \gamma_s}U_{ff}^{(sm)} \\
=&1
\end{aligned}
$$

We have shown all the diagonal element of the product matrix is 1. Now let's look at the other elements of the product matrix, e.g. the $(s,\ r)$ element, where $s \neq r$. It is the product of row s of the sensitivity matrix and column r of the Hessian of utility. If s and r both are even:

$$
\begin{aligned}
&\left(\tfrac{\partial BW_{j1}}{\partial \beta_s}\ \dots\ \tfrac{\partial BF_{jn_j}}{\partial \beta_s}\right)\left(U_{ww}^{(r1)}\ \dots\ U_{wf}^{(rn)}\right)^t \\
=&\frac{\partial BW_{js}}{\partial \beta_s}U_{ww}^{(rs)}+\frac{\partial BF_{js}}{\partial \beta_s}U_{wf}^{(rs)}+\sum_{m\neq s}\frac{\partial BW_{jm}}{\partial \beta_s}U_{ww}^{(rm)} \\
&+\sum_{m\neq s}\frac{\partial BF_{jm}}{\partial \beta_s}U_{wf}^{(rm)} \\
=&0
\end{aligned}
$$

Similarly, we can show all the elements except for the diagonal elements of the product matrix are 0, which indicates the product matrix is an identity matrix. □

We are particularly interested in the signs of these sensitivities:

Corollary 2. *$\partial BW_{jl}/\partial\beta_l < 0$ and $\partial BF_{jl}/\partial\gamma_l < 0$ always holds, and $\partial BF_{jl}/\partial\beta_l = \partial BW_{jl}/\partial\gamma_l > 0$ iff*

$$U_{gg}^{(j)} g_{wf}^{(jl)} \sum_{k\neq l}(K^{(jk)} \prod_{i\neq l,k} h_g^{(ji)}) + U_{wf}^{(jl)} \prod_{i\neq l} h_g^{(ji)} < 0 \tag{21}$$

Proof. As mentioned above, $h_g^{(jl)} > 0\ \forall l$, $K^{(jl)} > 0\ \forall l$, $g_{ww}^{(jl)} > 0$, and $g_{ff}^{(jl)} > 0$, since $g_{jl}(BW_{jl}, BF_{jl})$ is jointly strictly convex in (BW_{jl}, BF_{jl}). Also, $U_g^{(j)} < 0$ and $U_{gg}^{(j)} < 0$, since U_j is decreasing and strictly concave in loss. It follows that $H > 0$. In addition, $U_{ww}^{(jl)} < 0$ and $U_{ff}^{(jl)} < 0$. Therefore, $\partial BW_{jl}/\partial\beta_l < 0$ and $\partial BF_{jl}/\partial\gamma_l < 0$. Finally, $\partial BF_{jl}/\partial\beta_l = \partial BW_{jl}/\partial\gamma_l > 0$, iff the condition given, is trivial when $H > 0$. □

We note that when $n_j = 1$, the condition (21) reduces to $\partial BF_{jl}/\partial\beta_l =$

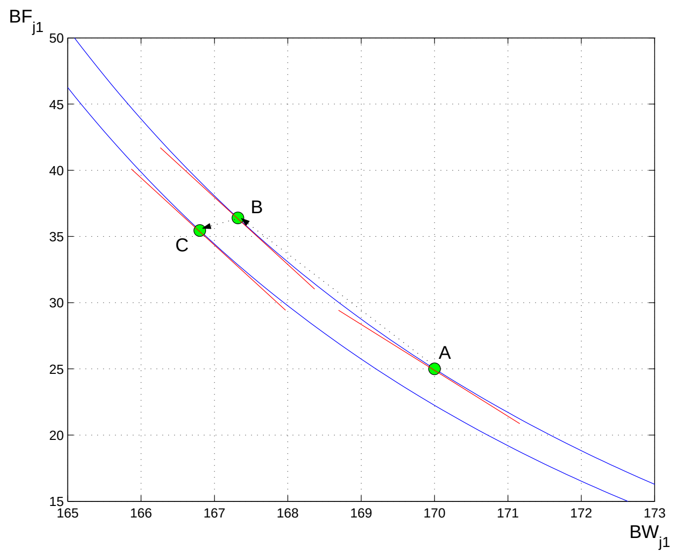

Fig. 5. Effect of price increase upon resources on same link

$\partial BW_{jl}/\partial\gamma_l > 0$ iff $U_{wf}^{(jl)} < 0$, namely iff U_j is submodular in buffer and bandwidth.

In economics terminology, $\partial BW_{jl}/\partial\beta_l < 0$ and $\partial BF_{jl}/\partial\gamma_l < 0$ means that neither buffer and bandwidth are *Giffen goods*. In addition, buffer and bandwidth on a single link are called *substitutes*, since the slope of the loss contour is negative. For two resources which are substitutes, when the price of one rises it causes two effects. The *substitution effect* causes a decrease in demand for the resource whose price rose and a corresponding increase in demand for the other resource, reflecting the new price ratio. The *income effect* causes a decrease in the demand for both resources, reflecting higher costs. Given that buffer and bandwidth are substitutes, $\partial BF_{jl}/\partial\beta_l = \partial BW_{jl}/\partial\gamma_l > 0$ when the substitution effect is greater than the income effect. Such a situation in pictured in figure 5. Again, we consider an aggregate of data flows traversing two links in the network. The original optimal allocation on link 1 is denoted as point **A**. After an increase in β_1 from 3.5 to 3.55, the optimal allocation on link 1 moves to point **C**. This change can be decomposed into two pieces. First, the substitution effect trades off bandwidth for buffer, moving from **A** to **B**. Second, the income effect causes an increase in loss on link 1, moving from **B** to **C**.

Corollary 3. *If $l' \neq l$, $\partial BW_{jl'}/\partial\beta_l > 0$ if $IBW^{(jl')}$ and $IBW^{(jl)}$ hold, $\partial BF_{jl'}/\partial\gamma_l > 0$ if $IBF^{(jl')}$ and $IBF^{(jl)}$ hold, $\partial BW_{jl'}/\partial\gamma_l > 0$ if $IBW^{(jl')}$ and $IBF^{(jl)}$ hold, and $\partial BF_{jl'}/\partial\beta_l > 0$ if $IBF^{(jl')}$ and $IBW^{(jl)}$ hold.*

Proof. Recall that $k_1^{(jl)} < 0$ iff $IBW^{(jl)}$ and that $k_2^{(jl)} < 0$ iff $IBF^{(jl)}$. Previously we proved that $h_g^{(jl)} > 0\ \forall l$, $U_{gg}^{(j)} < 0$, and $H > 0$. The corollary follows. $\square$

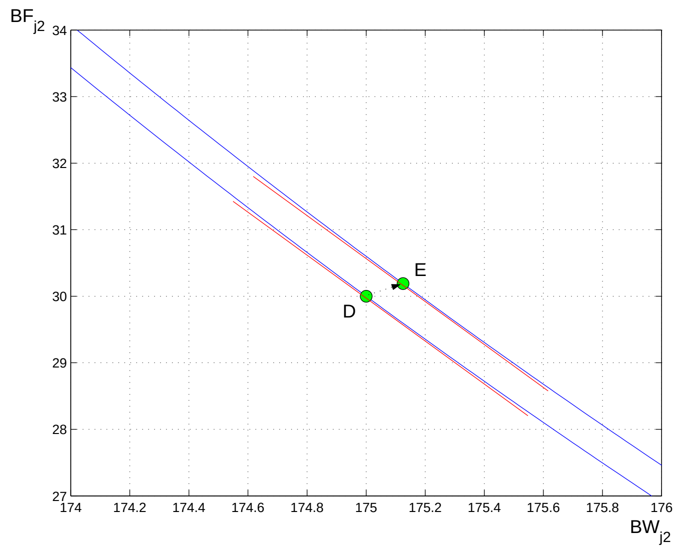

Fig. 6. Effect of price increase upon resources on other links

These results can be interpreted to mean that if resources on two different links are Normal goods, then an increase in the price of one causes in increase in demand for the other. Such a situation in pictured in figure 6. Continue with the previous example. The original optimal allocation on link 2 is denoted as point **D**. After an increase in β_1, the optimal allocation on link 2 moves to point **E**.

Finally, we can examine the sensitivity of loss on each link and total loss to changes in price:
Theorem 8. *The following sensitivity holds for all links l:*

$$\frac{\partial L_{jl}}{\partial \beta_l} = \frac{k_1^{(jl)}}{H}\{U_{gg}^{(j)} \sum_{k \neq l}(K^{(jk)} \prod_{i \neq k,l} h_g^{(ji)}) + U_g \prod_{i \neq l} h_g^{(ji)}\}$$
$$\frac{\partial L_{jl}}{\partial \gamma_l} = \frac{k_2^{(jl)}}{H}\{U_{gg}^{(j)} \sum_{k \neq l}(K^{(jk)} \prod_{i \neq k,l} h_g^{(ji)}) + U_g \prod_{i \neq l} h_g^{(ji)}\}$$

If $n_j \geq 2$, the following sensitivities hold for all pairs of links l and $l' \neq l$:

$$\frac{\partial L_{jl'}}{\partial \beta_l} = -\frac{U_{gg}^{(j)}}{H} K^{(jl')} k_1^{(jl)} \prod_{i \neq l} h_g^{(ji)}$$
$$\frac{\partial L_{jl'}}{\partial \gamma_l} = -\frac{U_{gg}^{(j)}}{H} K^{(jl')} k_2^{(jl)} \prod_{i \neq l} h_g^{(ji)}$$
$$\frac{\partial L_j}{\partial \beta_l} = \frac{U_g^{(j)}}{H} k_1^{(jl)} \prod_{i \neq l} h_g^{(ji)}$$
$$\frac{\partial L_j}{\partial \gamma_l} = \frac{U_g^{(j)}}{H} k_2^{(jl)} \prod_{i \neq l} h_g^{(ji)}$$

Proof. We can expand each of the sensitivities of loss as follows:

$$\frac{\partial L_{jl}}{\partial \beta_l} = g_w^{(jl)} \frac{\partial BW_{jl}}{\partial \beta_l} + g_f^{(jl)} \frac{\partial BF_{jl}}{\partial \beta_l}$$
$$\frac{\partial L_{jl}}{\partial \gamma_l} = g_w^{(jl)} \frac{\partial BW_{jl}}{\partial \gamma_l} + g_f^{(jl)} \frac{\partial BF_{jl}}{\partial \gamma_l}$$
$$\frac{\partial L_{jl'}}{\partial \beta_l} = g_w^{(jl')} \frac{\partial BW_{jl'}}{\partial \beta_l} + g_f^{(jl')} \frac{\partial BF_{jl'}}{\partial \beta_l}$$
$$\frac{\partial L_{jl'}}{\partial \gamma_l} = g_w^{(jl')} \frac{\partial BW_{jl'}}{\partial \gamma_l} + g_f^{(jl')} \frac{\partial BF_{jl'}}{\partial \gamma_l}$$
$$\frac{\partial L_j}{\partial \beta_l} = \sum_{k=1}^{n_j} \frac{\partial L_{jk}}{\partial \beta_l}, \quad \frac{\partial L_j}{\partial \gamma_l} = \sum_{k=1}^{n_j} \frac{\partial L_{jk}}{\partial \gamma_l}$$

Substituting the corresponding expressions from Theorems 7 gives the desired expressions. □

The signs of these sensitivities were considered in theorem 6.

5 Sensitivity to price with a binding delay constraint

In this section, we consider the sensitivity of loss and delay to changes in the price of bandwidth, in the case in which a delay constraint is binding for the traffic class. As with the previous case in which the delay constraint was not binding, we start by examining the set of minimum cost resource allocations under a range of loss, and the resulting minimum cost function. We then use the properties established to prove the form of the sensitivities.

When the delay constraint for class j is not binding, we previously found that buffer and bandwidth on link l are Normal goods iff $IBW^{(jl)}$ and $IBF^{(jl)}$

hold, respectively. When the delay constraint is binding, it is helpful to jointly consider loss and delay on each link, as outlined in section 2.2. With a fixed delay on link l, any decrease in the loss probability on link l can only be accomplished by simultaneously increasing buffer and bandwidth on link l (in a proportion determined by the delay on link l). The *expansion path* now follows the delay constraint (while it is binding), and correspondingly we denote the sensitivity of bandwidth and buffer on link l along the expansion path to loss on link l, as $dBW_{jl}/dL_{jl}|_{dc}$ and $dBF_{jl}/dL_{jl}|_{dc}$ respectively, and the corresponding slope of the expansion path as $dBW_{jl}/dBF_{jl}|_{dc}$.

Theorem 9.

$$\begin{aligned}
\frac{dBW_{jl}}{dL_{jl}}|_{dc} &= \frac{1}{g_w^{(jl)} + D_{jl}g_f^{(jl)}} < 0 \\
\frac{dBF_{jl}}{dL_{jl}}|_{dc} &= \frac{D_{jl}}{g_w^{(jl)} + D_{jl}g_f^{(jl)}} < 0 \\
\frac{dBF_{jl}}{dBW_{jl}}|_{dc} &= D_{jl} > 0
\end{aligned}$$

The proof is straightforward, and is omitted here. When the delay constraint is binding, define $C_{jl}(L_{jl}, D_{jl})$ as the minimum cost to achieve a loss of L_{jl} and a delay of D_{jl} on link l, and $C_j(L_j, D_j)$ as the minimum cost to achieve a total loss of L_j and a total delay of D_j. We say route j is convex and submodular in loss and delay if $\sum_{l\in R} C_{jl}(L_{jl}, D_{jl})$ is jointly convex and submodular in $(\sum_{l\in R} L_{jl}, \sum_{l\in R} D_{jl})$, where R is any subset of the links on route j.

When the delay constraint for class j is not binding, we previously found that if a resource is a Normal good, then an increase in the price of that resource causes the loss on that link to increase and the loss on all other links to decrease. When the delay constraint is binding and route j is convex and submodular in loss and delay, we find that an increase in the price of bandwidth causes *both* the loss and delay on that link to increase and both the loss and delay on all other links to decrease.

Theorem 10. *If route j is convex and submodular in loss and delay, then $\partial L_{jl}/\partial\beta_l > 0$, $\partial D_{jl}/\partial\beta_l > 0$, $\partial L_{jl'}/\partial\beta_l < 0$, and $\partial D_{jl'}/\partial\beta_l < 0$, $\forall l' \neq l$.*

Proof. The proof follows a similar procedure to that for theorem 6. Due to space restrictions, we only outline the differences here. We aggregate all links other than link l into a single virtual link denoted $\bar{l}$. The surplus for class j now becomes:

$$\begin{aligned}
f_1(L_{jl}, D_{jl}, -L_{j\bar{l}}, \beta_l) \equiv\, & U_j(L_{jl} + L_{j\bar{l}}) \\
& - C_{jl}(L_{jl}, D_{jl}) - C_{j\bar{l}}(L_{j\bar{l}}, D_{j\bar{l}})
\end{aligned}$$

Provided that route j is convex and submodular in loss and delay, it can be shown that the cross-derivatives for all six combinations of $\{L_{jl}, D_{jl}, -L_{j\bar{l}}, \beta_l\}$ are nonnegative. Following a similar approach as before, this can be used to prove that f_1 is supermodular in $\{L_{j1}, D_{jl}, -L_{j\bar{l}}, \beta_1\}$, and hence that f_1 is parameter monotonic in β_l. This establishes that $\partial L_{jl}/\partial\beta_l > 0$ and $\partial D_{jl}/\partial\beta_l > 0$.

To show that $\partial L_{jl'}/\partial\beta_l < 0$ and $\partial D_{jl'}/\partial\beta_l < 0$ $\forall l' \neq l$, we must still show that $\partial L_{j\bar{l}}/\partial\beta_l < 0$ and $\partial D_{j\bar{l}}/\partial\beta_l < 0$ implies that $\partial L_{jl'}/\partial\beta_l < 0$ and $\partial D_{jl'}/\partial\beta_l < 0$ $\forall l' \neq l$. Following a similar approach as in the proof for theorem 6, we aggregate all links in $\bar{l}$ except for link l' into a virtual link $\overline{l'}$. We consider the function:

$$\begin{aligned}
&f_2(L_{jl'}, D_{jl'}, L_{j\bar{l}}, D_{j\bar{l}}) \equiv \\
&-C_{jl'}(L_{jl'}, D_{jl'}) - C_{j\overline{l'}}(L_{j\bar{l}} - L_{jl'},\ D_{j\bar{l}} - D_{jl'})
\end{aligned}$$

It can be shown that f_2 is parameter monotonic in $(L_{jl'}, D_{jl'})$ with respect to $(L_{j\bar{l}}, D_{j\bar{l}})$. Combining this with $\partial L_{j\bar{l}}/\partial\beta_l < 0$ and $\partial D_{j\bar{l}}/\partial\beta_l < 0$ proves that $\partial L_{jl'}/\partial\beta_l < 0$ and $\partial D_{jl'}/\partial\beta_l < 0$ $\forall l' \neq l$. $\square$

Finally, if an increase in the price of bandwidth causes an increase in the loss and delay on that link, and a decrease in loss and delay on other links, then we can establish the corresponding changes in bandwidth usage on each link:
Corollary 4. *If route j is convex and submodular in loss and delay, then $\partial BW_{jl}/\partial\beta_l > 0$ and $\partial BW_{jl'}/\partial\beta_l < 0$ $\forall l' \neq l$.*

The proof uses theorems 9 and 10, and is omitted here. Changes in buffer usage on each link are indeterminate.

6 Conclusion

We have assumed the existence of a reservation-based QoS architecture that uses shadow-cost pricing. In particular, we considered a pricing policy which implements a distributed resource allocation to provide guaranteed bounds on packet loss and end-to-end delay for real-time applications. When the delay constraints are absent or not binding, we have given closed-form expressions for the sensitivity of buffer and bandwidth to changes in loss, and corresponding necessary and sufficient conditions for buffer and bandwidth to be Normal goods. We have shown that the cost required to achieve a target loss probability is decreasing and convex. We have given sufficient conditions to establish that an increase in the price for a resource results in an increase in the loss at that node, a decrease in the loss at all other nodes, and an increase in the total loss. We have also given closed-form expressions for all of the resource and QoS sensitivities, and given sufficient conditions to establish that an increase in the price for a resource results in a decreased demand for that resource, an increased demand for the other resource at that node, and an increased demand for resources at all other hops. Finally, when the delay constraint is binding, we have given sufficient conditions to establish that an increase in the price of bandwidth at one node results in increased loss and delay at that node, and decreased loss and delay at all other nodes.

Of course, much additional work would have to be done to make any such pricing approach practical, e.g. design of signalling protocols, reservation protocols, measurement algorithms, scheduling policies, user agents to automate the user response, and feedback algorithms to guarantee convergence. The sensitivity results derived here could be used to guide development of pricing feedback algorithms, and to adjust buffer and bandwidth in the optimal proportion in response to changes in desired levels of loss, as opposed to methods which may either increase only bandwidth (with fixed buffer) or may increase both buffer and bandwidth in a fixed linear proportion.

References

[1] J. K. MacKie-Mason, H. Varian, Pricing congestible resources, IEEE Journal of Selected Areas in Communications 13 (1995) 1141–49.

[2] R. Cocchi, D. Estrin, S. Shenker, L. Zhang, Pricing in computer networks: motivation, formulation, and example, IEEE/ACM Transactions on Networking 1 (6) (1993) 614–627.

[3] R. J. Gibbens, F. P. Kelly, Resource pricing and the evolution of congestion control, Automatica 35 (1999) 1969–1985.

[4] J. Murphy, L. Murphy, E. C. Posner, Distributed pricing for embedded ATM networks, in: International Teletraffic Congress, 1994.

[5] H. Jiang, S. Jordan, The role of price in the connection establishment process, European Transactions on Telecommunications 6 (4) (1995) 421–429.

[6] H. Jiang, S. Jordan, A pricing model for high speed networks with guaranteed quality of service, in: Proceedings of the Conference on Computer Communications (IEEE Infocom), San Francisco, CA, 1996, pp. 888–895.

[7] S. Low, P. Varaiya, Burstiness bounds for some burst reducing servers, in: Proceedings of the Conference on Computer Communications (IEEE Infocom), IEEE, 1993.

[8] S. Low, Equilibrium bandwidth and buffer allocations for elastic traffics, IEEE/ACM Transactions on Networking 8 (2000) 373–383.

[9] N. J. Keon, G. Anandalingam, Optimal pricing for multiple services in telecommunications networks offering quality-of-service guarantees, IEEE/ACM Transactions on Networking 11 (2003) 66–80.

[10] A. Lazar, A. Orda, D. Pendarakis, Virtual path bandwidth allocation in multiuser networks, IEEE/ACM Transactions on Networking 5 (6) (1997) 861–871.

[11] H. Yaiche, R. R. Maxumdar, C. Rosenberg, A game theoretic framework for bandwidth allocation and pricing in broadband networks, IEEE/ACM Transactions on Networking 8 (2000) 667–678.

[12] X. Cao, H. Shen, R. Milito, P. Wirth, Internet pricing with a game theoretical approach: Concepts and examples, IEEE/ACM Transactions on Networking 10(2) (2002) 208–216.

[13] R. J. Edell, N. McKeown, P. P. Varaiya, Billing users and pricing for TCP, IEEE Journal on Selected Areas in Communications 13 (7).

[14] J. Altmann, P. Varaiya, INDEX project: User support for buying QoS with regard to user's preferences, in: 1998 Sixth International Workshop on Quality of Service, 1998, pp. 101–104.

[15] S.-Q. Li, Study of information loss in packet voice systems, IEEE Transactions on Communications 37(1) (1989) 1192 –1202.

[16] D. Heyman, A. Tabatabai, T. Lakshman, Statistical analysis and simulation study of video teleconference traffic in ATM networks, IEEE Transactions on Circuits and Systems for Video Technology 2(1) (1992) 49–59.

[17] P. Skelly, M. Schwartz, S. Dixit, A histogram-based model for video traffic behavior in an ATM multiplexer, IEEE/ACM Transactions on Networking 1(4) (1993) 446–459.

[18] W. Leland, M. Taqqu, W. Willinger, D. Wilson, On the self-similar nature of ethernet traffic (extended version), IEEE/ACM Transactions on Networking 2(1) (1994) 1–15.

[19] J. Hui, Resource allocation for broadband networks, IEEE Journal on Selected Areas in Communications (1988) 1598–1608.

[20] F. Kelly, Effective bandwidths at multi-service queues, Queueing Systems 9 (1991) 5–16.

[21] G. Kesidis, J. Walrand, C.-S. Chang, Effective bandwidths for multiclass markov fluids and other atm sources, IEEE/ACM Transactions on Networking 1 (4) (1993) 424–428.

[22] M. Montgomery, G. DeVeciana, On the relevance of time scales in performance oriented traffic characterizations, in: Proceedings of Infocom, IEEE, 1996, pp. 513–520.

[23] A. W. Berger, W. Whitt, Effective bandwidths with priorities, IEEE/ACM Transactions on Networking 6 (4) (1998) 447–460.

[24] J. Morrison, Asymptotic analysis of a data-handling system with many sources, SIAM J. Appl. Math 49 (2) (1989) 617–637.

[25] S. Jamin, P. Danzig, S. Shenker, L. Zhang, A measurement-based admission control algorithm for integrated service packet networks, IEEE/ACM Transactions on Networking 5 (1) (1997) 56–70.

[26] M. Grossglauser, D. Tse, A framework for robust measurement-based admission control, IEEE/ACM Transactions on Networking 7 (3) (1999) 293–309.

[27] J. Qui, E. W. Knightly, Measurement-based admission control with aggregate traffic envelopes, IEEE/ACM Transactions on Networking 9 (2001) 199–210.

[28] C. A. Courcoubetis, A. Dimakis, G. D. Stamoulis, Traffic equivalence and substitution in a multiplexer with applications to dynamic available capacity estimation, IEEE/ACM Transactions on Networking 10(2) (2002) 217–231.

[29] L. Breslau, S. Shenker, Best-effort versus reservations: A simple comparative analysis, in: SIGCOMM Symposium on Communications Architectures and Protocols, 1998.

[30] J. Long-Tae Park; Jong-Wook Baek; Woon-Ki Hong, Management of service level agreements for multimedia internet service using a utility model, IEEE Communications Magazine 39 (2001) 100–106.

[31] A. Elwalid, D. Mitra, Effective bandwidth of general markovian traffic sources and admission control of high-speed networks, IEEE/ACM Transactions on Networking 1 (3) (1993) 329–344.

[32] A. Elwalid, D. Mitra, R. H. Wentworth, A new approach for allocating buffers and bandwidth to heterogeneous, regulated traffic in an atm node, IEEE Journal on Selected Areas in Communications 13 (6) (1995) 1115–1127.

[33] A. M. FP Kelly, D. Tan, Rate control for communication networks: Shadow prices, proportional fairness and stability, Journal of the Operational Research Society 49 (1998) 237–252.

[34] S. Low, D. Lapsley, Optimization flow control, i: Basic algorithm and convergence, IEEE/ACM Transactions on Networking (1999) 7(6):861–75.

[35] S. Jordan, Pricing of buffer and bandwidth in a reservation-based QoS architecture, in: Conference Record of the International Conference on Communications (ICC), IEEE, 2003.

[36] G. D. Veciana, C. Courcoubetis, J. Walrand, Decoupling bandwidths for networks: A decomposition approach to resource management, Proceedings of INFOCOM, IEEE (1994) 466 – 473 vol.2.

[37] S. Shenker, Fundamental design issues for the future internet, IEEE Journal on Selected Areas in Communications, Volume: 13 , Issue: 7 (1995) 1176 – 1188.

[38] N. Jin, G. Venkitachalam, S. Jordan, Dynamic congestion-based pricing of bandwidth and buffer, Tech. rep., Department of EECS, University of California, Irvine (2003).

[39] R. K. Sundaram, A First Course in Optimization Theory, Cambridge University Press, 1996.