QoS Extension to BGP *

Li Xiao‡ King-Shan Lui**† Jun Wang‡ Klara Nahrstedt‡

‡Dept. of Computer Science
University of Illinois at Urbana-Champaign
{lixiao, junwang3, klara}@cs.uiuc.edu

**Dept. of Electrical and Electronic Engineering
University of Hong Kong
kslui@eee.hku.hk

Abstract

To enable the end-to-end Quality of Service (QoS) guarantees in the Internet, based on the Border Gateway Protocol (BGP), inter-domain QoS advertising and routing are important. However, little research has been done in this area so far. Two major challenges, scalability and heterogeneity, make the QoS extension to BGP difficult. Two existing approaches, the Link Capacity Routing (LCR) and the Available Bandwidth Routing (ABR), address QoS advertising and routing in BGP with respect to bandwidth metric. But neither of them can solve the two challenges well.

In this paper, BGP is extended to advertise bandwidth information. But, instead of using link capacities or instantaneous available bandwidth values, a novel QoS metric, Available Bandwidth Index (ABI), is defined and used to perform bandwidth advertising and routing. The two major contributions of ABI are: (1) ABI dynamically abstracts available bandwidth into a probability interval, therefore, it is very flexible to represent heterogenous and dynamic bandwidth values; (2) By capturing the statistical property of the detailed available bandwidth distribution, ABI is so efficient that it can highly decrease the message overhead in routing, thereby making the QoS advertising and routing very scalable. Our extensive simulations confirm both contributions of the ABI extension to BGP very well.

1 Introduction

Quality-of-Service (QoS) routing is essential for providing end-to-end QoS guarantees. The Internet routing is divided into two levels hierarchically, the intra-domain routing and the inter-domain routing. Routing protocols have to be QoS-aware in both levels in order to provide end-to-end QoS support. There are many solutions for intra-domain QoS routing protocols, such as [11]. However, little work has been done so far to put QoS information into the context of inter-domain routing. In this paper, based on the border gateway protocol (BGP)[16], we will discuss the mechanisms and extensions to enable QoS inter-domain routing.

The Internet consists of Autonomous Systems (AS). Each AS is an independently managed network unit. Interior Gateway Protocol (IGP) is used inside an AS, such as OSPF. BGP is the de facto standard of inter-domain routing. Essentially, BGP is a distance vector protocol for hop-by-hop routing. The basic function is to exchange network reachability information between autonomous systems. The network reachability information, which is formatted in the UPDATE message, can advertise or withdraw a path to a network destination. The UPDATE message, also called advertisement, mainly contains the address of the network destination, the path represented in AS numbers (AS_PATH), and the next hop address (NEXT_HOP). Each AS calculates the degree of preference for each path it has received according to some path selection policies, installs the most preferred one into the local forwarding table, and propagates such routing decision to neighboring ASes.

In the existing BGP path selection process, many policies are involved, such as the commercial relationships between ASes, the number of hops in terms of AS, the multiple exit discriminator, etc.. In [1], the path selection algorithm from Cisco Systems is given. Gao[9] presents the path selection policies for achieving 'inherently safe backup routing'. However, because BGP routers can only infer limited QoS information from the advertisement they receive, the inter-domain routing decisions consider almost nothing about the real end-to-end QoS metrics, such as delay and bandwidth. Usually the route with the least AS hops is preferred. Therefore, the routing results could be quite different from the optimal paths in the sense of QoS, such as the path with the largest bottleneck bandwidth or the path with the minimum delay. As a result, it is necessary to take QoS metrics into

*This work was supported by DARPA under contract number F30602-97-2-0121, and NSF under contract number NSF ANI 00-73802 and NSF EIA 99-72884. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DARPA or NSF.

†Much of the work was done while the author was at UIUC.

1092-1628/02 $17.00 © 2002 IEEE

consideration in the BGP path selection process.

There are mainly three advantages in bringing QoS information into BGP. First, it will optimize the inter-domain packet forwarding performance. The paths with higher available bandwidth or lower traffic load can be identified and installed into the forwarding table by using the QoS information which is advertised in BGP messages. Second, it will make inter-domain traffic engineering [7] more effective. Local IP traffic can be better controlled and tuned if the global traffic condition is known. Third, it can provide services for other inter-domain related protocols which need QoS support from the routing layer. For example, in the inter-domain resource reservation protocol BGRP[15], the block rate will be decreased if the signal messages are distributed according to some QoS information.

Two major difficulties exist when QoS is introduced into BGP. (1) The extension has to be scalable. BGP is originally designed to exchange pure reachability information. If the QoS metrics are added, the scalability of Internet routing should not be compromised by the dynamic nature of QoS information. (2) The QoS representation should be able to handle the heterogeneity in the inter-domain routing. The connections between BGP routers may be of different types. For example, some connections may use direct physical links, while some may use the paths provided by the intra-domain routing, i.e. IGP routes. Moreover, the route updating periods may be different in different ASes. Therefore, the QoS information obtained from different ASes has different degrees of precision.

In order to cope with the two difficulties above, QoS metrics have to be appropriately selected. As we know, there exist two types of QoS metrics: the static QoS metrics and dynamic ones. The static metric is constant all the time, such as the link capacity and AS hop count. The dynamic metric varies according to different traffic load, such as the available bandwidth of a link or a path.

Routing using static metrics has low message overhead. After the routing table is set up, route will not be changed for QoS reason, because the value of the static QoS metric is fixed. However, static QoS metrics usually can not reflect the instantaneous network status. For example, even if the link capacity is high, the real available bandwidth could be low due to high traffic load. On the other hand, dynamic QoS metrics can represent the instantaneous network status. However, high routing message overhead is incurred due to the fluctuation of dynamic QoS metrics over time. Routing based on the dynamic QoS metrics without any control is not scalable in the global Internet. Some simple statistics based on the dynamic values, such as average values, can reduce the message overhead, but they are too coarse-grained to model the instantaneous information well.

As the main contribution of this paper, a novel QoS metric is proposed for inter-domain QoS routing to make QoS extension scalable and achieve satisfactory routing optimality[1]. Based on the histogram information of the available bandwidth, we define Available Bandwidth Index (ABI) to model the instantaneous available bandwidth. Basically, ABI is a compound metric which consists of an interval $\varpi = [l, u]$ and a probability ρ, meaning the real available bandwidth belongs to the interval ϖ with probability ρ.

The instantaneous value of the available bandwidth changes from time to time. However, in the Internet backbone, since a large number of flows is aggregated on each link, the statistical distribution of available bandwidth is far more stable than the instantaneous value. Thus, by using ABI in BGP advertising and routing, the routing message overhead can be reduced to a level which is close to the cost of routing using static QoS metrics. This approach makes ABI routing scalable to large networks. Although the instantaneous bandwidth information is not included in ABI, using simulations, we show that the optimality of ABI routing is much higher than static metric routing by taking the advantage of statistical bandwidth information.

ABI is also flexible to cope with the heterogeneity in the inter-domain routing, which neither the static nor dynamic metric could achieve. ABI can represent the bandwidth properties of either a direct physical link or an IGP route. Furthermore, different precision levels of QoS information can be represented by ABI. For example, a more imprecise QoS parameter may have a larger interval ϖ or a smaller probability ρ. In later sections, we will show that ABI can also represent the bandwidth on a path which contains some legacy routers that do not support BGP QoS extension.

The rest of the paper is organized as follows. In Section 2, the network model and the ABI are defined. In Section 3, we present BGP QoS extension based on ABI. Section 4 shows the simulation results. Section 5 describes the related work and Section 6 concludes the paper.

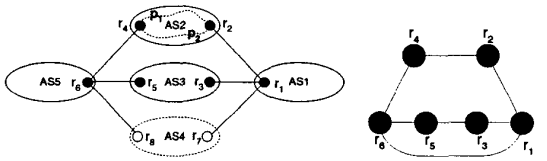
2 Network Model and ABI

2.1 Network Model

We consider a typical network with BGP routers and ASes, where BGP routers can be either QoS-aware or without any QoS extension, as shown in Fig. 1(a). In Fig. 1(a), the BGP routers in AS1, AS2, AS3 and AS5 are QoS-aware, and routers in AS4 are not. (We call those BGP routers without QoS extension the legacy BGP routers.) Our network model, representing the BGP routers and ASes, is then defined as a graph $G = (V, E)$, where V is the set of *QoS-aware* BGP routers and E is the set of logical links that connect QoS-aware BGP routers. Fig. 1(b) shows an example which is abstracted directly from the network in Fig.

[1] Routing optimality means the ability to find the path with the best QoS. We will give a rigorous definition for routing optimality in section 4.

1(a). With respect to different abstraction origins in the real network, there are three different types of logical links in E: (1) **TYPE-1:** A TYPE-1 logical link in E represents a real physical link which connects two BGP routers directly. Typically, this type of links exists between two neighboring ASes (e.g., the link between r_4 and r_6 in Fig. 1(b)). (2) **TYPE-2:** A TYPE-2 logical link stands for an IGP route inside an AS, connecting two BGP routers within the same AS (e.g., the link between r_2 and r_4 in Fig. 1(b)). (3) **TYPE-3:** A TYPE-3 logical link encapsulates a physical route across multiple ASes, along which all the intermediate routers are legacy BGP routers. For example, the link between r_1 and r_6 shown in Fig. 1(b) is a TYPE-3 logical link. This type of links corresponds to the scenario, where QoS-aware BGP routers are only incrementally or partially deployed.



(a) Network of BGP routers. (b) Network Model for BGP QoS extension.

Figure 1. Network Models.

In the network model, each link $e \in E$ could be associated with some QoS parameters. In this paper, we concentrate only on bandwidth metric. Two kinds of bandwidth information can be used for QoS routing: link capacity and instantaneous available bandwidth. The link capacity, focusing on the static aspect, describes the maximum data transferring rate of a link. The instantaneous available bandwidth, on the other hand, is a dynamic parameter and represents the instantaneous data transferring rate on a link at a certain time. The capacity of a link, being the upper bound of the available bandwidth on that link, is usually much larger than the available bandwidth due to existing traffic or bandwidth reservation.

Now, assuming the bandwidth information is available for each logical link in E, our focus is to bring QoS extensions to the original BGP so that inter-domain paths are optimized in the sense of bandwidth. To be more specific, in addition to the reachability information, a new QoS-aware BGP router should be capable of: (1) calculating and then advertising the bandwidth property of a path, and (2) using the bandwidth as one of the path selection metrics. The details of getting available bandwidth information for a logical link in E is thereafter discussed as follows.

In order to obtain available bandwidth for different types of links in E (TYPE-1, 2, or 3), three different ways are applied accordingly. If $e \in E$ is a TYPE-1 link, its available bandwidth can be simply obtained by monitoring its traffic directly. If e is of TYPE-2, then we can get the available bandwidth information of e from the IGP running in that AS (we assume the IGP to be QoS-enabled, such as the OSPF with QoS extensions [11]). If e is a TYPE-3 link, since e actually represents a route that consists of legacy BGP routers, we have to initiate an end-to-end bandwidth measuring process to obtain the available bandwidth information of e. Notice that: (1) For a TYPE-1 link, changes of its available bandwidth are caused by traffic fluctuations on the physical link. On the other hand, for a TYPE-2 or TYPE-3 link, since it may represent an entire path rather than a single physical link in the real network, its available bandwidth changes not only because of the traffic fluctuations on the path, but also due to route changes. For example, in Fig. 1(a) the IGP routing in AS2 may change from path $p1$ to path $p2$. This will result in the change of link metrics between r_2 and r_4 in Fig. 1(b), if the QoS properties of $p1$ and $p2$ are different. (2) Measurements of the end-to-end bandwidth are used to obtain the available bandwidth for TYPE-3 links. However, we do not rely on this technique to obtain bandwidth information for all links. The reasons are: (a) end-to-end measurements are very imprecise, and (b) large communication and computing overheads are involved.

2.2 The new bandwidth-related metric: ABI

In this section, we introduce a novel QoS metric – *Available Bandwidth Index* (ABI), which is scalable and can handle heterogeneity while providing good routing optimality.

2.2.1 ABI Definition

In the Internet, the instantaneous value of the available bandwidth varies over time. Different links may have very different precision and fluctuation patterns. Due to such high dynamics and heterogeneity, a single value is no longer sufficient to capture link bandwidth property. Therefore, in our new bandwidth-related metric, ABI, we bring in statistical properties of the instantaneous available bandwidth. Let us assume that the available bandwidth on each link is a random variable that follows a certain distribution. The instantaneous values (samples) fall into an interval $\varpi = [l, u]$ with probability ρ. The interval ϖ and its corresponding probability ρ can be used as a new compound statistic for these instantaneous values. Following this idea, we define the ABI metric as follows.

Definition 1 (Available Bandwidth Index (ABI)) *The Available Bandwidth Index $\hat{b}$ is defined as $\hat{b} = \{b_m, \delta, \rho\}$, meaning that the probability for the instantaneous available bandwidth b belonging to the interval $\varpi = [b_m - \delta, b_m + \delta]$ is no less than ρ, i.e., $Pr\,[b \in \varpi = [b_m - \delta, b_m + \delta]] \geq \rho$.*

In the definition of ABI, the b_m is related to the average value. δ and ρ are related to the dynamic scope of the instantaneous value. There are several advantages of using ABI as the routing metric.

First, ABI can represent the fine-grained statistical property of the available bandwidth efficiently. For the distribution of the available bandwidth, the most detailed representation is the probability density function or whole histogram information. However, using that information incurs too much processing overhead. In the other extreme, link capacity and some simple statistics of the available bandwidth can be used at low cost, such as the average value. But they are too coarse-grained to represent the detailed distribution of dynamic information. While, with ABI, the major statistical property of the instantaneous bandwidth values can be captured with acceptable processing overhead.

Second, ABI makes BGP QoS extension scalable. The available bandwidth of a link may vary frequently over time, but its statistical properties change much less frequently. If the available bandwidth is directly used as the routing metric in BGP, a large number of route update messages could flood over the whole network, thus the routing message overhead is unacceptable. On the contrary, since the ABI reflects the major statistical properties of the available bandwidth, it is far more stable than the instantaneous value. If we look for better paths in terms of ABI[2], most of the instantaneous bandwidth changes are filtered out to avoid unnecessary route updates. Therefore, adopting ABI as the routing metric makes the BGP with QoS extension more scalable.

Third, ABI can accommodate the link heterogeneity. Different link types may have very different bandwidth distribution. But ABI is flexible enough to handle such heterogeneity. For example, if the bandwidth information of a TYPE-3 logical link is imprecise due to some legacy BGP routers in the middle, it can still be represented by using ABI with a large interval length δ or small probability ρ.

2.2.2 ABI calculation

ABI representation is flexible. We do not need to know the analytical distribution function, and we also do not assume it follows certain well-known distribution.

For TYPE-1 links and TYPE-2 links, the calculation of ABI is based on a list of sample values of the available bandwidth in the history. Assume that n samples, $\overrightarrow{b} = \{b(t_1), b(t_2), \ldots, b(t_{n-1}), b(t_n)\}$, can be kept for each link, which represent the available bandwidth samples at time $t_1, t_2, \ldots, t_n$ respectively. The values of the samples can be obtained from direct physical link monitoring or IGP QoS routing. The samples are updated as new bandwidth information is available, and the old records are overwritten.

[2]We will address the comparison of ABIs later in Section 3.2.

we want to find a certain b_m and a certain δ for a corresponding ρ. Based on the bandwidth vector, the ABI with confidence interval $1-\alpha$ is calculated as follows: b_m equals to the median element in $\overrightarrow{b}$. In order to find δ, we need to find k such that there are k elements out of n samples of the instantaneous available bandwidth that fall into $[b_m-\delta, b_m+\delta]$. k is constrained by α, ρ, and n to guarantee that the instantaneous bandwidth belongs to $[b_m-\delta, b_m+\delta]$ with $1-\alpha$ confidence interval. Therefore, we can compute k for given α, ρ and n, and then calculate δ.

Intuitively, it is necessary that $k \geq n\rho$. If we consider the confidence interval $1-\alpha$ which reflects the accuracy of ABI calculation, we have the following theorem for an arbitrary bandwidth distribution. Let us assume z_α is the value of the standard normal curve above which we can find an area of α.

Theorem 1 *Given the available bandwidth vector $\overrightarrow{b}$, the number of samples n, probability ρ, and the confidence interval $1-\alpha$, if*

$$k = \frac{nz_\alpha^2 + 2n^2\rho + nz_\alpha\sqrt{4n\rho - 4n\rho^2 + z_\alpha^2}}{2\,(n + z_\alpha^2)} = g(n, \rho, z_\alpha) \tag{1}$$

and interval $\varpi = [b_m - \delta, b_m + \delta]$ contains k elements of $\overrightarrow{b}$, then the probability, that the instantaneous bandwidth belongs to the interval ϖ, is no less than ρ with confidence interval $1-\alpha$.

Proof Sketch: Let us define $p = Pr\,[b_m - \delta \leq b \leq b_m + \delta]$. From the proportion estimation theory[13], for any bandwidth distribution, we have $Pr\left[p \geq \frac{k}{n} - z_\alpha\sqrt{\frac{k/n(1-k/n)}{n}}\right] \simeq 1-\alpha$. Because it is required that $Pr\,[p \geq \rho] = 1-\alpha$, we get the requirement on k: $\frac{k}{n} - z_\alpha\sqrt{\frac{k/n(1-k/n)}{n}} = \rho$. By solving this equation, we get equation 1. (Please refer to [17] for the detailed proof.) ■

Discussions: (1) ρ is a tunable parameter for each link, and its value can be chosen according to the specific link properties. In order to capture the major portion of the samples, usually ρ should be close to 1, such as 90%. (2) α is set to be a small value, such as 0.05, to get a good confidence interval. (3) n should be a large number to make the ABI calculation more precise. A rule often used is $n\rho \geq 5$ and $n(1-\rho) \geq 5$ [13]. Since ρ is close to 1, the number of bandwidth samples n is required to be larger than $\frac{5}{1-\rho}$. For example, if $\rho = 90\%$, $n \geq 50$.

Based on the assumption that n is a large number, ρ is close to 1, and z_α is usually in $[0, 2]$ (because α is a small number), $g(n, \rho, z_\alpha)$ in Equation 1 can be simplified as

$$g(n, \rho, z_\alpha) \simeq n\rho + \frac{z_\alpha}{2} + z_\alpha\sqrt{n\rho(1-\rho)} \tag{2}$$

For the logical TYPE-3 links, we assume the end-to-end bandwidth measurement techniques, such as [12], can provide the approximate bandwidth range and the precision rate

of the measurement. We use the range and precision rate as the interval ϖ and probability ρ of ABI.

2.2.3 ABI Join Operations

A path is formed when the links are connected together in a sequence. Because bandwidth is a concave metric, the available bandwidth of a path is the minimum available bandwidth of all the links on that path. To obtain ABI, we can find the available bandwidth on that path first, and then calculate the ABI according to the definition. However, this method is not practical in BGP protocol. Instead, we calculate the ABI of a path by joining the ABIs of individual links or sub-paths directly.

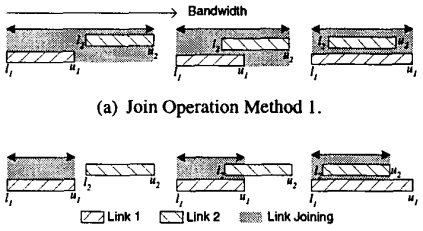
Given two ABIs $\hat{b}_1$ and $\hat{b}_2$, we define the ABI join operation as $\hat{b} = \hat{b}_1 \oplus \hat{b}_2$. Thus, the ABI of path $v_1 v_2 \ldots v_n$ is $\hat{b}_{v_1 v_2} \oplus \ldots \oplus \hat{b}_{v_{n-1} v_n}$, where $\hat{b}_{v_i v_j}$ is the ABI of link $v_i v_j$.

We make two assumptions in the join operation of ABI. First, ABIs of different links are independent. Similar assumption is made by Lorenz and Guerin in [14][10]. Two facts support this assumption. First, a large number of flows are aggregated on each link. Second, the intra-domain traffic is the major portion of the total network traffic. Thus, the correlation between two different links can be ignored. Second, the bandwidth distribution outside $\varpi = [b_m - \delta, b_m + \delta]$ is approximately symmetric around ϖ, i.e. $Pr[b < b_m - \delta] \simeq Pr[b > b_m + \delta] \simeq (1-\rho)/2$. This assumption holds well for the links of TYPE 1 and TYPE 2, because $1-\rho$ is very close to 0 according to the calculation of ABI, and the bandwidth distribution outside ϖ has small value. For TYPE 3 links, ρ could also be close to 1, if δ is large enough. If ρ is small due to the imprecision in bandwidth measurement, the symmetric assumption may not hold well. We will discuss this special case in the later part of this section.

In order to compute the interval ϖ and probability ρ for $\hat{b} = \hat{b}_1 \oplus \hat{b}_2$, we can set the $\hat{b}.\varpi$ by the combination of $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$, and compute $\hat{b}.\rho$ based on $\hat{b}.\varpi$. For convenience, let l_1 and u_1 denote the lower bound and upper bound of interval $\hat{b}_1.\varpi$, i.e., $l_1 = \hat{b}_1.b_m - \hat{b}_1.\delta$, $u_1 = \hat{b}_1.b_m + \hat{b}_1.\delta$. l_2 and u_2 are the lower bound and upper bound of interval $\hat{b}_2.\varpi$. Let b_1, b_2, and b denote the instantaneous available bandwidth related to $\hat{b}_1, \hat{b}_2$ and $\hat{b}$ respectively. $\rho_1 = \hat{b}_1.\rho$, and $\rho_2 = \hat{b}_2.\rho$.

We give two join operation methods for computing $\hat{b}_1 \oplus \hat{b}_2$, which are used in different situations. They are demonstrated in Fig. 2. The small rectangle boxes with cross-line patterns represent $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$, which are marked with l_1, u_1, l_2 and u_2, respectively. The gray area $\hat{b}.\varpi$ stands for the interval of the resulting ABI. For each join operation method, there are three cases shown in the figure based on the value of u_1.

Join Operation Method 1: Given that $\hat{b_1}$ and $\hat{b_2}$ are two ABIs for link 1 and link 2, and $\hat{b} = \hat{b}_1 \oplus \hat{b}_2$, then $\hat{b}.\varpi = [\min(l_1, l_2), \max(u_1, u_2)]$, i.e., $\hat{b}.b_m = [\min(l_1, l_2) + \max(u_1, u_2)]/2$, and $\hat{b}.\delta = [\max(u_1, u_2) - \min(l_1, l_2)]/2$.



(a) Join Operation Method 1.

(b) Join Operation Method 2.

Figure 2. ABI Join Operation Methods.

The join method 1, which is shown in Fig. 2(a), is a straightforward method to compute the interval of the joined link. $\hat{b}.\varpi$ covers the whole interval delimited by both $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$. The following theorem gives the value of $\hat{b}.\rho$.

Theorem 2 *Under the condition of join operation method 1,* $\hat{b}.\rho = (\hat{b}_1.\rho + \hat{b}_2.\rho)/2$

Proof: Denote b_1, b_2 and b as instantaneous available bandwidth on link 1, link 2, and the joined links, respectively.
(1) $u_1 \leq u_2$:
$Pr[b \in \varpi]$
$= Pr[l_1 \leq b_1 \leq u_2]Pr[b_2 \geq l_1] + Pr[b_1 > u_2]Pr[l_1 \leq b_2 \leq u_2]$
$\geq \rho_1 Pr[b_2 \geq l_1] + \rho_2(Pr[u_1 < b_1 \leq u_2] + Pr[b_1 > u_2])$
$\geq \rho_1(1+\rho_2)/2 + \rho_2(1-\rho_1)/2 = (\rho_1 + \rho_2)/2$
(2) $u_1 \geq u_2$:
$Pr[b \in \varpi]$
$= Pr[l_1 \leq b_1 \leq u_1]Pr[b_2 \geq l_1] + Pr[b_1 > u_1]Pr[l_1 \leq b_2 \leq u_1]$
$\geq \rho_1 Pr[b_2 \geq l_2] + Pr[b_1 > u_1]\rho_2$
$= \rho_1(1+\rho_2)/2 + \rho_2(1-\rho_1)/2 = (\rho_1 + \rho_2)/2$
From these two cases, we proved that $(\rho_1 + \rho_2)/2$ is the lower bound of probability of b belonging to ϖ. Therefore $\hat{b}.\rho = (\hat{b}_1.\rho + \hat{b}_2.\rho)/2$. ■

One of the advantages of join operation method 1 is that ρ of the joined ABI $\hat{b}$ is never less than ρ_1 and ρ_2 simultaneously. Thus, probability for the instantaneous value belonging to ϖ will not decrease substantially, if more links are joined to the path. However, the length of the interval ϖ may become larger and larger. It is acceptable if $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$ overlap or close to each other, but it is not appropriate when those two intervals are disjoint and separated with large distance. In the second join operation method, we shrink the length of the interval ϖ.

Join Operation Method 2: Given that $\hat{b_1}$, $\hat{b_2}$ are two ABIs for link 1 and link 2, and $\hat{b} = \hat{b}_1 \oplus \hat{b}_2$, then $\hat{b}.\varpi = [\min(l_1, l_2), \min(u_1, u_2)]$, i.e., $\hat{b}.b_m = [\min(l_1, l_2) + \min(u_1, u_2)]/2$, and $\hat{b}.\delta = [\min(u_1, u_2) - \min(l_1, l_2)]/2$.

Fig. 2(b) shows the join operation method 2. The following theorem gives the value of $\hat{b}.\rho$.

Theorem 3 *Under the condition of join operation method 2, if* $u_1 < u_2$, $\hat{b}.\rho = \rho_1(1+\rho_2)/2$; *otherwise,* $\hat{b}.\rho = \rho_2(1+\rho_1)/2$.

Proof: Similar to Theorem 2. Please refer to [17]. ■

In the join operation method 2, the length of $\hat{b}.\varpi$ is never larger than the lengths of $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$ simultaneously. When more links are joined to the path, the length of the resulting bandwidth interval will not increase substantially. However, the probability ρ of the resulting ABI is smaller than ρ_1 and ρ_2.

These two join operation methods are used by BGP routers to calculate the ABI of a route. Notice that: (1) ABI join operation methods 1 and 2 can be used alternatively depending on the relationship of $\hat{b}_1$ and $\hat{b}_2$. In general, if $\hat{b}_1.\rho$ or $\hat{b}_2.\rho$ is small, method 1 is preferred; if $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$ are disjointed and are separated with a large distance, method 2 is preferred. (2) These two join operations are both based on the symmetric distribution assumption. If this assumption does not hold well, b falls into the interval defined by join operation method 2 with probability $\hat{b}_1.\rho \times \hat{b}_2.\rho$. Therefore, we can use method 2 to calculate $\hat{b}.\varpi$, and let $\hat{b}.\rho$=$\hat{b}_1.\rho \times \hat{b}_2.\rho$. In most cases, especially when the link type is TYPE 1 or TYPE 2, where ρ is usually close to 1, these two join operations give satisfying precision in calculating the ABI of the joined links.

3 Protocol Extensions of BGP

In order to enable the inter-domain QoS routing, we make three modifications to BGP: (1) extend BGP UPDATE messages to record QoS information; (2) select paths based on the QoS information stored in the extended BGP UPDATE messages; (3) monitor and update the QoS state of the advertised routes.

3.1 BGP UPDATE Message Extension

QoS information has to be recorded in the UPDATE message, which represents the ability of a domain to provide the route with such QoS. In [6] a new attribute QoS_NLRI is proposed for this purpose. Similar attempt can be taken here. We require that bandwidth information can be put into the Path Attribute field, which represents QoS status of the advertised route. Accordingly, the BGP routing table is extended to keep the QoS information, as well. Extended BGP routers will use the ABI calculation methods and joining operations in Section 2.2 to obtain the ABI for links and paths.

In order to cope with legacy BGP routers, the QoS attribute is optional and transitive, which means QoS attribute may not be recognized by some legacy BGP routers and this attribute should be passed on even if it is not recognized.

The QoS-aware BGP router needs to know whether or not a BGP message is directly from a QoS-aware router, and where the last QoS-aware router is if not. For this purpose, in an UPDATE message, a new optional and transitive attribute is created to record the IP address of the last QoS-aware BGP router. Each QoS-aware BGP router records its IP address in this attribute when the UPDATE message passes by. Thus, QoS-aware routers get to know if a TYPE-3 link is needed.

3.2 QoS Path Selection

In the BGP path selection process, QoS-based path selection policy has to be involved. Because there are multiple policies effecting the path ranking, the priority of QoS metrics can be determined flexibly by the local network administration. In general, it can be put below the policies that specify the peer relationship between ASes defined in [9], so that routing protocol always converges. Moreover, because the QoS advertising in this paper is used to optimize end-to-end performance, its priority can be lower than IGP distance metric which is used to optimize the traffic inside a domain. To identify a path with better QoS, we need to compare the ABIs of the paths. We now present the method of comparing two ABIs.

3.2.1 Normalization of ϖ

The value of ρ influences the length of ϖ in the ABI definition. For example, a large probability ρ may lead to a large interval ϖ. Thus, if two ABIs have different $\rho's$, they can not be compared directly. Unfortunately, we can not always have the same ρ for any ABIs. There are two reasons: (1) ρ, as a tunable parameter, may be chosen differently on different links; (2) ρ of a path is the joining result of all the links on the path. Therefore, ϖ of different ABIs have to be normalized to remove the impact of ρ, so that two ABIs are comparable.

To solve this problem, we can scale the δ based on the value of ρ. Intuitively, the larger the ρ, the larger the δ, and vise versa. Because the primary objective of the normalization method is to provide a way to make two different ABIs comparable, it is not necessary to find the analytical relation between δ and ρ for any distribution. For simplicity of analysis, we use normal distribution as an approximation to find the relation between ρ and δ, and use the result for a general case. In Section 4.3, our simulation results demonstrate this approximation works well for other distributions.

Let us assume b is the instantaneous value of bandwidth, $b - b_m$ follows normal distribution $\mathcal{N}(0, \sigma^2)$, where σ is the standard variance. $F(x)$ is the distribution function of x.

$$\begin{aligned}\rho &= F(\delta) - F(-\delta) = 2\int_0^{\delta} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}\,\mathrm{d}x \\ &= \frac{2}{\sqrt{2\pi}\sigma}\int_0^{\delta} 1 - \frac{1}{2}\left(\frac{x}{\sigma}\right)^2 + O\left(\frac{x}{2\sigma}\right)^4 \mathrm{d}x\end{aligned}$$

Because $F(2\sigma) - F(-2\sigma) \simeq 95\%$, we can assume that $\delta < 2\sigma$. $\rho \simeq \frac{2}{\sqrt{2\pi}\sigma}\delta\left[1 - \left(\frac{\delta}{\sqrt{6}\sigma}\right)^2\right] \simeq c\frac{\delta}{\sigma}$ where $c = 2/\sqrt{2\pi}$ is a constant. Because approximately $\rho \propto \delta$, we can remove the effect of different ρ by normalizing δ with ρ. Denote the normalized result to be δ', and $\delta' = \frac{\delta}{\rho}$. Then the normalized $\varpi' = [b_m - \delta', b_m + \delta']$. Note that the normalization is only for ABI comparison in this paper. The original ABI is exchanged between routers for more accurate calculation.

3.2.2 Weight for Path Selection

With respect to QoS path selection, the optimization target is to find a route which has the largest instantaneous available bandwidth. Base on this target, if we use capacity or available bandwidth as the metric, the path weight $\mathcal{W}$ can be defined as the value of capacity or available bandwidth, respectively. The larger the weight is, the better the path will be.

For ABI, the quality of a path is determined by the interval ϖ and probability ρ jointly. Based on the normalized ϖ', the path with larger b_m, smaller δ' is more preferable, because statistically this path has larger available bandwidth over a long time. Thus, we can define the weight as $\mathcal{W} = b_m - \delta'$ (other weight definitions are also possible, but we only discuss this one in the paper.)

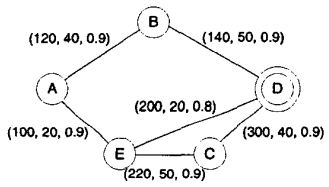


Figure 3. An Example: BGP QoS Extension.

3.2.3 An Example

Figure 3 shows a simple example of BGP QoS routing using ABI. Assume all the links are bidirectional and the numbers beside each link are the ABI parameters (b_m, δ, ρ). The nodes represent QoS-aware BGP routers. We assume that each node is in an independent AS. For simplicity, we only consider one destination, D. At first, D sends advertisements to B, C and E, respectively. Upon receiving the advertisement, C installs the path CD into its routing table with ABI $\hat{b}_{CD} = (300, 40, 0.9)$ and passes an advertisement to E. When E receives both advertisements from D and C, it first joins the ABI of link EC and path CD to get the ABI for the path ECD as $\hat{b}_{ECD} = (220, 50, 0.9) \oplus (300, 40, 0.9) = ((340 + 170)/2, (340 - 170)/2, (0.9 + 0.9)/2) = (255, 85, 0.9)$ (ABI join operation method 1 is used here). In this example, if the ϖ of two ABIs is disjointed, join method 2 is used; otherwise join method 1 is used. E then compares the weight of path ED and ECD. Since $\mathcal{W}_{ED} = 200 - 20/0.8 = 175$ and $\mathcal{W}_{ECD} = 255 - 85/0.9 = 160.6$, E selects path ED and passes this information to A and C via advertisements. After the routing process becomes stable, the routing table at each node is shown in Table 1. An 'Active Route' value ('Yes' or 'No') in the table indicates whether the corresponding route is being used or not. All routes marked with 'No' are candidate routes[3]. 'AS_PATH' is the full path from the source to node D. 'Next_Hop' is the next hop, in terms of node number, of the path. '(b_m, δ, ρ) / $\mathcal{W}$' are the ABI and the weight of the path. Recall that we have two ABI join methods (Section 2.2.3).

Source	Active Route	AS_PATH	Next_Hop	(b_m, δ, ρ) / $\mathcal{W}$
A	Yes	(E D)	E	(100, 20, 0.81) / 75.3
	No	(B D)	B	(135, 55, 0.9) / 73.8
B	Yes	(D)	D	(140, 50, 0.9) / 84.4
	No	(A E D)	A	(120, 40, 0.86) / 73.5
C	Yes	(D)	D	(300, 40, 0.9) / 255.6
	No	(E D)	E	(220, 50, 0.85) / 161.2
E	Yes	(D)	D	(200, 20, 0.8) / 175.0
	No	(C D)	C	(255, 85, 0.9) / 160.6

Table 1. The content of BGP Routing Tables at Each Node for Destination D.

Source	Active Route	AS_PATH	Next_Hop	(b_m, δ, ρ) / $\mathcal{W}$
A	Yes	(E C D)	E	(100, 20, 0.86) / 76.7
	No	(B D)	B	(135, 55, 0.9) / 73.8
B	Yes	(D)	D	(140, 50, 0.9) / 84.4
	No	(A E C D)	A	(120, 40, 0.88) / 74.5
C	Yes	(D)	D	(300, 40, 0.9) / 255.6
E	Yes	(C D)	C	(255, 85, 0.9) / 160.6
	No	(D)	D	(100, 20, 0.9) / 77.8

Table 2. After ABI of link ED changes from $(200, 20, 0.8)$ to $(100, 20, 0.9)$, the content of BGP Routing Tables.

3.3 QoS Information Update

In the conventional BGP, the path selection process is triggered by a BGP router whenever it detects a new route or a change (removal or update) of an existing route. If the selected path is different from what is currently used, the forwarding table will be updated with the new path, and

[3] An active route is the route installed in the forwarding table of a router. Candidate routes are all routes received by a router, which can potentially be used as an active route.

UPDATE messages will be sent to the neighboring BGP routers. In the QoS-aware BGP, route updates may also be caused by the changes of QoS status. In order to process such QoS-related changes, both the path selection process and the UPDATE message handling process in the original BGP should be slightly modified, while the BGP state machine model remains the same as [16]. There are two cases in which a QoS-aware BGP router may detect the change of QoS information. We will handle them separately as follows.

In the first case, the bandwidth information on a logical link has changed. We design a new process, called *linkChangeHandler*, to handle such change. *linkChangeHandler* will check all the entries in the BGP routing table which use this link as the next hop. If necessary, the QoS information of the route is updated and the path selection process is triggered. For example, in Fig. 3, if the available bandwidth on link ED changes from previous $(200, 20, 0.8)$ to $(100, 20, 0.9)$, router E will recalculate the weight for path ED as $\mathcal{W}_{ED} = 100 - 20/0.9 = 77.8$. Because $\mathcal{W}_{ED}$ is smaller than $\mathcal{W}_{ECD} = 160.6$, which is a candidate route, router E will change its route to D by replacing the route ED with ECD. E will also send UPDATE messages to A and C to withdraw previous route ED and advertise the new route ECD with its ABI $(255, 85, 0.9)$.

In the second case, an UPDATE message is received, which contains the route change or QoS change information. In the above example, after C receives UPDATE messages from E, C will simply withdraw its candidate route CED and keep its active path CD unchanged. When A receives UPDATE messages from E, it will withdraw path AED, and calculate the ABI and weight for the new route $AECD$: $\hat{b}_{AECD} = (100, 20, 0.9) \oplus (255, 85, 0.9) = (100, 20, 0.86)$, and $\mathcal{W}_{AECD} = 100 - 20/0.86 = 76.7$. Because $\mathcal{W}_{AECD} > \mathcal{W}_{ABD}$, A will choose route $AECD$ and further send UPDATE messages to B accordingly. After the routing is stabilized, the routing table of each router is shown in Table 2.

The example above shows that additional routing message overhead is incurred due to the QoS extension to BGP. In order to keep the QoS extension scalable, the rate of QoS-related route changes should be strictly controlled. In addition to the use of ABI instead of available bandwidth, setting up update thresholds is also an effective way to keep the routing message overhead low. Two types of thresholds, in terms of the path weight, are used: (1) *Link State Threshold ($\mathcal{T}_l$):* The small bandwidth fluctuation at the logical link should not trigger the *linkChangeHandler*. Only when the change of the weight[4] is greater than $\mathcal{T}_l$, will the *linkChangeHandler* process be called. (2) *Route Update Threshold ($\mathcal{T}_r$):* In the path selection process, only when the weight of the newly selected path is greater than the previously installed path by $\mathcal{T}_r$, will the new path be installed as the substitution of the previous path.

[4]The weight of the bandwidth of a link is defined the same as the weight of the bandwidth of a path in Section 3.2.2.

4 Simulation Results

In order to evaluate the performance of QoS extension to BGP, extensive simulations have been conducted. Three BGP-based routing protocols, link capacity routing (LCR), available bandwidth routing (ABR), and ABI routing (ABIR) are compared. We demonstrate that ABIR can find much better routes than LCR can and has much lower message overhead than ABR.

4.1 Simulation Model

The purpose of the simulation is to compare the optimality and message overhead of these three QoS extensions, based on the BGP routing protocol in [16], some simplifications are made: (1) Each AS is simplified as a single node; (2) We ignore the address aggregation; (3) We consider bandwidth information as the only path selection metric and ignore other BGP routing policies.

A BGP protocol simulator is implemented based on the simplified inter-domain routing model. In LCR, the capacity information is advertised; in ABR, the instantaneous available bandwidth is advertised; in ABIR, available bandwidth index is advertised. We use the advertised information to find the path which has the largest available bandwidth, i.e. the target of routing is to find the widest path in terms of the available bandwidth.

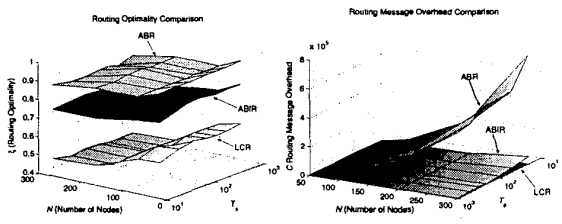
Internet topology generator BRITE [2] is used to generate flat AS level topologies for simulation. The Waxman model is used and nodes are placed according to the heavy-tail distribution. Denote the number of nodes in network as N. Four topologies are used in the simulation, with N equals 50, 100, 200, and 300, respectively.

The dynamic behavior of the available bandwidth is modeled with three different distributions: normal, uniform, and pareto. A random variable, e.g. normal random variable $\mathcal{N}(\mu, \sigma)$, is assigned to each link for generating the instantaneous values of the available bandwidth. In each time unit, a new bandwidth value is generated following this distribution, i.e. the available bandwidth is sampled for routing purpose on each link. In every T_s units of time, the parameters of the distributions, such as μ and σ in normal distribution, are changed randomly. Note: (1) T_s is an average value for all the links. Different links may have different periods and may change asynchronously. (2) T_s is the ratio between the change rates of the available bandwidth and its statistical distribution, and T_s is usually a large number. We assume $T_s \geq 20$ in simulations.

Two metrics are defined below to quantify the performance of routing protocols. (1) *Routing Optimality* ξ: Denote $\beta(\mathcal{R})$ as the average available bandwidth between all pairs of nodes based on the result of a routing protocol $\mathcal{R}$. The routing optimality of $\mathcal{R}$ is defined as $\xi = \frac{\beta(\mathcal{R})}{\max \beta}$, where $\max \beta$ can be obtained by running Dijkstra's algorithm on the network graph with the instantaneous available bandwidth as the link weight. (2) *Routing Message Overhead* C: C is the total number of BGP UPDATE messages exchanged in the network per time unit, which shows the cost and convergence speed of a routing protocol. Because the routing table could be set up by BGP or by static installation, we only consider the messages which are caused by the bandwidth information change.

4.2 Performance Comparisons

The performance of the three protocols, LCR, ABR and ABIR, is compared in Fig.4. The routing optimality and routing message overhead are shown in Fig.4(a) and Fig.4(b) with respect to different network topologies and values of T_s. Normal distribution is used to model the link available bandwidth.



(a) Routing optimality comparison. (b) Message overhead comparision.

Figure 4. Performance Comparisions.

In term of finding the path with the maximum available bandwidth, ABR protocol has the best performance among the three. If the thresholds ($\mathcal{T}_l$ and $\mathcal{T}_r$) in ABR are zero and assume the routing protocol converges fast enough in one time unit, ABR can achieve 100% optimality. The ABR curves, shown in Fig.4, have non-zero thresholds: $\mathcal{T}_l = 20$ and $\mathcal{T}_r = 80$. Its optimality ξ is about 85%. However, message overhead of ABR is very large and it increases substantially as the network size increases. Therefore, ABR is not a practical protocol.

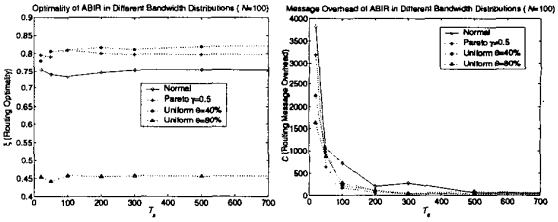
On the contrary, LCR only selects path by the static QoS metric – link capacity. Thus, there is no route change due QoS in LCR after the network is set up, i.e. $C = 0$. However, because LCR does not adapt to the real available bandwidth, its optimality ξ is only about 50%.

ABIR makes a good compromise between the routing message overhead and the routing optimality. Its routing optimality ξ is about 75%. Its routing message overhead is far less than the ABR protocol. In the worst case of our simulations, where $T_s = 20$ time units and the number of node is 300, the routing message used in ABIR is only 6.8% of those in ABR. When the T_s is larger, ABIR has even less message overhead. The advantage of ABIR comes from the routing based on the statistical properties of the available bandwidth instead of using instantaneous values. ABIR achieves higher routing optimality than LCR with much lower routing message overhead than ABR.

Furthermore, more simulation results in [17] demonstrate that ABR performs worse than ABIR if they are controlled to have similar message overhead by increasing $\mathcal{T}_l$ and $\mathcal{T}_r$ in ABR.

4.3 ABIR in different traffic distributions

In section 3.2.1, we use normal distribution to derive an ABI normalization method as an approximation for any general distribution. From the simulation results below, it is also demonstrated that this approximate method works well for other distributions. Two bandwidth distributions are tested: pareto and uniform. D is the link capacity. For pareto distribution $F(x) = 1 - (k/x)^a$, k is a random number in $[0.1D, 0.9D]$, and γ is the average value of the shape parameters a on all links. The uniform distribution is set to the interval $[s, s+d]$, where $d = \theta D$ and s is a random value in $[0, D-d]$. θ stands for the range of the bandwidth in the uniform distribution.



(a) Optimality comparison in different bandwidth distributions. (b) Message overhead comparison in different bandwidth distributions.

Figure 5. The performance of ABIR in different bandwidth distributions.

The simulation results are shown in Fig. 5. If the available bandwidth follows pareto or uniform($\theta = 0.4$) distribution, ABIR has similar optimality and message overhead than in the case of normal distribution. However, in the uniform distribution of $\theta = 0.8$, the optimality of ABIR is almost the same as LCR (shown in Fig.4(a)). This can be explained as follows. If the available bandwidth is distributed uniformly in interval $[0, D]$, the ABI tends to have $\varpi = [0, D]$, given ρ is close to 1. Thus, the bandwidth distribution will not change, and routing by ABI has no dif-

ference from routing by capacity only. Therefore, ABIR performs better than LCR if the distribution of the available bandwidth has a mode, i.e. the value of available bandwidth most likely occurs in an interval whose length is smaller than D. If $\theta = 0.4$, as shown in the simulation, the optimality of ABIR is about 80%, much higher than the optimality of LCR.

5 Related Work

Several related works have been done on inter-domain QoS routing. Bonaventure[4] focuses on how to distribute QoS information flexibly by BGP in different network scenarioes. Cristallo and Jacquenet[6] propose a new attribute for BGP UPDATE message, QOS_NLRI, to record QoS related information. Abarbanel and Venkatachalam[3] utilize BGP to propagate Traffic Engineering Weight, which represent the summary of the traffic condition in an AS. These three Internet drafts use either static QoS metrics or simple statistics of dynamic metrics, such as the average value or minimum value. Therefore, they can not advertise fine-grained properties of dynamic QoS information. They also can not address the heterogeneity problems introduced by IGP routing and incremental QoS deployment. Fei and Gerla[8] extend MBGP (Multiprotocol Extension to BGP4) for inter-domain QoS Multicast. However, the authors do not give an effective method to control the overhead of exchanging QoS update.

With respect to using statistical property in QoS routing, some related research work exists. Lorenz and Guerin proposed QoS routing algorithms based on the probability density function in [14][10]. However, obtaining and processing such density function would bring too much computation and communication overhead. Actually, in practice, it is not realistic to assume the distribution function is known. Chen and Nahrstedt[5] model the imprecise QoS value by an interval which is calculated from exponential average. Being different from ABI, their interval is a deterministic bound. It can be viewed as a special case of ABI model, where $\rho = 1.0$.

6 Conclusion

In this paper, we investigate a very challenging problem in the area of inter-domain routing – extension of the existing BGP to support QoS. Two challenges, scalability and heterogeneity, make this problem very difficult to solve. We propose a novel compound bandwidth metric, the Available Bandwidth Index (ABI), to perform QoS advertisement and route selection in BGP, which can accommodate the heterogenous bandwidth values and provide satisfiable performance with low message overhead.

The basic idea behind ABI can be extended to other types of metrics. In the future, we will study additive QoS metrics, such as delay or cost, in the context of BGP QoS routing by using similar ideas.

References

[1] BGP best path selection algorithm. In *Cisco Systems Inc. http://www.cisco.com/warp/public/459/25.shtml.*

[2] Boston university representative internet topology generator, http://cs-www.bu.edu/brite/.

[3] B. Abarbanel and S. Venkatachalam. *BGP-4 Support for Traffic Engineering*. Internet Draft draft-abarbanel-idr-bgp4-te-00.txt. Work in progress, September 2000.

[4] O. Bonaventure. *Using BGP to distribute flexible QoS information*. Internet Draft draft-bonaventure-bgp-qos-00.txt. Work in progress, February 2001.

[5] S. Chen and K. Nahrstedt. Distributed QoS routing with imprecise state information. In *Proceedings of IEEE ICCCN*, pages 614–621, October 1998.

[6] G. Cristallo and C. Jacquenet. *Providing Quality of Service Indication by the BGP-4 Protocol: the QOS_NLRI attribute*. Internet Draft draft-jacquenet-qos-nlri-03.txt. Work in progress, March 2002.

[7] N. Feamster, J. Borkenhagen, and J. Rexford. Controlling the impact of BGP policy changes on IP traffic. Tech report, AT&T Research Labs, Nov 2001.

[8] A. Fei and M. Gerla. Extending BGMP for shared-tree inter-domain QoS multicast. In *Proceedings of IWQoS*, 2001.

[9] L. Gao, T. Griffin, and J. Rexford. Inherently safe backup routing with BGP. In *Proceedings of INFOCOM*, pages 547–556, April 2001.

[10] R. Guerin and A. Orda. QoS-based routing in networks with inaccurate information: Theory and algorithms. In *Proceedings of INFOCOM*, 1997.

[11] R. A. Guerin, A. Orda, and D. Williams. QoS routing mechanisms and OSPF extensions. In *Proceedings of IEEE Globecom*, pages 1903–1908, 1997.

[12] M. Jain and C. Dovrolis. End-to-end available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput. In *Proceedings of ACM SIGCOMM*, 2002.

[13] R. Larsen and M. Marx. *An Introduction to Mathematical Statistics and Its Applications*. Prentice Hall, 2001.

[14] D. H. Lorenz and A. Orda. QoS routing in networks with uncertain parameters. *IEEE/ACM Transactions on Networking*, 6(6):768–778, December 1998.

[15] P. P. Pan, E. L. Hahne, and H. G. Schulzrinne. BGRP: A tree-based aggregation protocol for inter-domain reservations. *Journal of Communications and Networks*, 2(2):157–167, June 2000.

[16] Y. Rekhter and T. Li. A border gateway protocol 4 (BGP-4) rfc 1771, March 1995.

[17] L. Xiao, K.-S. Lui, J. Wang, and K. Nahrstedt. QoS extension to BGP. Tech report, Department Computer Science, University of Illinois, UIUCDCS-R-2002-2295, 2002.