Advertising Interdomain QoS Routing Information

Li Xiao, *Student Member, IEEE*, Jun Wang, *Student Member, IEEE*, King-Shan Lui, *Member, IEEE*, and Klara Nahrstedt, *Member, IEEE*

***Abstract*—To enable end-to-end quality-of-service (QoS) guarantees in the Internet, based on the border gateway protocol (BGP), interdomain QoS information 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. In the existing routing schemes, static and instantaneous QoS metrics, such as link capacity and available bandwidth, are used to represent QoS routing information, but neither of them can solve the two challenges well.**

In this paper, BGP is extended to advertise available bandwidth and delay information of routes, but, instead of using the traditional deterministic metrics, a series of statistical metrics, *available bandwidth index* (ABI), *delay index* (DI), *available bandwidth histogram* (ABH), and *delay histogram* (DH), are defined and applied to QoS information advertising and routing. Two major contributions of the proposed statistical metrics are: 1) QoS information is abstracted into one or several probability intervals and, thus, the heterogeneous and dynamic QoS information can be represented more flexibly and precisely and 2) by capturing the statistical property of the detailed distribution of QoS information, these new metrics are efficient and they can highly decrease the message overhead in routing, thereby making the QoS advertising and routing scalable. Our extensive simulations confirm both contributions of the QoS extension to BGP very well. Moreover, besides BGP, these statistical metrics can be applied to other networks and protocols to represent QoS information in a more scalable and precise way.

***Index Terms*—Border gateway protocol (BGP), network routing, quality-of-service (QoS).**

I. 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 intradomain routing and the interdomain 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 intradomain QoS routing protocols, such as OSPF QoS extension [1]. However, little work has been done so far to put QoS information into the context of interdomain routing. In this paper, based on the de facto interdomain routing standard, the border gateway protocol (BGP) [2], we will discuss the mechanisms and extensions to enable QoS information advertising and routing in the interdomain level.

The Internet consists of autonomous systems (AS). Interior gateway protocol (IGP) is used inside an AS, such as OSPF. BGP is essentially a hop-by-hop distance vector routing protocol for exchanging network reachability information between ASes. The network reachability information, which is formatted in the `UPDATE` messages, can advertise or withdraw a route to a network destination. The `UPDATE` messages, also called advertisements, mainly contain the addresses of the network destinations, the paths represented in AS numbers (`AS_PATH`), and the next-hop addresses (`NEXT_HOP`). Each AS calculates the degree of preference for each route it has received according to some path selection policies, installs the most preferred one into the local forwarding table, and propagates such routing decisions to neighboring ASes.

BGP is a policy-based routing protocol. Some BGP path selection rules from Cisco Systems are presented in [3], such as the number of hops in terms of ASes, the multiple exit discriminator, etc. In general, the relationship between ASes plays an important role in BGP route selection, route exporting and route importing policies. These policies reflect the business relation between different domains [4], and some of them are also essential to prevent BGP routing divergence [5], [6]. For example, the routes from the customers are preferable to the routes from the peers and the providers.

Recently, in order to increase network reliability, multihoming is becoming a more and more practiced technique in the Internet community. This solution provides us with multiple options to reach a network destination, even after the above routing policies are applied. Thus, it is possible to improve the routing performance by taking into account the QoS information in the path selection process. However, because BGP routers can only infer limited QoS information from the advertisement they receive, the current interdomain routing decisions consider almost nothing about the real end-to-end QoS metrics, such as delay and bandwidth. As a result, suboptimal routes may be selected in terms of QoS. Therefore, it would be beneficial to extend BGP for advertising interdomain QoS routing information.

There are mainly three advantages in bringing QoS information into BGP. First, it will optimize the interdomain packet forwarding performance. By properly using the QoS routing information in BGP messages, we can identify routes with higher available bandwidth or lower traffic load to forward data packets. Second, it will make interdomain traffic engineering [7] more effective. Local Internet protocol (IP) traffic can be better controlled if the global Internet traffic condition is known. Third, it can provide necessary information for other

Manuscript received September 3, 2003; revised April 14, 2004. This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under Contract F30602-97-2-0121, and in part by the National Science Foundation (NSF) under Contract NSF ANI 00-73802 and Contract 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.

L. Xiao, J. Wang, and K. Nahrstedt are with the Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana, IL 61801 USA (e-mail: lixiao@cs.uiuc.edu; junwang3@cs.uiuc.edu; klara@cs.uiuc.edu).

K.-S. Lui is with Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong, China (e-mail: kslui@eee.hku.hk).

Digital Object Identifier 10.1109/JSAC.2004.836004

interdomain related protocols which need QoS support from the routing layer. For example, in the interdomain resource reservation protocol BGRP [8], the block rate will be decreased if the signal messages are distributed according to appropriate QoS metrics.

However, there are two major difficulties when QoS information advertising is introduced into BGP. First, the extension has to be scalable. BGP is originally designed to exchange pure reachability information. If QoS metrics are added, the scalability of Internet routing should not be compromised by the dynamic nature of the QoS information. Second, the QoS representation should be able to handle the heterogeneity of links or routes in the interdomain 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 intradomain routing, i.e., IGP routes. Moreover, the route refreshing periods may vary in different domains. Thus, the QoS information obtained from different ASes has different degrees of precision.

In order to cope with the two difficulties described above, QoS metrics have to be appropriately selected. As we know, there exist two types of QoS metrics: the static QoS metrics and the dynamic ones. The static metrics are deterministic all the time, such as the link capacity and AS hop count. The dynamic metrics vary according to different traffic loads, 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, QoS information of routes will not be further exchanged, because the values of the static QoS metrics are constant. 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, but high routing message overhead is incurred due to the fluctuation of dynamic QoS metrics over time. Routing based on the instantaneous QoS metrics without any control is not scalable in the global Internet. Some simple statistics based on the instantaneous values, such as average available bandwidth (AAB), can reduce the message overhead, but they are too coarse-grained to model the instantaneous information well.

The main contribution of this paper is the proposal of four novel statistical QoS metrics, which make the QoS extension of interdomain routing scalable and achieve satisfactory routing optimality.[1] Based on the samples of available bandwidth and delay, we define available bandwidth index (ABI) and delay index (DI) to model the instantaneous values of the available bandwidth and delay. Basically, ABI or DI is a composite metric which consists of an interval $\varpi = [l, u]$ and a probability ρ, meaning that the instantaneous value belongs to the interval ϖ with probability ρ. In order to increase the precision of the QoS information being advertised, we further extend the concept of ABI and DI to available bandwidth histogram (ABH) and delay histogram (DH) by making use of multiple probability intervals.

The instantaneous values of the available bandwidth and delay fluctuate from time to time. However, in the Internet backbone, since a large number of flows are aggregated on each link, the statistical distributions of bandwidth and delay are far more stable than the instantaneous values. Thus, by using the statistical QoS metrics 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. Thus, this approach makes QoS information advertising scalable to large networks. On the other hand, although the instantaneous information is not advertised, using simulations, we show that the statistical metrics lead to much higher routing optimality than the static metrics.

Our new metrics are also flexible to cope with the heterogeneity in the interdomain routing, which neither the static nor the dynamic metric could achieve. The variation patterns of QoS information of a direct physical link are different from those of a group of IGP routes. Different updating periods of IGP also result in different precision levels of QoS information. Our statistical metrics can handle the heterogeneity properly. For example, a less precise QoS parameter may have a larger interval ϖ or a smaller probability ρ. This is helpful to represent the QoS routing information of a path which contains some legacy routers that do not support the BGP QoS extension.

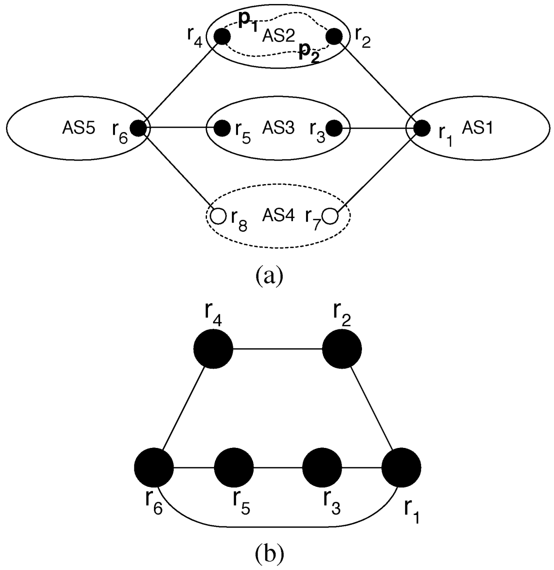
Furthermore, our statistical QoS representations are not just limited to the BGP application. They can also be applied to other networks and protocols, in which the resource information is dynamic, and the scalability and precision of the metrics are concerned. The rest of the paper is organized as follows. In Section II, the network model is defined. In Section III, we present two new metrics, ABI and DI, and their join operations. In Section IV, we present BGP QoS extension based on ABI and DI. In Section V, we extend ABI and DI to histogram information, ABH and DH, respectively. Section VI shows the simulation results. Section VII describes the related work and Section VIII concludes the paper.

II. 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, while 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(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)].

[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 VI.



(a)

(b)

Fig. 1. Network model. (a) Network of BGP routers. (b) Network model for BGP QoS extension.

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.

Each link $e \in E$ can be associated with some QoS parameters. In this paper, we concentrate on two types of QoS parameters: bandwidth and delay. In traditional QoS routing, link capacity and instantaneous available bandwidth are used to characterize the bandwidth information of links or routes. The instantaneous available bandwidth is a dynamic parameter and represents the instantaneous available data transferring rate on a link at a certain time. On the other hand, the link capacity, focusing on the static aspect, describes the maximum data transferring rate of a link, which is usually much larger than the available bandwidth due to the existing traffic or bandwidth reservations. The delay metric, which we address in this paper, is the aggregated information of processing delay, propagation delay and queuing delay. Likewise, the instantaneous delay changes frequently over time.

Now, assuming the bandwidth or delay information is available for each logical link in E, our focus is to bring QoS extensions to the original BGP so that interdomain routing, traffic engineering and other related protocols can be optimized based on appropriate QoS parameters.

In order to obtain available bandwidth for different types of links in E, three different ways are applied accordingly (the delay can be obtained similarly). 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 [1]). 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 the following.

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, the variation of the available bandwidth may be caused by traffic fluctuations and also by 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) The technique of end-to-end bandwidth measuring is used to obtain the available bandwidth for TYPE-3 links. However, we do not rely on this technique to obtain bandwidth information for TYPE-1 and TYPE-2 links. It is because end-to-end measurements may be very imprecise. Moreover, there are large communication and computing overheads involved.

III. New QoS Metrics: ABI and DI

In order to characterize the instantaneous bandwidth and delay information, we introduce two novel QoS metrics—ABI and DI, which are scalable and can handle heterogeneity, while providing good routing optimality.

A. *Definitions of ABI and DI*

In our new QoS metrics, ABI and DI, we bring in statistical properties of the dynamic QoS parameters. Let us assume that the available bandwidth on a link or a route 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 composite statistic for these instantaneous values. Following this idea, we define the ABI metric as follows.

Definition 1 (ABI): The ABI $\hat{b}$ is defined as $\hat{b} = \{b_l, b_u, \rho\}$, meaning that the probability for the instantaneous available bandwidth b belonging to the interval $\varpi = [b_l, b_u]$ is no less than ρ, i.e., $Pr[b \in \varpi = [b_l, b_u]] \geq \rho$.

Similarly, we have the definition of DI.

Definition 2 (DI): The DI $\hat{d}$ is defined as $\hat{d} = \{d_l, d_u, \rho\}$, meaning that the probability for the instantaneous delay d belonging to the interval $\varpi = [d_l, d_u]$ is no less than ρ, i.e., $Pr[d \in \varpi = [d_l, d_u]] \geq \rho$.

In the definitions of ABI and DI, ϖ represents the dynamic range of the instantaneous values. ρ is related to the statistical coverage of ϖ and the precision of the measurement. There are several advantages of using ABI and DI as routing metrics.

First, ABI and DI can represent the fine-grained statistical property of the available bandwidth and delay efficiently. With ABI and DI, the major statistical property of the instantaneous values can be captured with acceptable processing overhead. ABI and DI avoid the processing overhead that is incurred by using probability density functions, which theoretically can model QoS parameters distributions completely. On the other hand, ABI and DI are much more fine-grained than the static QoS parameters and some simple statistics, such as link capacity and AAB.

Second, ABI and DI make BGP QoS extension scalable. The instantaneous bandwidth or delay of a link may vary frequently over time, but its statistical distribution changes much less frequently. If the instantaneous values are directly used as the routing metric in BGP, a large number of route update messages could flood over the whole network, and the routing message overhead is unacceptable. On the contrary, since the ABI and DI reflect the major statistical properties of QoS, it is far more stable than the instantaneous values. For instance, if we look for better routes in terms of ABI,[2] most of the instantaneously changes are filtered out to avoid unnecessary route updates. Therefore, employing ABI and DI as routing metrics makes the BGP with QoS extension more scalable.

Third, ABI and DI can accommodate the link heterogeneity which is caused by different link types and different measuring precisions. For example, if the bandwidth information of a TYPE-3 logical link is imprecise due to some legacy BGP routers, this imprecision is reflected by a large interval length $|\varpi|$ or a small probability ρ in ABI.

B. Calculations of ABI and DI

ABI and DI calculations of a link depend on the link types. In this section, we mainly discuss the calculation of ABI. The technique for calculating DI is the same.

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. The detailed investigation of the traffic sampling techniques is out of the scope of this paper. We only briefly discuss some related parameters of sampling. Two parameters are related to the effectiveness of available bandwidth sampling.

1) Time Scale. Many research results have shown that the Internet traffic has self-similar property [9], which means that the time scale of measurement does not affect the traffic distribution too much. The traffic measured in the time scale of tens of seconds or hundreds of seconds can give us useful results.
2) Sampling Frequency. The sampling frequency influences the calculation precision of the statistical metrics. A very high sampling frequency can obtain precise bandwidth distribution, even if the distribution varies over time. However, in the Internet backbone, due to a large number of aggregated flows, the statistical distribution of the traffic changes very slowly. Thus, in our QoS extension, the sampling time intervals can be at minute levels.[3] This sampling frequency delivers satisfactory precision with acceptable overhead.

We want to find a certain interval $[b_l, b_u]$ and a corresponding ρ for the available bandwidth of a link. Based on the bandwidth vector $\overrightarrow{b}$, the ABI with confidence interval $1-\alpha$ is calculated as follows: Suppose b_m is the median element in $\overrightarrow{b}$. Then, $b_l = b_m - \delta$ and $b_u = b_m + \delta$, where δ is the half-length of interval $[b_l, b_u]$. We adjust δ so that k elements out of n samples in $\overrightarrow{b}$ fall into interval $[b_m - \delta, b_m + \delta]$. k is, thus, constrained by α, ρ, and n, in order to guarantee that the instantaneous bandwidth belongs to $[b_m - \delta, b_m + \delta]$ with probability ρ and the confidence interval is $1-\alpha$. Therefore, we can first compute k for given α, ρ and n, then calculate δ, and lastly obtain $[b_l, b_u]$.

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 to be 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 $\varpi = [b_l, b_u]$ contains k elements of $\overrightarrow{b}$, then the probability, with which the instantaneous bandwidth belongs to the interval ϖ, is no less than ρ with confidence interval $1-\alpha$.

Proof: Appendix includes the proof. ■

Theorem 1 yields several observations.

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 $n(1-\rho) \geq 5$ [10]. Since ρ is close to 1, the number of bandwidth samples n is required to be larger than $5/(1-\rho)$.

For example, if $\rho = 90\%$, $n \geq 50$.

Based on the assumptions 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 (1) can be simplified as

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

The DI for TYPE-1 and TYPE-2 links can be similarly calculated based on the samples of link delay or IGP routing delay. For TYPE-3 links, we assume the end-to-end bandwidth or delay measurement techniques, such as [11], can provide approximate ranges for available bandwidth and delay, as well as the precision rate of the measurement. We use the range and precision rate as the interval ϖ and probability ρ of ABI and DI.

A route is formed when the links are connected together in sequence. In the next two sections, we discuss how to compute the ABI and DI of a route by joining the ABIs and DIs of the links on the route together.

[2] We will address the comparison of ABIs and DIs later in Section IV-B.

[3] A similar case is SNMP, which reports the link load every 5 min.

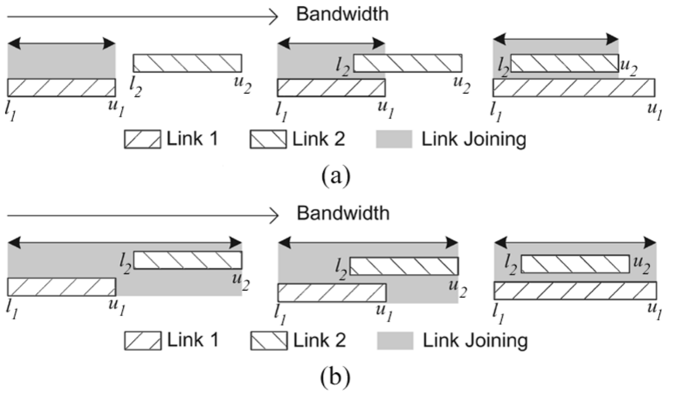


Fig. 2. ABI join operation methods. (a) Join operation method 1. (b) Join operation method 2.

C. ABI Join Operations

Because bandwidth is a concave metric, the available bandwidth of a route is the minimum available bandwidth of all links on that route. To obtain ABI of a route, a straightforward way is to find the available bandwidth of that route, 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 route by joining the ABIs of individual links or subroutes 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 route $v_1 v_2 \ldots v_n$ 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. Because a large number of flows are aggregated on each link, the correlation between two different links in interdomain level is small. Second, the bandwidth distribution outside $\varpi = [b_l, b_u]$ is approximately symmetric around ϖ, i.e., $Pr[b < b_l] \simeq Pr[b > b_u] \simeq (1-\rho)/2$. 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 $|\varpi|$ 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 last part of this section.

In order to compute the interval and probability for $\hat{b} = \hat{b}_1 \oplus \hat{b}_2$, we can set the $\hat{b}.\varpi$ by the combinations of $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$, and then compute $\hat{b}.\rho$ based on $\hat{b}.\varpi$. Obviously, there are multiple options of setting $\hat{b}.\varpi$. A smaller interval length corresponds to a smaller $\hat{b}.\rho$, which means that $\hat{b}.\varpi$ models the instantaneous available bandwidth less precisely. On the contrary, a large interval can cover more instantaneous values, but it may over estimate the dynamic range. We will present two ABI join operation methods, which can be used in different circumstances. For convenience, let $l_1 = \hat{b}_1.b_l$, $u_1 = \hat{b}_1.b_u$, $l_2 = \hat{b}_2.b_l$, and $u_2 = \hat{b}_2.b_u$ in the following discussions.

The two join operation methods for computing $\hat{b}_1 \oplus \hat{b}_2$ are illustrated in Fig. 2. Small rectangle boxes with cross-line patterns represent $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$. The shaded area stands for the interval of the resulting ABI $\hat{b}.\varpi$. For each join operation method, there are three subcases shown in the figure based on the value of u_1.

ABI Join Operation Method 1: 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 $\hat{b}.\varpi = [\min(\hat{b}_1.b_l, \hat{b}_2.b_l), \min(\hat{b}_1.b_u, \hat{b}_2.b_u)]$.

The ABI join method 1, which is shown in Fig. 2(a), applies the minimum operations on the intervals of the two links. The following lemma gives the value of $\hat{b}.\rho$.

Lemma 1: Under the condition of ABI join operation method 1, if $\hat{b}_1.b_u < \hat{b}_2.b_u$, then $\hat{b}.\rho = \hat{b}_1.\rho(1+\hat{b}_2.\rho)/2$; otherwise, $\hat{b}.\rho = \hat{b}_2.\rho(1+\hat{b}_1.\rho)/2$.

Proof: Denote b_1, b_2 and b as instantaneous available bandwidth values on link 1, link 2, and the joined links, respectively. $\rho_1 = \hat{b}_1.\rho$, and $\rho_2 = \hat{b}_2.\rho$. Without loss of generality, assume $l_1 \le l_2$. There are two cases based on the relation between u_1 and u_2.

1) $u_1 \le u_2$:

$$\begin{aligned} Pr[b \in \hat{b}.\varpi] &= Pr[l_1 \le b_1 \le u_1] Pr[b_2 \ge l_1] \\ &\quad + Pr[b_1 > u_1] Pr[l_1 \le b_2 \le u_1] \\ &\ge \rho_1 Pr[b_2 \ge l_2] = \frac{\rho_1(1+\rho_2)}{2} \end{aligned}$$

2) $u_1 \ge u_2$

$$\begin{aligned} Pr[b \in \hat{b}.\varpi] &= Pr[l_1 \le b_2 \le u_2] Pr[b_1 \ge l_1] \\ &\quad + Pr[b_2 > u_2] Pr[l_1 \le b_1 \le u_2] \\ &\ge \rho_2 Pr[b_1 \ge l_1] = \frac{\rho_2(1+\rho_1)}{2} \end{aligned}$$

Thus, we prove the lemma. ■

In the ABI join operation method 1, 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 route, the length of the resulting bandwidth interval will not increase substantially. However, the probability ρ of the resulting ABI may be smaller than $\hat{b}_1.\rho$ and $\hat{b}_2.\rho$. In the following join operation method 2, we enlarge the interval $\hat{b}.\varpi$ so as to increase $\hat{b}.\rho$.

ABI Join Operation Method 2: 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 $\hat{b}.\varpi = [\min(\hat{b}_1.b_l, \hat{b}_2.b_l), \max(\hat{b}_1.b_u, \hat{b}_2.b_u)]$.

The ABI join method 2 [in Fig. 2(b)] covers the whole range including both $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$. The following lemma gives the value of $\hat{b}.\rho$.

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

Proof: Denote b_1, b_2 and b as instantaneous available bandwidth on link 1, link 2, and the joined links, respectively. $\rho_1 = \hat{b}_1.\rho$, and $\rho_2 = \hat{b}_2.\rho$. Without loss of generality, assume $l_1 \le l_2$.

1) If $u_1 \le u_2$:

$$\begin{aligned} Pr[b \in \hat{b}.\varpi] &= Pr[l_1 \le b_1 \le u_2] Pr[b_2 \ge l_1] \\ &\quad + Pr[b_1 > u_2] Pr[l_1 \le b_2 \le u_2] \\ &\ge \rho_1 Pr[b_2 \ge l_1] \\ &\quad + \rho_2 \left(Pr[u_1 < b_1 \le u_2] + Pr[b_1 > u_2] \right) \\ &\ge \frac{\rho_1(1+\rho_2)}{2} + \frac{\rho_2(1-\rho_1)}{2} = \frac{(\rho_1+\rho_2)}{2} \end{aligned}$$

2) If $u_1 \geq u_2$:

$$\begin{aligned} Pr[b \in \hat{b}.\varpi] &= Pr[l_1 \leq b_1 \leq u_1]Pr[b_2 \geq l_1] \\ &\quad + Pr[b_1 > u_1]Pr[l_1 \leq b_2 \leq u_1] \\ &\geq \rho_1 Pr[b_2 \geq l_2] + Pr[b_1 > u_1]\rho_2 \\ &\geq \frac{\rho_1(1+\rho_2)}{2} + \frac{\rho_2(1-\rho_1)}{2} = \frac{(\rho_1+\rho_2)}{2}. \end{aligned}$$

From these two cases, we conclude that $(\rho_1 + \rho_2)/2$ is the lower bound of probability of b belonging to ϖ. Therefore, $\hat{b}.\rho = (\hat{b}_1.\rho + \hat{b}_2.\rho)/2$. ■

The advantage of join operation method 2 over the previous one is that ρ of the joined ABI $\hat{b}$ is never less than ρ_1 and ρ_2 simultaneously. If more links are joined to the route, ρ of the resulting ABI will not decrease substantially. However, the length of the interval ϖ may become larger and larger. Thus, method 2 is more appropriate for the cases where the ranges of $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$ are close to each other. It is not appropriate, when those two intervals are disjoint and separated with large distance.

These two join operation methods discussed above are used by BGP routers to calculate the ABI of a route. Notice the following.

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, method 1 is preferred, especially when the two intervals $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$ are disjoint; however, if $\hat{b}_1.\varpi$ and $\hat{b}_2.\varpi$ are largely overlapped, method 2 is preferred.

2) These two join operations are both based on the symmetric distribution assumption. If this assumption does not hold well, the probability that instantaneous available bandwidth belongs to the interval defined by join operation method 1 is at least $\hat{b}_1.\rho \cdot \hat{b}_2.\rho$. Therefore, we can use method 1 to calculate $\hat{b}.\varpi$, and let $\hat{b}.\rho = \hat{b}_1.\rho \cdot \hat{b}_2.\rho$. In most cases, especially when the link type is TYPE 1 or TYPE 2, ρ is close to 1, and the distribution outside ϖ is approximately symmetric. Simulation results show that these two join operations can give satisfactory precision in calculating the ABI of the joined links.

D. DI Join Operations

Delay is an additive metric, i.e., the delay of a route is the summation of delays of all links on the route. Similarly, calculating a route's DI by using the instantaneous values directly is not practical in BGP. Instead, we calculate DI of a route by joining the DI of each link on the route.

The DI join operation is defined as $\hat{d} = \hat{d}_1 \oplus \hat{d}_2$, where $\hat{d}_1$ and $\hat{d}_2$ are two given DIs. The DI of route $v_1 v_2 \ldots v_n$ is, thus, $\hat{d}_{v_1 v_2} \oplus \ldots \oplus \hat{d}_{v_{n-1} v_n}$, where $\hat{d}_{v_i v_j}$ is the DI of link $v_i v_j$. Following the same idea in ABI join operation, we can first set $\hat{d}.\varpi$ by different combinations of $\hat{d}_1.\varpi$ and $\hat{d}_2.\varpi$, and then compute $\hat{d}.\rho$. We also assume, the delays of different links are independent.

DI Join Operation Method 1: Given that $\hat{d}_1$, $\hat{d}_2$ are two DIs of link 1 and link 2, and $\hat{d} = \hat{d}_1 \oplus \hat{d}_2$, then $\hat{d}.\varpi = [\hat{d}_1.d_l + \hat{d}_2.d_l, \hat{d}_1.d_u + \hat{d}_2.d_u)]$.

The following lemma gives $\hat{d}.\rho$.

Lemma 3: Under the condition of DI join operation method 1, $\hat{d}.\rho = \hat{d}_1.\rho \cdot \hat{d}_2.\rho$.

Proof: Denote d_1, d_2 and d as the instantaneous values of delay on link 1, link 2, and the joined links, respectively

$$\begin{aligned} Pr[d \in \hat{d}.\varpi] &\geq Pr[\hat{d}_1.d_l \leq d_1 \leq \hat{d}_1.d_u]Pr[\hat{d}_2.d_l \leq d_2 \leq \hat{d}_2.d_u] \\ &= \hat{d}_1.\rho \cdot \hat{d}_2.\rho. \end{aligned}$$

Thus, $\hat{d}_1.\rho \cdot \hat{d}_2.\rho$ is the lower bound of the probability that the route delay belongs to $\hat{d}.\varpi$. ■

We can also increase $\hat{d}.\rho$ by enlarging the length of the joined interval. The following join operation method assumes that the delay is symmetrically distributed around the interval defined by DI.

DI Join Operation Method 2: Given that $\hat{d}_1$, $\hat{d}_2$ are two DIs of link 1 and link 2, and $\hat{d} = \hat{d}_1 \oplus \hat{d}_2$, then $\hat{d}.\varpi = [\max(\hat{d}_1.d_l, \hat{d}_2.d_l), \hat{d}_1.d_u + \hat{d}_2.d_u)]$.

$\hat{d}.\rho$ is calculated by using the following lemma.

Lemma 4: Under the condition of DI join operation method 2, if $\hat{d}_1.d_l < \hat{d}_2.d_l$, $\hat{d}.\rho = \hat{d}_2.\rho(1 + \hat{d}_1.\rho)/2$; otherwise, $\hat{d}.\rho = \hat{d}_1.\rho(1 + \hat{d}_2.\rho)/2$.

Proof: Denote d_1, d_2 and d as instantaneous delay values on link 1, link 2, and the joined links, respectively. $\rho_1 = \hat{d}_1.\rho$, and $\rho_2 = \hat{d}_2.\rho$. Without loss of generality, assume $\hat{d}_1.d_l < \hat{d}_2.d_l$. Then, $\hat{d}.\varpi = [\hat{d}_2.d_l, \hat{d}_1.d_u + \hat{d}_2.d_u]$

$$\begin{aligned} Pr[d \in \hat{d}.\varpi] &\geq Pr[\hat{d}_1.d_l \leq d_1 \leq \hat{d}_1.d_u]Pr[\hat{d}_2.d_l \leq d_2 \leq \hat{d}_2.d_u] \\ &\quad + Pr[d_1 < \hat{d}_1.d_l]Pr[\hat{d}_2.d_l \leq d_2 \leq \hat{d}_2.d_u] \\ &= \rho_1\rho_2 + \frac{\rho_2(1-\rho_1)}{2} = \frac{\rho_2(1+\rho_1)}{2}. \end{aligned}$$

The lemma is proved. ■

In the above two DI join operation methods, method 1 is used in general. When $\hat{d}_1.d_l$ is much smaller than $\hat{d}_2.d_l$ or vice versa, method 2 is preferred, because it generates a larger $\hat{d}.\rho$ than method 1 does, and the resulted intervals from both methods are approximately comparable in this scenario.

IV. Protocol Extensions of BGP

In order to enable the interdomain 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; and 3) monitor and update the QoS state of the advertised routes. We mainly focus on the routing of the best effort traffic. There is no essential difference between the best effort traffic and other types of traffic in terms of QoS representation and, thus, our new QoS metrics and the corresponding join operations can handle multiple traffic classes as well (such as the DiffServ architecture). However, the first priority of QoS extension on BGP is to control the extra overhead, so that it is scalable in the global Internet. Therefore, we currently only apply the QoS extension to the single class traffic for BGP in this paper.

A. 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 [12], a new attribute QoS_NLRI is proposed for this purpose. Similar attempt can be taken here. We require

QoS information to be put into the Path Attribute field. Accordingly, the BGP routing table is extended to keep the QoS information. Extended BGP routers use the ABI and DI calculation methods and the join operations, which is presented in Section III, to obtain the ABI and DI for links and paths.

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

Therefore, 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. 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 can decide if a TYPE-3 link is needed for exchanging QoS information.

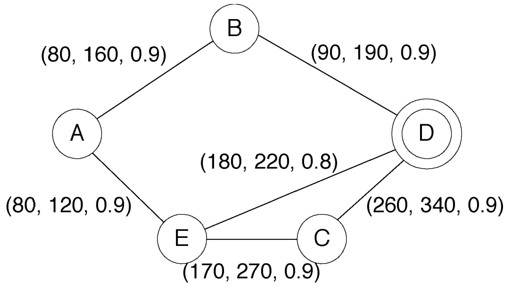


Fig. 3. BGP QoS extension example. ABIs are shown beside links.

B. QoS Path Selection

In the BGP path selection process, QoS-based path selection policy is involved. Because there are multiple policies affecting 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 [6], so that BGP routing 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.

Since ABI or DI is no longer a simple metric, we need to find methods to compare ABIs or DIs of different paths, so that a path with better QoS can be identified.

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 ρ's, they can not be compared directly. Unfortunately, we cannot 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 and 2) ρ of a path is the join result of all the links on the path. Therefore, ϖ has to be normalized to remove the impact of ρ, so that the intervals of two ABIs or DIs are comparable.

To solve this problem, we can scale the length of the interval $|\varpi|$ based on the value of ρ. Intuitively, the larger the ρ, the larger the $|\varpi|$, and vice versa. The primary objective of the normalization method is to provide a way to make two different ABIs comparable and, thus, it is not necessary to find the analytical relationship between $|\varpi|$ and ρ for any distribution. For simplicity of analysis, we use normal distribution as an approximation to find the relation between ρ and $|\varpi|$, and use the result for a general case. In Section VI-B3, our simulation results demonstrate that this approximation works well for other distributions.

Let us assume b is the instantaneous value of bandwidth, and b follows normal distribution $\mathcal{N}((b_l + b_u)/2, \sigma^2)$. $F(x)$ is the cumulative distribution function of b. Because $F(((b_l+b_u)/2)+2\sigma) - F(((b_l + b_u)/2) - 2\sigma) \simeq 95\%$, we can assume that $0 < (|\varpi|/2\sigma) < 2$. Thus

$$\rho = F(b_u) - F(b_l) = 2\int_0^{\frac{|\varpi|}{2\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$$
$$\simeq 0.2\frac{|\varpi|}{2\sigma}\left(4.4 - \frac{|\varpi|}{2\sigma}\right) \simeq \frac{0.44|\varpi|}{\sigma}.$$

Since $|\varpi|/2\sigma$ is relatively small in comparison with 4.4, approximately $\rho \propto |\varpi|$. We can remove the effect of different ρ by normalizing $|\varpi|$ with ρ, i.e., the normalized interval length is $|\varpi|/\rho$. Then, the normalized interval is $\varpi' = [b_m - \delta, b_m + \delta]$, where $b_m = (b_l + b_u)/2$ and $\delta = (b_u - b_l)/2\rho$. Note that the normalization is only for ABI comparison in this paper. The original ABI is exchanged between routers for more accurate calculation. The normalization of DI is also defined the same as ABIs.

2) Weight for Path Selection: With respect to QoS path selection, the paths with large available bandwidth or small delay are preferred. We also favor the paths which have stable available bandwidth and delay. In terms of the normalized ABI, the quality of a path is determined by the interval b_m and δ jointly, which reflects the average bandwidth and the bandwidth variance, respectively. Thus, the weight related to ABI is defined as $\mathcal{W}_b = b_m - \eta\delta = (b_u + b_l - \eta(b_u - b_l)/\rho)/2$, where $\eta > 0$ and it adjusts the tradeoff between the average bandwidth and the stability of the bandwidth. Similarly, the weight for DI (d_l, d_u, ρ) is defined as follows: $\mathcal{W}_d = (d_u + d_l + \eta(d_u - d_l)/\rho)/2$. The path with a larger $\mathcal{W}_b$ or a smaller $\mathcal{W}_d$ is more preferable. In general, the definitions of path weights based on ABI or DI are not limited to these two forms. Other definitions are also possible, depending on the purposes of routing, resource reservation, or traffic engineering. Due to space limitation, we only discuss the above two definitions in this paper.

3) An Example: Fig. 3 shows a simple example of BGP QoS routing using ABI. Let us assume that all links are bidirectional and the numbers beside each link are the ABI parameters (b_l, b_u, 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 $\hat{b}_{CD} = (260, 340, 0.9)$ and passes an advertisement to E. When E receives both advertisements

TABLE I
CONTENT OF BGP ROUTING TABLES AT EACH NODE FOR DESTINATION D

Source	Active Route	AS_PATH	Next_Hop	(b_l, b_u, ρ) / $\mathcal{W}_b$ $(\eta = 1)$
A	Yes	(E D)	E	(80, 120, 0.81) / 75.3
	No	(B D)	B	(80, 190, 0.9) / 73.8
B	Yes	(D)	D	(90, 190, 0.9) / 84.4
	No	(A E D)	A	(80, 160, 0.86) / 73.5
C	Yes	(D)	D	(260, 340, 0.9) / 255.6
	No	(E D)	E	(170, 270, 0.85) / 161.2
E	Yes	(D)	D	(180, 220, 0.8) / 175.0
	No	(C D)	C	(170, 340, 0.9) / 160.6

TABLE II
AFTER ABI OF LINK ED CHANGES FROM (180, 220, 0.8) TO (80, 120, 0.9), THE CONTENT OF BGP ROUTING TABLES

Source	Active Route	AS_PATH	Next_Hop	(b_l, b_u, ρ) / $\mathcal{W}$ $(\eta = 1)$
A	Yes	(E C D)	E	(80, 120, 0.86) / 76.7
	No	(B D)	B	(80, 190, 0.9) / 73.8
B	Yes	(D)	D	(90, 190, 0.9) / 84.4
	No	(A E C D)	A	(80, 160, 0.88) / 74.5
C	Yes	(D)	D	(260, 340, 0.9) / 255.6
E	Yes	(C D)	C	(170, 340, 0.9) / 160.6
	No	(D)	D	(80, 120, 0.9) / 77.8

from D and C, it first joins the ABI of link EC and path CD to get the ABI for the path ECD as $\hat{b}_{ECD} = (170, 270, 0.9) \oplus (260, 340, 0.9) = (\min(170, 260), \max(270, 340), (0.9 + 0.9)/2) = (170, 340, 0.9)$.[4] E then compares the weights of paths ED and ECD. We let $\eta = 1$. Since $\mathcal{W}_{ED} = (180 + 220 - (220 - 180)/0.8)/2 = 175$ and $\mathcal{W}_{ECD} = 160.6$, 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 I. 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.[5] "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. $(b_l, b_u, \rho)/\mathcal{W}_b$ are the ABI and the weight of the path.

C. 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 `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 [2]. There are two cases in which a QoS-aware BGP router may detect the change of QoS information. We will handle them separately.

In the first case, the QoS information on a logical link has changed. We design a new process, called *linkChangeHandler*, to handle such changes. *linkChangeHandler* will check all 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 (180, 220, 0.8) to (80, 120, 0.9), router E will recalculate the weight for path ED as $\mathcal{W}_{ED} = 77.8$. 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 (170, 340, 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 messages from E, it will withdraw the path AED, and calculate the ABI and weight for the new route $AECD$: $\hat{b}_{AECD} = (80, 120, 0.9) \oplus (170, 340, 0.9) = (80, 120, 0.86)$, and $\mathcal{W}_{AECD} = 76.7$. Because $\mathcal{W}_{AECD} > \mathcal{W}_{ABD}$, A will choose the 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 II.

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 and DI instead of the instantaneous values, 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* $(\mathcal{T}_l)$: A small bandwidth or delay fluctuation at a logical link should not trigger the *linkChangeHandler*. Only when the change of the weight 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.

Another advantage of using update thresholds is to adjust the tradeoff between the routing optimality and message overhead. In Section VI, we will use simulation to show the quantitative relations between the routing optimality and the routing message overhead which is controlled by the thresholds $\mathcal{T}_l$ and $\mathcal{T}_r$.

V. EXTENSION OF ABI AND DI TO HISTOGRAM INFORMATION

ABI and DI both use one interval to represent the dynamic QoS information. A natural extension is to employ multiple intervals and the corresponding probabilities to model available bandwidth and delay. In this section, we introduce ABH and DH to characterize QoS dynamics more precisely.

[4] ABI join operation method 2 is used here. In this example, if the ϖ's of two ABIs are disjoint, join method 1 is used; otherwise join method 2 is used.

[5] 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.

A. Definitions of ABH and DH

ABH and DH can be uniformly defined as a set $\{(\varpi_i, \rho_i)\}$, where $\varpi_i = [l_i, u_i]$ is the ith interval and ρ_i is the ith probability that the available bandwidth or delay falls into ϖ_i. We use set $\{(\varpi_i, \rho_i)\}$ to approximate the distribution density functions of bandwidth and delay.

The length of each interval, $|\varpi_i|$, reflects the tradeoff between metric precision and processing overhead. A smaller $|\varpi_i|$ can model the distributions of QoS information with finer granularity, but more system resources have to be consumed in histogram computing and communication. On the other side, when $|\varpi_i|$ is large enough that the available bandwidth or delay is represented by only one interval, ABH and DH will degrade to ABI or DI.

In the ABH and DH join operation methods which will be discussed shortly, we require that the set of intervals $\{\varpi_i\}$ is constructed by dividing the space of bandwidth or delay evenly, i.e., 1) any two different ϖ_i are not overlapping; 2) any two neighboring intervals share the same boundary; and 3) all intervals have the same length $|\varpi|$. Therefore, we can label all intervals using positive integers according to their ranks. That is, $\varpi_i = [(i-1)|\varpi|, i|\varpi|]$. Examples of ABH and DH are shown in Fig. 5, where the horizontal axes represent series of intervals and the vertical axes represent probabilities. In practice, in order to save storage space in the `UPDATE` messages, ABH and DH can be compressed. For example, the neighboring intervals which have identical ρ can be merged together, and also we only need to record the intervals with nonnegligible probabilities ρ_i. In the join operation, the compressed ABH and DH can be restored into the complete forms.

B. Join Operations of ABH and DH

We follow the same notation: the join operator of histogram is $\oplus$. Let us assume that the available bandwidth of two links or subpaths are independent random variables b_1 and b_2. The joined available bandwidth is $b = \min(b_1, b_2)$. It can be shown that the cumulative density function of b is as follows:

$$F_b(x) = F_{b_1}(x) + F_{b_2}(x) - F_{b_1}(x)F_{b_2}(x) \tag{3}$$

where F represents the cumulative density function. Thus, if we know the exact distributions of b_1 and b_2, the above equation can be used to calculate the available bandwidth distribution of the joined links.

In ABH, the bandwidth distribution information is represented approximately by histograms. From (3), we get the probability that the joined bandwidth falls into interval $[x, x+\Delta x]$:

$$\begin{aligned} Pr\,[b \in [x, x+\Delta x]] = {} & Pr\,[b_1 \in [x, x+\Delta x]]\,(1 - F_{b_2}(x)) \\ & + Pr\,[b_2 \in [x, x+\Delta x]]\,(1 - F_{b_1}(x)) \\ & - Pr\,[b_1 \in [x, x+\Delta x]] \\ & \times Pr\,[b_2 \in [x, x+\Delta x]]\,. \end{aligned} \tag{4}$$

According to (4), we have the algorithm of ABH join operation, which is presented in Fig. 4. The time complexity for computing the joined ABH is $O(N)$, where N is the maximum number of intervals in a histogram. An example of ABH join is shown in Fig. 5(a), where the ABHs of link 1, link 2, and the joined links are displayed. The available bandwidth of link 1 and link 2 follows normal distribution. We use 100 samples to generate the histogram. The "Joined ABH" is obtained by using the "ABH-Join" algorithm and the samples. The "Exact Value" is calculated by using (3). This example shows that ABH (with small number of intervals) and its join operation method approximate the bandwidth distribution of the joined links very well.

```
ABH-JOIN($abh_1, abh_2, abh$)
    $s \leftarrow \min\{i : abh_1.\rho_i > 0$ or $abh_2.\rho_i > 0\}$
    $t \leftarrow \max\{i : abh_1.\rho_i > 0$ or $abh_2.\rho_i > 0\}$
    $w_1 \leftarrow 0.0$
    $w_2 \leftarrow 0.0$
    for $i \leftarrow s$ to $t$
    do
        $abh.\rho_i \leftarrow (1 - w_2)abh_1.\rho_i + (1 - w_1)abh_2.\rho_i$
                $- abh_1.\rho_i \cdot abh_2.\rho_i$
        $w_1 \leftarrow w_1 + abh_1.\rho_i$
        $w_2 \leftarrow w_2 + abh_2.\rho_i$
```

Fig. 4. Algorithm for computing $abh = abh1 \oplus abh2$.

The distribution of the delay can similarly be computed by using the delay probability density functions of the links or subpaths

$$f_d(x) = \int_0^x f_{d_1}(t) f_{d_2}(x-t)\mathrm{d}t \tag{5}$$

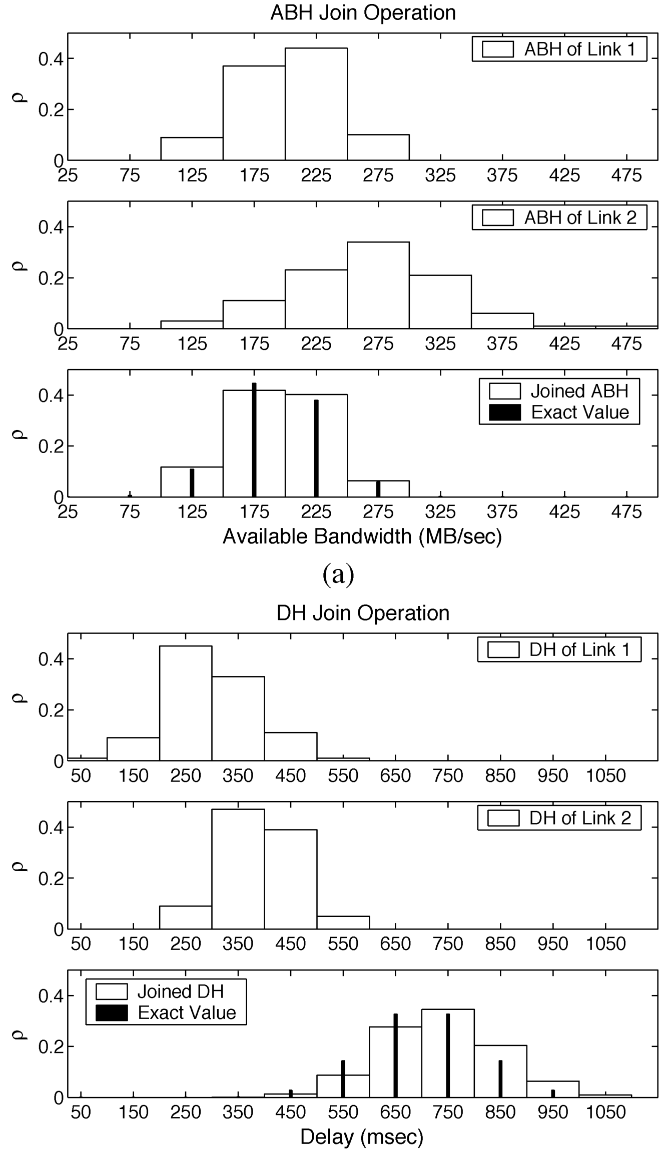
where f is the probability density function. d_1, d_2, and d are the delays of two subpaths and the joined path, respectively. DH of the joined path can be calculated by the discrete version of the above convolution equation. Suppose $dh = dh_1 \oplus dh_2$, where dh_1 and dh_2 are DHs of corresponding subpaths. Then

$$dh.\rho_k = \sum_{i=0}^{k} dh_1.\rho_i \cdot dh_2.\rho_{k-i}. \tag{6}$$

The time complexity of DH join is $O(N^2)$, where N is the maximum number of intervals in a histogram. If we apply fast fourier transformation to calculate the above convolution [13], the time complexity decreases to $O(N \log N)$. The example of DH join is shown in Fig. 5(b). The delay of link 1 and link 2 follows normal distribution. The "Exact Value" is calculated by using (5). The "Joined DH" is computed with (6), where 100 samples are used to generate the histograms. Again, our method approximates the delay distribution well.

C. Discussions

1) Using ABH and DH is a practical and flexible way to advertise the detailed statistical information of bandwidth and delay in routing protocols. The precise probability density function, which is difficult to obtain, causes too much processing overhead and, thus, is impractical to be used. While, the overhead of histogram metrics can be limited by adjusting the number of sample points and the interval length $|\varpi|$ appropriately. Suppose there are n historical samples recorded from the link bandwidth or



(a)

(b)

Fig. 5. Examples of ABH and DH join operations. (a) ABH join operation. (b) DH join operation.

delay. The samples are stored in a FIFO queue and updated periodically, which costs n units memory. A single scan can convert these samples into histogram metrics, and the computing complex is $O(n)$. The storage needed by histogram metrics is controlled by $|\varpi|$, and can be further decreased by using the compressed forms.

Compared with ABI and DI, the histogram metrics introduce extra processing overhead, but better performance can be obtained, which is the tradeoff we have to make. It is also interesting to note that the optimality of ABI may be better than ABH, when $|\varpi|$ in ABH is very large. This is because ABI does not fix the position and the length of the interval, while the intervals in ABH are defined in a fixed structure, dividing the bandwidth space evenly. Thus, ABI can represent the distribution of bandwidth more efficiently than an ABH that has only a few intervals. The simulation results in Fig. 9(a) of Section VI-C verify this observation. On the other hand, because of the fixed interval structure in ABH, the intervals from two ABHs are aligned to each other. This simplifies the ABH join operation. For example, there is only one ABH join operation method; while, we define two methods for ABI, based on the relationships of two intervals.

2) By using ABH and DH, we can acquire, from the advertised information, not only the expected values of the QoS parameters but also the variance and even higher order moment information, which characterizes the stability of the QoS. The route weight is, thus, defined to reflect the average value and the stability together: $\mathcal{W}_b = \mathbf{E}[abh] - \eta\mathbf{STD}(abh)$ and $\mathcal{W}_d = \mathbf{E}[dh] + \eta\mathbf{STD}(dh)$, where $\mathbf{E}[\cdot]$ and $\mathbf{STD}(\cdot)$ stand for the expectation and the standard deviation, respectively, which are calculated from the histograms.

In addition to ABH and DH, a simplified approach is to advertise average values and variances of the link QoS parameters directly.

With respect to bandwidth information advertising, AAB of each link can be used in BGP. The AAB of a route is defined as the minimum AABs over all links on the route. However, due to the concave property of the minimum operation, this join method based on AAB can overestimate the real average values of the route available bandwidth and, thus, result in incorrect routing decision. The reason is, from Jensen inequality, $\min(\mathbf{E}[b_1], \mathbf{E}[b_2]) \geq \mathbf{E}[\min(b_1, b_2)]$, where b_1 and b_2 are the available bandwidth of two links. Our simulation results in Section VI-C show that ABH is better than AAB in providing higher routing optimality and lower resource reservation rejection ratio.

In terms of delay, because it is an additive metric, applying average delay (AD), delay variance (DV) and the addition operations to BGP advertising can obtain correct results under the independent link assumption, i.e., $\mathbf{E}[d_1] + \mathbf{E}[d_2] = \mathbf{E}[d_1 + d_2]$ and $\mathbf{Var}(d_1) + \mathbf{Var}(d_2) = \mathbf{Var}(d_1 + d_2)$, where d_1 and d_2 are the delays of two links. If we only consider the average value and the variance, ADH performs approximately the same as AD and DV.

3) *Convergence Property*: Average bandwidth, AD, and DV, all satisfy the monotonicity condition of the convergence of path vector routing protocols [14], and they do not influence the convergence property of BGP. Thus, the BGP routing is preserved to be convergent, if $\mathcal{W}_d$ or $\mathcal{W}_b$ (subject to $\eta = 0$) is used. On the other hand, available bandwidth variance of a route is not monotone, i.e., the variance of the route bandwidth may decrease when a new link is added to the route.[6] Therefore, if the bandwidth variance is involved in the route weight $\mathcal{W}_b$, i.e., $\eta \neq 0$, the path vector routing protocols may not be convergent. However, in this scenario, our extended

[6] For example, a bottleneck link with a small bandwidth variance is added to a route which previously has a large bandwidth variance.

BGP QoS routing protocol can still converge by compounding the route weight, $\mathcal{W}_b$, to the path selection polices of AS peer relationships, according to the path selection policies described in Section IV-B.

VI. Simulation Results

In order to evaluate the performance of the QoS extension to BGP, extensive simulations have been conducted. Based on the same routing protocol (BGP) and the format of weight definitions, we simulate the performance of bandwidth information advertising by using the following metrics: link capacity (LC), available bandwidth (AB), ABI, AH, AAB. The related QoS routing protocols are called LCR, ABR, ABIR, AHR, and AABR, respectively, by adding a suffix "-R" to the names of the metrics. Similarly, in order to test delay related metrics, DI, DH, AD, and DV are simulated.

In this section, we discuss three aspects of our simulation results as follows: 1) the relationship between routing optimality and routing message overhead; 2) the performance of histogram metrics; and 3) the routing results if QoS stability is considered. We demonstrate that our new statistical routing metrics can find much better routes than static metrics and have much lower message overhead than instantaneous metrics.

A. Simulation Model

The purpose of the simulation is to study routing optimality and message overhead in BGP QoS extensions. Based on the BGP routing protocol in [2], three simplifications are made: 1) each AS is simplified as a single node; 2) we ignore address aggregations; and 3) we consider bandwidth or delay information as the only path selection metric and ignore other BGP routing policies.

A BGP protocol simulator is implemented based on the simplified interdomain routing model. The bandwidth and delay information are advertised using our proposed metrics. The routes are selected according to the route weight defined in previous sections. For performance comparison, we also simulate the scenarios where traditional QoS metrics are used, and the route with the largest bandwidth or the least delay is preferred.

Internet topology generator BRITE [15] 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 capacity of each link is generated randomly from the interval [10, 1050].

The dynamic behaviors of the available bandwidth and delay are 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 or delay. In each time unit, a new value is generated following this distribution, i.e., the available bandwidth or delay 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 and 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 our simulations.

Two metrics are defined to quantify the performance of routing protocols.

1) *Routing Optimality* ξ: Denote $\beta(\mathcal{R})$ as the AAB between all pairs of nodes based on the result of a routing protocol $\mathcal{R}$. The routing optimality of $\mathcal{R}$ is then defined as $\xi = \beta(\mathcal{R}) / \max \beta$, where $\max \beta$ can be obtained by running Dijkstra's algorithm on the network graph with the instantaneous available bandwidth as the link weight. In terms of delay related metrics, the routing optimality is similarly defined as $\xi = \Delta(\mathcal{R}) / \min \Delta$, where $\Delta(\mathcal{R})$ is the average delay of all source-destination pairs as the results of $\mathcal{R}$, and $\min \Delta$ is the optimum results, standing for the minimum delay that can be achieved.

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 QoS information update.

B. Optimality and Routing Message Overhead

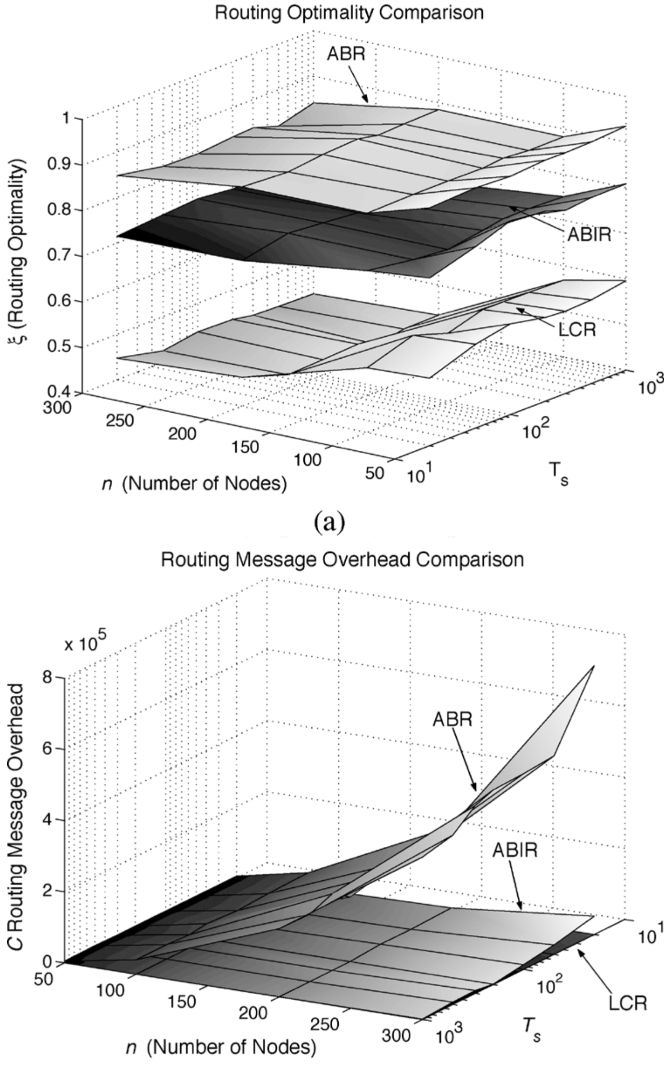
In order to show the advantages of using our proposed statistical metrics, we present the simulation results of LCR, ABR, and ABIR to study the relationship between the routing optimality and the message overhead. In the route weight calculation of ABI, $\eta = 1$.

1) Performance Overview: The routing optimality and routing message overhead are shown in Fig. 6(a) and (b) with respect to different network topologies and values of T_s. Normal distribution is used to model the link available bandwidth.

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 we assume that the routing protocol converges fast enough in one time unit, ABR can achieve 100% optimality. The ABR curves, shown in Fig. 6, have nonzero 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 to 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%.

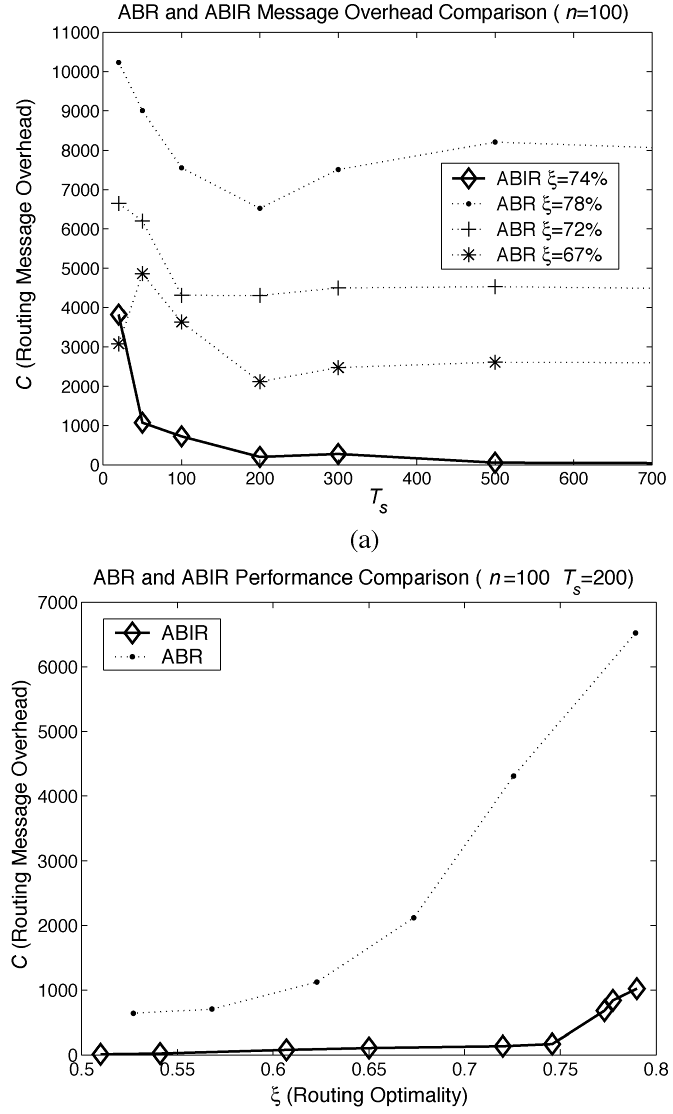
ABIR makes a good compromise between the routing message overhead and the routing optimality. Its routing optimality ξ is about 75%. Its routing message overhead is far less than the ABR protocol. In the worst case of our simulations, where $T_s = 20$ time units and the number of node is 300, the routing message overhead incurred in ABIR is only 6.8% of that in ABR. When 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. In summary, ABIR achieves higher



(a)

(b)

Fig. 6. Performance comparisons. (a) Routing optimality comparison. (b) Message overhead comparison.



(a)

(b)

Fig. 7. ABIR demonstrates better optimality than ABR when incurring the same amount of routing message overhead. (a) Routing message overhead of ABR and ABIR. (b) Performance comparison of ABR and ABIR.

routing optimality than LCR with much lower routing message overhead than ABR.

2) Comparison Between ABIR and ABR With Large Threshold: ABIR can substantially reduce routing message overhead by decreasing only slightly the routing optimality. Although ABR can also control the routing message overhead to a low level by using sufficiently large thresholds, $\mathcal{T}_l$ and $\mathcal{T}_r$, simulation results show that the optimality of ABR degrades quickly as the thresholds increase. If the same amount of message overhead is incurred, ABIR performs much better than ABR in terms of routing optimality.

Fig. 7(a) presents the simulation results of ABR and ABIR in a network of 100 nodes and the link available bandwidth follows normal distribution. The upper three curves show the message overhead C and optimality ξ of ABR with respect to T_s. When C is reduced (by increasing $\mathcal{T}_l$ or $\mathcal{T}_r$), ξ decreases, e.g., when $C \simeq 2000, \xi = 67\%$. On the contrary, ABIR, which is shown as the lowest curve, can achieve 74% optimality with much lower message overhead.

In Fig. 7(b), the relationship between the routing optimality and the routing message overhead is shown in one curve directly. Higher routing optimality is obtained at the price of larger message overhead. In the range of the optimality which can be achieved by ABIR, the routing message overhead incurred by ABIR increases much more slowly than that of ABR.

3) ABIR in Different Traffic Distributions: In Section IV-B1, we use normal distribution to derive an ABI normalization method as an approximation for any general distribution. The simulation results below support that this approximate method also 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 the shape parameter a is picked randomly from [0, 1]. 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.

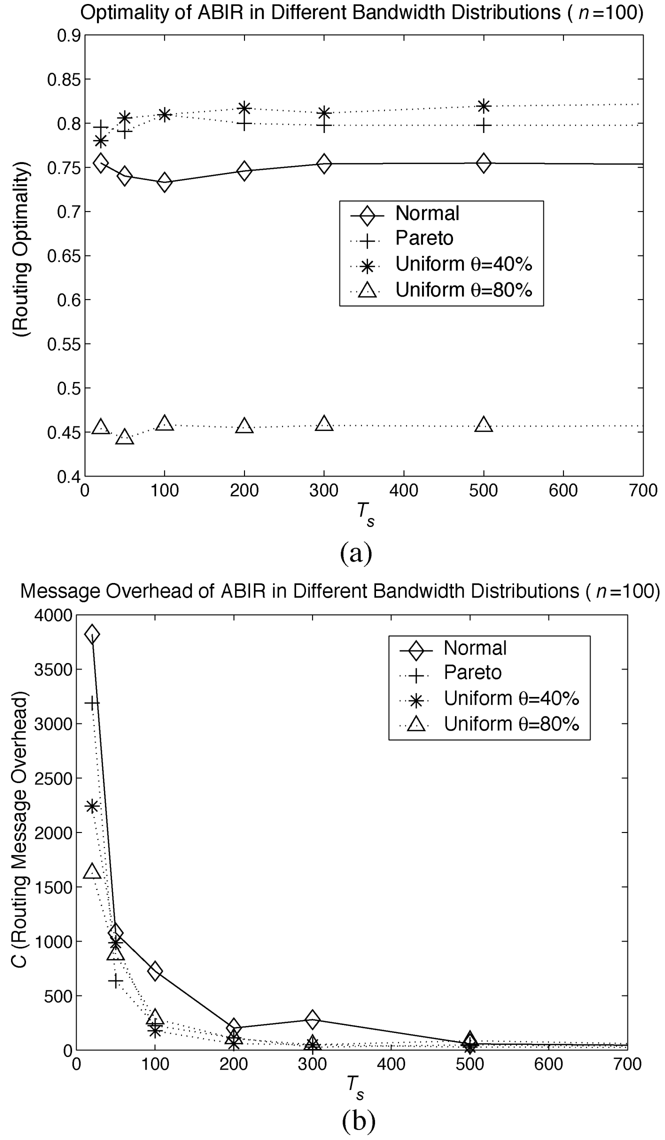


Fig. 8. Performance of ABIR in different bandwidth distributions. (a) Optimality comparison in different bandwidth distributions. (b) Message overhead comparison in different bandwidth distributions.

The simulation results are shown in Fig. 8. If the available bandwidth follows Pareto or uniform ($\theta = 0.4$) distribution, ABIR has similar optimality and message overhead in all three distributions. However, in the uniform distribution of $\theta = 0.8$, the optimality of ABIR is almost the same as LCR [shown in Fig. 6(a)]. This can be explained as follows. If θ is close to 1, ABI of each link tends to have $\varpi = [0, D]$, because ρ is chosen to be around 0.9. Thus, in this scenario, the ABIs of links actually only reflect link capacities. We conclude that ABIR performs better than LCR, if the available bandwidth is mainly distributed in an interval whose length is smaller than D. For example, if $\theta = 0.4$, as shown in the simulation, the optimality of ABIR is about 80%, much higher than the optimality of LCR.

C. Evaluations of Histogram QoS Metrics

In this section, we use ABH as an example to study the granularity of histograms and its advantages over other metrics. In route weight calculation, η equals 0. ABHR is compared with ABIR and AABR (AABR advertises link average bandwidth directly). The available bandwidth of each link is modeled by normal distribution.

Fig. 9(a) demonstrates the impact of the histogram granularity $|\varpi|$ on the optimality of ABHR. It is clearly shown that smaller $|\varpi|$ results in higher routing optimality. When $|\varpi|$ is less than about 200, ABHR has better performance than both ABIR and AABR. The selection of $|\varpi|$ really depends on the tradeoff between the processing overhead and the routing performance. In the following simulations, we let $|\varpi|$ equal to 50. From the figure, we can also observe that ABIR shows better performance than AABR and even better than ABHR when histogram granularity is large. Although ABI uses only one interval to model bandwidth distribution (actually ABI divides bandwidth space into three intervals), ABI can represent the distribution more efficiently and flexibly than ABH does, if they use the same number of intervals. It is because ABI does not fix the position and length of the interval when defining the distribution of bandwidth.

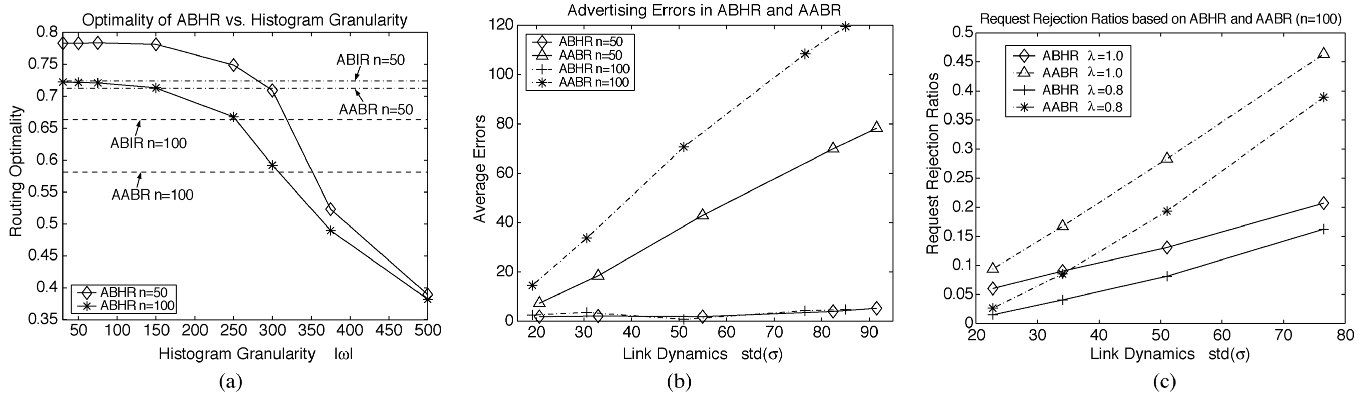
Both ABHR and AABR can provide average bandwidth information on a route. In Section V-C, we argue that the bandwidth values provided by ABHR are more precise than those from AABR. The conclusion is also demonstrated in Fig. 9(b). The horizontal axis is the averaged standard deviation of the σ parameters of normal distributions on all link, standing for the link dynamics. In order to obtain errors of bandwidth advertising, time average of the instantaneous available bandwidth between any pair of routers is calculated as the precise value. This result is subtracted from the expected value computed from the ABH in the routing table of ABHR and the average bandwidth in AABR, respectively. The advertising errors of all router pairs are averaged and the results are shown in Fig. 9(b). ABHR can advertise available bandwidth information much more precisely than AABR, especially when link available bandwidth changes more dynamically. Our further simulation also shows that the advertising error of ABIR is between those of ABHR and AABR.

Due to the precise available bandwidth information provided by ABHR, bandwidth reservations can also benefit. A reservation request will be accepted to further signaling process, if the required bandwidth is below λb, where b is the AAB obtained from the routing table and $0 < \lambda \leq 1$. Because AABR over-advertises the AAB information, it incurs much more false positive acceptances than ABHR does. The simulation results are demonstrated in Fig. 9(c).

D. Routing Considering QoS Stability

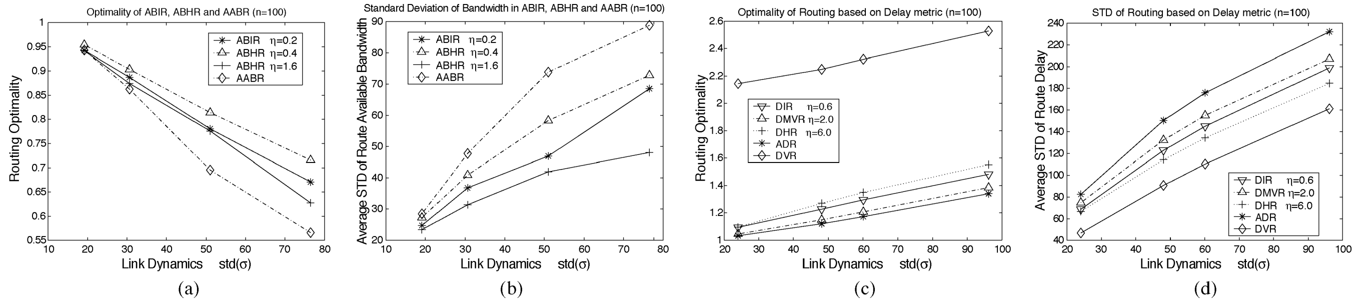
In this section, we show the routing results in scenarios where both the QoS stability of route and the routing optimality are considered in the path selection process. By adjusting η, we can change the influence of QoS stability on route weight calculation.

Fig. 10(a) and (b) presents the routing performance of ABIR, ABHR, and AABR. A larger η leads to preferring routes which have more stable available bandwidth. As has been analyzed in Section V-C, due to the precise bandwidth information advertising in ABI and ABH, they can find routes with smaller



(a) (b) (c)

Fig. 9. Evaluations of ABH. (a) Histogram granularity affects the optimality of ABHR. (b) ABHR advertises more precise bandwidth information than AABR. (c) ABHR leads to less request rejection ratios than AABR.



(a) (b) (c) (d)

Fig. 10. Performance of routing by using ABI, ABH, DI and DH. (a) Optimality comparisons of bandwidth related metrics. (b) Standard deviations of route available bandwidth. (c) Optimality comparison of delay related metrics. (d) Standard deviations of route delay.

bandwidth variance and with better optimality than AABR, especially when the link dynamics is large.

Fig. 10(c) and (d) shows the routing results related to the delay metrics. ADR is the routing using average link delay; DVR is the routing using the variance of link delay. In DMVR, the link average delay and the standard deviation are combined together. Because delay expectation and delay variance are additive metrics, they can be correctly advertised in ADR, DVR, and DMVR. Not surprisingly, DMVR has similar performance as DIR and DHR in the simulation. By making use of both average delay and delay variance information in routing, DIR, DHR, and DMVR provide flexible ways to select routes which satisfy routing optimality and QoS stability constraints. Also, the figures show that we can increase the QoS stability of routes by sacrificing a little bit performance in routing optimality. If average delay is the only metric considered, such as ADR, the resulting routes may have large delay variance; on the other hand, DVR considers the delay variance only, and the resulting routes have the worst optimality among all.

VII. Related Work

The related work on interdomain QoS routing is discussed as follows. Bonaventure [16] focuses on how to distribute QoS information flexibly by BGP in different network scenarios. Cristallo and Jacquenet [12] propose a new attribute for BGP UPDATE message, QoS_NLRI, to record QoS related information. Abarbanel and Venkatachalam [17] utilize BGP to propagate traffic engineering weight, which represents 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 cannot 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 [18] extend multiprotocol extension to BGP4 (MBGP) for interdomain 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 propose QoS routing algorithms based on the probability density function in [19], [20]. 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 [21] model the imprecise QoS value by an interval which is calculated from exponential average. Being different from ABI and DI, their interval is a deterministic bound. It can be viewed as a special case of our model, where ρ equals 1.0.

VIII. Conclusion and Future Work

In this paper, we investigate a very challenging problem in the area of interdomain routing—extending the existing BGP to support QoS. Two challenges, scalability and heterogeneity, make this problem very difficult to solve. Finding efficient and effective QoS metrics is important to tackle these challenges. Thus, we propose two novel composite QoS metrics, the ABI and DI, to perform QoS advertisement and route selection in BGP. We further extend ABI and DI to ABH and DH, respectively. Our simulation results show that these statistical metrics can accommodate heterogeneous QoS information and provide satisfying performance with low message overhead. Also, they are more informative and can represent dynamic QoS information more precisely than traditional static metrics (such as link capacity) and some simple statistics based on the instantaneous values (such as AAB).

It is observed that the changing patterns of the Internet traffic are similar in the period of days or months. In the future, we will take advantage of this periodical behavior to represent the QoS routing information more efficiently. Suppose that $\mathcal{Q}(t)$ denotes the statistical metrics (such as ABI, ABH, DI, and DH) at time t in one period. We can advertise a time series of QoS metrics $\{\mathcal{Q}(t)\}$ to summarize the QoS routing information of the whole time period. Several interesting problems need to be explored. For example, how to obtain $\mathcal{Q}(t)$, which is not directly advertised, by interpolation or estimation from the known metrics of time around t? As another example, how to join and compare a time series of QoS metrics? Effective solutions to these problems give us more powerful representations to the interdomain QoS routing information.

Appendix I
Proof of Theorem 1

Let us assume that the instantaneous available bandwidth b is a random variable. Denote p as the probability for b belonging to $\varpi = [b_l, b_u]$, i.e., $p = Pr[b_l \leq b \leq b_u]$. If k elements from all the samples in $\overrightarrow{b}$ belong into interval ϖ, by the proportion estimation theory [10], for any bandwidth distribution, we have, $Pr[p \geq (k/n) - z_\alpha \sqrt{(k/n(1-k/n))/n}] \simeq 1-\alpha$, where $1-\alpha$ is the confidence interval and z_α is the value of the standard normal curve above which we can find an area of α. According to the definition of ABI, p is required to be greater than ρ with confidence $1-\alpha$, i.e., $Pr[p \geq \rho] = 1-\alpha$, we get the requirement on k: $(k/n) - z_\alpha \sqrt{(k/n(1-k/n))/n} = \rho$. By solving this equation, we obtain (1). Therefore, if k satisfies (1), the probability for the instantaneous bandwidth b falling into ϖ is greater than ρ with confidence $1-\alpha$. ■

References

[1] R. A. Guerin, A. Orda, and D. Williams, “QoS routing mechanisms and OSPF extensions,” in *Proc. IEEE GLOBECOM*, 1997, pp. 1903–1908.

[2] Y. Rekhter and T. Li, “A border gateway protocol 4 (BGP-4),” RFC 1771, Mar. 1995.

[3] Cisco Systems Inc. BGP best path selection algorithm. [Online]. Available: http://www.cisco.com/warp/public/459/25.shtml

[4] G. Huston, “Interconnection, peering and settlements, part I and II,” *Internet Protocol J.*, vol. 2, no. 1–2, Jun. 1999.

[5] T. G. Griffin, F. B. Shepherd, and G. Wilfong, “The stable paths problem and interdomain routing,” *IEEE/ACM Trans. Networking*, vol. 10, pp. 232–243, Apr. 2002.

[6] L. Gao, T. Griffin, and J. Rexford, “Inherently safe backup routing with BGP,” in *Proc. INFOCOM*, Apr. 2001, pp. 547–556.

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

[8] P. P. Pan, E. L. Hahne, and H. G. Schulzrinne, “BGRP: A tree-based aggregation protocol for interdomain reservations,” *J. Commun. Networks*, vol. 2, no. 2, pp. 157–167, Jun. 2000.

[9] W. E. Leland, M. S. Taqqu, W. Willinger, and D. Wilson, “On the self-similar nature of ethernet traffic (extended version),” *IEEE/ACM Trans. Networking*, vol. 2, pp. 1–15, Feb. 1994.

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

[11] M. Jain and C. Dovrolis, “End-to-end available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput,” in *Proc. ACM SIGCOMM*, 2002, pp. 295–308.

[12] 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, Mar. 2002.

[13] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[14] J. L. Sobrinho, “Network routing with path vector protocols: Theory and applications,” in *Proc. ACM SIGCOMM*, 2003, pp. 49–60.

[15] Boston University Representative Internet Topology Generator. [Online]. Available: http://cs-www.bu.edu/brite/

[16] O. Bonaventure, “Using BGP to distribute flexible QoS information,” Internet Draft, draft-bonaventure-bgp-qos-00.txt. Work in Progress, Feb. 2001.

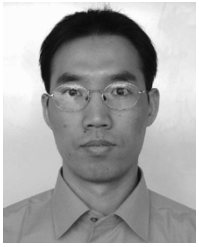
[17] B. Abarbanel and S. Venkatachalam, “BGP-4 support for traffic engineering,” Internet Draft, draft-abarbanel-idr-bgp4-te-00.txt. Work in Progress, Sept. 2000.

[18] A. Fei and M. Gerla, “Extending BGMP for shared-tree interdomain QoS multicast,” in *Proc. IWQoS*, 2001, pp. 123–139.

[19] D. H. Lorenz and A. Orda, “QoS routing in networks with uncertain parameters,” *IEEE/ACM Trans. Networking*, vol. 6, pp. 768–778, Dec. 1998.

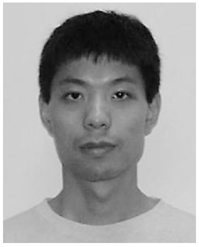
[20] R. Guérin and A. Orda, “QoS routing in networks with inaccurate information: Theory and algorithms,” *IEEE/ACM Trans. Networking*, vol. 7, pp. 350–364, June 1999.

[21] S. Chen and K. Nahrstedt, “Distributed QoS routing with imprecise state information,” in *Proc. IEEE ICCCN*, Oct. 1998, pp. 614–621.



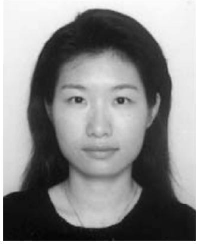
Li Xiao (S’02) received the B.S. and M.Eng. degrees in automatic control from Tsinghua University, Beijing, China, and the M.S. degree in computer science from the University of Illinois at Urbana–Champaign. He is currently working towards the Ph.D. degree from the University of Illinois at Urbana–Champaign.

His research interests are computer networks and data communication, with focus on network routing, quality-of-service (QoS), and network resilience.



Jun Wang (S’01) received the B.S. and M.Eng. degrees in computer science and technology from Tsinghua University, Beijing, China, and the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign in 2003.

He is a Postdoctoral Associate with the National Center for Supercomputing Applications (NCSA) and the Department of Computer Science, University of Illinois at Urbana–Champaign. His research interests include computer networks and data communications, network survivability and security, network quality-of-service (QoS), multimedia systems, and distributed systems.



King-Shan Lui (M'02) received the B.Eng. and M.Phil. degrees in computer science from the Hong Kong University of Science and Technology, Hong Kong, China, and the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign.

She is now an Assistant Professor in the Department of Electronic and Electrical Engineering, University of Hong Kong. Her research interests are network protocol design and analysis.

Klara Nahrstedt (M'95) received the B.A. degree in mathematics and the M.Sc. degree in numerical analysis from Humboldt University, Berlin, Germany, in 1984 and 1985, respectively, and the Ph.D. degree in computer and information science from the University of Pennsylvania, Philadelphia, in 1995.

From 1985 to 1990, she was a Research Scientist in the Institute for Informatik, Berlin, Germany. She is an Associate Professor in the Computer Science Department, University of Illinois at Urbana–Champaign. Her research interests are directed toward multimedia middleware systems, quality-of-service (QoS), QoS routing, QoS-aware resource management in distributed multimedia systems, and multimedia security. She is the coauthor of the widely used multimedia book *Multimedia: Computing, Communications and Applications* (Englewood Cliffs, NJ: Prentice Hall, 1995). Since 2001, she is the Editor-in-Chief of the *ACM/Springer Multimedia Systems Journal*, and the Ralph and Catherine Fisher Associate Professor.

Dr. Nahrstedt is the recipient of the Early NSF Career Award, the Junior Xerox Award, and the IEEE Communication Society Leonard Abraham Award for Research Achievements.