Route Optimization among a Group of Multihomed Stub Networks

Yong Liu, A. L. Narasimha Reddy

Department of Electrical Engineering
Texas A&M University
College Station, TX 77843
Email: {yongliu,reddy}@ee.tamu.edu
Phone: (979)845-7598

***Abstract*— Multihoming is used by stub networks to improve the reliability of their Internet connectivity. In recent years, commercial "intelligent route control" devices are used by multihomed stub networks to optimize the routing of their Internet traffic.**

In this work, we first conduct measurement of qualities of alternate paths through multihoming. In the remaining part, we study the route control of traffic among a group of multihomed stub networks which may belong to an organization and exchange data regularly. This type of route control is special because the access links of such networks may not be over-provisioned and are shared by traffic controlled by multiple route control devices. Such shared bottlenecks may affect the effectiveness of uncoordinated "intelligent route control". We propose a global optimization based approach to coordinate the route control among such a group of networks. Our approach can avoid oscillations which may be caused by uncoordinated route control. Simulation results show that our approach has advantages over equal-splitting based static load-balancing.

Keywords: Multihoming, Intelligent Route Control, Optimal Routing

I. Introduction

Multihoming is used as a method to improve the reliability of Internet connectivity of "stub networks"(networks that do not provide transit service for other networks). The emergence of commercial "intelligent route control" devices (e.g. [1]) in recent years has spurred much interest in multihoming. Because alternate paths via different upstream ISPs may have different qualities at a given time, route control devices can improve the performance of stub networks by selecting the best available path according to active or passive measurement results. Measurement based analysis of the benefit of multihoming [2] shows that route control may improve performance significantly for both enterprises and large data centers. Guo et al [3] have discussed the design space of multihoming route control systems. Goldenberg et al [4] have studied the optimization of cost and performance for multihoming.

Most current route control devices do not consider the interaction among themselves. Each route control device chooses the best path based on its own view of the qualities of alternate paths. When the traffic controlled by multiple route control devices shares common bottleneck links and is not negligible on the links, uncoordinated route control may cause oscillations because of the dynamics of the traffic and the variation of the queuing delays and loss rates on the links. Since today's backbone links are usually over-provisioned and the traffic controlled by route control devices are widely distributed on the Internet, these problems haven't become evident yet.

However, unlike links of Internet backbone networks, access links of stub networks are usually not over-provisioned. When a few multihomed stub networks exchange a large amount of traffic among themselves (e.g. an enterprise may consist of a few multihomed branches which may exchange a large amount of data regularly), the route control among such a group of networks may cause problems.

In this paper, we propose an optimal routing [5] based approach for coordinating the route control among such a group of multihomed stub networks. We consider the case when all the stub networks are under the control of a single administration. Because the number of stub networks controlled by such an administration is not large (e.g. less than 10), the cost of global coordination is not high.

We consider stub networks that have a block of IP addresses independent of its ISPs. When a stub network has independent IP addresses, it uses BGP to advertise its IP addresses to its ISPs. The stub network can send outgoing traffic via either of its ISPs, but it cannot decide which ISP the ingress traffic comes from. The routes of ingress traffic are decided by the BGP relationships between ASes(Autonomous Systems) on the Internet. While it is possible to control the incoming traffic direction through selective advertisement of addresses to different ISPs, such a control is only possible over longer timescales and may leave the stub network vulnerable to network failures. Hence, we consider the case when a stub network advertises all its addresses to its ISPs.

Our approach can also be applied to stub networks multihomed using NAT(Network Address Translation). The major difference is that there are more alternate paths for stub networks multihomed using NAT and the calculation of the optimal routing solution takes more time. However, when the additional benefit of using more alternate paths is not large, the stub networks can use a subset of the alternate paths.

This paper is organized as follows. In section II, we describe

our measurement experiment of qualities of alternate paths provided by multihoming and summarize the results. We propose our global coordination approach in section III and evaluate the performance of our approach by comparing it with a static load-balancing approach in section IV. We conclude this paper and describe our future work in section V.

II. Measurement of Qualities of Alternate Paths Provided by multihoming

A. Measurement method

To study the characteristics of alternate paths possible through multihoming, we conduct measurement on the Internet. First, we identify a number of multihomed stub networks using the BGP RIBs from the Route Views website [6]. Then, we identify the edge routers of these stub networks using the traceroute utilities on a number of Scriptroute servers [7]. Finally, we measure the alternate paths by "pinging"(using "traceroute" packets) the IP addresses of the edge routers that belong to the ISPs of the stub networks from a number of Scriptroute servers.

In this way, we measure the alternate "round trip paths" from a Scriptroute server to a multihomed stub network. We compare the qualities of two "round trip paths" by calculating the differences of the average RTTs and average loss rates of the "ping" packets on the two paths. If the IP addresses we "pinged" belong to one router, the reverse paths of the acknowledgement packets from the edge router to the Scriptroute server are same. In this case we get the difference between the forwarding alternate paths. Otherwise, we get difference between alternate "round trip paths" to the stub networks.

In our measurement, we find that the RTTs and loss rates of the "ping" packets to the edge routers sometimes fluctuate significantly and have a daily pattern. We suspect that it is because the access links of the remote networks are busy, so the characteristics of the paths may be concealed by the queuing delay and packet losses on the access links. To get rid of the queuing delay on access links of the stub networks, we calculate the RTT of a "route trip path" from a Scriptroute server to an edge router as $RTT1_{min} + RTT2 - RTT2_{min}$, where $RTT1$ is the RTT of a "ping" packet to the edge router (say N hops away from the Scriptroute server), $RTT2$ is the RTT of a "ping" packet to the router $N-1$ hops away from the Scriptroute server on the same path. The minimum of $RTT1$ and the minimum of $RTT2$ are expected to be the propagation delay (without queuing delays on links) of the "round trip paths". We use the loss rates of the packets to routers $N-1$ hops away from the Scriptroute server as the loss rates for the "round trip path".

B. Measurement results

We carried out measurement for more than 2 days in June, 2004. Here, we give some results from our measurement of 33 pairs of alternate paths in the United States.

Fig. 1 shows the RTT differences and loss rate differences of 3 pairs of alternate paths. The RTTs and loss rates are averaged over 5 minute durations. From the figure, we see significant performance differences between a pair of alternate paths for both RTTs and loss rates. The differences change over time. A better path may become worse some time later. The RTT differences persist over both small time scales (e.g. 5 minute) and over long time scales (e.g. a few hours). Similar differences are observed for more than half of the paths we measured. These observations indicate: (1) route control is possible to improve the performance of multihomed network significantly, and it should be implemented in a dynamic manner; (2) both small time scale route control and large time scale route control have the potential to improve the routing of multihomed stub networks.

III. Route optimization among a group of multihomed stub networks

In this paper, we propose an approach for coordinating the route control among a group of multihomed stub networks. We abstract each stub network as a "node". Thus, the network topologies we studied consist of (1) nodes representing the the stub networks and ISP edge routers connected to the stub networks, (2) edges representing access links of the stub networks and the paths between ISP edge routers taken by traffic between two stub networks. We call the traffic among the stub networks as "inhouse traffic" and the traffic between the stub networks and the rest of the Internet as "Internet traffic". We assume that the volume of inhouse traffic on a link of an ISP network or a link between two ISP networks is much smaller than the capacity of the link. Under this assumption, the quality of a path between two ISP edge routers in the topology is independent of the routing decisions of the stub networks. Thus we can abstract the path between two ISP edge routers as a link with given quality metrics, e.g. delay and loss rate considered in this paper. Fig. 2 shows a partial topology for a group of 4 stub networks, where we draw only the paths from stub network 2, 3 and 4 to stub network 1. In this paper, we assume the numbers of ISPs of stub networks are the same, but it is not required for our approach. Each stub network in the sample topology in Fig. 2 connects to 2 ISPs.

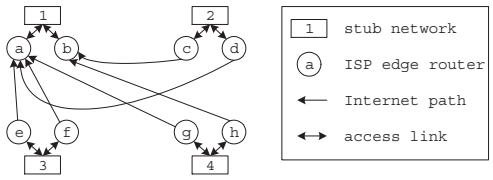

Fig. 2. Topology of 4 stub networks where each stub network has 2 ISPs

We formulate the problem as a variant of the optimal routing problem [5] as (1) to (5) and assume network traffic can be split arbitrarily using hashing based classification or other methods. Notations are listed in Tab. I. The objective is to minimize the sum of the "path delays" of inhouse traffic(item 1 of (1)) and the access link queuing delays of inhouse traffic(item 2) and Internet traffic(item 3) weighted by their traffic volume. The formulation can be understood as follows: the first item is for selecting a low delay path; the second

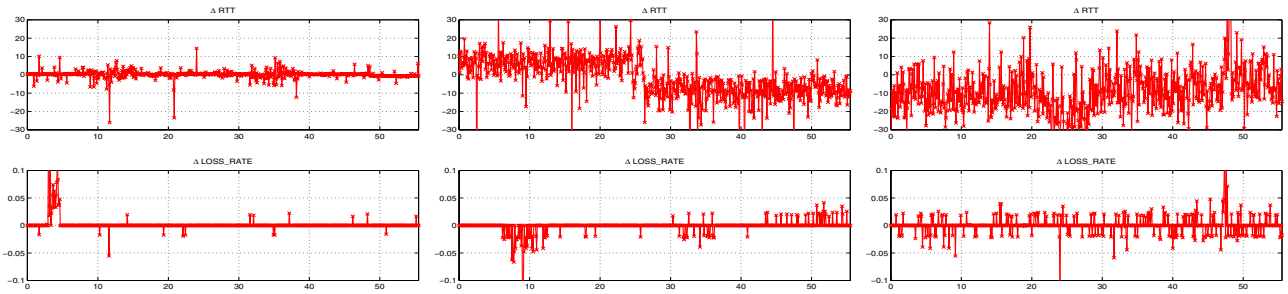

Fig. 1. Average RTT differences and loss rate differences of 3 pairs of alternate paths over 5 minute durations. (Y axis: ΔRTT in milliseconds and $\Delta loss\_rate$, X axis: time since the start of measurement in hours)

and the third item are for balancing load on access links. The output of the algorithm is x_{ijkl}, the fraction of inhouse traffic from i to j sent along path (i, j, k, l).

Symbol	Definition
N, K_n	stub networks in the topology and ISPs of stub network n
(i, j, k, l)	path from i to j via ISP k of i and ISP l of j
s_{ijkl}	= 1 (if path (i, j, k, l) exists); =0 (otherwise)
p_{ijkl}	one way delay of path (i, j, k, l)
x_{ijkl}	fraction of traffic from i to j sent along path (i, j, k, l)
d_{ij}	traffic demand from i to j
$u_{ik}(v_{ik})$	egress (ingress) Internet traffic volume of i via ISP k
$U_{ik}(V_{ik})$	total egress (ingress) traffic volume of i via ISP k
$q(x)$	queuing delay on a link, x is the load on the link

TABLE I
NOTATIONS

Minimize:

$$\sum_{i,j\in N,k\in K_i,l\in K_j:s_{ijkl}=1} x_{ijkl}\cdot d_{ij}\cdot p_{ijkl} + \sum_{i\in N,k\in K_i} U_{ik}\cdot q(U_{ik}) + \sum_{j\in N,l\in K_j} V_{jl}\cdot q(V_{jl}) \quad (1)$$

Subject to:

$$x_{ijkl} \geq 0, (i,j \in N, k \in K_i, l \in K_j : s_{ijkl} = 1) \quad (2)$$

$$\sum_{k\in K_i,l\in K_j:s_{ijkl}=1} x_{ijkl} = 1, (i,j \in N, i \neq j) \quad (3)$$

$$U_{ik} = \sum_{j\in N,l\in K_j:s_{ijkl}=1} x_{ijkl}\cdot d_{ij} + u_{ik}, \quad (i \in N, k \in N_i) \quad (4)$$

$$V_{jl} = \sum_{i\in N,k\in K_i:s_{ijkl}=1} x_{ijkl}\cdot d_{ij} + v_{jl}, \quad (j \in N, l \in K_j) \quad (5)$$

Our approach works as follows: The demand matrices of inhouse traffic and Internet traffic and path qualities are measured in a distributed manner. After a fixed period, say one minute, or when there are significant changes of the demand matrix or path qualities, the measured path characteristics and the demand information are exchanged among all the stub networks. After receiving the updated information, each stub network predicts the demand matrix and the path characteristics using a prediction model (see section IV-A for details), and calculates the optimal routing decision using an optimal routing algorithm for all the stub networks. The routing decision is adopted until next update. We also assume that the queuing delay function can be measured and is known to the optimization algorithm. As long as the queuing delay is a non-decreasing convex function, our algorithm can find an optimal solution [5]. When path qualities change too rapidly and become hard to predict, our approach uses the mean values over a longer time window to get an improvement in a statistical sense.

Packet losses are not considered in the above formulation. However, we can take loss rates into account by substituting p_{ijkl} with a "virtual delay function", $vdelay(p_{ijkl}, r_{ijkl})$, where r_{ijkl} represents the loss rate on the path. We leave the study of our approach on lossy networks as future work.

IV. EVALUATION

A. Simulation scenarios

We analyze the performance of our approach in two aspects:

1) Static analysis: we study the performance of our approach for a set of model-based random demand matrices on a set of randomly generated topologies with different path characteristics. In this analysis, we assume the demand matrix and path characteristics are static. This analysis gives an upper bound of the performance of our approach.
2) Dynamic analysis: We also study the performance of our algorithm with changing demand matrices. Specifically, we evaluate the performance of our algorithm with a time series of demand matrices generated from an Internet trace. In this analysis, the path characteristics are fixed. We study a simple demand prediction method, i.e. using the average demands of the last period as the prediction of demands for the next period. Empirical study [8] shows this simple prediction model is as good as other complex models for prediction period in order of minutes. We study optimization periods of 1, 2, 3,

5 and 10 minutes. We compare the performance of our approach using predicted demand matrices with the ideal performance, i.e. when the demand matrix of current period is available to the algorithm.

We compare the average end to end delay of inhouse traffic and the average queuing delay of Internet traffic with a static load-balancing approach that distributes egress inhouse traffic evenly on each link. We assume egress Internet traffic is distributed evenly on all egress links. Both our approach and the static load-balancing approach do not have control over the routes of ingress Internet traffic.

B. Topologies and path characteristics

A topology in our simulations is decided by three factors: number of stub networks (N), number of ISPs of each stub network (K) and type of paths between two stub networks. With regard to the last factor, depending on whether the stub networks multihome to the same set of ISPs, there are two types of topologies for inhouse traffic:

1) Symmetric topology: When all the stub networks multihome to the same set of ISPs, a path from a stub network i via ISP k to any other stub network in the topology, say j, must reach j via ISP k. Thus, the alternate paths from a stub network to another stub network in the topology are "parallel", they merge only inside the destination stub network.
2) Asymmetric topology: When the stub networks multihome to different sets of ISPs, the alternate paths from a stub network to another stub network in the topology are not necessarily "parallel". Two paths to a stub network may merge in one ISP of the destination stub network or in an AS between the two stub networks depending on the BGP relationship of the ASes in between.

We generate path characteristics as follows: we randomly map the stub networks onto 18 major cities of the United States. We generate "one way delay" of a path for inhouse traffic between two ISP edge routers by multiplying half of the RTT between the the two cities on the AT&T backbone [9] by a random factor, uniformly distributed from 1.1 to 1.5. The "one way delay" used here means the end to end delay minus the queuing delays on the access links of the stub networks. If more than one stub networks are mapped to the same city, we assigned the "one way delay" of a path for inhouse traffic between these stub networks as 1.5 millisecond multiplied by the above random factor.

C. Traffic demands:

Traffic demands of networks are usually not available to the public. Therefore, in our study, we first generate demand matrices using a simplified version of the Gravity model [10]. Gravity model is shown to work well in demand matrix estimation [11]. In our simulations, the expected volume of inhouse traffic is 50% of the total traffic on a link. We generate 20 random demand matrices for each set of N and K.

For dynamic analysis, we generate time series of demand matrices using the Leipzig-II trace from NLANR [12] website. We generate a time series of demand matrices by classifying packets with same source (or destination) IP addresses into one of a number of flows by a probability that is proportional to the expected volume of the flow.

D. Queuing delay models

We study the performance of our approach using two representative queuing models: 1) M/D/1 queuing model [13] that represents queues with Poisson packet arrivals; 2) P/M/1 queuing model [14] that represents queues with Pareto (one form of heavy-tailed distribution) packet arrivals. The queuing delay function of M/D/1 is: $q(x) = \frac{0.5}{\mu - x} + \frac{0.5}{\mu}$, where μ is the link capacity, x is the load on the link. In our simulations, we assume packet size is 1000 bytes. Because there is no close form of queuing delay function for P/M/1 queue, we use a piece-wise linear function to approximate the function according to the numerical result of [14] (with $\beta = 1.5$). Because both of these two models are defined in $x \in [0, 1)$ while the demand in our simulation may exceed the capacity, similar to previous work [15], we extend the queuing delay of the two models linearly for utilizations over 99%.

E. Implementation of the optimal routing algorithm

The nonlinear optimization problem can be solved using a linear programming approximation. In our evaluation, we use piece-wise linear function $f_{ik}(U_{ik})$ and $g_{jl}(V_{jl})$ to approximate the item $U_{ik} \cdot q(U_{ik})$ and $V_{jl} \cdot q(V_{jl})$ in (1).

F. Simulation Results

1) Static analysis: Fig. 3 shows the "performance improvement ratios" of our approach compared to the static load-balancing approach for topologies of 10 stub networks where each stub network has 2 ISPs. Here, we define "performance improvement ratio" of our approach compared compared to the static load-balancing approach as $(L_{lb} - L_{opt})/L_{lb}$, while L_{lb} is the average delay for the static load-balancing approach, L_{opt} is the average delay for our global optimization approach. The ratios are calculated for both the access link queuing delay of Internet traffic and end to end delay of inhouse traffic under different link utilizations. We fix a demand matrix and change the access link capacities to change the average utilizations. The four curves show the affects of queuing model and type of network topologies.

The main observations are:

1). The improvement for Pareto queuing model is more evident than Poisson queuing model. This is because the queuing delay of Poisson model is lower than the queuing delay of Pareto model. While our approach gets improvement mainly by exploiting path diversity for the Poisson queuing model case, it can get more improvement by balancing load on access links for the Pareto queuing model case.

2). The improvement for asymmetric topologies is more evident than symmetric topologies. This is because for symmetric topologies, inhouse traffic can be balanced on links using the static load-balancing.

3). The improvement is larger at higher average utilizations. This is because the queuing delay of both Poisson and Pareto

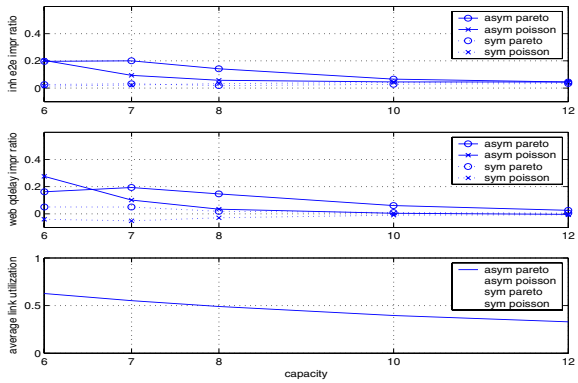

Fig. 3. Performance improvement ratios: topologies of 10 stub networks where each stub network has 2 ISPs, capacities are in Mbps

model is dramatically increased as utilization approaches 100%.

4). The average access link queuing delays of Internet traffic are also improved, except for symmetric networks under Poisson queuing model where the queuing delay is increased by 5% at most.

Results for more ISPs are similar, but as the number of ISPs increases, the improvement ratios become larger than the above 2-ISP case. The explanation is that more ISPs provide more opportunities for global optimization. Results for network of 20 stub networks and 4 stub networks are similar.

2) Dynamic analysis: Fig. 4 shows the average improvement ratios of end to end delay of inhouse traffic of our approach compared to the static load-balancing approach for asymmetric topologies with Pareto-arrival queuing model. Results for topologies of 9 different sizes are shown in groups of bars from left to right. The sizes of topologies can be represented by $xx - y$, where $xx = 04, 10, 20$, is the number of stub networks, $y = 2, 3, 4$, is the number of ISPs of each stub network. For topologies of a same size, from left to right, we plot the improvement ratios for the ideal optimization and optimizations using predicted demand matrices with optimization periods of 1, 2, 3, 5 and 10 minutes. In this set of simulations, the resulting average link utilization is about 50%.

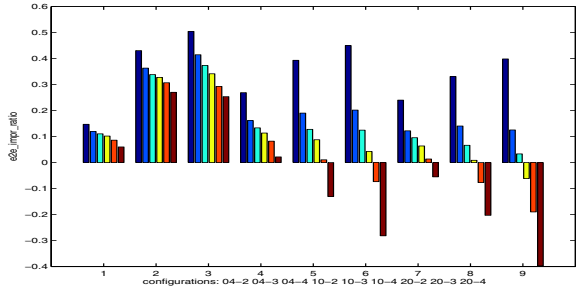

Fig. 4. Inhouse traffic end to end delay improvement

The results show that the performance improvements are larger over shorter periods of optimization. Shorter periods enable more accurate prediction of demand and path characteristics and hence provide larger performance gains. We also found that as the number of stub networks increases, the performance gains become smaller. This is primarily due to the artifact of how the demand matrices are generated in our simulations. In our simulations, the total expected volume of inhouse traffic on a link is fixed, as the number of stub networks increases, the volume of traffic from a stub network to another becomes smaller making the demand less predictable. As in earlier experiments, symmetric topologies and Poisson-arrival queuing model show smaller performance gains.

V. Conclusions and Future Work

In this work, we proposed an approach for coordinating the route control of traffic among a group of multihomed stub networks. We studied the static and dynamic performance of our approach using simulations. The results show that our approach can significantly improve the performance of route control among a group of multihomed stub networks when the access links are not over-provisioned. We also conducted measurement of the qualities of alternate paths through multihoming. The results show that dynamic route control of both small and large time scales can improve the performance of multihomed stub network.

Further analysis of the effect of the dynamics of real world demand matrices and the variation of path qualities is desirable.

References

[1] NetVmg, http://www.davidwriter.com/netvmgw/, Aug. 2005.
[2] A. Akella, B. Maggs, S. Seshan, A. Shaikh, and R. Sitaraman, "A measurement-based analysis of multihoming," in *Proceedings of SIGCOMM '03*. ACM Press, 2003, pp. 353–364.
[3] F. Guo, J. Chen, W. Li, and T. Chiueh, "Experiences in building a multihoming load balancing system."
[4] D. K. Goldenberg, L. Qiu, H. Xie, Y. R. Yang, and Y. Zhang, "Optimizing cost and performance for multihoming," in *Proceedings of SIGCOMM '04*. ACM Press, 2004, pp. 79–92.
[5] D. P. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Prentice-Hall, 1992.
[6] The Route Views Project, http://www.routeviews.org, Jun. 2004.
[7] Scriptroute, Jun. 2004, http://www.cs.washington.edu/research/networking/scriptroute/.
[8] Y. Qiao, J. Skicewicz, and P. A. Dinda, "An empirical study of the multiscale predictability of network traffic." in *HPDC*, 2004, pp. 66–76.
[9] AT&T U.S. Network Latency, Feb. 2005, http://ipnetwork.bgtmo.ip.att.net/pws/network_delay.html.
[10] J. Kowalski and B. Warfield, "Modeling traffic demand between nodes in a telecommunications network."
[11] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg, "Fast accurate computation of large-scale ip traffic matrices from link loads," in *SIGMETRICS '03*. ACM Press, 2003, pp. 206–217.
[12] NLANR/MNA, http://mna.nlanr.net/, Jun. 2004.
[13] D. Gross and C. Harris, *Fundamentals of Queuing Theory*, 3rd ed. John Wiley, 1998.
[14] C. M. Harris, P. H. Brill, and M. J. Fischer, "Internet-type queues with power-tailed interarrival times and computational methods for their analysis," *INFORMS J. on Computing*, vol. 12, no. 4, pp. 261–271, 2000.
[15] L. Qiu, Y. R. Yang, Y. Zhang, and S. Shenker, "On selfish routing in internet-like environments." in *SIGCOMM*, 2003, pp. 151–162.