Available online at www.sciencedirect.com
ScienceDirect
Computer Communications 30 (2007) 351–357
computer communications
www.elsevier.com/locate/comcom

On-the-fly TCP path selection algorithm in access link load balancing

Ying-Dar Lin, Shih-Chiang Tsao *, Un-Pio Leong

Department of Computer and Information Science, National Chiao Tung University, HsinChu, Taiwan

Received 17 April 2006; received in revised form 31 August 2006; accepted 2 September 2006
Available online 2 October 2006

Abstract

Many enterprises install multiple access links for fault tolerance or bandwidth enlargement. Dispatching connections through good links is the ultimate goal in utilizing multiple access links. The traditional dispatching method is only based on the condition of the access links to ISPs. It may achieve fair utilization on the access links but poor performance on connection throughput. In this work, we propose a novel approach to maximize the per-connection end-to-end throughput by the on-the-fly round trip time (RTT) probing mechanism. The end-to-end RTTs through all possible links are probed by duplicating the SYN packet during the three-way handshaking stage of a TCP connection. The experiment results show that the ratio to choose the best outgoing access link is 79% on the average. If the second best link is chosen, it is usually very close to the best, thus averagely achieving 94% of the maximum possible throughput. The ratio of the traditional round-robin (RR) algorithm is only 35%, and the link selected by RR algorithm could provide 69% of throughput.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Path selection; RTT; TCP three-way handshaking

1. Introduction

Today, the Internet is a major medium for enterprises to collect or provide information. To connect to the Internet, the enterprise may rent the *access link* from the Internet Service Provider (ISP). To ensure the availability of accessing to the Internet, the enterprise has to rent multiple *access link*s. When one link breaks down, the enterprise may still connect with the Internet by using the other links. It is suggested that an enterprise should rent the links from multiple ISPs [1–3], because by following the suggestion, the enterprise would connect to the Internet even when one of the ISPs breaks down.

Under such a topology, a network with multiple ISP links, a device usually called the load balancer is necessary to intelligently select a link for an establishing connection. The load balancer in this work focuses on the outbound traffic handling [1], which is different from that of handling the inbound traffic [4]. Generally speaking, the load balancer prefers to assign the link with the lowest utilization for an establishing outbound connection because such a link seems to provide the new connection the most available bandwidth.

However, the remote target site of a connection is usually far from the local host. The path from the local host to the remote site consists of multiple links. The access link is only the heading link of the end-to-end path. Thus, the policy preferring lowest-utilization link would not select the path with the most available bandwidth. For example, when a user residing in Taiwan surfs on a website located in Germany, the device should select the access link heading the *fastest* path to Germany. Unfortunately, the traditional load balancer only spends effort on determining the link with the lowest loading, but not the path with the most available bandwidth to the destination. Obviously, the traditional balancer strays from its aim.

Only a few load balancers are aware of the stray and collect more path information to improve the link selection. They

* Corresponding author. Tel: +886 3 5731899; fax: +886 3 5721490.
E-mail addresses: ydlin@cs.nctu.edu.tw (Y.-D. Lin), weafon@cs.nctu.edu.tw (S.-C. Tsao).

0140-3664/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2006.09.001

retrieve the information beyond the next hop by relying on SNMP or ICMP. However, the retrieve by SNMP requires authorization, while the ICMP packet is often filtered out for security concerns. A path selection mechanism (PSM) proposed in [5] measures the conditions over the end-to-end path during the connection with *specific-format* packets, which implies the PSM has to *modify the user protocol* and cannot make the selection before a connection is established.

This work aims to select an access link in a load balancer before a connection is established while the selected link provides the optimal path for a TCP connection. The optimal path indicates the path providing the most available bandwidth for a TCP connection. Generally, a path with the lowest loss ratio and the shortest RTT would be such an optimal path since the mean rate of a TCP connection is determined on the packet loss ratio and RTT of a path [6]. However, the loss ratio is not considered in this work because the loss ratio in the Internet is variant and hard to estimate upon establishing a connection. In fact, it is difficult even when the connection is running. Thus, the same ratio of packet losses in all access WAN links is assumed in this work although the assumption may not be true in the real Internet. Such an assumption sometimes leads to the slight inaccuracy of the best selection, as shown in the evaluation later.

To select the optimal path, a novel *on-the-fly RTT probing (OFRP) mechanism* is proposed in Section 3 and implemented into the NetBSD [7] kernel. By the probing strategy, the load balancer can select a link leading the shortest-RTT Internet path to the destination. The strategy probes the RTT upon establishing a connection. The TCP SYN segment of this establishing connection is employed as the probing packet. It is duplicated and sent concurrently along all access links. Next, the access link where the SYN/ACK segment returns the earliest is regarded as the link heading the shortest-RTT path. Such an *on-the-fly RTT probing mechanism*, therefore, provides a TCP connection the optimal path under the assumption that the access WAN links are equal to the packet loss ratio. Notably, although the mechanism is only suitable for the TCP connections, it is still desirable to be employed in the load balancer. After all, the TCP connections still dominates the Internet traffic [8].

For the rest of this work, the organization is as follows: Section 2 reveals the problem of the traditional load balancer. Section 3 describes our on-the-fly RTT probing mechanism and the concern of the implementation into the NetBSD kernel. The performance evaluation in Section 4 shows how well this link load balancing approach behaves. Finally, Section 5 wraps up with the conclusion and recommended future work.

2. Problems in traditional link selection

A link with the lowest loading is preferred for establishing a new outbound connection in load balancers [9,10]. A simple way to select the lowest-loading link is weighted round-robin. The next selection in round-robin way just is the lowest one since the loading between links is balanced. Generally speaking, the policy to select a link with the lowest loading is easy to implement. However, the link selected by such a policy does not ensure providing a path with the most available bandwidth to the remote site.

New policies adopted in [10] further collect or probe the information beyond the next hop to select a link that is the head of the links composed of the optimal path. The following exposes the problems in these policies in terms of *used protocol*, *collection distance*, and *measured moment*:

2.1. Used protocol

SNMP and ICMP are popular protocols to collect the condition on path [11]. Unfortunately, because ICMP is used for hacking and probing security holes, most firewalls today filter out ICMP packets for security concerns. Further, even without being filtered out, ICMP packets may gain a low priority when being forwarded or replied, affecting the accuracy of the measurement. SNMP is another protocol for collection. By it, the load balancer can directly retrieve the statistic information of the hops on the path. However, accessing a hop by SNMP requires authorization for security concerns. It is impossible for a local load balancer to own the authorization to access all hops between the local host and the destination.

2.2. Collection distance

The traditional load balancing algorithms fail for only collecting the condition of the access link. The ideal dis-

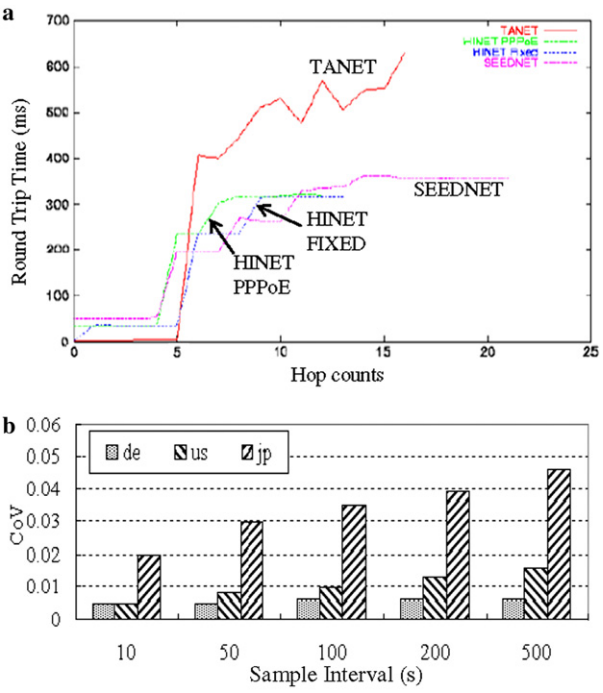

Fig. 1. (a) The RTT probing results of each hop for four links (b) The coefficient-of-variance (CoV) of RTT measured in different sample intervals.

tance of the collection should cover the overall path. Fig. 1(a) shows the RTT probing results of each hop in four connection paths provided by four WAN links rented from TANET [12], HINET [13] (PPPoE), HINET (Fixed IP), and SEEDNET [14], respectively. Fig. 1(a) exposes the inaccuracy selection if only probing the near-end hops. The whole path is broken down into a sequence of hops and the time spent between hops is displayed. Obviously, when considering only the RTT of the near hops and given that the packet loss ratio is equivalent between all links, the access link of TANET is the best link for a TCP connection, which makes the TCP connection send packets with a highest rate. However, when considering the RTT of the end-to-end path, the path through the TANET link has the largest RTT and provides the worst transmitting throughput for a TCP connection.

2.3. Measured moment

To collect the condition of the whole path, the only moment is upon establishing a connection, because the destination of a connection is unexpected and given at real time. It is impossible to collect *in advance* the conditions of the paths to each destination in the Internet, which is another reason, besides authorization, that limits the off-line collections to retrieve the conditions from near hops.

3. On-the-fly RTT probing mechanism

This work proposes an on-the-fly RTT probing (OFRP) mechanism to assist WAN load balancer (WLB) in selecting the link heading the path with the shortest RTT for an establishing TCP connection.

3.1. RTT is short-time stable

The following reveals the RTT is short-time stable so as to be probed on-the-fly. The RTT in the Internet is dynamic but not violently oscillatory on a path as shown in Fig. 1(b). The RTTs between local and the three websites are measured through the four links in 4000 s. The result reveals the largest Coefficient-of-variance (CoV) is only 0.045 even in a time scale of 500 s. That is, the value of RTT shakes falls in a 4.5% range of the mean. Obviously, the RTT of the whole session is stable. Based on such a stable character, it is believed that the average RTT can be represented only by the value averaged from a few samples. Similar observations could be found in [15,16].

3.2. On-the-fly probing on RTT

The subsection describes how the OFRP mechanism on-the-fly probes the RTTs of multiple paths. The mechanism employs the three-way handshaking of TCP, the necessary procedure for establishing a TCP connection. The handshaking procedure states that to establish a connection the client should send a SYN packet to the server and wait the server for the SYN/ACK packet. After receiving the SYN/ACK packet, the client responds a SYN/ACK packet to the server and the connection is established. During the procedure, the RTT of the path can be retrieved by measuring the difference on timestamp between the SYN packet and the first SYN/ACK packet.

However, it is inefficient to sequentially retrieve the RTTs of multiple paths. To retrieve these RTTs *simultaneously*, WLB duplicates the SYN packet and then sends out these copies via different WAN links, respectively, as shown in Fig. 2. Each returned SYN/ACK packet represents the probing results of each path via each WAN link, respectively. The SYN/ACK packet returning the earliest represents the corresponding path has the shortest RTT. Then, the link heading the path is selected. An ACK packet is replied through the link to establish the TCP connection while RST packets are sent via other links to close out the probing connection. Finally, all the duplicated handshaking are closed, and the selected connection behaves like a normal connection.

The overhead of such an on-the-fly RTT probing is that the duplication of the SYN packet. It increases the load of the destination server's accepting state, but no overhead for the rest of data transmission.

Although WLB sends multiple SYN packets parallel to a server, we believe it could not be treated as a potential denial of a service attack. First, each SYN packet has its individual IP address. They are sent out from individual access links and an individual IP is allocated for each link by ISP. Second, the useless connections are terminated by sending the RST packets as soon as the shortest one is decided.

3.3. Packet loss ratio

Actually, both RTT and packet loss ratio are necessary to estimate the bandwidth occupied by a TCP connection on a path [6]. Unfortunately, the packet loss ratio in the Internet is variant and hard to gather a stable estimation even by a long-term measurement. In other words, in establishing connection, it is impossible in the real time to obtain the loss ratio of the path. Thus, here we assume the Internet paths are equal in the loss ratio although it may be not true in the real world. Such assumption implies that the

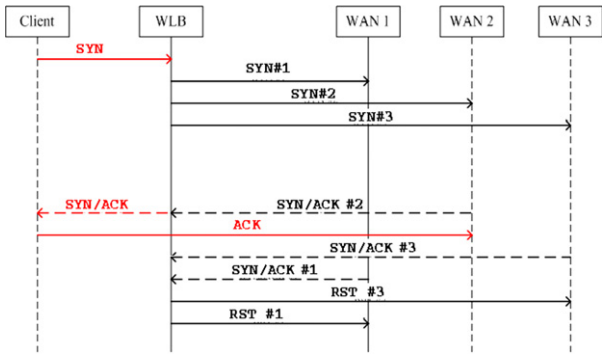

Fig. 2. On-the-fly RTT probing mechanism.

RTT is the only determinant in our TCP path selection. In fact, it brings the selection slight inaccuracy, which was shown in the evaluation later (Section 4.4).

3.4. Implementation issues

The following discusses two concerns of the implementation of the WLB with the OFRP mechanism in the NetBSD kernel.

WHERE. This work implements the WLB into the Network Address Translation (NAT) module because two operations, *IP modification* and *stateful redirection*, done in the NAT are required by WLB. The WLB redirects the traffic to the selected outgoing access link; thus, the associated source IP address of outgoing data packets must be modified to the IP of the selected link. Otherwise, their ACKs cannot return from the selected link. Moreover, after selecting the particular link, the WLB should redirect the following packets in one connection to the same link. Thus, the stateful redirection is necessary for WLB.

HOW. For a received SYN packet, NAT handles three major operations, including (a) modifying the IP of the SYN (b) keeping the state (c) sending out the SYN. Three modifications are required to integrate OFRP-embedded WLB into NAT. First, the SYN duplication should perform before the operation (a). Second, the operation (b) is delayed until receiving the first SYN/ACK, the time WLB can know which link is selected. Third, on receiving the SYN/ACK, as stated in Section 3.2, the WLB needs to send ACK packet through the selected link and RST packets through other links.

4. Evaluation

4.1. Testbed configuration

The section demonstrates the OFRP-embedded WLB (OFRP-LB) can select the link heading the path with the maximum available bandwidth for a TCP connection by the experiments run in Internet. The testbed was built as shown in Fig. 3 and there are three WAN links providing the load balancer accessing the Internet. To establish a connection, one of the three links is chosen according to the probing results of the OFRP mechanism implemented in the device. In the following tests, there are three destinations including ftp.de.freebsd.org (de), ftp.freebsd.org (us), and ftp.jp.freebsd.org (jp), which are located in Germany, US, and Japan, respectively. For each destination, in one testing round we first forecast the best link by OFRP-LB and then retrieve a file with 8 MB through the three links, respectively. By comparing their mean throughput, we can verify whether the link selected by OFRP-LB is the best one or not. Each round of this test is performed repeatedly every 10 min. At last, the result is compared to the traditional round-robin (RR) link selection algorithm.

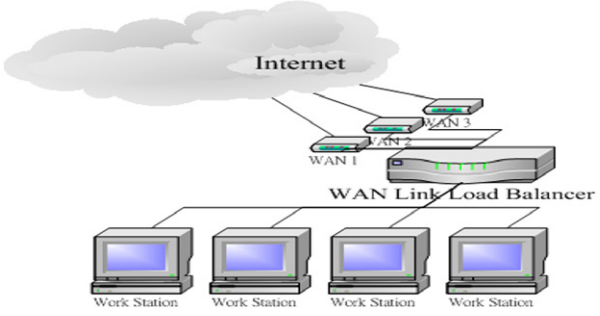

Fig. 3. WAN link load balancing testbed configuration.

4.2. The link selected by OFRP-LB

Fig. 4(a) displays the throughput results of 140-rounds transferring files from ftp.de.freebsd.org through the three WAN links, TANET, HINET PPPoE, and SEEDNET PPPoE. In the figure, for each round, the link selected by OFRP-LB is highlighted with the curve *WLB Selected*. Obviously, the curve always overlaps the result with the highest throughput, indicating OFRP-LB can select the link providing the maximum throughput for TCP connections. The similar results are displayed in Figs. 4(b) and (c), where the destinations are ftp.freebsd.org and ftp.jp.freebsd.org, respectively. Note that Figs. 4(a)–(c) also reveal that the Internet status does change from time to time. According to our observation, the unusual large RTT seriously degrades the throughput provided by one links.

4.3. The accuracy of OFRP-LB

Next, for each destination, the hit ratio of OFRP-LB is averaged from 100 rounds of selecting results, where the hit represents the selected link providing the highest bandwidth for a TCP connection among the three links and thus

$$\text{the hit ratio} = \frac{\text{the number of the hit rounds}}{\text{the number of the total rounds}}.$$

Fig. 5 shows OFRP-LB determines the best links for connecting to de, us, and jp with a hit ratio of 89.78%, 71.32%, and 76.64%, respectively. Compared to the result of the RR link selection, OFRP-LB provides a higher hit ratio ($79\% > 35\%$) in the determination of the best access link.

4.4. The comparison between selected and maximum links

Instead of the hit ratio of selection, the following reveals that the link selected by OFRP-LB provides the bandwidth similar to the maximum bandwidth provided by one of the three links. In Fig. 6, there are four bars for each destination. The bars *Maximum* and *Minimum* indicate the mean throughput averaged from the maximum and minimum throughput in each round, respectively. The bar *Selected* and *Round-Robin* represent the mean throughputs provided by the links selected by OFRP-LB and the RR link selec-

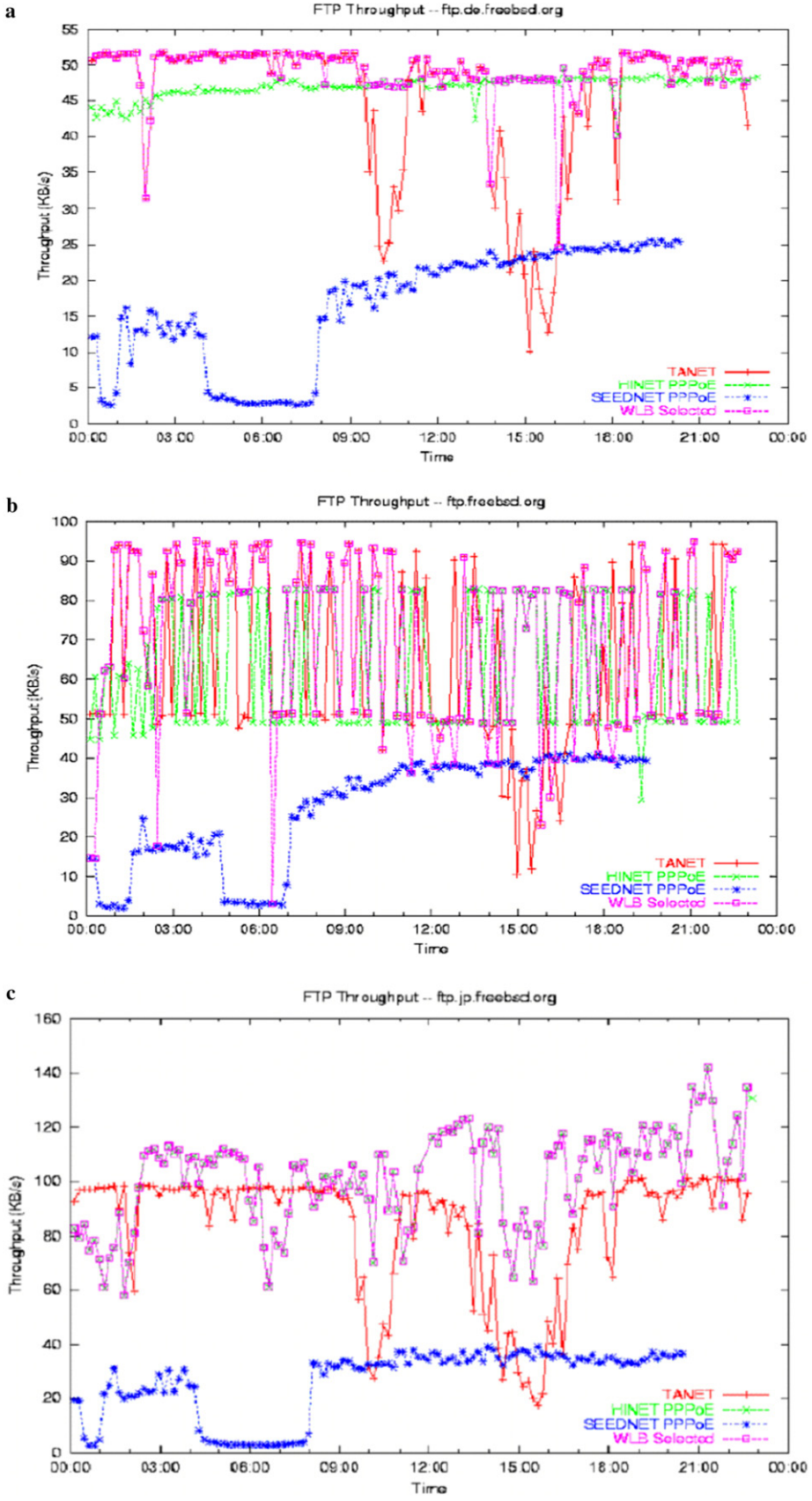

Fig. 4. (a) FTP Throughput – ftp.de.freebsd.org (b) FTP Throughput – ftp.freebsd.org (c) FTP Throughput – ftp.jp.freebsd.org.

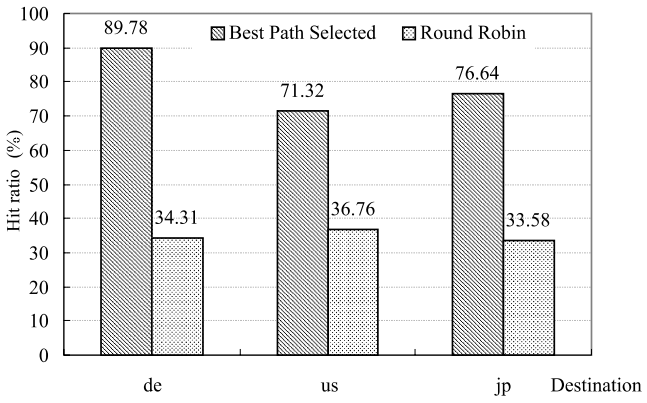

Fig. 5. The hit ratio of TCP path selection algorithm.

tion algorithm. The chart reveals that the selected link to the destination *de* provides 49 KBps bandwidth, which equals to 98% of the maximum bandwidth, 50 KBps, the best link could provide. It is far better than the link selected by the RR, which only provides 36/50, or 72%, of the maximum bandwidth.

The result in Fig. 6 outperforms than that in Fig. 5, revealing an interesting phenomenon. That is, even on the case that the OFRP-LB does not correctly forecast the best link, the connection through over the link selected by the OFRP-LB still obtains 94% of bandwidth that the best path can provide on average. It implies for the incorrect forecasting cases, the conditions between the best link and our selected link is similar.

Since the link selected by the OFRP-LB already provides the bandwidth near that of the best path, it is demonstrated that the OFRP mechanism is enough to select a link for TCP even as it considers only one sample of RTT. Such a result also reveals that ignoring the packet loss ratio in the selection merely brings slight inaccuracy. Assume that our WLB will always select the best link if additionally considering the loss ratio. Then, the best link provides at most 6% more bandwidth than that selected by the WLB without considering the loss ratio.

5. Conclusion and future work

The goal of link load balancing is to select the best link for establishing a connection. The best link should provide the best quality of network conditions over an end-to-end path. This work focuses on the best link selection for TCP connection and proposes an on-the-fly RTT probing mechanism. The mechanism duplicates the SYN packet during the three-way handshaking stage of a TCP connection to simultaneously retrieve the end-to-end RTT through each access link. *The link heading the shortest-RTT Internet path* is selected for establishing the connection since the short path is expected to provide a TCP connection high available bandwidth.

The result of our experiments on the NetBSD implementation shows that the on-the-fly probing mechanism indeed picks a link heading the shortest path to provide the highest available bandwidth with an average probability close to 79%. For the cases not selecting the best path (21%), the further observation reveals that only a slight difference exists in per-connection throughput between the best path and the selected path. The path headed by our selected link averagely provides 94% of per-connection throughput the best path could provide. Such a result demonstrates that an easy-to-implement on-the-fly RTT probing mechanism is enough to select an ideal link even as it ignores the packet loss ratio and takes only one measured sample of RTT to select a link.

According to our implementation experience, the on-the-fly RTT probing mechanism proposed in this work can be implemented at the gateway without modifying the client protocol. Thus, it is easy for deployment and is compatible with existing applications. In the future, to improve the accuracy of the best TCP path selection, we will continue to study how to obtain a suitable estimation

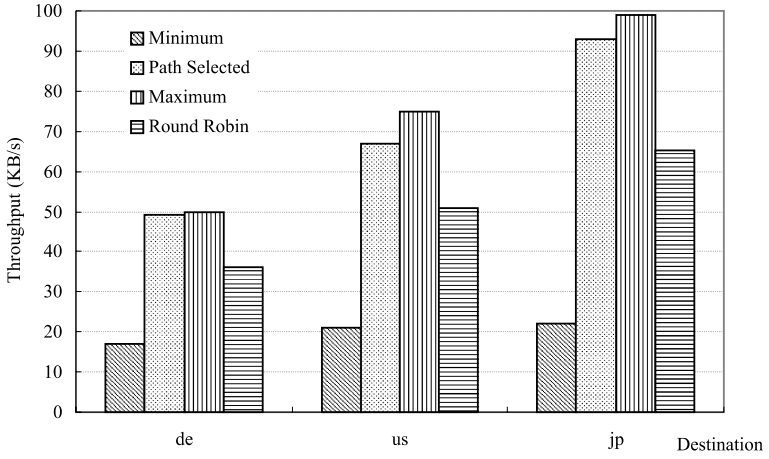

Fig. 6. The selected throughput compared to the minimum and maximum possible throughput utilization.

on the packet loss ratio. Besides, since the WAN link load balancing gateway requires to access information at the transport layer, we will further discuss the possible conflict if IP SEC is implemented.

References

[1] Fanglu Guo, Jiawu Chen, Wei Li, Tzi-cker Chiueh, Experiences in building a multihoming load balancing system, in: Proceedings of INFOCOM 2004, vol. 2, 2004, pp. 1241–1251.

[2] T. Bates, Y. Rekhter, Scalable Support for Multi-homed Multi-provider Connectivity, RFC 2260, 1998.

[3] F5 Network, Conquering Multi-Homed ISP Link Challenges, White Paper, <http://www.f5.com/solutions/technology/pdfs/lc_multihome_wp.pdf/>.

[4] A. Mihailovic, G. Leijonhufvud, T. Suihko, Providing multi-homing support in IP access networks, in: 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, vol. 2, 2002, pp. 540–544.

[5] Kashihara, S, etc. Path selection using active measurement in multi-homed wireless networks, in: Proceedings of 2004 International Symposium on Application and the Internet, 2004, pp. 273–276.

[6] J. Padhye, V. Firoiu, D. Towsley, J. Kurose, Modeling TCP throughput: a simple model and its Empirical Validation, in: Proceedings of ACM SIGCOMM'98, September 1998.

[7] The NetBSD Project, <http://www.netbsd.org//>.

[8] Fraleigh, c, etc. Packet-level traffic measurements from the Sprint IP backbone, in: IEEE Network, vol. 17, No. 6, November 2003, pp. 6–16.

[9] Radware LinkProof, Internet Link Application Switching. <http://www.radware.com/content/products/lp/default.asp/>.

[10] A Link Load Balancing Solution for Multi-Homed Networks. <http://www.f5.com/solutions/tech/multi_homing.html/>.

[11] Load Balancing, Radware Ltd., United State Patent, US006249801B1, January 19, 2001.

[12] TANET, <http://www.edu.tw/EDU_WEB/Web/MOECC/home.htm/>.

[13] HINET, <http://www.hinet.net/english/index.htm/>.

[14] SEEDNET, <http://www.digitalunited.com/>.

[15] F. Chatte, B. Ducourthial, S.I. Niculescu, Robustness issues of fluid approximations for congestion detection in best effort networks, Seventh International Symposium on ISCC 2002, July 2002, pp. 861–866.

[16] Tsunyi Tuan, Kihong Park, Multiple time scale redundancy control for QoS-sensitive transport of real-time traffic, in: Proceedings of Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOMM 2000, vol.3, March 2000, pp. 1683–1692.

Ying-Dar Lin received the Bachelor's degree in Computer Science and Information Engineering from National Taiwan University in 1988, and the M.S. and Ph.D. degrees in Computer Science from the University of California, Los Angeles (UCLA) in 1990 and 1993, respectively. He joined the faculty of the Department of Computer and Information Science at National Chiao Tung University (NCTU) in August 1993 and is Professor since 1999. From 2005, he is the director of the newly established graduate Institute of Network Engineering. He is also the founder and director of Network Benchmarking Lab (NBL), co-hosted by Industrial Technology Research Institute (ITRI) and NCTU since 2002, which reviews the functionality, performance, conformance, and interoperability of networking products ranging from switch, router, WLAN, to network and content security, and VoIP. His research interests include design, analysis, implementation and benchmarking of network protocols and algorithms, wire-speed switching and routing, quality of services, network security, content networking, and embedded hardware software co-design. He can be reached at ydlin@cs.nctu.edu.tw.

Shih-Chiang Tsao received the B.S. and the M.S. in Computer & Information Science from National Chiao Tung University, Hsinchu, Taiwan, in 1997 and 1999, respectively. He worked as an Associate Researcher in Chung-Hwa Telecom from 1999 to 2003, mainly to capture and analyze switch performance. He is currently pursuing a Ph.D. in Computer Science at National Chiao Tung University and advised by Dr. Ying-Dar Lin. His research interests include TCP-friendly congestion control algorithms, fair-queuing algorithms, and Web QoS.

Un-Pio Leong received the B.S. in Computer Science and Information Engineering and the M.S. in Computer & Information Science from National Chiao Tung University, Hsinchu, Taiwan, in 1997 and 2003, respectively.