An Adaptive Flow-level Load Control Scheme for Multipath Forwarding *

Youngseok Lee and Yanghee Choi
E-mail : {yslee, yhchoi}@mmlab.snu.ac.kr

School of Computer Science and Engineering, Seoul National University

Abstract. Compared with the traditional single path routing model, multipath routing increases total network utilization and end-to-end performance. When disseminating traffic into multiple paths, routers should adaptively allocate flows to each path in order to achieve load balancing among multiple paths, as most IP flows are short-lived and the flow size is not normally distributed. Moreover, routers should distribute packet streams belonging to a flow into the same next-hop not to cause end-to-end performance degradation. This paper proposes an adaptive multipath load control method using a flow classifier which detects long-lived flows through the flow characteristics of the duration and the size. By dividing flows into long-lived and short-lived, congestion from the bursty transient flows may be avoided. It is shown by simulation experiments with the real packet trace that the proposed algorithm adaptively controls the load of multiple paths satisfying the given load ratio, and the minimal per-flow states at routers can be maintained by aggregating flows with the destination network prefix.
Keywords: Flow, load control, multipath

1 Introduction

A router capable of multipath routing maintains multiple next-hop nodes for the same destination in its routing table. Multipath routing provides increased bandwidth and enhances the utilization of network resources more than the traditional Internet routing mechanism based on the single shortest path algorithm.

Multipath routing has been incorporated in several routing protocols. The best-known one is the Equal-Cost Multi-Path(ECMP) routing. This is explicitly supported by Open Shortest Path First(OSPF) and Intermediate System to Intermediate System(IS-IS). Some router implementations allow equal-cost multipath for Routing Information Protocol(RIP). In the Multi-Protocol Label Switching(MPLS) network, where IP datagrams are switched by looking up the fixed-size label, paths between an ingress router and an egress router are explicitly set up by Explicitly Routed Label Distribution Protocol(ER-LDP) or

* This work was supported in part by the Brain Korea 21 project of Ministry of Education, in part by the National Research Laboratory project of Ministry of Science and Technology, 2001, Korea.

the Resource ReSerVation Protocol(RSVP). Therefore, multiple explicit Label Switched Paths(LSPs) between an ingress router and an egress router can be set up and there can be even non-shortest paths for multipath routing.

When forwarding packets to multiple paths, routers should have an adaptive load control function for load balancing across parallel paths in order to support dynamic traffic behaviors and varying link/path characteristics(available bandwidth, delay, and packet loss rate). Otherwise, some of multiple paths may experience significant congestion due to the high traffic load.

In this paper, we propose a simple flow classifier based algorithm as the flow-aware adaptive multipath load control scheme. The proposed algorithm has the following features.

- Flow-level load control of multiple paths when the load ratio for each path is given: The input traffic can be split to satisfy the pre-defined load ratio of each path in a flow-level multipath forwarding mode. The sequence of IP packet streams should be maintained within a flow. Otherwise, the receiver must handle out-of-order packet arrivals with a large buffer, and end-to-end performance will be degraded.
- Minimal per-flow states: The number of per-flow states retained by a router should be as small as possible.
- Differentiation between long-lived flows and short-lived ones: [3] suggests that long-lived flows have less bursty arrival characteristics than short-lived flows. The bursty transient flows can abruptly increase the queue length at routers, causing packet losses.

The organization of this paper is as follows. In Section 2, the related work for the multipath and traffic engineering is explained. Section 3 presents the flow-level adaptive load control problem in multipath forwarding. Then, Section 4 describes the proposed flow-level load control algorithm. The results of the performance evaluation are discussed in Section 5 , and the conclusion and future work are given in Section 6.

2 Related Work

There have been many studies on multipath routing. [4] proposes a multipath forwarding extension scheme for the distance vector and the link state routing protocol. In [7], Quality-of-Service(QoS) routing via multiple paths for the time constraint is proposed when the bandwidth can be reserved, assuming all the re-ordered packets are recovered by the optimal buffer at the receiver, which causes the overhead of the dynamic buffer adjustment at the receiver. In connection-oriented networks, [8] has analyzed the performance of multipath routing algorithms and has shown that the connection establishment time for multipath reservation is significantly lowered. [9] has proposed a dynamic multipath routing algorithm in connection-oriented networks, where the shortest path is used under light traffic conditions and multiple paths are utilized as the shortest path becomes congested.

To avoid the negative effects by the bursty short-lived flows, the enhanced routing scheme separating long-lived and short-lived flows is proposed in [10] where long-lived flows are dynamically routed whereas transient flows are forwarded on the pre-provisioned paths. However, the flow trigger is considered only under the static network provisioning policy. In [12], a hashing-based load control method without flow states is proposed, but the load adaptation scheme for the dynamic network and traffic behavior is not well presented. In [11], it is shown that the quality of services can be enhanced by dividing the transport-level flows into UDP and TCP flows. Yet, it does not consider the aggregated flows.

3 The Flow-level Load Control Problem

In this section, we examine how packet-level multipath forwarding may degrade the end-to-end throughput and why the adaptive flow-level load control method should be devised for multipath forwarding.

3.1 Negative Impacts on End-to-End Performance by Packet-level Multipath Forwarding

Packet-level multipath forwarding in a round-robin fashion may cause the end-to-end performance degradation.

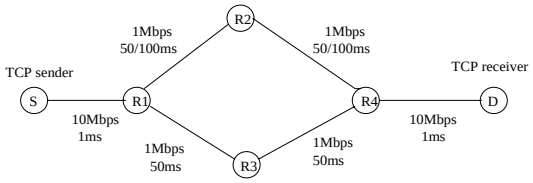
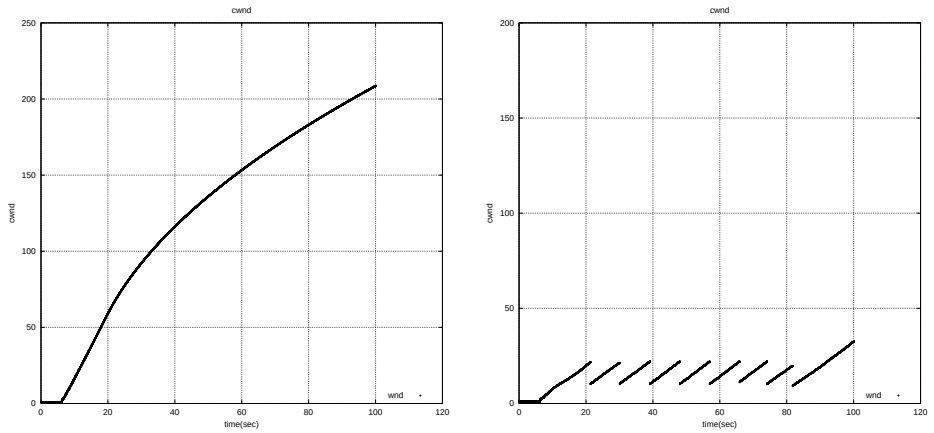


Fig. 1. The simulation topology

When the router $R1$ distributes incoming packets destined for D to two next-hops($R2$ and $R3$) concurrently(Fig. 1), the effect of the different delays on TCP performance is illustrated in Fig. 2. Fig. 2-(a) represents the case where both the upper path($R1 - R2 - R4$) and the lower path($R1 - R3 - R4$) are set to 100 ms, and S sends packets to D after opening an FTP connection[1]. Fig. 2-(b) is for the same FTP connection run under different delays(the upper path set to 200 ms). In Fig. 2-(b), the congestion window(*cwnd*) at S periodically decreases by half due to fast retransmit and fast recovery algorithms, resulting in the poor TCP throughput. When two paths have different delays, packets with higher sequence number may arrive at the receiver too early, causing the

[1] This simulation was tested for TCP Reno with NS-2[13]



(a) Same link delays (b) Different link delays

Fig. 2. TCP congestion window behavior under different path delays in packet-level multipath forwarding

receiver to send duplicate ACKs. After receiving three duplicate ACKs, the sender retransmits the late arrived packet again and reduces *cwnd*. In addition to the three duplicate ACK problem, the increasing speed of *cwnd* is slow because the ACKs with lower sequence numbers, which may arrive at the sender later than the ACKs with higher sequence numbers, are ignored.

3.2 Skewed Flow Characteristics

Most IP flows[2] are shown to be short-lived and small, whereas a few ones have long duration and large traffic loads, dominating the total traffic load in a link or path[2]. Hence, we examine the load balancing condition in general flow-level multipath forwarding.

Assuming that packet arrivals are modeled as packet trains[5], multiple paths are identical, and the packet size is normally distributed, then the flow-level round-robin load balancing can be explained by the following lemma, which is defined for load balancing by multiple identical servers in [6].

Lemma 1. *(Flow-level Round-Robin Load Balancing): Let l_i be a random variable describing the total delivery time required for all the flows mapped to a given path P_i. Let r' be a random variable of the delivery time for a flow, N be the number of packets, and N' be the number of flows in the batch or train. If N' flows are assigned to m paths in a round-robin manner, then the square of the*

[2] IP flows are defined by packet arrivals satisfying the end point specification(network addresses, transport protocol, and application port) within a time interval.

coefficient of variation of l_i is given by

$$CV[l_i]^2 = \left(\frac{m}{N'}\right) CV[r']^2 \tag{1}$$

and hence, when r' has finite variance and $N' \approx N$

$$\lim_{N\to\infty} CV[l_i] = 0. \tag{2}$$

From the above lemma it is concluded that for sufficiently large packet train size, the loads in a multiple path set are balanced if the coefficient of the variation of this normal distribution tends to zero.

N' and r' are dependent on the flow organization in a batch or train. The number of flows N' for the given N packets varies from 1 to N. Therefore, when a few flows carry most of the packets($N' \ll N$), the load balancing can not be achieved, because N' is quite small compared to the large N especially when the flow granularity is coarse.

When f_i and r_i denote the number of packets and the delivery time of a packet for a flow i respectively, the flow delivery time r'_i will be $f_i \cdot r_i$. The expectation of the flow delivery time r'_i is as follows:

$$E[r'] = E[f] \cdot E[r] \tag{3}$$

Therefore, the square of the coefficient of variation of the flow delivery time will be as follows.

$$CV[r']^2 = \frac{Var[r']}{E[r']^2} = \frac{Var[f] + Var[r']}{E[f]^2 \cdot E[r']^2} \tag{4}$$

Thus, $Var[f]$ should be finite in order for the flow delivery time r' to have a finite coefficient of variation. However, the skewed flow size distribution may result in a very large variation. In Fig. 3[3], for example, even 1 % of flows contain 65 - 90 % of the load in byte percentage, and 57 - 88 % in packet percentage.

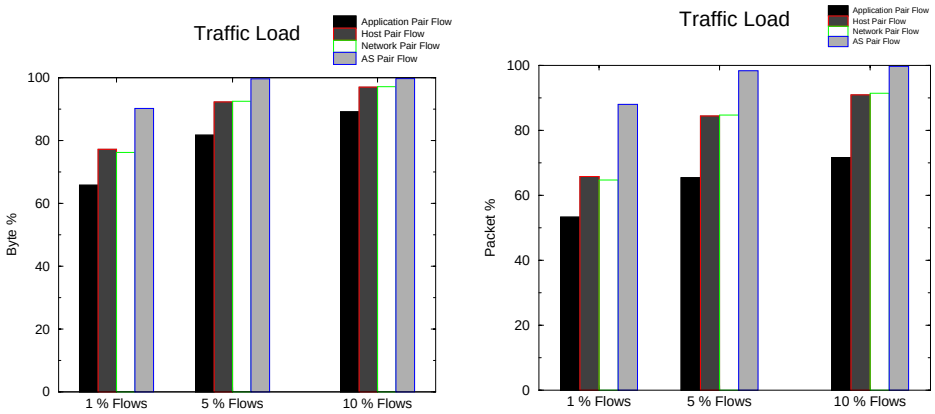
4 The Proposed Load Control Scheme

We develop control scheme for routers with two next-hops(a primary path and a secondary one) for the same destination. This scheme can be easily extended to multiple next-hop cases.

4.1 Flow Classification

For flow assignment, flows which have long duration, high-bit rate, and large flow size(called "base" flows) are distinguished from short-lived transient ones, and assigned to the primary path. Fig. 4 depicts the ingress router with the flow classifier. Packets not belonging to base flows(called "transient" flows) are forwarded to the secondary path.

[3] This trace was measured for one hour on KORNET, a commercial Korean Internet backbone, by Cisco NetFlow[14].



(a) Byte percentage (b) Packet percentage

Fig. 3. Traffic load distribution of 1/5/10 % flows

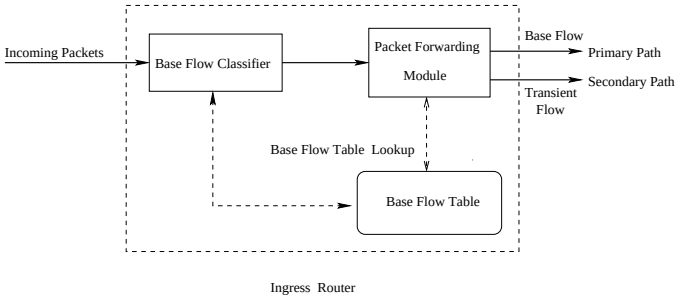


Fig. 4. The ingress router with load control

The base flow detection is based on the X/Y(X: packet count, Y: timeout) flow classifier used in IP switch[1]. In the X/Y flow classifier, a flow is detected when X packets with the same flow specification arrive within Y seconds. This means that the initial X packets of a base flow are forwarded to the transient path. By adjusting X and Y, we can easily control the load assigned to each path. If we increase X(or decrease Y), then less flows will be detected and the load to the primary path will decrease. Decreasing X(or increasing Y) will do the opposite. Thus, adaptive X/Y flow classifier can adapt to dynamic path and traffic behaviors. The packet forwarding module delivers an incoming packet to an appropriate next-hop by looking up the flow table.

4.2 Load Control Algorithm

The load ratio of the primary path is measured by the number of packets sent along the primary path over the total number of packets. The load control algorithm uses the adaptive base flow classifier to meet the given load ratio of the primary path. Although there are two possible adaptive parameters in the X/Y base flow classifier, the flow size, X, is chosen to be variable. The flow size X of the adaptive base flow classifier is adjusted according to the most recent base flow load($BFL(t)$). If $BFL(t)$, which is smoothed by the previous value($BFL(t-1)$) and the recent sample($SampleBFL$), is greater than the given base flow load threshold(BFL_{thr}), the flow size X is increased by Δ. Otherwise, the flow size is decreased multiplicatively by the pre-defined constant, C. Δ is set such that the base flow size estimator X does not increase too quickly. A constant k is used to adjust the increasing amount of Δ.

$$\Delta = \frac{X}{k}, (k > 1) \tag{5}$$

The most recent base flow load in the interval [t-1, t] uses the first-order filter to dampen the abrupt fluctuation of the base flow load. When α approaches $1 (0 < \alpha < 1)$, abrupt changes are suppressed.

Algorithm 1 Adaptive Load Control Algorithm

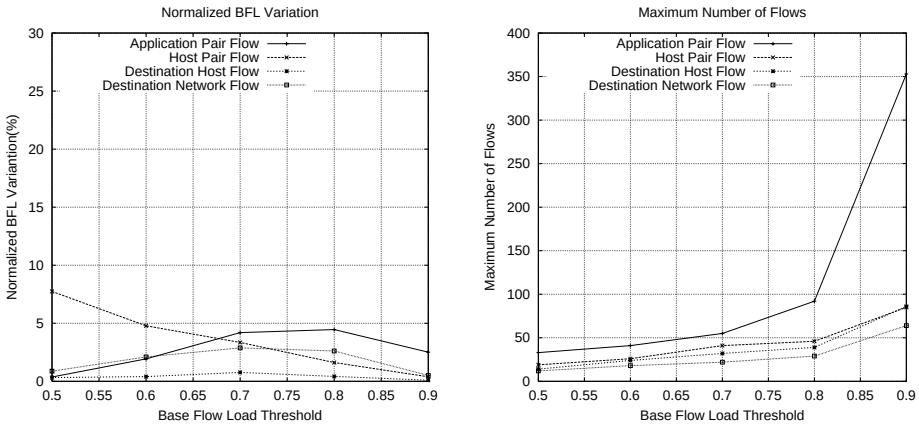
1: $BFL(t) = \alpha \cdot BFL(t-1) + (1-\alpha) \cdot SampleBFL$
2: **if** $(BFL(t) \geq BFL_{thr})$ **then**
3: $\quad \Delta = \frac{X}{k}$
4: $\quad X = X + \Delta$
5: **else**
6: $\quad X = \frac{X}{C}$
7: **end if**

5 Performance Evaluation

To evaluate the load control algorithm, packet traces at the border router of our campus network were captured with tcpdump, and the full routing table of the border router was used. The traffic through the border router shows an average of 4 - 5 Mbps and the traditional traffic pattern of TCP(FTP, WWW) applications.

To compare the pre-defined load ratio of the primary path and the detected base flow load, we define the normalized base flow load ratio variation, $\hat{B}(\%)$,

$$\hat{B} = 100 * \frac{|BFL_a - BFL_{thr}|}{BFL_{thr}} \tag{6}$$



(a) Normailized BFL variation (b) Maximal number of flows

Fig. 5. Normalized base flow load ratio variation and maximal number of flows

where BFL_{thr} and BFL_a are the threshold and the acquired base flow load ratio for the primary path, respectively.

The proposed algorithm requires pre-flow state at routers, and not scalable. This is overcome by aggregating flows going to the same destination. The aggregation can be done in different levels: application, host pair, destination host, or destination network.

From Fig. 5-(a) we can see that the proposed algorithm satisfies the base flow load threshold within 10 %. Among four flow aggregation types, the destination host flow mode shows the lowest normalized variation(1 %) under various base flow load thresholds. This is because the variation of the flow size and the flow duration is rather high except the destination host flow aggregation which generates normally distributed flows.

For the proposed algorithm, the ingress router should maintain the entire per-flow states. The maximum number of flows will affect the scalability of the proposed algorithm. In Fig. 5-(b), the destination network prefix aggregated flows require the minimum number of per flow states even at high threshold. In conclusion, we can see that base flow load assigned to the primary path does not deviate much from the given threshold with the minimal memory requirement.

6 Conclusion

In this paper, we proposed an adaptive flow-level load control algorithm for practical multipath forwarding. It is shown by experiment that the proposed algorithm, which uses the adaptive X/Y flow classifier, divides the input traffic in a way to satisfy the pre-defined load ratio of multiple paths in order to absorb the dynamic flow characteristics. The number of per-flow states required for

a multipath packet forwarding router can be minimized by aggregating flows with the destination host or network prefix. Through this load control scheme, the network resource can be fully utilized and the congestion from the bursty transient flows can be avoided. The proposed load control scheme will be useful for multipath packet forwarding without much additional overhead at routers.

References

1. P. Newman, T. Lyon, and G. Minshall, "Flow Labeled IP: A Connectionless Approach to ATM," IEEE INFOCOM'96, 1996
2. W. Fang, and L. Peterson, "Inter-AS Traffic Patterns and Their Implications," Princeton University TR-598-99, March 1999
3. A. Feldmann, J. Rexford, and R. Caceres, "Efficient Policies for Carrying Web Traffic over Flow-switched Networks," IEEE/ACM Transactions on Networking, pp. 673-685, Dec. 1998
4. J. Chen, P. Druschel, and D. Subramanian, "An Efficient Multipath Forwarding Method," INFOCOM'98, 1998
5. R. Jain, and S. A. Routhier, "Packet Trains - Measurements and a New Model for Computer Network Traffic," IEEE JSAC, vol. 4, no. 9, Sept. 1986
6. D. Thaler, and C.V. Ravishankar, "Using Name-Based Mappings to Increase Hit Rates," IEEE/ACM Transactions on Networking, February 1998.
7. N. S. V. Rao, and S. G. Batsell, "QoS Routing Via Multiple Paths Using Bandwidth Reservation," INFOCOM'98, 1998
8. I. Cidon, R. Rom, and Y. Shavitt, "Analysis of Multi-Path Routing," IEEE/ACM Transactions on Networking, vol. 7, no. 6, pp. 885 - 896, Dec. 1999
9. S. Bahk, M. Zarki, "Dynamic Multi-path Routing and How it Compares with other Dynamic Routing Algorithms for High Speed Wide Area Networks," ACM Computer Communications Review, vol. 22, no. 4, pp. 54-64, Oct. 1992
10. A. Shaikh, J. Rexford, and K. G. Shin, "Load-Sensitive Routing of Long-Lived IP Flows," SIGCOMM'99, 1999
11. P. Bhaniramka, W. Sun, R. Jain, "Quality of Service using Traffic Engineering over MPLS: An Analysis," Globecom'99, 1999
12. Z. Cao, Z. Wang, and E. Zegura, "Performance of Hashing-Based Sche mes for Internet Load Balancing," INFOCOM'2000, 2000
13. The Network Simulator- NS-2, http://www.isi.edu/nsnam/ns/
14. Cisco, http://www.cisco.com