ON THE PERFORMANCE ANALYSIS OF TRAFFIC SPLITTING ON LOAD IMBALANCING AND PACKET REORDERING OF BURSTY TRAFFIC

Sumet Prabhavat[†], Hiroki Nishiyama[†], Nirwan Ansari[‡], Nei Kato[†]

[†] Graduate School of Information Sciences, Tohoku University, Japan
[‡] Advanced Networking Laboratory, ECE Department, NJIT, USA
[†] {sumetp, bigtree, kato}@it.ecei.tohoku.ac.jp, [‡] Nirwan.Ansari@njit.edu

Abstract

Owing to the heterogeneity and high degree of connectivity of various networks, there likely exist multiple available paths between a source and a destination. To be able to simultaneously and efficiently use such parallel paths, it is essential to facilitate high quality network services at high speeds. So, traffic splitting, having a significant impact on quality of services (QoS), is an important means to achieve load balancing. In general, most existing models can be classified into flow-based or packet-based models. Unfortunately, both classes exhibit some drawbacks, such as low efficiency under the high variance of flow size in flow-based models and the phenomenon of packet reordering in packet-based models. In contrast, Table-based Hashing with Reassignment (THR) and Flowlet Aware Routing Engine (FLARE), both belonging to the class of flow-based models, attempt to achieve both efficient bandwidth utilization and packet order preservation. An original flow can be split into several paths. As compared to the traditional flow-based models, load balancing deviation from ideal distribution decreases while the risk of packet reordering increases. In this paper, we introduce analytical models of THR and FLARE, and derive the probabilities of traffic splitting and packet reordering for each model. Our analysis shows that FLARE is superior to THR in packet order preservation. Also, the performance of FLARE on bursty traffic is demonstrated and discussed.

1 Introduction

Effective exploitation of multiple available paths between a source and a destination, in network provisioning, is essential to maximize high quality network services at high speed [1] such as real-time communications under a strict delay requirement. Traffic splitting, which is an important means for load balancing, has significant impact on quality of services (QoS). Since traditional flow-based splitting models, e.g., Direct Hashing (DH), Table-based Hashing (TH), and Highest Random Weight (HRW) [2-6], forward all packets intended for the same destination via the same path, packet reordering is prevented. However, there exists a gap between the actual load and desired load when the sizes of flows are not equal. A variety of flow-based models has been proposed to mitigate the deviation problem [7-9]. As compared to the model proposed by Ref. [7], Table-based Hashing with Reassignment (THR) [8] and Flowlet Aware Routing Engine (FLARE) [9] facilitate control of the load of each path by adjusting some parameters according to a traffic distribution policy. A key point is to split a flow into subflows, which are switched to the lowest utilized path. Ideal load balancing is to ensure that the actual load is equal to the desired load in all paths and the sequence of packets is kept from end to end for all flows. THR and FLARE can balance the load among all paths better as compared to traditional models, but at the expense of preservation of packet order. So, the ability to preserve packet order is the major issue for THR and FLARE to tackle. Unfortunately, THR does not have a mechanism to efficiently prevent packet reordering, whereas FLARE with appropriate control parameters can mitigate the risk of packet reordering because a flow can be split only when the packet interarrival time is larger than the maximum difference of delays among the parallel paths. However, burstiness, which may induce rather small packet-interarrival time, may potentially affect the performance of FLARE. Therefore, we also address its performance on bursty traffic, after having derived the probability of packet reordering in THR and FLARE.

The paper is organized as follows: Section 2 describes major issues concerning load balancing, i.e., packet reordering and the deviation between the actual load and the desired load. Section 3 presents overviews of THR and FLARE. Section 4 introduces analytical models of both schemes. Analytical results are discussed in Section 5; and concluding remarks are presented in Section 6.

2 Issues on load balancing

When the size of each flow and/or the number of flows in each path varies significantly, network

978-1-4244-4900-2/09/$25.00 ©2009 IEEE

congestion may occur in some paths while other paths experience underutilization as a result of imbalanced load distribution among the paths. Splitting a flow into subflows and then moving each subflow to the lowest utilized path is a solution to mitigate the load imbalance [7-9]; however, ability to prevent packet reordering which is the key advantage of flow-based schemes can be compromised when flow splitting and path switching are performed without concern to arrival time at the destination. The two major issues are further discoursed next.

2.1 Load imbalance issue

By using the boundary condition of a path's deficit load stated in Ref. [9], the probability of having a certain degree of load imbalance can be roughly quantified as follows. Let w_p be the normalized desired load of path p. Assume that, over an interval (0, t], the number of active flows is $N(t)$ and a total number of packets over all $N(t)$ flows is $L(t)$. The deficit load of path p, $D_p(t)$ can be calculated as $L_p(t) - w_pL(t)$, where $L_p(t)$ is the actual load of path p. The probability of experiencing the deviation larger than ξ can be expressed as follows:

$$\Pr[|D_p(t)| > \xi] < (\gamma^2 + 1)/(4\xi^2\,\mathrm{E}[N(t)]), \quad (1)$$

where γ is the coefficient of variation of the flow size. As stated in Ref. [9], γ can be reduced by splitting an original flow into several subflows; the probability of observing the deficit load larger than ξ becomes smaller by splitting the flow. From the above discussion, it is clear that load imbalance can be improved by flow splitting, and thus the degree of load imbalance is a function of the probability of flow splitting. In other words, larger splitting probability can achieve better load balancing. From this view point, the probability of splitting is used as an indicator showing how the actual load sharing is close to the ideal load distribution.

2.2 Packet reordering issue

While the sequence of packets is always preserved in most flow-based schemes, it has not been proved that other schemes such as THR and FLARE do preserve packet order. Owing to the different characteristics of parallel paths, the order of packets arriving at a destination may be different from those transmitted via different parallel paths. Packet reordering can lead to a significant performance degradation in applications because it may take a long time to recover packets in the correct order in the IP layer by waiting for an arrival of delayed packets or retransmitted packets. Several works have addressed the packet reordering problem [10-15]. As stated in Refs. [12,13], packet reordering may occur more frequently with higher probabilities of splitting a flow and switching a path. On the other hand, the possibility of packet reordering will decrease if the interarrival time of two successive packets belonging to the same flow is greater than the maximum time required to deliver a packet via the parallel paths. The packet interarrival time must be greater than the difference between the maximum and the minimum delays among parallel paths to ensure preservation of packet order [9].

3 Overviews of THR and FLARE

Due to space limitation, the detailed descriptions of THR and FLARE having been clearly described in Refs. [8,9], respectively, will be omitted in this paper. We will only give overviews of these two that are necessary to understand our analysis. THR and FLARE are flow-based load balancing models attempting to improve their efficiency by splitting a flow into subflows. The packet interarrival time is used to decide whether an original flow should be split or not in both schemes. However, the criterion of splitting an original flow is slightly different from each other. THR allows a flow to be split in between two successive packets with the longest interarrival time observed during a certain time period. Meanwhile, in FLARE, a flow is split when the interarrival time of successive packets exceeds a certain threshold. In both schemes, a path selection may be changed upon splitting a flow. As compared to other flow-based models such as DH, TH, and HRW, the probability of splitting has increased in THR and FLARE, implying the improvement of load balance. We assume that the parameters of THR are chosen such that it aims to achieve the avoidance of packet reordering, in order to compare its performance to FLARE's.

In FLARE, each subflow referred to as a flowlet in Ref. [9] is a group of packets having their interarrival time smaller than an interarrival time threshold, δ_{th}. A packet arriving within δ_{th} is part of an existing flowlet and will be sent via the same path as the previous one. Otherwise, the packet arriving beyond the threshold corresponds to the head of a new flowlet, and is assigned to a path with the largest amount of deficit load. For a smaller δ_{th}, the deviation from the desired load distribution can be decreased at the price of higher risk of packet reordering, and vice versa. In order to guarantee that the two consecutive packets can be assigned to different paths without the risk of packet reordering, the threshold, δ_{th}, needs to be larger than the value of Mean Time Before Switch-ability (MTBS) [9], Δ_{max}, which is the maximum delay difference among all available parallel paths. To estimate the value of Δ_{max}, periodically, FLARE executes the ping operation to measure the round trip delay of each path and calculates the maximum delay difference among the parallel paths from the measured delays; it uses the obtained value, $\Delta^{\text{e}}_{\text{max}}$, as an estimate of MTBS. As $\Delta^{\text{e}}_{\text{max}}$ includes estimation error, $\Delta^{e}_{\max}=\Delta_{\max}+\varepsilon$, where $\varepsilon\in[-\Delta_{\max},\infty)$, the performance of FLARE largely depends on the estimation accuracy. More frequent measurement may reduce $|\varepsilon|$, but this causes additional overheads. An overestimation error ($\varepsilon > 0$) causes a flow not to be split even if it should be, thus leading to the decrease of the opportunity of splitting. On the other hand, an underestimation error ($\varepsilon < 0$) causes a flow to be split more than necessarily, thus causing packet reordering. In the bursty traffic environment, a sudden increase in packet arrival rate can cause underestimation errors. Especially, when the splitting model makes an adaptation decision based on the packet interarrival time, a decision error tends to occur.

4 Analytical models

Let π_s and π_r be the probability that a flow is split upon an arrival of a packet and the probability that packet reordering occurs due to the split, respectively. Assume that they are statistically independent of each other; the relation between them can be expressed as follows:

$$\pi_r = \pi_s \sum_{i\in\mathbf{P}} \sum_{j\in\mathbf{P}} \Phi(i,j)\Omega(\Delta_{i,j}) , \qquad (2)$$

where $\mathbf{P}$ is a set of parallel paths; $\Phi(i, j)$ denotes the probability of the path switching from path i to path j, depending on a path selection strategy; Ω denotes the probability of packet reordering when the path is switched from path i to path j, and is a function of $\Delta_{i,j}$, i.e., the difference of delays between path i and path j. If we assume that path p can be selected with probability w_p for a given arrival packet, $\Phi(i, j)$ is equal to w_iw_j. Especially in a random-based path selection scheme, where all paths have the same selection probability, $\Phi(i, j)$ is a simple function of $|\mathbf{P}|$ as $1/|\mathbf{P}|^2$. Unfortunately, $\Phi(i, j)$ in THR or FLARE cannot be expressed by such a simple function. On the other hand, we do not need to know the exact form of $\Phi(i, j)$ in order to derive the characteristics of π_r. Note that Ω corresponds to the upper bound of π_r. Therefore, in our analysis, the probability of packet reordering is evaluated by using Ω instead of π_r. Besides, the load imbalance is evaluated by using the probability of splitting, π_s, since there exists a direct correlation between them as previously described.

Before proceeding to the analysis on Ω and π_s in THR and FLARE, we shall first formulate the interarrival time of packets by using an interarrival process of the Interrupted Poisson Process (IPP) traffic model, a special case of a Markov Modulated Poisson Process (MMPP) and a widely used renewal process to model bursty traffic and correlated packet-arrivals. Since IPP is stochastically equivalent to a hyperexponential renewal process, the cumulative distribution function of the interarrival time can be simplified to a hyperexponential distribution [16] as follows.

$$H(\delta) = \Pr[X \le \delta; \lambda, \sigma_1, \sigma_2]$$
$$= \begin{cases} p(1-e^{-\mu_1\delta}) + (1-p)(1-e^{-\mu_2\delta}) & ; \ \delta \ge 0 \\ 0 & ;\text{otherwise,} \end{cases} \qquad (3)$$

where $p = (\lambda - \mu_2)/(\mu_1 - \mu_2)$,

$$\mu_1 = \tfrac{1}{2}(\lambda + \sigma_1 + \sigma_2 + \sqrt{(\lambda + \sigma_1 + \sigma_2)^2 - 4\lambda\sigma_2}) \quad ,$$

and

$$\mu_2 = \tfrac{1}{2}(\lambda + \sigma_1 + \sigma_2 - \sqrt{(\lambda + \sigma_1 + \sigma_2)^2 - 4\lambda\sigma_2}) .$$

Here, λ is the arrival rate averaged over a burst period. The mean burst period and mean idle period are $1/\sigma_1$ and $1/\sigma_2$, respectively. Eq. (3) presents the probability that the interarrival time, X, of a packet is not larger than the value of δ.

5 Analysis and discussions

5.1 Probability of splitting

First, we analyze the probability of splitting in THR aiming to tackle the packet reordering issue. The interval between control phases, T, is chosen such that $T \ge 1/(\lambda_{\text{avg}}\mathrm{E}[N(t)])$ to assure the arrival of at least one packet during the interval. $\mathrm{E}[N(t)]$ and λ_{avg} are the expected number of currently active flows and the average packet arrival rate in each flow, respectively. The probability of splitting, π_s^{THR}, derived from the counting process of IPP [16].

$$\pi_s^{\text{THR}} = 1/(\lambda_{\text{avg}} T \mathrm{E}[N(t)]) , \qquad (4)$$

where $\lambda_{\text{avg}} = \lambda / (1 + \sigma_1/\sigma_2)$. It is clear from Equation (4) that a smaller T allows THR to achieve a larger opportunity of splitting. Second, we consider the probability of splitting in FLARE, π_s^{FLARE}, with $\delta_{\text{th}} \ge 0$. From Equation (3), π_s^{FLARE} is

$$\pi_s^{\text{FLARE}} = \Pr[X > \delta_{\text{th}}] = pe^{-\mu_1\delta_{\text{th}}} + (1-p)e^{-\mu_2\delta_{\text{th}}} . \ (5)$$

In Figure 1, for the validation of the above analytical model, π_s^{FLARE} calculated from Eq. (5) and the experimental data, collected in US and Europe [9], are compared in cases with different burst/idle periods. The results confirm the validity of our proposed analytical models. Figure 2 presents the comparison between π_s^{THR} and π_s^{FLARE}. As evident from the values at $T = 0.08$s in THR and $\delta_{\text{th}} = 0.05$s in FLARE as recommended in Ref. [8] and [9], respectively, π_s^{THR} is greater than π_s^{FLARE}. Regardless of the average packet arrival rate, smaller values of T and δ_{th} in THR and FLARE, respectively, are required to achieve higher probability of splitting which results in better balanced load with less deviation from the desired load. Regarding the performance of FLARE, we can also see that the probability of splitting, which infers the efficiency of load balancing, tends to degrade from bursty traffic causing a long burst period.

5.2 Probability of packet reordering

The probability of packet reordering in THR, $\Omega_{\mathrm{THR}}(\Delta_{i,j})$, caused by changing the path from path i to path j, is

$$\Omega_{\mathrm{THR}}(\Delta_{i,j}) = H(\Delta_{i,j})\,. \qquad (6)$$

Eq. (6) shows that the probability of packet reordering is not zero as long as $\Delta_{i,j} > 0$. It is impossible to guarantee preservation of packet order if any two of the available parallel paths exhibit different delays. The probability of packet reordering in FLARE, $\Omega_{\mathrm{FLARE}}(\delta_{\mathrm{th}}, \Delta_{i,j})$, is a function of not only $\Delta_{i,j}$ but also the threshold, δ_{th}.

$$\Omega_{\mathrm{FLARE}}(\delta_{\mathrm{th}}, \Delta_{i,j}) = \Pr[\delta_{\mathrm{th}} < X \le \Delta_{i,j}]$$
$$= \begin{cases} H(\Delta_{i,j}) - H(\delta_{\mathrm{th}}) & ; \quad \delta_{\mathrm{th}} < \Delta_{i,j} \\ 0 & ; \quad \text{otherwise}. \end{cases} \qquad (7)$$

Since FLARE allows only a flow (with $\delta > \delta_{\mathrm{th}}$) to be split, the probability of packet reordering can be controlled by adjusting the value of δ_{th}. In fact, it is possible to reduce the risk of packet reordering to almost zero regardless of the probability of splitting when δ_{th} is set to $\Delta^{\mathrm{e}}_{\max}$ which is close to $\Delta_{i,j}$. While comparing between Eqs. (6) and (7), it is obvious that FLARE outperforms THR because it determines the splitting of flows according to the network condition, and thus reducing the risk of packet reordering. Eq. (7) demonstrates that the risk of packet reordering can be reduced by increasing the value of δ_{th}. A large value of δ_{th} is required to maintain a certain probability of packet reordering under the conditions where parallel paths have a large variance in their delays. FLARE with properly chosen δ_{th} can mitigate the risk of packet reordering; whereas THR does not have such an advantage.

Estimation errors causing $\Delta^{\mathrm{e}}_{\max}$ to be far from $\Delta_{i,j}$ are a factor affecting the performance of FLARE; however, it should be noted that its performance in terms of packet reordering is always better than that of THR even in the existence of estimation errors. Since THR has no more factor to be taken into account, we will further discuss only the effect (of estimation error) on the performance of FLARE.

5.3 Effect of estimation errors in FLARE

Without an estimation error, i.e., ε=0 and $\Delta^{\mathrm{e}}_{\max}$=$\Delta_{\max}$, FLARE can achieve almost ideal performance with negligible risk of packet reordering by setting δ_{th} to $\Delta^{\mathrm{e}}_{\max}$. However, in general, the estimated $\Delta^{\mathrm{e}}_{\max}$ includes estimation error. An overestimation error leads to the loss of the opportunity of splitting, which results in deteriorated load balancing. On the other hand, an underestimation error increases the risk of packet reordering while load balancing can be promoted. Figure 3 demonstrates the effect of underestimation error on the probability of packet reordering, Ω_{FLARE}. It can be confirmed that the increase of underestimation error leads to the increase of the risk of packet reordering, regardless of the value of δ_{th}, and the length of the burst period of incoming traffic. It should be noted that, to decrease the risk of packet reordering under the condition with the existence of a certain estimation error, δ_{th} needs to be set to a large value, while a large δ_{th} jeopardizes the opportunity to improve load imbalance.

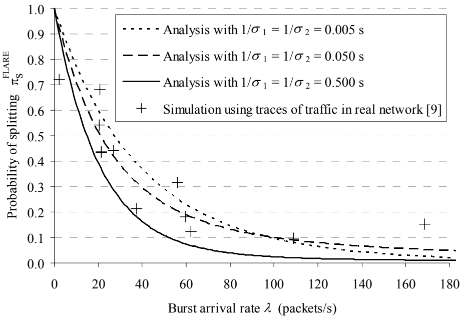


Figure 1. Analytical model validation.

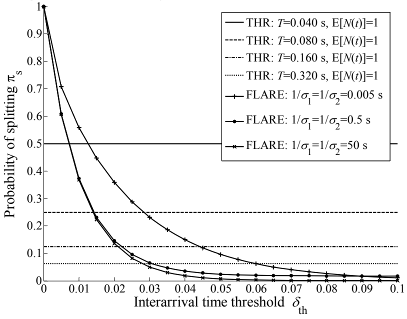


Figure 2. Comparison between probability of splitting in THR and that in FLARE, for λ_{avg} = 50 packets/s.

Meanwhile, as derived from Eqs. (3) and (7), Ω_{FLARE} can be affected by the length of the burst/idle periods, $1/\sigma_1$ and $1/\sigma_2$, of incoming traffic. In fact, a clear difference can be observed in between Figures 3(a) and 3(b). From the comparison between these figures, it is clear that, to maintain a certain probability of packet reordering, the threshold, δ_{th}, needs to be dynamically adjusted according to the burst period of the incoming traffic.

6 Concluding remarks

Since effective traffic splitting is critical to efficiently utilize multiple available paths, we have proposed analytical models to conduct performance analysis and comparison between recently introduced models such as THR and FLARE, in terms of the probability of traffic splitting and packet reordering, under varying control parameters as well as bursty conditions. These analytical results demonstrate that in order to perform load balancing, better load distribution with smaller deviation can be achieved by setting smaller update interval times in THR or by setting smaller interarrival time thresholds in FLARE, at the expense of preservation of packet order, and vice versa. By using the analytical models, we are able to conduct a detailed comparative analysis of THR and FLARE; and to explain why FLARE can outperform THR in mitigating the packet reordering problem.

Moreover, we illustrate the degraded performance of FLARE in preventing packet reordering when an underestimation error occurs. The results indicate that a larger error causes a higher risk of packet reordering; in addition, the probability of packet reordering may be affected by the length of the burst/idle periods of the incoming traffic. In order to alleviate the probability of packet reordering, the interarrival time threshold needs to be dynamically adjusted according to the incoming traffic. By setting a relatively larger threshold, the effect of underestimation error causing packet reordering can be suppressed regardless of the lengths of the burst/idle periods, at the expense of the loss of the opportunity to further improve load imbalance.

References

[1] L. Golubchik, J. Lui, T. Tung, A. Chow, W. Lee, G. Franceschinis, and C. Anglano, “Multi-path continuous media streaming: What are the benefits?,” Performance Evaluation, vol. 49, pp. 429–449, Sep. 2002.

[2] C. Villamizar, “OSPF optimized multipath (OSPF-OMP),” Internet draft draft-ietf-ospf-omp-02.txt, Feb. 1999.

[3] D. Thaler and C. Hopps, “Multipath issues in unicast and multicast next-hop selection,” RFC 2991, Nov. 2000.

[4] C. Hopps, “Analysis of an equal-cost multi-path algorithm,” RFC 2992, Nov. 2000.

[5] Z. Cao, Z. Wang, and E. Zegura, “Performance of hashing based schemes for Internet load balancing,” in *Proc. IEEE INFOCOM*, Mar. 2000, pp. 332–341.

[6] D. G. Thaler and C. V. Ravishankar, “Using name-based mappings to increase hit rates,” IEEE/ACM Trans. Networking, vol. 6, no. 1, pp. 1–14, Feb. 1998.

[7] W. Shi, M. H. MacGregor, and P. Gburzynski, “Load balancing for parallel forwarding,” IEEE/ACM Trans. Networking, vol. 13, no. 4, pp. 790–801, Aug. 2005.

[8] T. W. Chim, K. L. Yeung, and K.-S. Lui, “Traffic distribution over equal-cost-multi-paths,” Computer Networks, vol. 49 (4), pp. 465–475, Nov. 2005.

[9] S. Kandula, D. Katabi, S. Sinha, and A. Berger, “Dynamic load balancing without packet reordering,” ACM SIGCOMM Computer Communication Review, vol. 37, no. 2, Apr. 2007, pp. 53–62.

[10] V. Paxson, G. Almes, J. Mahdavi, and M. Mathis, “Framework for IP performance metrics,” RFC 2330, May. 1998.

[11] C. Demichelis, P. Chimento, “IP packet delay variation metric for IP performance metrics (IPPM),” RFC 3393, Nov. 2002.

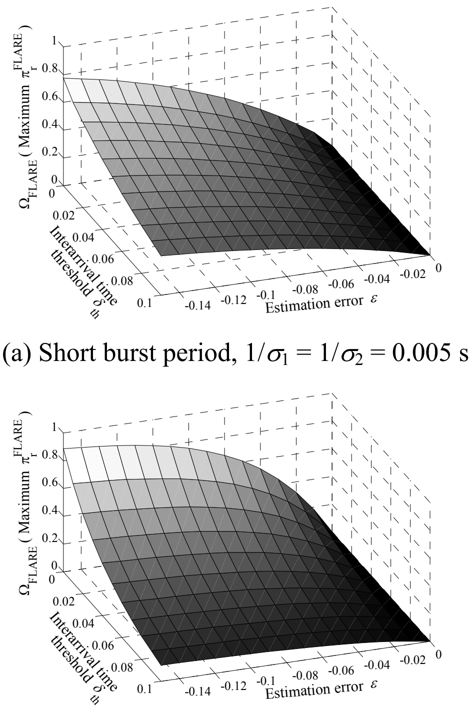
[12] N. M. Piratla, A. P. Jayasumana , A. A. Bare , and T. Banka, “Reorder buffer-occupancy density and its application for measurement and evaluation of packet reordering,” Computer Communications, vol. 30 no.9, pp.1980–1993, Jun. 2007.

[13] N. M. Piratla and A. P. Jayasumana, “Reordering of packets due to multipath forwarding – An analysis,” in *Proc. IEEE ICC*, Jun. 2006, pp. 829–834.

[14] A. Morton, L. Ciavattone, G. Ramachandran, S. Shalunov, and J. Perser, “Packet reordering metrics,” RFC 4737, Nov. 2006.

[15] A. Jayasumana, N. Piratla, T. Banka, R. Whitner, “Improved packet reordering metrics,” RFC 5236, Jun. 2008.

[16] W. Fischer and K. Meier-Hellstern, “The Markov-modulated Poisson process (MMPP) cookbook,” Performance Evaluation, vol. 18 (2), pp. 149–171, Sep. 1993.



(a) Short burst period, $1/\sigma_1 = 1/\sigma_2 = 0.005$ s.

(b) Long burst period, $1/\sigma_1 = 1/\sigma_2 = 0.5$ s.

Figure 3. Probability of packet reordering occurrence in FLARE, for $\lambda_{avg} = 10$ packets/s.