DOI: 10.3745/KIPSTC.2008.15-C.4.321

혼잡 망에서의 큐 제어 방식과 전송지연시간에 대한 웹 반응 시간 분석

석 우 진†

요 약

본 논문에서는, 혼잡 망에서의 큐 제어 방식과 전송 지연 시간에 대한 웹 반응 시간을 분석하였다. FIFO 방식에서는, 웹 반응 시간이 큐 크기에 대해서 거의 일정한 성능을 보여주었으나, 트래픽 부하가 높아질수록 웹 반응 시간은 길어졌다. 80%의 트래픽 부하보다, 90%와 98%의 트래픽 부하일 경우에, 큐 크기가 달라짐에 따라 웹 반응 시간이 더 뚜렷하게 다르게 나타났다. 특히 전송 지연 시간이 짧을 경우, 웹 반응 시간의 차이가 더욱 뚜렷하게 나타났다. RED 방식에서는, 상대적으로 큰 큐 크기일 경우, 3가지 서로 다른 설정의 RED 방식이 웹 반응 시간에 뚜렷한 영향을 미치지 못하였다. 큐 크기가 작아졌을 경우, 짧은 전송 지연에 대하여 각 RED 방식의 웹 반응 시간이 서로 뚜렷하게 다르게 나타났다. FIFO와 RED의 비교에서, 긴 전송 지연 시간일 경우, RED 방식이 FIFO보다 작은 웹 반응 시간을 보여주었다.

키워드: FIFO, RED, 웹 반응 시간

Analysis of Web Response Time on Queue Managements and Transmission Latency in Congested Network

Woojin Seok†

ABSTRACT

In this paper, we analyze web response time depending on queue managements and transmission latencies in highly congested network situation. Under FIFO scheme, the response times are for three different sizes of queue are almost the same, but the response time increases as traffic intensity increases. The performance between different queue sizes shows more different in 90% and 98% traffic intensity than in 80% traffic intensity. Especially the difference becomes bigger in short latency case than long latency case. Under RED scheme, three differently tunned RED schemes do not impact on the response time when the size of queue is relatively large. When the queue size becomes smaller, the response time of the differently tunned RED schemes becomes different for short latency case while the response times are almost same for long latency case. When comparing FIFO and RED schemes under same size of queue, RED scheme shows less response time than that of FIFO for the long latency case in high traffic intensity.

Key Words: FIFO, RED, Web Response Time

1. Introduction

There is increasing access to web in current network. Most of information is provided via web and the amount of traffic accessing web is voluminous. When accessing a web, every users want to get quick response from the web. The response time depends on a lot of factors such as the number of users, accessing behaviors, server performance, and network performance. The network performance also includes transmission latency, bandwidth, handoff management, queue management, and so on.

Among the factors, we consider queue managements and latency and their influences on the network performance. In the equipments in core or access networks, they may run different queue management policies to handle bulky traffic efficiently in a limited network resources. First In First Out(FIFO)

† 종신회원 : 한국과학기술정보연구원 연구망개발팀 선임연구원
논문접수 : 2008년 1월 16일
수 정 일 : 2008년 6월 20일
심사완료 : 2008년 6월 21일

scheme and Random Early Detection(RED) are popular queue management schemes studied so far.

The RED algorithm, first described over ten years ago [1], inspired a new focus for congestion control research on the area of active queue management. The study in [9] showed the effects of RED on the performance of web access with an aspect of response time for HTTP request-response pairs. They concluded that RED queue management appeared to provide no clear advantage over tail-drop FIFO for end-user response times. Concerning network load, it was previously shown that Adaptive RED(ARED) did not improve response time performance (compared to tail-drop FIFO scheme) at offered loads up to 90% of link capacity [2]. The other active queue management such as Proportional Integrator (PI), the Random Exponential Marking (REM) also did not show better response time over tail-drop FIFO. The study in [3] presented an empirical study of the effects of ARED and explicit congestion notification(ECN) on the distribution of response times, and the results were that the combined RED and ECN did not give performance improvement. The studies for RED or active queue management showed that the response time to end-user for web access did not give significant improvement over tail-drop FIFO.

In current and future network, wireless access is popular and will become more popular. In the wireless network, the transmission latency is varied in wide range. The round trip time in General Packet Radio Service(GPRS) of cellular network is between 500 and 3000 ms [4][5]. There is also increasing need to use communications satellites to carry Internet traffic, whose transmission latency is very long [6]. Long propagation delay is one of the main causes of negative effects on TCP's performance which plays important role for web response time. The one-way propagation delays of satellite environment are from 110 to 150 ms for medium earth orbit systems (MEO) and from 250 to 280 ms for geostationary satellites (GEO) [6][7][8].

As differently with the previous studies, we analyze the web response time in highly congested network for FIFO and RED and for two different sets of transmission latency to emulate the various transmission latency including wireless access networks. We measure the response times with the two different sets of transmission latencies, short latency and long latency, under FIFO and RED schemes. The "short" and "long" are relative expressions with each other. In order to measure them, we establish a testbed using Intel machine running Linux operation system. In the routing node in the testbed, two queue management schemes run. Traffic is generated and responded to represent short and long latency.

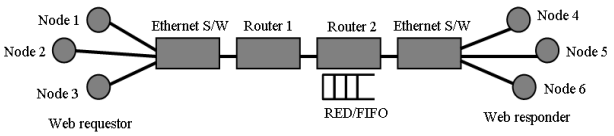
2. Experiments Environment

2.1 Testbed Environment

In this paper, we measure response time to analyze the influence of different transmission latencies and queue managements. We also measure queue length to augment the analysis of response time. To measure response time, we configure a testbed as shown in (Figure 1). The three nodes of web requestors generate web traffic destined for three nodes of web responder. The web requestor emulates the browsing behavior of thousands of human users (described in Section 2.2). The web responder creates the traffic in response to the browsing requests.

Web requestor and responder runs on FreeBSD 2.2.8, and are sey by 10/100Mbps Ethernet interface. The Ethernet SW is Cisco Systems Catalyst 5000. The router transfers packets to other domain, and runs on FreeBSD Intel machines.

The parameters of RED schemes are shown in <Table 1>, each of which means best overall re-



(Fig. 1) Testbed Diagram

〈Table 1〉 Three Sets of RED Parameters(th: threshold, max: max probability)

min_th	max_th	weight	max	qlen	
5	90	1/128	1/20	480	RED-A; best overall response
30	180	1/512	1/10	480	RED-B; highest link utilization
90	150	1/512	1/20	480	RED-C; lowest drop rate

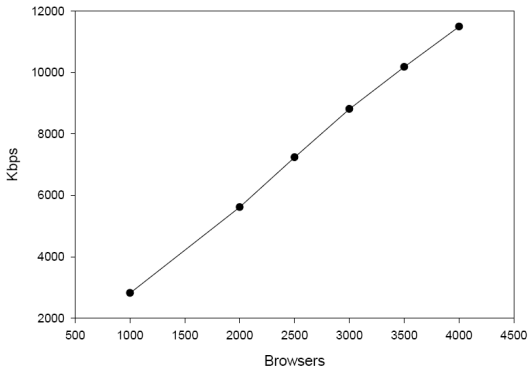
〈Table 2〉 Short Transmission Latency (ms)

	node4	node5	node6
node 1	-	48	109
node 2	38	-	7
node 3	91	10	-

sponse(RED-A), highest link utilization(RED-B), and lowest drop rate(RED-C) as studied in [9]. The parameters of FIFO schemes have the three different lengths of queue, 60, 120, and 180. The transmission latencies have two sets, called short and long latency. The long latency is three times larger than short latency. The short latencies are shown in <Table 2>.

2.2 Calibration

The web requestor and responder generate web traffic which emulates human behavior to access web. In order to do this, the real web traffic is accumulated and generated by program. The real web traffic was extracted from more than 230 hours of traces collected on the UC-Berkeley campus over 1.6 million HTTP protocol, and Christiansen [9] wrote program re-generating traffic emulating the collected traces. The program creates an instance of browser which generates web traffic, and is used in this paper. As many as the program generates browsers, the whole traffic becomes larger. In order to see how many browsers generates a specific traffic intensity, we run the program in 100Mbps environment where every links in the testbed are set by 100Mbps. We measure the throughput between the nodes of web requestor and responder. The results are shown in (Figure 2). From the results of experiments under 100Mbps, we can get the linear formula, such as, y = 2.9327x - 125.0857 by



(Fig. 2) Calibration

〈Table 3〉 Three Numbers of Browsers

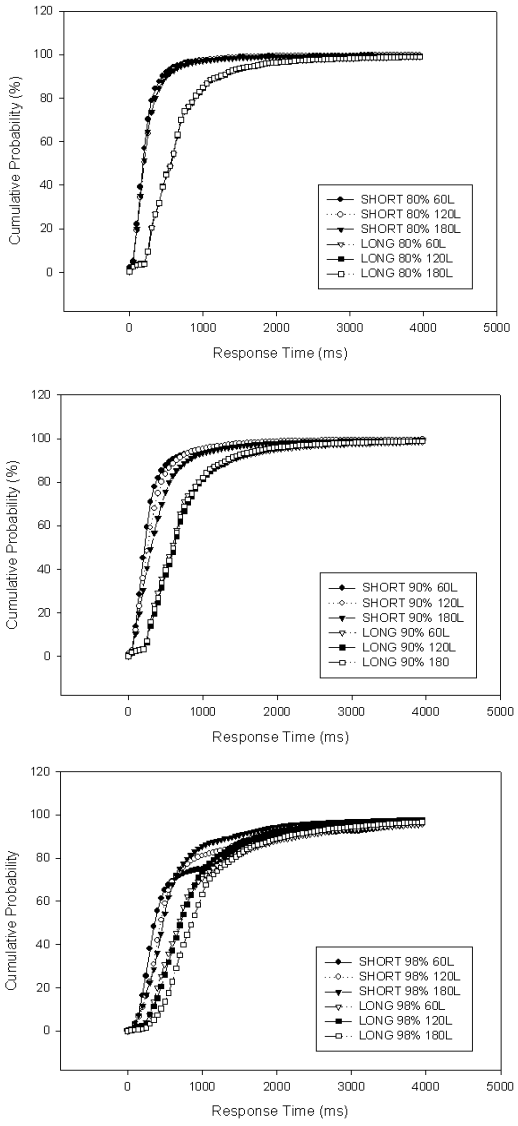
Traffic Load	The Number of Browsers
80%	2770
90%	3111
95%	3384

regression function provided in SigmaPlot 2000 which is a mathematical tool. From the linear formula, the number of browsers to generate 80%, 90%, and 98% traffic load turns out in <Table 3>.

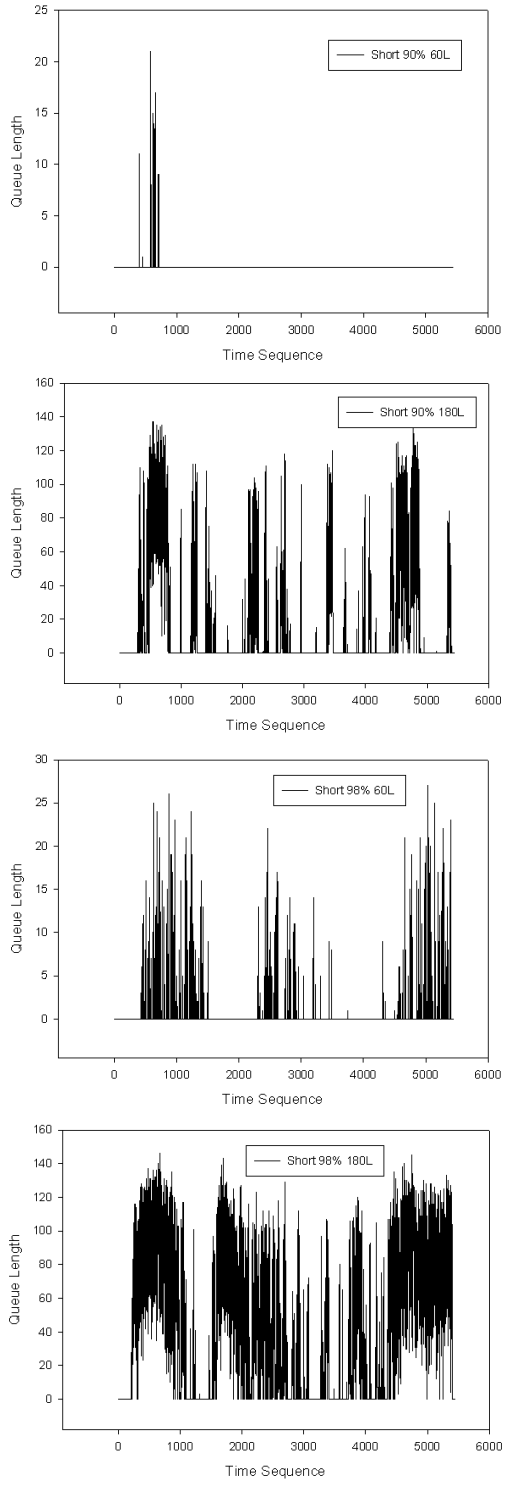
3. Results and Discussions

3.1 FIFO Results

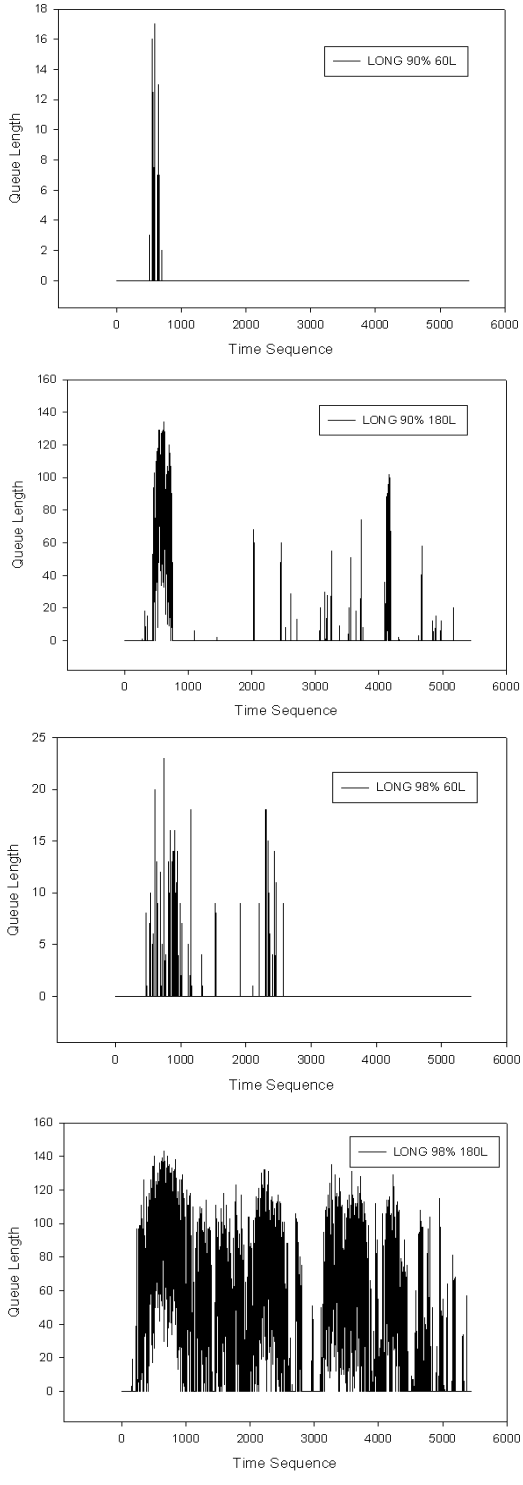
In (Figure 3), the response time for 80% traffic



(Fig. 3) Response Time on Traffic Intensity 80%, 90%, 98%(60L, 120L, and 180L means Queue Size)



(Fig. 4) Queue Occupancy for Short Latency Case



(Fig. 5) Queue Occupancy for Long Latency Case

intensity of short latency case shows that 90% of response time is less than around 500 ms, while the response time for 80% traffic intensity of long latency case shows that 90% of response time is less than around 1200 ms. Response time does not depend on the size of queue on 80% traffic intensity, that is, 3 different queue sizes show almost same response time in short and long latency case. As traffic intensity increases, the response time increases too, because the queuing latency increases as traffic in-

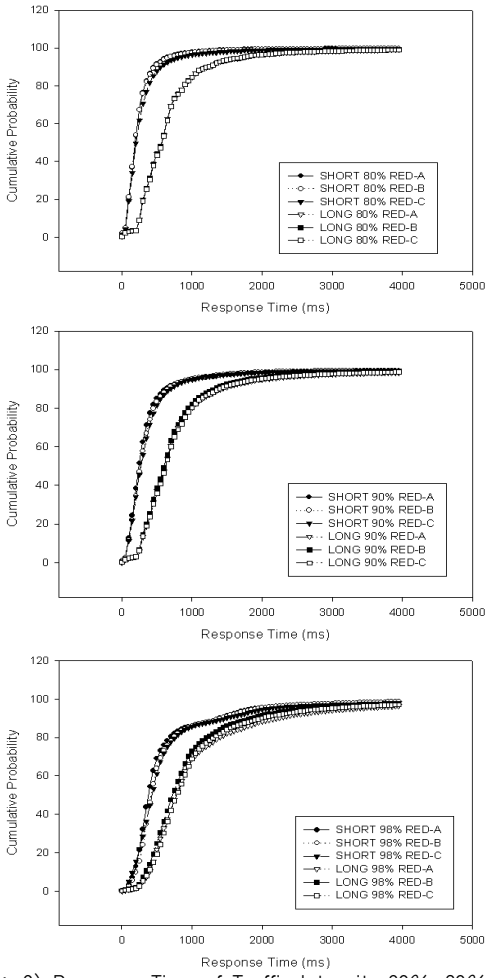
tensity increases. That is, the traffic has to be queued as the arrival rate gets higher, so the longer queue cause longer delay to response time. This feature is appeared in both short and long latency case. In much congested environment, 98% traffic intensity, the response time becomes longer as queue size becomes larger because high traffic intensity almost occupies the small queue size and causes packet drop which causes retransmission for the dropped packet. The retransmission requires additional time to complete a web request and this makes the response time to become long. The performance difference between different queue sizes shows more apparently in 90% and 98% traffic intensity than in 80% traffic intensity. Especially short latency case shows the difference more apparently than long latency case.

The performance of response time depends on the queue occupancy. (Figure 4 and 5) show the queue occupancy for short and long latency. The results show two features; 1) higher traffic intensity shows much higher occupancy of queue, 2) longer queue shows much higher occupancy of queue. The reason of higher occupancy of queue for the higher traffic intensity is that link speed is not enough fast to transmit all packets generated by higher traffic intensity, so queuing is not evitable. The reason of higher occupancy of longer queue size is that the packets are not discarded but queued while they may be discarded on less size of queue.

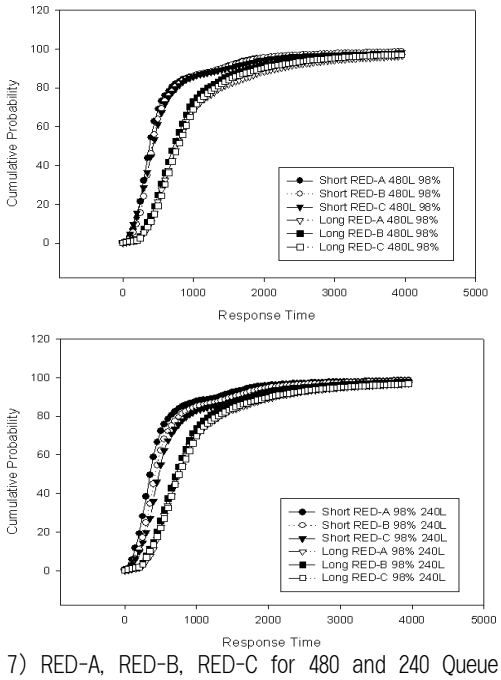
Under the same size of queue, long latency case with 90% traffic intensity has fewer packets than that of short latency case. It is because the link itself plays buffer-like role in transmitting data. For the same incoming traffic and the same size of queue, the long latency case can provide less queue occupancy for the traffic than the short latency. This produces packet drop more frequently in short latency case, and this results in that response time more depends on queue size in case of short latency than long latency.

3.2 RED Results

As mentioned in <Table 1>, RED-A, RED-B, and RED-C are optimized to have best overall response time, highest link utilization, and lowest drop rate as expected in [9]. In the (Figure 6), the response time to the 80% traffic intensity shows that 90% of response time is less than around 500 ms and around



(Fig. 6) Response Time of Traffic Intensity 80%, 90%, 95%



(Fig. 7) RED-A, RED-B, RED-C for 480 and 240 Queue Size under 98% Traffic Intensity

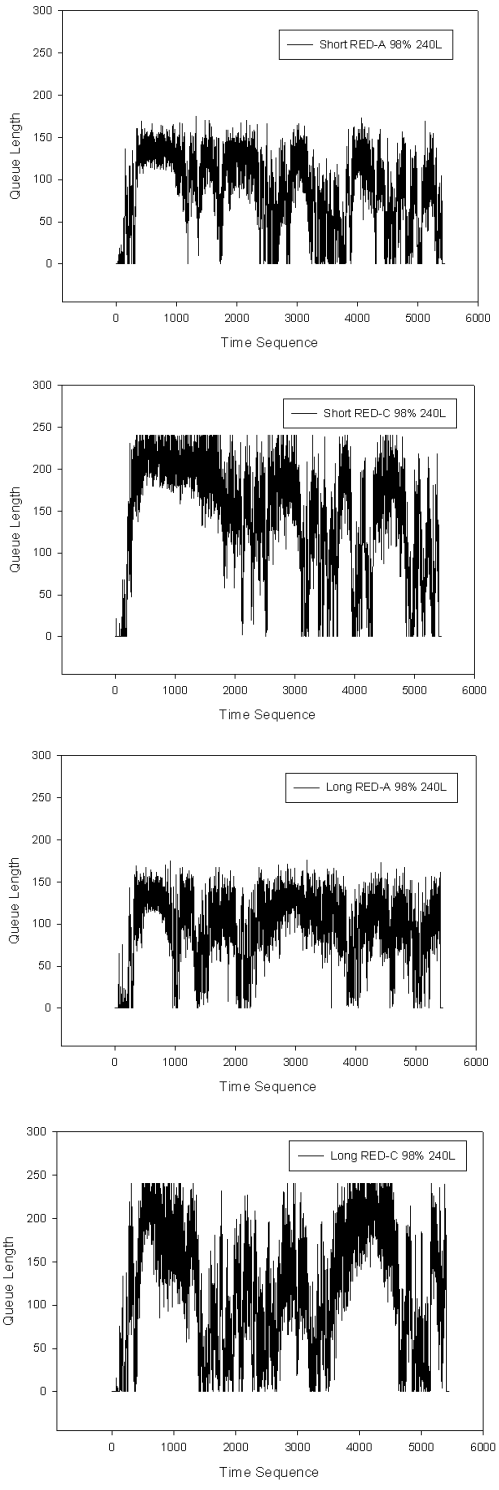
1200 ms in short and long latency case, respectively. The response time to the 98% traffic intensity is around 700 ms and 1500 ms in short and long latency case, respectively. That is, the response time increases as the traffic intensity increases in RED scheme. But the different sets of RED parameters shows almost same response time in all of traffic intensities, in both of short and long latency case.

The (Figure 7) shows the response time of different RED schemes for 480 queue size and 240 queue sizes. The result with 480 queue size shows almost same for three different RED schemes. In 240 queue sizes, however, RED-C scheme shows apparently longer response time than the RED-A and RED-B only in short latency.

The (Figure 8) shows the queue length of RED-A and RED-C with short RTT and long RTT of 240 queue sizes. Two figures, (Figure 8-A) and (Figure 8-B), have much different queue occupancy, that is, the average size of queue length of (Figure 8-B) seems to be much higher than 150 while (Figure 8-A) has a little less than 150. So the response time can be much different between the two cases. On the other hand, the average of queue length of (Figure 8-D) seems to be around 150, and the average of (Figure 8-C) seems to be around 100. So the response time may not much different between two cases as compared with difference between (Figure 8-A) and (Figure 8-B).

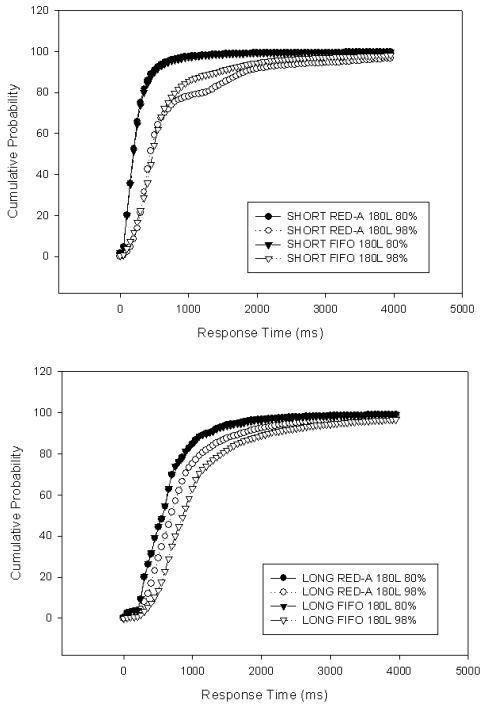
In the (Figure 9), we compare the response time of FIFO and RED scheme for the same size of queue. The response time of 80% traffic intensity of short latency case shows that FIFO and RED are almost same response times. But the response time of FIFO and RED under 98% traffic intensity of long latency is apparently different with each other. In long latency case, the queue must not be full as much as short latency case, and RED can utilize the queue more efficiently than FIFO does. In the highly congested and small queue size shown in the left figure of (Figure 9), RED and FIFO almost the same results because there always are many packets arriving at queue which is almost full. In the long latency case, in highly congested situation, RED shows better performance because it has enough time to handle incoming traffic.

RED shows better performance than FIFO in case that highly congested and long latency situation. The



(Fig. 8) Queue Length of 98% Traffic Intensity on Short RTT and Long RTT in 240 Queue Size (Figure 8-A, Figure 8-B, Fig 8-C, Figure 8-D in clockwise from left top to right bottom)

differently tunned RED shows almost similar performance for traffic intensities and for short and long latency case. In the case that small queue size and



(Fig. 9) Response Time of FIFO and RED-A on 180 queue size(left: Short RTT, right: Long RTT)

highly congestion is provided, RED-A shows better performance than RED-B and RED-C. As a result, RED-A would be much beneficially used in the situation that high congestion, small queue size, and long latency are provided.

5. Conclusions

We conduct two sets of experiments to measure web response time for short latency case and long latency case with two different queue managements. We observe that the response time of long latency case is much larger than short latency and this property conserves for both of FIFO and RED.

Under FIFO scheme, the response time is almost the same to three different sizes of queue, but the response time increases as the traffic intensity increases, which means that the traffic intensity impacts on the response time more than queue size does.

In particular, the performance between different queue sizes is more differently in 90% and 98% traffic intensity than in 80% traffic intensity. Especially short latency case shows the difference become bigger than long latency case.

Under RED scheme, three different RED schemes do not impact on the response time when the size of queue is 480. When the queue size becomes 240, the response time of different RED schemes of short latency case becomes different. In long latency, however, the response times of three RED schemes are still almost same for 240 queue size.

When comparing FIFO and RED schemes under same size of queue, the short latency case does not show much different response time, but the long latency case shows that RED scheme shows less response time than that of FIFO in high traffic intensity. Therefore, RED could be used beneficially in the situation that high congestion and long latency are provided.

References

[1] S. Floyd and V. Jacobson, “Random Early Detection Gateways for Congestion Avoidance,” IEEE/ACM Transactions on Networking, Vol.1, No.4, pp.397-413, August 1993.

[2] L. Le, J. Aikat, K. Jeffay, and F.D. Smith, “The Effects of Active Queue Management on Web Performance,” ACM SIGCOMM 2003, pp.265-276, August 2003.

[3] L. Le, J. Aikat, K. Jeffay, and F. D. Smith, “The Effects of Active Queue Management and Explicit Congestion Notification on Web Performance,” IEEE/ACM Transactions on Networking, Vol.15, pp.1217-1230, December. 2007

[4] R. Chakravorty, P. Vidales, K. Subramanian, I. Pratt, and J. Crowcroft, “Performance Issues with Vertical Handovers - Experiences from GPRS Cellular and Wireless LAN Hot-spots Integration,” in Proc. IEEE PERCOM, pp.155-164, March 2004.

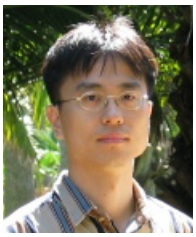
[5] A. Gurtov and J. Korhonen, “Effect of vertical handovers on performance of TCP-friendly rate control,” ACM SIGMOBILE Mobile Computing and Communications Review, Vol.8, No.3, pp.73-87, July 2004.

[6] M. Allman, D. Glover, and L. Sanchez, “Enhancing TCP over Satellite Channels using Standard Mechanisms,” IETF RFC 2488, January 1999.

[7] N. Ghani and S. Dixit, “TCP/IP Enhancemetns for Satellite Networks,” IEEE Communications Magazine, July 1999.

[8] M. Allman, et al., “Ongoing TCP Research Related to Satellites,” IETF RFC 2760, February 2000.

[9] M. Christiansen, J. Kevin, D. Ott, and F. D. Smith, “Tunning RED for Web Traffic,” IEEE/ACM Transactions on Networking, Vol.9, No.3, pp.249-264, June 2001.



석 우 진

e-mail : wjseok@kisti.re.kr

1996년 경북대학교 컴퓨터공학과(학사)

2003년 Univ. North Carolina, Computer Science(이학석사)

2003년~현 재 한국과학기술정보연구원 선임연구원

관심분야 : 무선/이동 QoS, TCP 성능 분석