DBpia

논문 19-44-01-13 The Journal of Korean Institute of Communications and Information Sciences '19-01 Vol.44 No.01
https://doi.org/10.7840/kics.2019.44.1.81

Cut-through 스위칭의 지연성능 분석: 큐기반 근사화

주 창 희◆°, 이 승 현*

Delay Performance of Cut-through Switching: A Queueing-Based Approximation

Changhee Joo◆°, Seunghyun Lee*

요 약

데이터센터 네트워크에서 초저지연 성능이 중요해지면서 이를 지원하기 위한 기술로서 cut-through 스위칭이 많은 주목을 받고 있다. 패킷 전체의 수신을 완료하기 전부터 일부 수신된 패킷을 전송하기 시작하는 cut-through 스위칭은 이미 상용제품이 개발되었다. 그러나, 최근의 관심에도 불구하고 다양한 네트워크 환경에서 cut-through 스위칭의 실질적인 지연성능은 잘 알려져 있지 않다. 본 논문에서는 큐기반의 분석을 통해 cu-through 스위칭의 지연성능을 근사화할 수 있는 분석 프레임워크를 제안한다. 기존의 큐 모델과는 다르게, 패킷의 마지막 비트의 동작을 함께 고려함으로써 cut-through 전송의 주요 특성을 이해할 수 있다.

Key Words : switching, cut-through, delay, low latency, approximation.

ABSTRACT

Cut-through switching has attracted much attention as a promising approach to achieve ultra low latency in datacenter networks. Several cut-through switches that can forward a packet before its reception completes are already commercially available. However, despite the growing interest, the performance of cut-through switching in various network environments is not well understood. In this work, we propose a queueing-based framework to understand the behaviors of cut-through switching and to analyze its delay performance. Unlike traditional queueing models, we take into account the first-bit arrival of a packet as well as the last-bit departure of a packet to capture the essential features of cut-through transmissions.

Ⅰ. Introduction

Cut-through switching (or cut-through forwarding) is a method for packet transmission such that a packet is forwarded before being completely received. The technique has been first proposed by Kermani and Kleinrock[1], to avoid unnecessary buffering delay in front of idle outgoing link. However, most practical up-to-date switches transmit packets in a store-and-forward method: a packet is

※ 본 연구는 2018년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No. 2015-0-00278, Research on Near-Zero Latency Network for 5G Immersive Service)

◆° First Author and Corresponding Author : (ORCID 0000-0003-1690-2298)UNIST, School of Electrical and Computer Engineering, cjoo@unist.ac.kr, 정회원

* UNIST, School of Electrical and Computer Engineering, seunghyunlee@unist.ac.kr, 정회원

논문번호 : 201817-222-B-RN, Received July 8, 2018; Revised October 25, 2018; Accepted November 13, 2018

www.dbpia.co.kr

completely received and then transmitted to the next hop. Recently, as the demand for ultra-low latency service emerges for healthcare, transport, entertainment, and manufacturing[2,20], cut-through transmission has attracted much attention to achieve ultra-low latency.

In contrast to the rapidly growing interest, there are few analytical results to analyze delay performance of cut-through switching. In [1], the authors developed an analytical technique to estimate average packet delay under cut-through switching under limited scenarios, e.g., when all link rates and traffic amount are identical. The authors in [3] have used the M/D/1 queueing model under assumption of fixed packet length. They have investigated the end-to-end behavior under cut-through when the channel is nosy, and show that cut-through switching outperforms store-and-forward switching. The M/G/1 queueing model has been also used in [4] to understand the behaviors of hybrid switching, where only a message whose length is beyond certain threshold is segmented into multiple packets. The authors of [5] have analyzed the performance of cut-through switching in a special network topology of banyan networks. They have shown that cut-through switching has great advantage in terms of delay, in particular, with large switch size.

Besides analytical approaches, the authors in [6], [7] provide a simulation-based performance comparison between cut-through switching and store-and-forward transmission. The performance of routing algorithms under cut-through switching has been compared in [6], and several switch designs for cut-through switching has been evaluated in [7]. Nowadays cut-through switching has been implemented[8] and available in the market[9]. Some test results[10] show that cut-through switching can significantly improve delay performance in limited environments such as datacenter networks.

In wireless networks, as the techniques of interference cancellation advance, there has been much effort to exploit cut-through switching to reduce wireless relay delay. In [11], the authors proposed shifting time alignment for relay through cut-through switching. The authors have shown that by shifting the alignment of frame duration, the forwarding latency can reduce by a half. Wireless interference constraints have been further studied in [12, 17, 18], where the authors considers simultaneous bidirectional wireless transmission denoted by full-duplex transmission, and studies complex interference relationship between transmissions. In [13], a wireless network architecture has been proposed to enable cut-through switching in multi-channel wireless environments.

The previous works rely heavily on limited network environments, and applicable only to a few network scenarios, e.g., datacenter networks. In this work, we investigate the end-to-end behavior of cut-through switching, and develop an analysis framework that can be extended to more general scenarios; specifically, mobile network environments where the traffic enters backbone networks through access links. Through approximation with reasonable assumptions, we can successfully obtain a closed-form result for the end-to-end delay performance, and evaluated the results through simulations. Our main contribution includes:

- Develop a new queueing framework to understand forwarding performance under cut-through switching.
- Design an approximated queueing model to capture the essential features of cut-through switching.
- Estimate packet delays under cut-through switching with high accuracy and low complexity.

The paper is organized as follows. We overview cut-through switching technology, describe our system, and briefly overview previous analytical results to understand the delay performance of cut-through switching in Section II. We provide Markov chain models that capture the system behaviors with accuracy and develop an analysis framework that provides approximated delay performance under cut-through switching in Section III. After evaluation of our model through simulations in Section IV, we conclude our paper in Section V.

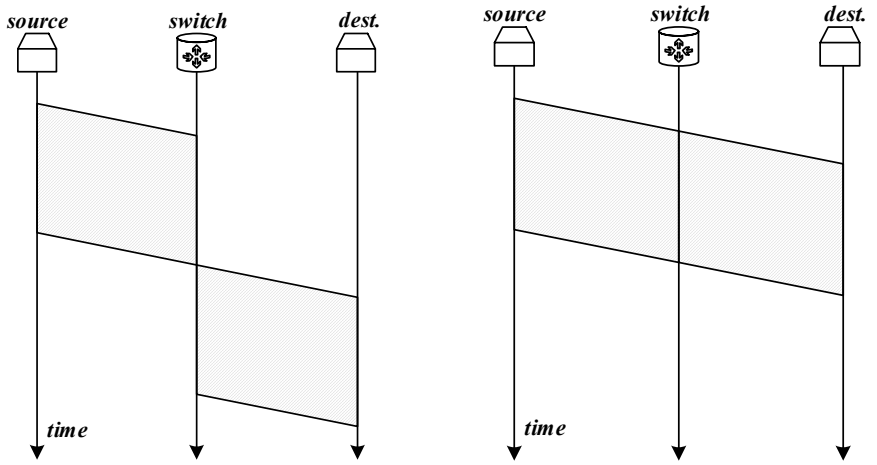
www.dbpia.co.kr

Ⅱ. Cut-through Switching and System Model

Cut-through switching is similar to message switching or packet switching, except that when a packet (or a frame) arrives in an intermediate node and its corresponding outgoing link is available, then the packet starts being forwarded immediately without waiting for complete reception. In comparison with conventional store-and-forward switching that incurs 5.7 usec on 1 Gbps link to transmit a packet of 500 byte at every hop[14], cut-through switching can reduce transmission latency significantly as shown in Fig. 1. In case of multi-hop forwarding, the differences in delay performance will become more significant as it is accumulated over hops.

There are two different types of cut-through transmission: a full cut and a partial cut. Suppose that an incoming packet finds non-empty queue, and before it is completely received, the queue becomes empty. Then the partially-received packet can be partially cut-through transmitted, which is called *partial cut*[4]. In practice, however, it would be hard to change the transmission mode during the packet reception, and the partially received packet should wait until full reception before being forwarded, which is called *full cut*. Fig. 2 shows the difference between the partial cut and the full cut, when two packets P1 and P2 (from different sources) arrive in order and their receiving times are partially overlapped. Under the full cut, the second packet P2 should wait until its full reception. Since current commercial cut-through switches support only full cut, we assume full cut-through. Performance evaluation for partial cut-through switching remains as an interesting open problem.

Another constraint of cut-through switching is *rate-mismatching*[8]. Cut-through switches on the market require that the rate of incoming port be matched with that of outgoing port. Specifically, cut-through transmission is available only if the incoming link rate equals to the outgoing link rate, and otherwise, the cut-through function is disabled. Note that the rate-mismatching is common at the edge of networks, i.e., in a relatively large network that consists of backbone links and access links with different rates. Aiming to understand the performance of cut-through in such environments, we assume that cut-through switching can support rate-mismatching with restriction that the outgoing transmission rate is no greater than the incoming

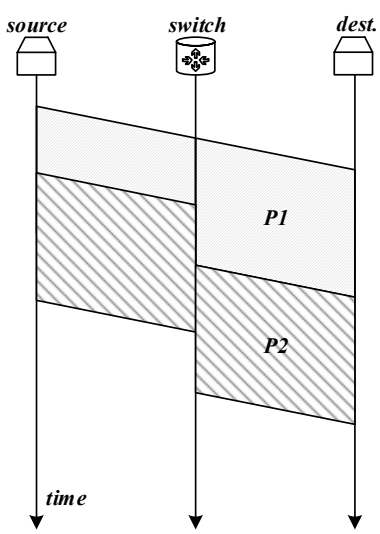


(a) Store-and-Forward switching
(a) Store-and-Forward 스위칭

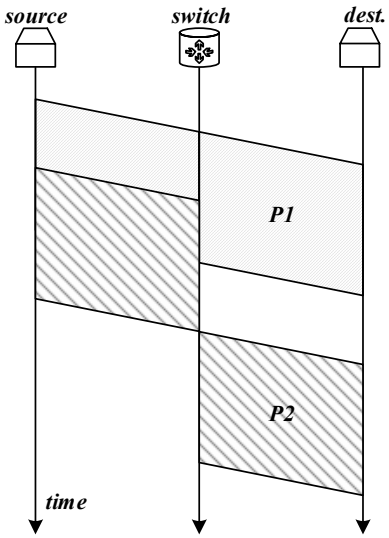
(b) Cut-through switching
(b) Cut-through 스위칭

그림 1. 일반적인 store-and-forward 스위칭과 저지연 cut-through 스위칭
Fig. 1. Conventional store-and-forward and low-latency cut-through switching

www.dbpia.co.kr



(a) Partial cut of P2
(a) P2가 부분 컷이 되는 경우



(b) Full cut (non-partial cut of P2)
(b) P2가 전체 컷이 되는 경우

그림 2. 두 가지 cut-through 스위칭: 부분 컷과 전체 컷
Fig. 2. Two different cut-through switching: partial cut and full cut

transmission rate, i.e., the outgoing link may transmit at a lower rate than its capacity under cut-through switching.

We consider a backbone network with graph $G=(N,E)$, where N denotes the set of nodes and E denotes the set of edges. For each flow s, packets are generated at user u_s, following a Poisson distribution with arrival rate λ_s, and their length is exponentially distributed with mean 1. The packets of flow s enter the network through ingress node n_s with access rate b_s, and arrive at their destination d_s through M backbone links as shown in Fig. 3. We assume that the route is fixed, each backbone link l has identical bandwidth $\mu_l=\mu$ and infinite buffer capacity. We define the end-to-end

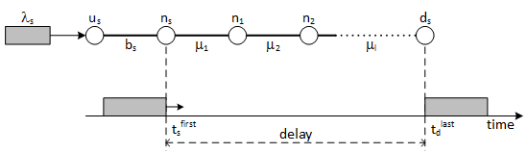


그림 3. 고정된 라우트 위에서 패킷의 도착과 패킷 지연시간의 정의
Fig. 3. Packet arrival with a fixed route and the definition of delay

(network) transmission delay of a packet as the interval from the time t_d^{first} when the first-bit of the packet arrives at ingress node n_s, to the time t_s^{end} when the last-bit of the packet arrives at the destination, i.e., $t_s^{last}-t_d^{first}$. This is equivalent to the typical definition of transmission delay. As in [1], we adopt the assumption of Jackson networks, where all the queues evolve independently in a Markovian manner, which typically holds under high-level traffic multiplexing (see [1] and references therein). Also, we assume that the propagation delay is negligible.

2.1 Delay performance of Store-and-Forwarding transmission

We start with basic performance evaluation of the system under conventional store-and-forwarding transmissions. It provides a good starting point, and also serves as the reference to highlight the novelty of our proposed framework. We consider delay performance of a particular flow s, and for ease of exposition, omit subscript s if there is no confusion.

For backbone link l, let K_l denote the set of

www.dbpia.co.kr

flows that have route on link l and $|K_l|$ denote its cardinality. Owing to the assumption of Jackson networks, packets arrive following a Poisson process with mean rate λ_l, which is the sum of flow arrival rates on the link (i.e., $\lambda_l = \sum_{s \in K_l} \lambda_s$), and packets have a length exponentially distributed with mean $\frac{1}{\mu_l}$. We can model the number of packets in service for each link as a Markov Chain. Under store-and-forwarding transmission, a state transition should occur when the last-bit of a packet arrives or departs the link (or the associated queue). From the traditional M/M/1 queueing model, the stationary probability p_n that there are n packets in the queue can be obtained as $p_n = (1-\rho_l)\rho_l^n$ where $\rho_l := \frac{\lambda_l}{\mu_l}$. Then average packet delay $\overline{T}_l^{SF}$ of link l (from the last-bit arrival to the last-bit departure) is known[19] as

$$\begin{aligned} \overline{T}_l^{SF} &= p_0 \cdot \frac{1}{\mu_l} + p_1 \cdot \frac{2}{\mu_l} + \cdots \\ &= \sum_{n=0}^{\infty} p_n \cdot \frac{n+1}{\mu_l} = \frac{1}{\mu_l - \lambda_l} \end{aligned} . \quad (1)$$

From the assumption of Jackson networks, the end-to-end packet delay $\overline{T}^{SF}$, from the departure of the first bit at the ingress node to the arrival of the last bit at the destination, can be obtained as

$$\overline{T}^{SF} = \frac{1}{b} + \sum_l \overline{T}_l^{SF}, \quad (2)$$

where the first term denotes the time between the first-bit arrival and the last-bit arrival at the ingress node through access link, which equals to average packet length 1 divided by access rate b. If packets go through statistically identical M backbone links, we have $\overline{T}_l^{SF} = \frac{1}{\mu_l - \lambda_l}$ and

$$\overline{T}^{SF} = \frac{1}{b} + \frac{K}{\mu - \lambda}.$$

2.2 Extension to cut-through switching

Let us consider an ideal cut-through system under which a packet starts being forwarded upon its first-bit arrival if the queue is empty, denoted as immediate forwarding model S. In practice, before forwarding, we need to receive the header of a packet to find the corresponding next hop. However, it is applicable when the reception time for the header is negligible comparing to the reception time of the whole packet, and also shows the achievable performance gain of cut-through forwarding. For further accurate estimation, we may subtract the header-reading time in its calculation.

In [1], the authors have analyzed the performance gain of cut-through transmissions based on traditional queueing models. It is worth giving a brief description. Under the well-known Jackson's assumption, the end-to-end delay under store-and-forwarding transmission can be obtained from the well-known Little's law. Hence, the authors calculate average per-link delay gain of cut-through transmission and subtract it from forwarding delay at each link. To elaborate, if the packets arrive following a Poisson process under the assumption of Jackson networks, then due to PASTA, the probability that a packet is cut-through transmitted is $p_o = 1 - \frac{\lambda_l}{\mu_l}$, and the delay gain per cut-through forwarding is $\frac{1}{\mu_l}$. At each link l, the delay gain of cut-through is $\left(1 - \frac{\lambda_l}{\mu_l}\right)\frac{1}{\mu_l}$. Hence, the end-to-end delay $\overline{T}^{CT,old}$ can be obtained as

$$\overline{T}^{CT,old} = \frac{1}{b} + \sum_l \left(\overline{T}_l^{SF} - \left(1 - \frac{\lambda_l}{\mu_l}\right)\frac{1}{\mu_l} \right). \quad (3)$$

Although this result takes into account the delay gain of cut-through transmission, it assumes that all the packets can be transmitted at rate μ at each backbone link, which is not true due to the causality

www.dbpia.co.kr

constraint. For example, suppose that all the queues are empty. When a packet is generated and transmitted to n_s at access rate b, it will be cut-through forwarded to n_1, and the forwarding rate should be b since the outgoing rate cannot be higher than the incoming rate. Such network is called as not balanced, e.g., the network has different rate (or different rate share) for the incoming link and the outgoing link. This restricts the application of (3) to a very limited set of scenarios.

Ⅲ. Performance Models for Cut-through Switching

We develop a precise Markovian model of system S under cut-through switching, and then approximate it to estimate packet delays with low complexity. For simplicity, we also omit subscript l if there is no confusion.

3.1 State probability distribution

We first note that when the network is not balanced, the Markov Chain cannot be fully described by the queue length as in the traditional analysis. For example, for a backbone link (n_s, n_1) shown in Fig. 3, we consider the following scenario with $t_0 < t_1 < t_2 < t_3$. Suppose that the first bit of a packet arrives at t_0 and it is forwarded to n_1 by cut-through transmission with access rate b due to the causality constraint. At t_1, another packet's first-bit arrives to the link (from different flow). The packet has to be queued up due to the current on-going cut-through transmission. At t_2, the second packet has been received completely, and at t_3, the on-going cut-through transmission finished, and the second packet starts being transmitted at rate μ. In this scenario, we can find only one packet in the system during time interval of $[t_0, t_1]$ and after t_3, and thus will be represented by the same state in the previous models. However, during the two intervals, the system has a different behavior: for the former, the link transmission rate is b, and for the latter, it is μ. This implies that unlike typical queueing models[15, 16], the state cannot be fully described only by the number of packets in the system.

We develop a Markov Chain model that can fully describe the state of a cut-through capable link. We describe the state of a link with three variables (x, y, z) as shown in Fig. 4, where $x \in \{0, 1\}$ denotes on-going cut-through transmission, where $x = 0$ if cut-through switching is active and $x = 1$ otherwise, $y \in \mathbb{N}$ denotes the number of receiving packets, and $z \in \mathbb{N}$ denotes the number of packets that complete reception and wait in the queue. In the previous example, the state in $[t_0, t_1]$ is (1,1,0), and transits to (1,2,0) at t_1, to (1,1,1) at t_2, and to (0,0,1) at t_3. We emphasize that, in this model, the state transitions are triggered either by the first-bit arrival of a packet or by the last-bit departure.

Although the model fully describes the behavior of the system, its evolution is quite complicated, and it is difficult to obtain a closed-form result. To this end, we develop a low-complexity approximation model that captures the essential features of the system, while providing accurate system performance. We simplify the model with the following assumptions.

- Assumption 1[1]: If a packet arrival finds non-empty queue, it is assumed that the whole packet is immediately received.

Under this assumption, the system can be described without parameter y, i.e., by (x, z), as shown in Fig. 5. Note that it may result in

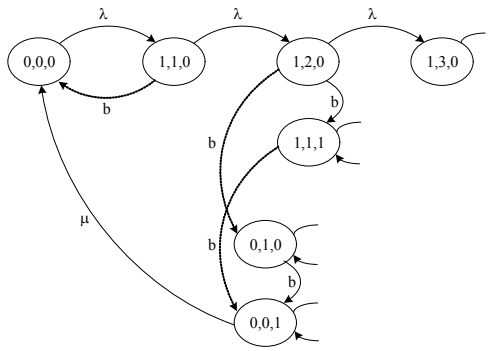


그림 4. Cut-through 시스템의 마르코프 체인 모델
Fig. 4. Markov Chain model of cut-through system.

www.dbpia.co.kr

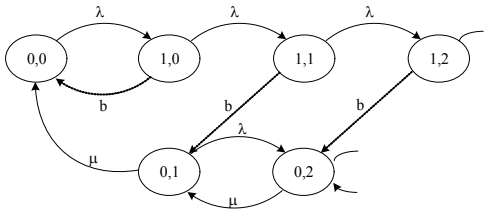


그림 5. Cut-through 시스템 근사화를 위한 마르코프 체인 (중간) 모델
Fig. 5. Intermediate Markov Chain model of approximated cut-through system

underestimation of packet delays, since the model causes early reception for some packets than the actual system. However, we claim that the error is insignificant, since a mismatch occurs only when the first bit of a packet arrives in the middle of on-going cut-through transmission and the cut-through transmission ends before the complete reception of the incoming packet.

- Assumption 2: Each server has dynamic service rate, specifically, μ if there are more than 1 packet in the system, and b if there is only 1 packet in the system.

The second assumption effectively changes the service rate of state (0,1) in Fig. 5 to b and the service rate of each $(1,j)$ to μ for all $j>1$. Let us consider a sequence of events that the queue becomes empty, and a time interval characterized by two consecutive queue-empty events. All the packets in the interval will be transmitted at rate μ except the first packet that is transmitted by cut-through switching at rate b. In terms of the delay performance of these packets, we may consider that all packets are served at rate μ except the last packet that is served at rate b, which does not change the total delay sum. We incorporate this idea into the approximate state transition diagram of Fig. 6. We note that, however, this model has additional approximation errors in calculating the steady-state distribution. We emphasize that the Markov Chain in Fig. 6 is different from the conventional Markov Chain based on the last-bit arrival and the last-bit departure. Our model presents the number of the packets whose first bit arrives at the queue and last bit does not leave the queue yet in the system. In

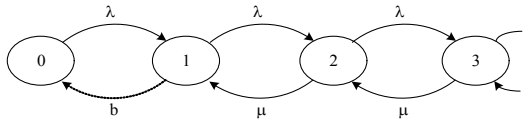


그림 6. Cut-through 시스템을 위한 근사화된 마르코프 체인 모델 (최종)
Fig. 6. Final Markov Chain model of approximated cut-through system.

the sequel, we denote it by the number of partial packets. Also, we denote the approximation model under Assumptions 1 and 2 by S'.

We consider the steady-state probability of system S'. The probability p_n that a packet arrival finds n partial packets in the system S' can be obtained from the standard Markov Chain analysis as

$$p_n = \begin{cases} p_{0,} & n=0, \\ p_0 \cdot c \cdot \rho^n, & n \geq 1, \end{cases}$$

where $c := \frac{\mu}{b}$ and $p_0 := \frac{1}{1+c\rho/(1-\rho)}$. We use it to estimate the number of partial packets in the system when a packet of interest arrives.

In contrast to the steady-state probability, the calculation of average packet delay under S' is different from that under the conventional Markov Chain, since our model describes the system behavior based on the first-bit arrival as well as the last-bit departure. In the following, we calculate average delay performance of cut-through systems.

We consider the waiting time of a packet in S' before its first-bit departure. Accurate calculation of the waiting time is challenging due to the complex operation of cut-through switching when the service rates are different for access links and backbone links. To this end, we approximate the delay performance under the following assumption:

- **Assumption 3:** When a packet finds $n(\geq 1)$ packets in the system, it waits for $(n-1)$ normal transmissions and 1 cut-through transmission.

Note that a packet can observe at most one cut-through transmission under Assumption 1, while it waits for service, and some packets may observe

www.dbpia.co.kr

no cut-through transmission. Thus, Assumption 3 overestimates the waiting time. Nonetheless, we claim that it well approximates packet delay for the following reasons. i) Under light traffic loads, the estimation will be close to the actual value since cut-through transmissions will occur frequently. ii) Under heavy traffic, the queueing delay will dominate and the delay due to few cut-through transmissions is negligible.

Under Assumption 3, expected packet delay $\overline{T}_l^{CT,new}$ of link l (from the first-bit arrival to the first-bit departure) can be written as

$$\overline{T}_l^{CT,\neq w} = p_0 \cdot 0 + \sum_{n=1}^{\infty} p_n \cdot \left(\frac{1}{b} + \frac{n-1}{\mu}\right) = \frac{1-p_0}{b} + \frac{p_0}{b} \cdot \frac{\rho^2}{(1-\rho)^2} \quad (4)$$

Hence, we can write the end-to-end packet delay $\overline{T}_l^{CT,new}$ as

$$\overline{T}^{CT,\neq w} = \sum_l \overline{T}_l^{CT,\neq w} + \left(\frac{1}{b} \cdot p_0^{last} + \frac{1}{\mu^{last}} \cdot (1-p_0^{last})\right), \quad (5)$$

where p_0^{last} denotes the probability that the last-hop queue is empty and μ^{last} denotes the service rate of the last-hop link. The last term in the parenthesis takes into account the transmission time at the last-hop link, i.e., from the time when the first bit arrives at the destination to the time when the last bit arrives at the destination. The last-hop link rate depends on the system state when the last-bit of the packet departs, and from the reversibility of the Markov Chain, the queue is empty with probability p_0^{last} and non-empty with probability $1-p_0^{last}$. For the network with statistically identical M links, we have that

$$\overline{T}^{CT,\neq w} = M \cdot \left(\frac{1-p_0}{b} + \frac{p_0}{b} \cdot \frac{\rho^2}{(1-\rho)^2}\right) + \left(\frac{1}{b} \cdot p_0 + \frac{1}{\mu} \cdot (1-p_0)\right). \quad (6)$$

It is interesting to see different methods in calculating (3) and (6). In (3), the delay is computed based on the last-bit arrivals and departures at each node and the transmission delay $\frac{1}{b}$ is calculated at the access link. In contrast, we compute the delay based on the first-bit arrivals and departures at each node and the transmission delay is calculated at the last-hop link. By taking the latter approach, we could obtain more accurate delay analysis results as shown in the next section.

Ⅳ. Numerical Result

We provide preliminary evaluation of our model through simulations. We consider a network with $M=3$ backbone links of identical rate $\mu = 10\,\mathrm{Kbps}$ as shown in Fig. 7. There are 4 independent flows, each of which has Poisson packet arrivals with the same mean λ and inputs packets to the backbone network through a dedicated access link of rate b. Packet length is exponentially distributed with mean 400 bytes. We fix the routes of flows such that $|K_l| = 2$ for all links l. In our simulations, a packet is forwarded at the incoming link rate if it finds empty queue (i.e., rate-mismatching is allowed). Otherwise, it waits until the reception is completed, and then forwarded at the outgoing link rate (i.e., full-cut cut-through). We measure average packet delay of flow 0 whose packets are forwarded through multiple wired links. We run simulation 10 times, each for 0.8 second, and present their average. The results for longer simulation time are similar and thus omitted.

Fig. 8 shows the delay performance when $b = \mu = 10\,\mathrm{Kbps}$. We can observe that, for

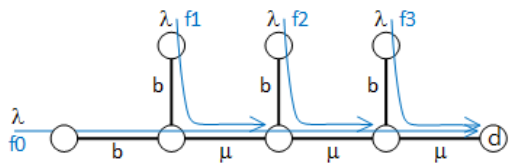


그림 7. 선도적 실험을 위한 네트워크 토폴로지
Fig. 7. Network topology for preliminary simulations.

www.dbpia.co.kr

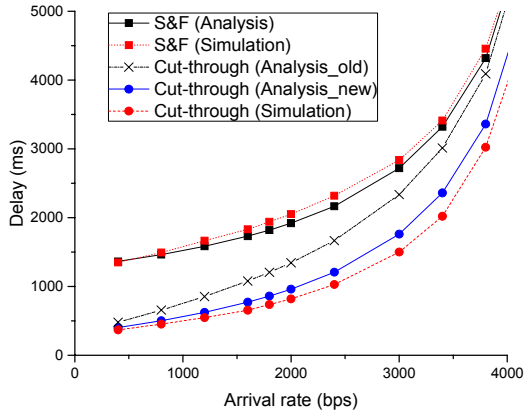


그림 8. 분석과 실험 결과 비교 ($b=10\,\mathrm{Kbps}$).
Fig. 8. Comparison of analysis results and simulation results, when $b=10\,\mathrm{Kbps}$.

store-and-forwarding (S&F), the analysis results well match the simulation results. Similarly, under cut-through forwarding (Cut-through), our analysis results (Analysis_new) are very close to the simulation results, and the previous approach of [1] (Analysis_old) shows substantial differences when $\lambda > 1000\,\mathrm{bps}$. The results not only verify our model, but also show that cut-through switching significantly outperforms store-and-forward switching.

We further compare the delay performance of store-and-forward switching and cut-through switching with different number of hops. Fig. 9

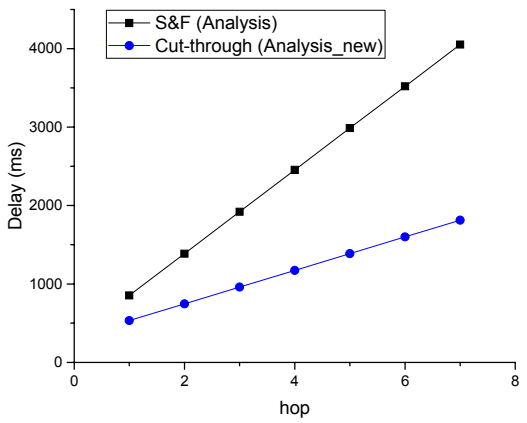


그림 9. Store-and-forward 스위칭과 cut-through 스위칭의 성능 비교
Fig. 9. Performance comparison of store-and-forward switching and cut-through switching with different number of hops.

illustrates that average delay of both store-and-forward switching and cut-through switching increases linearly. However, the performance gap between them enlarges as the number of hop increases as expected.

Fig. 10 shows the delay performance of both switching schemes (with $M=3$) with access rate $b=5\,\mathrm{Kbps}$, where a flow has the same share $\frac{\mu}{|K_l|}$ of backbone link rate throughout its route. Hence, it is a balanced network as the case studied in [1]. In this case, both our analysis results and the results of [1] provide accurate delay estimation as expected. Note that at the ingress node, the physical port rates are mismatched, which incurs inefficiency in cut-through transmissions and reduces the performance gain of cut-through switching. As arrival rate increases, more packets will be forwarded in the store-and-forwarding manner, and thus the delay gain of cut-through reduces.

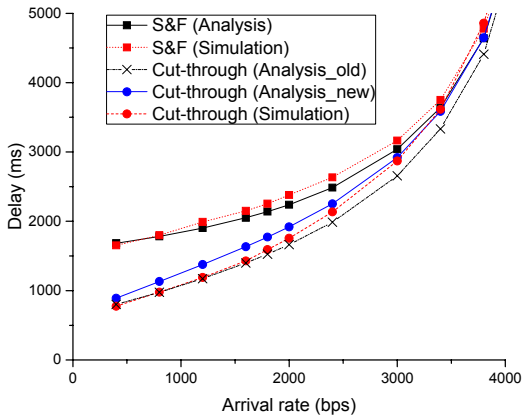


그림 10. 분석과 실험 결과 비교 ($b=5\,\mathrm{Kbps}$).
Fig. 10. Comparison of analysis results and simulation results, when $b=5\,\mathrm{Kbps}$.

V. Conclusion

In this work, we provide an analysis framework to evaluate the delay performance of cut-through transmission. We develop a queueing-based framework to understand cut-through switching, and then estimate packet delays under approximated model with reasonable assumptions. Unlike the

www.dbpia.co.kr

known results in balanced-network scenarios, our results can be applied to more general cases with the settings of backbone networks with access links. We validate our model through simulations.

References

[1] P. Kermani and L. Kleinrock, "Virtual cut-through: A new computer communication switching technique," *Comput. Netw.*, vol. 3, no. 4, pp. 267-286, Sep. 1979.

[2] M. A. Lema, A. Laya, T. Mahmoodi, M. Cuevas, J. Sachs, J. Markendahl, and M. Dohler, "Business case and technology analysis for 5G low latency applications," *CoRR*, vol. abs/1703.09434, 2017. [Online]. Available: http://arxiv.org/abs/1703.09434

[3] B. Shin and C. Un, "Performance analysis of a Quasi-M/D/1 cut-through switching network with noisy channels," *IEEE Trans. Commun.,* vol. 34, no. 9, pp. 882-889, Sep. 1986.

[4] M. Ilyas and H. M. Ouftah, "Quasi cut-through: New hybrid switching technique for computer communication networks," in *ZEE Proc.*, vol. 131, part E, Jan. 1984.

[5] I. Widjaja, A. Leon-Garcia, and H. Mouftah, "The effect of cut-through switching on the performance of buffered banyan networks," *Comput. Netw. and ISDN Syst.*, vol. 26, no. 1, Sep. 1993.

[6] N. M. A. Ayad and F. A. Mohamed, "Performance analysis of a cut-through vs. packet-switching techniques," in *IEEE Symp. Comput. and Commun.*, Jul. 1997.

[7] R. G. Bubenik and J. S. Turner, "Performance of a broadcast packet switch," *IEEE Trans. Commun.,* vol. 37, no. 1, Jan. 1989.

[8] "*Cisco to acquire kalpana, leading ethernet s witching company*," Retrieved Jul. 16, 2018, ht tp://www.webcitation.org/5qaWOQRdn?url=htt p://newsroom.cisco.com/dlls/1994/corp_10249 4.html.

[9] "*Cut-through and wtore-and-forward ethernet switching for low-latency environments*," Retrieved Jul. 16, 2018, http://www.cisco.co m/c/en/us/products/collateral/switches/nexus-50 20-switch/white_paper_c11-465436.html

[10] "*Arista, Blade share top spot in switch test*," Jan. 2010.

[11] W. K. Jia and Y. C. Chen, "A cut-through forwarding scheme for delay optimization in IEEE 802.16j simultaneous transmit and receive multihop relay networks," in *Int. Conf. Netw. and Distrib. Comput.*, Oct. 2010.

[12] Y. Yang and N. B. Shroff, "Scheduling in wireless networks with full-duplex cut-through transmission," in *IEEE INFOCOM*, Apr. 2015.

[13] R. McTasney, D. Grunwald, and D. Sicker, "Low-latency multichannel cut-through vs. CSMA/CA wireless mesh networking," in *IEEE MILCOM*, Nov. 2008.

[14] "*Ethernet v. Infiniband*," Retrieved Jul. 16, 2018, https://www.informatix-sol.com/docs/Eth ernetvInfiniBand.pdf

[15] Y. Kim, "Queueing performance analysis of CDF-Based scheduling over markov fading channels," *J. KICS*, vol. 41, no. 10, Oct. 2016.

[16] J. Choi, S. Ahn, and S. Lee, "Effective vehicle queue length estimation mechanism based on V2I communications," *J. KICS*, vol. 43, no. 4, Apr. 2018.

[17] C. Nam, C. Joo, and S. Bahk, "Joint subcarrier assignment and power allocation in full-duplex OFDMA networks," *IEEE Trans. Wireless Commun.(TWC)*, vol. 14, no. 6, Jun. 2015.

[18] C. Nam, C. Joo, S. G. Yoon, and S. Bahk, "Resource allocation in full-duplex OFDMA networks: Approaches for full and limited CSIs," *J. Commun. and Netw. (JCN)*, vol. 18, no. 6, Dec. 2016.

[19] C. Joo and N. B. Shroff, "A novel coupled queueing model to control traffic via QoS-Aware collision pricing in cognitive radio networks," *IEEE INFOCOM*, May 2017.

[20] M. Kim and C. Joo, "Prediction-based reliable data forwarding method in VANET," *J. KICS*, vol. 42, no. 1, 2017.

www.dbpia.co.kr

주 창 희 (Changhee Joo)

1998년 2월 : 서울대학교 전기공학부 졸업
2000년 2월 : 서울대학교 전기공학부 석사
2005년 2월 : 서울대학교 전기공학부 박사

<관심분야> 컴퓨터 네트워크

이 승 현 (Seunghyun Lee)

2014년 2월 : 울산과학기술원 전기전자컴퓨터공학부 졸업
2016년 8월 : 울산과학기술원 전기전자컴퓨터공학부 석사
2016년 9월~현재 : 울산과학기술원 전기전자컴퓨터공학부 박사과정

<관심분야> 컴퓨터 네트워크

www.dbpia.co.kr