Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2024.47.3.001

Vol. 47, No. 3 : 1-7, September 2024
ISSN : 2005-0461 / EISSN : 2287-7975

Reexamining Net Neutrality in a Queuing Model

Jeong-Yoo Kim* · Seung J. Noh**†

*Department of Economics, Kyung Hee University
**Department of Business, Myongji University

대기모형을 이용한 망중립성 효과 분석

김정유* · 노승종**†

*경희대학교 경제학과
**명지대학교 경영학과

In an influential paper, Choi and Kim (2010) derived waiting times in an $M/M/1$ queuing model under net neurality and under prioritization. In this short paper, we argue that the waiting times of content transmission that Choi and Kim (2010) derived by using the $M/M/1$ gueuing model under the non-preemptive priority rule are miscalculated. We provide corrected waiting times in the $M/M/1$ queuing model in the prioritization case. We also show that this correction does not affect their main results on the delay time and the incentive to invest in the network capacity qualitatively.

Keywords : Net Neutrality, Queuing Theory, Congestion, Delay, Invariance Result

1. Introduction

In 2017, U.S. Federal Communications Commission (FCC) dismantled the net neutrality regulations that prohibited Internet service providers (ISP) from blocking websites or charging for higher-quality service or certain content. This means that ISPs are given free rein to deliver service at their own discretion, that is, to discriminate contents by blocking, throttling and prioritization etc.

There are many important static and dynamic issues regarding net neutrality. First of all, it is an important issue whether repealing net neutrality and giving priorities to some content providers may help alleviate delays in content transmission due to heavy traffic. From a dynamic perspective, it is also important to examine whether ISPs will have a stronger incentive to improve the transmission quality by increasing the network capacity.

For years, queuing models have been adopted to address congestion problems in areas of priority pricing and network neutrality, since Choi and Kim [2] and Cheng et al.[1]. It is well known that queuing models give a good approximation for the arrival process in communication networks where many service requests compete transmission under limited bandwidth. Among many queuing models, $M/M/1$ models have found the widest range of application in most areas. In $M/M/1$ system, Poisson arrivals with rate λ are assumed. That is, interarrival times between two adjacent arrivals are exponentially distributed with mean $1/\lambda$. This feature well suits the random nature inherent in the consumer requests in communications networks. The system has one server, service times of which are exponentially distributed with mean

Received 23 June 2024; Finally Revised 25 July 2024;
Accepted 26 July 2024.
† Corresponding Author : sjnoh@mju.ac.kr

$1/\mu$. The term μ is thus the service rate, or the average number of services performed in unit time. Choi and Kim [2] interpret it as bandwidth in communications networks.[1)]

By using an $M/M/1$ model, Choi and Kim [2] derive the waiting time elegantly and beautifully and established the invariance result that given a fixed network capacity, the average waiting times are identical regardless of net neutrality. They also showed that the ISP's incentive to invest may be weaker under prioritization, because an increase in the network capacity reduces the relative value of prioritized contents. They call this the rent extraction effect.

In this short paper, we argue that the waiting times of content transmission that Choi and Kim [2] derived by using the $M/M/1$ queuing model under the non-preemptive priority rule are miscalculated. We will provide corrected waiting times in the $M/M/1$ queuing model in the prioritization case. We also show that this correction does not alter their static and dynamic results on the delay time and the incentive to invest in the network capacity qualitatively.

2. Model

We closely follow the model of Choi and Kim [2]. End users are uniformly distributed over [0, 1]. So, the mass of end users is normalized to one. CP1 and CP2 are located in $x=0$ and $x=1$ respectively. Each consumer request contents from one of the content providers and gets some valuation v. The unit transportation cost is t.

A monopolistic Internet service provider (ISP) has n servers with identical capacity.

Under net neutrality, both the interarrival time of content requests and the service time of each server of the ISP are assumed to follow exponential distributions with λ and μ. That is, the mean of the time between content requests is $1/\lambda$ and the mean of the service time is $1/\mu$. As usual, we assume that $\mu > \lambda$ to avoid the possibility that the waiting time will explode. Under no net neutrality, let λ_1 and λ_2 be the rate parameters of the exponential distributions that the interarrival times of content requests from CP1 and CP2 follow respectively.

Our main interest in this paper is how the repeal of net neutrality could affect the average waiting time (transmission time) of data that end users request. To compute the waiting times under two different regimes (net neutrality vs. no net neutrality), we borrow some established results in queuing theory.

3. Preliminaries

We consider a system in which customers arrive at rate λ according to a Poisson process and a server serves customers one at a time from the front of the queue on a first-come, first-served basis. Service times of each server have an exponential distribution with rate parameter μ.

It is useful to define server utilization by $\rho=\lambda/\mu$ to characterize waiting times. It has an interpretation as the probability that a server is occupied at arbitrary point in time. We assume that $0<\rho<1$. Otherwise, the number of customers in the system, and in turn the average waiting time will eventually explode.

No Priority The average number of customers in the system, denoted by L, is calculated as $L=\sum_{k=0}^{\infty} k p_k$, where p_k is the probability that there are k customers in the system at arbitrary point in time. It is well known for $M/M/1$ that $\{p_k\}$ is a geometric sequence with the common ratio ρ,[2)] and that the resulting sum is

$$L=\frac{\rho}{1-\rho}=\frac{\lambda}{\mu-\lambda} \quad . \tag{1}$$

The average waiting time in system, denoted by W, is obtained using Little's Law, $L=\lambda W$, as

$$W=\frac{L}{\lambda}=\frac{1}{\mu-\lambda} \quad . \tag{2}$$

Note that W is the average total time spent in the system.[3)]

1) Economides and Hermalin [3] model bandwidth and prioritized service in a different way. Instead of a queuing model, they use a sub-bandwidth model in which ISP allocates a wider sub-bandwidth to a prioritized CP. In other words, unlike the $M/M/1$ queuing model in which the ISP processes prioritized contents before unprioritized contents, prioritized contents and unprioritized contents are assumed to use separate portions of the bandwidth (fast lane and slow lane, respectively) in their model.

2) See, for example, Gross and Harris [4].

3) W is also called system time, sojourn time, or response time, etc. in literature.

Since a customer has to wait not only in queue but also for his own service time, the average total waiting time consists of two components: the average time in queue, W_q, and the average time in service, W_s. That is, $W = W_q + W_s$. From this relation, we can compute W alternatively as:

$$W = W_q + W_s = \frac{L}{\mu} + \frac{1}{\mu} = \frac{1}{\mu - \lambda} \quad . \tag{3}$$

The first term, L/μ, is computed by using the PASTA (Poisson Arrivals See Time Averages) property saying that an arrival (arriving customer) from a Poisson process sees L customers on average, each of which is expected to have service time with mean $1/\mu$.

Since $W_s = 1/\mu$, we can obtain the average time in queue as

$$W_q = W - W_s = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu - \lambda} . \tag{4}$$

Little's Law holds in any queuing system, and particularly, it applies to systems within a system. Since queue (waiting line) itself is one subsystem and a server is another, Little's Law could be applied to each one, as well as the whole system.

The intuition underlying behind Little's Law is quite clear. It simply states that the average number of customers in a system or a subsystem (total quantity) must be equal to the average arrival rate of the customers (speed), multiplied by the average time that a customer spends in that system (total time). It always holds regardless of the interarrival time distribution or service time distribution, nor does it depend upon the number of servers in the system or queuing discipline within the system. For example, if we apply Little's Law to the waiting time (queue) only, we can find the average number of customers in queue as

$$L_q = \lambda W_q = \frac{\rho^2}{1 - \rho} \quad . \tag{5}$$

This implies that we can also find the average number of customers who are served as

$$L_s = L - L_q = \frac{\rho}{1 - \rho} - \frac{\rho^2}{1 - \rho} = \rho \ . \tag{6}$$

Priority Now, we consider an $M/M/1$ system serving two different types of customers, type 1 and type 2. Type 1 and type 2 customers arrive according to independent Poisson processes with rate λ_1 and λ_2 respectively. We assume that $\rho_1 + \rho_2 < 1$, where $\rho_i = \lambda_1/\mu$, i.e., the occupation rate by type i customers. Type 1 customers are prioritized over type 2 customers.

First, we assume the non-preemptive priority system *à la* Choi and Kim [2]. Under this priority system, even prioritized type 1 customers are not allowed to interrupt the service of unprioritized type 2 customers.

The average waiting time for prioritized type 1 customers is computed as

$$W_1 = \frac{L_1}{\mu} + \frac{\rho_2}{\mu} + \frac{1}{\mu} \ , \tag{7}$$

where L_i is the average number of type i customers in the system. The first term is the waiting time for the services for type 1 customers to be completed, i.e., the service time type 1 customers in the system. This is again due to the PASTA property. The second term comes from the fact that when an arriving type 1 customer finds a type 2 customer in service, he has to wait until the service of the type 2 customer is completed. According to the PASTA property, the probability that he finds a type 2 customer in service is equal to the fraction of time the server spends on type 2 customer, ρ_2. The third term is the average service time of his own. Thus, we can rewrite (7) as

$$W_1 = W_{q1} + W_{s1} . \tag{8}$$

where $W_{q1} = \frac{L_1}{\mu} + \frac{\rho_2}{\mu}$ and $W_{s1} = \frac{1}{\mu}$.

Equation (7), together with Little's Law given by $L_1 = \lambda_1 W_1$, leads to

$$L_1 = \frac{(1 + \rho_2)\rho_1}{1 - \rho_1} , \tag{9}$$

$$W_1 = \frac{(1 + \rho_2)/\mu}{1 - \rho_1} . \tag{10}$$

Equation (10) implies that the waiting time of prioritized customers must depend on λ_2. This is due to the second term of (7). Intuitively, a newly arriving customer must wait in the front of the queue while a type 2 customer is in service (if any), even if the arriver is prioritized, under the non-pre-

emptive priority rule. Note that λ_2 does not enter equation (2) of Choi and Kim [2] that assumes the non-preemptive priority rule.

For the waiting time of unprioritized type 2 customers, we need the following relation:

$$L = L_1 + L_2 = \frac{\rho_1 + \rho_2}{1-\rho_1-\rho_2}, \tag{11}$$

which can be similarly derived as (1). Then, from (9) and (11), we obtain

$$L_2 = L - L_1 = \frac{(1-\rho_1(1-\rho_1-\rho_2))\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)}. \tag{12}$$

Therefore, by applying Little's Law we obtain the average waiting time of unprioritized customers as

$$W_2 = \frac{(1-\rho_1(1-\rho_1-\rho_2))/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}. \tag{13}$$

Equation (13) can be rewritten as

$$W_2 = \phi(\lambda_1, \lambda_2; \mu)\, W_1, \tag{14}$$

where $\phi(\lambda_1, \lambda_2; \mu) = \dfrac{\mu^2 - \lambda_1\mu + \lambda_1(\lambda_1+\lambda_2)}{\mu^2 - \lambda_1\mu - \lambda_2(\lambda_1+\lambda_2)} > 1$ for any λ_1, $\lambda_2 > 0$. This implies that $W_2 > W_1$ for any $\lambda_1, \lambda_2 > 0$.

On the other hand, in the preemptive priority rule under which a newly arriving type 1 customer is allowed to interrupt the service time of an unprioritized type 2 customer, the computation of the average waiting time for type 1 customers is relatively simple, because an arriving type 1 customer does not need to worry about the possibility that type 2 customers are in queue or in service. Therefore, the situation is the same for a type 1 customer as if there were only type 1 customers in the system. This leads to

$$L_1 = \frac{\rho_1}{1-\rho_1} = \frac{\lambda_1}{\mu-\lambda_1}, \tag{15}$$

$$W_1 = \frac{1/\mu}{1-\rho_1} = \frac{1}{\mu-\lambda_1}. \tag{16}$$

Then, by using (11), we get

$$L_2 = L - L_1 = \frac{\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)}. \tag{17}$$

By applying Little's Law, we obtain

$$W_2 = \frac{L_2}{\lambda_2} = \frac{\rho_2}{\lambda_2(1-\rho_1)(1-\rho_1-\rho_2)} = \frac{W_1}{(1-\rho_1-\rho_2)} = \frac{\mu}{\mu-\lambda} W_1 \tag{18}$$

where $\lambda = \lambda_1 + \lambda_2$.

While the waiting time of the prioritized customers depends on λ_2 in the non-preemptive priority rule, because they cannot interrupt the services for unprioritized customers, the waiting time for the prioritized customers does not depend on λ_2 in the preemptive priority rule, because they can always interrupt the service for type 2 customers and regard them as nonexistent null players.

The upshot is that under the non-preemptive priority rule which Choi and Kim [2] assumed, their result on the waiting times is not correct, while their result is correct if the preemptive priority rule is assumed.

4. The Effects of Prioritization on Waiting Times

Under net neutrality, all contents from either CP are equally treated by ISP. So, the average waiting time of contents from either CP is identically $W = \dfrac{1}{\mu-\lambda}$.

A consumer chooses to request contents from a CP that yields a lower total cost defined by the sum of waiting time and the transportation cost. Let x_N be the location of the consumer who is indifferent between two CPs. Then, it is clear that $x_N = \dfrac{1}{2}$, because the waiting times of CP1 and CP2 are the same.

If net neutrality is repealed, ISP can grant priority to either CP. We will consider both priority rules, the preemptive priority rule and the non-preemptive priority rule.

Let x_{PP} be the borderline consumer under the preemptive priority rule. Since x_{PP} is the ratio of users who request contents from CP1, the total amounts of content requests from CP1 and CP2 are $\lambda_1 = \lambda x_{PP}$, and $\lambda_2 = \lambda(1-x_{PP})$ respectively. Then x_{PP} is determined by

$$W_1^{PP} + t\tilde{x} = W_2^{PP} + t(1-x_{PP}), \tag{19}$$

where

$$W_1^{PP} = \frac{1}{\mu - \lambda_1}, \tag{20}$$

$$W_2^{PP} = \frac{W_2}{\mu - \lambda}. \tag{21}$$

This leads to

$$x_{PP} = \frac{1}{2} + \frac{1}{t}(W_2^{PP} - W_1^{PP}). \tag{22}$$

It is clear that $x_{PP} > \frac{1}{2}$, because $W_2^{PP} > W_1^{PP}$.

If we denote the average waiting time under the preemptive priority rule by W_{PP}, we have $W_{PP} = x_{PP} W_1^{PP} + (1 - x_{PP}) W_2^{PP}$. Then, the invariance result of Choi and Kim [2] follows.

Proposition 1. (Choi and Kim [2]). $W_{PP} = W$.

It says that the invariance result holds in the case of the preemptive priority rule.

Now, consider the non-preemptive priority rule. Let x_{NP} be the borderline consumer under the non-preemptive priority rule. Then, the total amounts of content requests from CP1 and CP2 are $\lambda_1 = \lambda x_{NP}$ and $\lambda_2 = \lambda(1 - x_{NP})$ respectively. Similarly, x_{NP} is determined by

$$x_{NP} = \frac{1}{2} + \frac{1}{t}(W_2^{NP} - W_1^{NP}), \tag{23}$$

where

$$W_1^{NP} = \frac{(1 + \rho_2)/\mu}{1 - \rho_1} = \frac{\mu + \lambda_2}{\mu(\mu - \lambda_1)}, \tag{24}$$

$$W_2^{NP} = \phi W_1^{NP}. \tag{25}$$

Similarly, the average waiting time under the non-preemptive priority rule is $W_{NP} = x_{NP} W_1^{NP} + W_2^{NP}(1 - x_{NP}) W_2^{NP}$. Then, we have

Proposition 2. $W_{NP} = W$.

This proposition says that the invariance result also holds in the case of the non-preemptive priority rule.

It is not surprising that the invariance result holds in both priority regimes, because the repeal of net neutrality simply affects the order of serving content requests as long as the total amount of traffic remains the same.[4]

5. The Effect on the Investment in the Network Capacity

Choi and Kim [2] showed that the ISP's incentive to invest may be weaker in the case of prioritization under the preemptive priority rule,[5] because an increase in the network capacity reduces the relative value of prioritized contents. They call this the rent extraction effect.[6]

Proposition 3. *In M/M/1 queuing model,* $\frac{d(W_2 - W_1)}{d\mu} < 0$ *under both the preemptive priority rule and the non-preemptive priority rule.*

This proposition says that as ISP invests more in network capacity, the quality difference between prioritized contents and unprioritized contents becomes smaller, i.e., the relative value of prioritized contents becomes lower. This implies that ISP will have less incentive to invest in the network capacity.

This result was first established by Choi and Kim [2] only for the case of the preemptive priority rule. In Appendix, we provide a proof for the case of the non-preemptive priority rule as well. Therefore, Proposition 3 complements their result in the sense that it extends the result to the case of the non-preemptive priority rule.

6. Conclusion

In this paper, we showed that the invariance result that Choi and Kim [2] obtained holds both under the preemptive rule and under the non-preemptive rule. We also showed that

4) Kim [5] argues that the invariance result does not hold if the request rate for prioritized contents is higher than the request rate for unprioritized contents or contest under net neutrality.

5) What they actually showed was that the effect on the investment in the network capacity due to the rent extraction effect is negative under the preemptive priority rule, not under the non-preemptive priority rule.

6) In addition to the rent extraction effect, they identified another effect of prioritization which they call access fee effect. If an ISP invests in network capacity, it can raise the access fee because it increases the utility of end users. Since the improvement in the transmission time differs across CPs under prioritization, the effect is ambiguous. So, we focus only on the rent extraction effect by assuming that the access fee is zero.

the rent extraction effect of an investment in the network capacity that Choi and Kim [2] identified is valid both under the preemptive rule and under the non-preemptive rule. We believe that it will be worthwhile to extend this model to the case that there are several servers by using the $M/M/n$ queuing model to provide an analysis for a more natural interpretation of the network capacity as the number of servers.

Acknowledgments

Jeong-Yoo Kim gratefully acknowledges the financial support from the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S 1A5A2A0304932311).

References

[1] Cheng, H.K., Bandyopadhyay, S., and Guo, H., The Debate on Net Neutrality: A Policy Perspective, *Information Systems Research*, 2011, Vol. 22, pp. 60-82.

[2] Choi, J.P. and Kim, B.C., Net Neutrality and Investment Incentives, *Rand Journal of Economics*, 2010, Vol. 41, pp. 446-471.

[3] Economides, N. and Hermalin, B., The Economics of Network Neutrality, *Rand Journal of Economics*, 2012, Vol. 43, pp. 602-629.

[4] Gross, D. and Harris, C.M., *Fundamentals of Queueing Theory*, 3rd ed., Wiley, New York, 1998.

[5] Kim, J.-Y., On the Invariance Result of Net Neutrality, *Review of Network Economics*, 2022, Vol. 20, pp. 139-157.

ORCID

Jeong-Yoo Kim | http://orcid.org/0009-0003-4137-4175

Seung J. Noh | http://orcid.org/0009-0009-9353-7424

Appendix

Proof of Proposition 1: The waiting time under the preemptive priority rule is

$$\begin{aligned} W_{PP} &= x_{PP} W_1^{PP} + (1-x_{PP})\, W_2^{PP} \\ &= W_1^{PP}[x_{PP} + (1-x_{PP})\frac{\mu}{\mu-\lambda_1-\lambda_{21}}] \\ &= W_1^{PP}\frac{x_{PP}(\mu-\lambda_1-\lambda_2)+(1-x_{PP})\mu}{\mu-\lambda_1-\lambda_2} \\ &= W_1^{PP}\frac{\mu-x_{PP}(\lambda_1+\lambda_2)}{\mu-\lambda_1-\lambda_2} \\ &= \frac{1}{\mu-x_{PP}\lambda}\frac{\mu-x_{PP}\lambda}{\mu-\lambda} \\ &= \frac{1}{\mu-\lambda} \\ &= W. \end{aligned}$$

Proof of Proposition 2: Under the preemptive priority rule, we have

$$\begin{aligned} W_{NP} &= x_{PP} W_1^{NP} + (1-x_{NP})\, W_2^{NP} \\ &= W_1^{NP}[x_{NP} + (1-x_{NP})\phi] \\ &= W_1^{NP}[x_{NP} + (1-x_{NP})\frac{\mu^2-\lambda_1\mu+\lambda_1\lambda}{\mu^2-\lambda_1\mu+\lambda_2\lambda}] \\ &= W_1^{NP}\frac{\mu^2-\lambda_1\mu-x_{NP}\lambda_2\lambda+(1-x_{NP})\lambda_1\lambda}{\mu^2-\lambda_1\mu-\lambda_2\lambda} \\ &= \frac{\mu+\lambda_2}{\mu(\mu-\lambda_1)}[\frac{\mu^2-\lambda_1\mu}{\mu^2-\lambda_1\mu-\lambda_2\lambda}] \\ &= \frac{\mu+\lambda_2}{\mu^2-\lambda_1\mu-\lambda_2\lambda} \\ &= \frac{\mu+(1-x_{NP})\lambda}{(\mu+(1-x_{NP})\lambda)(\mu-\lambda)} \\ &= \frac{1}{\mu-\lambda} \\ &= W. \end{aligned}$$

Proof of Proposition 3: Under the preemptive priority regime, we have

$$\Phi_{PP}(\mu;x) = W_2^{PP} - W_1^{PP} = W_1^{PP}\frac{\lambda}{\mu-\lambda} = \frac{1}{\mu-x\lambda}\frac{\lambda}{\mu-\lambda}. \quad (26)$$

Note that $\Phi'_{PP}(\mu;x) < 0$ for any $x \in [0,1]$. Also, note that $\Phi_{PP}(\mu;0) > 0$, for any $\mu > 0$. Assuming the uniqueness of the solution for x_{PP}, we obtain $\frac{dx_{PP}}{d\mu} < 0$, i.e., $\frac{d(W_2^{PP} - W_1^{PP})}{d\mu} < 0$.

Similarly, under the non-preemptive priority rule, we have

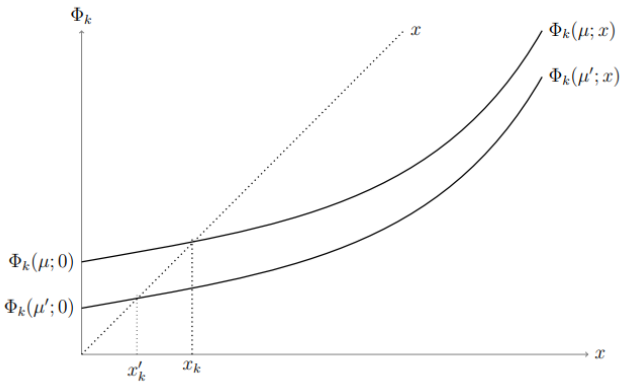
$$\Phi_{NP}(\mu;x) = W_2^{NP} - W_1^{NP} = W_1^{NP}[\phi - 1]. \quad (27)$$

We have

$$\frac{\partial W_1^{NP}}{\partial \mu} = -\frac{\mu^2+\lambda_2(2\mu-\lambda_1)}{[\mu(\mu-\lambda_1)]^2} < 0, \quad (28)$$

$$\frac{\partial(\phi-1)}{\partial \mu} = -\frac{(2\mu-\lambda_1)\lambda^2}{(\mu^2-\lambda_1\mu-\lambda_2\lambda)^2} < 0. \quad (29)$$

Since $W_1 > 0$ and $\phi - 1 > 0$, we have $\frac{\partial \Phi_{NP}}{\partial \mu} < 0$. Also, note $\Phi_{NP}(\mu;0) = \frac{1}{\mu}\frac{\lambda^2}{\mu^2-\lambda^2} > 0$ for any $\mu > 0$. Therefore, it follows from the uniqueness of the solution that $\frac{d(W_2^{NP} - W_1^{NP})}{d\mu} < 0$. (See <Figure 1>.)



<Figure 1> The Effect of an Increase in $\mu\,(\mu' > \mu)$ on $x_k\,(k = PP, NP)$