Simulation Studies on Asymptotic Approximations Analysis of M/M/s and M/D/s Queues*

Jinho Lee**

Contents

Abstract

This paper deals with asymptotic approximations analysis of M/M/s and M/D/s queues. For M/M/s queue, we observe "economies of scale" under the fixed utilization ρ and the fixed probability α that customer waits in system, how the average system size vary according to the number of servers s increasing. Simulation results show that as s increases, the number of servers who are idling increases, that is, the slack $n-E[Q_n]$ diverges. In addition, through changing the waiting probability α under the M/M/s system, α was not highly sensitive to the behavior of the system size. And, it is shown that using $\rho_n = 1-k/\sqrt{n}$ to handle heavy-traffic regime is only appropriate for $k=1$ by observing the effect on the performance of the system with different values of k. For the M/D/s queue, two approximations are used to evaluate the expected system size under the fixed ρ and α. Simulations and comparison of these two approximations show that Cosmetatos' approximation performs quite well when the number of servers is small and traffic intensity is heavy, but it overestimates the true value for the large number of servers. Meanwhile, the modified approximation gives good results for the steady state count of the system although the number of servers grows large.

Keywords: M/M/s, M/D/s, Queueing System, Asymptotic Approximation, Cosmetatos' Approximation, Economies of Scale, Heavy-traffic Regime

접수일 (2024년 07월 25일), 수정일 (2024년 08월 22일), 게재확정일 (2024년 09월 11일)
* This work was supported by 2024 Hongik University Research Fund.
** Associate Professor, College of Business Management, Hongik University, jinholee@hongik.ac.kr

1. Introduction

Multi-server queue has been important models for evaluating the performance of various service systems (Jeong, 2018; Lee, 2023; Wang et al., 2020) in many fields such as computer/communications, transportation, manufacturing and so on. Call center (Shim, 2022) is one of the simplest and most widely used such models with s operators (servers) and arrival rate λ. Corresponding to the systems such as call centers that have multi-servers with any service distribution, and customers arriving with Poisson process, the analysis of multi-server queueing systems has been paid attention. In this study, asymptotic approximations analysis under heavy traffic regime is dealt. *Heavy traffic* is the regime in which the utilization of the servers approaches the maximum permissible, i.e., ρ goes to 1 as close as possible, but this does not imply that ρ is exactly equal to 1. In order to analyze asymptotic approximation under the heavy traffic regime, we focus on the two systems that have different service distributions, respectively, one of which has independent and identical exponential service time, so-called M/M/s queue, and the other of which has deterministic service time (i.e., constant service time for all customers), so-called M/D/s queue. Note that both systems have independent and identical exponential inter-arrival times, i.e., Poisson arrival process. We carry out the approximation in the regime where the number of servers s becomes increasingly large, the utilization ρ has the value very close to one, the expected value of system size increases according to s. The point here is to observe the expected value of system size as s increases under the assumption of system stability with fixing close to one and fixing the waiting probability in system denoted by α. For the numerical approach of M/M/s queue, this research basically follows the mechanism that was used in Kumar (2008) in which the asymptotic analysis of this system was conducted in heavy traffic regime. M/D/s queue has no difference from M/M/s queue except for having the deterministic service time instead of exponential one, and it is often the case that is accepted as an approximation to some real system, either because the distribution of service times (or inter-arrival times) has a very small coefficient of variation or in order to obtain a bound on some measure of practical interest; see Cosmetatos (1975). However, in practice, dealing with M/D/s queue that has exact formulas such as expected value of system size and mean waiting time before starting service is non-trivial, as it is shown in the formulation (1) that has been proposed by Crommelin (1934):

$$EW = \frac{1}{\mu}\sum_{j=1}^{\infty}\sum_{k=js+1}^{\infty}\left[\frac{(js\rho)^{k-1}}{(k-1)!} - \frac{(js\rho)^{k}}{(\rho k)!}\right]e^{-js\rho} \quad (1)$$

In this formula, EW is the mean waiting time of M/D/s system before beginning service, assuming that the system is stable

and is in steady state, i.e., $\rho < 1$. Unfortunately, this formula needs us to calculate EW by solving an infinite system of linear equations, which implies that this is obviously non-trivial to solve. Consequently, such a numerical cumbersomeness in calculating EW has been a strong motivation for developing simple and accurate approximations. Among approximations that have been proposed so far, two of M/M/s based approximations will be used to deal with M/D/s queue, and the results from each approximation will be compared. First, we use Cosmetatos' approximation (1975) which is not accurate for large s. And, later we use the approximation proposed by Kimura (1991) which uses the Cosmetatos' approximation to obtain better accuracy for both large s and in light traffic. Numerical results by these two approximations will be also shown in later section.

2. Literature Review

A lot of research has been done on the asymptotic analysis of the different queuing systems. Kumar (2008) analyzes asymptotic approximation in heavy traffic regime for M/M/s queue, which gives us a strong motivation for this study. The steady state distribution for M/M/s queue is presented by Cooper (1972). And, Halfin and Whitt (1981) justify limit theorems for M/M/s queue. Shifting gears to M/D/s queue, Cosmetatos (1975) presents two formulas for the approximate evaluation of the average queueing time not only in the process of M/D/s queue but in one of D/M/s. Kimura (1991) deals with refining Cosmetatos' approximation for the mean waiting time in M/D/s queue. Although Cosmetatos' approximation performs quite well in heavy traffic, it overestimates the true value when the number of servers gets large or the traffic is light. Kimura's refining approximation from Cosmetatos' one shows through his numerical tests that the relative percentage error is less than 1% for almost all cases with $s < 20$ and at most 5% for other cases. In this paper, for the comparison of approximation performance, we consider Consmetatos' and Kimura's approximations for M/D/s system. Boxma et al. (1979) present a direct approximation from Crommelin (1934), and Kimura (1994) shows indirect approximation for M/D/s queue based on M/M/s and M/G/s queues. In addition, many researchers have proposed various approximate methods for M/G/s cases. Hokstad (1978) uses the method of supplementary variables to obtain difference-differential equations for a joint distribution of the number of customers in system and remaining service times. With some additional approximate assumptions, Hokstad solves these equations to generate an approximation for the probability generating function of the queue-length distribution. A few different methods in the same spirit have been developed by Tijms et al. (1981) and Miyazawa (1986). Some heuristic methods similar to this paper can be found in Maaloe (1973), Smith (1985) and

Takahashi (1977). For the GI/G/s case, however, possible approaches are quite limited: Halachmi and Franta (1978) and Wu (1990) develop diffusion approximations for the queue-length distribution. More recently, Cruz et al.(2017) use a Bayesian technique and the sampling/importance resampling method to estimate the parameters of M/M/s queues. Park et al.(2018) derive a product-form stationary joint probability distribution under M/M/1 queue to consider the (s, S) inventory policy. As one of the most recent studies, Nakamura and Phung-Duc(2024) show exact and asymptotic analysis of M/M/∞ queues. In this study, we focus on asymptotic approximations for the M/M/s and M/D/s queues.

3. M/M/s Queues

Consider a system with s servers. Suppose that the arrival of customers follows a Poisson process with rate λ. Further suppose that each customer requires an amount of service for a period, that is exponentially distributed with mean $1/\mu$. These service times are independent of each other and of the arrival process. Customers who cannot find a free server on arrival wait in a queue until a server is free, and always have room to do so. As soon as a server completes service it begins work on the customer at the head of the queue if there is one, and idles otherwise. Servers are indexed with 1, 2, ..., s, and for concreteness assume that an arriving customer who finds two idle servers is served by the server with the lower index. The number of customers in the system $Q(t)$, either being served or waiting for a server at any point of time t, forms a continuous-time Markov chain on the non-negative integers. In particular it forms a Birth-Death process with birth rate λ. The following result is obtained using standard Birth-Death process theory.

The steady state probability of k customers in the system is shown in (2):

$$P(Q=k)=\begin{cases}\eta\dfrac{(s\rho)^k}{k!}, \text{if } k<s\\ \eta\dfrac{s^s\rho^k}{s!}, \quad otherwise\end{cases} \tag{2}$$

where η is a constant with

$$\eta=\frac{1}{\sum_{k=0}^{s-1}\dfrac{(s\rho)^k}{k!}+\dfrac{(s\rho)^s}{s!(1-\rho)}} \tag{3}$$

The probability that customer will wait in the system, i.e., $Q\geq s$, will be denoted by α and formulated as follows, which is known as the Erlang-C formula (4):

$$\alpha=\eta\frac{(s\rho)^s}{s!(1-\rho)} \tag{4}$$

The mean (5) and variance (6) of the steady state customer count Q can be expressed in terms of α as (from Kumar' s notes):

$$E[Q]=\rho s+\frac{\alpha\rho}{1-\rho} \tag{5}$$

$$Var[Q]=\rho s(1+\alpha)+\frac{\alpha\rho+\alpha(1-\alpha)\rho^2}{(1-\rho)^2} \tag{6}$$

3.1 Asymptotic analysis

As a part of the research, we have performed simulation studies on the asymptotic

behavior of the M/M/s queue as the number of servers s tends to infinity. In order to carry out this asymptotic analysis as the number of servers becomes large, we resort to the following device. We look at a sequence of systems, with each system in the sequence having one server more than the previous system. Clearly, if we can characterize the limiting behavior of this sequence, we will be able to analyze the asymptotic behavior of the system. Every M/M/s queue, and thus, each system in this sequence, is denoted by three parameters: the number of servers s, which we know is increasing sequentially, the arrival rate λ, and the rate at which customers can be served by a server with service rate μ. Here, we keep service rate μ fixed and $\mu=1$, i.e., we are increasing the number of servers, not making them work faster. Depending on how we choose λ across systems in the sequence, we get different limiting behavior. That is, we can create different asymptotic regimes based on choice of λ. We can pick two regimes: (i) By scaling λ and s proportionally, which is also called *economies of scale*, (ii) regime in which the utilization of the servers approaches the maximum permissible. This is called *heavy traffic regime*.

In this section, we analyze both economies of scale and heavy traffic regime with the simulation settings summarized in Tab. 3-1.

3.2 Economies of scale

To analyze the asymptotic behavior of the system as s increases, we build a sequence of M/M/s queues indexed by n with the number of servers in the n-th system s_n being set to n. Let Q_n denote the steady state count in the n-th system. Fixing $\rho=0.9$, we plot $E[Q_n]$ against n. By fixing $\rho=0.9$ we are scaling the arrival rate λ and the number of servers s proportionally.

Tab. 3-1: Simulation settings

Economies of scale		
Input		Values
Parameter	Utilization ρ	[0.75, 0.95] * change with interval of 0.05
	Probability that customer waits α	[0.05, 0.40] * change with interval of 0.05
Variable	# servers n	Start from $n=500$ Increase by 500 Stop when $n=5000$
Output		$E[Q_n]$, $E[Q_n]\pm 3\sigma$ where $\sigma=\sqrt{Var[Q]}$
Heavy-traffic regime		
Input		Values
Parameter	Probability that customer waits α	0.1
Variable	# customers k	Start from $k=1$ Increase by 1 Stop when $k=5$
Variable	# servers n	Start from $n=500$ Increase by 500 Stop when $n=5000$
Output		$E[Q_n]$, $E[Q_n]\pm 3\sigma$ by using $\rho_n=1-k/\sqrt{n}$

As a reference it also plots n itself. As easily found from Fig. 3-1, $E[Q_n]$ has a slope that is smaller than 1. We can observe that the slack $n-E[Q_n]$, which is the proxy for the number of idle servers diverges. We also plot a $3-\sigma$ confidence interval around the mean count. This confidence interval is calculated using the variance formula. It is evident that

the confidence interval eventually excludes the reference n line and diverges from it. That is, it becomes increasingly unlikely that Q_n exceeds the number of servers n, i.e., all servers are busy, as n increases. Of course, we relied on explicit computation at $\rho = 0.9$ to come up with this explanation.

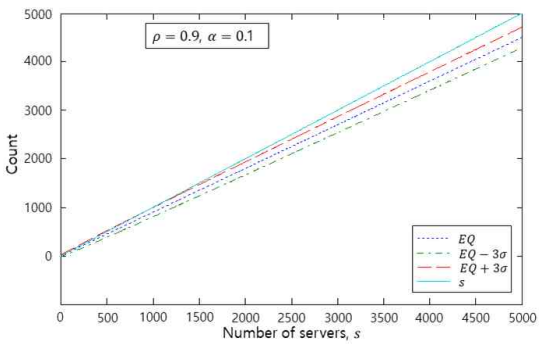

Fig. 3-1: Economies of scale when $\rho_n = \rho$

We analyze the influence of α, the probability of customer waiting in the system, in the behavior of the system. Fig. 3-2 shows the graphs for two different values of α.

We can observe that for different values of α, the performance of the system is not changed much, i.e., α is not highly sensitive to the behavior of the system. We carry a data sensitivity analysis of how $E[Q_n]$ varies for different values of α as number of servers s increases. Tab. 3-2 is the results of the sensitivity analysis (for $\rho = 0.9$).

It is clear that for $\alpha = 0.05$ and comparatively large value of α (= 0.4), the difference in $E[Q_n]$ is not high. This result can be made sense when we recall the formula of $E[Q] = \rho s + \alpha\rho/(1-\rho)$ from (5).

Since we fix α and ρ(=0.9), $E[Q_n]$ is mostly affected by n, as n increases. That is, $E[Q_n]$ is less sensitive than the case where n is relatively small. In this sensitivity analysis, the term $\alpha\rho/(1-\rho) = \alpha \times 0.9/(1-0.9) = 9\alpha$. Hence, $\alpha\rho/(1-\rho)$ is always in [0.4, 3.6].

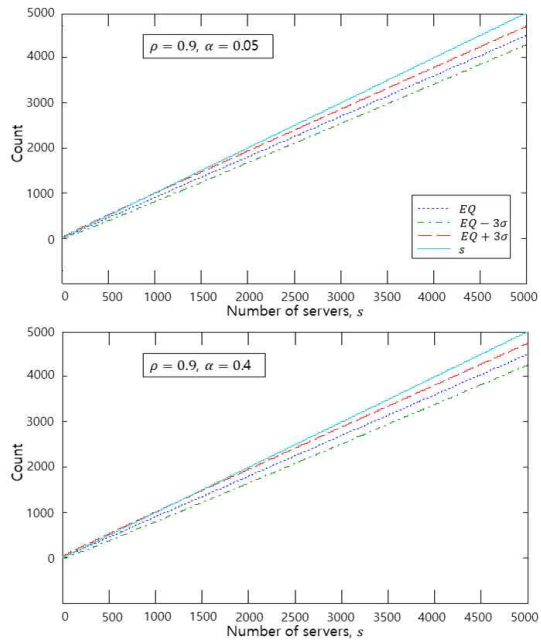

Fig. 3-2: Economies of scale for two different values of α

Fig. 3-3 shows the comparison of two graphs for $\rho = 0.9$ and $\rho = 0.95$.

As it is evident from the figure that for small difference in ρ there is considerably high variation in the performance of the system, i.e., ρ is sensitive to the performance of the system. Tab. 3-3 is the data sensitivity analysis of $E[Q_n]$ for different values of ρ against s ($\alpha = 0.1$).

Tab. 3-2: Sensitivity analysis of $E[Q_n]$ for $\rho = 0.9$

α	s									
	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
0.05	450.45	900.45	1350.45	1800.45	2250.45	2700.45	3150.45	3600.45	4050.45	4500.45
0.1	450.9	900.9	1350.9	1800.9	2250.9	2700.9	3150.9	3600.9	4050.9	4500.9
0.15	451.35	901.35	1351.35	1801.35	2251.35	2701.35	3151.35	3601.35	4051.35	4501.35
0.2	451.8	901.8	1351.8	1801.8	2251.8	2701.8	3151.8	3601.8	4051.8	4501.8
0.25	452.25	902.25	1352.25	1802.25	2252.25	2702.25	3152.25	3602.25	4052.25	4502.25
0.3	452.7	902.7	1352.7	1802.7	2252.7	2702.7	3152.7	3602.7	4052.7	4502.7
0.35	453.15	903.15	1353.15	1803.15	2253.15	2703.15	3153.15	3603.15	4053.15	4503.15
0.4	453.6	903.6	1353.6	1803.6	2253.6	2703.6	3153.6	3603.6	4053.6	4503.6

Tab. 3-3: Sensitivity analysis of $E[Q_n]$ for different values of ρ

ρ	s									
	500	1000	1500	2000	2500	3000	3500	4000	4500	5000
0.75	375.3	750.3	1125.3	1500.3	1875.3	2250.3	2625.3	3000.3	3375.3	3750.3
0.8	400.4	800.4	1200.4	1600.4	2000.4	2400.4	2800.4	3200.4	3600.4	4000.4
0.85	425.6	850.6	1275.6	1700.6	2125.6	2550.6	2975.6	3400.6	3825.6	4250.6
0.9	450.9	900.9	1350.9	1800.9	2250.9	2700.9	3150.9	3600.9	4050.9	4500.9
0.95	476.9	951.9	1426.9	1901.9	2376.9	2851.9	3326.9	3801.9	4276.9	4751.9

3.3 Heavy traffic regime

For heavy traffic regime which is also called as Halfin and Whitt regime, we use the above formulas (5) and (6) to calculate $E[Q_n]$ and $Var[Q_n]$ by using $\rho_n = 1 - k/\sqrt{n}$. In this case, as the number of servers $s \to \infty$, $\rho \to 1$. Fig. 3-4 represents the graph for $k = 1$ and $\alpha = 0.1$.

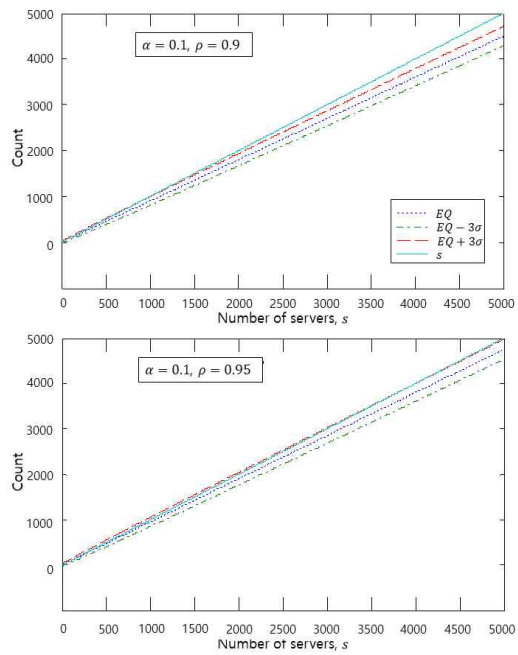

Fig. 3-3: Economies of scale for two different values of ρ

We also analyze the system for different values of k. As in Fig. 3−5, we observe that the constant k has a significant effect on the performance of the system. The heavy traffic regime is appropriate only for $k=1$. It is evident from Fig. 3−4 that the confidence interval of the count $E[Q_n]\pm 3\sigma(Q_n)$ always contains the reference line s, i.e., the number of servers is never more (or less) than fixed number of standard deviations above (or below) the mean count. We can see that as k increases, the reference line s is diverging from the count.

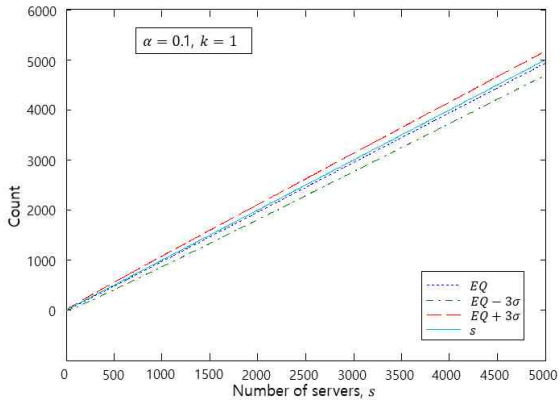

Fig. 3−4: Heavy traffic regime $\rho_n=1-1/\sqrt{n}$

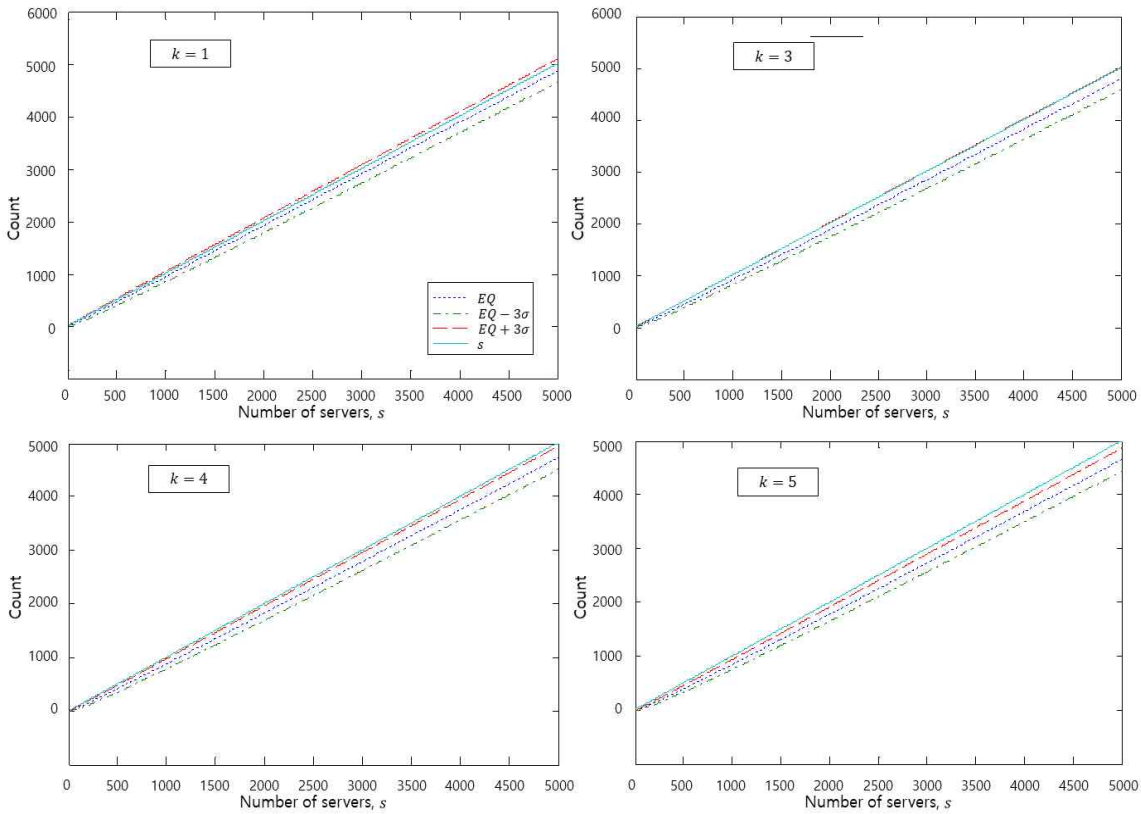

Fig. 3−5: Heavy traffic regime for different k

4. M/D/s Queues

4.1 Expected waiting time in M/D/s queues

Consider the M/D/s queuing system with s homogeneous parallel servers, independent and identically distributed exponential inter-arrival times and constant service time. The waiting room is infinite, i.e., there is no upper bound on queue length, and service policy is first-come first-served. Sticking to the conventions, we denote λ and μ as the arrival and service rates, respectively. Therefore, traffic intensity is given as $\rho = \frac{\lambda}{s\mu}$. Let $EW(M/D/s)$ denote the mean waiting time in this system assuming that the system is in steady state, i.e., $\rho < 1$. The calculation of $EW(M/D/s)$ is numerically cumbersome and therefore a lot of simple and accurate approximations have been derived. Among these approximations, Cosmetatos' approximation (1975),

$$EW(M/D/s) \cong \frac{1}{2}[1+f(s)g(\rho)]EW(M/M/s)$$

where

$$f(s) = \frac{(s-1)(\sqrt{4+5s}-2)}{16s}$$

$$g(\rho) = \frac{1-\rho}{\rho}$$

is evaluated as having the best quality for most of the practical purposes. Here, $EW(M/M/s)$ denote the mean waiting time in the corresponding M/M/s queue with same mean arrival rate and service time as the M/D/s queue.

We know that

$$EW(M/M/s) = \frac{(s\rho)^s\left[\sum_{j=0}^{s-1}\frac{(s\rho)^j}{j!} + \frac{(s\rho)^s}{s!(1-\rho)}\right]^{-1}}{s!s\mu(1-\rho)^2}$$

which can be written as

$$EW(M/M/s) = \frac{\alpha}{s\mu(1-\rho)}$$

where α refers to the probability that a customer will wait and is calculated using the formula (4).

Thus, for given values of α, ρ, s and μ we can easily calculate $EW(M/M/s)$ and hence can find out expected waiting time in queue for the corresponding M/D/s queue.

4.2 Expected steady state customer count in M/D/s system

We can use the aforementioned equation for calculating $EW(M/D/s)$ to find out the expected steady state customer count of M/D/s system using Little's Law. Since EW is the mean waiting time, $W = EW + \frac{1}{\mu}$. By Little's Law,

$$EQ(M/D/s) = \lambda \cdot W$$
$$= \frac{\lambda}{2}[1+f(s)g(\rho)]EW(M/M/s) + \frac{\lambda}{\mu}$$

which can again be calculated for given values of α, ρ, s and μ using spreadsheet or other computational tools.

We perform simulations to find out the behavior of approximation with different varying values of the aforementioned parameters and the results are displayed in Fig. 4-1 and Fig. 4-2. Fig. 4-1 show that

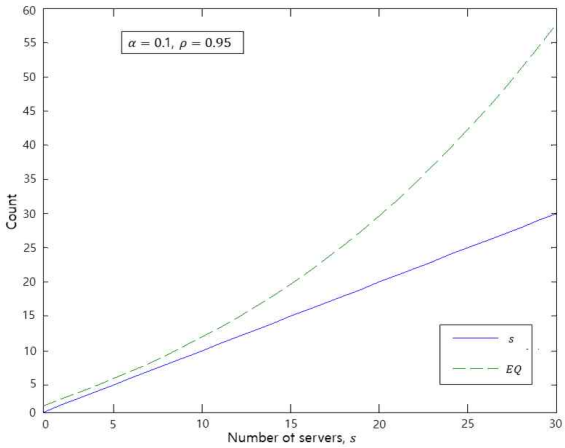

Fig. 4−1: Cosmetatos' approximation for Heavy traffic regime

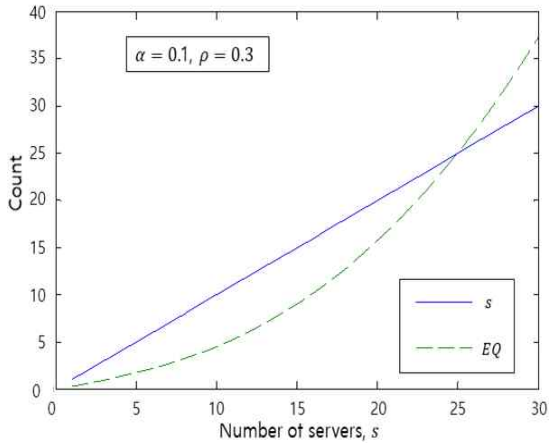

Fig. 4−2. Cosmetatos' approximation for low traffic intensity

Cosmetatos' approximation performs quite well when the number of servers s is small and traffic intensity is heavy. However, it also shows that for large number of servers the approximation overestimates the true value (as mentioned by Kimura(1991)).

Fig. 4−2 shows the performance of Cosmetatos' approximation for low traffic intensity. Again it can be seen that the approximation overestimates the true value.

4.3 Modified Cosmetatos' approximation

Kimura(1991) propose modifications in Cosmetatos' approximation to take care of the two defects mentioned above. The proposed approximation gives the asymptotically exact value of expected waiting time in queue (and simultaneously the expected steady state count) when $s\to\infty$ or $\rho\to 0$ or $\rho\to 1$. Even for some other values of s and ρ it gives a good approximation. The approximation is given as:

$$EW'(M/D/s) \cong \frac{1}{2}[1+f(s)g(\rho)]EW(M/M/s)$$

where $h(s,\rho)$ is a correction function given by

$$h(s,\rho) = \xi(s,a(\rho))\eta(b(s),\rho)$$

$$\xi(s,x) = \sqrt{1-\exp\left(-\frac{2x}{s-1}\right)}, x \geq 0$$

$$\eta(y,\rho) = 1-\exp\left(-\frac{\rho y}{1-\rho}\right), y \geq 0$$

The functions $a(\rho)$ and $b(s)$ are defined as

$$a(\rho) = \frac{25.6}{[g(\rho)\eta(\beta_1,\rho)]^2}$$

$$b(s) = \frac{s-1}{(s+1)f(s)\xi(s,\alpha_1)}$$

Here, α_1 and β_1 are constants linked through this relation: $\alpha_1\beta_1^2 = 25.6$.

Kimura(1991) also suggests that the approximations are fairly insensitive to these constants though performs best when $\alpha_1 = 2.2$ and $\beta_1 = 3.41$ (we use this value throughout in

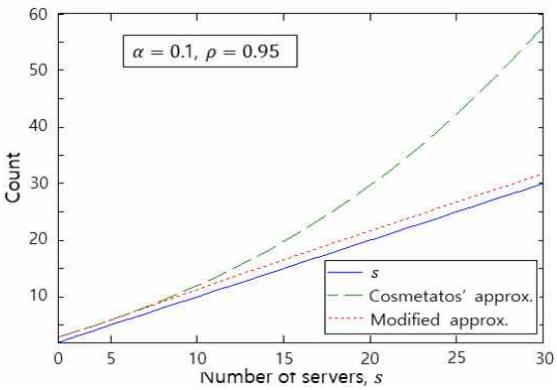

Fig. 4-3: Modified Cosmetatos' approximation vs. Cosmetatos' approximation for heavy traffic regime

simulation in order to get the best approximation).

Using $EW'(M/D/s)$ and Little's law we derive the following approximation for Expected Steady state customer count in M/D/s system:

$$EQ(M/D/s) = \frac{\lambda}{2}[1+f(s)g(\rho)h(s,\rho)]EW(M/M/s)+\frac{\lambda}{\mu}$$

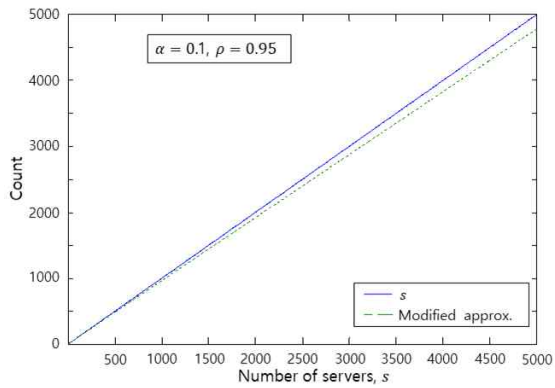

Fig. 4-4: Performance of modified Cosmetatos' approximation for large s under heavy traffic regime

We also simulate the modified Cosmetatos' approximation for different values of servers and ρ, and compare them to the results we obtained from Cosmetatos' approximation. Fig. 4-3 shows that the modified Cosmetatos' approximation gives good approximation for the steady state count of the system even when the number of servers is large against Cosmetatos' approximations which becomes increasingly diverge with increasing number of servers (which appears to be violating the concept of economies of scale). The modified Cosmetatos' approximation also holds (with small error) even when the number of servers are too large as shown in Fig. 4-4.

5. Conclusions

Throughout this study, we performed asymptotic approximation analysis for M/M/s and M/D/s queues. For the M/M/s queue, we observed "Economies of Scale", i.e., under the fixed utilization ρ and the fixed probability that customer wait in system, α, how the average system size vary according to the number of servers s increasing. Simulation results showed that as s increases, the number of servers who are idling increased, that is, the slack $n-E[Q_n]$ diverged. In addition, through changing the waiting probability α under the M/M/s system, α was not highly sensitive to the behavior of the system size. And, it was shown that using $\rho_n = 1-k/\sqrt{n}$ to handle Heavy-traffic regime was only appropriate for $k=1$ by observing

the effect on the performance of the system with different values of k. For the M/D/s queue, we used two approximations, both of which are M/M/s based approximations. Simulations and comparison of these two approximations showed that Cosmetatos' approximation performs quite well when the number of servers is small and traffic intensity is heavy, but it overestimates the true value for the large number of servers. Meanwhile, the modified approximation gave good results for the steady state count of the system although the number of servers becomes large.

M/M/s and M/D/s queueing models offer valuable insights into the management and optimization of service systems in various industries. M/M/s model is more suited to environments where adaptability is critical due to high variability in both demand and service times, such as call centers, hospitals, and customer service centers (e.g., retailers, banks, etc,). In contrast, M/D/s model excels in predictable environments where efficiency and consistency are the primary goals such as manufacturing and assembly lines, fast-food restaurants, transportation and logistics. In addition, M/M/s models often require more dynamic resource allocation, which can increase operational costs but also improve service levels during peak times, while M/D/s models allow for more streamlined operations, potentially reducing costs, but require a stable environment where service times are consistent. This study is deemed to contribute to enhancing service levels by enabling more accurate predictions of customer wait times based on industry characteristics in actual service areas where queuing theory can be applied, and by preparing appropriate countermeasures accordingly.

[References]

[1] Boxma, O.J., Cohen, J.W., and Huffels, N. (1979), Approximations of the mean waiting time in an M/G/s queueing system, *Operations Research*, 7(6), 1115-1127.

[2] Cooper, R.B.(1972), Introduction to Queueing Theory, Macmilan, New York.

[3] Cosmetatos, G.P.(1975), Approximate explicit formulae for the average queueing time in the processes (M/D/r) and (D/M/r), *INFOR: Information Systems and Operational Research*, 13, 328-331.

[4] Crommelin, C.D.(1934), Delay probability formulae, *Post Office Electrical Engineer's Journal*, 26, 266-274.

[5] Cruz, F.R.B., Quinino, R.C., and Ho, L.L. (2017), Bayesian estimation of traffic intensity based on queue length in a multi-server M/M/s queue, *Communications in Statistics - Simulation and Computation*, 46(9), 7319-7331.

[6] Halachmi, B., and Franta, W.R.(1978), A diffusion approximation to the multi-server queue, *Management Science*, 24, 522-529.

[7] Halfin, S., and Whitt, W.(1981), Heavy-traffic limits for the queues with many exponential servers, *Operations Research*, 29(3), 567-588.

[8] Hokstad, P.(1978), Approximations for the M/G/m queue, *Operations Research*, 26, 510-523.

[9] Jeong, Y.(2018), The effects of fairness and quality on the trust and loyalty in the R&D process, *Journal of Service Research and Studies*, 8(3), 115-136. (정용길(2018), 연구개발 과정에서 공정성과 품질이 신뢰와 충성도에 미치는 영향, *서비스연구*, 8(3), 115-136)

[10] Kimura, T.(1991), Refining Cosmetatos' approximation for the mean waiting time in the MID/s queue, *Journal of Operations Research Society*, 42(7), 595-603.

[11] Kimura, T.(1994), Approximations for multi-server queues: System interpolations, *Queueing Systems*, 17, 347-382.

[12] Kumar, S.(2008), Stochastic Networks, Graduate School of Business, Stanford University.

[13] Lee, J.W.(2023), Classification of servic quality for HMR unmanned store business, *Journal of Service Research and Studies*, 13(2), 41-61. (이종원 (2023), HMR 무인매장 서비스 품질 분류에 관한 연구, *서비스연구*, 13(2), 41-61)

[14] Miyazawa, M.(1986), Approximation for the queue-length distribution of an M/GI/s queue by the basic equations, *Journal of Applied Probability*, 23, 443-458.

[15] Maaloe, E.(1973), Approximation formulae for estimation of waiting-time in multiple-channel queueing systems, *Management Science*, 19, 703-710.

[16] Nakamura, A., and Phung-Duc, T.(2024), Exact and asymptotic analysis of infinite server batch service queues with random batch sizes, *Queueing Systems*, 106, 129-154.

[17] Park, J., Lee, H.G., Kim, J.H., Yun, E.H., and Baek, J.W.(2018), Analysis of an M/M/1 Queue with an Attached Continuous-type (s,S)-inventory, *Journal of the Korea Industrial Information Systems Research*, 23(5), 19-32. (박진수, 이현근, 김종현, 윤은혁, 백정우 (2018), 정책하의 연속형 내부재고를 갖는 M/M/1 대기행렬모형 분석, *한국산업정보학회논문지*, 23(5), 19-32)

[18] Shim, J.(2022), The effects of traditional leadership recognized by members of the telemarketing organization on orgnaizational perfomance-centered on the moderating effect of followship, *Journal of Service Research and Studies*, 12(3), 45-59. (심지현 (2022), 텔레마케팅 조직구성원이 인식한 변혁적 리더십이 조직성과에 미치는 영향-팔로워십의 조절효과를 중심으로, *서비스연구*, 12(3),

45-59)

[19] Smith, V. L.(1985), Approximating the distribution of customers in M/En/s queues, *Journal of Operations Research Society*, 36, 327-332.

[20] Takahashi, Y.(1977), An approximation formula for the mean waiting time of an M/G/c queue, *Journal of Operations Research Society of Japan*, 20, 150-163.

[21] Tijms, H. C., van Hoorn, M. H., and Federgruen, A.(1981), Approximations for the steady-state probabilities in the M/G/c queue, *Advances in Applied Probability*, 13, 186-206.

[22] Wang, X., Kim, Y., and Park, J.S.(2020), A study on effect of service characteristic factors of theme park on customer satisfaction and revisit intention, *Journal of Service Research and Studies*, 10(2), 43-57. (왕효뢰, 김영길, 박정수(2020), 테마파크의 서비스 특성 요인이 관람객의 만족과 재방문 의도에 미치는 영향에 관한 연구, *서비스연구*, 10(2), 43-57)

[23] Wu, J.S. (1990), Refining the diffusion approximation for the G/G/c queue, *Computational Mathematics and. Applications*, 20, 31-36.

Jinho Lee (jinholee@hongik.ac.kr)

Jinho Lee is an Associate Professor at College of Business Management, Hongik University. He received the B.S. degree in Electrical Engineering from the Republic of Korea Naval Academy in 2002, the M.S. degree in Industrial Engineering from Yonsei University in 2006, and the Ph.D. degree in Operations Research & Industrial Engineering from The University of Texas at Austin, Austin, TX, USA, in 2012. His research interests include stochastic & network optimization, combinatorial optimization, systems modeling & simulation and their application to industry, defense, and security.

M/M/s와 M/D/s 대기행렬의 점근 근사법 분석을 위한 시뮬레이션 연구

이진호*

초 록

본 연구는 M/M/s와 M/D/s 대기행렬의 점근 근사법 분석을 수행한다. M/M/s 대기행렬 분석을 위해, 활용률 ρ와 고객의 시스템 대기확률 α가 특정값을 가질 때 서버수 s의 증가에 따라 평균 시스템 대기자의 크기가 변화하는 양상을 통해 "규모의 경제"를 관찰하였다. 시뮬레이션 결과, s가 증가함에 따라 유휴시간을 갖는 서버의 수인 $n-E[Q_n]$은 발산함을 보여주었다. 그리고 고객의 수 k의 변화에 따른 시스템 성능을 관찰한 결과, heavy-traffic regime(활용률이 점점 증가하는 상태)을 살펴보기 위하여 $\rho_n = 1-k/\sqrt{n}$을 이용하는 것은 고객의 수가 1($k=1$)인 경우에만 유효함을 확인하였다. M/D/s 대기행렬의 경우 고정된 ρ와 α 하에서 평균 시스템 대기자의 크기 분석을 위해 두 가지의 근사법을 이용하였다. 시뮬레이션 및 비교 분석 결과, 서버의 수가 작고 heavy-traffic인 경우에는 Cosmetatos 근사법이 좋은 성과를 보여 주지만, 서버의 수가 큰 경우에는 실제값을 과대평가하는 경향이 있음을 보여주었다. 반면에 수정 근사법(modified approximation)은 서버의 수가 증가할 때에도 시스템 안정상태에 대한 보다 정확한 근사치를 제공하였다.

핵심어: M/M/s, M/D/s, 대기행렬, 점근 근사법, Cosmetatos 근사법, 규모의 경제, 고-교통량 영역

* 홍익대학교 상경대학 부교수, jinholee@hongik.ac.kr