



NUMERICAL STUDY OF OPTIMAL BUFFER SIZE AND VACATION LENGTH IN M/G/1/K QUEUES WITH MULTIPLE VACATIONS

Kilhwan Kim ^{*1}

^{*1} Department of Management Engineering, Sangmyung University, South Korea



Abstract:

Nowadays, due to the advent of clouding computing, buffer size can be readily extended in a couple of minutes for computing servers, where the buffer size should not be considered as given when optimizing the system performance. In this context, we explore optimal combinations for the buffer size and the length of vacation time in M/G/1/K queues with multiple vacations numerically. We consider the cases of deterministic and exponentially distributed vacation and service times. In order to do this, we also formulate an optimal problem and define cost factors: the customer loss cost, the buffer holding cost, and the server operating cost. We present some numerical examples to investigate the impact of the system parameters such as the buffer size, the length of the vacation time, and the distribution of the service time, to performance measures and the total cost. We also investigate optimal combinations for the buffer size and the vacation length for various values of the cost factors.

Keywords: M/G/1/K Queues; Multiple Vacations; Optimal Buffer Size; Optimal Vacation Length.

Cite This Article: Kilhwan Kim. (2019). "NUMERICAL STUDY OF OPTIMAL BUFFER SIZE AND VACATION LENGTH IN M/G/1/K QUEUES WITH MULTIPLE VACATIONS." *International Journal of Engineering Technologies and Management Research*, 6(2), 1-13. DOI: <https://doi.org/10.29121/ijetmr.v6.i2.2019.350>.

1. Introduction

In this paper, we consider M/G/1/K queues with multiple vacations and explore optimal combinations for the buffer size and the length of the vacation time through a numerical study.

M/G/1/K queues with multiple vacations have been employed in various studies to analyze the performance of systems with finite buffer and server vacations in the telecommunication, computing network, and manufacturing areas [1-4]. As a result, M/G/1/K queues with multiple vacations have been extensively studied [5-12]. However, in many of those studies, only the performance measures of the system were derived for a given buffer size and a given distribution of the vacation time, and neither optimal buffer size nor optimal length of the vacation time was considered [5-10]. In some of those studies, the optimal length of the vacation time was considered for a fixed buffer size [11, 12], but both optimal buffer size and optimal length of the vacation time were not considered simultaneously. This is partly because the buffer size was not able to be adjusted dynamically in a short period of time in many real-world problems. Also, the fact that the

performance measures such as the customer loss probability do not have a closed-form expression for M/G/1/K queues hindered analytic studies on the optimality of the performance measures.

Nowadays, due to the advent of clouding computing, buffer size can be readily extended in a couple of minutes for computing servers, where the buffer size should not be considered as given when optimizing the system performance. In this context, we believe that studies on optimal combinations for the buffer size and the length of the vacation time in M/G/1/K queues with multiple vacations are needed. However, since there are no closed expressions for the performance measures of M/G/1/K queues, here we simply tackle the problem in a numerical approach as a preliminary to future studies.

This paper is organized as follows: In Section 2, we define the queueing model that are analyzed in this paper, and present a couple of performance measures which will be used to formulate an optimization problem. In Section 3, we formulate an optimization problem. In Sections 4 and 5, we present some numerical examples to investigate the impact of the system parameters such as the buffer size, the length of the vacation time, and the distribution of the service time, to performance measures and the total cost. In Section 6, we investigate optimal combinations for the buffer size and the vacation length for various values of the cost factors.

2. Queueing Model

We consider an M/G/1/K queue with multiple vacations: Customers arrive at the system according to a Poisson process with rate λ . The service times S of customers are independent and identically distributed random variable with an arbitrary general distribution. Define $S(x)$ as the distribution function of S . Define also ρ as the offered load, i.e., $\rho = \lambda E[S]$. There is a single server in the system. When customers are present in the system, the server keeps servicing customers until the system becomes empty of customers. As soon as the system becomes empty of customers, the server leaves for vacation for a vacation time D , which is a random variable, and the distribution function of D is denoted by $D(x)$. If the server finds any customers in the system when it returns from its vacation, then it restarts servicing customers. Otherwise, it leaves for another vacation until it finds customers at the end of a vacation. At a given time, at most K customers can be accommodated in the system. If there are already K customers in the system when a customer arrives, then that customer will be lost immediately.

For the proposed queueing model, various performance measures such as the mean system size, the loss probability, and the mean waiting time have been derived in several studies (5; 6; 7; 9; 10). Here we present several key results from those studies with brief explanation, which will be used for our optimization model.

Let q_n denote the probability that the number of customers is n just after a service is completed, $n, n = 0, \dots, K - 1$. Then, we have the following global balance equations:

$$q_n = \sum_{j=1}^{n+1} \left(q_j + \frac{q_0 v_j}{-v_0 + 1} \right) a_{-j+n+1}, \quad 0 \leq n \leq K - 2 \tag{1}$$

$$q_{K-1} = q_0 \sum_{j=K}^{\infty} \frac{v_j}{-v_0 + 1} + \sum_{j=1}^{K-1} \left(q_j + \frac{q_0 v_j}{-v_0 + 1} \right) \sum_{i=K-j}^{\infty} a_i \tag{2}$$

where a_i and v_i are the probabilities that i customers arrive during the service time and the vacation time, respectively. That is,

$$a_i = \int_0^{\infty} \frac{(\lambda x)^i}{i!} e^{-\lambda x} dS(x)$$

And

$$v_i = \int_0^{\infty} \frac{(\lambda x)^i}{i!} e^{-\lambda x} dD(x).$$

From (1), (2) and the normalization condition

$$\sum_{n=0}^{K-1} q_n = 1, \tag{3}$$

we can numerically calculate $q_n, n = 0, \dots, K - 1$.

Let p_n denote the probability that there are n customers in the system at an arbitrary time. Also, let ρ_e denote the carried load of the M/G/1/K queue. Since the effective arrival rate is $\lambda(1 - p_K)$, we have

$$\rho_e = \lambda(-p_K + 1)E(S). \tag{4}$$

Also, if we let I and B denote the length of an idle period and the length of a busy period of the M/G/1/K queue, then we have

$$\rho_e = \frac{E(S)}{(E(I) + E(S))q_0 + (-q_0 + 1)E(S)}. \tag{5}$$

Since the expected length of the idle period is expressed as

$$E(I) = \frac{E(D)}{-v_0 + 1},$$

from (4) and (5) we have

$$p_K = 1 + \frac{v_0 - 1}{\lambda(E(D)q_0 - E(S)v_0 + E(S))}, \quad (6)$$

which is the customer loss probability because we assumed a Poisson arrival process. From (6) and the following relationship

$$\lambda(-p_K + 1)E(S) = \frac{E(B)}{E(B) + E(I)},$$

we can also derive the expected busy period length $E[B]$ as

$$E(B) = \frac{\lambda(p_K - 1)E(D)E(S)}{(v_0 - 1)(\lambda E(S)p_K - \lambda E(S) + 1)}. \quad (7)$$

3. Optimization Problem

In this section, we formulate an optimization problem to find an optimal combination for the buffer size and the length of the vacation time in the M/G/1/K queue with multiple vacations. We first introduce three cost factors: customer loss cost, buffer holding cost, and server operating cost. The customer loss cost represents sales opportunity loss or customer dissatisfaction that incurs when a customer who arrives at the system is lost. The buffer holding cost represents a cost for accommodating customers who are waiting and being serviced in the system. It might be a rent for physical facility or memory space in a clouding computing environment. The server operating cost represents a cost incurring when the server is on duty or preparing its job. For example, in many telecommunication devices, the server works in two modes: the wakeup or sleep mode. When the server is in the wakeup mode, the server is available to service and its power consumption is high (thus high in the operating cost). In contrast, when the server is in the sleep mode, it is unavailable to service, and its power consumption is low (thus low in the operating cost). However, while it is in the sleep mode, it needs to inspect the buffer periodically to check whether any customers are present. As the time interval between these inspections is short, then the server's preparation cost will be high.

Let C_l denote the cost that incurs whenever a customer is lost, C_b the cost that incurs for buffer space for a single customer per unit time, C_h the cost that incurs per unit time when the server is on duty, and C_v the cost that incurs for every inspection at the end of a server vacation because of the server's buffer inspection activity. We also let C denote the expected total cost that incurs per unit time. Then, the expected total cost C per unit time is expressed as

$$C = C_b K + \frac{C_h E(B)}{E(B) + E(I)} + C_l \lambda p_K + \frac{C_v}{(E(B) + E(I))(-v_0 + 1)} \quad (8)$$

because the rate of customer loss is

$$\lambda p_K,$$

the fraction of time when the server is on duty is

$$\frac{E(B)}{E(B) + E(I)}$$

and the vacation rate is

$$\frac{1}{(E(B) + E(I))(-v_0 + 1)}$$

which is calculated as the expected number of vacations that the server takes during one cycle (composed of one idle period and the following busy period), which is $1/(1 - v_0)$, divided by the expected length of the one cycle, which is $E(I) + E(B)$.

Suppose that the vacation time is deterministic and is set to a positive value T . Then, the optimization problem to find an optimal combination for the buffer size and the length of the vacation time of the M/G/1/K queue with multiple vacations can be formulated as

$$\min_{K,T} C(K, T).$$

Also, if the vacation time is assumed to be exponentially distributed with rate μ_D , then the optimization problem to find an optimal combination for the buffer size and the length of the vacation time of the M/G/1/K queue can be formulated as

$$\min_{K,\mu_D} C(K, \mu_D).$$

For other distributions, the optimization problems can be formulated, but we only deal with the above two cases for the sake of simplicity. Note that the terms p_K , v_0 , $E(I)$ and $E(B)$ in (8) changes according to the decision variables of the optimization problems, so we see the behaviors of these terms as well as that of the total cost numerically, changing the decision variables, in the next section.

4. Numerical Study - The Case of Deterministic Vacation Times

In this section, we explore optimal combinations for the maximum buffer size K and the length T of the vacation time when the vacation time is deterministic. Throughout our numerical study, we set the cost factors C_l, C_v, C_b, C_h to be 20, 1, 1, 1.

We first consider the case when the service time is also deterministic and the offered load (traffic) is light (i.e., $\rho = 0.25$). Figure 1 shows various performance measures and the total cost when the arrival rate is set to 1 and the service rate is set to 4 (i.e., $E(S) = 0.25$). In 1, we can see the changes of the loss probability, the time fraction with the server being on duty, the vacation rate, and the total cost as a function of the buffer size, which changes from 1 to 10 for each value of the vacation time T , which changes from 0.25 to 2 times the mean interarrival time and is distinguished with a different gradation.

The left top panel of Figure 1 displays the loss probability as a function of the buffer size K for each T value. It clearly shows that the loss probability drops as the buffer size increases for all the T values. Also, the impact of the length of the vacation time T on the loss probability becomes prominent when the buffer size is relatively small, while it becomes insignificant when the buffer size is large enough for the loss probability to drop to near zero.

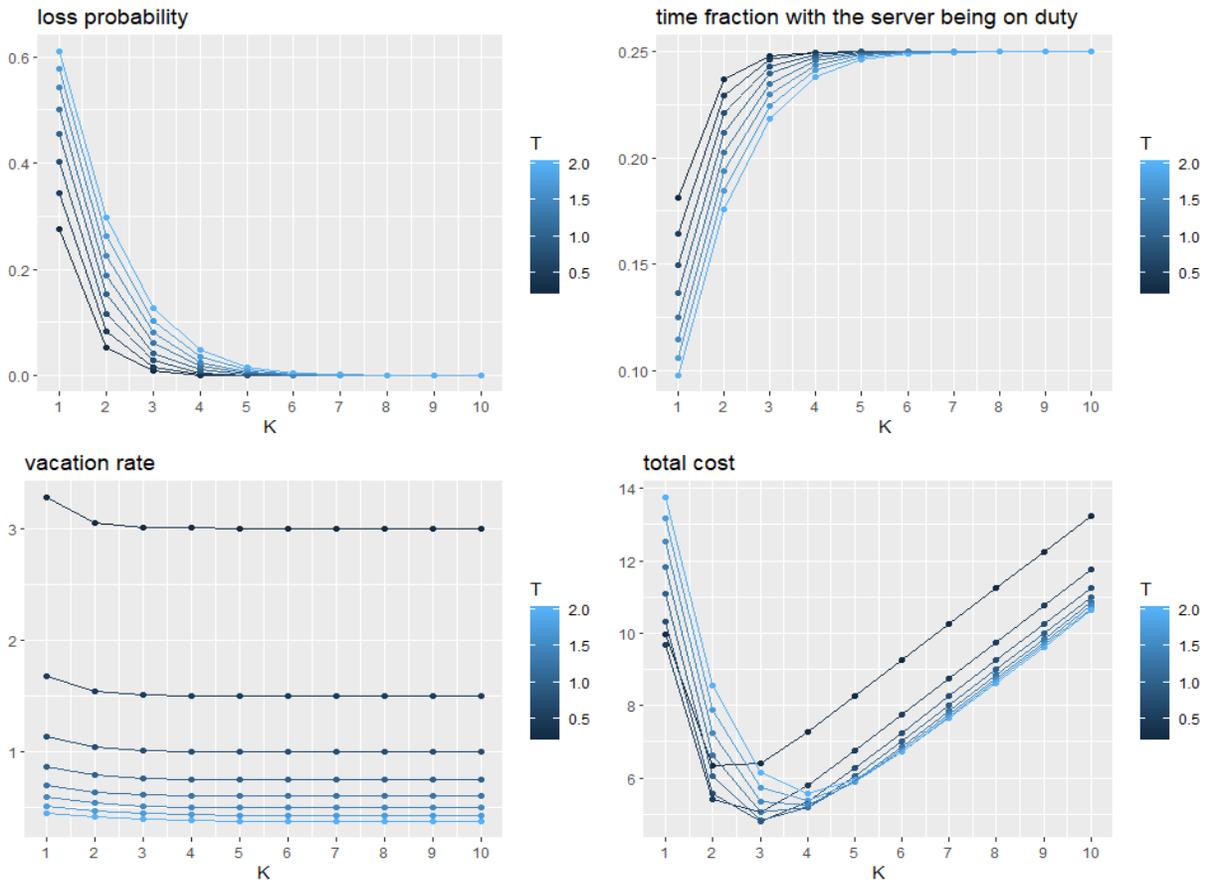


Figure 1: performance measures and cost for a light traffic when the vacation and service times are deterministic

The right top panel of Figure 1 displays the average time fraction with the server being on duty as a function of the buffer size K for each T value. It clearly shows that the time fraction with the server being on duty rises as the buffer size increases for all the T values. This is because, when the buffer size becomes large, the customer loss probability drops so that the total effective offered load rises and the server has more jobs to do. Like the loss probability, the impact of the length of the vacation time T on the time fraction with the server being on duty becomes prominent when the buffer size is relatively small, while it becomes insignificant when the buffer size is large enough for the loss probability to drop to near zero.

The left bottom panel of Figure 1 displays the server vacation rate as a function of the buffer size K for each T value. It shows that the vacation rate first slowly drops as the buffer size increases for all the T values, but the slope then becomes flat. Thus, the impact of the buffer size K on the vacation rate is not as significant as on the loss probability or the time fraction with the server

being on duty. When the buffer size is small, the busy period becomes relatively short so the server tends to go on a vacation more often, but this effect is a secondary factor of the vacation rate under the light traffic. Unlike the loss probability or time fraction with the server being on duty, the impact of the length of the vacation time T on the vacation rate is prominent and a primary factor when the traffic is light.

Finally, the right bottom panel of Figure 1 displays the total cost as a function of the buffer size K for each T value. It shows that, for the given cost factors, the total cost tends to first drop and then rise as the buffer size increases for all the T values. The main reason why the total cost drops when the buffer size is small is that the customer loss cost rapidly decreases in this interval. When the buffer size is large, the customer loss cost slowly decreases so the increment of the buffer holding cost exceeds it, thus the total cost increases. Also, when the buffer size is small, the total costs for large T values are higher than those for small T values. In contrast, when the buffer size is large, the total costs for large T values are lower than those for small T values. This is because, when the buffer size is small, the customer loss cost dominates and a long vacation time has a significant negative impact on the total cost, but, when the buffer size is large, the loss probability drops to near zero so the vacation rate dominates, thus a long vacation time has a positive impact on the total cost. Overall, in this case, the optimal K and T are 3 and 0.75, respectively, and the minimum total cost is 4.8.

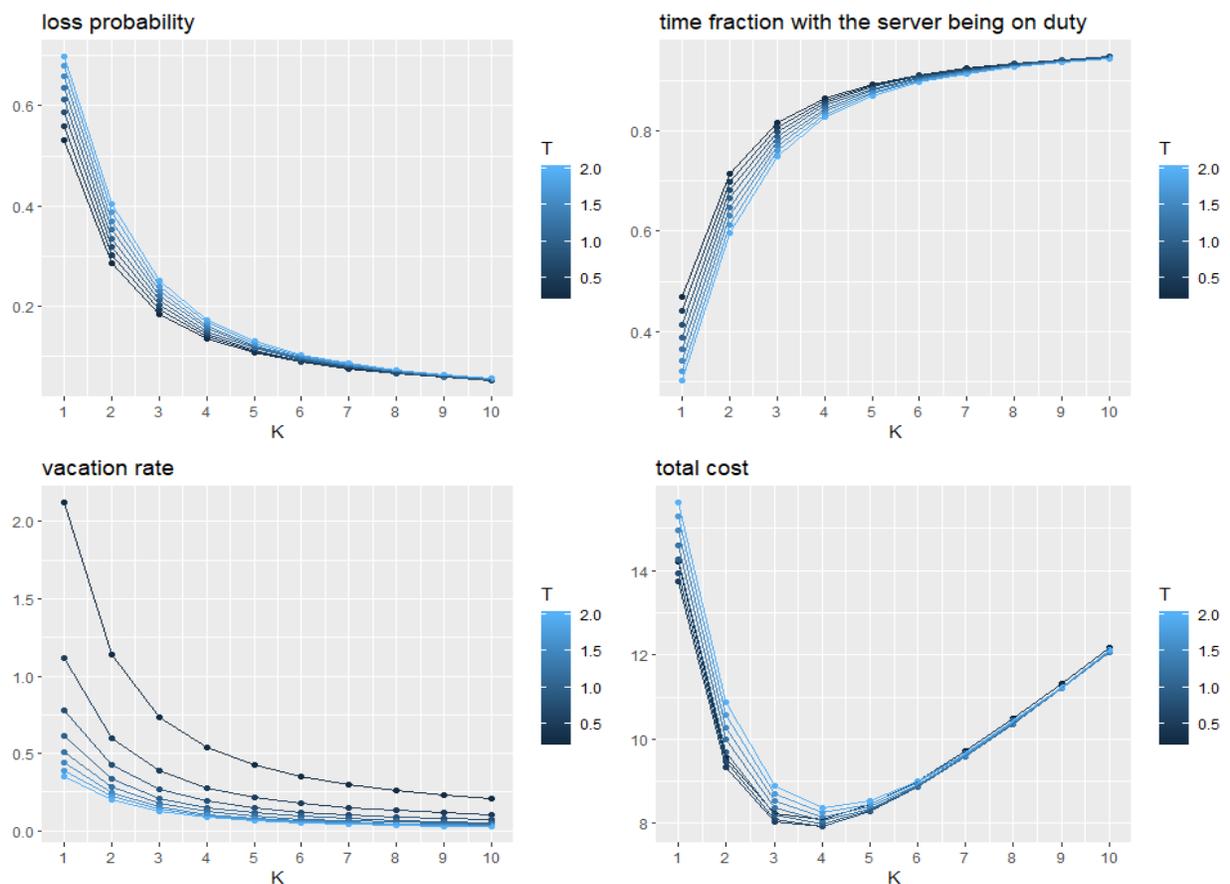


Figure 2: performance measures and cost for a heavy traffic when the vacation and service times are deterministic

We now consider the case when the service time is deterministic and the offered load is heavy (i.e., $\rho = 1$). Figure 2 shows various performance measures and the total cost when the arrival rate is set to 1 and the service rate is set to 1 (i.e., $E(S) = 1$). We can see the same trend of the performance measures and total cost as in the case of the light traffic (compare with 1). The main difference is that the impact of the buffer size appears more prominent but the impact of the vacation length less prominent under the heavy traffic, compared to the light traffic case. This is because the customer loss cost dominates more when the traffic becomes heavier. In the heavy traffic case, the optimal K and T are 4 and 0.5, respectively, and the minimum total cost is 7.93.

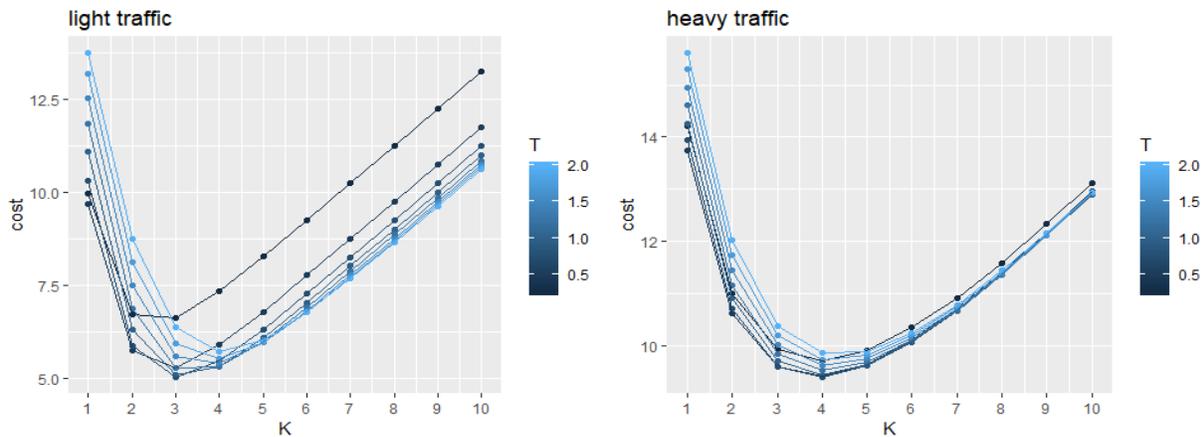


Figure 3: performance measures and cost for deterministic vacation and exponential service times

We now consider the case when the service time is exponentially distributed. The left panel of Figure 3 shows the total cost when the offered load is light (i.e., $\rho = 0.25$). In fact, the arrival rate is set to 1 and the service rate is set to 4. The right panel of Figure 3 shows the total cost when the offered load is heavy (i.e., $\rho = 1$). In fact, the arrival rate is set to 1 and the service rate is set to 1. The cost curves in the cases of the exponential service times show quite the same pattern as in the cases of the deterministic service times. In the light traffic case, the optimal K and T are 3 and 0.75, respectively, and the minimum total cost is 5.02. In the heavy traffic case, the optimal K and T are 4 and 0.75, respectively, and the minimum total cost is 9.4. Note that, compared to the cases of the deterministic service times, the minimum total costs rise due to the randomness of the service time.

5. Numerical Study - The Case of Exponential Vacation Times

In this section, we explore optimal combinations for the maximum buffer size K and the length T of the vacation time when the vacation time is exponentially distributed. We also use the same cost factors C_l, C_v, C_b, C_h as in the case of deterministic vacation times. In order to make a comparison with the cases of deterministic vacation times, we also let T denote the mean vacation time, which means that the service rate $\mu_D = 1/T$ and the density function of the vacation time is expressed as

$$\frac{1}{T} e^{-t/T}.$$

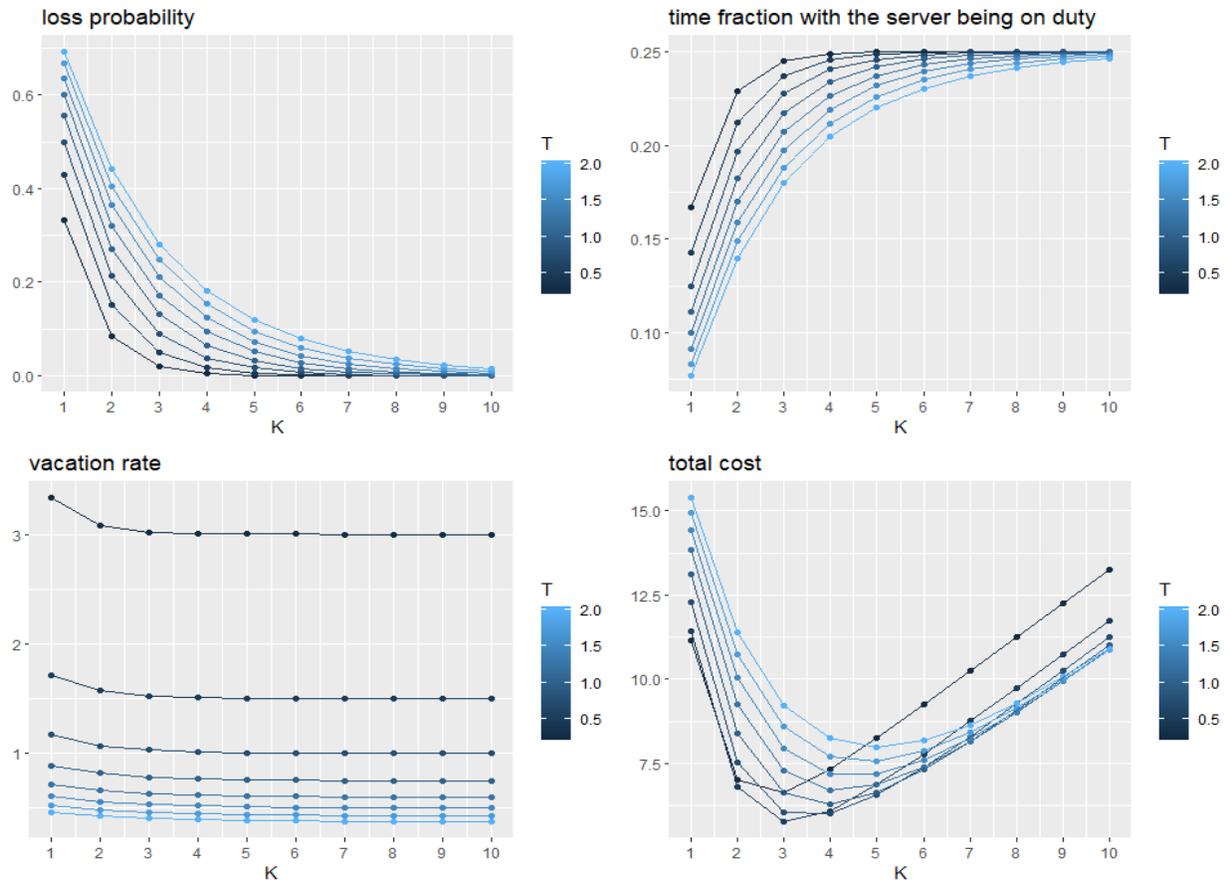


Figure 4: performance measures and cost for a light traffic when the vacation time is exponentially distributed and the service time is deterministic

We first consider the case when the service time is also deterministic and the offered load is light (i.e., $\rho = 0.25$). Figure 4 shows various performance measures and the total cost when the arrival rate is set to 1 and the service rate is set to 4 (i.e., $E(S) = 0.25$). In 4, we can see the changes of the loss probability, the time fraction with the server being on duty, the vacation rate, and the total cost as a function of the buffer size, which changes from 1 to 10 for each length of the mean vacation times T , which changes from 0.25 to 2 times the mean interarrival time and is presented with a different gradation. As you can see in Figure 4, the performance measures and the cost curves reveal similar patterns to the case of deterministic vacation times (compare with Figure 1). However, due to the randomness of the vacation time, the loss probability and thus the total cost tend to be significantly larger when the mean vacation time is large and the buffer size is small than those in the corresponding case of deterministic vacation times. Overall, in this case, the optimal K and T are 3 and 0.5, respectively, and the minimum total cost is 5.78.

When the service time is deterministic and the offered load is heavy (i.e., $\rho = 1$), we can see the same trend of the performance measures and total cost as shown in Figure 2 of the case of deterministic vacation times with the heavy traffic so we omit graphs for this case. Like the light traffic case, due to the randomness of the vacation time, the loss probability and thus the total cost tend to be larger when the mean vacation time is large and the buffer size is small than those in the

corresponding case of deterministic vacation times. Overall, in the heavy traffic case, the optimal K and T are 4 and 0.5, respectively, and the minimum total cost is 8.13.

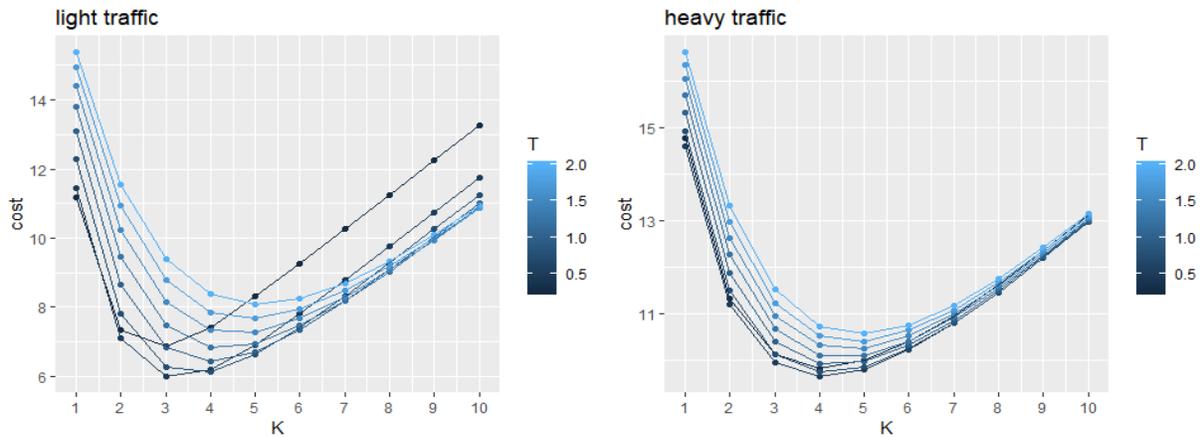


Figure 5: performance measures and cost for exponential vacation and service times

We now consider the case when the service time is exponentially distributed. The left panel of Figure 5 shows the total cost when the offered load is light (i.e., $\rho = 0.25$). The right panel of Figure 5 shows the total cost when and the offered load is heavy (i.e., $\rho = 1$). The cost curves show quite the same pattern as in the cases of deterministic service times. In the light traffic case, the optimal K and T are 3 and 0.5, respectively, and the minimum total cost is 5.99. In the heavy traffic case, the optimal K and T are 4 and 0.5, respectively, and the minimum total cost is 9.66. Compared to the cases of deterministic service times, the minimum total costs rise due to the randomness of the service time (compare to Figure 4). Also, compared to the cases of deterministic vacation times, the minimum total costs rise due to the randomness of the vacation time (compare to Figure 3).

6. Numerical Study - Optimal Buffer Size and Vacation Length

In Sections 4 and 5, we explored the behavior of the performance measures and the total cost as a function of the buffer size K and the vacation length T for deterministic and exponentially distributed vacation and service times. In this section, we explore the changes of optimal combinations for the buffer size K and the vacation length T for various values of the cost factors when the traffic is moderate (i.e., $\rho = 0.5$)

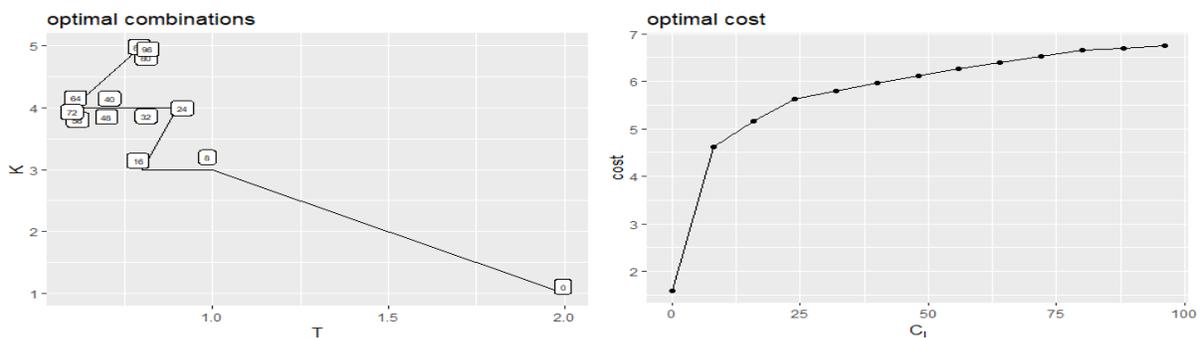


Figure 6: Optimal combinations and the optimal cost for various customer loss costs C_l

We first see the impact of the customer loss cost C_l on the optimal combination of K and T , and the optimal cost. We employ the grid search method to find an optimal combination, where K changes from 1 to 10 by 1, and T changes from 0.1 to \$2 times the mean interarrival time by 0.1 the mean interarrival time. The left panel of Figure 6 shows the optimal combination K and T for various C_l , which is displayed in the labels in the plot, when the other cost factor remain the same. The right panel of Figure 6 shows the optimal cost as a function of C_l . As seen in Figure 6, as the customer loss cost C_l increases, the optimal K tends to increases and the optimal T decreases for the same K but jumps to a higher value when the optimal K changes to one step higher values. The optimal cost first increases rapidly, but then slowly moves.

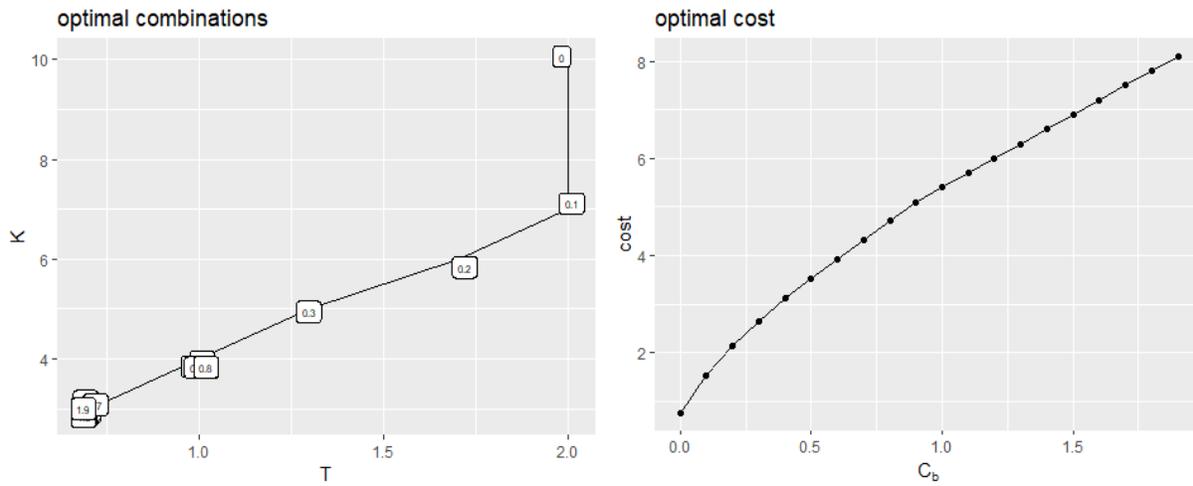


Figure 7: Optimal combinations and the optimal cost for various buffer holding costs C_b

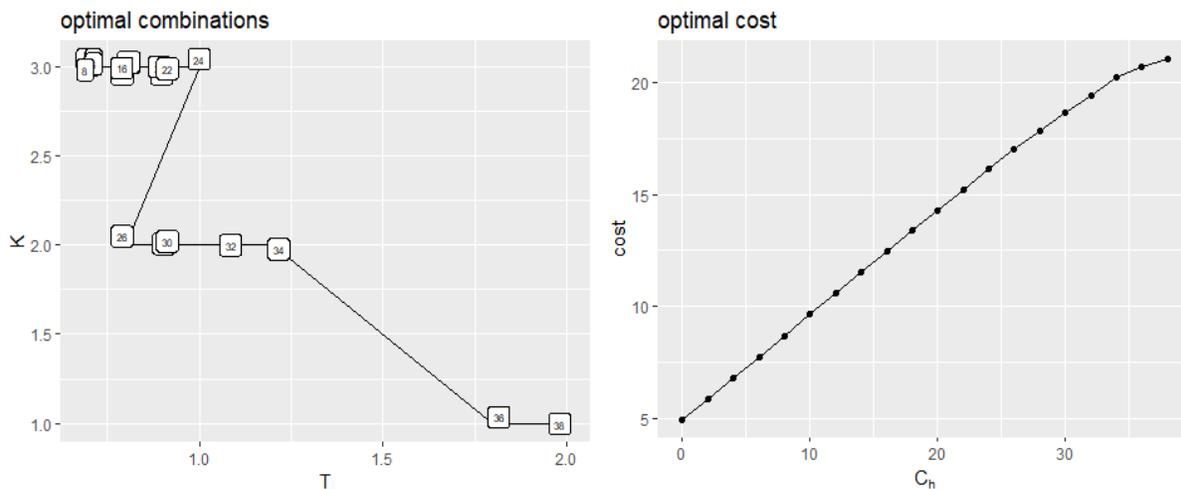


Figure 8: Optimal combinations and the optimal cost for various server up costs C_h

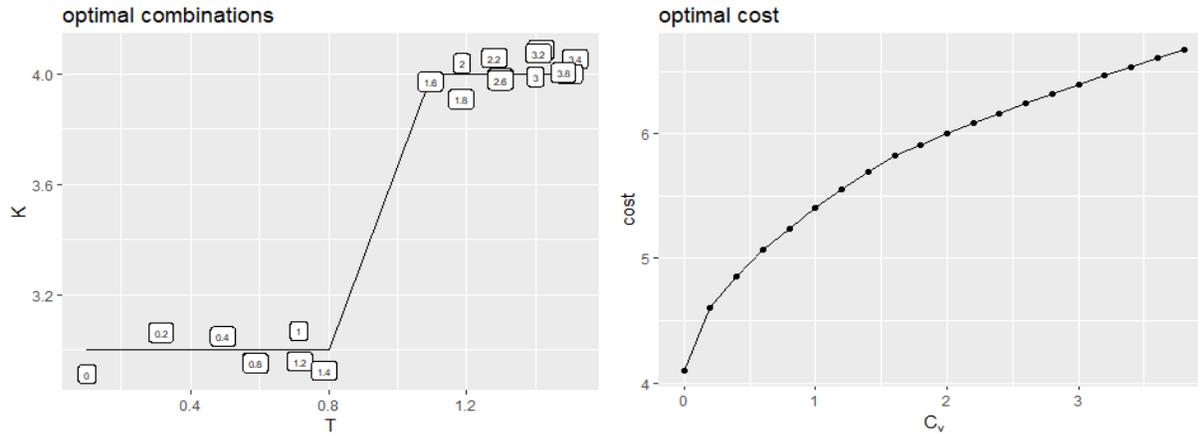


Figure 9: Optimal combinations and the optimal cost for various server vacation costs C_v

Figures 7, 8, and 9 shows optimal combinations of K and T and the corresponding optimal cost for various buffer holding cost C_b , server up cost C_h , and server vacation cost C_v , respectively. The optimal K and T tend to linearly decreases first as the buffer holding cost C_b but the moving becomes slow (see Figure 7). When the server up cost is not high, the optimal K does not change and only the optimal T increases as the server up cost rises, while, when the server up cost becomes too high, the optimal K drops toward 1, resulting in blocking most customers (see Figure 8). As the server vacation cost C_v rises, the optimal K tends not to change, but the optimal T increases rapidly because the vacation rate is mainly determined by the vacation length T (see Figure 9). For all the three cases, the optimal cost increases as the cost factor rises.

7. Conclusion

In this paper, we explored optimal combinations for the buffer size and the length of vacation time in $M/G/1/K$ queues with multiple vacations numerically. We considered the cases of deterministic and exponentially distributed vacation and service times. In order to do this, we also formulated the optimal problem and defined the cost factors: the customer loss cost, the buffer holding cost, and the server operating cost.

Regardless of the distributions of the service and vacation times, the customer loss probability drops as the buffer size increases, and so does it as the length of the vacation time decreases. Also, when the traffic is heavy, the impact of the buffer size dominates. The fraction of the time fraction with the server being on duty acts in the opposite way. It rises as the buffer size increases, and so does it as the length of the vacation time decreases. The vacation rate rises as the length of the vacation time decreases. In addition, when the traffic is light, the impact of the buffer size tends to be less significant. The total cost represents the combined effect of the customer loss probability, the fraction of the time fraction with the server being on duty, the vacation rate and the buffer size. It tends to first drop and then rise as the buffer size increases. When the buffer size is relatively small, the total cost with a shorter vacation time is lower than that with a longer vacation time. In contrast, when the buffer size is relatively large, the total cost with a shorter vacation time is higher than that with a longer vacation time. This is because the customer loss probability drops to near zero when the buffer becomes large enough so the efficiency of the server operating becomes more significant factor. Even though we only considered the deterministic and exponentially distributed

cases, the optimal total cost tends to increase as the randomness of the service time and the vacation time increases.

As mentioned earlier, there are no closed expressions for the performance measure of M/G/1/K queues, studies on the optimal combination of the buffer size and length of the vacation time in M/G/1/K queues with multiple vacations are needed in order to understand the detailed behavior of the performance of those queues. Also, due to the advent of clouding computing, the size of buffer can be readily extended in a couple of minutes for computing servers. In this context, we believe that the numerical study of optimal combinations for the buffer size and the length of the vacation time can help system engineers understand the system behaviors when optimizing the system performance.

References

- [1] R. Zheng, J. C. Hou, and L. Sha, "Performance analysis of power management policies in wireless networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 6, pp. 1351–1361, 2006.
- [2] K. De Turck, S. De Vuyst, D. Fiems, S. Wittevrongel, and H. Bruneel, "Performance of the sleep-mode mechanism of the new iee 802.16 m proposal for correlated downlink traffic," in *International conference on network control and optimization*, 2009, pp. 152–165.
- [3] C.-Y. Chen, C.-H. Hsu, and K.-T. Feng, "Performance analysis and comparison of sleep mode operation for iee 802.16 m advanced broadband wireless networks," in *Personal indoor and mobile radio communications (pimrc), 2010 iee 21st international symposium on*, 2010, pp. 1425–1430.
- [4] H. K. Aksoy and S. M. Gupta, "Optimal management of remanufacturing systems with server vacations," *The International Journal of Advanced Manufacturing Technology*, vol. 54, nos. 9-12, pp. 1199–1218, 2011.
- [5] T. T. Lee, "M/g/1/n queue with vacation time and exhaustive service discipline," *Operations Research*, vol. 32, no. 4, pp. 774–784, 1984.
- [6] T. T. Lee, "M/g/1/n queue with vacation time and limited service discipline," *Performance Evaluation*, vol. 9, no. 3, pp. 181–190, 1989.
- [7] H. Takagi, "M/g/1/k queues with n-policy and setup times," *Queueing systems*, vol. 14, nos. 1-2, pp. 79–98, 1993.
- [8] S. Kasahara, Y. Takahashi, and T. Hasegawa, "Analysis of waiting time of m/g/1/k system with vacations under random scheduling and lcf," *Performance evaluation*, vol. 21, no. 3, pp. 239–259, 1995.
- [9] A. Frey and Y. Takahashi, "A note on an m/gi/1/n queue with vacation time and exhaustive service discipline," *Operations research letters*, vol. 21, no. 2, pp. 95–100, 1997.
- [10] K.-H. Wang and J.-C. Ke, "A recursive method to the optimal control of an m/g/1 queueing system with finite capacity and infinite capacity," *Applied Mathematical Modelling*, vol. 24, no. 12, pp. 899–914, 2000.
- [11] K.-H. Wang, C.-C. Kuo, and W. Pearn, "Optimal control of an m/g/1/k queueing system with combined f policy and startup time," *Journal of Optimization Theory and Applications*, vol. 135, no. 2, pp. 285–299, 2007.
- [12] Y. Park and G. U. Hwang, "An efficient power saving mechanism for delay-guaranteed services in iee 802.16 e," *IEICE transactions on communications*, vol. 92, no. 1, pp. 277–287, 2009.

*Corresponding author.

E-mail address: khkim@ smu.ac.kr