Cost Optimization of the $GI/GI/n$ Queueing Model

Mohamed Ghazali, Abdelghani Ben Tahar

***Abstract*—Queueing systems play a pivotal role in modeling and analyzing various real-world scenarios across diverse domains. In this paper, we focus on the GI/GI/n queue, a more general and flexible model that accommodates independent, non-exponential interarrival and service time distributions. Our research explores cost optimization within GI/GI/n queue systems, with a primary objective of identifying the optimal service rate and number of servers. Leveraging advanced performance metrics and numerical techniques to address the complexities of general distributions, we develop strategies to enhance system efficiency while minimizing operational costs.**

***Index Terms*—Queuing theory, cost optimization, GI/GI/n queue, optimal number of servers, optimal service rate.**

I. Introduction

Queuing theory is a mathematical framework for the study of waiting lines or queues. It is used in many fields including healthcare [1], [2], supply chain [3], telecommunication [4], computer networks [5], and cloud and fog computing [6], [7], [8].

In today's dynamic and interconnected world, where the demand for efficient service delivery is paramount, the optimization of queueing systems has garnered significant attention across industries. One prominent queueing model, the $GI/GI/n$ queue system, stands out for its generality and ability to capture the variability in both arrival and service processes with multiple servers. In this context, cost optimization becomes a crucial objective, as organizations strive to balance service quality with resource utilization and operational expenses.

While the $M/M/n$ queue model has been widely studied for cost optimization, including applications in call center operations [9], supermarket checkout lines [10], and hospital emergency departments [11], the $GI/GI/n$ queue model presents additional challenges due to the general nature of its arrival and service distributions. The flexibility of the $GI/GI/n$ model allows it to represent more realistic scenarios where inter-arrival and service times are not restricted to exponential distributions, making it suitable for complex and variable environments, such as cloud computing and network systems.

Several studies have addressed optimization problems in various queueing system scenarios. Wang and Zijm [12] develop a cost model for the $M/M/R$ queueing system, considering finite capacity alongside balking, reneging, and server breakdowns to determine the optimal number of servers. Yang et al. [13] analyze the F-policy $M/M/1/K$ queueing system, incorporating working vacations and an exponential startup time, to minimize costs by determining the optimal system capacity K, threshold F, and service rates. Similarly, Sethi and Sivakumar [14] investigate customer impatience in an unreliable $M/M/1$ queueing system operating under N-policy, formulating a cost function and optimization problem to identify optimal repair and service rates. Bouchentouf et al. [15] focus on an $M^X/M/c$ Bernoulli feedback queueing system with waiting servers, examining synchronous vacations with single and multiple vacation policies and employing the quadratic fit search method (QFSM) to optimize service rates and minimize costs.

However, while these studies provide valuable insights into broader queueing models and their steady-state behavior, few have tackled the optimization problem for the $GI/GI/n$ queue model. Existing literature often focuses on deriving steady-state distributions and performance metrics for such systems without delving into detailed cost optimization. Unlike these prior works, which typically address only specific elements such as the optimal service rate or number of servers, there is a clear gap in comprehensive cost analysis for the $GI/GI/n$ model.

In this paper, we address this gap by extending the methodology used for the $M/M/n$ queue to the more general $GI/GI/n$ system. Our work formulates a detailed cost function that considers both variable service rates and the number of servers concurrently. We also account for the variability inherent in the inter-arrival and service time distributions, making the optimization framework applicable to more realistic and complex queueing environments.

This paper is structured as follows: Section 2 outlines the performance metrics of the $GI/GI/n$ queue model, which are essential for the discussion in Section 3. In Section 3, we propose strategies for cost optimization in $GI/GI/n$ queue systems by determining the optimal number of servers and the optimal service rate.

II. Performance metrics

The $GI/GI/n$ queueing model generalizes the $M/M/n$ queue by allowing both interarrival and service times to follow arbitrary, independent distributions (GI stands for "General Independent"). This system consists of n identical servers working in parallel. Customer interarrival times are independent and identically distributed (i.i.d.) random variables with a mean of $1/\lambda$ (λ represents the arrival rate) and a squared coefficient of variation (SCV) of c_a^2 (variance divided by the mean squared). Similarly, service times are also i.i.d. random variables with a mean of $1/\mu$ (μ represents the

Manuscript received November 20, 2024; revised February 22, 2025.

M. Ghazali is a PhD student at Hassan First University of Settat, Faculty of Science and Technology, B.P. 577, Settat 26000, Morocco. (e-mail: m.ghazali@uhp.ac.ma).

A. Ben Tahar is a professor at Hassan First University of Settat, Faculty of Science and Technology, B.P. 577, Settat 26000, Morocco. (e-mail:abdelghani.bentahar@uhp.ac.ma).

service rate) and an SCV of c_s^2. All interarrival and service times are mutually independent. Customers are served on a first-come, first-served (FCFS) basis. For stability of the system, we assume that the traffic intensity $\rho = \frac{\lambda}{n\mu} < 1$.

For the $M/M/n$ queue, the steady-state probability π_k of having k customers in the system is defined as:

$$\pi_k = \begin{cases} \dfrac{(n\rho)^k}{k!}\pi_0, & \text{for } 0 \le k < n, \\ \dfrac{(n\rho)^k}{n! \cdot n^{k-n}}\pi_0, & \text{for } k \ge n, \end{cases}$$

where π_0 is the probability that the system is empty:

$$\pi_0 = \left[\sum_{k=0}^{n-1} \frac{(n\rho)^k}{k!} + \frac{(n\rho)^n}{n!} \cdot \frac{1}{1-\rho}\right]^{-1}.$$

The average number of waiting customers in the $M/M/n$ queue is given by

$$L_q^{M/M/n} = \sum_{j \ge n+1} (j-n)\pi_j = \left(\frac{(n\rho)^n \rho}{n!(1-\rho)^2}\right)\pi_0.$$

Therefore, by Little's law; the average waiting time in the queue is given by

$$W_{M/M/n} = \frac{1}{\lambda}\left(\frac{(n\rho)^n \rho}{n!(1-\rho)^2}\right)\pi_0.$$

In contrast to the $M/M/n$ system, which assumes exponential (memoryless) interarrival and service time distributions, the $GI/GI/n$ model accommodates more general distributions, thereby capturing real-world variability more effectively. However, this generalization introduces analytical complexity, as the absence of the memoryless property hinders the derivation of closed-form expressions for key performance metrics.

Sakasegawa [16] proposed the following closed form approximation for the expected waiting time in the $M/M/n$ queue:

$$W_{M/M/n} \simeq \frac{\rho^{\sqrt{2(n+1)}-1}}{n\mu(1-\rho)}.$$

Building upon this, Whitt [17] suggested the following approximation for the expected waiting time in the $GI/GI/n$ queue:

$$W_{G/G/n} \simeq \frac{c_a^2 + c_s^2}{2} W_{M/M/n}.$$

These approximations allow us to derive performance indicators for the $GI/GI/n$ queue. For conciseness, we omit the $GI/GI/n$ subscript in the following. The average waiting time W is:

$$W = \frac{1}{\mu}\left(\frac{c_a^2 + c_s^2}{2}\right)\frac{\rho^{-1+\sqrt{2(m+1)}}}{m(1-\rho)}. \tag{1}$$

Consequently, the average sojourn time R, which includes both waiting and service times, is:

$$\begin{aligned} R &= 1/\mu + W \\ &= \frac{1}{\mu}\left[1 + \left(\frac{c_a^2 + c_s^2}{2}\right)\frac{\rho^{-1+\sqrt{2(m+1)}}}{m(1-\rho)}\right]. \end{aligned} \tag{2}$$

Applying Little's Law yields the average number of waiting customers L_q and the average number of customers in the system L:

$$L_q = \lambda W = \frac{\lambda}{\mu}\left(\frac{c_a^2 + c_s^2}{2}\right)\frac{\rho^{-1+\sqrt{2(m+1)}}}{m(1-\rho)}, \tag{3}$$

$$L = \lambda R = \frac{\lambda}{\mu}\left[1 + \left(\frac{c_a^2 + c_s^2}{2}\right)\frac{\rho^{-1+\sqrt{2(n+1)}}}{n(1-\rho)}\right]. \tag{4}$$

III. Cost Optimization

To calculate the total cost of the $GI/GI/n$ queue system, we consider two main components: the cost associated with operating the servers (C_s) and the cost incurred due to customer waiting time, including the cost of waiting while being served (C_w). The cost of operating the servers involves expenses such as server maintenance and infrastructure costs, which are typically incurred regardless of the system's queue dynamics. On the other hand, the cost of customer waiting time reflects the impact of queueing delays on customer satisfaction, potential revenue loss. By combining these costs, we obtain the total cost of the system, which can be expressed as

$$C_T = nC_s + L(n, \lambda/\mu, c_a^2, c_s^2)C_w.$$

where $L(n, \lambda/\mu, c_a^2, c_s^2)$ denotes the expected number of customers in the system in function of the parameters. It is important to note, however, that we do not account for additional charges unrelated to the parameters of the queue system. Using the explicit formula given in (4), the cost function become

$$C_T = nC_s + C_w\left[\left(\frac{c_a^2 + c_s^2}{2}\right)\frac{\rho^{\sqrt{2(n+1)}}}{(1-\rho)} + \lambda/\mu\right]. \tag{5}$$

Define the constant $C = C_s/C_w$, the cost function is now given by

$$C_T = C_w\left[nC + \frac{n\mu\,(c_a^2 + c_s^2)}{2\,(n\mu - \lambda)}\,(\lambda/n\mu)^{\sqrt{2(n+1)}} + \lambda/\mu\right]. \tag{6}$$

The constant C measures how C_s relates to C_w and reflects the service quality of your queuing system. When C_w is much larger than C_s (i.e., $C \gg 1$), it suggests a lower service quality. On the flip side, if C_s significantly outweighs C_w ($C \ll 1$), it indicates a strong emphasis on customer satisfaction, resulting in better service but higher costs. Striking the right balance ensures optimal service quality while managing expenses effectively. These assessments are context-specific, and the actual values for C_s and C_w depend on factors such as the nature of the business or the industry, and customer expectations.

A. Optimal number of servers

The number of servers in a queuing system has a significant impact on its cost dynamics. As the number of servers n increases, the capacity of the system to process entities concurrently also increases, potentially reducing the average waiting time for customers. This decrease in waiting time can lead to a reduction in the cost associated with customer waiting C_w, as customers spend less time waiting

in queues, thereby decreasing dissatisfaction and potential revenue loss. However, increasing the number of servers also incurs additional cost of service C_s.

Given specific values of C and λ/μ, we want to know the optimal value that minimize the objective function. Minimizing the cost function in (6) is equivalent to minimize the following function

$$Z(n; C, \lambda/\mu, c_a^2, c_s^2) = nC + \lambda/\mu + \frac{n\mu\,(c_a^2 + c_s^2)}{2\,(n\mu - \lambda)}\,(\lambda/n\mu)^{\sqrt{2(n+1)}}, \quad (7)$$

with C is a fixed parameter. Considering the fact that the number of servers is discret, therefore we can not use the analytic approach to find the minimum values where we derive the objective function and look for its root. For this, we follow an heuristic approach. For fixed values of C and λ/μ, then the optimal number of servers k^* satisfy the following inequality

$$Z(k^*; C, \lambda/\mu, c_a^2, c_s^2) \le Z(k^*+1; C, \lambda/\mu, c_a^2, c_s^2),$$

and

$$Z(k^*; C, \lambda/\mu, c_a^2, c_s^2) \le Z(k^*-1; C, \lambda/\mu, c_a^2, c_s^2).$$

This is equivalent to say that

$$k^*C + L_q(k^*, \lambda/\mu, c_a^2, c_s^2) \le (k^*+1)C + L_q(k^*+1, \lambda/\mu, c_a^2, c_s^2)$$

and

$$k^*C + L_q(k^*, \lambda/\mu, c_a^2, c_s^2) \le (k^*-1)C + L_q(k^*-1, \lambda/\mu, c_a^2, c_s^2).$$

Therefore, k^* is optimal if

$$L_q(k^*, \lambda/\mu, c_a^2, c_s^2) - L_q(k^*+1, \lambda/\mu, c_a^2, c_s^2) \le C$$

and

$$C \le L_q(k^*-1, \lambda/\mu, c_a^2, c_s^2) - L_q(k^*, \lambda/\mu, c_a^2, c_s^2).$$

1) Numerical results for the optimal number of servers: Tables I-IV present the optimal solutions across various values of λ/μ and C. Table I specifically addresses the $M/M/n$ case, while Tables II-VII consider the $GI/GI/n$ case under different coefficients of variation for interarrival and service times (c_a^2 and c_s^2). These solutions are obtained using Algorithm 1. Across all tables, a consistent trend emerges: as λ/μ increases, so does the optimal number of servers n^*. This is intuitive, as a higher arrival rate relative to the service rate necessitates more servers to maintain acceptable performance.

Looking at the impact of the cost ratio C, we observe an inverse relationship with n^*. As C increases, n^* tends to decrease. This suggests a trade-off: while more servers improve performance, they also increase costs. The optimal solution balances these competing factors.

The influence of the squared coefficients of variation of interarrival and service times (c_a^2 and c_s^2, respectively) is evident in Tables II, III, and IV. Comparing these tables reveals that higher variability (Table III) generally leads to a higher optimal number of servers for the same λ/μ and C values compared to lower variability (Table II). This is because increased variability in arrival and service times creates more unpredictable queueing behavior, requiring additional server capacity to mitigate performance degradation. Table IV, with moderate variability, falls between the other two, further supporting this relationship.

In summary, the optimal number of servers is positively correlated with the system load λ/μ and variability (c_a^2, c_s^2) and negatively correlated with the cost ratio C. The system cost generally increases with increasing λ/μ and decreases with decreasing C. The interaction of these parameters determines the optimal operating point for minimizing cost while maintaining acceptable performance.

Algorithm 1 Heuristic algorithm for finding optimal n

Require: C_{val}, γ, n_{upper}, c_a^2, c_s^2

procedure $L_q(n, \gamma, c_a^2, c_s^2)$
  $c \leftarrow \left(c_a^2 + c_s^2\right)/2$
  $A \leftarrow (\gamma/n)^{\sqrt{2(n+1)}}$
  $B \leftarrow 1 - \gamma/n$
  return $c \times (A/B)$
end procedure

procedure $C_{diff}(n, \gamma, c_a^2, c_s^2)$
  return $L_q(n, \gamma, c_a^2, c_s^2)$ - $L_q(n+1, \gamma, c_a^2, c_s^2)$
end procedure

procedure FIND_OPTIMAL_N($C_{val}, \gamma, n_{upper}, c_a^2, c_s^2$)
  for $n \leftarrow \lfloor \gamma \rfloor + 1$ to n_{upper} **do**
    if $C_{diff}(n, \gamma, c_a^2, c_s^2) < C_{val} < C_{diff}(n-1, \gamma, c_a^2, c_s^2)$ **then**
      break
    end if
  end for
  return n
end procedure

B. Service rate optimization

1) Cost function with varying service rate: In the scenario of varying service rates, our objective is to minimize the following function with respect to μ:

$$C_T(\mu) = nC_s(\mu) + C_w\Big(\frac{n\mu\,(c_a^2 + c_s^2)}{2\,(n\mu - \lambda)}\,(\lambda/n\mu)^{\sqrt{2(n+1)}} + \lambda/\mu\Big).$$

subject to the constraint $0 < \mu_0 \le \mu \le \mu_1$. In contrast to the fixed-cost definition presented in (5), the cost of service C_s is now a variable function of μ. This adjustment is logical, as the cost of service inherently depends on and correlates with the service rate; an increase in the service rate corresponds to a proportionate increase in the cost of service. The cost of service function $C_s(\mu)$ can manifest in various forms, but in this context, we adopt a linear model. Consequently, $C_s(\mu)$ is an increasing function satisfying $C_s(\mu_0) = C_0$ and $C_s(\mu_1) = C_1 > C_0$. The function $C_s(\mu)$ can be expressed as:

$$C_s(\mu) = a\mu + b$$

TABLE I: Optimal number of servers n for an $M/M/n$ queue under various values of λ/μ and C.

	C	n^*	$Z(n^*;C,\lambda/\mu)$
$\lambda/\mu=3.5$	0.9	5	8.882
	0.5	6	6.748
	0.1	7	4.276
$\lambda/\mu=4.5$	1	6	11.765
	0.8	7	10.491
	0.15	8	5.834

	λ/μ	n^*	$Z(n^*;C,\lambda/\mu)$
$C=0.15$	2	4	2.774
	5.2	9	6.685
	7.4	12	9.342
$C=0.38$	5.8	9	9.523
	7.5	11	12.057
	9	13	14.294

TABLE II: Optimal number of servers n for a $GI/GI/n$ queue with $c_a^2=0.2$ and $c_s^2=0.3$, under various values of λ/μ and C.

	C	n^*	$Z(n^*;C,\lambda/\mu)$
$\lambda/\mu=3.5$	0.9	5	8.242
	0.5	5	6.242
	0.1	6	4.18
$\lambda/\mu=4.5$	1	6	10.841
	0.8	6	9.641
	0.15	7	5.67

	λ/μ	n^*	$Z(n^*;C,\lambda/\mu)$
$C=0.15$	2	4	2.656
	5.2	8	6.515
	7.4	10	9.134
$C=0.38$	5.8	8	9.072
	7.5	10	11.559
	9	11	13.694

TABLE III: Optimal number of servers n for a $GI/GI/n$ queue with $c_a^2=1.5$ and $c_s^2=2.5$, under various values of λ/μ and C.

	C	n^*	$Z(n^*;C,\lambda/\mu)$
$\lambda/\mu=3.5$	0.9	6	9.539
	0.5	6	7.139
	0.1	8	4.407
$\lambda/\mu=4.5$	1	7	12.456
	0.8	7	11.056
	0.15	9	6.03

	λ/μ	n^*	$Z(n^*;C,\lambda/\mu)$
$C=0.15$	2	5	2.889
	5.2	10	6.894
	7.4	13	7.4
$C=0.38$	5.8	10	9.97
	7.5	12	12.545
	9	14	14.818

TABLE IV: Optimal number of servers n for a $GI/GI/n$ queue with $c_a^2=1.2$ and $c_s^2=1.4$, under various values of λ/μ and C.

	C	n^*	$Z(n^*;C,\lambda/\mu)$
$\lambda/\mu=3.5$	0.9	5	9.26
	0.5	6	6.915
	0.1	7	4.362
$\lambda/\mu=4.5$	1	7	12.122
	0.8	7	10.722
	0.15	8	5.959

	λ/μ	n^*	$Z(n^*;C,\lambda/\mu)$
$C=0.15$	2	5	2.841
	5.2	9	6.815
	7.4	12	9.488
$C=0.38$	5.8	9	9.732
	7.5	11	12.306
	9	13	14.544

with

$$a=\frac{C_1-C_0}{\mu_1-\mu_0}\quad b=\frac{C_0\mu_1-C_1\mu_0}{\mu_1-\mu_0}.$$

The cost of service function $C_s(\mu)$ is intuitively positive, necessitating that $a\mu_0+b\geq 0$. With $a>0$, the function is confirmed to be increasing. Attention is directed towards the determination of b, specifically the selection of C_0 and C_1. In particular, we seek to satisfy the inequality:

$$1<\frac{C_1}{C_0}\leq\frac{\mu_1}{\mu_0}.$$

Now, we have the following optimization problem:

$$\begin{aligned}\min_{\mu}\quad & C_T(\mu)\\ \text{s.t}\quad & g_i(\mu)\geq 0\quad \text{for}\quad i=1,2,3\end{aligned}$$

with $g_1(\mu)=\mu-\lambda/n,\ g_2(\mu)=\mu_1-\mu,\ g_3(\mu)=\mu-\mu_0$. We define the Lagrangian for this problem

$$\mathcal{L}(\mu,s)=C_T(\mu)+sg(u),$$

where $s\in\mathbb{R}_+^3$ and $g(\mu)=(g_1(\mu),g_2(\mu),g_3(\mu))^T$. Using Newton-Raphson method, we search for the solution of $\nabla\mathcal{L}(\mu,s)=0$, by iterating the following equation

$$\begin{bmatrix}\mu_{k+1}\\ s_{k+1}\end{bmatrix}=\begin{bmatrix}\mu_k\\ s_k\end{bmatrix}-\begin{bmatrix}\nabla^2_{\mu\mu}\mathcal{L}(\mu_k) & \nabla g(\mu_k)\\ \nabla g(\mu_k)^T & 0\end{bmatrix}^{-1}\begin{bmatrix}\nabla_\mu\mathcal{L}(\mu_k)\\ g(\mu_k)\end{bmatrix}$$

This is equivalent to

$$\begin{bmatrix}\mu_{k+1}\\ s_{k+1}\end{bmatrix}=\begin{bmatrix}\mu_k\\ s_k\end{bmatrix}+d_k,$$

where $d_k=-(\nabla^2\mathcal{L}(\mu_k,s_k))^{-1}\nabla\mathcal{L}(\mu_k,s_k)$. Since the hessian matrix $\nabla^2\mathcal{L}$ is non invertible, we use a different approach to calculate the direction d_k. The SQP (sequential quadratic programming) algorithm suggest that the direction d_k is found by solving the following QP subproblem

$$\begin{aligned}\min_{d}\quad & \nabla C_T(\mu_k)d+\tfrac{1}{2}d^T\nabla^2_{\mu\mu}\mathcal{L}(\mu_k,s_k)d\\ \text{s.t.}\quad & g(\mu_k)+\nabla g(\mu_k)^Td\geq 0,\end{aligned}$$

- Choosing the initial iterate (μ_0,s_0)
- Calculate $\nabla\mathcal{L}(\mu_0,s_0)$ and $\nabla^2\mathcal{L}(\mu_0,s_0)$.

TABLE V: Optimal service rate μ^* for an $M/M/n$ queue with $(C_0, C_1) = (1.2, 3)$ and $C_w = 0.5$, under various λ, n, and (μ_0, μ_1) values.

	λ	(μ_0, μ_1)	μ^*	$C_T(\mu^*)$
$n = 1$	3.5	(4.14, 11.42)	6.159	2.357
	7.6	(8.24, 15.52)	11.520	2.980
	9.2	(9.84, 17.12)	13.513	3.175
$n = 2$	3.5	(2.39, 9.67)	3.142	3.598
	7.6	(4.44, 11.72)	5.809	4.242
	9.2	(5.24, 12.52)	6.803	4.441
$n = 3$	3.5	(1.807, 9.087)	2.149	4.863
	7.6	(3.173, 10.453)	3.914	5.534
	9.2	(3.707, 10.987)	4.572	5.739

TABLE VI: Optimal service rate μ^* for a $GI/GI/n$ queue with $(C_0, C_1) = (1.2, 3)$ and $C_w = 0.5$, under various λ, n, and (μ_0, μ_1) values and different squared coefficients of variation $(c_a^2,\ c_s^2)$.

			$(c_a^2, c_s^2) = (0.2, 0.3)$		$(c_a^2, c_s^2) = (1.5, 2.5)$		$(c_a^2, c_s^2) = (1.2, 1.4)$	
	λ	(μ_0, μ_1)	μ^*	$C_T(\mu^*)$	μ^*	$C_T(\mu^*)$	μ^*	$C_T(\mu^*)$
$n = 1$	3.5	(4.14, 11.42)	4.999	1.967	7.018	2.657	6.457	2.461
	7.6	(8.24, 15.52)	9.694	2.307	12.904	3.491	12	3.157
	9.2	(9.84, 17.12)	11.481	2.410	15.063	3.755	14.048	3.376
$n = 2$	3.5	(2.39, 9.67)	2.619	3.322	3.515	3.815	3.273	3.672
	7.6	(4.44, 11.72)	4.932	3.701	6.445	4.649	6.027	4.383
	9.2	(5.24, 12.52)	5.819	3.812	7.520	4.912	7.053	4.604
$n = 3$	3.5	(1.807, 9.087)	1.846	4.67	2.365	5.021	2.225	4.917
	7.6	(3.173, 10.453)	3.357	5.098	4.304	5.864	4.051	5.649
	9.2	(3.707, 10.987)	3.941	5.218	5.019	6.128	4.729	5.874

TABLE VII: Optimal service rate μ^* for an $M/M/n$ queue with $n = 5$ and $(C_0, C_1) = (1.2, 3)$, under various λ and (μ_0, μ_1) values.

	λ	(μ_0, μ_1)	μ^*	$C_T(\mu^*)$
$C_w = 0.2$	3.5	(0.78, 8.06)	1.077	7.145
	7.6	(1.6, 8.88)	2.043	7.572
	9.2	(1.92, 9.2)	2.410	7.701
$C_w = 0.6$	3.5	(0.78, 8.06)	1.469	8.369
	7.6	(1.6, 8.88)	2.520	9.209
	9.2	(1.92, 9.2)	2.915	9.454
$C_w = 1$	3.5	(0.78, 8.06)	1.791	9.267
	7.6	(1.6, 8.88)	2.915	10.452
	9.2	(1.92, 9.2)	3.328	10.792

TABLE VIII: Optimal service rate μ^* for a $GI/GI/n$ queue with $n = 5$ and $(C_0, C_1) = (1.2, 3)$, under various λ and (μ_0, μ_1) values and different squared coefficients of variation $(c_a^2,\ c_s^2)$.

			$(c_a^2, c_s^2) = (0.2, 0.3)$		$(c_a^2, c_s^2) = (1.5, 2.5)$		$(c_a^2, c_s^2) = (1.2, 1.4)$	
	λ	(μ_0, μ_1)	μ^*	$C_T(\mu^*)$	μ^*	$C_T(\mu^*)$	μ^*	$C_T(\mu^*)$
$C_w = 0.2$	3.5	(0.78, 8.06)	0.945	7.013	1.167	7.248	1.109	7.181
	7.6	(1.6, 8.88)	1.817	7.27	2.203	7.792	2.1	7.649
	9.2	(1.92, 9.2)	2.157	7.342	2.591	7.961	2.474	7.792
$C_w = 0.6$	3.5	(0.78, 8.06)	1.361	8.292	1.559	8.446	1.5	8.394
	7.6	(1.6, 8.88)	2.227	8.948	2.726	9.423	2.593	9.282
	9.2	(1.92, 9.2)	2.562	9.117	3.157	9.721	3	9.546
$C_w = 1$	3.5	(0.78, 8.06)	1.714	9.216	1.867	9.325	1.816	9.286
	7.6	(1.6, 8.88)	2.651	10.251	3.121	10.638	2.987	10.514
	9.2	(1.92, 9.2)	2.988	10.521	3.580	11.032	3.416	10.873

- Build the QP subproblem and find the solution d_0
- Deduce (μ_1, s_1), and then repeat the same process until reaching the convergence.

In practice, implementing this is not straightforward. Challenges such as numerical instability can arise, particularly if the Hessian matrix of the Lagrangian is poorly conditioned. Additionally, selecting a suboptimal initial guess can lead to convergence to a local minimum or even failure to converge. There is also the risk of encountering an infeasible solution.

The QP subproblem is solved using a QP solver (in our case we use SLSQP solver in the subpackage optimize of library Scipy in Python.)

2) Numerical results for the optimal service rate: Our numerical results in Tables V–VIII are obtained using the sequential least squares programming (SLSQP) algorithm. Tables V through VIII analyze the optimal service rate μ^* and its associated cost $C_T(\mu^*)$ under various conditions. Let's break down the key observations and relationships:

Table V establishes the baseline for the $M/M/n$ queue. As expected, for a fixed number of servers n, both μ^* and $C_T(\mu^*)$ increase with λ. This is because a higher arrival rate necessitates a faster service rate to maintain stability and acceptable performance, which in turn increases the cost. Furthermore, as n increases, μ^* decreases for a fixed λ. This reflects the economies of scale offered by multiple servers: with more servers, each individual server can operate at a lower rate while still achieving the desired overall service capacity. However, the total cost $C_T(\mu^*)$ tends to increase with n for a given λ, indicating the trade-off between increased server capacity and the associated expenses.

Table VI introduces the impact of arrival and service time variability (c_a^2 and c_s^2) on μ^* and $C_T(\mu^*)$. Comparing the different variability scenarios, we see that higher variability necessitates a higher μ^* for the same λ and n. This is because increased variability leads to more unpredictable queueing behavior, requiring a faster service rate to compensate and maintain performance targets. Consequently, higher variability also leads to a higher $C_T(\mu^*)$.

Table VII focuses on the influence of waiting cost (C_w) relative to service cost (C_s), represented by the ratio C. As C_w increases (and therefore $C = C_s/C_w$ decreases), both μ^* and $C_T(\mu^*)$ increase for a fixed λ and n. This is because a higher C_w places greater emphasis on minimizing waiting times, driving the need for a faster service rate, even at a higher cost.

Table VIII combines the effects of variability and waiting cost. Similar to Table VII, increasing C_w leads to higher μ^* and $C_T(\mu^*)$ for a given λ and variability scenario. Furthermore, the impact of variability is consistent with Table VI: higher variability leads to higher μ^* and $C_T(\mu^*)$. This table demonstrates the complex interplay between arrival rate, service rate, variability, and cost parameters in determining the optimal service rate and the associated total system cost.

In summary, the optimal service rate μ^* and its associated cost $C_T(\mu^*)$ are positively correlated with arrival rate λ, variability c_a^2, c_s^2, and waiting cost C_w. The number of servers n has an inverse relationship with μ^* but a generally positive relationship with $C_T(\mu^*)$, reflecting the economies of scale versus the cost of additional servers. These tables provide a comprehensive picture of how these factors interact to influence the optimal operating point of a queueing system.

Figures 1-3 provide a visual analysis of the cost function $C_T(\mu)$ within a $GI/GI/n$ queue, specifically for $n = 5$ and $\lambda = 7.6$. They explore the interplay of service rate μ, waiting cost weight (C_w), and variability on total system cost.

Figure 1, a 2D representation, demonstrates the U-shaped cost curves for varying C_w values under fixed variability ($c_a^2 = 1.2$, $c_s^2 = 1.4$). This U-shape signifies the existence of an optimal μ^* that minimizes cost, with this optimal point shifting towards higher μ^* values as C_w increases. This shift reflects the need for faster service to mitigate increased waiting costs as their importance grows.

Figures 2 and 3 extend this analysis into 3D, visualizing C_T as a function of both μ and C_w. Figure 2, representing low variability ($c_a^2 = 0.2$, $c_s^2 = 0.3$), shows a curved cost surface with a distinct valley tracing the optimal μ^* path for different C_w. Figure 3, depicting high variability ($c_a^2 = 1.5$, $c_s^2 = 2.5$), exhibits a similar trend but with a steeper surface and higher overall costs, indicating increased cost sensitivity to μ and the inherent cost penalty of high variability. Across all figures, the consistent upward trend of the cost surface with increasing C_w underscores the significant role of waiting costs in the total system cost. These visualizations emphasize the importance of optimizing μ in response to varying waiting cost considerations and system variability to achieve minimal overall costs.

IV. Conclusion

This paper conducts a comprehensive cost analysis of the $GI/GI/n$ queue, developing a cost function based on established performance metrics. The primary goal is to determine the optimal number of servers and the optimal service rate. A heuristic approach optimizes the server count, while the sequential least squares programming (SLSQP) algorithm is employed to optimize the service rate. Future research could extend this analysis to more general scenarios incorporating practical constraints such as limited system capacity, server breakdowns, server vacations, or customer impatience.

References

[1] A. Komashie, A. Mousavi, P. J. Clarkson, and T. Young, "An integrated model of patient and staff satisfaction using queuing theory," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 3, pp. 1–10, 2015.

[2] M. Singer and P. Donoso, "Assessing an ambulance service with queuing theory," *Computers & Operations Research*, vol. 35, no. 8, pp. 2549–2560, 2008.

[3] W. Zhou, R. Zhang, and Y. Zhou, "A queuing model on supply chain with the form postponement strategy," *Computers & Industrial Engineering*, vol. 66, no. 4, pp. 643–652, 2013.

[4] G. Giambene, *Queuing Theory and Telecommunications*, vol. 585. Springer, 2014.

[5] T. G. Robertazzi, *Computer Networks and Systems: Queueing Theory and Performance Evaluation*. Springer Science & Business Media, 2000.

[6] J. Vilaplana, F. Solsona, I. Teixidó, J. Mateo, F. Abella, and J. Rius, "A queuing theory model for cloud computing," *The Journal of Supercomputing*, vol. 69, pp. 492–507, 2014.

[7] L. Mas, J. Vilaplana, J. Mateo, and F. Solsona, "A queuing theory model for fog computing," *The Journal of Supercomputing*, vol. 78, no. 8, pp. 11138–11155, 2022.

[8] Mohamed Ghazali, and Abdelghani Ben Tahar, "A Queuing Theory Approach to Task Scheduling in Cloud Computing with Generalized Processor Sharing Queue Model and Heavy Traffic Approximation," *IAENG International Journal of Computer Science*, vol. 51, no. 10, pp. 1604–1611, 2024.

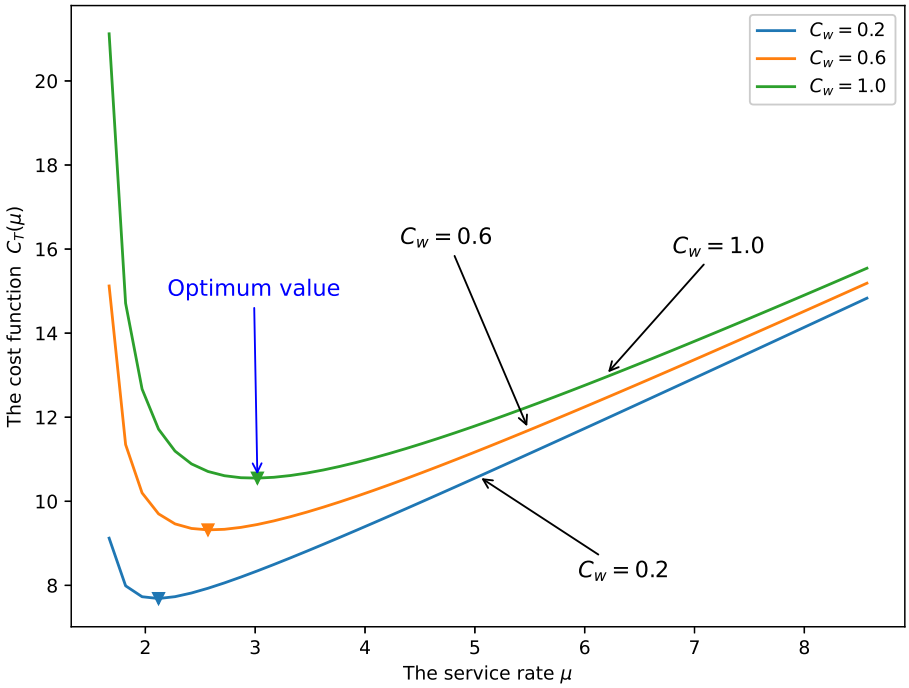

Fig. 1: Cost function $C_T(\mu)$ for a $GI/GI/n$ queue with $n = 5$, $\lambda = 7.6$, $(C_0, C_1) = (1.2, 3)$, $c_a^2 = 1.2$, and $c_s^2 = 1.4$.

[9] S. Zeltyn and A. Mandelbaum, "Call centers with impatient customers: Many-server asymptotics of the M/M/n+ G queue," *Queueing Systems*, vol. 51, pp. 361–402, 2005.

[10] C.-f. Chai, "Problem analysis and optimizing of setting service desks in supermarket based on M/M/C queuing system," in *The 19th International Conference on Industrial Engineering and Engineering Management: Assistive Technology of Industrial Engineering*, pp. 833–841, Springer, 2013.

[11] H. Vass and Z. K. Szabo, "Application of queuing model to patient flow in emergency department. case study," *Procedia Economics and Finance*, vol. 32, pp. 479–487, 2015.

[12] K.-H. Wang and Y.-C. Chang, "Cost analysis of a finite M/M/R queueing system with balking, reneging, and server breakdowns," *Mathematical Methods of Operations Research*, vol. 56, pp. 169–180, 2002.

[13] D.-Y. Yang, K.-H. Wang, and C.-H. Wu, "Optimization and sensitivity analysis of controlling arrivals in the queueing system with single working vacation," *Journal of Computational and Applied Mathematics*, vol. 234, no. 2, pp. 545–556, 2010.

[14] R. Sethi, M. Jain, R. Meena, and D. Garg, "Cost optimization and ANFIS computing of an unreliable M/M/1 queueing system with customers' impatience under N-policy," *International Journal of Applied and Computational Mathematics*, vol. 6, pp. 1–14, 2020.

[15] A. A. Bouchentouf and A. Guendouzi, "Cost optimization analysis for an $M^X/M/c$ vacation queueing system with waiting servers and impatient customers," *SeMA Journal*, vol. 76, pp. 309–341, 2019.

[16] H. Sakasegawa, "AN APPROXIMATION FORMULA L,=-a-p$^{\alpha}$/(1-p)," *Ann. Inst. Statist. Math*, vol. 29, no. Part A, pp. 67–75, 1977.

[17] W. Whitt, "The queueing network analyzer," *The Bell System Technical Journal*, vol. 62, no. 9, pp. 2779–2815, 1983.

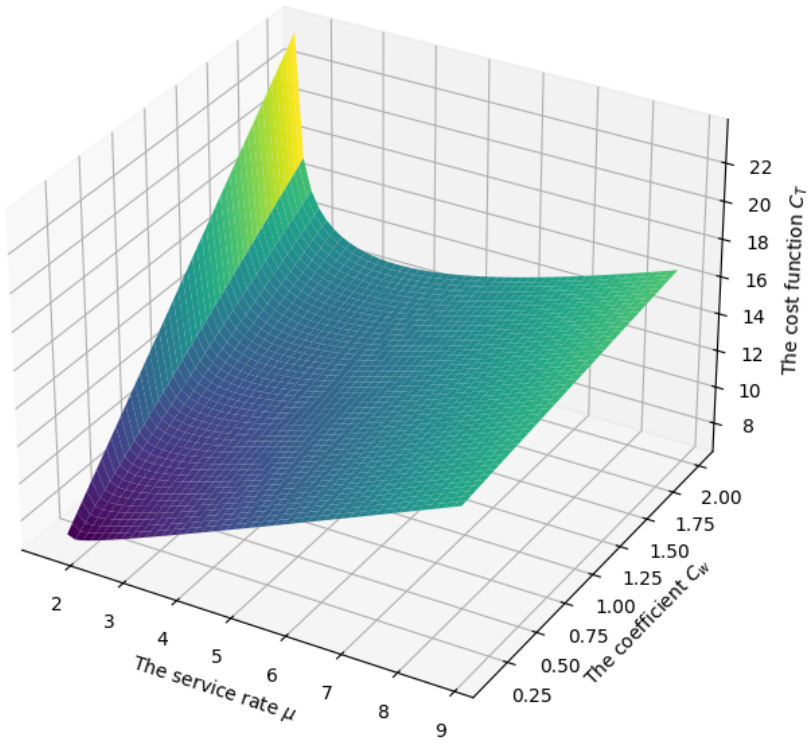

Fig. 2: Cost function C_T as a function of μ and C_w for a $GI/GI/n$ queue with $n = 5$, $\lambda = 7.6$, $(C_0, C_1) = (1.2, 3)$, $c_a^2 = 0.2$, and $c_s^2 = 0.3$.

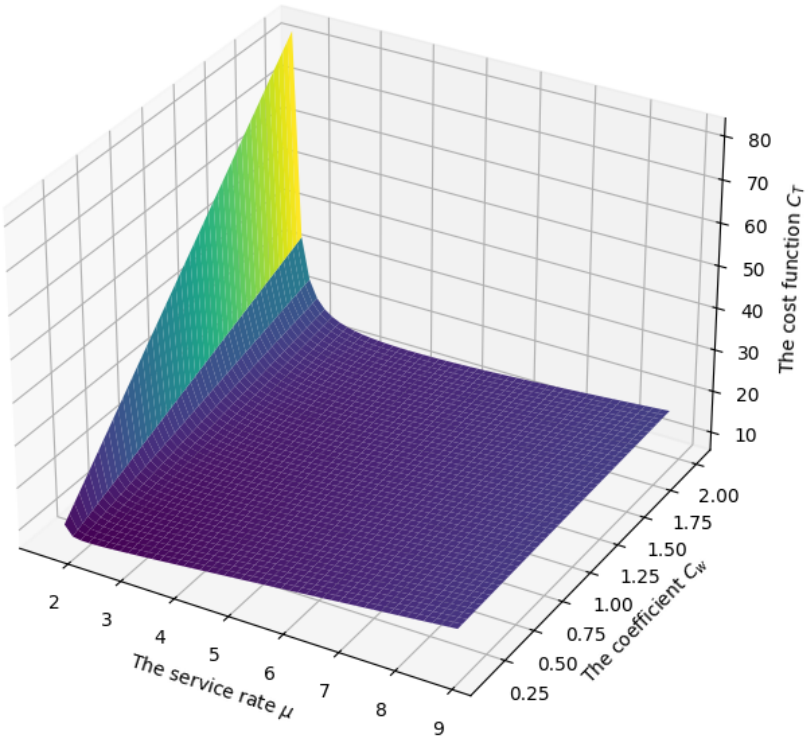

Fig. 3: Cost function C_T as a function of μ and C_w for a $GI/GI/n$ queue with $n = 5$, $\lambda = 7.6$, $(C_0, C_1) = (1.2, 3)$, $c_a^2 = 1.5$, and $c_s^2 = 2.5$.