# Performance assessment of queuing networks with intermittently accessible servers, failed servers, corrective maintenance, feedback, and customer abandonment.

**Abstract**

The objective of this work is to evaluate the performance of M/M/K (K>2) multi-server queuing networks with intermittently accessible servers, failed servers, corrective maintenance, feedback, and customer abandonment. As the model is more complex to analyze numerically, we proceeded by the algorithmic method. First, we used the PSO algorithm to minimize operational costs by dynamically adjusting the $\lambda$ arrival rate, the $\mu$ service level and the number of $K$ servers then in a second step, we used the PSO algorithm to minimize the average waiting time and abandonment rate in order to maximize customer satisfaction. This can therefore be beneficial to both the operator and customers.

**Keywords:** *Queue networks, M/M/K model, PSO algorithm, intermittent servers, servers down, corrective maintenance, aborting, operational costs*

**MSC:** 60K20, 60K25, 60K30

# 1 Introduction

Queuing systems with heterogeneous servers, customer feedback and abandonment have several applications, including manufacturing systems, computer systems, telecommunications systems, etc.

Several recent studies have dealt with queue models with heterogeneous servers, feedback and customer abandonment.

Agassi Melikov and Sevinj Aliyeva [1] studied a system with heterogeneous servers, a Markov modulated Poisson flow and instantaneous feedback. They used approximate algorithms to obtain steady-state probabilities of models with finite and infinite queues and presented results from numerical experiments.

Seenivasan [5] analyzed the performance of two heterogeneous server queue models with one server that can be accessed intermittently. By introducing the bivariate process, he obtained steady-state probabilities using the matrix geometric method. Some numerical results were obtained.

Seenivasan [7] studied a queue model with single working holidays and disasters in which he

considered a system with a single holiday and different service levels. He used the matrix geometric method (MGM) to obtain steady-state probability vectors.

Toky Basilide Ravaliminoarimalalason [6] used game theory to optimize resource sharing in queue networks based on customer aspirations. She used analytical methods and algorithms for static and dynamic models applicable to distributed systems such as cloud computing.

K. Divya [10] studied a model of queues with 2 heterogeneous servers with feedback, one of which is accessible intermittently and the other is assumed to always be accessible without disturbances. He used the matrix geometry method to determine the probability vectors that allowed him to evaluate metrics such as server state probabilities and the average number of customers in the system.

In order to fill the shortcomings of this recent work, we will proceed as follows:

- first, we will introduce intermittencies or periods of downtime for the server that is assumed to be always accessible, provide a detailed analysis of performance metrics and explore strategies to optimize this performance, add a component to model customer impatience and evaluate by numerical illustration, the influence of parameters on system performance.

- second, we will extend the model to include a variable number of servers by integrating optimization algorithms to minimize operational costs or maximize customer satisfaction.

### 1.0.1 Model construction

Consider a system of queues of 2 heterogeneous servers with feedback and customer abandonment. Customers arrive in the system through a Poisson process of $\lambda$ rate. The service levels for servers 1 and 2 are $\mu_1$ and $\mu_2$ respectively and follow an exponential distribution. Server 1 is prone to $\gamma_1$ rate failures and benefits from $\rho_1$ rate corrective maintenance. Server 2 is either idle or intermittently accessible. Customers enter the system based on the status of the servers, i.e. whether they are active or inactive or intermittently accessible. Inactive Server 2 can either become active to provide service at a rate $\eta_0$, or enter a period of intermittent accessibility to provide a service at rate $\eta_1$. After receiving service, a dissatisfied customer may decide to request additional service with a $\overline{\theta}$ rate (feedback) or permanently abandon the system with a $\theta = 1 - \overline{\theta}$ rate.

The state of the system is described by: $S(t) = (N(t), S_1(t), S_2(t))$ with:

- $N(t)$, the number of customers in the system.
- $S_1(t)$ describes the state of server 1 and:

$$S_1(t) = \begin{cases} 0; \text{inactive i.e. no customer in service} \\ 1; \text{active, i.e. serving a customer} \\ 2; \text{i.e. down} \\ 3; \text{i.e. under corrective maintenance} \end{cases}$$

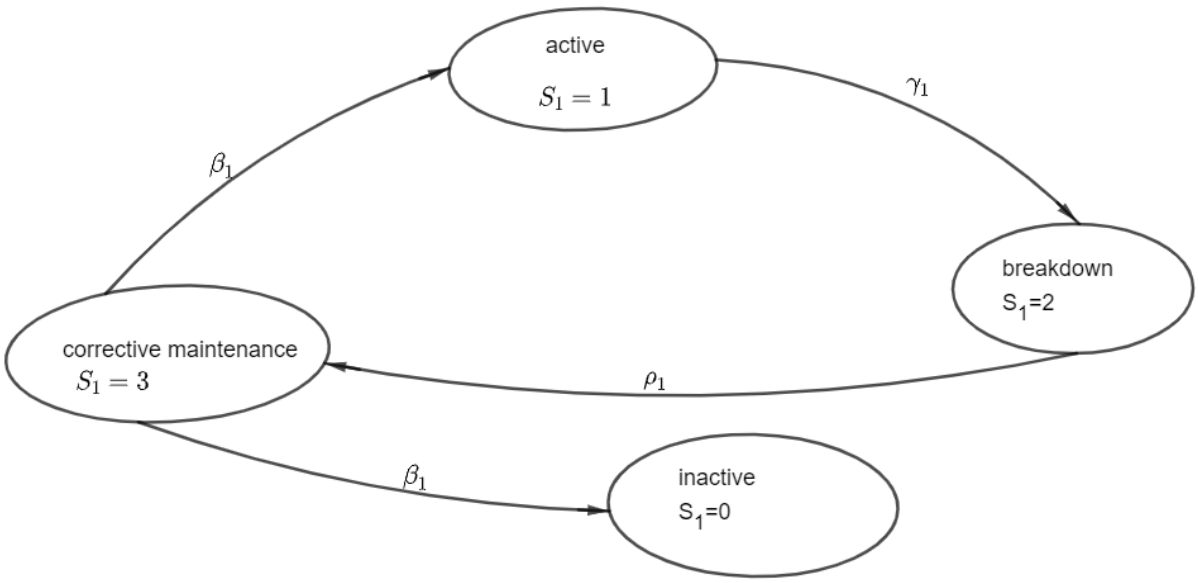
- $S_2(t)$ describes the state of server 2 and:

$$S_2(t) = \begin{cases} 0; \text{inaccessible i.e. cannot serve a customer} \\ 1; \text{accessible i.e. can serve a customer} \end{cases}$$

Hence the state space is:

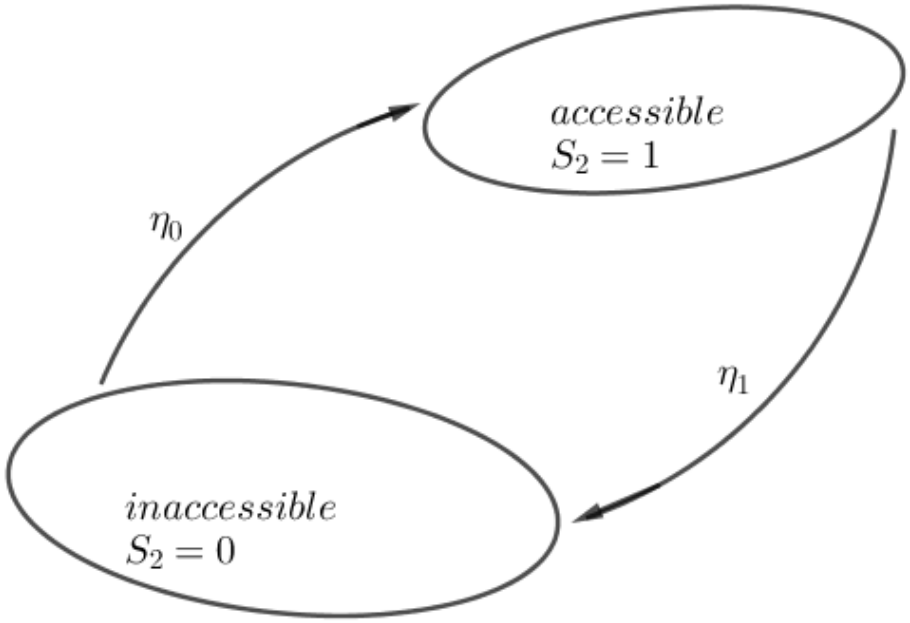
$\Omega = \{(N, S_1, S_2) : N \geq 0, S_1 \in \{0, 1, 2, 3\}, S_2 \in \{0, 1\}\}$

graph 1 : **Internal transition graph for server 1 (subject to failures with corrective maintenance)**




The graph shows that server 1 goes from the active state ($S_1 = 1$) to the failed state with a failure rate of $\gamma_1$ then from the failure state to the corrective maintenance state ($S_1 = 3$) with a rate of $\rho_1$ and then from the corrective maintenance state ($S_1 = 3$) to the idle state ($S_1 = 0$) or active ($S_1 = 1$) with a repair rate of $\beta_1$. When $S_1 = 1$, it can serve a customer with a rate $\mu_1$ reducing the number of customers in the system by 1.

graph 2 : **Internal transition graph for server 2 (intermittent)**




The graph shows that server 2 goes from the accessible state ($S_2 = 1$) to the inaccessible state ($S_2 = 0$) with a rate of $\eta_1$ and then from the inaccessible state ($S_2 = 0$) to the accessible state ($S_2 = 1$) with a rate of $\eta_0$. When Server 2 is accessible, it can serve a customer with a $\mu_2$ rate, but inaccessible, it cannot serve customers.

The infinitesimal generator G is defined by:

$$G = \begin{pmatrix} A_0 & B_0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ B_2 & B_1 & B_0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & B_2 & B_1 & B_0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & B_2 & B_1 & B_0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & B_2 & B_1 & B_0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

with:

• $A_0$, the matrix of internal transitions of servers. It takes into account failures, repairs, corrective maintenance, and alternations between accessible and inaccessible states for Server 2.

• $B_0$, the matrix of customer arrivals that occur at a $\lambda$ rate.

• $B_1$, the matrix of customer departures after service. These departures depend on the active servers $S_1 = 1$ or $S_2 = 1$.

• and $B_2$, the matrix of customer abandonments that occur at a $\theta$. Applying the matrix

geometric method, we get:

$$A_0 = \begin{pmatrix} -(\beta_1+\eta_0) & \eta_0 & \beta_1 & \beta_1+\eta_0 \\ \eta_1 & -(\beta_1+\eta_1) & \beta_1+\eta_1 & \beta_1 \\ \beta_1 & \beta_1+\eta_0 & -(\gamma_1+\eta_0) & \eta_0 \\ \beta_1+\eta_1 & \beta_1 & \eta_1 & -(\gamma_1+\eta_1) \end{pmatrix}$$

$$B_0 = \begin{pmatrix} 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$B_1 = \begin{pmatrix} -(\mu_1+\mu_2+\theta) & \eta_1 & \gamma_1 & 0 \\ \eta_0 & -(\mu_1+\theta) & 0 & 0 \\ \rho_1 & 0 & -(\mu_2+\theta) & 0 \\ \rho_1 & \beta_1 & 0 & -(\mu_2+\theta) \end{pmatrix}$$

$$B_2 = \begin{pmatrix} -\theta & 0 & 0 & 0 \\ 0 & -\theta & 0 & 0 \\ 0 & 0 & -\theta & 0 \\ 0 & 0 & 0 & -\theta \end{pmatrix}$$

Let be $B$ the matrix of the system's global transitions.

We have:

$$\begin{aligned} B &= B_0 + B_1 + B_2 \\ &= \begin{pmatrix} 0 & \lambda & 0 & 0 \\ 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \lambda \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} -(\mu_1+\mu_2+\theta) & \eta_1 & \gamma_1 & 0 \\ \eta_0 & -(\mu_1+\theta) & 0 & 0 \\ \rho_1 & 0 & -(\mu_2+\theta) & 0 \\ \rho_1 & \beta_1 & 0 & -(\mu_2+\theta) \end{pmatrix} + \begin{pmatrix} -\theta & 0 & 0 & 0 \\ 0 & -\theta & 0 & 0 \\ 0 & 0 & -\theta & 0 \\ 0 & 0 & 0 & -\theta \end{pmatrix} \end{aligned}$$

Hence:

$$B = \begin{pmatrix} -(\mu_1+\mu_2+2\theta) & \lambda+\eta_1 & \gamma_1 & 0 \\ \eta_0 & -(\mu_1+2\theta) & \lambda & 0 \\ \rho_1 & 0 & -(\mu_2+2\theta) & \lambda \\ \rho_1 & \beta_1 & 0 & -(\mu_2+2\theta) \end{pmatrix}$$

**Theorem 1.1.** *Let $P = (p_0, p_1, p_2, p_3)$ be the probability vector. The probability vector is ob-*

*tained by solving the equations* $PB = 0$ *and* $\sum_{i=0}^{3} p_i = 1$. *We have the following system:*

$$-\left(\mu_1 + \mu_2 + 2\theta\right) p_0 + \eta_0 p_1 + \rho_1 p_2 + \rho_1 p_3 = 0 \tag{1}$$

$$\left(\lambda + \eta_1\right) p_0 - \left(\mu_1 + 2\theta\right) p_1 + \beta_1 p_3 = 0 \tag{2}$$

$$\gamma_1 p_0 + \lambda p_1 - \left(\mu_2 + 2\theta\right) p_2 = 0 \tag{3}$$

$$\lambda p_2 - \left(\mu_2 + 2\theta\right) p_3 = 0 \tag{4}$$

**Proof 1.1.** *Solving the system by substitution, we get:*

$$p_0 = \frac{\left(\mu_1 + 2\theta\right)\left[\left(\mu_1 + \mu_2 + 2\theta\right)\left(\mu_2 + 2\theta\right)^2 - \gamma_1\rho_1\left(\mu_2 + 2\theta + \lambda\right)\right] - \beta_1\lambda\left[\lambda\left(\mu_1 + \mu_2 + 2\theta\right) + \eta_0\gamma_1\right]}{C + D} \tag{5}$$

*with:*

$$C = \left(\mu_1 + 2\theta + \lambda + \eta_1\right)\left[\left(\mu_1 + \mu_2 + 2\theta\right)\left(\mu_2 + 2\theta\right)^2 - \rho_1\gamma_1\left(\mu_2 + 2\theta + \lambda\right)\right]$$

*and:*

$$D = \left[\lambda\left(\mu_1 + \mu_2 + 2\theta\right) + \eta_0\gamma_1\right]\left[-\beta_1\lambda + \left(\lambda + \eta_1\right)\left(\mu_2 + 2\theta\right) + \lambda\left(\lambda + \eta_1\right)\right]$$

$$p_1 = \frac{\left(\lambda + \eta_1\right)\left[\left(\mu_1 + \mu_2 + 2\theta\right)\left(\mu_2 + 2\theta\right)^2 - \rho_1\gamma_1\left(\mu_2 + 2\theta + \lambda\right)\right]}{C + D} \tag{6}$$

$$p_2 = \frac{\left(\lambda + \eta_1\right)\left(\mu_2 + 2\theta\right)\left[\lambda\left(\mu_1 + \mu_2 + 2\theta\right) + \eta_0\gamma_1\right]}{C + D} \tag{7}$$

$$p_3 = \frac{\lambda\left(\lambda + \eta_1\right)\left[\lambda\left(\mu_1 + \mu_2 + 2\theta\right) + \eta_0\gamma_1\right]}{C + D} \tag{8}$$

Using these probabilities, the following performance measures of the system are deduced:

- Let $L_q$, the average number of customers in the queue, and $\rho$, the occupancy rate of the system.

  We have:

$$L_q = \frac{p_2\rho}{2\left(1 - \rho\right)^2} \tag{9}$$

  Since:

$$\rho = \frac{\lambda}{2\mu} \tag{10}$$

Then:

$$L_q = \frac{p_2 \lambda}{4\mu \left(1 - \frac{\lambda}{2\mu}\right)^2} \tag{11}$$

Replacing $p_2$ with its expression, we have:

$$L_q = \frac{\lambda \left(\lambda + \eta_1\right) \left(\mu_2 + 2\theta\right) \left[\lambda \left(\mu_1 + \mu_2 + 2\theta\right) + \eta_0 \gamma_1\right]}{4\mu \left(1 - \frac{\lambda}{2\mu}\right)^2 (C + D)} \tag{12}$$

- Let be $L$, the average number of customers in the system.

  We have:

$$L = L_q + \frac{\lambda}{\mu} \tag{13}$$

  with $\frac{\lambda}{\mu}$, the average number of customers in service. Replacing $L_q$ with its expression, we have:

$$L = \frac{\lambda \left(\lambda + \eta_1\right) \left(\mu_2 + 2\theta\right) \left[\lambda \left(\mu_1 + \mu_2 + 2\theta\right) + \eta_0 \gamma_1\right]}{4\mu \left(1 - \frac{\lambda}{2\mu}\right)^2 (C + D)} + \frac{\lambda}{\mu} \tag{14}$$

- Let $W_q$, the average wait time in the queue.

  We have:

$$W_q = \frac{L_q}{\lambda} \tag{15}$$

  Replacing $L_q$ with its expression, we have:

$$W_q = \frac{\left(\lambda + \eta_1\right) \left(\mu_2 + 2\theta\right) \left[\lambda \left(\mu_1 + \mu_2 + 2\theta\right) + \eta_0 \gamma_1\right]}{4\mu \left(1 - \frac{\lambda}{2\mu}\right)^2 (C + D)} \tag{16}$$

- Let be $R$, the customer abandonment rate and $T_{threshold}$, the threshold time set by customers.

  We have:

$$R = \begin{cases} 1 \text{ if } W_q > T_{threshold} \\ \dfrac{W_q}{T_{threshold}} \quad \text{otherwise} \end{cases} \tag{17}$$

  Replacing $W_q$ with its expression, we have:

$$R = \begin{cases} 1 \text{ if } W_q > T_{threshold} \\ \dfrac{\left(\lambda + \eta_1\right) \left(\mu_2 + 2\theta\right) \left[\lambda \left(\mu_1 + \mu_2 + 2\theta\right) + \eta_0 \gamma_1\right]}{4\mu \left(1 - \frac{\lambda}{2\mu}\right)^2 (C + D)\, T_{threshold}} \quad \text{otherwise} \end{cases} \tag{18}$$

# 2 M/M/K Network Model (K>2)

## 2.1 Model Construction

Consider a multi-server network of queues of type M/M/K (K>2). Customers arrive in the network through a Poisson process of global rate $\lambda$ and service times are exponentially distributed by global rate $\mu$ and there are $K$ servers (K>2) divided into two categories: some are prone to outages with corrective maintenance and other servers that can be accessed intermittently. Let $\theta$ be the overall rate of customer abandonment when they are dissatisfied with the quality of service. A subset of the servers ($K_{outages}$) can fail with a probability of $p_{outage}$ and the repair time follows an exponential distribution of rates $\mu_r$. Another subset of servers ($K_{intermittent}$) can become temporarily unavailable with a probability of $p_i$. The servers are restored to efficiency after repair.

## 2.2 Modeling Objectives

As the model is more complex to analyze digitally, this requires the use of algorithms in order to:

- minimize operational costs, i.e. reduce costs related to servers (operating or under repair), reduce costs related to breakdowns, reduce costs related to corrective maintenance, reduce costs related to intermittencies, reduce costs related to customer losses i.e. abandonments,
- maximize customer satisfaction, i.e. minimize the abandonment rate and the average waiting time.

We will therefore first use the Particle Swarm Optimization (PSO) algorithm to dynamically adjust the $\lambda$ arrival rate, the $\mu$ service level, the number of servers $K$ and minimize the total operational cost $C$. In a second step, we will use the PSO to minimize the abandonment rate and the average waiting time in order to maximize customer satisfaction, which will therefore be beneficial for both the operator and the consumers (or customers).

## 2.3 Minimizing Operational Costs

Using the PSO algorithm and Using Matlab software, we obtain the optimal values for the arrival rate $\lambda$, the service level $\mu$, the number of servers $K$, the total operational cost $C$ and the performance parameters of the M/M/K networks.

### 2.3.1 Results obtained

Table 1 : The results obtained are recorded in the following table

| | |
|---|---|
| Arrival rate $\lambda$ | 10.00 |
| Service Rates $\mu$ | 10.00 |
| Number of Servers $K$ | 2 |
| Minimum total cost$C$ | 273.54 |
| Occupancy rate $\rho$ | 0.50 |
| Average number of customers in queue $L_q$ | 0.00 |
| Average Wait Time $W_Q$ | 0.00 Hour |
| Average number of customers in the system$L$ | 1.00 |
| Average total time W | 0.10 hour |

This table gives the best configurations of the arrival rate, service level, and number of servers to minimize operational costs. The minimum total operational cost is 273.54 and the optimal values of the arrival rate, service level and number of servers to achieve this minimum cost are 10, 10 and 2 respectively. The average number of customers in the queue as well as the average waiting time are zero.

### 2.3.2 Numerical simulations

By numerical simulation, the PSO algorithm gives the evolution of the minimum total operational cost and the performance parameters of the M/M/K network as a function of the arrival rate, the service rate and the number of servers. The following curves are obtained:

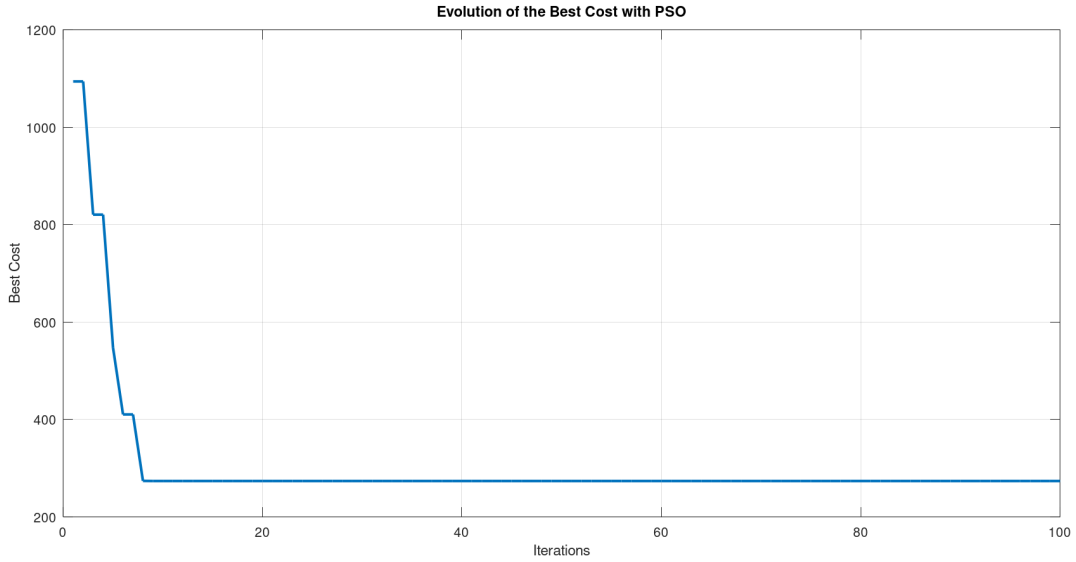
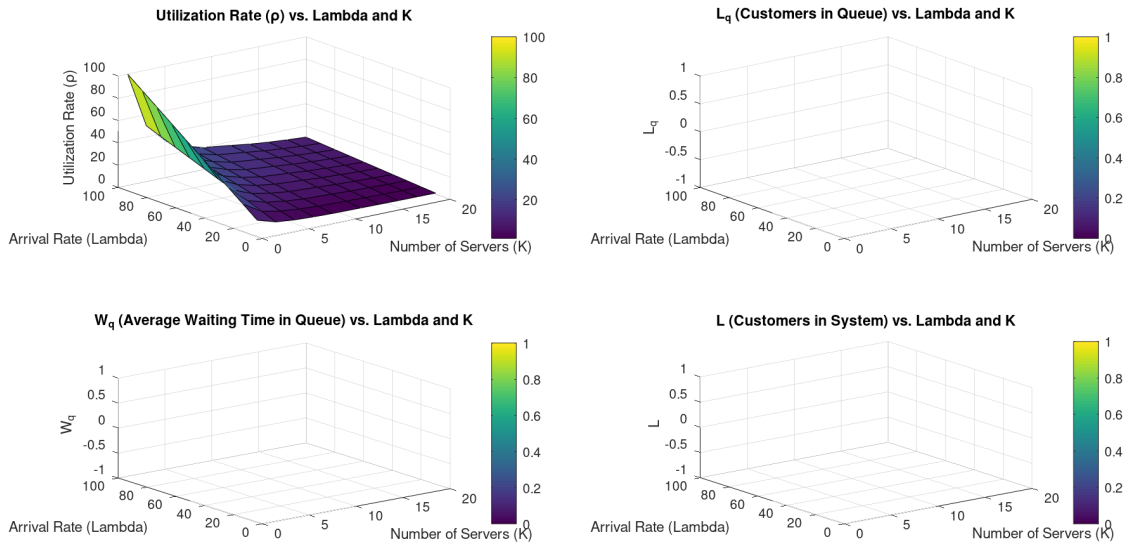



Figure 1:Evolution of the Best cost with PSO

The figure shows that from the first to the tenth iteraton, the minimum total operational cost decreases and from the tenth to the hundredth iteration, the cost remains constant. On the one hand, this indicates better operational efficiency. Increasing the number of servers allows for better load management, reducing outages and downtime. This reflects value for money that improves the overall profitability of infrastructure. On the other hand, the cost remaining fixed despite the increase in the arrival rate, the service level and the number of servers, can be explained by inefficiency in the management of resources.

Figure 2




In these graphs, the blue and green areas indicate a high-performing system with low values for the average number of customers in the queue $L_q$, the average wait time $W_q$, the average number of customers in the system $L$, and moderate occupancy. Orange areas indicate increasing load and high pressure on the system. These areas are therefore points of attention. Black areas indicate critical conditions where performance is unstable or infinite ($\rho \geq 1$).

## 2.4 Maximizing Customer Satisfaction

To maximize customer satisfaction, we use the PSO algorithm to reduce wait time and abandonment rate by using a weighted cost function to combine them.

### 2.4.1 Results obtained

Using the R software, we obtain the optimal values for the arrival rate $\lambda$, the service level $\mu$, the number of servers $K$, the minimum total cost $C$, the minimum average wait time ($W_q$) and the minimum abandonment rate ($R$). These results are recorded in the following table:

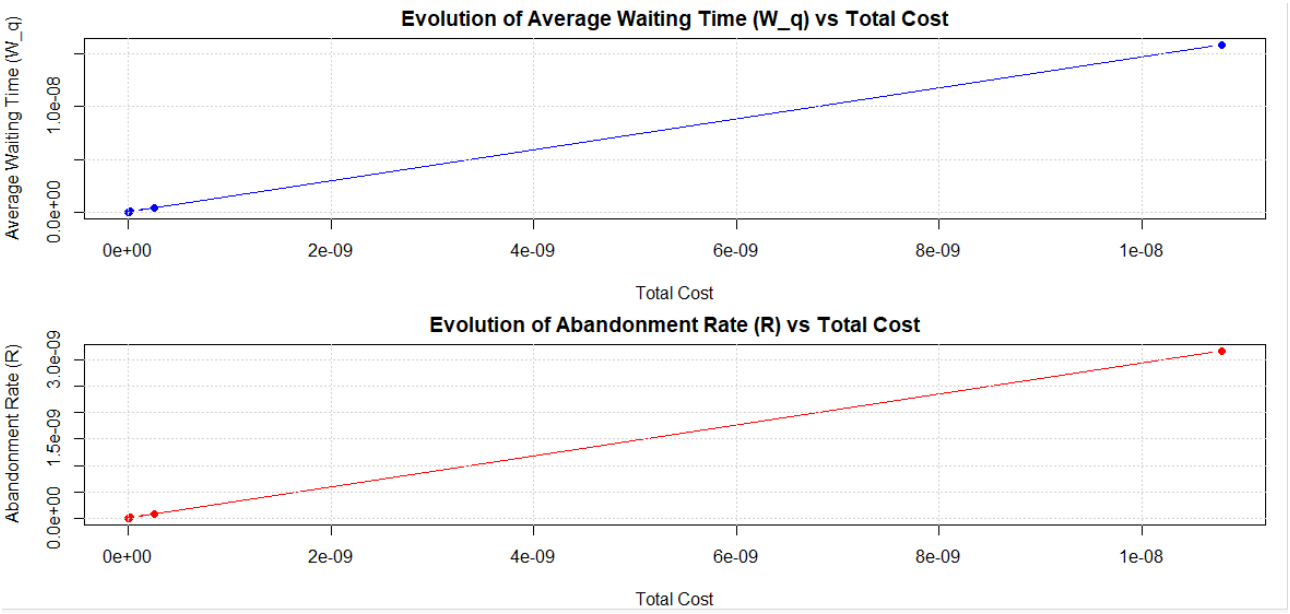
| | |
|---|---|
| Arrival rate $\lambda$ | 10 |
| Service Rates $\mu$ | 10 |
| Number of Servers $K$ | 20 |
| Minimum total cost $C$ | 0 |
| Minimum average wait time $W_q$ | 0 |
| Minimum Abandonment Rate $R$ | 0 |

Table 2. Best configurations of different parameters

This table gives the best configurations of arrival rate, service rate, number of servers, and minimum total cost to minimize abandonment rate and average wait time. The optimal values for the arrival rate, service level, number of servers and minimum total cost are respectively 10; 10; 20 and 0. The minimum average wait time and minimum dropout rate are 0 and 0, respectively.

### 2.4.2 Numerical simulations

By numerical simulation, the PSO algorithm gives the evolution of the abandonment rate and the average waiting time as a function of the minimum total cost. The following curves are obtained: Figure 3:




According to the figure, the increase in total cost leads to an increase in wait time and abandonment rate. Increased wait time leads to an increased risk of customer dissatisfaction as they spend more time in the queue. The simultaneous increase in total cost, average wait time, and abandonment rate is due to a mismatch between demand and resources. This requires the right sizing to maintain acceptable service levels while keeping costs under control.

### 2.4.3 Some proposals

- Increase the number of servers to reduce overhead, which directly decreases the average wait time,

- Ensure that the waiting time remains below the impatience threshold by adjusting the service level or the number of servers,
- Manage demand by implementing strategies to smooth the load,
- Optimize weights, i.e. increase the weight for wait time and weight for abandonment rate so that the OSP prioritizes customer satisfaction.

# 3 Conclusion and Perspective

In this paper, we investigated the performance of M/M/K (K>2) queuing networks with intermittently accessible servers, failed servers with corrective maintenance, and customer abandonment. The model being more complex to analyze numerically, we first used the Particle Swarm Optimization (PSO) algorithm to dynamically adjust the arrival rate $\lambda$, the service level $\mu$, the number of servers $K$ in order to minimize the total operational cost $C$. This cost is related to servers, outages, corrective maintenance, intermittents, and customer abandonment. Second, we used the PSO to minimize the abandonment rate and average wait time to maximize customer satisfaction. Finally, we made proposals to help further maximize customer satisfaction. In perspective, we want to make an extension to the M/G/K model, i.e., by considering general service distributions (G) instead of the exponential distribution (M).

# References

[1] **Agassi Melikov, Sevinj Aliyeva and Janos Sztrik** *Analysis of Instantaneous Feedback Queue with Heterogeneous Servers* In: Mathematics 2020, 8, 2186
doi: 10.3390/math8122186

[2] **T.B. Ravaliminoarimalalason, F. Randimbindrainibe** *Distribution of Occupied Resources on A Discrete Resources Sharing in A Queueing System* In: International Journal of Engineering Research and Technology (IJERT), ISSN 2278-0181, pp 638-643, Vol. 10 Issue 1, 2021

[3] **C. Shekhar, S. Varshney and A. Kumar** *Matrix-geometric solution of multi-server queueing systems with Bernoulli scheduled modified vacation and retention of reneged customers: a meta-heuristic approach* In: Qual. Technol. Quant. Manage. 18 (2021) 39–66.

[4] **S. Thakur, A. Jain and M. Jain** *ANFIS and cost optimization for markovian queue with operational vacation* In: Int. J. Math. Eng. Manage. Sci. 6 (2021) 894.

[5] **M.Seenivasan, M.Indumathi and V. J. Chakravarthy** *Performance Analysis Of Two Heterogeneous Server Queueing Model with Intermittently Obtainable Server Using Matrix Geometric Method* In: Journal of Physics: Conference Series 1724 (2021) 012001
doi: 10.1088/1742-6596/1724/1/012001

[6] **T.B. Ravaliminoarimalalason, F. Randimbindrainibe,** *Occupied resources in a system with resource sharing between service-based customers* In: International Journal of Engineering Research and Technology (IJERT), ISSN 2278-0181, pp 158-164, Vol. 10 Issue 10, 2022.

[7] **M. Seenivasan, R. Senthilkumar and K.S. Subasri** *M/M/2 heterogeneous queueing system having unreliable server with catastrophes and restoration* In: Mater. Today: Proc. 51 (2022) 2332–2338.

[8] **S. Wang, Z. Ma, X. Niu and Y. Liu** *Performance analysis of a Queueing system based on vacation with fault repairable and spare servers in the MP2P network* In: Wireless Networks 29 (2023) 2321–2336.

[9] **N. Singh, M. Jain and S. Dhibar** *ANFIS computing for $M/M/\infty$ queue with two types of service interruption and balking* In: Proc. Nat. Acad. Sci. India Sect. A: Phys. Sci. 94 (2024) 63–73

[10] **K. Divya and K. Indhira** *Performance analysis and ANFIS computing of a markovian queuing model with intermittently accessible server under a hybrid vacation policy* In: RAIRO Operations Research 58 (2024) 1257–1279
https://doi.org/10.1051/ro/2024044