Multi server queuing model with dynamic power shifting and performance efficiency factor

Sreelatha V[1], **Elliriki Mamatha**[1,*], **Chandra S Reddy**[2] and **Krishna Anand S**[3]

[1] *Department of Mathematics, GITAM University, Bengaluru, India*
E-mail: ⟨*vsreelat@gitam.in, mellirik@gitam.edu*⟩

[2] *Department of Engineering, Garden City University, Bengaluru, India*
E-mail: ⟨*saisrimax@gmail.com*⟩

[3] *Department of AI&DS, Shridevi Institute of Engineeing and Technoogy, Tumkur, India*
E-mail: ⟨*skanand86@gmail.com*⟩

Abstract. The need for high performance models in the queuing systems; availability in the fields of computing and communication systems and logistics management poses extensive challenges in design and development of the appropriate modeling. In this work, we propose a robust probabilistic model to mitigate these problems by appropriately choosing a multi-server queuing system with dynamic service facility. The arrival substances, such as packets or jobs or customers or consumers, follow a Poisson arrival process and these arrivals enter a queuing system according to FIFO discipline. The system designed in the system is armed with a fixed number of service stations, in which some servers are capable of allocating additional service capacity by adjusting dynamically. When a customer arrives at the system, if a free server is available, it is immediately served by one of the free processors; if no server is free, that is all are busy, the arrived job is accommodated in the queue and waits for service in the system. In this paper, we proposed a stochastic model to handle peak loads efficiently, boosting service capacity during burst arrivals. Numerical results presented in this paper are generated by the spectral expansion method to demonstrate the model's performance, offering insights into its efficiency and accuracy. Furthermore, we derived some important special cases of speedup factor, which provide the mathematical estimation of system performance in terms of computation and communication times.

Keywords: dynamic service facility and load distribution, mean waiting time, multi-server system, QBD process, spectral expansion method

Received: August 11, 2023; accepted: May 6, 2024; available online: October 7, 2024

DOI: 10.17535/crorr.2024.0014

1. Introduction

In real-life queuing systems, especially during peak periods, it's essential to dynamically adjust server power to accommodate increased customer arrival rates [1, 12, 16]. This flexibility is crucial across various industries, including manufacturing and management. During service time, workload variations are often substantial. At odd times arrivals may be minimal so that servers sit idle during lean periods. Hence, for cost-effectiveness optimizing server utilization is imperative which leads to prolonged lifespan of the servers. Identifying a suitable strategy to allocate power dynamically to the server at the peak period enhances the resources efficiently. It tends to minimize resource wastage, optimize service delivery and responsiveness in queuing systems.

[*] Corresponding author.

The pursuit of high-performance levels has led to numerous challenges in modeling, designing, and developing fault-tolerant wireless systems [15]. Despite the rapid growth of communication services, customers' expectations for availability and performance remain unchanged [21]. Serving signal/data packets and evaluating their performance involves complex computations, leading to scientific challenges in computing, network communication, and logistics systems [6, 17]. This complexity paves the way for designing models where a single unbounded queue evolves as the environment's state changes over time . Instantaneous service to the customer and its arrival largely depends on the environment's state and, to some extent, on the number of customers arriving at the system. Intrinsically, developing a competent model to accommodate this dynamic service facility is essential to meet consumer requirements and safeguarding system safety and reliability.

Addressing additional power allocation requirements and optimizing the service systems necessitates tackling the recent technology-scaling issues. These service issues are closely connected to the great challenges to handle during peak times [19]. To mitigate these issues Power Shifting mode is one viable solution that works efficiently. In this model, during peak arrival, service capacity is bolstered to selected servers by enhancing its service capacity. This approach, proposed in this paper ensures seamless, uninterrupted time bounded service and optimizes the system performance [5, 18]. To alleviate peak power consumption, one of the strategies available in the field is dynamically increasing service capacity [11]. This can be achieved through integrating activity-related power estimation techniques and real-time performance feedback [20].

This process can be easily extended to a multi-server computer system to solve complex problems. We can observe that dynamically power shifting mode successfully improves power management, alleviating monetary burdens for various industrial organizations.

The contributions of this paper encompass the following:

- ▶ Exemplifying the greater efficiency of providing dynamic power allocation over static budgeting.
- ▶ Analyzing critical system and workload factors is essential for the success of power shifting proposals.
- ▶ Proposing performance-sensitive power budget enforcement mechanisms to ensure system reliability.

Present available optimization models in the multiprocessor queuing system addresses various resource allocation strategies in computing, and communication [13]. Its significance is exemplified through a case study where optimal resource allocation for computing is attained by minimizing energy usage while accounting for constraints such as average response time, response time reliability, queuing system stability, and the maximum allowable quantity of resources [14]. This study underscores the importance of incorporating response time reliability to ensure service quality in cloud computing resource allocation [8].

The model describes a network comprising power-constrained nodes transmitting over channels, such as wireless links with adaptive transmission rates, as outlined in [15, 22]. Customers randomly enter the system and await service in the queue at each node, with their data transmitted through the network to respective destinations. To examine various traffic organization levels, a mathematical model is formulated using rate matrices for support. The design involves power allocation and joint routing distribution to stabilize the system and ensure bounded average service guarantees when input rates fall within the capacity region [23]. This performance holds for both centralized and decentralized implementations, considering general arrival and channel state processes. The network system is monitored, and the system stability of decentralized algorithms is studied concerning a mobile relay strategy.

In the work [2], a dynamic power control problem is addressed, focusing on two similar-level service states subject to random variations in connectivity and switchover server delays between queues. In each time slot, the server determines whether to maintain a constant level of service capacity or switch to an additional power level, thus increasing the service capacity [9]. This decision is based on the current connectivity and queue length information. The introduction of switchover time as a modeling parameter adds a new layer of complexity, enhancing the overall interest in the problem.

To describe system stability, a novel approach is proposed. In this method employs state-action frequencies to identify stationary solutions of the Markov Process and formulate a corresponding structured plan [3, 7]. The stability region is characterized with respect to connectivity parameters. This characterization aids in the development of a new framework for throughput-optimal network policies, supported by state-action frequencies.

2. Mathematical Model and QBD Process System

A queuing network system is modeled in terms of a discrete two-dimensional Markov process on a semi-infinite lattice strip. The process follows a Markovian property, and the transition state of the system at observation time t can be expressed by two random variables $I(t)$ and $J(t)$ used to study the system state at any time t is represented by a two-dimensional random vector. $I(t)$ represent the system's operative state, and its values belong to $[0, N]$ interval of integers, whereas $J(t)$ represents the number of customers present in the queue (including served customers) that may be finite or infinite depending on the queuing system. The Markov process for the QBD queuing system is denoted by: $X(t) = [I(t), J(t); t \geq 0]$ with its state space $[0, 1, 2, ...N] \times [0, 1...]$ for infinite queue size. Let $X(t)$ represents the customer's position at discrete time t with mean arrival and service rates. The number of births at an i^{th} time interval $(t, t + \Delta t)$ with time Δt would be $\sigma \Delta t + o(\Delta t)$.

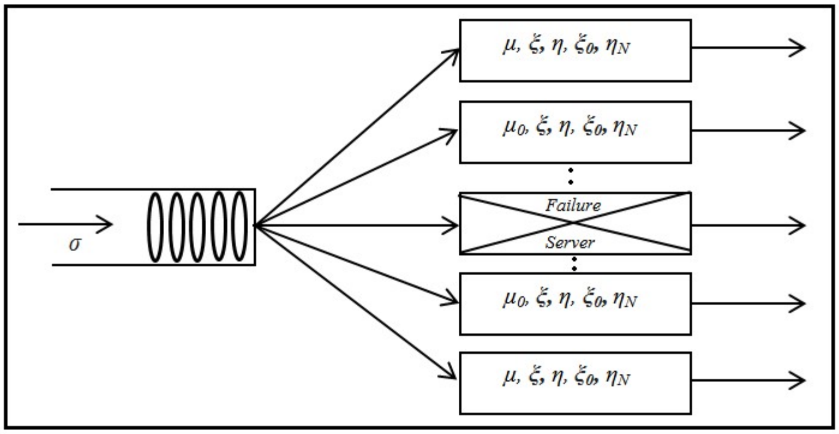
$\mu, \xi, \eta, \xi_0, \eta_N$
$\mu_0, \xi, \eta, \xi_0, \eta_N$
Failure
Server
σ
$\mu_0, \xi, \eta, \xi_0, \eta_N$
$\mu, \xi, \eta, \xi_0, \eta_N$

Figure 1: *Parallel processor servicing system with dynamic power.*

Let there be $N + 1$ processor configurations, represented by the values $I(t) = 0, 1, \ldots, N$, denoting operative states of the multiprocessor system. These configurations constitute the operative states of the model, and the model assumptions ensure that $I(t)$, for $t \geq 0$, forms an irreducible Markov process. Subsequently, $X(t) = \{I(t), J(t)]; t \geq 0\}$ represents an irreducible Markov process on a lattice strip (a QBD process) that model the system. This system has been scrutinized for exact performability [4, 10], even under infinite waiting time, i.e., for $L \to \infty$.

Matrix A is the instantaneous transition rates from operative state i to operative state k, with zeros on the main diagonal, indicating purely lateral transitions within the model X. Matrices B and C serve as transition matrices for one-step upward and one-step downward transitions, respectively. Moreover, when the transition rate matrices do not depend on j, the system reaches a steady state for $j \geq M$, where M, an integer constant, represents a threshold value, and the process X evolves through the following instantaneous transitions:

- A_j: Purely lateral transition rate, from state (i, j) to state (k, j), $(i = 0, 1, \ldots, N, k = 0, 1, \ldots, N; i \neq k; j = 0, 1, \ldots, L)$, caused by a change in the operative state (i.e. servers sleeping or break-down followed by service up during arrival, and a repair time).
- B_j: One-step upward transition rate matrix, from state (i, j) to state $(k, j + 1)$, $(i = 0, 1, \ldots, N, k = 0, 1, \ldots, N,$ and $j = 0, 1, \ldots, L)$, caused by a job arrivals into the system.
- C_j: One-step downward transition rate matrix, from state (i, j) to state $(k, j - 1)$, $(i = 0, 1, \ldots, N, k = 0, 1, \ldots, N,$ and $j = 0, 1, \ldots, L)$, caused by the departure of a serviced job.

Let the power be allocated dynamically to the system at the stages where the system needs more power to process customers' delays during peak time. At peak times, it requires more customer service; hence it requires extra power to accomplish the work. During this busy period, additional servers are dynamically assigned to the system. In the regular period, the system operates with a fixed service capacity. The parameter μ is the normal mean service rate and μ_0 is the dynamically allocated power to service for the system during a busy time. It can be observed that $J(t)$ is a process that may move up or down depending on the customer's arrival or departure from the queue. $I(t)$ is the service state corresponding to the arrivals that takes values $0, 1, 2, \ldots, p, q, q + 1, q + 2, \ldots, N$. Here the states 0 to p represent the system is working in normal mode, whereas from state q to N the system moves work with dynamic mode.
The QBD Markov system can be expressed in mathematical form as:

$$
\begin{aligned}
x_1(t+1) &= S_{p+1} x_{p+1}(t) + ... + S_{p+q} x_{P+q}(t) - (\mu_1 + \sigma_1) x_1(t) \\
x_2(t+1) &= \sigma_1 x_1(t) - (\mu_2 + \sigma_2) x_1(t) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
x_{p+q+r-1}(t+1) &= \sigma_{p+q+r-2}(t) x_{p+q+r-2}(t) - (\mu_{p+q+r-1} + \sigma_{p+q+r-1}) x_{p+q+r-1}(t) \\
x_{p+q+r}(t+1) &= \sigma_{p+q+r-1} x_{p+q+r-1}(t) - \mu_{p+q+r} x_{p+q+r}(t)
\end{aligned}
\tag{1}
$$

$x_i(t)$ indicates the system state at i^{th}transient position with arrival and service rates σ_i, μ_i respectively. At the r^{th} state the system moves from the transient state to the steady state, so that the system will not depend on arrivals. Equation (1) can be modeled by the spectral expansion method with transient matrices Aj, Bj, and Cj where the mean arrival rates, the service rates, and the additional dynamic allocations are denoted by σ μ and μ_0, respectively. The system is prone to breakdown either a single server randomly, with a mean rate ξ or bulk servers at a rate ξ_0. The service rates to repair these servers are represented by η and η_N for single server and all servers, respectively. It is explained in Figure 1. The service for the packets arrived at the system is followed by FIFO (first in - first out) discipline. Once the service is completed, packets are dispatched from the system. Matrix A_j is purely phase transitions representing services, and B_j is the upward transitions matrix representing the customers' new arrival. Matrix C_j represents the downward transition matrix. It represents the number of customers serviced during the system up.

As previously stated, matrix A_j represents purely phase transitions representing services, whereas matrix B_j denotes the upward transitions matrix depicting customers' new arrivals. Conversely, matrix C_j represents the downward transition matrix, indicating the number of customers serviced during the system's operation.

$$A_j = \begin{bmatrix} 0 & N\eta_0 & 0 & \cdots & 0 & N\eta_N \\ \xi + \xi_0 & 0 & (N-1)\eta_0 & \cdots & 0 & N\eta_N \\ \xi_0 & N\eta_0 & 0 & \cdots & 0 & N\eta_N \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \xi_0 & N\eta_0 & 0 & \cdots & 0 & \eta_N + \eta_0 \\ \xi_0 & 0 & 0 & \cdots & N\xi & 0 \end{bmatrix} \tag{2}$$

$$B_j = \begin{bmatrix} \sigma & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma \end{bmatrix} \tag{3}$$

$$C_j = \begin{bmatrix} min(0,j)\mu & 0 & \cdots & \mu_0 & \cdots & \mu_0 & 0 & \cdots & 0 \\ 0 & min(1,j)\mu & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & min(p+1,j)\mu & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & min(N,j)\mu \end{bmatrix} \tag{4}$$

At threshold point M, the transition matrices reach steady states and become level independent. In this steady state, these matrices no longer depend on the parameter j, transitioning to steady state irreducible matrices A, B, and C respectively.
In matrix from

$$\begin{aligned} A_j &= A, \text{ for all } j \geq M \\ B_j &= B, \text{ for all } j \geq M \\ C_j &= C, \text{ for all } j \geq M \end{aligned} \tag{5}$$

For calculating various parameters, the probability of state (i, j) in the steady state is computed using the probability coefficient $P_{i,j}$ which has been introduced and defined as follows:

$$P_{i,j} = lim_{t \to \infty}[I(t) = i, J(t) = j], \; 0 \leq i \leq N, \& \, 0 \leq j \leq L \tag{6}$$

Where L can be finite or infinite.
The probability row vectors at the j^{th}stage are defined as

$$V_j = [P_{0,j}, P_{1,j}, ...P_{N,j}], \;\; j = 0, 1, 2... \tag{7}$$

The probability vectors, along with transient state matrices, have been represented with the help of balance equations.

$$V_0[D_0^A + D_0^B] = V_0 A_0 + V_1 C_1 \tag{8}$$

$$V_j[D_j^A + D_j^B + D_j^C] = V_{j-1}B_{j-1} + V_j A_j + V_{j+1}C_{j+1} \tag{9}$$

$$V_j[D^A + D^B + D^C] = V_{j-1}B + V_j A + V_{j+1}C \; M \leq j \leq L \tag{10}$$

$$V_L[D^A + D^C] = V_{L-1}B + V_L A \tag{11}$$

For an infinite state space, the balance equation is further simplified.

$$V_{j-1}B_{j-1} + V_j[A_j - D_j] + V_{j+1}C_{j+1} = 0, \; j = 0, 1 ... M - 1 \tag{12}$$

In equation (12), $D_j = D_j^A + D_j^B + D_j^C$ where D_j^R ($R = A$ or$B$ or C) represents the diagonal matrix, whose diagonal elements are the sum of each corresponding row of the matrix R_j.

The threshold condition for the system attains at $M = N$. After reaching this threshold condition, the system enters a steady state, i.e., the system's behavior no longer depends on j. Consequently, balance equations for $j = M, M + 1, \ldots$ are as follows.

$$V_{j-1}B + V_j[A - D] + V_{j+1}C = 0 \tag{13}$$

Furthermore, the total probability of the system always remains at 1, i.e.,

$$\sum_{j=0}^{\infty} V_j.e = 1 \tag{14}$$

Here $'e'$ represents the n^{th}- order column matrix with elements equal to 1.

To find the probability vectors V_j, the balance equations can be reformulated in terms of eigenvalues and eigenvectors. Equation (14) leads to a quadratic equation from which eigen values and their corresponding left eigen vectors can be derived. These values will be essential in computing performance measures.

Let's define diagonal matrices $Q_0, Q_1,$ and Q_2 with sizes $(N + 1) \times (N + 1)$ from the study state matrices $A, B, C,$ as $Q_0 = B, Q_1 = A - D^A - D^B - D^C, Q_2 = C$.
Then the balance equations can be expressed in terms of quadratic form.

$$V_j Q_0 + V_{j+1}Q_1 + V_{j+2}Q_2 = 0; \quad \text{where } (M - 1) \leq j \leq (L - 2) \tag{15}$$

From this, the characteristic matrix polynomial further can be expressed as:

$$Q(\lambda) = Q_0 + Q_1(\lambda) + Q_2\lambda^2 \tag{16}$$

Where

$$\psi Q(\lambda) = 0; \; |Q(\lambda)| = 0. \tag{17}$$

Here λ, and ψ are the eigenvalues and left eigenvectors of the quadratic polynomial $Q(\lambda)$ respectively.

To compute eigenvectors and their corresponding eigenvalues, we further simplify the quadratic form in matrix form.

$$Q = \begin{bmatrix} 0 & -T_0 \\ I & T_1 \end{bmatrix} \tag{18}$$

where $T_0 = B/C$ and $T_1 = [A–D_A–D_B–D_C]/C$

It can be further expressed in quadratic notation form as:

$$\psi[T_0 + T_1\lambda + \lambda^2] = 0 \tag{19}$$

It can be expressed in a square matrix form of order $2N$ to compute eigen values and their corresponding eigen vectors.

$$\psi \begin{bmatrix} 0 & -T_0 \\ I & T_1 \end{bmatrix} = \lambda\psi \tag{20}$$

Since the matrix $\begin{bmatrix} 0 & -T_0 \\ I & T_1 \end{bmatrix}$ is a $2N$ size matrix, hence it has a $2N$ set of eigen values and their corresponding left eigen vectors. Out of these $2N$ eigenvalues, N eigenvalues are exactly less than one in magnitude.

Now, by considering these N eigenvalues and their corresponding half of left eigenvectors, the probability vectors can be computed as:

$$V_j = \sum_{(k=0)}^{N} a_k\psi_k\lambda_k^{j-M+1} \tag{21}$$

Inverting the next N eigenvalues that are greater than one and considering their corresponding half of left eigenvectors, the probability vector can be defined for the finite state as:

$$V_j = \sum_{(k=0)}^{N} a_k\psi_k\lambda_k^{(j-M+1)} + \sum_{k=1}^{N} b_k\phi_k\beta_k^{L-j} \tag{22}$$

Where a and b are arbitrary constants. β and ϕ are the reciprocal eigenvalues with magnitude greater than or equal to one and their left eigenvectors of the matrix Q, respectively.

Which can be further resolved in the form of state probabilities as follows:

$$p_{i,j} = \sum_{k=0}^{N} a_k\psi_k(i)\lambda_k^{j-M+1} + b_k\phi_k(i)\beta_k^{L-j} \text{ where } M-1 \leq j \leq L \tag{23}$$

The arbitrary constants a_k and b_k, $(k = 0, 1, \ldots, N)$ are either scalar real constants or complex values that need to be computed. These constants can be found with balance equations to compute various performance measures such as service probability, mean waiting time and loss probability of the customers, and other measures.

3. Performance Efficiency Factor

In the contemporary world, the ever-growing need for enhanced computational speed is fueled by technological advancements. One approach to achieving this goal involves partitioning computational tasks among multiple servers. This can be realized through the implementation of a coupled system or by leveraging cloud computing techniques.

While pursuing the development of a new system, it is crucial to consider the limitations of network communications and potential delays in work arising from increased computational and transmission times. Additionally, the system should be designed to be cost-effective. This section delves into the discussion of boosting speed through the utilization of multiple servers, acknowledging certain constraints and employing parallel server configurations.

The computation of the parallel computing performance factor has been considered. The speedup factor ($S_{up}(p)$) is defined to measure adequate performance by adding additional servers.

$$\begin{aligned} S_{up}(p) &= \frac{t_s}{t_p} = \frac{\text{Execution time of single server}}{\text{Execution time of multiple servers}} && (24) \\ &= \frac{2n^2}{2\frac{n^2}{p} + p(t_{startup} + t_{data}) + n^2(t_{startup} + 4t_{data})} && (25) \\ &= \frac{n(2n+4)}{\frac{n}{p}(2n+4) + p(t_{startup} + nt_{data})} && (26) \end{aligned}$$

The effect of computation and communication times play crucial role in performance of a system. As the size of the system increases, exchanging information among the servers gradually increases. The rate between these two factors can be expressed as:

$$t_{p/c} = \frac{2\frac{n^2}{p}}{p(t_{startup} + 2t_{data}) + 4n^2(t_{startup} + t_{data})} \tag{27}$$

Where $t_{p/c}$ represents the ratio processing time over communication time.

$$t_{p/c} = \frac{t_{comp}}{t_{comm}} = \frac{\frac{n}{p}(2n+4)}{pt_{startup} + nt_{data}} = O\left(\frac{n^2/p}{p + n^2}\right) \tag{28}$$

Which suggests improvement with larger n (scalable).

Several factors affect the maximum not performing speed of the system. These factors include:

- All servers are not performing effectively, and in the meantime, some servers are idle.
- Communication between processes is another factor in reducing speed.
- Effective utilization of other peripherals in multiple service systems.

Anticipating heightened customer arrivals during peak times, it is rational to allocate additional power, while reverting to standard services during non-peak periods. Initially, N servers are designated for service. When faced with peak arrivals, additional M servers are dynamically assigned to bolster power for seamless task execution. The fraction of work, denoted as f, is undertaken by the extra power servers (M servers), while the remaining fraction of work $(1-f)$ is handled by uniform servers ($N - M$ servers). Let t_s be the total time required to complete the work, p represent the single server performance speed factor, and k be the dynamic power increasing factor from servers. This graphical representation is illustrated in Figure 2, depicting the relationship between the performance efficiency factor and dynamic power allocation for a portion of the work.

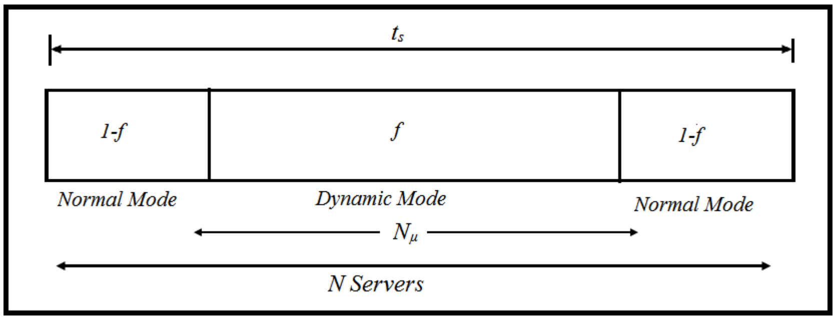

Figure 2: *Performance efficiency factor and dynamic power allocation for part of the work.*

Then the speedup performance factor represents.

$$S_{up}(p) = \tfrac{t_s}{f t_s \frac{M}{P} + (1-f) t_s \frac{N-M}{kp}} \tag{29}$$

$$= \tfrac{kp}{N-M+f[(k+1)M-N]} \tag{30}$$

From this, the following cases can be derived.

Case 1: If a customer arrives uniformly throughout the period, the service is performed typically. Hence, the fraction of work computed with dynamic mode becomes zero.

In this case, $M = 0$. Hence, the speedup performance factor becomes:

$$S_{up}(p) = \frac{kp}{N} \tag{31}$$

Case 2: If the service is expected with extra power throughout the system life, every server works dynamically. In this case, M becomes N. Hence:

$$S_{up}(p) = \frac{p}{N} \tag{32}$$

4. Result Analysis

This section presents numerical results in graphical format, utilizing various parameters to assess the performance of queuing systems when dynamic services are employed during busy periods. The comprehensive performance measures are provided for the proposed approach to compare with traditional method. The results indicate that the performance is notably improved with the proposed dynamic mode power shifting method. It demonstrates effective job handling during peak periods, while normal mode service can be seamlessly executed with uniform service during the system's general arrivals. Simulation results, depicted graphically, offer insights into the model's performance as described in the preceding section, with modifications to various parameters.

In presenting the results, certain parameters have been kept constant unless explicitly mentioned for a specific experiment. Unless otherwise it is not mentioned, the parameters are fixed as $= 1.8$, $\mu_0 = 0.1$, $\eta_N = 0.01$, $\xi_0 = 0.02$, $\xi = 0.8$, and the maximum number of servers operating in the system, $N = 10$, have been consistently maintained throughout the experiments. A substantial effort has been invested in ensuring a high degree of accuracy in this work.

Figures 3 and 4 depict a queuing system with varying service rates while maintaining a fixed number of servers and uniform service distribution. The illustrations demonstrate that as the arrival rate increases, both the mean queue length and the waiting time for service also increase in response to the influx of job arrivals. These observations suggest an exponential relationship between the mean queue length, waiting time, and the rate of job arrivals. Figure 3 specifically portrays the queue length, while Figure 4 represents the time required for service.

In Figure 5, the attention is directed towards dynamic service stations, showcasing different power factors assigned to servers spanning from station 2 to station 8. Additional powers are distributed across three distinct levels: server 2 to 5, server 2 to 6, and server 2 to 8. Through this depiction, it can be deduced that the inclusion of additional servers with dynamic service capacities results in a decrease in the average waiting time for arrivals, as visually demonstrated in the graph. Figure 6 illustrates the escalating number of customers awaiting service as the arrival throughput steadily increases. This graph showcases dynamic service stations ranging from 2 to 8, each operating at different levels of service capacity.

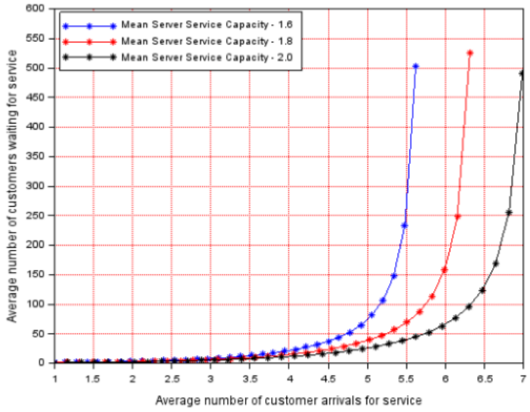

Figure 3: *Number of customers waiting for service in the queuing system over time.*

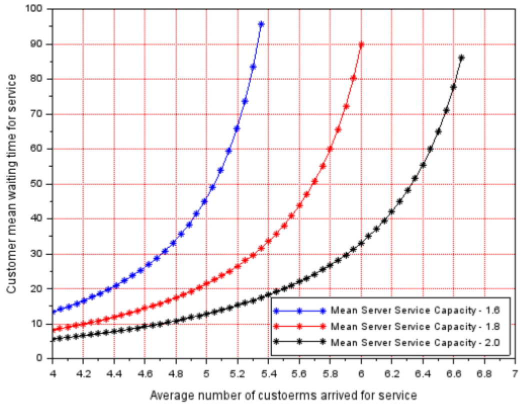

Figure 4: *Expected time of the service to the customer with mean arrival rate increases.*

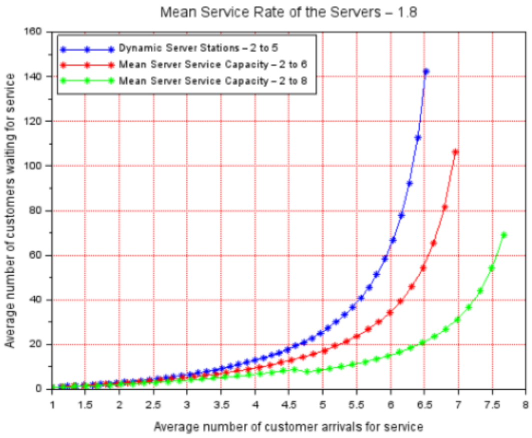

Figure 5: *Expected number of customers waiting for service with dynamic power service facility.*

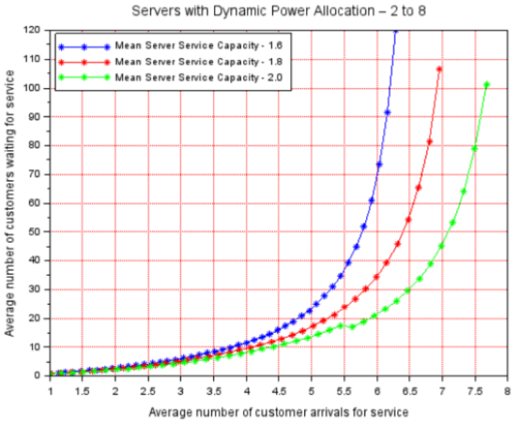

Figure 6: *The rate of increase in the number of customers varies with different levels of service capacity and arrival rates.*

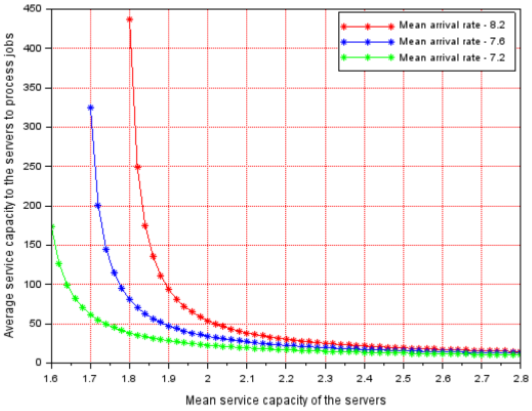

Figure 7: *The anticipated number of customers waiting for service upon joining the queue, as service capacity increases.*

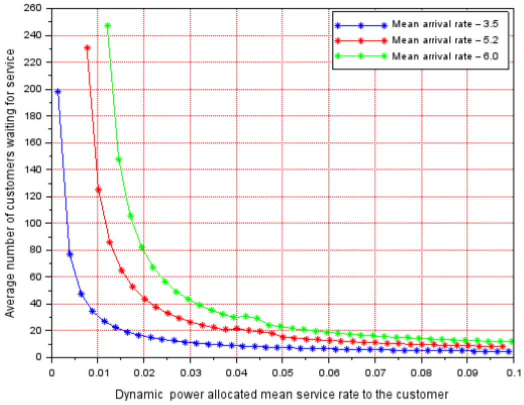

Figure 8: *Expected number of customers waiting for service as service capacity increases with server power adjustment.*

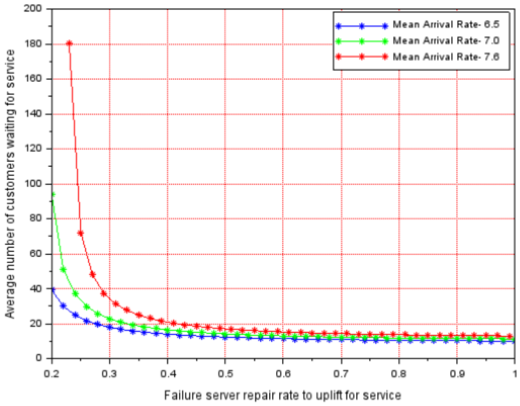

Figure 9: *Number of customers waiting for services across various arrival rates, concerning the improvement after uplifting failure server.*

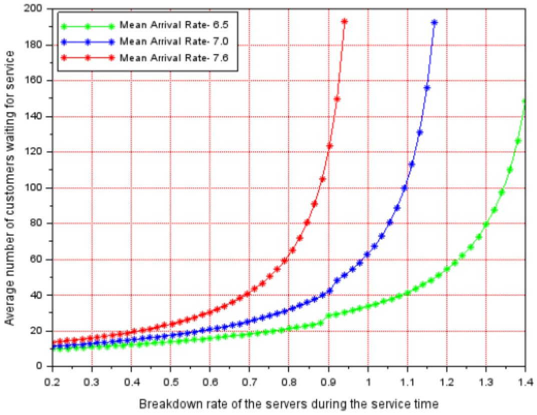

Figure 10: *Correlation between the number of customers waiting for service with servers prone to breakdowns.*

Both Figure 7 and Figure 8 depict constant arrival rates across different service capacities. In Figure 7, arrival rates of 8.2, 7.6, and 7.2 are examined within a general service system, while Figure 8 explores arrival rates of 3.5, 5.2, and 6.0 with dynamic service power shifting. The results illustrate the influence of additional power allocation to the server on the mean queue length. It's observed that as service capacity dynamically increases, there's a corresponding reduction in waiting time until it stabilizes at a constant level. This stabilization point indicates the optimal utilization level, which can be calculated from these observations. Overall, it's noted that as service capacity increases, waiting time decreases.

Referring to Figures 9 and 10, the graphs illustrate outcomes under varying arrival rates, mirroring the patterns observed in the preceding figures. In both scenarios, the machine service capacity remains constant, set with appropriate parameter values. Figure 9 reveals an intriguing observation: during periods of fast services for failure servers, service is promptly completed. Notably, an escalation in repair rate results in a significant reduction in waiting time initially, until stability is achieved. Turning to Figure 10, the findings depict the impact of varying levels of server breakdowns on service time. It becomes evident that as the breakdown rate increases, the time taken to serve waiting customers also increases. This relationship underscores a direct proportionality between service time and failure rate.

5. Conclusion

As global customer demand continues to surge, many systems that were once efficient have become obsolete. Sustained survival requires ongoing improvements in methodologies across various real-time applications. This work addresses the need for continuous enhancement in one such application, focusing on the development of new strategies to minimize customer waiting time in queues during peak periods. The novelty of this approach lies in its dual objectives: reducing waiting time during peak hours while ensuring servers remain active during less busy periods. The validation of these methodologies with numerical values has been performed, demonstrating the practical applicability of the proposed theories in everyday scenarios.

References

[1] Akyildiz, I. F., Lee, W. Y., Vuran, M. C. and Mohanty, S. (2006). NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey. Computer networks, 50(13), 2127–2159. doi: 10.1016/j.comnet.2006.05.001

[2] Ata, B. and Shneorson, S. (2006). Dynamic control of an M/M/1 service system with adjustable arrival and service rates. Management Science, 52(11), 1778–1791. doi: 10.1287/mnsc.1060.0587

[3] Bouchentouf, A. A., Guendouzi, A. and Majid, S. (2020). On impatience in Markovian M/M/1/N/DWV queue with vacation interruption. Croatian Operational Research Review, 11(1), 21–37. doi: 10.17535/crorr.2020.0003

[4] Chakka, R. (1995). Performance and reliability modelling of computing systems using spectral expansion. Doctoral dissertation, Newcastle University. Retrieved from: theses.ncl.ac.uk/jspui/handle/10443/2112

[5] Chakka, R. and Van Do, T. (2007). The MM $\Sigma_{k=1}^{K}$CPPk/GE/c/L G-queue with heterogeneous servers: Steady state solution and an application to performance evaluation. Performance Evaluation, 64(3), 191–209. doi: 10.1016/j.peva.2006.05.001

[6] Chan, C. W., Huang, M. and Sarhangian, V. (2021). Dynamic server assignment in multiclass queues with shifts, with applications to nurse staffing in emergency departments. Operations Research, 69(6), 1936–1959. doi: 10.1287/opre.2020.2050

[7] Chen, D. and Trivedi, K. S. (2005). Optimization for condition-based maintenance with semi-Markov decision process. Reliability engineering and system safety, 90(1), 25–29. doi: 10.1016/j.ress.2004.11.001

[8] Devi, K. L. and Valli, S. (2021). Multi-objective heuristics algorithm for dynamic resource scheduling in the cloud computing environment. The Journal of Supercomputing, 77(8), 8252–8280. doi: 10.1007/s11227-020-03606-2

[9] Diamantoulakis, P. D., Kapinas, V. M. and Karagiannidis, G. K. (2015). Big data analytics for dynamic energy management in smart grids. Big Data Research, 2(3), 94–101. doi: 10.1016/j.bdr.2015.03.003

[10] Elliriki, M., Reddy, C. S., Anand, K. and Saritha, S. (2022). Multi server queuing system with crashes and alternative repair strategies. Communications in Statistics-Theory and Methods, 51(23), 8173–8185. doi: 10.1080/03610926.2021.1889603

[11] Gandhi, A., Harchol-Balter, M., Das, R. and Lefurgy, C. (2009). Optimal power allocation in server farms. ACM SIGMETRICS Performance Evaluation Review, 37(1), 157–168. doi: 10.1145/2492101.1555368

[12] Harrison, J. M. and Zeevi, A. (2004). Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. Operations Research, 52(2), 243–257. doi: 10.1287/opre.1030.0084

[13] Mamatha, E., Sasritha, S. and Reddy, C. S. (2017). Expert system and heuristics algorithm for cloud resource scheduling. Romanian Statistical Review, 65(1), 3–18.

[14] Mamatha, E., Saritha, S., Reddy, C. S. and Rajadurai, P. (2020). Mathematical modelling and performance analysis of single server queuing system-eigenspectrum. International Journal of Mathematics in Operational Research, 16(4), 455–468. doi: 10.1504/IJMOR.2020.108408

[15] Moridi, E., Haghparast, M., Hosseinzadeh, M. and Jassbi, S. J. (2020). Fault management frameworks in wireless sensor networks: A survey. Computer communications, 155, 205–226. doi: 10.1016/j.comcom.2020.03.011

[16] Sadana, U., Chenreddy, A., Delage, E., Forel, A., Frejinger, E. and Vidal, T. (2024). A survey of contextual optimization methods for decision-making under uncertainty. European Journal of Operational Research. doi: 10.1016/j.ejor.2024.03.020

[17] Saritha, S., Mamatha, E. and Reddy, C. S. (2019). Performance Measures of Online Warehouse Service System with Replenishment Policy. Journal Europeen Des Systemes Automatises, 52(6), 631–638. doi: 10.18280/jesa.520611

[18] Saritha, S., Mamatha, E., Reddy, C. S. and Anand, K. (2019). A model for compound poisson process queuing system with batch arrivals and services. Journal Europeen des Systemes Automatises, 53(1), 81–86. doi: 10.18280/jesa.530110

[19] Saritha, S., Mamatha, E., Reddy, C. S. and Rajadurai, P. (2022). A model for overflow queuing network with two-station heterogeneous system. International Journal of Process Management and Benchmarking, 12(2), 147–158. doi: 10.1504/IJPMB.2022.121592

[20] Shirdel, G. H. and Abdolhosseinzadeh, M. (2016). The critical node problem in stochastic networks with discrete-time Markov chain. Croatian Operational Research Review, 7(1), 33–46. doi: 10.17535/crorr.2016.0003

[21] Sreelatha, V., Mamatha, E., Anand, S. K. and Reddy, N. H. (2022). Markov Process Based IoT Model for Road Traffic Prediction. In International Conference on Modeling, Simulation and Optimization, 329–338. doi: 10.1007/978-981-99-6866-4_24

[22] Sreelatha, V., Mamatha, E., Reddy, C. S. and Rajdurai, P. S. (2022). Spectrum relay performance of cognitive radio between users with random arrivals and departures. In Mobile Radio Communications and 5G Networks: Proceedings of Second MRCN 2021, 533–542. doi: 10.1007/978-981-16-7018-3_40

[23] Yang, T., Zhao, L., Li, W. and Zomaya, A. Y. (2021). Dynamic energy dispatch strategy for integrated energy system based on improved deep reinforcement learning. Energy, 235, 121377. doi: 10.1016/j.energy.2021.121377