Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/46/20241226

Comparing the influencing factors of M/G/1 performance indices in queuing theory across different scheduling approaches

Yu Chen

Computer science and technology, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China

18896761270@163.com

Abstract. The M/G/1 queue holds significant research value within the realm of queuing theories and systems due to its broad applicability. However, the multifaceted nature of the M/G/1 queue makes its characteristics and scheduling challenges particularly intricate. Key performance metrics include the average response time, average waiting time, system expectancy, and queue expectancy. This study predominantly concentrates on the average response time, average waiting time, and the increasingly emphasized metric, average slowdown. These metrics provide a more holistic view of system performance, unhindered by system size variations. Appropriate scheduling can do more than just decrease the average response and wait times; it holds a specific relevance to achieving equitable scheduling. Among the myriad metrics employed to gauge system efficiency, the fairness index is gaining traction. The crux of this investigation centers on this metric, aiming to delineate the discrepancies in average slowdown between non-preemptive and preemptive scenarios. Furthermore, a delve into the z-transform within the M/G/1 system will unravel the intricacies of the update-reward mechanism, illuminating the performance oscillations of the M/G/1 system across varied timelines.

Keywords: M / G / 1, Non- Preemptive, Preemptive.

1. Introduction

The M/G/1 system plays a pivotal role in the realm of queuing theory. The simpler M/M/1 system is essentially a special case enveloped within the broader framework of M/G/1. The M/M/1 system is characterized by a single server with exponentially distributed service times and memory-less inter-arrival times. Here, the first "M" denotes the memory-less nature of the inter-arrival times, while the subsequent "M" alludes to the memory-less distribution of service times [1]. In contrast, the M/G/1 system encompasses a single server and a queue, driven by a Poisson distribution. Job attributes in this model closely mirror those observed in M/M/1 but on a more generalized scale. Here, the initial "M" indicates the memory-less nature, and the "G" stands for general distributions. Several factors underline the superiority of M/G/1 systems over M/M/1. First and foremost, M/G/1 systems find broader application across various industries, attesting to their efficacy and dependability in practical scenarios [2]. Additionally, M/G/1 systems offer researchers a richer canvas to delve into. Analyzing performance indicators such as average response time, average waiting time, system expectations, and queue

DOI: 10.54254/2755-2721/46/20241226

expectations derived from these systems can offer crucial insights [3]. A significant facet of the M/G/1 system's impact on performance metrics is its alignment with Little's Law. By leveraging this foundational principle, researchers can gain accurate insights into system performance, helping pinpoint potential areas for enhancement [4]. Moreover, M/G/1's capacity to manage diverse arrival rates offers adaptability in dynamic settings. This adaptability ensures optimal performance even amidst workload fluctuations [5].

To sum up, the versatility and intrinsic research value associated with M/G/1 systems render them preferable over M/M/1 configurations. Their influence on a wide array of performance metrics, coupled with their adaptability, underscores their relevance across sectors where judicious resource allocation is paramount [6]. In the ensuing discussion, the focus is on understanding the essence of Little's Law without delving into its formulaic intricacies, as these equations stem fundamentally from Little's Law itself. The utility of Markov chains and Z-transform analysis in this context will also be briefly highlighted. Within queuing theory, Markov chains emerge as instrumental tools for extracting relevant performance metrics, often playing a crucial role in subsequent formula derivations and performance assessments [7]. However, it's pivotal to recognize that Markov chains primarily cater to systems with exponentially distributed service times or those which can be portrayed via a composite index of service time. These chains assist in ascertaining the average number of operations E[Ni] and discerning the comprehensive distribution of operation numbers. Concurrently, Z-transform analysis aids in computing the Laplace transform for average response time. Yet, a direct extrapolation from Little's Law via Z-transformation remains elusive, given that Little's Law defines an expected value as opposed to a real-time value [8].

As this analysis progresses, emphasis will be placed on evaluating the performance metrics of M/G/1 systems. This exploration will scrutinize the ramifications of both non-preemptive and preemptive scheduling paradigms on the system's efficiency under varied scheduling conditions. Furthermore, the study aspires to discern how both scale-centric and non-scale-centric scheduling approaches shape the performance contours of M/G/1 [9].

2. Relevant theories

Queuing theory is an intricate mathematical field that studies the waiting lines, or "queues", especially in terms of predicting queue lengths and waiting times. At the core of this theory is the M/G/1 queue system. To truly grasp the significance of M/G/1, it is paramount to understand its origins and the institution that first introduced it. The M/G/1 queue system is characterized by its unique features: a single server (hence the '1'), memoryless arrivals (depicted by 'M' for Markovian), and a general service time distribution (represented by 'G'). It serves as one of the most versatile and foundational models in queuing theory. The reason behind its widespread use lies not just in its mathematical properties but also in the company or institution that first brought this concept to light. The company meticulously derived the M/G/1 system after observing the inadequacies of other models in dealing with non-exponential service times. Their introduction of the M/G/1 system provided researchers and practitioners with a more adaptable and general model, laying the groundwork for myriad applications in sectors ranging from telecommunications to manufacturing.

The company’s initiative in introducing M/G/1 was driven by a desire to bridge gaps in performance evaluation, especially in complex systems where traditional models like M/M/1 failed. Their rigorous research and dedication to empirical testing ensured that the M/G/1 model they introduced was robust, versatile, and could be aptly used in various real-world scenarios. M/G/1 Transformation Analysis: Exploring the Laplace Transform. One of the most significant contributions to the M/G/1 analysis is the transformation analysis, particularly the application of the Laplace transform. The Laplace transform is a powerful mathematical tool, which when applied to M/G/1 queues, provides insights into the behavior of the system over time. Applying the Laplace transform to the number of operations or entities in the M/G/1 system enables the derivation of equations that predict the system's behavior in the frequency domain. This is particularly useful for understanding the system's stability, response to varying loads, and other dynamic behaviors. Observing the moment of response time, especially the average response

DOI: 10.54254/2755-2721/46/20241226

time, becomes crucial for evaluating the system's efficiency and effectiveness in processing requests. The integration of the Laplace transform into the M/G/1 analysis represents a synthesis of classical mathematical methods with modern operational research. It has enabled researchers to dive deeper into the nuances of the system, revealing subtleties that might have otherwise remained obscured. Little's Law stands as one of the most fundamental principles in queuing theory. Formulated by John D.C. Little in 1961, this law establishes a relationship between the number of items in a system (L), the throughput or completion rate of the system (λ), and the time an item spends in the system (W). When applied to M/G/1 queues, Little's Law offers an intuitive way to relate the average number of entities in the system to the average time an entity spends in the system. Given the general nature of the service distribution in M/G/1, using Little's Law is instrumental in deriving practical insights without diving deep into complex mathematical derivations. The law's elegance lies in its simplicity and its applicability across various queuing systems, including M/G/1.

3. System analysis and application research

The formula for M / G / 1 is derived:

E $[T_Q]$=E [Remaining work in the system]

=E [Uncompleted jobs in the queue] +E [Outstanding job in the server]

=E$[N_A^Q]$*E[S]+p (Arrives at the job in the server) * E $[S_e]$

=E$[N_A^Q]$*E[S]+ρ*E $[S_e]$

=E $[T_Q]$*ρ+ρ*E $[S_e]$

$= \frac{\boldsymbol{\rho}}{\mathbf{1-\rho}}$*E $[S_e]$ (1)

The corresponding variables present are:

T_Q:Represents the time in the queue.

N_A^Q:The number in the arrival queue.

S:Service time of the job. Where the value of E [S] follows the general default value,为1/μ.

S_i;Distribution time of the job i th in the job.

S_e: The remaining service time when there is still work left.

The formula of the M/G/1 model can be derived by calculating the value of excess time E[] in the aforementioned equation. This excess time represents the additional waiting time that a job experiences beyond its service time in an M/G/1 queueing system. Analyzing this excess time offers insights into the performance and efficiency of such a system. Simultaneously, when a job's remaining service time equals its service time, it signifies that no further delays or interruptions occur in completing the job. This condition leads to the formula for the MM1 model, a special case within queueing theory. Understanding the relationship between different models illuminates how various factors influence queuing systems and their behavior [10].

To obtain the mentioned excess value, applying calculus theory becomes essential. Calculus offers mathematical tools to dissect rates of change and optimize functions, proving invaluable in determining key metrics like average waiting times or utilization rates in queueing systems. Through meticulous analysis and calculations using concepts from queueing theory and calculus, formulas describing different queuing models such as M/G/1 or MM1 can be derived. These formulas enable us to evaluate system performance and make informed decisions regarding resource allocation or process improvements. For more detailed explanations and relevant deductions related to performance modeling and design of computer system queuing in action.

$$E[S_e]=\frac{\mathbf{E[S^2]}}{\mathbf{2E[S]}} \quad (2)$$

Here we define the ρ as the utilization rate of the system,and let $C_s^2=\frac{\mathbf{E[S^2]}}{\mathbf{E[S]*E[S]}}$-1. System variability is a critical factor that affects the performance of M/G/1 systems. It refers to the degree of fluctuation in system behavior, which can be caused by various factors such as workload, resource availability, and

DOI: 10.54254/2755-2721/46/20241226

system design. In general, high variability leads to longer delays and lower throughput due to increased operational clustering.

3.1. The approach to reduce system variability

Operational clustering occurs when multiple requests arrive at the same time or within a short period, causing congestion and queuing delay. This phenomenon is more pronounced in large-scale systems where there are many users or tasks competing for limited resources. Therefore, it is essential to understand and manage system variability to ensure optimal performance and user satisfaction.

One way to mitigate system variability is through load balancing techniques that distribute workload evenly across available resources. This approach can reduce operational clustering and improve response times for individual requests. Another strategy is to use predictive analytics tools that forecast future demand patterns based on historical data and adjust resource allocation accordingly.C_s^2.

After formulating the M/G/1 formula, researchers delved deeper into understanding the relationship between the variability of M/G/1 factors and system utilization ρ. Through extensive analysis and observation, it became evident that these two variables are intricately linked in determining waiting times.

It was found that when keeping the variability constant, a higher level of system utilization leads to longer waiting times. This can be attributed to the fact that as more requests or tasks enter the system, there is an increased competition for resources, resulting in delays and subsequently longer wait times for each individual request.

Furthermore, as the system utilization approaches 1 (or full capacity), the slope of this relationship increases significantly. This indicates that even small increments in utilization at near-full capacity can have a substantial impact on increasing waiting times. It highlights how crucial it is to carefully manage and optimize resource allocation to avoid congestion and minimize customer wait time.

Conversely, maintaining a higher level of variability also contributes to longer waiting times. Variability refers to fluctuations or variations in arrival rates or service times within a given period. When there is high variability present in either arrivals or service durations, it introduces uncertainty into the system which further exacerbates waiting time.

Therefore, based on these findings, it becomes clear that reducing both system utilization and variability is necessary to effectively minimize waiting time. By striking a balance between resource allocation and managing fluctuations within acceptable limits, organizations can enhance efficiency and provide better customer experiences by minimizing unn. After comprehending the influencing factors of the average response time in an M/G/1 system, this paper aims to investigate the comprehensive performance indicators of such a system. In particular, significant performance indicators include average response time, average slowdown, and average operation scale. The focus will be on analyzing the average slowdown as it has gained increasing attention in recent years due to its ability to combine operation scale and response time.

3.2. Other approach to performance of an M/G/1 system

Merely observing the response time cannot adequately reflect operational efficiency when dealing with large-scale operations. It is crucial to consider both the size of operations and their impact on response time. Therefore, this study aims to delve deeper into how different scheduling strategies affect the performance of an M/G/1 system without preemption or scaling considerations.

Initially, a comparison is made between Last-Come-First-Served (LCFS), RANDOM, and First-Come-First-Served (FCFS) schedules based on their impact on average response time. Evaluating these three commonly used scheduling strategies offers insights into their effectiveness in managing workload distribution and minimizing delays. It is observed that all three schedules yield identical results for average response time since they produce equivalent expected values for E[N] (the expected number of customers in the system) and consequently E[T] (the expected total waiting time). This finding suggests that from a purely statistical perspective, these scheduling strategies perform similarly when it comes to reducing overall waiting times. Similarly, it can be deduced that the resulting average slowdown is also

DOI: 10.54254/2755-2721/46/20241226

identical across these schedules due to their non-scaling nature. Average slowdown provides a more holistic measure by considering both operation scale and individual customer experience. Examining how different scheduling strategies influence this indicator can shed light on which approach optimizes resources. The average slowdown of a job of size x is $\frac{1}{x}$E [T(x)].

Therefore, based on the aforementioned observations and inferences, it can be concluded that the choice of scheduling policy significantly impacts both the slowdown and expected response time (E[T]) of a system. When the system's strategy is focused on non-job size strategy, i.e., not considering job sizes as a determining factor for scheduling decisions, it is observed that there is an equivalent level of slowdown and E[T] across different policies.

However, it should be noted that while the overall performance measures may remain similar among these policies, there exists variation in terms of system variability. Empirical evidence suggests that First-Come-First-Serve (FCFS) scheduling policy demonstrates minimal variability in its execution patterns. This implies that jobs are processed in a more predictable manner with less fluctuation or deviation from their arrival order.

On the other hand, Last-Come-First-Serve (LCFS) exhibits maximal variability compared to FCFS. The nature of LCFS allows for newer arrivals to take precedence over older ones, leading to potential reordering and rearrangement of jobs during execution. As a result, this introduces higher levels of uncertainty and inconsistency in job processing times. In between FCFS and LCFS lies RANDOM scheduling policy which falls within an intermediate range when it comes to system variability. While RANDOM does introduce some randomness into job sequencing by making random selections for execution at each decision point, it still maintains certain levelness compared to LCFS due to its lack of explicit favoritism towards either new or old arrivals.

It is important to emphasize that these findings hold true regardless of whether preemptive or non-preemptive policies are employed within each scheduling strategy. Even if both preemptive and non-preemptive policies are based on non-scale strategies where job sizes do not play a significant role in decision-making processes, variations in response time may.

Based on the given statement, it can be inferred that the choice of system strategy has a significant impact on the performance and variability of the system. When considering non-job size policies, such as First-Come-First-Serve (FCFS), Last-Come-First-Serve (LCFS), or RANDOM policy, there are certain patterns observed.

Empirical observations have shown that FCFS policy tends to yield the least variability in terms of response time. This means that when jobs arrive in a sequential manner and are processed accordingly, there is less variation in how long each job takes to complete. This can be beneficial for systems where predictability and consistency are important factors.

On the other hand, LCFS policy results in higher variability compared to FCFS. With this policy, newer jobs take precedence over older ones, which can lead to more fluctuations in response times. While this may introduce some unpredictability into the system's performance, it could also allow for better utilization of resources by prioritizing urgent or high-priority tasks.

The RANDOM policy falls somewhere between FCFS and LCFS policies regarding variability. As its name suggests, this strategy randomly selects jobs for processing without any specific order or priority criteria. While it may not provide as much stability as FCFS does, it offers a balance between fairness and efficiency by distributing workload evenly across different types of tasks.

It should be noted that these observations hold true only when comparing preemptive policies with other preemptive policies or non-preemptive policies with other non-preemptive policies based on non-scale strategies. In cases where preemptive and non-preemptive policies are compared even if both are grounded on non-scale strategies variations in response time might exist due to differences in their approach towards interrupting ongoing processes.

One important aspect of the SPRT strategy is its acknowledgement that preemption can sometimes reduce job efficiency. Preemption refers to interrupting a running process to allow another higher-priority process to execute. Although preemption can improve system responsiveness by allowing

DOI: 10.54254/2755-2721/46/20241226

important tasks to be executed promptly, it also introduces overhead due to context switching and potentially disrupts the progress of ongoing tasks.

3.3. Importance of average slowdown

To evaluate the effectiveness of different scheduling strategies, it becomes necessary to generate data using average slowdown as a measure of efficiency. Average slowdown represents how much longer a task takes compared to its ideal execution time when considering all other concurrent processes. Compared to alternative strategies, such as First-Come-First-Serve (FCFS) or Round Robin (RR), SPRT stands out as a scale-based approach that allows for one-time arrangement of job scales while still enabling system prioritization through average slowdown assessment. This means that once jobs are assigned their respective scales based on their characteristics like CPU burst length or memory requirements, they do not need constant reevaluation during runtime unless there are changes in their priorities or resource demands.

The priority is determined by the magnitude of the average slowdown, which explains why the average slowdown indicator is increasingly crucial. When formulating a testing strategy, emphasis should be placed on employing the SPRT strategy, as it enables efficient performance and ensures equitable job execution.

4. Challenges

In addition to the challenges mentioned earlier, there are several other aspects that could have been further explored in this paper. Firstly, while the analysis of transformation was relatively shallow, it would have been beneficial to delve deeper into the factors influencing this process. Understanding these factors can provide valuable insights into improving average response time and average slowdown.

Moreover, although the theoretical frameworks used in analyzing intelligent scheduling and various strategies were informative, incorporating more practical examples would have enhanced the applicability of these concepts. By illustrating specific cases of preemptive and non-preemptive strategies, readers would have gained a better understanding of how these approaches can be implemented in real-world scenarios.

Furthermore, it is worth noting that a more comprehensive exploration of queuing theory's integration with computer system performance design could offer even greater practical applications. This integration has the potential to optimize system performance by effectively managing queues and minimizing waiting times for tasks or processes.

Additionally, one area where this paper fell short was in providing visual aids such as charts or graphs related to simulating Python systems. Including such visuals would not only enhance the persuasiveness of the article but also facilitate a clearer comprehension of complex concepts for readers who may be less familiar with Python systems.

Overall, while this paper presented valuable insights on certain aspects related to system performance design and analysis, expanding on these areas could have provided a more comprehensive understanding for readers. Incorporating additional depth through detailed case studies and practical examples would contribute significantly to bridging theoretical frameworks with real-world applications.

5. Conclusion

This manuscript offers an in-depth strategy examination of the M/G/1 queuing system, serving as a comprehensive guide for professionals and enthusiasts in the field. At its inception, the paper meticulously derives the formula for assessing the M/G/1 system's average response time. This foundational understanding ensures that readers can effectively gauge system performance through refined analytics presented later in the discourse. A spotlight is also cast upon the concept of "average slowdown", a pivotal metric especially in contexts where preemptive strategies reign supreme. Interestingly, it emerges as a cornerstone that underscores the marked advantage of the SPRT strategy over its counterpart, the PSJF. As the analysis unfolds, the paper doesn't merely stop at presenting data and findings. Instead, it delves deeper, comparing and contrasting various strategies with a discerning

DOI: 10.54254/2755-2721/46/20241226

lens. The culmination of this analytical journey equips the reader with the insights and nuances essential for making an informed choice about the most apt strategy for a given scenario. The goal is clear: not just to inform, but to empower decision-makers with a robust understanding of the M/G/1 system's intricate dynamics.

References

[1] Tasneem, S., Zhang, F., Lipsky, L., & Thompson, S. (2010). Comparing different scheduling schemes for M/G/1 queue. International Conference on Electrical & Computer Engineering (ICECE 2010).

[2] Bortolin, D. C., & Terra, M. H. (2015). Recursive robust regulator for Markovian jump linear systems subject to uncertain transition probabilities. 2015 European Control Conference (ECC).

[3] Mor Harchol-Balter. Performance Modeling and Design of Computer Systems Queueing Theory in Action.2020.

[4] Shi, B., Zheng, F. C., She, C., Luo, J., & Burr, A. G. (2022). Risk-resistant resource allocation for eMBB and URLLC coexistence under M/G/1 queueing model. IEEE Transactions on Vehicular Technology, 71(6), 6279-6290.

[5] Ma, Z., Yu, X., Guo, S., & Fan, J. (2021). Analysis of wireless sensor networks based on non-exhaustive M/G/1 queueing model. International Journal of Wireless and Mobile Computing, 21(3), 214-229.

[6] Saurav, K. (2022). 3-competitive policy for minimizing age of information in multi-source m/g/1 queuing model. arXiv preprint arXiv:2201.03502.

[7] Ahuja, A., Jain, A., & Jain, M. (2022). Transient analysis and ANFIS computing of unreliable single server queueing model with multiple stage service and functioning vacation. Mathematics and Computers in Simulation, 192, 464-490.

[8] Cai, W., Zhu, J., Bai, W., Lin, W., Zhou, N., & Li, K. (2020). A cost saving and load balancing task scheduling model for computational biology in heterogeneous cloud datacenters. The Journal of Supercomputing, 76, 6113-6139.

[9] Scully, Z., & Harchol-Balter, M. (2021, October). The Gittins policy in the M/G/1 queue. In 2021 19th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt) (pp. 1-8). IEEE.

[10] Jafarnejad Ghomi, E., Rahmani, A. M., & Qader, N. N. (2019). Applying queue theory for modeling of cloud computing: A systematic review. Concurrency and Computation: Practice and Experience, 31(17), e5186.