Impact of Service Time Distributions and Server Utilization on Tandem Queueing System Performance

Marko Matulin(✉), Štefica Mrvelj, and Luka Čop

Faculty of Transport and Traffic Sciences, University of Zagreb, Vukeliceva 4, 10000 Zagreb, Croatia
{mmatulin,smrvelj,0135259750}@fpz.unizg.hr

Abstract. The paper explores the performance dynamics of a tandem queueing system (TQS) comprising two interconnected nodes and assesses its responsiveness under varying service conditions. By using simulation tools, we conducted a range of tests to discover the variability of waiting times in the queues and the system, for different server utilization, as well as for cases when the service times follow exponential and normal distributions. We discovered the interplay between the nodes, i.e., how the performance of one node affects the other within the same TQS. Namely, we showed how the waiting times in the queues are crucial in understanding the TQS dynamics. We also compared the simulation results with the well-established analytical models and discovered that their applicability is limited. Particularly noteworthy was the disparity observed in the performance estimation of the second node within the TQS across all simulations, indicating potential over- or underestimation by analytical models.

Keywords: queueing system · queueing theory · system performance · service time · waiting time · Markovian system · non-Markovian system

1 Introduction

In the realm of performance analysis for queueing systems, queuing theory plays a pivotal role, providing a mathematical framework for understanding and optimizing the efficiency of systems that involve waiting times and service processes [1]. This significance extends beyond isolated queueing systems to encompass interconnected networks of nodes. One particular type of interconnected queueing systems, known as tandem queueing systems (TQS), holds particular importance. Tandem queueing systems represent scenarios where users must pass through multiple service systems sequentially, each with its queue.

Internet nodes represent a pertinent example of TQS in the digital landscape. For instance, in an online shopping scenario, a customer initiates a purchase and must navigate through multiple servers sequentially. First, their request may be processed by a web server, then directed to a payment gateway, followed by order fulfillment systems, and

D. Perakovic and L. Knapcikova (Eds.): FABULOUS 2024, LNICST 596, pp. 67–79, 2025.
https://doi.org/10.1007/978-3-031-72393-3_6

finally, shipping and delivery logistics. Additionally, in cloud computing environments, users often interact with distributed systems composed of interconnected nodes. When accessing a cloud-based application, requests may pass through various servers responsible for authentication, data processing, and storage. In the realm of networked gaming, players connecting to multiplayer servers encounter TQS as they navigate through authentication servers, game servers, and matchmaking services. The sequential handling of player requests across these servers illustrates the system dynamics, where users progress through multiple service points, each potentially impacting the overall gaming experience. Transitioning to the quantum communication network example, where secure communication is facilitated through the exchange of quantum keys, each step in the process involves cryptographic operations performed by specialized quantum nodes. These nodes, equipped with quantum cryptographic algorithms, handle the generation, transmission, and verification of quantum keys, which are essential for establishing secure communication channels.

The complexities of TQS pose unique challenges for performance analysis. Unlike single-server queues, tandem systems involve a sequence of interconnected nodes, each with its own characteristics and performance metrics. Evaluating the performance of individual nodes within tandem systems presents considerable difficulty due to the interplay between successive service systems [2]. The behavior of one node can significantly influence the performance of subsequent nodes, creating dependencies that are challenging to model analytically.

Queuing theory, rooted in the study of queues, offers valuable insights into the behavior of queueing systems, including tandem configurations. However, the complexity of tandem systems often exceeds the capabilities of traditional analytical methods, necessitating the use of simulation tools for accurate assessments. By constructing a simulated network of Markovian and/or non-Markovian queueing systems tailored for performance evaluations, researchers can approximate the behavior of tandem configurations more effectively [3]. This approach enables the assessment of key performance metrics, including service time, waiting time, queue length, and server utilization, within the context of tandem systems.

The importance of such research extends beyond theoretical considerations, impacting the practical efficiency of provided services. By gaining a deeper understanding of how individual nodes in tandem systems respond to varying server utilization and interactions, researchers can optimize resource allocation, improve response times, and enhance the overall user experience. Hence, we focus on analyzing a specific TQS consisting of two interconnected nodes. In our model, we conduct tests to determine the impact of different server utilization and service time distributions on the performance of the overall system, as well as individual nodes. We analyzed the TQS performance (namely, the waiting times in the queues and the system) in cases when service times follow exponential and normal distributions, and for different server utilization levels.

The paper is structured as follows. We continue to Sect. 2 by drafting the related research within the field of queueing systems and TQS performance analysis. In Sect. 3 we describe our research methodology and testing environment. Results are presented and discussed in Sect. 4 while Sect. 5 brings our concluding remarks.

2 Related Work

In [4], a comprehensive analysis of various queue disciplines was conducted utilizing an analytical model of multi-service systems. Employing a C++ based simulation model, the authors carried out a series of five simulations, each comprising 1,000,000 entities. The outcomes revealed notable disparities between analytical calculations and simulation experiments, particularly contingent upon the server utilization, thereby underscoring the significance of simulation testing.

Likhttsinder et al. [5] introduced an interval method for queue analysis, focusing on the arrival of applications at constant time intervals and the resultant queue sizes. This method yielded a formula for computing the average queue length in a single server queuing system. Similarly, in [6], a formula for average queue length was derived for cases with constant processing time.

Traditionally, the assumption of independent interarrival and service times is made when solving waiting time in a queue system. However, [7] explored waiting times in the M/G/1 queuing system and concluded that correlated interarrival and serving times could potentially decrease waiting times. Meanwhile, Kyritsis et al. [8] attempted to analyze and predict waiting times using a machine-learning approach. While they utilized an open-access database to develop the prediction model, the lack of server deployment information in the database prevented verification using queuing theory.

Beyond single queue systems, several studies investigate systems comprising multiple interconnected nodes (queues), known as TQS. As previously discussed, in TQS, a customer traverses a series of nodes before exiting the system. Hence, performance analysis of individual nodes and the interplay between interconnected subsystems are crucial for understanding TQS behavior [2]. The interconnectedness of TQS can lead to delays propagating as customers move through the network. Moreover, inadequate control policies may result in the underutilization of downstream resources, a phenomenon addressed in [9], where authors examine arrival management in queuing networks to maintain a target occupancy level (queue length) within the system.

Srinivasa et al. [10] addressed the development of a tandem queueing model with non-homogeneous Poisson service processes, aiming to analyze system performance measures such as queue content, customer waiting time, throughput, and system variability. The sensitivity analysis revealed the significant influence of time-dependent service rates on system performance measures, demonstrating the model's accuracy in predicting performance under such conditions.

In [11] a unified model for TQS analysis was proposed, treating TQS as bidirectional cascade systems. Machine learning was applied by Kudou et al. in [12] to evaluate TQS performance, demonstrating significant computational time reduction compared to discrete simulations. The authors focused on a specific TQS scenario where customers could exit the system after being served by a single node, without traversing the entire TQS.

The Age of Information (AoI) metric is used in [13] to assess the freshness of information in multi-hop networks with tandem queues. The authors observed age escalation patterns under different arrival and departure processes, noting implications for optimal arrival rates. The AoI is also used by Koukoutsidis et al. [14]. The authors presented

various configurations of M/M/1 FIFO queues and highlighted age profile maintenance within networks handling different update packet categories.

Recently, the TQS paradigm was employed to develop a routing algorithm for Quantum Key Distribution networks (QKD) [15]. The Tandem Queue Decomposition (TQD) policy aims to achieve secure and capacity-achieving routing in QKD networks, catering to diverse traffic types including unicast, broadcast, and multicast. In [16] Stamatiou et al. explore the application of quantum devices in computer system control, particularly utilizing model predictive control and quantum annealing for the stabilization and management of tandem queue systems.

3 Methodology and Test Environment

The research focused on analyzing a TQS consisting of two nodes, as illustrated in Fig. 1. Both nodes of our TQS were equipped with FIFO queues of unlimited capacity, ensuring a lossless system. Upon exiting the queue at the first node, entities proceeded to service. Following service at the first node, entities entered the queue of the second node, where they awaited service again before exiting the TQS. The assumption was that there were no other entries into the second node. A single server configuration was set at both nodes.

At the entrance of the first node, we employed an entity generator to regulate the input flow of the TQS, allowing for the creation of various traffic scenarios. The assumption was that the input flow into the TQS follows a Poisson process, meaning that interarrival times of entities adhere to an exponential distribution. We also experimented with the service times at both Node 1 and Node 2. Specifically, these times were either exponentially or normally distributed. Importantly, during each simulation, the service times at both nodes followed the same distribution type and were kept equal. For each combination of server utilization and service time, we conducted five simulations, totaling 200 simulations, with a minimum of 3000 entities generated in each. Table 1 lists the simulation parameters.

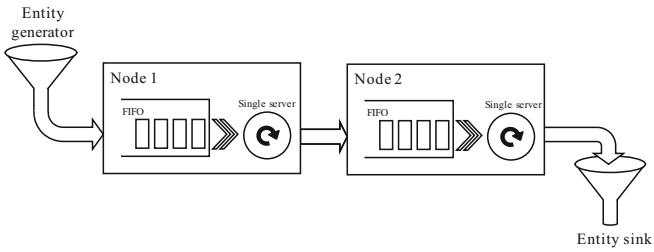

Fig. 1. A diagram describing the configuration of the TQS.

The simulation model was implemented using MATLAB, with parameters such as the number and generation time of entities, waiting time, and service time at both nodes tracked throughout. Based on these parameters, we calculated the total time an entity spends in TQS, and in each node, server utilization, and the queue state at both nodes.

The decision to vary the mean interarrival times and service times at both nodes aimed to simulate diverse operational scenarios representative of real-world systems. By generating entities multiple times and adjusting these parameters systematically, we were able to create a comprehensive set of test cases covering a wide range of server utilization and service demands. This approach enabled us to evaluate the TQS performance under various workload conditions, providing insights into system behavior across different operational regimes.

Table 1. Simulation parameters.

Service time distribution on both nodes	Server utilization
Exp. (mean = 1) or Norm. (mean = 1, st. dev. = 0.2)	0.1, 0.4, 0.7, 0.9
Exp. (mean = 3) or Norm. (mean = 3, st. dev. = 0.7)	0.1, 0.4, 0.7, 0.9
Exp. (mean = 5) or Norm. (mean = 5, st. dev. = 1.2)	0.1, 0.4, 0.7, 0.9
Exp. (mean = 7) or Norm. (mean = 7, st. dev. = 1.8)	0.1, 0.4, 0.7, 0.9
Exp. (mean = 9) or Norm. (mean = 9, st. dev. = 2.3)	0.1, 0.4, 0.7, 0.9

In addition to the simulation-based approach, we also conducted comparisons between simulation results and analytical results obtained using queueing theory expressions for calculating average waiting times in the queues (for the Markovian M/M/1 and non-Markovian M/G/1 models). This comparative analysis aimed to assess the degree to which analytical models accurately describe the system behavior. By applying established queueing theory formulas to our system configuration, we derived theoretical predictions for average waiting times in the queues at each node within the TQS. Discrepancies between simulated and analytical results provided insights into the limitations and applicability of analytical models in capturing the intricacies of TQS.

4 Results

4.1 The Average Waiting Times in the Queues

Figure 2 depicts the behavior of the average waiting time in the queues when service times follow an exponential distribution. Note that all variables related to time are measured in time units. As seen from the figure, the time spent in the queues (W_q) depends not only on the server utilization (ρ) but also on the service time distribution, i.e., the average service time denoted as T_s, consistent with the expression from [17]:

$$W_q = \frac{\rho \cdot T_s}{1 - \rho} \tag{1}$$

We can also observe that despite equal server utilization and identical service time distribution (exponential, with the same T_s), and interarrival time (t_a) distribution, the average time spent in the queues at Node 1 differs from Node 2. An example of t_a distributions at both nodes is shown in Fig. 3 for $\rho = 0.9$ and $T_s = 9$. All other simulation procedures yielded the same results.

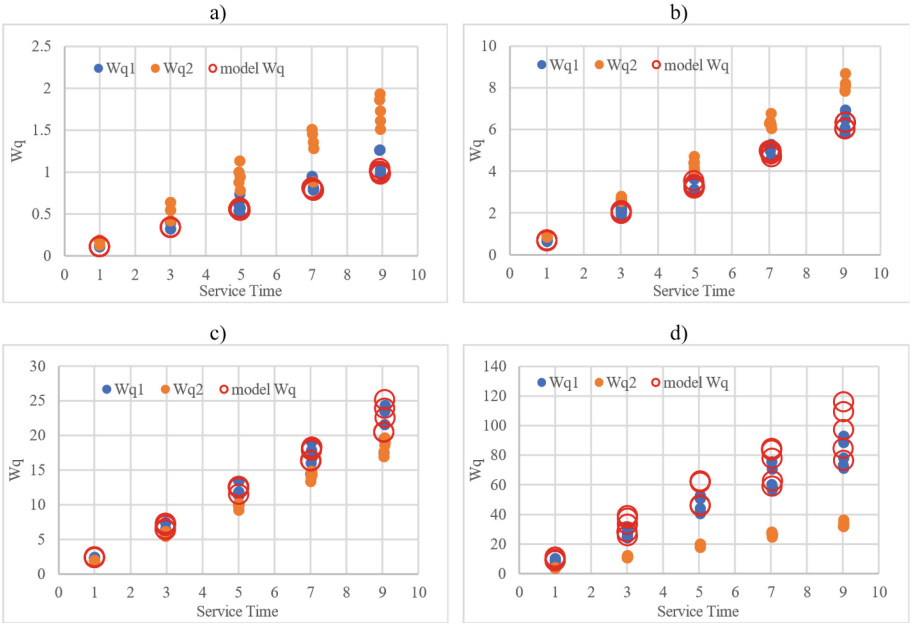

Fig. 2. Simulated average waiting times in Node 1 (Wq1) and Node 2 (Wq2), and modeled waiting time (model Wq) obtained using Eq. 1. The waiting times are correlated to service time where server utilization equals: a) 0.1, b) 0.4, c) 0.7, and d) 0.9.

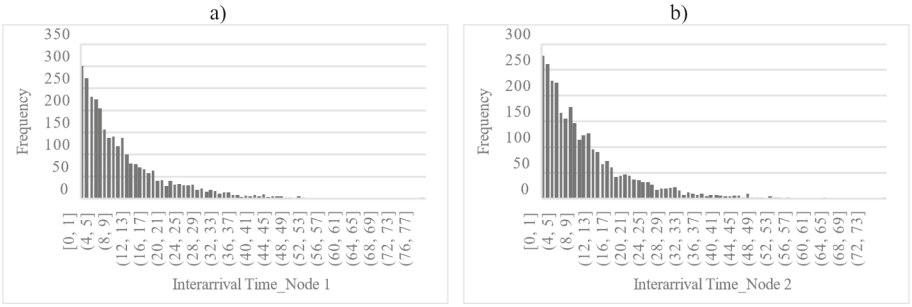

Fig. 3. The interarrival time distribution when $\rho = 0.9$ and $T_s = 9$ at: a) Node 1, b) Node 2.

For lower server utilization ($\rho < 0.4$), Node 1 exhibits shorter average queue waiting times compared to Node 2. Conversely, as server utilization increases, the trend reverses, with Node 2 experiencing shorter average waiting times. A notable contrast emerges in the queue waiting times between Node 1 and Node 2 at $\rho = 0.9$ while the average waiting time calculated using Eq. 1 closely aligns with the observed average waiting time at Node 1 (as indicated by the red circles). The largest deviation occurs at $\rho = 0.9$ and $T_s = 9$.

The investigation was extended to include scenarios where service times conform to a normal distribution. In this context, the dynamics of queue waiting times are contingent upon both server utilization and average service time. Notably, it is essential to emphasize

that the standard deviation remains consistent for a given T_s, irrespective of server utilization (refer to Table 1). In contrast to the prior analysis, under these conditions, the duration spent in the queue consistently surpasses that of Node 1 compared to Node 2. Consequently, the behavior of average queue waiting times is illustrated in Fig. 4, specifically for $\rho = 0.9$. Upon comparing waiting times under equivalent average service times, i.e., exponentially vs. normally distributed service times, it becomes evident that average queue waiting times are diminished when service times adhere to a normal distribution.

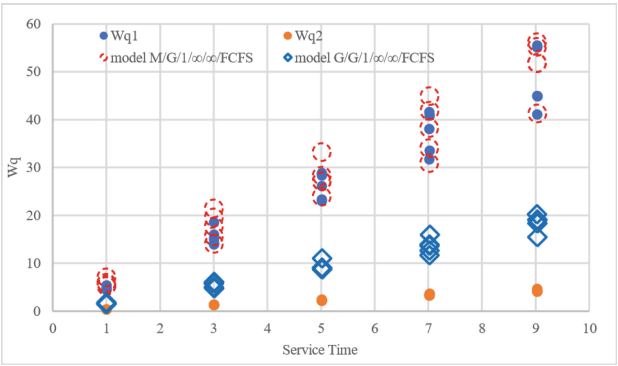

Fig. 4. Simulated average waiting times in Node 1 (Wq1) and Node 2 (Wq2), and modeled waiting time obtained using expression for M/G/ /∞/∞/ and G/G/ /∞/∞/ models [17]. The waiting times are correlated to service time. Server utilization equals 0.9.

4.2 Variations in the Waiting Times in the Queues

Given the potential limitations of relying solely on average values to gauge network capacities, a comprehensive analysis becomes imperative for a more precise evaluation of network performance. Consequently, Fig. 5 provides an insight into the analysis of maximum queue waiting times and standard deviation (SD) within scenarios characterized by service times conforming to an exponential distribution. Notably, the maximum values consistently surpass those of the first node, barring instances of lower ρ values (0.1). As the server utilization increases, the maximum waiting times at the first node become progressively higher, while they remain approximately the same at the second node.

For short average service time ($T_s = 1$), the standard deviation of the waiting time in the queue (σ_w) is approximately equal for both nodes regardless of the server utilization. In the cases of lower server utilizations, the standard deviations are comparable across both nodes, whereas, for higher server utilizations, the standard deviation is higher at the first node due to a greater dispersion of waiting time values.

The aforementioned highlights the inadequacy of relying solely on average waiting times in queues to comprehensively assess TQS performances. For example, when considering $\rho = 0.9$, $T_s = 9$, and service times following an exponential distribution, the

average waiting time at Node 2 is approximately 36 (Fig. 2d), yet it can escalate to 65 (as depicted in Fig. 5d). Node 1 exhibits even more substantial disparities. However, it is crucial to acknowledge that the elevated waiting times were only recorded for 10% of the generated entities. Figures 6a and b portray the 90th percentile of waiting times in queues, specifically for $\rho = 0.4$ and 0.9. Notably, at $\rho = 0.4$, a discernible contrast emerges compared to the maximum waiting time values. For instance, while the maximum waiting time values for $T_s = 9$ in the simulation reach up to 110 (as shown in Fig. 5b) for Node 1, surpassing those of Node 2, the 90th percentile stands at 24 (as depicted in Fig. 6a), which is lower than Node 2's 90th percentile. At $\rho = 0.9$, Node 2's 90th percentile (Fig. 6b) is marginally lower than the maximum values (Fig. 5d), yet notably inferior to the values recorded for Node 1.

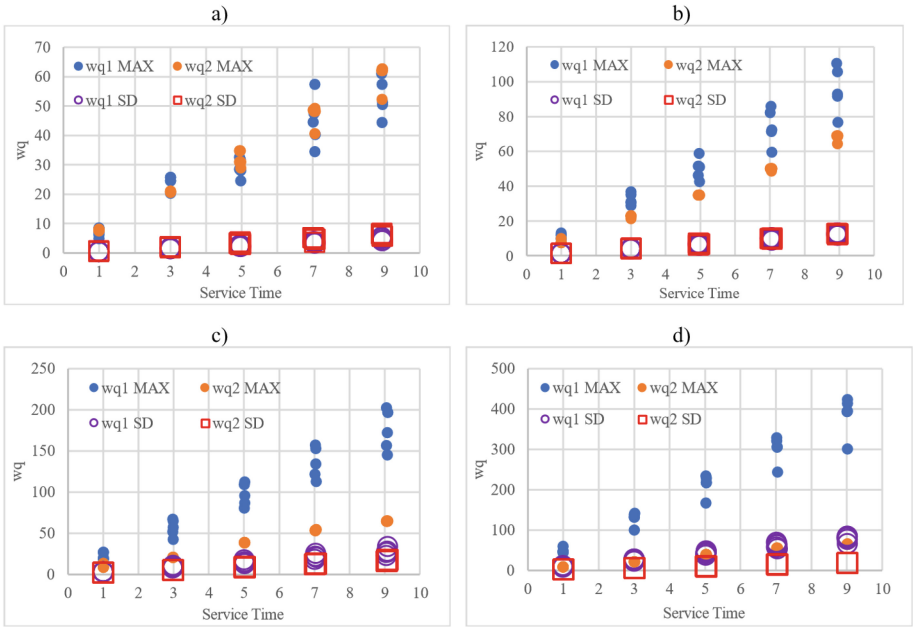

Fig. 5. Simulated maximum waiting times and corresponding SDs in Node 1 (wq1 MAX and wq1 SD) and Node 2 (wq2 MAX and wq2 SD). The waiting times are correlated to service time where server utilization equals: a) 0.1, b) 0.4, c) 0.7, and d) 0.9.

When comparing the metrics representing the extent of data dispersion derived from scenarios where service times adhere to an exponential distribution (Figs. 5d and 6b) with those of service times following a normal distribution (Figs. 7a and b), several conclusions can be inferred. Firstly, maximum waiting time values for the normal distribution of service times are notably lower compared to those observed for the exponential distribution of service times, holding constant T_s and ρ for both nodes. Secondly, the 90th percentile at Node 2 for the normal distribution of service times exhibits negligible deviation from the average waiting time value.

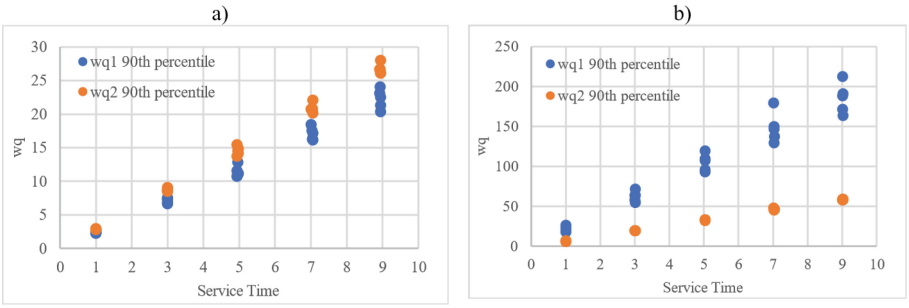

Fig. 6. The 90th percentile of waiting time correlated with the exponentially distributed service time when the server utilization equals: a) 0.4, b) 0.9.

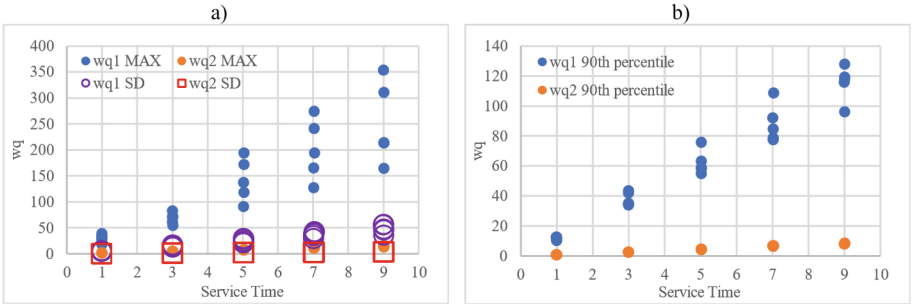

Fig. 7. The maximum waiting time values and SD values for normally distributed service time at both nodes for $\rho = 0.9$ (a), and the 90th percentile of the waiting time (b).

4.3 Interarrival Time Analysis

Besides server utilization and service time distribution, the behavior of waiting time in the queue can also be attributed to the interarrival times. When service times adhere to an exponential distribution, the interarrival times at the second node also follow suit, for the following reason. Under low server utilization, the intervals between entity arrivals at the second node mirror those at the first node, with minimal deviations in the standard deviations of interarrival times. Conversely, at high server utilization ($\rho = 0.9$), the intervals between arrivals at the second node approximate the service time at the first node (as depicted in Fig. 8a). Though minor disparities exist in the standard deviations of interarrival times at the first and second nodes (as illustrated in Fig. 8b).

A comparable scenario is evident in the remaining simulations. Nevertheless, across all conducted simulations, the ratios of standard deviation to the average, as delineated in Table 2, consistently lie within specific intervals. This suggests that the interarrival time at the second node adheres to an exponential distribution, aligning with the depiction in Fig. 3. It is worth noting that when the coefficient of variation, denoted as $V = \frac{\sigma_t}{t}\epsilon$ [0.84, 1.14], the data can be deemed to follow an exponential distribution [17, 18].

Upon analyzing the interarrival times at the second node in scenarios where service time conforms to a normal distribution, several observations emerge. For instances of

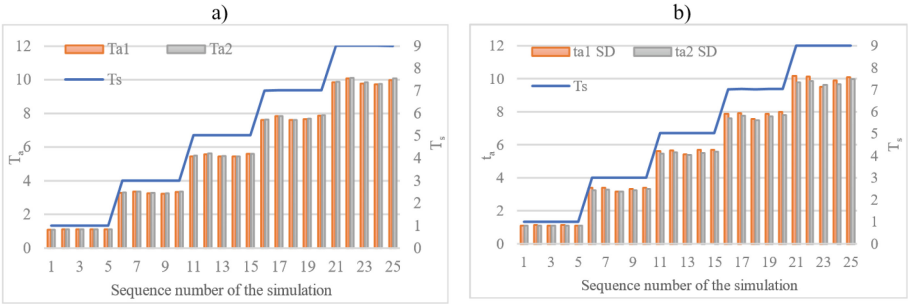

Fig. 8. The average interarrival time T_a (a) and standard deviation σ_a (b) at both nodes for $\rho = 0.9$.

Table 2. Coefficients of variation achieved in all simulations.

ρ	$min(\sigma_a/T_a)$	$min(\sigma_a/T_a)$
0.1	0.970	1.025
0.4	0.933	0.978
0.7	0.945	0.982
0.9	0.973	1.01

very low server utilization ($\rho = 0.1$), the interarrival times adopt an exponential distribution with equivalent average values to those at the first node. At moderate server utilization ($\rho = 0.4$), the interarrival times at the second node still exhibit characteristics of an exponential distribution (Fig. 9), given that the coefficient of variation (V) falls within the range of [0.84, 1.14].

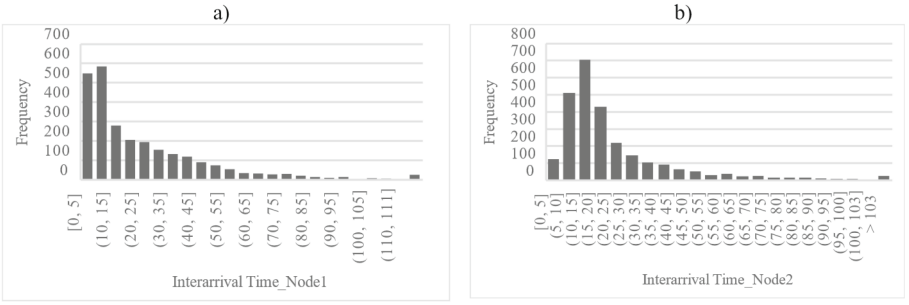

Fig. 9. Distribution of interarrival times at Node 1 (a) and Node 2 (b) for $\rho = 0.4$ and $T_s = 9$.

Nonetheless, disparities in the distributions of interarrival times between the two nodes amplify with increased service times (as depicted in Fig. 4), resulting in diminished waiting times at the second node, despite identical server utilization and interarrival times. Conversely, under heavy server utilization ($\rho = 0.9$), as previously mentioned,

the intervals between entity arrivals at the second node mirror a distribution akin to the service time distribution at the first node, as illustrated in Fig. 10 (first simulation for ρ = 0.9: $T_s = 9.03$, $\sigma_s = 2.32$, $T_{a1} = 9.79$, $\sigma_{a1} = 10.11$, $T_{a2} = 9.78$, $\sigma_{a2} = 5.03$).

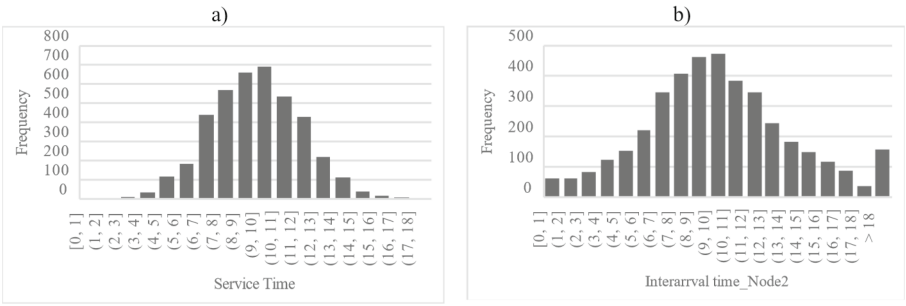

Fig. 10. Comparison of the distribution of service time and interarrival time at Node 2.

4.4 The Average Waiting Times in the System

Figure 11 illustrates the average waiting time in the system, derived from simulated data and the application of Jackson's formula [17, 19] for open networks. Notably, it becomes apparent that Jackson's formula inadequately estimates the average waiting time in the system under conditions of high server utilization. Simulation results indicate that this time is shorter in scenarios characterized by heightened server utilization.

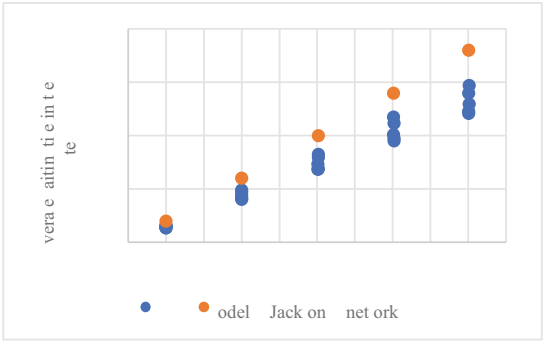

Fig. 11. The average waiting time in the system obtained from simulations (W) and Jackson's formula (model_Jackson's network).

Considering that different applications have varying sensitivities to network delays, especially to variations in delay, it is important to understand what happens at each node in the network. This understanding is crucial for implementing appropriate mechanisms to ensure satisfactory quality of service. However, the mentioned approach cannot be applied when the service time follows a normal distribution, as it does not meet the assumption that service times are exponentially distributed.

5 Conclusion

This study delved into the performance dynamics of a TQS with two interconnected nodes and scrutinized its responsiveness under diverse service conditions. Through extensive simulation tests, we unearthed the variance in average waiting times within the queues and the overall system across different levels of server utilization, as well as under exponential and normal service time distributions. Our analysis illuminated the interplay between nodes within the TQS, underscoring the significance of the waiting times in comprehending its dynamics. Furthermore, we compared simulation outcomes with established analytical models, revealing their limitations in accurately estimating system performance. Notably, our investigation highlighted significant disparities in performance estimations for the second node across all simulations, suggesting potential overestimation or underestimation by analytical models.

To enhance the relevance and applicability of our research, integration with realworld service systems is needed. Validating the developed TQS models in real-world scenarios by gathering data on actual traffic flows and service dynamics would strengthen our findings. Furthermore, our analyses relied on parameters such as waiting time, interarrival rate, and server utilization, commonly used in queueing theory to quantify system performance. Incorporating real-world data into our models would not only validate our findings but also enhance the accuracy of performance metrics in complex network environments.

This conclusion underscores the importance of integrating simulation tools and analytical models for a comprehensive understanding of intricate systems like TQS, while highlighting the necessity for further research to enhance the accuracy of performance estimations and develop advanced analytical frameworks reflecting real-world conditions in complex network environments. In our future work, we will continue examining different TQS configurations and their responsiveness to various input flows and server policies.

References

1. Vasilyev, S.A., Bouatta, M.A., Kanzitdinov, S.K., Tsareva, G.O.: Numerical analysis of shortest queue problem for time-scale queueing system with a small parameter. In: Dudin, A., Nazarov, A., and Moiseev, A. (eds.) Information Technologies and Mathematical Modelling. Queueing Theory and Applications, pp. 16–28. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-32990-6_2
2. Wu, K., Zhao, N.: Dependence among single queues in series. In: 2016 IEEE International Conference on Industrial Technology (ICIT), pp. 1347–1352 (2016). https://doi.org/10.1109/ICIT.2016.7474953
3. Klimenok, V., Dudin, A., Vishnevsky, V.: Tandem queueing system with correlated input and cross-traffic. In: Kwiecień, A., Gaj, P., Stera, P. (eds) Computer Networks. CN 2013. Communications in Computer and Information Science, vol 370. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38865-1_42
4. Hanczewski, S., Weissenberg, J.: The Impact of the adopted queue discipline on the accuracy of the analytical model in queuing systems with elastic and adaptive traffic. In: 2022 International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom), pp. 1–7 (2022). https://doi.org/10.1109/CoBCom55489.2022.9880653

5. Likhttsinder, B.Ya., Bakai, Yu.O.: Development of an interval method for queue analysis in queueing systems. In: 2022 Systems of Signals Generating and Processing in the Field of on Board Communications, pp. 1–4 (2022). https://doi.org/10.1109/IEEECONF53456.2022.9744357
6. Blatov, I., Likhttsinder, B., Kitaeva, E.: On estimates of average queue length for queueing systems in the case of correlated input flow. In: 2020 International Conference on Information Technology and Nanotechnology (ITNT), pp. 1–7 (2020). https://doi.org/10.1109/ITNT49337.2020.9253184
7. Kartashevskiy, I.: Waiting time analysis for the M/G/1 queuing system with correlated traffic. In: 2017 4th International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T), pp. 268–272 (2017). https://doi.org/10.1109/INFOCOMMST.2017.8246395.
8. Kyritsis, A.I., Deriaz, M.: A machine learning approach to waiting time prediction in queueing scenarios. In: 2019 Second International Conference on Artificial Intelligence for Industries (AI4I), pp. 17–21 (2019). https://doi.org/10.1109/AI4I46381.2019.00013
9. Badrinath, S., Balakrishnan, H.: Robust control of arrivals into a queuing network. IEEE Trans. Intell. Transp. Syst. **23**, 4474–4489 (2022). https://doi.org/10.1109/TITS.2020.3045030
10. Srinivasa Rao, K., Durga Aparajitha, J.: On two node tandem queueing model with time dependent service rates. Int. J. Syst. Assurance Eng. Manag. **10**, 19–34 (2019). https://doi.org/10.1007/s13198-018-0731-z
11. Li, L., Qian, Y., Yang, Y., Du, K.: A common model for the approximate analysis of tandem queueing systems with blocking. IEEE Trans. Automat. Contr. **61**, 1780–1793 (2016). https://doi.org/10.1109/TAC.2015.2478127
12. Kudou, T., Nii, S., Okuda, T.: A performance evaluation of tandem queueing systems by machine learning. In: 2022 IEEE International Conference on Consumer Electronics – Taiwan, pp. 389–390 (2022). https://doi.org/10.1109/ICCE-Taiwan55306.2022.9869030
13. Kam, C., Molnar, J.P., Kompella, S.: Age of information for queues in tandem. In: MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM), pp. 1–6 (2018). https://doi.org/10.1109/MILCOM.2018.8599728
14. Koukoutsidis, I.: Age of information in an overtake- free network of quasi – reversible queues. In: 2020 28th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), pp. 1–6 (2020). https://doi.org/10.1109/MASCOTS50786.2020.9285958.
15. B, V., Sinha, A.: Fast and secure routing algorithms for quantum key distribution networks. In: 2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), pp. 120–128 (2022). https://doi.org/10.1109/COMSNETS53615.2022.9668578
16. Stamatiou, G.T., Magoutis, K.: Quantum-enhanced control of a tandem queue system. In: Kalyvianaki, E. and Paolieri, M. (eds.) Performance Evaluation Methodologies and Tools. pp. 99–114. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-48885-6_7
17. Begovic, M.: Podvorbeni sustavi. Fakultet prometnih znanosti Sveucilišta u Zagrebu, Zagreb (2006)
18. Keilson, J.: Markov Chain Models — Rarity and Exponentiality. Springer New York, New York, NY (1979). https://doi.org/10.1007/978-1-4612-6200-8
19. Chen, H., Yao, D.D.: Fundamentals of Queueing Networks. Springer New York, New York, NY (2001). https://doi.org/10.1007/978-1-4757-5301-1