



Transient behaviour of time-varying tandem queueing networks

Anjale Ramesh¹ · M. Manoharan¹

Accepted: 15 May 2024

© The Author(s), under exclusive licence to Operational Research Society of India 2024

Abstract

Most of the large-scale service systems in real life are subject to time-varying conditions, such as arrival rates, service rates, and other factors that can affect system performance. These systems can be adequately modelled using time-varying queueing systems, where one or several parameters change over time. Analysis of transient behaviour in such time-varying queueing systems is more challenging than their steady-state analysis. This study deals with the transient analysis of Markovian queues connected in tandem, where both the service and arrival processes at each station depend on time. To begin with, we derive the transient distributional relationship between the average workload and the customer's waiting time in a single-server non-Markovian queue with time-varying arrival and service rates. We then generalise the transient laws for single-server queues to a k-station tandem network. Furthermore, we develop an algorithm for analysing transient performance measures in a k-station tandem queueing network and conduct a numerical study based on the algorithm. Numerical study supports the effectiveness of the algorithm, and the results provide insights into the transient behaviour of tandem networks, specifically in bottleneck scenarios. The study reveals that the location of the bottleneck station in a line has a significant impact on average workload in the stations.

Keywords Transient analysis · Tandem queues · Time-varying workload · Time-varying queues

✉ Anjale Ramesh
anjaleramesh@gmail.com

M. Manoharan
manumavila@gmail.com

¹ Department of Statistics, University of Calicut, Malappuram, Kerala 673635, India

1 Introduction

Time-varying queueing system has been widely used to model complex service systems such as communication [1–3], transportation [4–6], healthcare [7–9] etc. Steady-state analysis of time-varying general queueing systems in the literature mainly focuses on analytical approximation methods and simulation methods, since exact mathematical analysis of such systems is intractable. Transient analysis of general queueing models is considered challenging and becomes much more complex when we account for the arrival and/or service rates changing over time. Bertsimas and Mourtzinou [10] focused on transient version of Little's Law, which is one of the fundamental principles of queueing theory. Fralix and Riano [11] extended the time-varying generalisation of Little's Law and discussed its applications. Kim and Whitt [12] utilised time-varying Little's Law (TVLL) to estimate and reduce the average waiting time in a queueing system. Whitt [13] discussed stabilising performance of a single server queueing system with time-varying arrival rate.

Over the last few decades, tandem queueing network has attracted a great deal of attention due to its widespread practical applications in real world such as communication networks, manufacturing systems, tollbooths, supply chains, traffic flows etc. Recent research on applications of tandem queueing networks includes the works of [14–16] and others.

The transient behaviour of a single server two-station tandem network is investigated in [17] and transient analysis of k -station tandem queueing model with load dependent service rates is discussed in [18]. Zychlinski et.al. [19] analysed tandem networks with general time-varying arrival rate and blocking, using time-varying fluid models.

In this study, we primarily consider a single server queueing system, where the arrivals constitute a time-varying counting process with general distributional assumptions. We also assume that the waiting space for the queue is unlimited, the service is time-dependant and is provided in the order of arrival. In the first section, we derive a transient distributional law that relates the virtual workload in the system to the waiting times of customers. This formula is shown to subsume Brumelle's formula in [20] when we relax time-varying assumptions and go on with the stationary regime. In Sect. 3, we formulate transient performance measures such as, number of customers in the system and virtual workload, for a two station non-stationary Markovian tandem queueing network and we extend it to a k -station tandem queueing network in Sect. 4. In Sect. 5, we introduce an algorithm to obtain the transient performance measures discussed in previous section. In Sect. 6, we conduct a numerical study of three station tandem network model and analyse the transient behaviour of its performance measures.

2 Transient performance measures for single server queues

Consider a single server $G_t/G_t/1$ queueing system with general time-varying arrival rate $\lambda(t)$ and service rate $\mu(t)$ and with First Come First Serve (FCFS) service discipline.

2.1 Number of customers

Let $L(t)$ be the number of customers in the system at time $t, t \geq 0$ and let $W(s)$ be the waiting time of the customer who arrive at time $s, 0 \leq s \leq t$. Then the time-varying number of customers in the system or transient generalisation of Little’s Law in [10] is,

$$E(L(t)) = \int_0^t F_s^c(t - s)\lambda(s)ds \tag{1}$$

where $F_t(x) = P\{W(t) \leq x|A_t\}, x \geq 0$ is the cumulative distribution function of the waiting time for a customer who arrive at time t and $F_t^c(x) = P\{W(t) > x|A_t\}, x \geq 0. A_t = \{A(t), t \geq 0\}$ is the arrival process to the system.

2.2 Virtual workload

Virtual workload process $Z = \{Z(t), t \geq 0\}$ is the amount of work remaining in the system at time t . The following theorem gives an explicit integral formula for virtual workload in a single-server queueing system.

Theorem 1 *Let $Z(t)$ be the virtual workload in the system at time t and $V(s)$ be the service requirement of the customer who arrive at time s . Then expected virtual workload at time t is,*

$$E(Z(t)) = \int_0^t F_s^c(t - s)\rho(s)ds + \int_0^t \frac{E(V(s)^2)}{2} \lambda(s)ds \tag{2}$$

Proof Consider the interval $[0, t]$. We start with a reverse-time construction. Let $T_{-k}(t)$ be the k^{th} arrival before time t . i.e., $T_{-(k+1)}(t) < T_{-k}(t) \leq t, \forall n \geq k \geq 1$. Let $W_{-k}(t) = W(T_{-k}(t))$ be the waiting time of the customer who arrive at time $T_{-k}(t)$. Then the workload in the system at time t can be expressed as,

$$Z(t) = \sum_{k=1}^{\infty} \left(I_{\{W(T_{-k}(t)) \geq t - T_{-k}(t)\}} V(T_{-k}(t)) + \frac{V(T_{-k}(t))^2}{2} \right) \tag{3}$$

where $I_{\{W(T_{-k}(t)) \geq t - T_{-k}(t)\}} = \begin{cases} 1 & \text{if } W(T_{-k}(t)) \geq t - T_{-k}(t) \\ 0 & \text{otherwise} \end{cases}$. The first term in (3) gives the service time of the customer who arrive at $T_{-k}(t)$ and waiting for service at time t . $\frac{V(T_{-k}(t))^2}{2}$ is the remaining service time of the customer in the server who arrive at $T_{-k}(t)$. If the server is idle then this term becomes zero.

(3) can be written as,

$$Z(t) = \int_0^t \left(I_{\{W(s) > t - s\}} V(s) + \frac{V(s)^2}{2} \right) dA(s)$$

$$= \int_0^t (I_{\{W(s) > t-s\}} V(s)) dA(s) + \int_0^t \frac{V(s)^2}{2} dA(s) \tag{4}$$

By using Campbell–Mecke formula in [11] for taking expectations of stochastic integrals, we get

$$E(Z(t)) = \int_0^t P\{W(s) > t-s\} E(V(s)) \lambda(s) ds + \int_0^t \frac{E(V(s)^2)}{2} \lambda(s) ds \tag{5}$$

For $F_s^c(x) = P\{W(s) > x|A_s\}$, $x \geq 0$ and $\rho(s) = \frac{\lambda(s)}{\mu(s)}$

$$E(Z(t)) = \int_0^t F_s^c(t-s) \rho(s) ds + \int_0^t \frac{E(V(s)^2)}{2} \lambda(s) ds$$

By the definition of squared coefficient of variation of service time, this can also be written as

$$\begin{aligned} E(Z(t)) &= \int_0^t F_s^c(t-s) \rho(s) ds + \int_0^t \frac{c_s^2 + 1}{2} \frac{\lambda(s)}{\mu(s)^2} ds \\ &= \int_0^t F_s^c(t-s) \rho(s) ds + \int_0^t \frac{c_s^2 + 1}{2} \frac{\rho(s)}{\mu(s)} ds \end{aligned}$$

□

Remark 1 If we relax the time-varying conditions, we can see that the well-known Brumelle’s formula in [20] immediately follows from this result. i.e., If we assume that $\tilde{Z} = \{\tilde{Z}(t); t \geq 0\}$ is stationary, then

$$\begin{aligned} E(\tilde{Z}(0)) &= \int_{-\infty}^0 P_s(\tilde{W}(s) > -s) \rho ds + \int_{-\infty}^0 \frac{c_s^2 + 1}{2\mu} \rho ds \\ &= \rho \int_{-\infty}^0 P_0(\tilde{W}(0) > -s) ds + \int_{-\infty}^0 \frac{c_s^2 + 1}{2\mu} \rho ds \\ E(\tilde{Z}(0)) &= \rho E(\tilde{W}(0)) + \rho \frac{c_s^2 + 1}{2\mu} \end{aligned}$$

3 Transient performance measures for tandem queueing network of two stations

In this section, we develop transient performance measures for two-station tandem queueing network. We model a queueing network with two stations in tandem and unlimited waiting capacity at each queue, as illustrated in Fig. 1. Arrival and service processes in each queue in the system are time-dependant and we restrict the distributional assumptions associated with the system to Markovian($M_t/M_t/1$) supposition.

The model is characterised by the following parameters;

1. Arrival process $A_i = \{A_i(t), t \geq 0\}$, $A_i(t)$ represents number of arrivals at time t with $E(A_i(t)) = \lambda_i(t)$; where $\lambda_i(t)$ is the arrival rate of station i at time t , for $i = 1, 2$.
2. Service process $V_i = \{V_i(t), t \geq 0\}$, $V_i(t)$ is the service requirement of a customer who arrive at time t with service rate $\mu_i(t)$ (mean= $1/\mu_i(t)$), $i = 1, 2$.
3. Transition from station 1 to station 2 occurs with probability p , $0 \leq p \leq 1$ and customer departs from station 1 with probability $1 - p$.
4. Number of servers, $N_i = 1, i = 1, 2$.
5. $W_i(t)$ is the waiting time of a customer who arrive at time $t, i = 1, 2$.
6. $Z_i(t)$ is the virtual workload in the i^{th} station at time t .
7. The counting process $L_i = \{L_i(t), t \geq 0\}$, the number of customers present in station i waiting for service, $i = 1, 2$.
8. Departure process $D_i = \{D_i(t), t \geq 0\}$, $D_1(t)$ denote departure of customers from station 1 to station 2, therefore $D_1(t)$ is merely the pattern of arrivals to the station 2($A_2(t)$). $D_2(t)$ denote the departures from station 2 and $D_3(t)$ denote the departures from station 1 to out of the network.

$$\begin{aligned}
 L_1(t) &= \int_0^t I_{\{W_1(s) > t-s\}} dA_1(s) \\
 E(L_1(t)) &= \int_0^t (P\{W_1(s) > t-s\}) \lambda_1(s) ds \\
 &= \int_0^t F_1^c(t-s) \lambda_1(s) ds
 \end{aligned}
 \tag{6}$$

$$\begin{aligned}
 L_2(t) &= \int_0^t I_{\{W_2(s) > t-s\}} dD_1(s) \\
 E(L_2(t)) &= \int_0^t (P\{W_2(s) > t-s\}) \lambda_2(s) ds \\
 &= \int_0^t F_2^c(t-s) p \mu_1(s) ds
 \end{aligned}
 \tag{7}$$

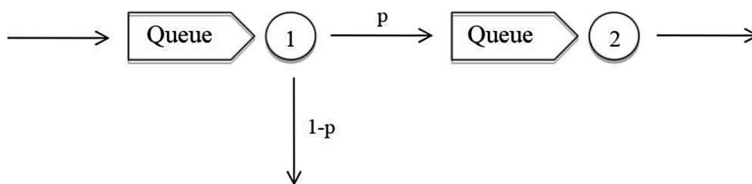


Fig. 1 A two station tandem queueing network model

where $I_{\{W_i(s) > t-s\}}$ represents the number of customers entered at station i at time s and waiting for service at time t , $0 \leq s \leq t$ and $F_i^c(t-s) = 1 - F_i(t-s) = P(W_i(s) > t-s)$. In equation (7) arrival rate to the station 2, $\lambda_2(t)$ is replaced by $p \mu_1(s)$.

$$\begin{aligned}
 Z_1(t) &= \int_0^t I_{\{W_1(s) > t-s\}} V_1(s) dA_1(s) + \int_0^t \frac{V_1(s)^2}{2} dA_1(s) \\
 E(Z_1(t)) &= \int_0^t F_1^c(t-s) \frac{\lambda_1(s)}{\mu_1(s)} ds + \int_0^t \frac{c_1^2 + 1}{2} \frac{\lambda_1(s)}{\mu_1(s)^2} ds
 \end{aligned}
 \tag{8}$$

Similarly, for the second station,

$$\begin{aligned}
 Z_2(t) &= \int_0^t I_{\{W_2(s) > t-s\}} V_2(s) dD_1(s) + \int_0^t \frac{V_2(s)^2}{2} dD_1(s) \\
 E(Z_2(t)) &= \int_0^t F_2^c(t-s) \frac{\lambda_2(s)}{\mu_2(s)} ds + \int_0^t \frac{c_2^2 + 1}{2} \frac{\lambda_2(s)}{\mu_2(s)^2} ds \\
 &= p \left(\int_0^t F_2^c(t-s) \frac{\mu_1(s)}{\mu_2(s)} ds + \int_0^t \frac{c_2^2 + 1}{2\mu_2(s)} \frac{\mu_1(s)}{\mu_2(s)} ds \right)
 \end{aligned}
 \tag{9}$$

where $c_i^2, i = 1, 2$ is the coefficient of variation of service process in station i .

4 A tandem network of k stations

We extend the two station non-stationary Markovian tandem model in section 3 to a k station tandem network model, as illustrated in Fig. 2. The transition probability $p_{i,i+1}$ denotes the probability that a customer transfer from station i to station $i + 1$ after service, $i = 1, 2, 3, \dots, k - 1$. Let $A_i = \{A_i(t), t \geq 0\}$ and $D_i = \{D_i(t), t \geq 0\}$, $i = 1, 2, 3, \dots, 2k - 1$ be the arrival and departure processes respectively. Here $D_i(t)$ for $i = 1, 3, 5, \dots, 2k - 3$ are departure processes as well as arrival processes and $D_i(t)$ for $i = 2, 4, 6, \dots, 2k - 2$ and $2k - 1$ are departures from the network.

$$\begin{aligned}
 E(L_1(t)) &= \int_0^t F_1^c(t-s) \lambda(s) ds \\
 E(L_i(t)) &= \int_0^t F_i^c(t-s) p_{i-1,i} \mu_{i-1}(s) ds \\
 E(L_k(t)) &= \int_0^t F_k^c(t-s) p_{k-1,k} \mu_{k-1}(s) ds
 \end{aligned}
 \tag{10}$$

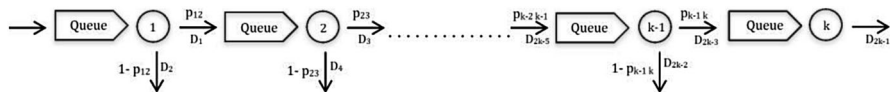


Fig. 2 A k station tandem queueing network model

$$\begin{aligned}
 E(Z_1(t)) &= \int_0^t F_1^c(t-s) \frac{\lambda_1(s)}{\mu_1(s)} ds + \int_0^t \frac{c_1^2 + 1}{2} \frac{\lambda_1(s)}{\mu_1(s)^2} ds \\
 E(Z_i(t)) &= p_{i-1,i} \left(\int_0^t F_i^c(t-s) \frac{\mu_{i-1}(s)}{\mu_i(s)} ds + \int_0^t \frac{c_i^2 + 1}{2\mu_i(s)} \frac{\mu_{i-1}(s)}{\mu_i(s)} ds \right) \quad (11) \\
 E(Z_k(t)) &= p_{k-1,k} \left(\int_0^t F_k^c(t-s) \frac{\mu_{k-1}(s)}{\mu_k(s)} ds + \int_0^t \frac{c_k^2 + 1}{2\mu_k(s)} \frac{\mu_{i-1}(s)}{\mu_i(s)} ds \right)
 \end{aligned}$$

where c_i^2 $i = 1, 2, 3, \dots, k$ is the coefficient of variation of service process in station i

Remark 2 The set of equations (10) and (11) derived in this section can be easily generalised to k-station tandem queueing network with non-stationary non-Markovian ($G_t/G_t/1$) queues by proceeding on similar lines as outlined above.

5 Algorithm for k station tandem model

In this section, we propose an algorithm to obtain transient performance measures such as number of customers and average virtual workload for a k-station tandem network model with non-stationary Markovian queues. Here we consider some prerequisites for developing the algorithm.

- We utilise the principle of rate-matching control, discussed in [13], for choosing the service rate function. In rate-matching control, we choose the service rate to be proportional to the arrival rate, for fixed traffic intensity ρ . i.e., for a given traffic intensity ρ_i , the service rate becomes,

$$\mu_i(t) \equiv \lambda_i(t)/\rho_i, \quad i = 1, 2, 3, \dots, k \quad t \geq 0 \quad (12)$$

- For an $M_t/M_t/1$ model, distribution of the waiting time $W(u)$, i.e., the probability that the waiting time of a customer who arrives at u , is larger than x is,

$$P(W(u) > x) = \rho e^{-(1-\rho)\Lambda_t(u)/\rho} \quad (13)$$

where $\Lambda_t(u) = \Lambda(t+u) - \Lambda(u)$, $\Lambda(\cdot)$ is the cumulative arrival rate function defined as,

$$\Lambda(u) = \int_0^u \lambda(r)dr, \quad r \geq 0$$

Table 1 Four cases of traffic intensities

Case	ρ_1	ρ_2	ρ_3
A	0.45	0.60	0.90
B	0.60	0.90	0.45
C	0.90	0.60	0.45
D	0.90	0.60	0.87

and $\Lambda_t(u)$ need to be strictly increasing and continuous, see [13]. Therefore, the expected waiting time, $E(W(t))$ in the $M_t/M_t/1$ model is,

$$E(W(t)) = \int_0^\infty P(W(u) > x) = \rho \int_0^\infty e^{-(1-\rho)\Lambda_t(u)/\rho} \tag{14}$$

- Probability that the waiting time of a customer who arrives at time s will be greater than $t - s$ is,

$$P(W(s) > t - s) = \rho e^{-(1-\rho)\Lambda_t(s)/\rho}$$

where $\Lambda_t(s) = \Lambda(t) - \Lambda(s)$. Here for station i , $\Lambda_{t,i}(s) = \Lambda_i(t) - \Lambda_i(s)$ and

$$P(W_i(s) > t - s) = \rho_i e^{-((1-\rho_i)\Lambda_{t,i}(s))/\rho_i} \quad i = 1, 2, 3, \dots, k$$

- The term squared coefficient of variation of service time(c^2) is involved in (11). Since we considered non-stationary Markovian queueing model, c^2 can be assumed to be 1.

5.1 Algorithm

A general framework of the algorithm for obtaining transient performance measures in a k -station tandem network model is presented in this section.

Algorithm 1

Require: External arrival rate $\lambda_1(t)$, transition probabilities $p_{i,j}$, $i = 1, 2, 3, \dots, k - 1$, $j = 2, 3, \dots, k$ and traffic intensity, $\rho = \{\rho_1, \rho_2, \dots, \rho_k\}$

Ensure: Transient performance measures, $E(L_i(t))$ and $E(Z_i(t))$ for $i = 1, 2, 3, \dots, k$.

- 1: Compute $\lambda_i(u) = p_{i-1,i} \frac{\lambda_{i-1}(u)}{\rho_{i-1}} \quad i = 1, 2, 3, \dots, k$
- 2: Compute $\mu_i(u) = \frac{\lambda_i(u)}{\rho_i} \quad i = 1, 2, 3, \dots, k$
- 3: Compute $\Lambda_i(t, s) = \int_0^t \lambda_i(u) du - \int_0^s \lambda_i(u) du. \quad i = 1, 2, 3, \dots, k$
- 4: Compute $P(W_i(s) > t - s) = \rho_i e^{-((1-\rho_i)\Lambda_{t,i}(s))/\rho_i} \quad i = 1, 2, 3, \dots, k$
- 5: Apply the formulas in (10) and (11) to obtain expected number of customers and virtual workload at each station.

6 Numerical study

In this section, we discuss two examples of tandem networks. We utilise the algorithm developed in previous section to compute performance measures and study the transient behaviour of the models. Specifically, we analyse the effects of traffic intensity on the transient performance measures.

6.1 A three-station example

Here we consider a three-station tandem network of non-stationary Markovian $M_t/M_t/1$ queues with following characteristics;

- Let the external arrival rate to station 1, $\lambda_1(t)$ be the identity function, $t, t \geq 0$.

$$\lambda_i(t) = p_{i-1,i}\mu_{i-1}(t), \quad i = 2, 3, \tag{15}$$

where $\mu_i(t)$, $i = 1, 2, 3$ is the service rate function obtained from (12) and $p_{i-1,i}$ is the transition probability from station i to $i + 1$, $i = 1, 2$. Here we take $p_{i-1,i} = p = 0.75$, ie., equal probability for all transitions.

- In this study, we consider four cases, A, B, C and D by taking arbitrary values for traffic intensities, as shown in Table 1. Stations with traffic intensities close to 1 are considered bottleneck stations. Case D is more challenging because of two bottleneck stations, whereas all the other cases have only one bottleneck station.

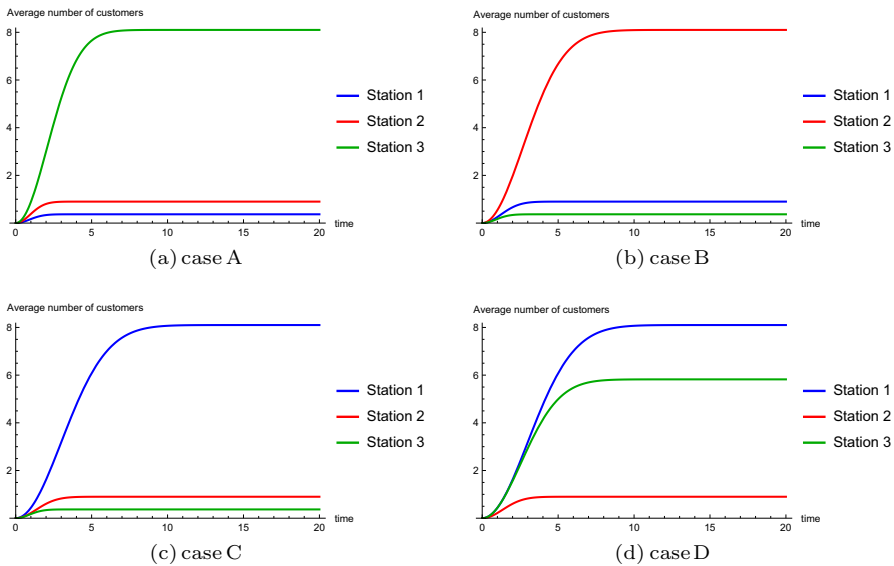


Fig. 3 Average number of customers at time t for four different cases considered in this study

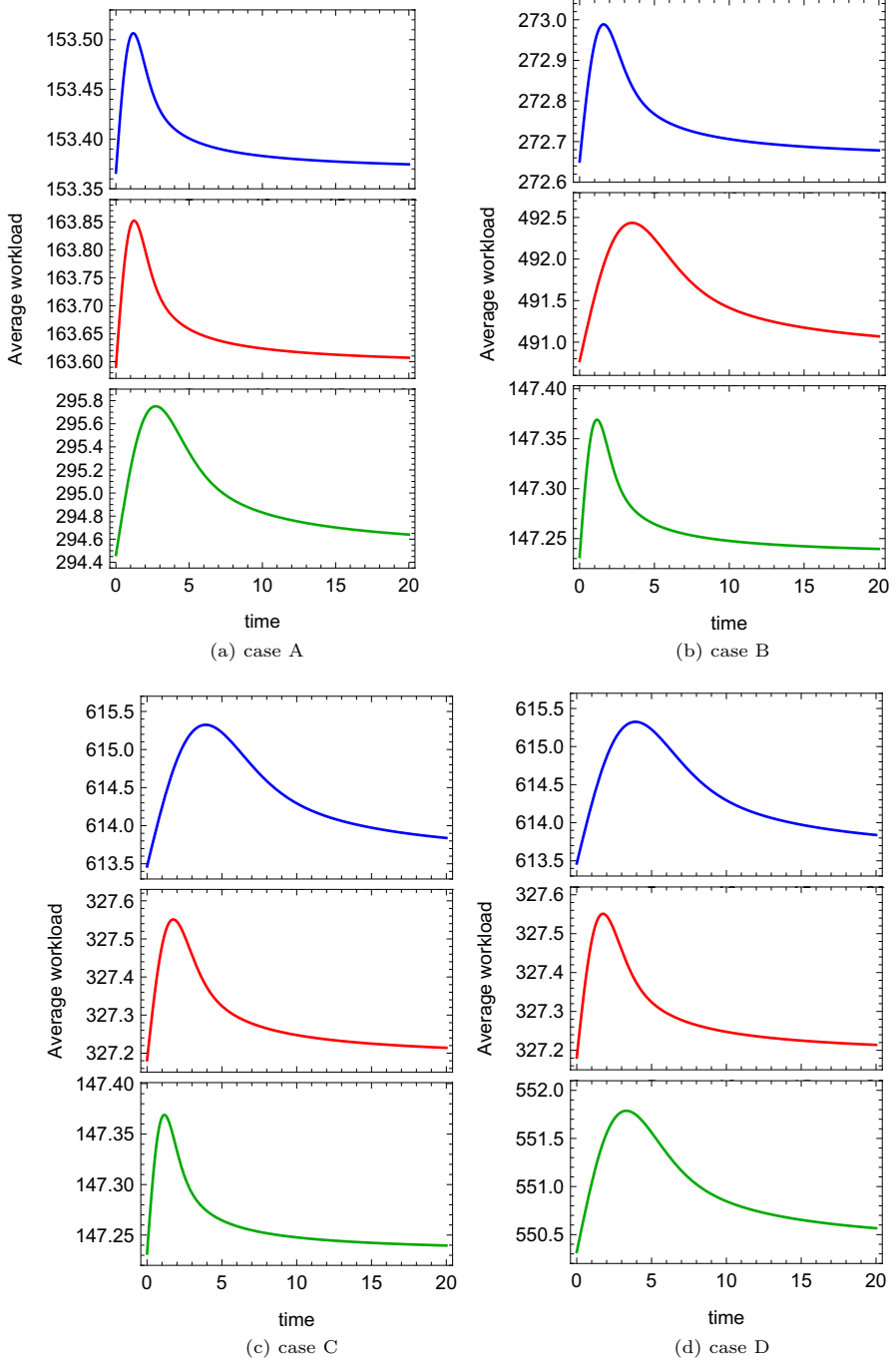


Fig. 4 Average workload at time t for different cases. Here blue, red and green figures represent station 1, station 2 and station 3 respectively

Table 2 Four cases of traffic intensities

case	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
A	0.45	0.50	0.75	0.60	0.90
B	0.60	0.45	0.90	0.50	0.75
C	0.90	0.75	0.60	0.50	0.45
D	0.90	0.45	0.85	0.60	0.90

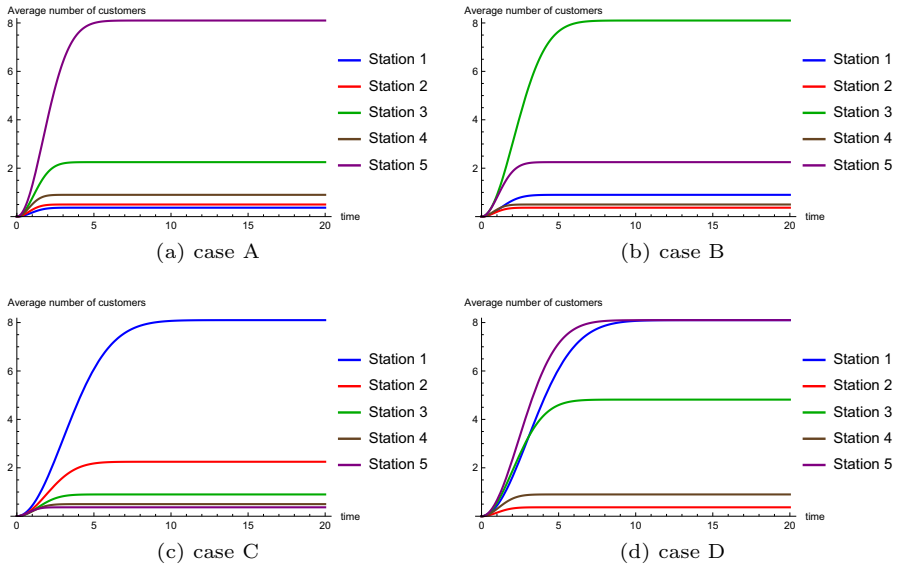


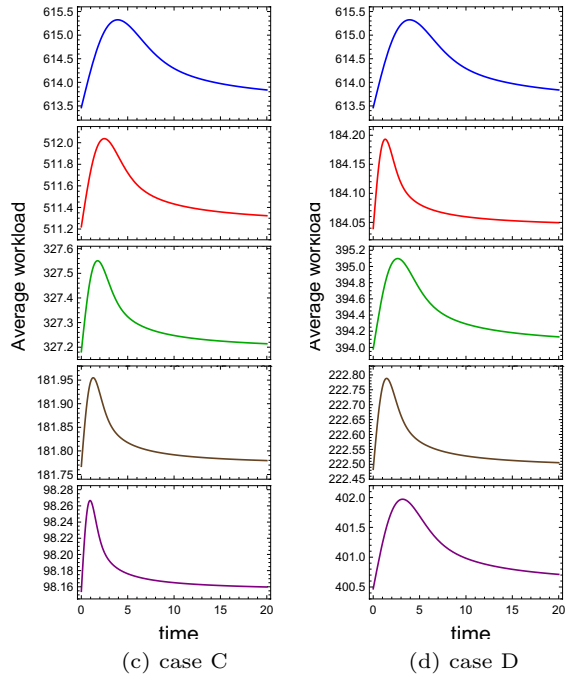
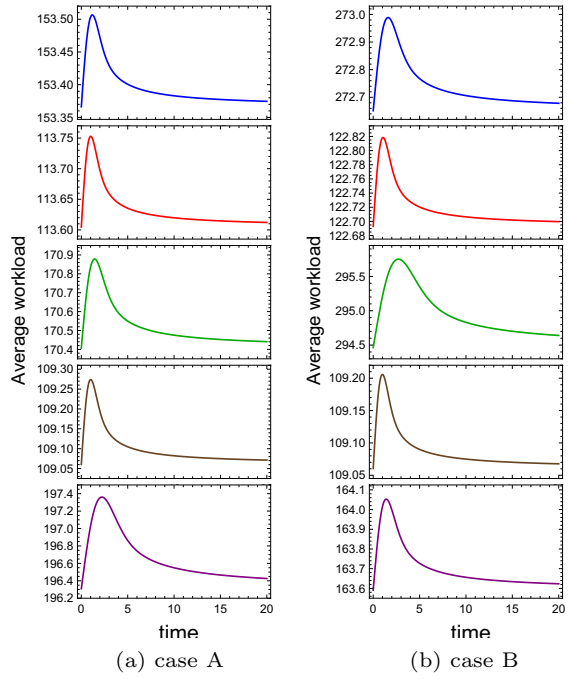
Fig. 5 Average number of customers at time t for four different cases considered in this study

The results are illustrated by implementing algorithm 1 in Wolfram Mathematica 12.3. Fig. 3 presents the time-varying number of customers in each station in the three-station tandem network. As can be seen from the figure, stations with high traffic intensity have a large number of customers in the queue. Fig. 4 shows the average virtual workload in each station. The workload is heavy for bottleneck stations when compared to other stations. If the bottleneck station is located in the first position, the workload becomes heavier than if it is located in the second or third position.

6.2 A five-station example

We consider another example of five-station tandem network model with similar external arrival rate and transition probability. We look at four different cases

Fig. 6 Average workload at time t for different cases. Here blue, red, green, brown and purple figures represent stations 1, 2, 3, 4 and 5 respectively



of traffic intensity for each station and summarised in Table 2. As in the previous example, Fig. 5 illustrates the time-varying number of customers and Fig. 6 presents the average virtual workload in the five-station tandem queueing network.

The results from the numerical study provide insight into the relation between location of bottleneck stations and transient performance measures. It is evident in this example that if the bottleneck station is located first, the workload in both the first station and the next stations will be heavy. As a result, customers will have to spend more time in the system. Interestingly, when the bottleneck station is at the end of the series, there is no effect on the average workload in the previous stations; hence, customers can obtain service without any delay. Whereas, when there are multiple bottleneck stations within the system, the average workload is consistently high.

7 Conclusion

In this paper, we discussed some transient distributional laws that characterise the performance of a time-varying tandem queueing network. To derive the virtual workload in the system, we initially considered a general single server queueing system with unlimited waiting capacity. Subsequently, we extended our model from single server queue to tandem queueing network of k servers and formulated transient performance measures such as, number of customers and average virtual workload in stations. We further introduced an algorithm that provides a general framework for obtaining transient performance measures in a k -station tandem network, and we implemented the algorithm through numerical study. The results clearly demonstrated the relationship between performance measures and traffic intensities.

Although time-varying queueing systems are widely used in many real life applications, their transient analysis requires more attention in the queueing systems literature. Through this paper, we have attempted to advance the field further. Implementation of finite buffer case and extending the results to different queueing disciplines are some interesting avenue for future research.

Acknowledgements We thank the editor and both anonymous reviewers for their comments and suggestions, leading to the improvement in the quality of this manuscript.

Author contributions All authors have contributed equally.

Funding This research received no external funding.

Data availability Not Applicable

Declarations

Conflict of interest The authors declare that they have no competing interests

Consent to participate Not Applicable

Consent for publication Not Applicable

References

1. Neely, M.J., Modiano, E., Rohrs, C.E.: Dynamic power allocation and routing for time-varying wireless networks. *IEEE J. Sel. Areas Commun.* **23**(1), 89–103 (2005). <https://doi.org/10.1109/jsac.2004.837349>
2. Leung, K.K., Massey, W.A., Whitt, W.: Traffic models for wireless communication networks. *IEEE J. Sel. Areas Commun.* **12**(8), 1353–1364 (1994). <https://doi.org/10.1109/49.329340>
3. Shakkottai, S., Srikant, R., Stolyar, A.L.: Pathwise optimality of the exponential scheduling rule for wireless channels. *Adv. Appl. Probab.* **36**(4), 1021–1045 (2004). <https://doi.org/10.1239/aap/1103662957>
4. Newell, G.F.: Approximation methods for queues with application to the fixed-cycle traffic light. *Siam Rev.* **7**(2), 223–240 (1965). <https://doi.org/10.1137/1007038>
5. Kurzhanskiy, A.A., Varaiya, P.: Active traffic management on road networks: a macroscopic approach. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **368**(1928), 4607–4626 (2010). <https://doi.org/10.1098/rsta.2010.0185>
6. Ran, B., Boyce, D.: *Dynamic Urban Transportation Network Models: Theory and Implications for Intelligent Vehicle-Highway Systems*. Springer Science & Business Media, Berlin (2012)
7. Yom-Tov, G.B., Mandelbaum, A.: Erlang-r: a time-varying queue with reentrant customers, in support of healthcare staffing. *Manuf. Serv. Op. Manag.* **16**(2), 283–299 (2014). <https://doi.org/10.1287/msom.2013.0474>
8. Shi, P., Chou, M.C., Dai, J.G., Ding, D., Sim, J.: Models and insights for hospital inpatient operations: time-dependent ED boarding time. *Manag. Sci.* **62**(1), 1–28 (2016). <https://doi.org/10.1287/mnsc.2014.2112>
9. Dai, J.G., Shi, P.: A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Op. Res.* **65**(2), 514–536 (2017). <https://doi.org/10.1287/opre.2016.1566>
10. Bertsimas, D., Mourtzinou, G.: Transient laws of non-stationary queueing systems and their applications. *Queueing Syst.* **25**(1–4), 115–155 (1997). <https://doi.org/10.1023/A:1019100301115>
11. Fralix, B.H., Riaño, G.: A new look at transient versions of little’s law, and m/g/1 preemptive last-come-first-served queues. *J. Appl. Probab.* **47**(2), 459–473 (2010). <https://doi.org/10.1239/jap/1276784903>
12. Kim, S.-H., Whitt, W.: Estimating waiting times with the time-varying little’s law. *Probab. Eng. Inf. Sci.* **27**(4), 471–506 (2013). <https://doi.org/10.1017/S0269964813000223>
13. Whitt, W.: Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Syst.* **81**, 341–378 (2015). <https://doi.org/10.1007/s11134-015-9462-x>
14. Gerum, P.C.L., Baykal-Gürsoy, M.: How incidents impact congestion on roadways: a queueing network approach. *EURO J. Transp. Logist.* **11**, 100067 (2022). <https://doi.org/10.10166/j.ejtl.2021.100067>
15. Gangadhar, N.D., Kadambi, G.R.: Delay distributions in discrete time multiclass tandem communication network models. *Int. J. Electr. Comput. Eng. Syst.* **13**(6), 417–425 (2022). <https://doi.org/10.32985/ijeces.13.6.1>
16. Sinu Lal, T., Krishnamoorthy, A., Joshua, V., Vishnevsky, V.: A two-stage tandem queue with specialist servers. *Appl. Probab. Stoch. Process.* (2020). https://doi.org/10.1007/978-981-15-5951-8_20
17. Prabhu, N.U.: Transient behaviour of a tandem queue. *Manag. Sci.* **13**(9), 631–639 (1967). <https://doi.org/10.1287/mnsc.13.9.631>
18. Murthy, M.B.R., Rao, K.S., Ravindranath, Rao, P.S.: Transient analysis of k-node tandem queueing model with load dependent service rates. *Int. J. Eng. Technol.* **7**, 141 (2018). <https://doi.org/10.14419/ijet.v7i3.31.18284>
19. Zychlinski, N., Mandelbaum, A., Momilović, P.: Time-varying tandem queues with blocking: modeling, analysis, and operational insights via fluid models with reflection. *Queueing Syst.* **89**, 15–47 (2018). <https://doi.org/10.1007/s11134-018-9578-x>
20. Brumelle, S.L.: On the relation between customer and time averages in queues. *J. Appl. Probab.* **8**(3), 508–520 (1971). <https://doi.org/10.2307/3212174>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.