Palmer, R., & Utley, M. (2020). On the modelling and performance measurement of service networks with heterogeneous customers. *Annals of Operations Research*, *293*, 237-268. https://doi.org/10.1007/s10479-019- 03391-z

Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to published version (if available):
10.1007/s10479-019- 03391-z

Link to publication record in Explore Bristol Research
PDF-document

This is the final published version of the article (version of record). It first appeared online via Springer at https://link.springer.com/article/10.1007/s10479-019-03391-z . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/

Annals of Operations Research (2020) 293:237–268
https://doi.org/10.1007/s10479-019-03391-z

S.I.: QUEUEING THEORY AND NETWORK APPLICATIONS II

On the modelling and performance measurement of service networks with heterogeneous customers

Ryan Palmer[1] · Martin Utley[2]

Published online: 2 December 2019
© The Author(s) 2019

Abstract

Service networks are common throughout the modern world, yet understanding how their individual services effect each other and contribute to overall system performance can be difficult. An important metric in these systems is the quality of service. This is an often overlooked measure when modelling and relates to how customers are affected by a service. Presented is a novel perspective for evaluating the performance of multi-class queueing networks through a combination of operational performance and service quality—denoted the "flow of outcomes". Here, quality is quantified by customers moving between or remaining in classes as a result of receiving service or lacking service. Importantly, each class may have different flow parameters, hence the positive/negative impact of service quality on the system's operational performance is captured. A fluid–diffusion approximation for networks of stochastic queues is used since it allows for several complex flow dynamics: the sequential use of multiple services; abandonment and possible rejoin; reuse of the same service; multiple customers classes; and, class and time dependent parameters. The scalability of the approach is a significant benefit since, the modelled systems may be relatively large, and the included flow dynamics may render the system analytically intractable or computationally burdensome. Under the right conditions, this method provides a framework for quickly modelling large time-dependent systems. This combination of computational speed and the "flow of outcomes" provides new avenues for the analysis of multi-class service networks where both service quality and operational efficiency interact.

Keywords Queueing · Multi-class networks · System performance · Fluid and diffusion approximation · Flow of outcomes

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10479-019-03391-z) contains supplementary material, which is available to authorized users.

✉ Ryan Palmer
ryan.palmer.14@ucl.ac.uk

[1] Department of Mathematics, UCL, Gower Street, London WC1E 6BT, UK

[2] Clinical Operational Research Unit, UCL, 4 Taviton Street, London WC1H 0BT, UK

1 Introduction

Throughout the modern world service systems such as health care services, telecommunications and computer networks are common and of significant importance to world economies. Typically these systems consist several, semi-autonomous services that each have a distinct function yet are linked by an overarching purpose to which they contribute to achieving. For such systems, the quality of the service provided/received by customers is important (Seth et al. 2005; Ghotbabadi et al. 2015). Particularly, the quality of service or the service outcomes are a key metric for gauging how well services are performing, individually and as a whole, and relate closely to the overarching system purpose. For example, in call centre systems the overarching purpose may be to sell goods and the service outcomes are the extent to which customer needs are met or what type of sale achieved. Within health care, the purpose of services is partly to maintain and improve patient health; thus, key service outcomes here may be the clinical impact on patient health (Palmer et al. 2017).

A combination of customer flow modelling and service quality is presented in this paper. In a model of customer flow, the system is viewed as comprising a set of distinct states through which discrete entities move. Often these systems are modelled from a purely operational perspective relating to how customers enter, leave and move between states that form the service process, and how queues build up and dissipate (Côté 2000). The operational modelling and measurement of such systems has a long history of research, but less work has been done considering in addition the quality of service.

Here, quality is quantified by customers moving between or remaining in classes as a result of receiving service or lacking service. Each class of customer may have different flow parameters representing differing resource/service requirements and different capacities to benefit from service. This combination broadens the perspective on how the performance of a queueing system may be understood. Namely, through the "flow of outcomes"—a perspective as to how individual services contribute to both the system's service outcomes, e.g. the output of customers in certain classes, and the system's operational performance. The model's output is thus informed by the effect of service, or absence of it, on customer class and the effect of customers with different service requirements.

At an individual level the "flow of outcomes" relates to how a single customer engages with/is engaged by the system and the resultant effect on them and their needs. At a population this aggregates to an understanding of how good and bad flow within the system may produce better or worse service outcomes. This is important when considering scarce resources, the possibility for multiple service interactions, and the subsequent effect that a changing mix of customers may have on operational performance. Thus, this method can help to inform resource allocation through new metrics that provide insight into service quality and operational efficiency. Whilst, the "flow of outcomes" may be applied to a single service, the application to a network of diverse yet similarly purposed services highlight its significant benefits.

When customers' use of a service relates to their service outcomes, the possible flow dynamics in such systems can become complex (Deo et al. 2013). Examples of such dynamics include the possibility for customers to use services multiple times or potentially use several services sequentially. As a result the system may quickly become intractable for traditional methods and the scale becomes large when considering multiple services and customer classes. In turn, this may greatly increase the number of operations required to model the system, thus increasing computational time and effort required.

To overcome these difficulties, presented in this paper is an application of a fluid–diffusion approximation for a network of services which serve several classes of customer. Several complex flow dynamics are considered for which a general model is presented (these dynamics are: the sequential use of multiple services; abandonment and possible rejoin; reuse of the same service; multiple customers classes; and, class and time dependent parameters). The classes in the model represent measurable aspects of a customer's service needs, requirements, behaviours and opinions (or some amalgam) that may be influenced by the receipt of service, or lack of it. Thus, when a customer leaves a service, they are modelled as being able to remain or transition in class, representing a service outcome which may mark the quality of service.

A fluid approximation is the limit in distribution for a stochastic process that is found by scaling the size of the system (number of servers and new arrivals) and applying the law-of-large-numbers. A diffusion limit can also be produced through an application of the functional central limit theorem to the scaled process (Remerova 2014). The variance within the queueing process of the system and performance measures can be calculated from the resulting limit, providing insight into the system's stochastic variability.

These approximations produce a continuous representation of the discrete process that overcomes the computational difficulty of traditional methods (Hillston 2005). The approach avoids blow-up of the state space in the analysis of large systems (Chen et al. 2016), which is a useful property when considering multiple services and classes. Due to the scaling process, the approximations produce more accurate results for large and heavily loaded systems (Ko and Gautam 2013) and are appropriate for analysing both transient behaviour in time-varying systems and the finite-horizon evolution of systems in steady-state (Yom-Tov and Mandelbaum 2014).

There is wide literature on fluid–diffusion approximations ranging from mathematical investigations to methodological developments/applications. In this paper, the method follows from and proceeds in a similar vein of Mandelbaum et al. (1998, 2002)—that is a fluid–diffusion approximation is developed for a Markovian service network using a strong law of large number limit theorem. Notably, in the presented work, the flow dynamics are similar to Ding et al. (2015) yet extended to apply to a network and include a further generalisation of the flow dynamics. It should be noted however that there are several other methods for producing fluid–diffusion approximations in general and for multi-class queues. Below are three examples—this is not exhaustive. Pang et al. (2007) present a "review" of martingale proof of many-server heavy-traffic limit theorems for Markovian queueing models working through several in-depth proofs of the underlying theory. Alternatively, non-Markovian approaches include the recent work by Pender and Ko (2017) in which fluid–diffusion approximations are derived for queues where the general interarrival and service times are approximated using phase-type distributions. Of note, a multi-class extension is possible in this latter case, however the dynamics presented here are not currently captured. A further approach, this time taking into account abandonment was produced by Massey and Pender (2013) in which they build upon the work of Mandelbaum et al. (1998), and that of Ko and Gautam (2013) to create a new three-dimensional dynamical system that is based on estimating the mean, variance, and third cumulant moment.

Furthermore, the mathematical properties and optimality of fluid approximations in modelling systems with multiple customer classes have been widely explored, in scenarios of varying dynamics. For single server queues of multiple customer classes, see Guo (2012) and Tahar and Jean-Marie (2012); systems with abandonment, see Whitt (2006) and Larrañaga et al. (2015); systems with multiple servers which may only serve specific customers, see Atar et al. (2011); and with non-fixed classes, see Tekin et al. (2012). Regarding applications of these methods, settings have included health care—in particular, acute care: Cohen et al. (2014), Yom-Tov and Mandelbaum (2014), Chen et al. (2016) and Zychlinski et al. (2017)—computing: Pender and Phung-Duc (2016) and Mukherjee et al. (2017)—and telecommunications: Mandelbaum et al. (2002) and Ding et al. (2015). However the scope and applicability of these methods can extend to several other settings including other service systems, computer systems and manufacturing.

Overall, the main aim of this work was to produce a method that both captures complex flow dynamics for a network of services and quantifies the impact of the operational performance and service quality. Our motivating context was in health care however the method is presented generally due to the potential for wider application. Given this, the main contribution is to the modelling of multi-class queueing systems and the performance measurement of service systems where service quality is important and may effect the future operation of the system. Building upon current approximation methods (Mandelbaum et al. 2002; Ding et al. 2015), a framework is produced that combines customer classes and flow modelling by using time and class dependent parameters. The method developed is a generalisation of this previous work and may be used to model systems with complex flow dynamics (such as service reuse; abandonment and rejoins; and several services). Further distinctions of this method are the inclusion of transitions between class, and the generalisation of customer movements post-abandonment/post-service.

In the following section the stochastic system is described for which fluid–diffusion approximations are produced and the dynamic server allocations are introduced. In Sect. 3, the approximations for this system and the output measures that may be calculated are presented. Finally, in Sect. 4 the pragmatic constraints for the possible use and limitation of these methods are discussed alongside the possible directions for future work.

2 Description of the stochastic system

Consider a network consisting of J multi-server services. For any service, during a time interval $[0, T]$, customers may: arrive as a new customer; abandon the queue and potentially rejoin it, seek to use an alternative service or leave the system as a loss (L); or, receive service and potentially reuse the same service, use another service within the network, or exit having completed service (E)—see Fig. 1. Each service therefore consists of five process orbits: the service and queue (Q), the rejoin process (R), the reuse process (U), the alternative service process (A), and the other service process (O). Note that the term *alternative service* always refers to a use of another service after abandonment, and that *other service* refers to a use of another service having just completed service.

Suppose that at any time, a customer belongs to a class $k \in \{1, 2, \ldots, K\} = Cla$. Each class represents a level of progressive measure of quality/customer need that customers move between as they proceed through the system. For a service $i \in \{1, 2, \ldots, J\} = Ser$ and class $k \in Cla$, at time $t \in [0, T]$, denote:

$$\begin{aligned}
Z_{k,Q,i}(t) &:= \text{number of customers in the queue or service,}\\
Z_{k,R,i}(t) &:= \text{number of customers in the rejoin orbit,}\\
Z_{k,U,i}(t) &:= \text{number of customers in the reuse orbit,}\\
Z_{k,A,i}(t) &:= \text{number of customers in the alternative service orbit,}\\
Z_{k,O,i}(t) &:= \text{number of customers in the other service orbit,}\\
Z_{k,L,i}(t) &:= \text{number of customers lost due to abandonment,}\\
Z_{k,E,i}(t) &:= \text{number of customers leave after service.}
\end{aligned}$$

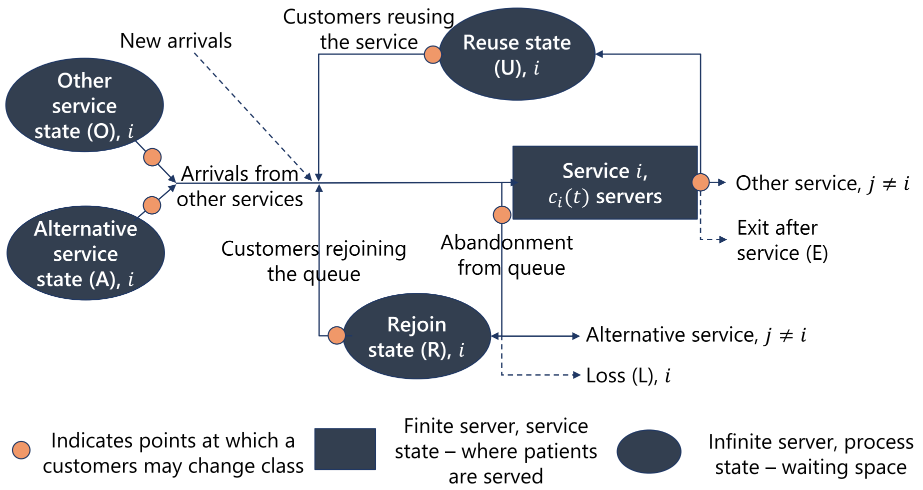
New arrivals
Customers reusing the service
Reuse state (U), i
Other service state (O), i
Arrivals from other services
Alternative service state (A), i
Service i, $c_i(t)$ servers
Other service, $j \neq i$
Exit after service (E)
Customers rejoining the queue
Abandonment from queue
Rejoin state (R), i
Alternative service, $j \neq i$
Loss (L), i
Indicates points at which a customers may change class
Finite server, service state – where patients are served
Infinite server, process state – waiting space

Fig. 1 Diagram of general customer flow in all classes through a single service within the stochastic queueing network

For each service $i \in Ser$, a time-varying number of servers, $c_i(t)$, are available. For tractability, customers are served in up to K parallel queues per service, each formed of a single class. Notably, classes can be defined such that customers have equal priority within each class; thus, they are served on a first come first served basis (FCFS).

The number of servers allocated to customers in class k queueing for service i at time t is denoted $C_{k,i}(t)$ such that $\sum_{k \in Cla} C_{k,i}(t) = c_i(t)$. In Sect. 2.1, two methods for allocating servers are suggested. Since the number of servers for each queue may vary with time, the number of servers may drop below the number of customers in service. This situation is handled in the model by pre-emptive resumption (Mandelbaum et al. 2002). That is, the number of customers in service is reduced to equal the number of servers by placing arbitrary customers into an infinite buffer space. Their service thus paused and later resumed once a server becomes available, with priority ahead of the queue.

New customers, in a class $k \in Cla$, arrive to a service $i \in Ser$ according to a time-inhomogeneous Poisson process of rate $\lambda_{k,i}(t)$. If a server is free, customers are served according to a time-inhomogeneous exponentially distributed process of rate $\mu_{k,i}(t)$. If no servers are available, they wait within an infinite buffer space. Whilst waiting, customers may lose patience and abandon the queue at a time-inhomogeneous exponentially distributed rate $\theta_{k,i}(t)$.

Upon abandoning, one of three events may occur: (1) a customer may rejoin the queue (seeking to access the service again) with probability $r_{k,L,i,i}(t)$. Such customers enter the rejoin orbit, where they spend a time-inhomogeneous exponentially distributed amount of time, rejoining the queue at rate $\delta_{k,R,i}(t)$; (2) a customer may seek to use an alternative service with probability $r_{k,L,i,j}(t),\ i \neq j,\ j \in Ser$. These customers enter the alternative service orbit for j, where they spend a time-inhomogeneous exponentially distributed amount of time, joining the queue at a rate $\delta_{k,A,j}(t)$; (3) a customer will leave the system as a loss with a probability $r_{k,L,i,J+1}(t) > 0$. Notably: $\sum_{j=1}^{J+1} r_{k,L,i,j}(t) = 1,\ \forall t \in [0, T]$.

Similarly, after completing service, one of three events may occur: (1) having used a service i, a customer may seek further service within i with probability $r_{k,S,i,i}(t)$, entering the reuse orbit. Customers spend a time-inhomogeneous exponentially distributed amount of time in this state, arriving to the service at rate $\delta_{k,U,i}(t)$; (2) with probability $r_{k,S,i,j}(t),\ i \neq j,\ j \in Ser$ a customer may seek to use another service, entering the orbit of arrivals from other services for j. They remain in this state for a time-inhomogeneous exponentially distributed amount of time, and join the queue at a rate $\delta_{k,O,j}(t)$; (3) there is a probability $r_{k,S,i,J+1}(t) > 0$ that a customer will not require any further service and leave the system. Notably: $\sum_{j=1}^{J+1} r_{k,S,i,j}(t) = 1,\ \forall t \in [0, T]$.

A customer's class may change throughout their interaction with the system and may occur at: the completion of service; the point of abandoning the queue; or, upon joining the queue as a rejoin, reuse, alternative service arrival or other service arrival. $s_{k,l,m,i}(t)$ is the probability that a customer transitions from a class k to a class l given that they are leaving a process state $m \in \{Q, R, U, A, O\}$ at time t, or abandoning the queue when $m = L$.

Given the above, the stochastic process for this system, $\{\mathbf{Z}(t), t \geq 0\}$, can be defined as a vector of length $7KJ$ such that:

$$\mathbf{Z}(t) := (\mathbf{Z}_{1,1}(t), \mathbf{Z}_{2,1}(t), \ldots, \mathbf{Z}_{K,1}(t), \mathbf{Z}_{1,2}(t), \ldots, \mathbf{Z}_{K,2}(t), \ldots, \mathbf{Z}_{K,J}(t))^T. \tag{1}$$

where for $k \in Cla, i \in Ser$:

$$\mathbf{Z}_{k,i}(t) := (Z_{k,Q,i}(t), Z_{k,R,i}(t), Z_{k,U,i}(t), Z_{k,A,i}(t), Z_{k,O,i}(t), Z_{k,L,i}(t), Z_{k,E,i}(t)).$$

This is a Markov process since the inter-arrival rates, service duration and orbit durations are exponentially distributed, and class/service state transitions are Markovian. The state space for this process is $\mathbb{Z}_{+}^{7KJ}$.

2.1 Dynamic multi-class server allocations

To ensure tractability whilst modelling the differentiated service of different class customers, each service is modelled using parallel queues that pertain to each class and share from a single pool of servers. To maintain the FCFS assumption, servers must be allocated to each queue.

The simplest method is to equally assign servers across queues. If K is not a factor of $c_i(t)$, define: $C_{k,i}(t) = \left\lfloor \frac{c_i(t)}{K} \right\rfloor$. If $\sum_{k=1}^{K} C_{k,i}(t) < c_i(t)$, assign $c_i(t) - \sum_{k=1}^{K} C_{k,i}(t)$ servers, one at a time, to arbitrary queues until all servers are assigned.

In using constant allocations, the only interaction between the queues is through the class transitions of customers, otherwise the queues act autonomously. However, in real world systems, queues may affect one another through how customers use servers, i.e. by using a server, a customer denies other customers the opportunity to be served by that server.

One way to model this is through a dynamic server allocation. There is a wide and extensive literature that is relevant to this type of allocation such as that for multi-class queues e.g. Federgruen and Groenevelt (1988), Maglaras (1999) and Ata (2006) and scenarios where the pool of servers are shared e.g. Zhang and Tian (2004) and Atar et al. (2004). Applied here is an allocation which continually updates in response to the changes in the overall demand for service, the attributes of different customer classes and customer mix. Within the stochastic system the number of servers allocated to each queue is updated each time an event occurs that changes the size of $Z_{k,Q,i}(t)$ e.g. an arrival (new or from a process state), a completion of service or an abandonment. Thus, the fluid approximation will provide a continuous, deterministic approximation.

One method is to assign servers to each queue according to the proportion of customers in each class k and in the queue/service state for a service i:

$$C_{k,i}(t) = C_{k,i}(\mathbf{Z}(t)) = \left\lfloor \frac{c_i(t) Z_{k,Q,i}(t)}{\sum_{l=1}^{K} Z_{l,Q,i}(t)} \right\rfloor. \tag{2}$$

Alternatively, a continuous weight or cost function, $B_{k,i}(t)$, could be used to favour customers in certain classes. For example, $B_{k,i}(t)$ may be defined as $1/\mu_{k,i}(t)$ to allocate servers to the queues that will take the longest time to serve. Or, if $B_{k,i}(t) = \theta_{k,i}(t)$, servers are allocated based on the potential for customers to abandon, seeking to mitigate losses in the system. Thus, for the stochastic process servers may be allocated by:

$$C_{k,i}(\mathbf{Z}(t)) = \left\lfloor \frac{c_i(t) B_{k,i}(t) Z_{k,Q,i}(t)}{\sum_{l=1}^{K} B_{l,i}(t) Z_{l,Q,i}(t)} \right\rfloor, \quad \text{for all } t \in [0, T]. \tag{3}$$

In both cases, the method introduced above may be implemented to ensure that all the servers are allocated when the total number does not divide evenly. Additionally, the system must never be empty to ensure that the allocation is well-defined. A further limitation of these allocations is that their fluid approximation must be continuously differentiable, a limitation introduced by the calculation of the virtual waiting time (VWT), as discussed later.

Notably each of the above allocations depend on $Z_{k,Q,i}(t)$ and thus depend on $C_{k,i}(t)$ by definition. This however is not a limitation since a fall in the number of allocated servers leads to customers who were formerly in service re-entering the queue such that $Z_{k,Q,i}(t)$ is unchanged. Furthermore, allocations may be defined based on different orbits, process states or queue as long as the definition remains continuously differentiable. In scenarios where $C_{k,i}(t)$ does not depend on the output of the stochastic system (e.g. a constant function) the input parameters may be piecewise continuous.

These dynamic allocations may be used to understand how the service requirements of customers in different classes and fluctuations in demand affect the operation of the system, since the allocations continuously update in response to customer mix and server occupancy. Likewise, this method overcomes the traditional inefficiency of parallel queues—that customers of one class may be waiting, whilst servers assigned to other queues are inactive. Overall, this method may be used to understand how the allocation of servers can help to mitigate negative process outcomes.

3 Fluid and diffusion approximations for stochastic queueing networks with heterogeneous customers

3.1 Fluid approximation for stochastic queueing networks with heterogeneous customers

Conservation equations are first formulated for the stochastic system (1) in a similar manner to Ding et al. (2015). The are here defined to include multiple classes, dynamic server allocations, multiple services and the new orbits these introduce. Notably, these definitions are also inline with the method set out by Mandelbaum et al. (2002).

Flux terms for modelling the movement of customers within the system must be defined for each class $k, l \in Cla$ and for each service $i, j \in Ser$. Firstly, the arrival process of new customers $\Pi_{\lambda_{k,i}(t)}$, is a Poisson process of rate $\lambda_{k,i}(t)$. Secondly, the number of customers leaving process states—service (S), abandonment from queue (L), rejoin (R), reuse (U), alternative service (A) and arrivals from other services (O)—are defined as independent Poisson processes of rate 1 such that:

$$D_{k,S,i}(t) = \Pi_{k,S,i}\left(\int_0^t \mu_{k,i}(u)\min\left(Z_{k,Q,i}(u), C_{k,i}(\mathbf{Z}(u))\right)du\right), \tag{4}$$

$$D_{k,L,i}(t) = \Pi_{k,L,i}\left(\int_0^t \theta_{k,i}(u)(Z_{k,Q,i}(u) - C_{k,i}(\mathbf{Z}(u)))^+du\right), \tag{5}$$

$$D_{k,m,i}(t) = \Pi_{k,m,i}\left(\int_0^t \delta_{k,m,i}(u)Z_{k,m,i}(u)du\right). \tag{6}$$

$m \in \{R, U, A, O\} = St$, $(x)^+ := \max(0, x)$. Proof of these statements may be produced along the lines of Lemma 2.1 in Pang et al. (2007).

Thirdly, multinomial random variables are used to model the movement of customers between classes. For customers who transition to a new process state according to $D_{k,m,i}(t)$, $m \in \{S, L, R, U, A, O\}$, a change in class is given by:

$$\mathbf{MS}_{k,m,i}(t) \sim \text{Mult}(D_{k,m,i}(t), \mathbf{s}_{k,m,i}(t)). \tag{7}$$

where $\mathbf{MS}_{k,m,i}(t)$ is a vector of length K. Its l-th element, denoted $MS^{(l)}_{k,m,i}(t)$, gives the number of customers who were in class k before moving to class l, according to $D_{k,m,i}(t)$. This process is governed by class transition parameters:

$$\mathbf{s}_{k,m,i}(t) = (s_{k,1,m,i}(t), s_{k,2,m,i}(t), \ldots, s_{k,K,m,i}(t)).$$

where $\sum_{l=1}^{K} s_{k,l,m,i}(t) = 1$ such that $\sum_{l=1}^{K} MS^{(l)}_{k,m,i}(t) = D_{k,m,i}(t)$.

Again, multinomial random variables are used to model the movement of customers after abandoning/completing service. For customers who, upon abandoning/completing service for $i \in Ser$, have moved to a class k, $\sum_{l=1}^{K} \mathbf{MS}^{(k)}_{l,n,i}(t)$, $n \in \{S, L\}$, their post abandonment/service movement is modelled by:

$$\mathbf{MR}_{k,n,i}(t) \sim \text{Mult}\left(\sum_{l=1}^{K} \mathbf{MS}^{(k)}_{l,n,i}(t), \mathbf{r}_{k,n,i}(t)\right). \tag{8}$$

where $\mathbf{MR}_{k,n,i}(t)$ is a vector of length $J+1$. Its j-th element, denoted $MR^{(j)}_{k,n,i}(t)$, gives the number of class k customers who enter the alternative/other service orbit for $j \in Ser,\ j \neq i$; or, for $j = i$, enter the rejoin/reuse orbit of i; or, for $j = J+1$, leave the system as a loss/exit the system upon completion of service. This process is governed by post-abandonment transition parameters:

$$\mathbf{r}_{k,n,i}(t) = (r_{k,n,i,1}(t), r_{k,ni,2}(t), \ldots, r_{k,n,i,J}(t), r_{k,n,i,J+1}(t)).$$

where $\sum_{j=1}^{J+1} r_{k,n,i,j}(t) = 1$ such that $\sum_{j=1}^{J+1} MR^{(j)}_{k,n,i}(t) = \sum_{l=1}^{K} \mathbf{MS}^{(k)}_{l,n,i}(t)$.

Given these flux terms, the conservation equations for customer flow in (1), for $t \in [0, T)$, for $k, l \in Cla$ and $i, j \in Ser$, are:

$$\begin{aligned} Z_{k,Q,i}(t) = Z_{k,Q,i}(0) + \Pi_{\lambda_{k,i}(t)} + \sum_{l=1}^{K}\sum_{m\in St} MS^{(k)}_{l,m,i}(t) \\ - D_{k,S,i}(t) - D_{k,L,i}(t), \end{aligned} \tag{9}$$

$$Z_{k,R,i}(t) = Z_{k,R,i}(0) + MR^{(i)}_{k,L,i}(t) - D_{k,R,i}(t), \tag{10}$$

$$Z_{k,U,i}(t) = Z_{k,U,i}(0) + MR^{(i)}_{k,S,i}(t) - D_{k,U,i}(t), \tag{11}$$

$$Z_{k,A,i}(t) = Z_{k,A,i}(0) + \sum_{j=1;\, j\neq i}^{J} MR^{(i)}_{k,L,j}(t) - D_{k,A,i}(t), \tag{12}$$

$$Z_{k,O,i}(t) = Z_{k,O,i}(0) + \sum_{j=1;\, j\neq i}^{J} MR^{(i)}_{k,S,j}(t) - D_{k,O,i}(t), \tag{13}$$

$$Z_{k,L,i}(t) = Z_{k,L,i}(0) + MR^{(J+1)}_{k,L,i}(t), \tag{14}$$

$$Z_{k,E,i}(t) = Z_{k,E,i}(0) + MR^{(J+1)}_{k,S,i}(t). \tag{15}$$

To formulate the fluid limit for Eqs. (9)–(15), consider a sequence of models where the η-th model—denoted by the superscript (η)—has a scaled arrival rate $\eta\lambda_{k,i}(t)$ for new customers and scaled number of servers $\eta c_i(t)$ for all $k \in Cla$ and $i \in Ser$. The *scaled fluid process* is defined as: $\overline{Z}^{(\eta)}_{k,m,i}(t) := \frac{Z^{(\eta)}_{k,m,i}(t)}{\eta}$, for $k \in Cla,\ i \in Ser,\ m \in \{Q, R, U, A, O, L\}$ and is gained by replacing (4)–(8) with:

$$\overline{D}^{(\eta)}_{k,S,i}(t) = \Pi_{k,S,i}\left(\eta\int_0^t \mu_{k,i}(u)\min\left(\overline{Z}^{(\eta)}_{k,Q,i}(u), C_{k,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)du\right)\Big/\eta,$$

$$\overline{D}^{(\eta)}_{k,L,i}(t) = \Pi_{k,L,i}\left(\eta\int_0^t \theta_{k,i}(u)\left(\overline{Z}^{(\eta)}_{k,Q,i}(u) - C_{k,i}\left(\overline{\mathbf{Z}}^{(\eta)}(u)\right)\right)^+ du\right)\Big/\eta,$$

$$\overline{D}^{(\eta)}_{k,m_1,i}(t) = \Pi_{k,m_1,i}\left(\eta\int_0^t \delta_{k,m_1,i}(u)\overline{Z}^{(\eta)}_{k,m_1,i}(u)du\right)\Big/\eta,$$

$$\overline{\mathbf{MS}}^{(\eta)}_{k,m_2,i}(t) = \frac{\mathbf{MS}^{(\eta)}_{k,m_2,i}(t)}{\eta},\ \ \mathbf{MS}^{(\eta)}_{k,m_2,i}(t) \sim \text{Mult}\left(\eta\overline{D}^{(\eta)}_{k,m_2,i}(t), \mathbf{s}_{k,m_2,i}(t)\right),$$

$$\overline{\mathbf{MR}}^{(\eta)}_{k,n,i}(t) = \frac{\mathbf{MR}^{(\eta)}_{k,n,i}(t)}{\eta},\ \ \mathbf{MR}^{(\eta)}_{k,n,i}(t) \sim \text{Mult}\left(\sum_{l=1}^{K}\mathbf{MS}^{(\eta)^{(k)}}_{l,n,i}(t), \mathbf{r}_{k,n,i}(t)\right).$$

where $m_1 \in St, m_2 \in \{Q, R, U, A, O, L\}, n \in \{S, L\}$.

Furthermore, to construct the fluid limit for a time period $[0, T]$, $k \in Cla$ and $i \in Ser$ the following initial conditions are required: $(z_{k,Q,i}(0), z_{k,R,i}(0),\ z_{k,U,i}(0), z_{k,A,i}(0), z_{k,O,i}(0), z_{k,L,i}(0), z_{k,E,i}(0))$. From the above definitions and following the Theorem 2.1 of Mandelbaum et al. (2002), the fluid approximation is as follows:

Theorem 1 *Assuming that for the given initial conditions,* $\lim_{\eta\to\infty}\overline{Z}^{(\eta)}_{k,m,i}(0) = z_{k,m,i}(0)$, *where* $k \in Cla$, $i \in Ser$, *and* $m = \{Q, R, U, A, O, L, E\}$. *Then, by the law of large numbers the fluid limit for (1) is* $\lim_{\eta\to\infty}\overline{Z}^{(\eta)}_{k,m,i}(t) = z_{k,m,i}(t)$*—where the convergence of* t. *This is uniquely determined by the initial conditions and the following system of equations where* $t \in [0, T)$*:*

$$\begin{aligned} z_{k,Q,i}(t) =\ & z_{k,Q,i}(0) + \int_0^t \lambda_{k,i}(u) - \theta_{k,i}(u)\big(z_{k,Q,i}(u) - c_{k,i}(\mathbf{z}(u))\big)^+ du \\ & + \int_0^t \sum_{l=1}^{K}\sum_{n\in St}\Big(s_{l,k,m,i}(u)\delta_{l,m,i}(u)z_{l,m,i}(u)\Big)\,du, \\ & - \int_0^t \mu_{k,i}(u)\min\big(z_{k,Q,i}(u), c_{k,i}(\mathbf{z}(u))\big)\,du, \end{aligned} \tag{16}$$

$$\begin{aligned} z_{k,R,i}(t) = {} & z_{k,R,i}(0) - \int_0^t \delta_{k,R,i}(u) z_{k,R,i}(u)\, du \\ & + \int_0^t r_{k,L,i,i}(u) \sum_{l=1}^{K} s_{l,k,L,i}(u)\theta_{l,i}(u)\big(z_{l,Q,i}(u) - c_{l,i}(\mathbf{z}(u))\big)^{+}\, du, \end{aligned} \tag{17}$$

$$\begin{aligned} z_{k,U,i}(t) = {} & z_{k,U,i}(0) - \int_0^t \delta_{k,U,i}(u) z_{k,U,i}(u)\, du \\ & + \int_0^t r_{k,S,i,i}(u) \sum_{l=1}^{K} s_{l,k,S,i}(u)\mu_{l,i}(u) \min\big(z_{l,Q,i}(u), c_{l,i}(\mathbf{z}(u))\big)\, du, \end{aligned} \tag{18}$$

$$\begin{aligned} z_{k,A,i}(t) = {} & z_{k,A,i}(0) - \int_0^t \delta_{k,A,i}(u) z_{k,A,i}(u)\, du \\ & + \int_0^t \sum_{j\neq i}^{J} \sum_{l=1}^{K} r_{k,L,j,i}(u) s_{l,k,L,j}(u)\theta_{l,j}(u) \\ & \times \big(z_{l,Q,j}(u) - c_{l,j}(\mathbf{z}(u))\big)^{+}\, du, \end{aligned} \tag{19}$$

$$\begin{aligned} z_{k,O,i}(t) = {} & z_{k,O,i}(0) - \int_0^t \delta_{k,O,i}(u) z_{k,O,i}(u)\, du \\ & + \int_0^t \sum_{j\neq i}^{J} \sum_{l=1}^{K} r_{k,S,j,i}(u) s_{l,k,S,j}(u)\mu_{l,j}(u) \\ & \times \min\big(z_{l,Q,j}(u), c_{l,j}(\mathbf{z}(u))\big)\, du, \end{aligned} \tag{20}$$

$$\begin{aligned} z_{k,L,i}(t) = z_{k,L,i}(0) + & \int_0^t \sum_{l=1}^{K} r_{k,L,i,J+1}(u)\, s_{l,k,L,i}(u)\theta_{l,i}(u) \\ & \times \big(z_{l,Q,i}(u) - c_{l,i}(\mathbf{z}(u))\big)^{+}\, du, \end{aligned} \tag{21}$$

$$\begin{aligned} z_{k,E,i}(t) = z_{k,E,i}(0) + & \int_0^t \sum_{l=1}^{K} r_{k,S,i,J+1}(u)\, s_{l,k,S,i}(u)\mu_{l,i}(u) \\ & \times \min\big(z_{l,Q,i}(u), c_{l,i}(\mathbf{z}(u))\big)\, du. \end{aligned} \tag{22}$$

Analytical expressions cannot be found for (16)–(22); however, these equations can be solved using common numerical schemes. By Theorem 1, fluid approximations are gained for server allocations (2) and (3). For (2), the continuous fluid approximation is:

$$c_{k,i}(\mathbf{z}(t)) = \frac{c_i(t) z_{k,Q,i}(t)}{\sum_{l=1}^{K} z_{l,Q,i}(t)}, \forall t \in [0, T]. \tag{23}$$

For (3), the weighted allocation, the continuous fluid approximation is:

$$c_{k,i}(\mathbf{z}(t)) = \frac{c_i(t) B_{k,i}(t) z_{k,Q,i}(t)}{\sum_{l=1}^{K} B_{l,i}(t) z_{l,Q,i}(t)}, \forall t \in [0, T]. \tag{24}$$

Notably, the server allocation algorithm does not need to be implemented.

As previously noted, to calculate the VWT (Sect. 3.2), $c_{k,i}(\mathbf{z}(t))$ needs to be continuously differentiable. Thus, $z_{k,Q,i}(t)$, $\forall k \in Cla, i \in Ser$ must be continuously differentiable throughout $[0, T]$, giving the requirement that all input parameters are continuous. The time varying server allocation $c_{k,i}(\mathbf{z}(t))$ is a further output for the model.

3.2 Diffusion approximation

Following the method set out by Mandelbaum et al. (1998, 2002), the diffusion limit can be formulated for (1). The diffusion limit quantifies deviations from the first order fluid approximation (Mandelbaum et al. 1998), providing a system of ODEs for calculating the mean and covariance of the diffusion process. Since all of the flow functions are continuous, the assumptions stated in Theorem 2.4 of Mandelbaum et al. (1998) are maintained; thus, the diffusion approximation may be formulated as in Mandelbaum et al. (2002).

By Theorem 1 of this paper, $\lim_{\eta\to\infty}\overline{\mathbf{Z}}^{(\eta)}(t) = \mathbf{z}(t)$ a.s. with uniform convergence on compact sets of t (Mandelbaum et al. 2002). Thus, the diffusion limit is gained by applying the functional central limit theorem to $\hat{\mathbf{z}}(t) = \{\hat{\mathbf{z}}(t)|T > t \geq 0\}$ (Mandelbaum et al. 1998). Hence, if $\lim_{\eta\to\infty}\sqrt{\eta}(\overline{\mathbf{Z}}^{(\eta)}(0) - \mathbf{z}(0)) = \hat{\mathbf{z}}(0)$ holds, with $\hat{\mathbf{z}}(0)$ constant, then:

$$\lim_{\eta\to\infty}\sqrt{\eta}\left(\overline{\mathbf{Z}}^{(\eta)}(t) - \mathbf{z}(t)\right) \overset{d}{=} \hat{\mathbf{z}}(t).$$

This is a convergence in distribution of the processes (Mandelbaum et al. 1998). If the set of time points $\{t \in [0, T)\,|z_{k,Q,i}(t) = c_{k,i}(\mathbf{z}(t))\}$ has zero measure, $\hat{\mathbf{z}}(t)$ is a Gaussian process (Mandelbaum et al. 2002). Thus, the mean vector and covariance matrix for the diffusion process are the unique solutions to autonomous differential equations. Furthermore, for a service $j \in Ser$ both $\min(z_{k,Q,j}(t), c_{k,j}(\mathbf{z}(t)))$ and $(z_{k,Q,j}(t) - c_{k,j}(\mathbf{z}(t)))^+$, are everywhere continuous. Also, they are everywhere differentiable, except when $z_{k,Q,j}(t) = c_{k,j}(\mathbf{z}(t))$.

Explicitly formulating this system for (1) the method in Mandelbaum et al. (2002) is extended. First note Eqs. (16)–(22) may be rewritten in terms of column transition vectors and locally integrable Lipschitz continuous rate functions $\alpha_{t,i}(\mathbf{x}(t))$. Since $\hat{\mathbf{z}}(t)$ is a column vector, for $0 \leq t < T$ and for all $\mathbf{x} \in \mathbb{R}^{(7KJ)}$, define:

$$\alpha_t(\mathbf{x}(t)) \equiv \sum_{i\in I}\alpha_{t,i}(\mathbf{x}(t))\mathbf{v}_i.$$

such that:

$$\frac{d}{dt}\mathbb{E}\left[\hat{\mathbf{z}}(t)\right] = \mathbf{A}_t^T\mathbb{E}\left[\hat{\mathbf{z}}(t)\right].$$
$$\frac{d}{dt}\mathbf{Cov}[\hat{\mathbf{z}}(t)] = \mathbf{Cov}[\hat{\mathbf{z}}(t)]\mathbf{A}_t^T + \mathbf{A}_t\mathbf{Cov}[\hat{\mathbf{z}}(t)] + \mathbf{B}_t.$$

where $\mathbf{A}_t = D\alpha_t(\mathbf{z}(t))$ is the Jacobian of $\alpha_t(\mathbf{z}(t))$ when differentiated at $\mathbf{z}(t)$ and $\mathbf{B}_t = \sum_{i\in I}\alpha_{t,i}(\mathbf{z}(t))\mathbf{v}_i \otimes \mathbf{v}_i$ is the tensor product of two vectors forming a symmetrical matrix. For $0 \leq t < T$, the matrices, $\mathbf{A}(t)$, $\mathbf{B}(t)$ and $\mathrm{Cov}[\hat{\mathbf{z}}(t)]$ are of dimension $7KJ \times 7KJ$.

Working with a more explicit notation to highlight how this method applies to the extended system, begin with the rate functions, for $k, l \in H$ and $i, j \in Ser$:

$$\begin{aligned}
\alpha_{k,i,1}(\mathbf{z}(t)) &= \lambda_{k,i}(t),\\
\alpha_{k,l,i,2}(\mathbf{z}(t)) &= s_{k,l,R,i}(t)\delta_{k,R,i}(t)z_{k,R,i}(t),\\
\alpha_{k,l,i,3}(\mathbf{z}(t)) &= s_{k,l,U,i}(t)\delta_{k,U,i}(t)z_{k,U,i}(t),\\
\alpha_{k,l,i,4}(\mathbf{z}(t)) &= s_{k,l,A,i}(t)\delta_{k,A,i}(t)z_{k,A,i}(t),\\
\alpha_{k,l,i,5}(\mathbf{z}(t)) &= s_{k,l,O,i}(t)\delta_{k,O,i}(t)z_{k,O,i}(t),\\
\alpha_{k,l,i,6}(\mathbf{z}(t)) &= s_{k,l,L,i}(t)r_{l,L,i,J+1}(t)\theta_{k,i}(t)(z_{k,Q,i}(t) - c_{k,i}(\mathbf{z}(t)))^+,\\
\alpha_{k,l,i,7}(\mathbf{z}(t)) &= s_{k,l,S,i}(t)r_{l,S,i,J+1}(t)\mu_{k,i}(t)\min\left(z_{k,Q,i}(t), c_{k,i}(\mathbf{z}(t))\right),
\end{aligned}$$

$$\alpha_{k,l,i,j,8}(\mathbf{z}(t)) = s_{k,l,L,i}(t) r_{l,L,i,j}(t)\theta_{k,i}(t)(z_{k,Q,i}(t) - c_{k,i}(\mathbf{z}(t)))^{+},$$
$$\alpha_{k,l,i,j,9}(\mathbf{z}(t)) = s_{k,l,S,i}(t) r_{l,S,i,j}(t)\mu_{k,i}(t) \min\left(z_{k,Q,i}(t), c_{k,i}(\mathbf{z}(t))\right).$$

such that:

$$\begin{aligned} z_{k,Q,i}(t) = z_{k,Q,i}(0) + \int_0^t \alpha_{k,i,1}(u) + \sum_{l=1}^{K} \Bigg(& \alpha_{l,k,i,2}(u) + \alpha_{l,k,i,3}(u) + \alpha_{l,k,i,4}(u) \\ & + \alpha_{l,k,i,5}(u) - \alpha_{k,l,i,6}(u) - \alpha_{k,l,i,7}(u) \\ & - \sum_{j=1}^{J} \left(\alpha_{k,l,i,j,8}(u) + \alpha_{k,l,i,j,9}(u)\right) \Bigg) du, \end{aligned}$$

$$z_{k,R,i}(t) = z_{k,R,i}(0) + \int_0^t \sum_{l=1}^{K} \Big(\alpha_{l,k,i,i,8}(u) - \alpha_{k,l,i,2}(u)\Big) du,$$

$$z_{k,U,i}(t) = z_{k,U,i}(0) + \int_0^t \sum_{l=1}^{K} \Big(\alpha_{l,k,i,i,9}(u) - \alpha_{k,l,i,3}(u)\Big) du,$$

$$z_{k,A,i}(t) = z_{k,A,i}(0) + \int_0^t \sum_{l=1}^{K} \left(\sum_{i=1;\, j\neq i}^{J} \alpha_{l,k,j,i,8}(u) - \alpha_{k,l,i,4}(u) \right) du,$$

$$z_{k,O,i}(t) = z_{k,O,i}(0) + \int_0^t \sum_{l=1}^{K} \left(\sum_{i=1;\, j\neq i}^{J} \alpha_{l,k,j,i,9}(u) - \alpha_{k,l,i,5}(u) \right) du,$$

$$z_{k,L,i}(t) = z_{k,L,i}(0) + \int_0^t \sum_{l=1}^{K} \alpha_{l,k,i,6}(u) du,$$

$$z_{k,D,i}(t) = z_{k,D,i}(0) + \int_0^t \sum_{l=1}^{K} \alpha_{l,k,i,7}(u) du.$$

Continuing with this notation, now form a basis of transition vectors of length $7KJ$. Denoting the m-th element of each vector as $\mathbf{v}_{k,i,1}^{(m)}$, the transition vectors are defined as:

$$\mathbf{v}_{k,i,1}^{(m)} = \begin{cases} 1, & \text{if } m = 7K(i-1) + 7(k-1) + 1, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathbf{v}_{k,l,i,2}^{(m)} = \begin{cases} 1, & \text{if } m = 7K(i-1) + 7(l-1) + 1, \\ -1, & \text{if } m = 7K(i-1) + 7(k-1) + 2, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathbf{v}_{k,l,i,3}^{(m)} = \begin{cases} 1, & \text{if } m = 7K(i-1) + 7(l-1) + 1, \\ -1, & \text{if } m = 7K(i-1) + 7(k-1) + 3, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathbf{v}_{k,l,i,4}^{(m)} = \begin{cases} 1, & \text{if } m = 7K(i-1) + 7(l-1) + 1, \\ -1, & \text{if } m = 7K(i-1) + 7(k-1) + 4, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathbf{v}_{k,l,i,5}^{(m)} = \begin{cases} 1, & \text{if } m = 7K(i-1) + 7(l-1) + 1, \\ -1, & \text{if } m = 7K(i-1) + 7(k-1) + 5, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathbf{v}_{k,l,i,6}^{(m)} = \begin{cases} -1, & \text{if } m = 7K(i-1) + 7(k-1) + 1, \\ 1, & \text{if } m = 7K(i-1) + 7(l-1) + 6, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathbf{v}_{k,l,i,7}^{(m)} = \begin{cases} -1, & \text{if } m = 7K(i-1) + 7(k-1) + 1, \\ 1, & \text{if } m = 7K(i-1) + 7(l-1) + 7, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathbf{v}_{k,l,i,j,8}^{(m)} = \begin{cases} -1, & \text{if} m = 7K(i-1) + 7(k-1) + 1, \quad \text{for } j = 1, \dots, J, \\ 1, & \text{if } m = 7K(i-1) + 7(l-1) + 2, \quad \text{for } j = i, \\ 1, & \text{if } m = 7K(j-1) + 7(l-1) + 4, \quad \text{for } j \neq i, \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathbf{v}_{k,l,i,j,9}^{(m)} = \begin{cases} -1, & \text{if } m = 7K(i-1) + 7(k-1) + 1, \quad \text{for } j = 1, \dots, J, \\ 1, & \text{if } m = 7K(i-1) + 7(l-1) + 3, \quad \text{for } j = i, \\ 1, & \text{if } m = 7K(j-1) + 7(l-1) + 5, \quad \text{for } j \neq i, \\ 0, & \text{otherwise.} \end{cases}$$

In this case:

$$\begin{aligned} \alpha_t(\mathbf{z}(t)) \equiv & \sum_{k=1}^{K}\sum_{i=1}^{J} \alpha_{k,i,1}(\mathbf{z}(t))\mathbf{v}_{k,i,1} + \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{J}\sum_{p=2}^{7} \alpha_{k,l,i,p}(\mathbf{z}(t))\mathbf{v}_{k,l,i,p} \\ & + \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{q=8}^{9} \alpha_{k,l,i,j,q}(\mathbf{z}(t))\mathbf{v}_{k,l,i,j,q}, \\ \mathbf{B}_t = & \sum_{k=1}^{K}\sum_{i=1}^{J} \alpha_{k,i,1}(\mathbf{z}(t))\mathbf{v}_{k,i,1} \otimes \mathbf{v}_{k,i,1} \\ & + \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{J}\sum_{p=2}^{7} \alpha_{k,l,i,p}(\mathbf{z}(t))\mathbf{v}_{k,l,i,p} \otimes \mathbf{v}_{k,l,i,p} \\ & + \sum_{k=1}^{K}\sum_{l=1}^{K}\sum_{i=1}^{J}\sum_{j=1}^{J}\sum_{q=8}^{9} \alpha_{k,l,i,j,q}(\mathbf{z}(t))\mathbf{v}_{k,l,i,j,q} \otimes \mathbf{v}_{k,l,i,j,q}. \end{aligned}$$

Both A_t and B_t are matrices of dimension $7KJ \times 7KJ$. For $k \in H$, $i \in Ser$, let $u = 7K(i-1) + 7(k-1) + 1$. For $m = 1, \dots, 7KJ$ define:

$$a_t^{(m,u)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,Q,i}(t)}, \quad a_t^{(m,u+1)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,R,i}(t)}, \quad a_t^{(m,u+2)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,U,i}(t)},$$

$$a_t^{(m,u+3)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,A,i}(t)}, \quad a_t^{(m,u+4)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,O,i}(t)},$$

$$a_t^{(m,u+5)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,L,i}(t)}, \quad a_t^{(m,u+6)} = \frac{d\alpha_t^{(m)}(\mathbf{z}(t))}{dz_{k,D,i}(t)}.$$

The above system of equations provides the diffusion limit for approximating the variance seen in stochastic system (1).

3.3 Virtual waiting time

A method for calculating the VWT for each service in (1) is now presented, adapting the method of Mandelbaum et al. (2002).

Definition 1 For an infinitely patient "virtual customer" arriving to the service and queue at a fixed time τ, $T > \tau \geq 0$, their **virtual waiting time** (VWT) is how long they wait until their service begins. This is denoted: $VWT_{k,i}(\tau)$ for each $i \in Ser$ and $k \in Cla$.

Due to the parallel queues and multiple services, to calculate the VWT over $[0, \infty)$, the following assumptions are required:

1. $c_{k,i}(\mathbf{z}(t))$ are continuously differentiable with respect to time;
2. All $\mu_{k,i}(t)$ are continuous;
3. $\delta_{k,R,i}(t)$, $\delta_{k,U,i}(t)$, $\delta_{k,A,i}(t)$, $\delta_{k,O,i}(t)$ and $\theta_{k,i}(t)$ are bounded on compact intervals.

The first assumption places the restriction that all input parameters are continuous, unless the capacity allocation is independent of $\mathbf{z}(t)$).

To calculate the VWT at time $\tau > 0$, (9)–(15) are modified giving a new system denoted $\mathbf{Z}^*$. For $\tau > t \geq 0$, $\mathbf{Z}^*(t) = \mathbf{Z}(t)$; thus, Theorem 1 and the diffusion limit still hold in this time period. For $t > \tau$, the time after a virtual customer has arrived, the process differs as follows:

1. There are no external arrivals, rejoins, reuses, uses of alternative service or arrivals from other services;
2. Only customers remaining in the queue and service are served after τ;
3. Any customer departing the service and queue process leaves the system;
4. There are no class transitions after τ.

Importantly, these assumptions simplify the calculation of the VWT, as each queue behaves independently of each other for $t > \tau$. Therefore, the VWT can be solved independently for each queue as in Mandelbaum et al. (2002).

3.4 Production of service outcomes

Within a given time period, (9)–(15) may be adapted to measure the production of service outcomes—the number of customers who leave a service at a point in time, in a given class. This includes those who leave after completing service, $P_{k,E,i}(t)$; leave the system as a loss, $P_{k,L,i}(t)$; remain in the system having completed service, $P_{k,S,i}(t)$; or remain in the system having abandoned the queue, $P_{k,A,i}(t)$. Thus, over a period of time $[t_s, t_e] \subseteq [0, T]$ the production of customers in class $k \in Cla$ from a service $i \in Ser$ is:

$$P_{k,E,i}(t) = \sum_{l=1}^{K} MR_{l,S,i}^{(J+1)}(t_e) - MR_{l,S,i}^{(J+1)}(t_s),$$

$$P_{k,L,i}(t) = \sum_{l=1}^{K} MR_{l,L,i}^{(J+1)}(t_e) - MR_{l,L,i}^{(J+1)}(t_s),$$

$$P_{k,S,i}(t) = \sum_{l=1}^{K}\sum_{j=1}^{J} MR^{(j)}_{l,S,i}(t_e) - MR^{(j)}_{l,S,i}(t_s),$$

$$P_{k,A,i}(t) = \sum_{l=1}^{K}\sum_{j=1}^{J} MR^{(j)}_{l,L,i}(t_e) - MR^{(j)}_{l,L,i}(t_s).$$

with the fluid approximation of:

$$p_{k,E,i}(t) = \int_{t_s}^{t_e} \sum_{l=1}^{K} r_{k,S,i,J+1}(u) s_{l,k,S,i}(u) \mu_{l,i}(u) \min\big(z_{l,Q,i}(u), c_{l,i}(\mathbf{z}(u))\big)\, du,$$

$$p_{k,L,i}(t) = \int_{t_s}^{t_e} \sum_{l=1}^{K} r_{k,L,i,J+1}(u) s_{l,k,L,i}(u) \theta_{l,i}(u) \big(z_{l,Q,i}(u) - c_{l,i}(\mathbf{z}(u))\big)^{+}\, du,$$

$$p_{k,S,i}(t) = \int_{t_s}^{t_e} \sum_{l=1}^{K}\sum_{j=1}^{J} r_{k,S,i,j}(u) s_{l,k,S,i}(u) \mu_{l,i}(u) \min\big(z_{l,Q,i}(u), c_{l,i}(\mathbf{z}(u))\big)\, du,$$

$$p_{k,A,i}(t) = \int_{t_s}^{t_e} \sum_{l=1}^{K}\sum_{j=1}^{J} r_{k,L,i,j}(u) s_{l,k,L,i}(u) \theta_{l,i}(u) \big(z_{l,Q,i}(u) - c_{l,i}(\mathbf{z}(u))\big)^{+}\, du.$$

This measure can help to understand how different capacity allocations and changes in time varying systems may affect the output of customers in certain classes from a system, and the system's impact on customers' classes.

4 Exploration of the accuracy of fluid–diffusion approximations

The accuracy of fluid–diffusion approximations have been widely discussed within the literature, including the work by Mandelbaum et al. (1998), Mandelbaum et al. (2002), Ko and Gautam (2013) and Remerova (2014); Ding et al. (2015). It is known that these methods are increasingly accurate for heavily loaded queues i.e. when long run demand is greater than a service's capacity such that queues grow infinitely long (when demand is not inhibited). Ding et al. (2015) identified that for queues with customers who rejoin or reuse a service, an effective traffic intensity (ETI) is required to account for the increased demand created by these re-entrant customers: $\tilde{\rho} = \frac{\lambda}{c\mu(1-q)}$, where q is the probability that a customer seeks to reuse a service.

An illustrative analysis of the system is now presented to show how the accuracy of the approximations developed in this paper is affected by multiple classes, dynamic server allocations and several services. Of particular concern is when the system is effectively heavily loaded. The analysis and insights gained add to those published in Mandelbaum et al. (2002) and Ding et al. (2015) since: the effect of changes in a range of different parameters is considered, the analysis is conducted over a split time interval, and time dependent behaviour is modelled. Analysing the system over two time intervals shows the accuracy of the model during queue formation and as the system reaches steady state, informing on the accuracy of modelling time varying behaviour. This also mitigates the bias that the length of the modelled time period introduces (discussed below).

Only hypothetical examples of the models were used since the purpose of these investigations is to find the pragmatic constraints that the need for heavy loading places on the input parameters. Hence several scenarios are now explored to test the approximations and identify key limitations.

The accuracy of the fluid–diffusion approximations is evaluated in comparison to the averaged solution of a discrete event simulation (DES) of the stochastic system, gained from 1000 runs. Due to the lack of comparable data, the simulated solution is taken to be "true". The fluid–diffusion approximation and relevant simulations were computed on a node with a Windows 10 operating system, a 2.4 GHz quad-core processor and 8 GB RAM.

The error is calculated for the average number of customers in each process state, the VWT and the variance of each. For $m \in \{Q, R, L, U, A, O\}$; $p, q \in \{0, 1, \ldots, T/dt\}$; $p < q$ such that $t_p = dt \times p$:

$$z_{err,m}(t_q - t_p) = \frac{\sum_{r=p}^{q} \mathbb{E}Z_m(t_r) - z_m(t_r)}{\sum_{r=p}^{q} \mathbb{E}Z_m(t_r)}, \tag{25}$$

$$z_{err,m}^{var}(t_q - t_p) = \frac{\sum_{r=p}^{q} \mathrm{Var}(Z_m(t_r)) - \mathrm{Var}(z_m(t_r))}{\sum_{r=p}^{q} \mathrm{Var}(Z_m(t_r))}, \tag{26}$$

$$VWT_{err}(t_q - t_p) = \frac{\sum_{r=p}^{q} WT^{\mathrm{Sim}}(t_r) - \mathrm{VWT}(t_r)}{\sum_{r=p}^{q} WT^{\mathrm{Sim}}(t_r)}, \tag{27}$$

$$VWT_{err}^{var}(t_q - t_p) = \frac{\sum_{r=p}^{q} \mathrm{Var}(WT^{\mathrm{Sim}}(t_r)) - \mathrm{Var}(\mathrm{VWT}(t_r))}{\sum_{r=p}^{q} \mathrm{Var}(WT^{\mathrm{Sim}}(t_r))}. \tag{28}$$

where WT^{Sim} indicates the waiting time gained from the simulation. Notably, the waiting time is computed for each simulated queue (rather than a simulated VWT) to determine whether and under what conditions the VWT is an reasonable measure of actual waiting time. This is because, in practice, actual waiting time is often the true metric of interest in service systems.

For models that begin with $z_Q(t) < c(t)$, two distinct phases occur within the solution relating to when the queues form and when they stabilise. In a real world system with available serves, new arrivals immediately enter service until the system reaches a critical point $z_Q(t) = c(t)$. When $z_Q(t) >= c(t)$ subsequent arriving customers form a queue from which they may abandon. Due to random variation in the arrival process for a stochastic system, the existence and size of the queue fluctuates in time such that abandonment may occur throughout the whole time frame. However, since the fluid approximation is deterministic, this variation does not occur. Instead, there is no queue or loss within the fluid system until the critical point is reached. This delay causes an initial inaccuracy.

The error may then diminish as the system reaches steady state, hence the size of T then affects the error measurements. As a result two errors are produced for the system by splitting the modelled time interval to mitigate this bias. Firstly, errors are calculated as the queue forms in the fluid approximation—for $[0, T_I]$, where $T_I = \max\{t + 1 | z_Q(t) \leq c(t)\}$—denoted as the "formation error". From this an initial error and the length of time over which this error occurs (T_I) are both gained, providing understanding of how the size of the system and the ETI affects the system. Secondly, errors are calculated for the remaining time period: $(T_I, T]$. T is set so that the system reaches steady state.

Now presented are three hypothetical examples to illustrate how the extensions incorporated in this work alter the understanding of when the system is heavily loaded. The first is a steady state case for two customer classes and a constant server allocation is used here. The aim is to show how class transitions affect when the queues become heavily loaded. Building on this, the second example introduce class and time dependent parameters and thus a

Table 1 Parameters used to assess the accuracy of the approximations—steady state analysis of a single service and two classes

Parameters	Class	
	$k = 1$	$k = 2$
μ_k	1	1
θ_k	1	1
λ_k	20	20
c_k	20	20
$r_{k,L,1,1}$	0.3	0.3
$r_{k,S,1,1}$	0.3	0.3
$\delta_{k,R}$	1	1
$\delta_{k,F}$	1	1

dynamic server allocation, again for two classes. This example is intended to show further how the extensions in this paper affect the accuracy of the model, particularly the definition of effective heavy loading. From the two cases pragmatic constraints are highlighted regarding the accuracy of this method, each of which inform the considerations that should be made in seeking to apply this method. Finally, a larger, multi-service example is shown to illustrate the findings gained from these two scenarios.

4.1 Single service and two classes: steady state analysis

This first steady state example uses a constant and equal allocation of servers across queues and two customer classes $k = 1$ or $k = 2$, to show how the accuracy of the approximations is affected by the multiple classes, class transitions and class dependent parameters. The relevant parameters are set out in Table 1. Notably, in this example, customers from either class have the same input parameters, such that the value of $\tilde{\rho}$ given by Ding et al. (2015) is the equivalent for both queues. By modelling two groups with the same input parameters, the effect that class transitions have on the accuracy of the approximations can be understood, adding to the understanding provided by the aforementioned papers.

$$S_S = \begin{bmatrix} 0.3, & 0.7 \\ 0, & 1 \end{bmatrix}, S_L = \begin{bmatrix} 1, & 0 \\ 0.6, & 0.4 \end{bmatrix}, S_R = \begin{bmatrix} 0.8, & 0.2 \\ 0.5, & 0.5 \end{bmatrix}, S_F = \begin{bmatrix} 0.8, & 0.2 \\ 0.2, & 0.8 \end{bmatrix}.$$

The transition matrices are defined such that $k = 2$ is representative of a preferable class (despite the lack of difference between their flow parameters). Thus it is more likely that customers more to state $k = 2$ after receiving service denoting a potentially beneficial, but not perfect, effect. Abandonment however has a similar reverse effect. Customers seeking to rejoin may have an improvement in their class after rejoin—representing a delayed benefit of service. Finally, for those who reuse the service, their class is assumed to change or stay the same after their time in the relevant orbit but are more likely to remain in the class in which they entered. These matrices highlight the differences in customers' capacities to benefit from service given the receipt or absence of service.

Table 2 presents the errors for this system over the formation period—$[0, T_I)$, and the error thereafter—$[T_I, 15]$. There is a clear difference in the accuracy of the approximations for the two classes with more accurate solutions for both $z_R(t)$ and VWT for $k = 2$ and $t > T_I$. This indicates that the $k = 2$ queue is "more effectively heavily loaded" since it is well known that the accuracy of these outputs increases for more heavily loaded queues

Table 2 Error between the approximations and simulation as a percentage of the simulated solution—parameters from Table 1

Parameters		z_{err}	z^{var}_{err}	VWT_{err}	VWT^{var}_{err}
$k = 1$	Q	0.34	2.12	75.32	64.77
	R	63.00	37.00	–	–
$t \leq T_I = 3.4$	U	2.60	3.97	–	–
$k = 1$	Q	0.46	1.45	12.03	6.30
	R	4.24	1.95	–	–
$t > T_I = 3.4$	U	2.67	2.33	–	–
$k = 2$	Q	0.24	1.79	52.56	54.32
	R	65.50	47.85	–	–
$t \leq T_I = 2.9$	U	3.48	4.22	–	–
$k = 2$	Q	0.42	1.30	6.71	1.55
	R	2.05	2.06	–	–
$t > T_I = 2.9$	U	1.41	1.59	–	–

Table 3 Parameters used to assess the accuracy of the approximations—time-varying analysis of a single service and two classes

Parameters	Class	
	$k = 1$	$k = 2$
μ_k	1/2	1
θ_k	1	1/2
λ_k	Below	15
p_k	0.5	0.3
q_k	0.5	0.3
$\delta_{k,R}$	1	1/2
$\delta_{k,F}$	1	1/2

(Mandelbaum et al. 2002; Ding et al. 2015). Furthermore, there is a difference between the length of the formation periods for the two class queues. A smaller T_I indicates that the system has a higher ETI since the queue grows faster due to a higher effective arrival rate. Thus, this example shows that the ETI is no longer only dependent on customers who reuse a service and ρ [as in Ding et al. (2015)], since, by their definition, $\tilde{\rho}$ is the same for both groups.

A reason for the difference in ETI for this system is that customers may join a queue for the class they did not arrive in through either the reuse or rejoin orbits. In Ding et al. (2015), rejoining customers would be captured by λ in the steady state system; now however, this is no longer true. Instead, arrivals at one queue may affect the others, such that class transitions are influential when formulating the ETI with multiple classes. As a final observation, the calculation of the approximations was over 250 times faster than the simulation with a CPU time for the simulation of 502.98 s and 1.79 s for the fluid–diffusion approximation.

4.2 Time varying analysis: dynamic server allocation

Extending the previous scenario, a time-varying system is now modelled with non-empty initial condtions ($z_{k,Q}(0) = 15$ for $k = 1, 2$) and two classes have different flow parameters, see Table 3. Customers in class $k = 1$ now have longer service times, a higher propensity to abandon, a higher likelihood of rejoin or reuse, and require sequential service sooner—representing more resource intensive customers. Here, a small spike in the

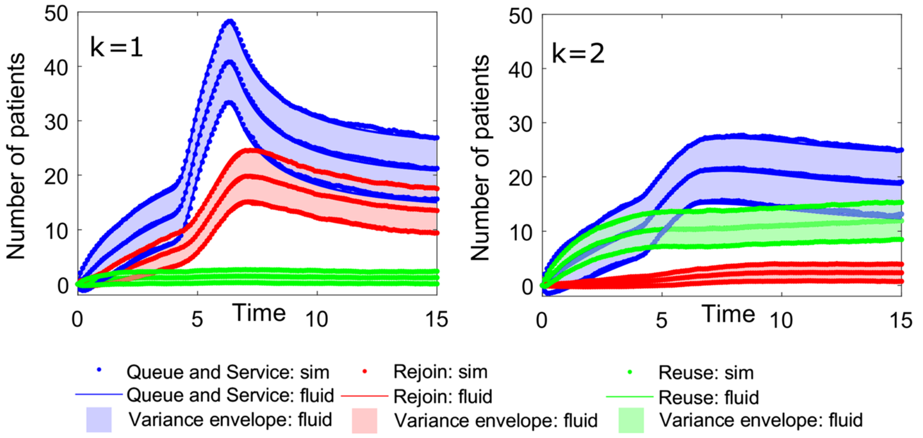

Fig. 2 Number of customers in process states with corresponding variance envelopes—two classes with dynamic server allocation, parameters from Table 3

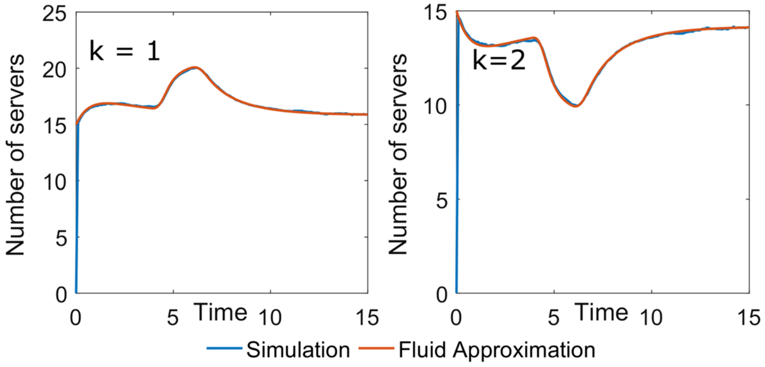

Fig. 3 Number of servers allocated to each queue over time—two classes with dynamic server allocation, parameters from Table 3

arrivals of customers in class $k = 1$ is also considered. Hence the proportional dynamic allocation of servers, Eq. (23), is employed with $c(t) = 30\ \forall t \in [0, 15]$. Since the input parameters are required to be continuous to ensure that $z_Q(t)$ is continuously differentiable, a continuous jump in arrivals is defined.

$$\lambda_1(t) = \begin{cases} 15, & t \in [0, 4) \cap [7, 15], \\ 15 + 15 \times (\sin(\pi(t-4) - \frac{\pi}{2}) + 1), & t \in [4, 5), \\ 45, & t \in [5, 6), \\ 15 + 15 \times (\sin(\pi(t-6) + \frac{\pi}{2}) + 1), & t \in [6, 7). \end{cases}$$

$$S_S = \begin{bmatrix} 0.3, & 0.7 \\ 0, & 1 \end{bmatrix}, S_L = \begin{bmatrix} 1, & 0 \\ 0.6, & 0.4 \end{bmatrix}, S_R = \begin{bmatrix} 0.8, & 0.2 \\ 0.5, & 0.5 \end{bmatrix}, S_F = \begin{bmatrix} 0.8, & 0.2 \\ 0.2, & 0.8 \end{bmatrix}.$$

Figure 2 shows that the approximations are accurate throughout the modelled time period for both the number of customers in each process orbit and the variance. Notably, the increased arrivals for $k = 1$ has little visual impact on the queue for $k = 2$, however there is a subtle effect. This is made clearer by the effect on the dynamic allocation of servers, shown in

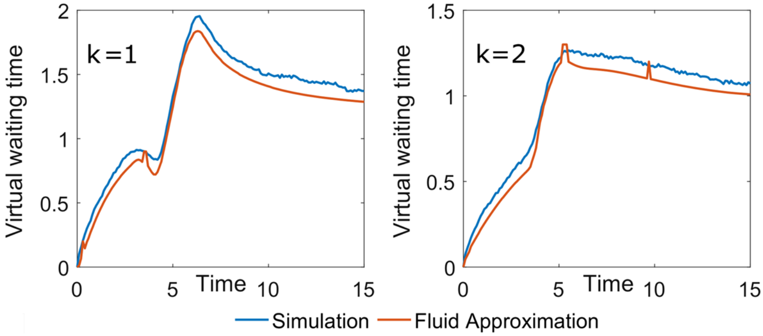

Fig. 4 VWT for two class system with dynamic server allocation, parameters from Table 3

Fig. 3. The influx of $k = 1$ customers causes an increase in the number of servers allocated to $k = 1$—hence, a longer queue exists for both customer classes. This results in a raised rate of abandonment for both classes of customer. Thus, since customers who abandon from either class are more likely to rejoin in class $k = 1$, there is a further increase the $k = 1$ demand. This is further confirmed by Fig. 4. For $k = 2$ the gradient of the VWT increases at $t = 4$, reflecting the increased queue lengths and longer waits that occurs due to the loss in allocated servers to $k = 1$. Additionally, there is a large increase in VWT for $k = 1$. Whilst this queue gains more servers, causing an initial dip, the increase in new arrivals and $k = 1$ rejoin customers raises the expected waiting time. There is a discrepancy between the two results in Fig. 4. Primarily this is due to a comparison being made between simulated waiting time and the VWT here. Qualitatively the VWT provides a good understanding of the waiting time profile matching the result with reasonable accuracy.

Considering the variance of the VWT, Fig. 5, the fluid–diffusion approximations match the behaviour but fail to capture the size of the simulated solution. For increased size and reuse, the results may improve. However, since the variance of the simulated waiting time has the most variability of the system outputs, when combined with the variability of the dynamic server allocation, this increases the inaccuracy. Further investigation would be valuable.

Finally, the production of outcomes, measured by the rate at which customers in each class leave the system over time, is affected—Fig. 6. The number of customers lost due to abandonment and are in the worst class $k = 1$ greatly increases, whilst the number of served customers leaving in class $k = 2$ decreases and the number lost in $k = 2$ increases. This is understandable due to the reduced service of $k = 2$ customers.

The interaction between the queues and these additional outputs are helpful for understanding the "flow of outcomes", in particular, how a service produces good and bad outcomes over time in light of customer mix, demand, available/allocated capacity and flow dynamics. The example above highlights the need for considered server allocations since an influx and preference towards $k = 1$ considerably affects the output and operational performance of the system—indicating a negative "flow of outcomes". This provides a perspective on the quality of service and operation of the system in relation to process outcomes (such as customer throughput and number of abandonments) and how the differing needs of customers impact the system.

Of note, the simulation for 1000 runs took 601.49 s to run, whilst the fluid–diffusion took 4.75 s to run. Additionally, there are small discontinuities in Figs. 4 and 5 of order dt and lasting for an amount of time of a similar order. To investigate these discontinuities, results

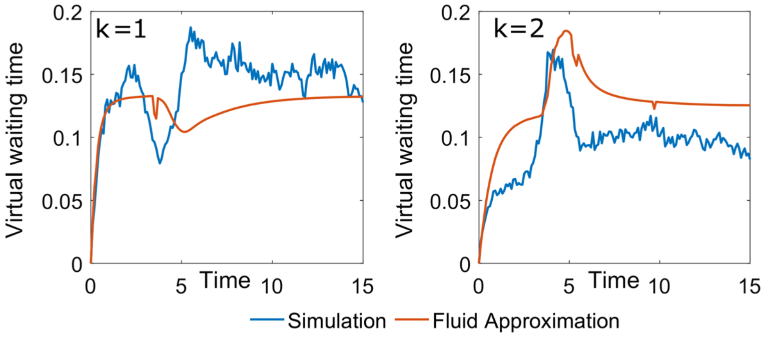

Fig. 5 Variance of VWT for two class system with dynamic server allocation, parameters from Table 3

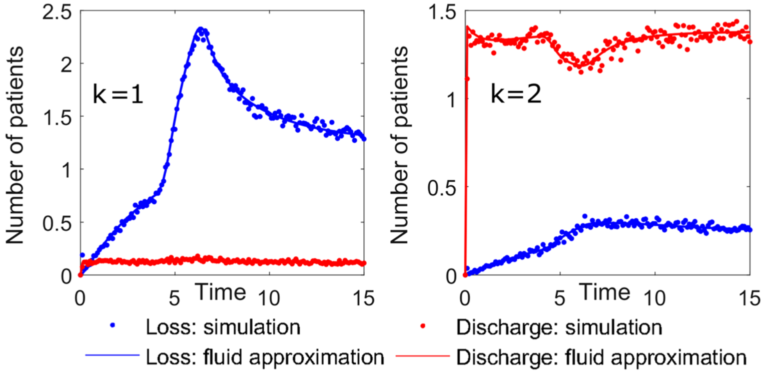

Fig. 6 The production of service outcomes for two class system with dynamic server allocation, parameters from Table 3

are compared for a scenario with $dt = 0.1, 0.05$ and 0.01, Fig. 7. Clearly, as dt decreases, the size of the jumps decrease; however, their frequency increases.

These discontinuities may be a result the fluid approximation becoming non-linear when using a dynamic server allocation [the combination of (16)–(22) with (23) or (24)]. Ultimately, these errors do not have a significant impact on the solution of the VWT or its variance since they last for short time intervals with potentially small magnitude, Fig. 7. By decreasing dt, the size of the errors reduces; however, the time required to solve the numerical scheme increases, in Table 4, creating a trade off between usability and accuracy. When the errors are small compared to the overall solution, there is little benefit in reducing dt.

Summary of single service and multi-class models

From this brief exploration, it has been shown that the appropriateness and applicability of the approximations is maintained when extending to multiple classes with class transitions. Importantly, there is an additional influence of the transition matrices on the ETI since services now have an effective arrival rate comprising new customers, reuse customers and rejoins, including those who previously queued within another class.

This is important for systems where reuse is low for a particular class, since these methods may be accurately applied if there is a significant flow of customers arriving from other classes. Therefore, the ETI for each $k \in Cla$ queue, when considering multiple classes, is dependent

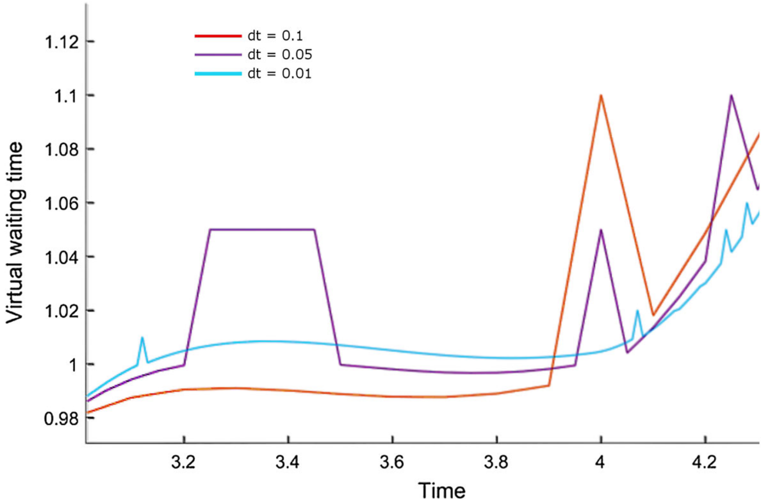

Fig. 7 Example of small errors in the solution of the VWT when using a dynamic server allocation. These errors are of order dt

Table 4 Time taken to solve the fluid approximation for different sizes of dt

dt	0.1	0.05	0.01
Time (s)	6.72	25.50	605.06

on the combination of: $\lambda_k(t)$, $c_k(\mathbf{z}(t))$, $\mu_k(t)$, $\mathbf{S}_{k,m}(t)$, $q_k(t)$ and $p_k(t)$, $\forall k \in Cla$ and $m \in \{S, L, R, U\}$.

4.3 Extending to multiple services

The analysis presented above indicates the parameters that determine accuracy of the fluid–diffusion model in comparison to simulation. Notably these findings hold when applying both the approximations and the simulation to larger systems since this is equivalent to modelling an amalgamation of these smaller models. As such, the modelling of larger systems may be implemented through a modular programming of the code which would increase its flexibility and scalability for modelling these scenarios.

By introducing multiple services, the flow dynamics of the other service orbits and alternative service orbits are introduced. A new comparison with simulation is not required to understand the accuracy of the systems since it is fundamentally similar to the previous. Rather, any further changes to how the ETI is understood may be inferred from the model's structure.

In the analysis of multiple services, reuses and rejoins are governed by $r_{k,m,i,i}$, $m \in \{S, L\}$ respectively. Since customers may use other services after completing service, or use alternative services having abandoned, $r_{k,m,i,j}$, $m \in \{S, L\}$, $j \neq i$ may be small for systems of multiple services. However, customers may now arrive from other/alternative services, increasing the number of arrivals to each queue.

Thus, in considering the effective traffic intensity of a service in the network, alongside the parameters previously noted, the values of $r_{k,m,i,j}$, $m \in \{S, L\}$, for all $k \in Cla$; $i, j \in Ser$ should also be considered, helping to understand when the approximations are accurate for the multiple service extension. Thus, in such scenarios, the size and value of $\mathbf{S}_{k,m,i}$, $\mathbf{R}_{k,Q,i}$, $\mathbf{R}_{k,L,i}$, for all $k \in H,\ i \in Ser$ and $m \in \{S, L, R, U, A, O\}$ may combine to increase the model's accuracy.

To illustrate the application to a larger system, a fluid–diffusion approximation for a three service and three class system with has all the dynamics described in Fig. 1 is now analysed. The input parameters used to populate this example are provided in the supplementary material. Service 1 is modelled to be likely to serve customers in classes considered to be worse and represent services that are short in length. From service 1 customers may then use service 2 or 3 depending on their needs. Service 2 has longer service durations and may serve customer in any class, whilst service 3 has the longest service duration and typically serves customers in classes that are considered to be better. Furthermore, a customers class is considered to improve only through service, and may decline in between service.

This scenario highlights how the model may be used to represent a system of diverse services that each have a different purpose, type of service (indicated by service rate) and customer mix. Notably, given its small initial condition and arrival rate, the effective traffic intensity for service 3 is significantly increased by the flow of customers from other services. Figure 8 shows the number of customers in each process orbit, the variance is not shown to improve the readability of the figure. Figure 9 gives the VWT and its variance for each queue, whilst Figure 10 shows the dynamic capacity allocation for each class and service in the system.

Figure 11 illustrates the benefits of the production output in this scenario. For each service the output of customers in different classes over time is given by the loss and service completion curves. Additional curves correspond to customers who remain in the system having completed service or abandoned the queue for each service. Together, these plots provide greater insight into the flow of customers and service outcomes in the system and may be used to identify negative and positive patterns of flow. For example, whilst service 3 has the highest rate of customers leaving the system in good classes, there is a significant flow of customers from other services who are in good classes. Thus, this service does not achieve good service outcomes in isolation. Rather, this shows how services combine to produce good service outcomes as customers participate in multiple interactions and use several services.

5 Summary and discussion

This paper makes two contributions to how multi-class service networks may be modelled and their performance measured—particularly when service quality is important. The first contribution is the extension of current fluid–diffusion approximation methods (Mandelbaum et al. 2002; Ding et al. 2015) to include multiple classes, class transitions and dynamic server allocations. These approximations provide an efficient method for modelling systems of queues with several complex flow dynamics including: the sequential use of multiple services, abandonment, rejoin, reuse, multiple classes, and class and time dependent parameters. Importantly, there is a dependency between overall demand and system capacity such that these dynamics introduce a feedback loop of delayed demand. For example, having arrived, queued and completed service, a customer seeking to reuse will wait for a period of time before re-entering the queue. Understanding the effect of these flow dynamics, and how resources may be managed in light of them, is important since ignoring them may lead to under or over staffing in scenarios where they are significant.

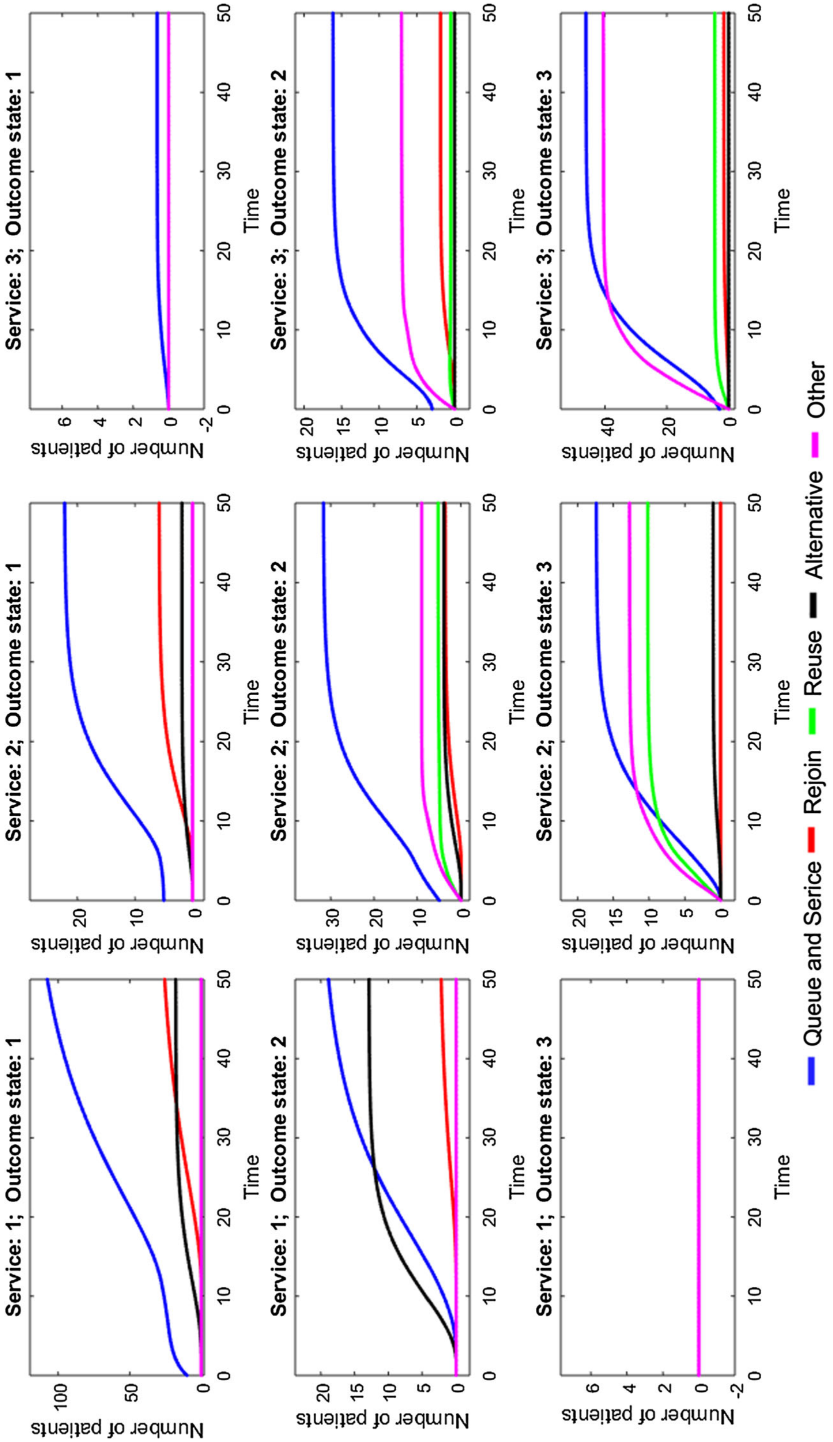

Fig. 8 Three service and three class system—number of customers in each process state

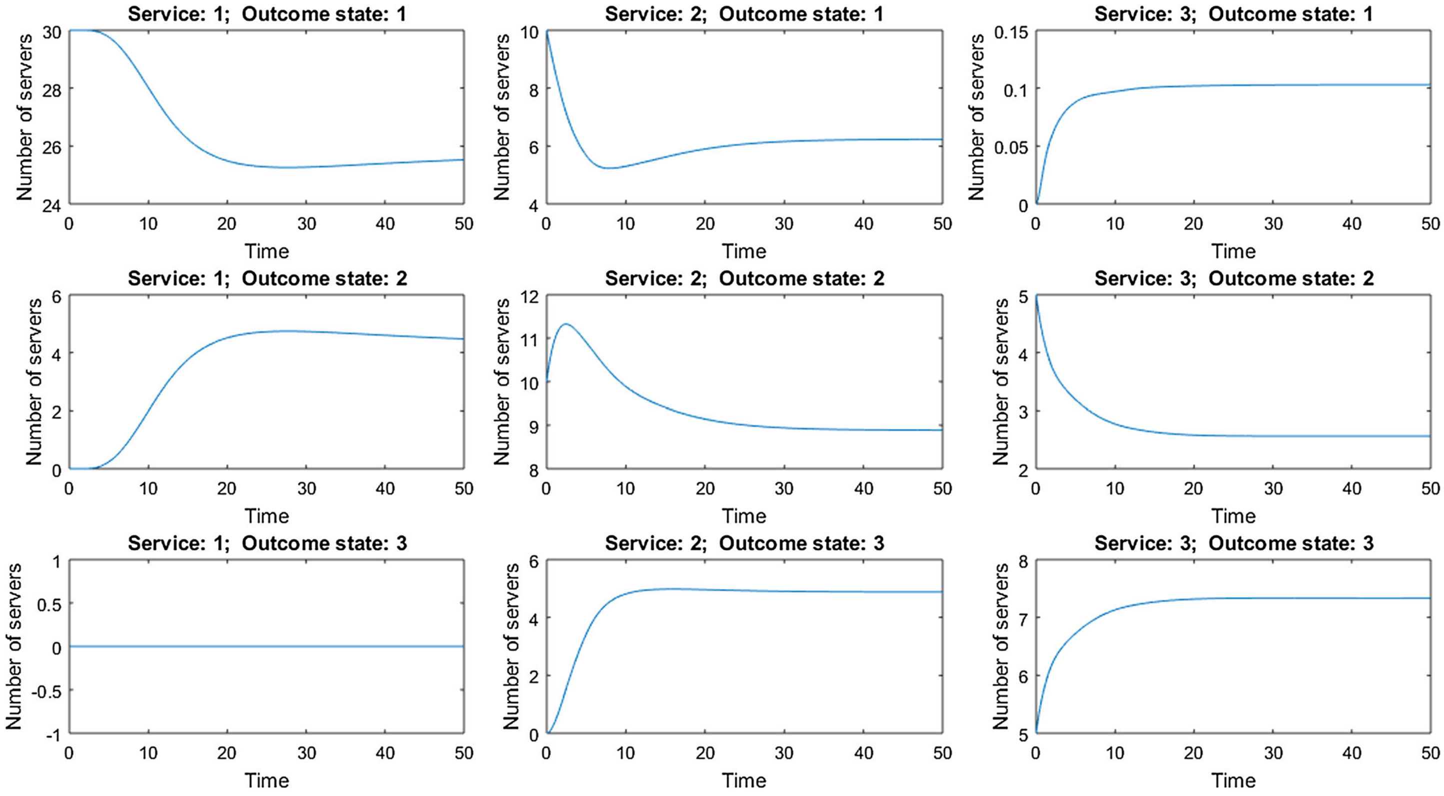

Fig. 9 Three service and three class system—dynamic server allocation

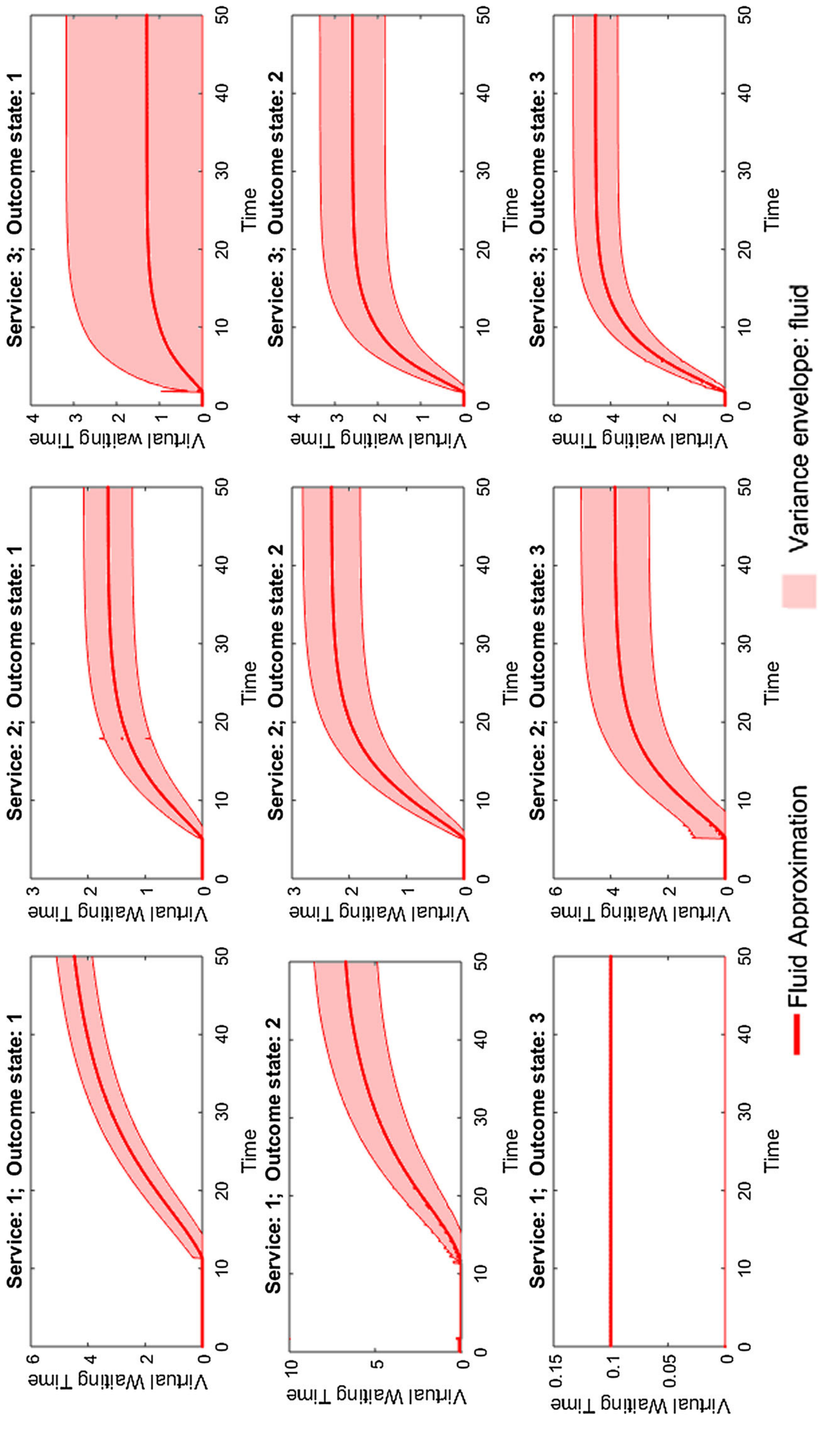

Fig. 10 Three service and three class system—the VWT and variance

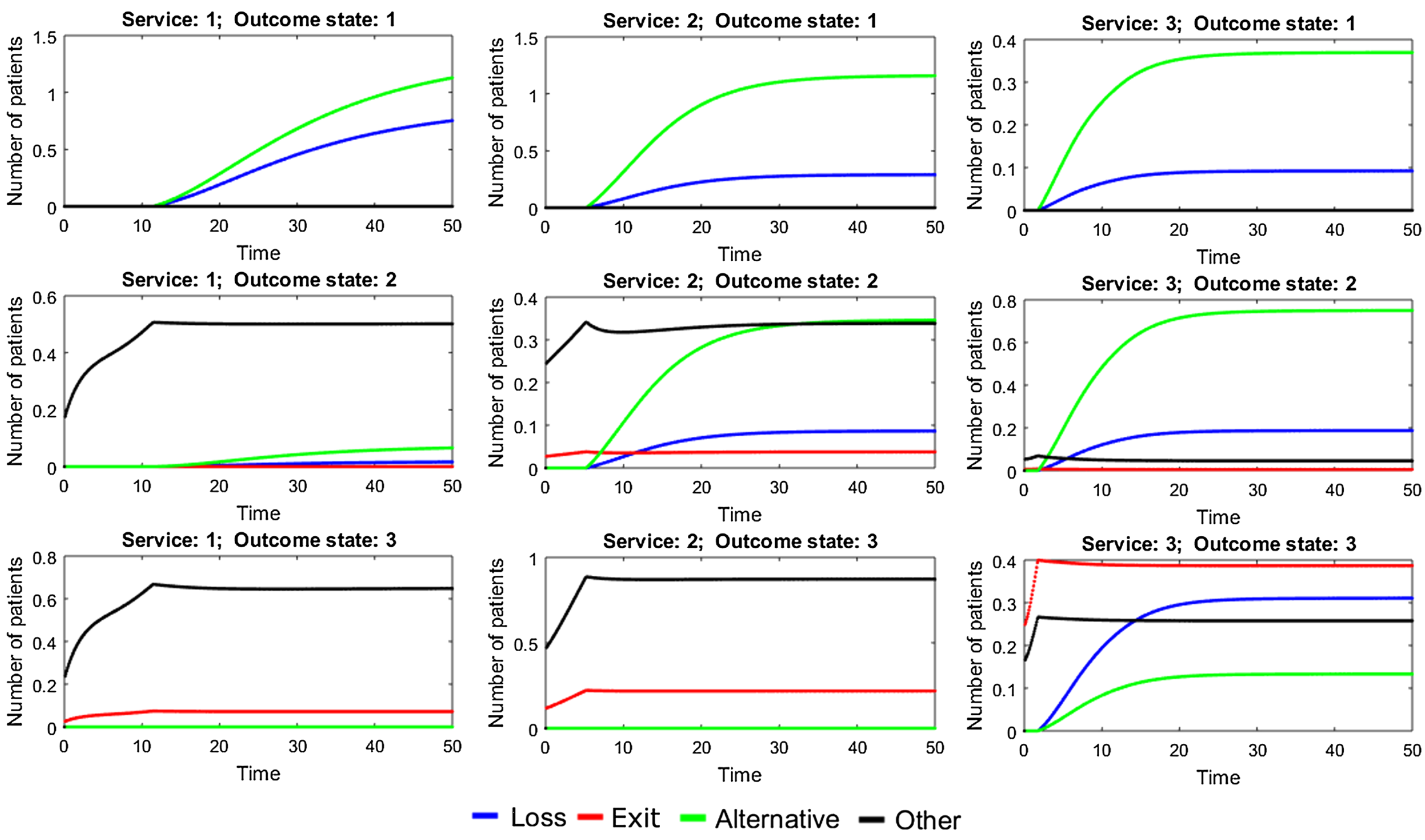

Fig. 11 Three service and three class system—the production of service outcomes

Whilst parallel queues are traditionally inefficient due to the possibility of inactive servers, this limitation is overcome by using a dynamic server allocation. Since servers are continuously reallocated, they cannot become inactive if there are customers in any class waiting for the service. Thus the benefits of using multiple queues to represent different classes may be fully utilised. That is, the flow of customers with differing service requirements and different capacities to benefit from service can be modelled.

This leads to the second contribution, the measuring of system performance from a new perspective, denoted the “flow of outcomes”. Through the inclusion of multiple classes and transitions, the flows for customers with differing resource/service requirements and different capacities to benefit from service may be modelled. Importantly, this may reflect real life scenarios where customers with varying needs have markedly different interactions with a service.

Thus, the system’s performance can be understood by how individual services contribute to the system’s service outcomes and the system’s operational performance. In particular, the effect of service, or absence of it, on customers—as measured by service outcomes—and the effect of customers with different requirements—e.g. service times—on the operation of the system can both be understood. This is captured by the production measure which provides insight into the positive and negative effects of a system’s process outcomes on the quality/impact of service over time.

For example, in scenarios where access to services is poor and there is high abandonment, the possible negative impact on customers may be better understood by using these methods. For instance, having abandoned, customers may re-enter the queue in a “worse” class than before and thus require a more resource intensive service, increasing the future burden on the system. This represents a scenario of poor “flow of outcomes” and would result in a high production of customers in worse classes and fewer customers being produced in classes that reflect good service outcomes.

Alternatively, the combination of transitions in class, reuse and uses of other services, helps to understand how multiple service interactions combine to produce good service outcomes. For example, the receipt of service may affect a customer’s future use of services through a positive service impact, such as an improvement in class. In this case customers may require fewer interactions, reducing their future demand and the intensity of service needed, reflecting a positive “flow of outcomes”. This can be understood by using the production output as the time varying output of customers in better classes may be higher throughout time due to a combination of services and interactions, as in the example in Fig. 11.

5.1 Limitations

Fluid and diffusion approximations are most accurate for large and heavily loaded systems, potentially limiting when and how these approximations may be used Ko and Gautam (2013). Furthermore, in considering multiple services, several classes and time dependence, the method can become more complex and unwieldy due to the number of input parameters. Thus, editing and changing the inputs can be time consuming depending on the implementation, especially if several configurations of the system are analysed. As a result a careful implementation of these methods is required. One way to overcome this is to use a configurable interface or scheme for entering the input parameters.

Another potential limitation may occur when compared to data for real world systems given the strict Markovian assumptions. Thus, these methods may be better used for a stylistic representation of a system to help understand the dynamics of service networks and the consequences of changes in the system. Alternative methods such as simulation and system dynamics have a greater flexibility in this respect; however, may be hindered by runtime. Depending on the desired analysis and requirements of the model computation time may not be an issue; however, it will limit how a model may be used. For example, heuristic approaches or scenario analysis both require a wide range of scenarios to be computed such that computational time will determine how many iterations may be run. Thus, despite the limitations that the Markovian assumptions introduce, the speed and efficiency of fluid–diffusion approximations facilitate their use for larger, more intensive analysis. This is discussed further below.

5.2 Possible avenues for future work

It would be beneficial to explore further the use of "flow of outcomes" in understanding system performance. Having developed an illustrative method of the potential benefits in this paper, it would be insightful to apply these methods to the large, multi-service real world systems for which they were intended. Likewise it would be worthwhile to explore their benefits and limitations in comparison to other modelling methods, such as system dynamic and Markov chain approaches. One particular direction for this would be to explore the combination of these methods with optimisation and heuristic approaches given the speed of calculation and ODE representation of the system.

In particular, the flexibility in the definition of $C_{k,i}(\mathbf{Z}(t))$ and the inclusion of classes introduces the possibility for novel constraints and objectives, such as: how best to allocate servers to maximise the production of good service outcomes in a system; or, to minimise the flow of customers through patterns of service that lead to poor service outcomes and that increase flow problems. This may lead to new avenues of analysis and insight for service networks where both operational efficiency and the quality of the service are important.

Additionally, future extensions to this work include the relaxation of the Markovian assumptions to form a more generalisable approximation to overcome the current limitations. Recent work by Pender and Ko (2017) and Aras et al. (2018) are both promising in this regard. As such it would also be beneficial to also explore different parameter definitions. This includes mechanisms for loss that are dependent on the number of customers in different parts of the system, or the introduction of finite waiting space. As well as different definitions of the dynamic server allocations to increase the range of possible analyses such as heuristic optimisation, capacity allocation and priority queueing. Likewise, extending the method by Ko and Gautam (2013) to a network setting, with service reuse and customer classes would be another promising direction for future work. This especially so given the greater accuracy of their method in transient settings and their direct comparison to the methods presented in Mandelbaum et al. (1998, 2002).

Similarly, alternative definitions of the class transition process should be explored in particular those that include a time dependence such that customers may change class whilst continuing to reside in a given process state. Two possible directions are to investigate the use of fluid–diffusion approximations for feedback queue and the state-dependent queues, for example (Cheah and Smith 1994; Mitchell and Smith 2001; Zhu et al. 2017). In such cases, the capacity allocation functions C may be defined to reflect how upstream nodes are dependent on the traffic congestion of downstream nodes—hence reflecting longer and shorter waiting times. Likewise, the time dependent service rate and routing matrices for customers moving between outcome classes may also be defined to this end—especially with regards to state-dependent queues where the service rate of a queueing system is dependent on the number of customers in it. It is also possible that the fluid–diffusion approximation may be used to model queueing networks with time-dependent proportional routing (Liu and Whitt 2011), this would be a prudent direction for new work.

6 Conclusions

In this paper we have presented a method for modelling queues of heterogeneous customers who may change class throughout the service process. The development of these methods has made several contributions to the way in which networks of services may be modelled, and how system performance may be understood.

There is a methodological benefit since the approximations form set of ODEs that are efficient to solve, even as the system grows large, providing informative performance measures. These include: the number of customers within different classes, process orbits and services in the systems; the virtual waiting time for each service; the variance of each; and, the production of service outcomes. Furthermore, complex dynamics may be modelled using these methods, such as: customers reusing a service; future uses of other services; and, the potential for customers to abandon and potentially rejoin the queue or use another service.

This leads to the benefit that the combination of classes and flow provides: new avenues for insightful analysis within service systems. The methods highlight how two key perspectives of performance in service networks may be united in a single modelling framework. These methods may be used to help understand: how customers use services; the effect of multiple interactions on customer class; the effect of delayed demand/reuse of services on the operation of the system and on customer class; and, how a dependency between capacity of the system and the future arrival process affects the system.

Finally, by extending existing fluid–diffusion approximation methods, the scope for the application and use of these approximations has been increased for various settings.

Acknowledgements The first author was supported by the Health Foundation as part of the Improvement Science PhD programme (unique reference number: 7192). The Health Foundation is an independent charity committed to bringing about better health and health care for people in the UK. The second is supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care (CLAHRC) North Thames at Barts Health NHS Trust. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Aras, A. K., Chen, X., & Liu, Y. (2018). Many-server Gaussian limits for overloaded non-Markovian queues with customer abandonment. *Queueing Systems*, *89*, 1–45.

Ata, B. (2006). Dynamic control of a multiclass queue with thin arrival streams. *Operations Research*, *54*(5), 876–892.

Atar, R., Mandelbaum, A., Reiman, M. I., et al. (2004). Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *The Annals of Applied Probability*, *14*(3), 1084–1134.

Atar, R., Giat, C., & Shimkin, N. (2011). On the asymptotic optimality of the cμ/θ rule under ergodic cost. *Queueing Systems*, *67*(2), 127–144.

Cheah, J. Y., & Smith, J. M. G. (1994). Generalized M/G/c/c state dependent queueing models and pedestrian traffic flows. *Queueing Systems*, *15*(1–4), 365–386.

Chen, X., Wang, L., Ding, J., & Thomas, N. (2016). Patient flow scheduling and capacity planning in a smart hospital environment. *IEEE Access*, *4*, 135–148.

Cohen, I., Mandelbaum, A., & Zychlinski, N. (2014). Minimizing mortality in a mass casualty event: Fluid networks in support of modeling and staffing. *IIE Transactions*, *46*(7), 728–741.

Côté, M. J. (2000). Understanding patient flow. *Decision Line*, *31*(2), 8.

Deo, S., Iravani, S., Jiang, T., Smilowitz, K., & Samuelson, S. (2013). Improving health outcomes through better capacity allocation in a community-based chronic care model. *Operations Research*, *61*(6), 1277–1294.

Ding, S., Remerova, M., van der Mei, R. D., & Zwart, B. (2015). Fluid approximation of a call center model with redials and reconnects. *Performance Evaluation*, *92*, 24–39.

Federgruen, A., & Groenevelt, H. (1988). M/g/c queueing systems with multiple customer classes: Characterization and control of achievable performance under nonpreemptive priority rules. *Management Science*, *34*(9), 1121–1138.

Ghotbabadi, A. R., Feiz, S., & Baharun, R. (2015). Service quality measurements: A review. *International Journal of Academic Research in Business and Social Sciences*, *5*(2), 267.

Guo, Y. J. (2012). On the fluid approximation for a multiclass queue under non-preemptive SBP service discipline. *Acta Mathematica Sinica, English Series*, *28*(2), 379–404.

Hillston, J. (2005). Fluid flow approximation of PEPA models. In *Second international conference on the quantitative evaluation of systems* (pp. 33–42). IEEE.

Ko, Y. M., & Gautam, N. (2013). Critically loaded time-varying multiserver queues: Computational challenges and approximations. *INFORMS Journal on Computing*, *25*(2), 285–301.

Larrañaga, M., Ayesta, U., & Verloop, I. M. (2015). Asymptotically optimal index policies for an abandonment queue with convex holding cost. *Queueing Systems*, *81*(2–3), 99–169.

Liu, Y., & Whitt, W. (2011). A network of time-varying many-server fluid queues with customer abandonment. *Operations Research*, *59*(4), 835–846.

Maglaras, C. (1999). Dynamic scheduling in multiclass queueing networks: Stability under discrete-review policies. *Queueing Systems*, *31*(3–4), 171–206.

Mandelbaum, A., Massey, W. A., & Reiman, M. I. (1998). Strong approximations for markovian service networks. *Queueing Systems*, *30*(1–2), 149–201.

Mandelbaum, A., Massey, W. A., Reiman, M. I., Stolyar, A., & Rider, B. (2002). Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, *21*(2–4), 149–171.

Massey, W. A., & Pender, J. (2013). Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems*, *75*(2–4), 243–277.

Mitchell, D. H., & Smith, J. M. G. (2001). Topological network design of pedestrian networks. *Transportation Research Part B: Methodological*, *35*(2), 107–135.

Mukherjee, D., Dhara, S., Borst, S. C., & van Leeuwaarden, J. S. (2017). Optimal service elasticity in large-scale distributed systems. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, *1*(1), 25.

Palmer, R., Fulop, N. J., & Utley, M. (2017). A systematic literature review of operational research methods for modelling patient flow and outcomes within community healthcare and other settings. *Health Systems*, *7*, 1–21.

Pang, G., Talreja, R., Whitt, W., et al. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*, *4*, 193–267.

Pender, J., & Ko, Y. M. (2017). Approximations for the queue length distributions of time-varying many-server queues. *INFORMS Journal on Computing*, *29*(4), 688–704.

Pender, J., & Phung-Duc, T. (2016). A law of large numbers for m/m/c/delayoff-setup queues with non-stationary arrivals. In *International conference on analytical and stochastic modeling techniques and applications* (pp. 253–268). Springer, Berlin.

Remerova, M. (2014). Fluid limit approximations of stochastic networks. Ph.D. thesis

Seth, N., Deshmukh, S., & Vrat, P. (2005). Service quality models: A review. *International Journal of Quality and Reliability Management*, *22*(9), 913–949.

Tahar, A. B., & Jean-Marie, A. (2012). The fluid limit of the multiclass processor sharing queue. *Queueing Systems*, *71*(4), 347–404.

Tekin, S., Andradóttir, S., & Down, D. G. (2012). Dynamic server allocation for unstable queueing networks with flexible servers. *Queueing Systems*, *70*(1), 45–79.

Whitt, W. (2006). A multi-class fluid model for a contact center with skill-based routing. *AEU-International Journal of Electronics and Communications*, *60*(2), 95–102.

Yom-Tov, G. B., & Mandelbaum, A. (2014). Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing and Service Operations Management*, *16*(2), 283–299.

Zhang, Z. G., & Tian, N. (2004). An analysis of queueing systems with multi-task servers. *European Journal of Operational Research*, *156*(2), 375–389.

Zhu, J., Hu, L., Jiang, Y., & Khattak, A. (2017). Circulation network design for urban rail transit station using a PH (n)/PH (n)/C/C queuing network model. *European Journal of Operational Research*, *260*(3), 1043–1068.

Zychlinski, N., Mandelbaum, A., Momcilovic, P., & Cohen, I. (2017). Bed blocking in hospitals due to scarce capacity in geriatric institutions—Cost minimization via fluid models. Working paper (pp. 1–41).

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