Operations Research Letters 45 (2017) 248–253

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Diffusion limits for the $(MAP_t/Ph_t/\infty)^N$ queueing network

Young Myoung Ko [a], Jamol Pender [b,*]

[a] Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, Gyeongbuk, 37673, South Korea
[b] School of Operations Research and Information Engineering, Cornell University, 228 Rhodes Hall, Ithaca, NY 14853, United States

ARTICLE INFO

Article history:
Received 13 April 2016
Received in revised form
5 March 2017
Accepted 9 March 2017
Available online 16 March 2017

Keywords:
Markovian Arrival Processes
Infinite server queueing networks
Fluid and diffusion limits
Strong approximations
Time-varying rates

ABSTRACT

In this paper, we prove strong approximations for the $(MAP_t/Ph_t/\infty)^N$ queueing network. These strong approximations allow us to derive fluid and diffusion limits for the queue length processes of the network. This extends recent work that provides fluid and diffusion limits in the single station setting.

1. Introduction

Queueing networks are very useful for analyzing and approximating real stochastic systems. Many queueing networks assume that the arrivals to the network follow a Poisson process. This is a natural assumption when there is no dependence or correlation between arrivals. However, an independence assumption is not warranted in many applications. Stochastic models for describing the dynamics of internet data traffic in telecommunication networks are notoriously difficult since they have dependencies. One example of this is when a user downloads a file from the internet; the arrival of the first packet often indicates that more packets are going to arrive subsequently. Despite the assertions of the Palm–Khintine theorem, which asserts that the superposition of a large number of renewal processes will converge to a Poisson process, it is well known in the teletraffic literature that arrival traffic is not renewal.

There are many applications and scenarios where data traffic is not renewal, see for example [9,8]. To this end, we construct a queueing network where the arrivals are not Poisson and are constructed from Markovian Arrival Processes (MAP's). MAP's, unlike phase type distributions, allow one to consider non-renewal processes for the arrival process and offer more flexibility when modeling arrival traffic. One main reason that the MAP is a generalization of a phase type distribution is that the MAP is not restarted independently of its past history. In a MAP, unlike phase type distributions, the next interarrival time is dependent on the exit state of the Markov chain and this feature allows one to capture *memory* into the arrival process.

For the analysis of queues involving the MAP and phase-type distributions, the matrix-geometric method (MGM) introduced in [16] is frequently used. We, however, cannot escape from state space explosion when we have a large number of states and servers. [14] propose the partial-moment differential equations for the analysis of $Ph_t/Ph_t/\infty$ queues and [15] extend the result to the $[Ph_t/Ph_t/\infty]^N$ networks. [5] use phase-type distributions for approximating small-size $G_t/G_t/n_t + G_t$ queues.

In this paper, we study the $(MAP_t/Ph_t/\infty)^N$ queueing network. As a result, we extend recent work by [23,11], which only considers the single station setting. To this end, we prove strong approximations for the $(MAP_t/Ph_t/\infty)^N$ queueing network and extend the Poisson process representation to the network setting. These strong approximations not only allow us to derive fluid and diffusion limits for the queueing network, but also they provide us with simple differential equations that can be integrated numerically to approximate the sample path behavior of the mean and variance of the queueing network. Lastly, there are already many useful algorithms available for fitting phase-type distributions and MAP's from data such as [2,7,1,3] and there we can exploit this feature to approximate very complicated arrival processes that arise in practice.

[*] Corresponding author.
E-mail addresses: youngko@postech.ac.kr (Y.M. Ko), jjp274@cornell.edu (J. Pender).

http://dx.doi.org/10.1016/j.orl.2017.03.006
0167-6377/© 2017 Published by Elsevier B.V.

1.1. *Main contributions of paper*

The contributions of this work can be summarized as follows.

- We derive a Poisson process representation for the $(MAP_t/Ph_t/\infty)^N$ queueing network and prove strong approximations for the network.
- Using strong approximations for Poisson processes, we develop fluid and diffusion limits for the $(MAP_t/Ph_t/\infty)^N$ queueing network to understand the sample path mean and variance dynamics of the network.

1.2. *Organization of paper*

The remainder of this paper is organized as follows. Section 2 describes the construction of a MAP and phase type distributions. Section 3 builds a mathematical model for describing the dynamics of the system for the $(MAP_t/Ph_t/\infty)^N$ queueing network via time-changed Poisson processes. Using the Poisson representation in Section 3, we also prove the fluid and diffusion limits for the $(MAP_t/Ph_t/\infty)^N$ queueing network. Finally, Section 4 concludes and offers suggestions for future research.

2. Markovian Arrival Processes (MAP's)

In this section, we give a brief description of MAP's. The reader should review Section 2 of [11] and Chapter 11 of [1] for a more extensive discussion on MAP's and their versatility in stochastic modeling and queueing theory.

In order to define a MAP, we will follow the construction given in [4]. We first consider an irreducible continuous time Markov chain (CTMC) with h transient states. At the end of a sojourn in state i, that is exponentially distributed with parameter λ_i, there are two possible events that can happen. The first possibility corresponds to an event or arrival and the CTMC can visit state j (including $j = i$) with probability p_{ij}. The second possible event corresponds to no arrival and the CTMC can visit state j ($j \neq i$) with probability q_{ij}. Therefore, the CTMC is able to go from state i to state $i$through an arrival. Then, we define matrices $\mathcal{D}_0$ where $[\mathcal{D}_0]_{ij} = d^0_{ij}$ and $\mathcal{D}_1$ where $[\mathcal{D}_1]_{ij} = d^1_{ij}$ where $d^0_{ii} = -\lambda_i$, $1 \leq i \leq h$; $d^0_{ij} = \lambda_i \cdot q_{ij}, j \neq i$, $1 \leq i, j \leq h$; $d^1_{ij} = \lambda_i p_{ij}$, $1, \leq i, j \leq h$, with $\left(\sum_{j=1}^h p_{ij} + \sum_{j\neq i}^h q_{ij}\right) = 1$, for $1 \leq i \leq h$. In our description of the MAP, we have suppressed its dependence on time. However, all of our results apply to the time varying setting when the parameters are locally integrable with respect to time and therefore, we suppress time for notational convenience.

With the above construction, a MAP is described by the two $h \times h$ matrices $\mathcal{D}_0$ and $\mathcal{D}_1$. The matrix $\mathcal{D}_0$ corresponds to transitions where there is no arrival and $\mathcal{D}_1$ corresponds to the transitions that generate an actual arrival. With this construction, it is also obvious that this is more general than the renewal process with phase type inter-event time distributions. Dependence between arrivals is created by the fact that when an arrival is generated, then the Markov chain can re-enter the same state, however, when no arrival is generated, it cannot re-enter the same state. Now that we have defined a MAP, it is now important to understand how the MAP is a generalization of some well known stochastic arrival processes.

2.1. *Phase-type distributions*

A very special case of MAP's is the phase type distribution. Unlike MAP's, phase type distributions can only approximate renewal processes with arbitrary precision. A phase-type distribution with h phases can be viewed as the time taken from an initial state to an absorbing state of a continuous time Markov chain with the following infinitesimal generator matrix:

$$\mathbf{Q} = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{s} & \mathbf{S} \end{pmatrix},$$

where $\mathbf{0}$ is a $1 \times h$ zero vector, $\mathbf{s}$ = is an $h \times 1$ vector, and $\mathbf{S}$ is an $h \times h$ matrix. Note $\mathbf{s} = -\mathbf{Se}$ where $\mathbf{e}$ is an $h \times 1$ vector of ones. The matrix $\mathbf{S}$ and the initial distribution $\boldsymbol{\beta}$ which is a $1 \times h$ vector identify the phase-type distributions.

We assume that our phase-type distributions for the service times have an initial distribution, $\boldsymbol{\beta}$ and infinitesimal generator matrix, $\mathbf{Q_S}$. The number of phases in $\mathbf{S}_S$ is h_S and the matrix $\mathbf{S_S}$ and vector $\mathbf{s_S}$ can be expressed as:

$$\mathbf{S_S} = \begin{pmatrix} \mu_{11} & \cdots & \mu_{1h_S} \\ \vdots & \vdots & \vdots \\ \mu_{h_S 1} & \cdots & \mu_{h_S h_S} \end{pmatrix}, \qquad \mathbf{s_S} = (\mu_{10}, \ldots, \mu_{h_S 0})', \tag{2.1}$$

where the μ_{il}'s agree with the definition of the infinitesimal generator matrix $\mathbf{Q_S}$. For notational consistency, we use a term *phase* to indicate the state of CTMC for both the MAPs and phase-type distributions throughout this paper.

3. Poisson construction of $(MAP_t/Ph_t/\infty)^N$ network

With the MAP's and phase-type distributions described in Section 2, we now build a mathematical queueing model to describe the dynamics of the $(MAP_t/Ph_t/\infty)^N$ queueing network. To this end, we need to provide the primitives of the queueing network. The network consists of N stations. For station m in the network, we assume that the external arrivals are generated by a MAP, $U_{jm}(\cdot), j \in \{1, \ldots, h_{A,m}\}$, where $h_{A,m}$ is the number of phases of the MAP. Similarly for service, given that we are in station m, we assume that the initial distribution for the phase-type service distribution is given by $\boldsymbol{\beta}_m = (\beta_{1m}, \ldots \beta_{h_{S,m}m})$. We let $X_{im}(t)$ denote the number of customers in phase i, $i \in \{1, \ldots, h_{S,m}\}$, of the phase-type service distribution at time $t \geq 0$. After a customer in station m is served, the customer moves to station i with probability p_{mi} and leaves the network with probability $p_{m0} = 1 - \sum_{i=1}^N p_{mi}$. Lastly, we assume that the queueing network starts with no customers. Fig. 3.1 illustrates an example of a $(MAP_t/Ph_t/\infty)^N$ queueing network with two stations.

Thus, in the infinite-server setting, we have the following Poisson process representation for the $(MAP_t/Ph_t/\infty)^N$ queueing network,

$$\begin{aligned} U_{jm}(t) = \quad & \underbrace{U_{jm}(0)}_{\text{Initial Value of Token}} \\ & + \underbrace{\sum_{k\neq j}^{h_{A,m}} \mathbf{\Pi}^{A0}_{kjm}\left(\int_0^t d^{A0}_{kjm} U_{km}(s)ds\right)}_{\text{MAP in station } m \text{ moves from state } k \text{ to } j \text{ (no arrival generated)}} \\ & + \underbrace{\sum_{k=1}^{h_{A,m}} \sum_{i=1}^{h_{S,m}} \mathbf{\Pi}^{A1}_{kjim}\left(\int_0^t d^{A1}_{kjm} \beta_{im} U_{jm}(s)ds\right)}_{\text{MAP in station } m \text{ moves from state } k \text{ to } j \text{ (arrival generated)}} \\ & - \underbrace{\sum_{k\neq j}^{h_{A,m}} \mathbf{\Pi}^{A0}_{jkm}\left(\int_0^t d^{A0}_{jkm} U_{jm}(s)ds\right)}_{\text{MAP in station } m \text{ moves from state } j \text{ to } k \text{ (no arrival generated)}} \end{aligned}$$

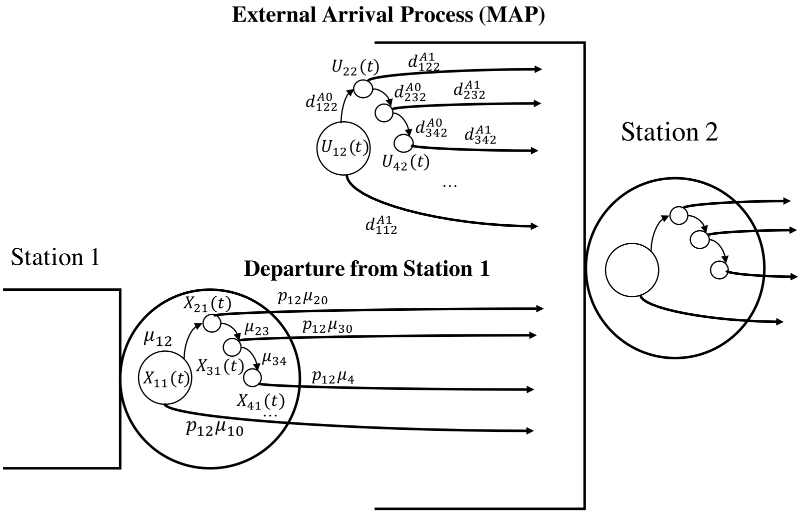

Fig. 3.1. A $(MAP_t/Ph_t/\infty)^N$ network with two stations.

$$- \underbrace{\sum_{k=1}^{h_{A,m}}\sum_{i=1}^{h_{S,m}} \mathbf{\Pi}^{A1}_{jkim}\left(\int_0^t d^{A1}_{jkm}\beta_{im}U_{jm}(s)ds\right)}_{\text{MAP in station } m \text{ moves from state } j \text{ to } k \text{ (arrival generated)}}$$

$$\text{for } 1 \le j \le h_{A,m}, \tag{3.1}$$

$$\begin{aligned}
X_{im}(t) = & \underbrace{\sum_{j=1}^{h_{A,m}}\sum_{k=1}^{h_{A,m}} \mathbf{\Pi}^{A1}_{jkim}\left(\int_0^t d^{A1}_{jkm}\beta_{im}U_{jm}(s)ds\right)}_{\text{External arrivals into phase } i \text{ of service}} \\
& + \underbrace{\sum_{l\neq i}^{h_{S,m}} \mathbf{\Pi}^{S}_{lim}\left(\int_0^t \mu_{lim}X_{lm}(s)ds\right)}_{\text{Internal transitions from phase } l \text{ to phase } i} \\
& - \underbrace{\sum_{l\neq i}^{h_{S,m}} \mathbf{\Pi}^{S}_{ilm}\left(\int_0^t \mu_{ilm}X_{im}(s)ds\right)}_{\text{Internal transitions from phase } i \text{ to phase } l \text{ of service}} \\
& - \underbrace{\mathbf{\Pi}^{D}_{im}\left(\int_0^t \mu_{i0m}p_{m0}X_{im}(s)ds\right)}_{\text{Departures of phase } i \text{ from network}} \\
& - \underbrace{\sum_{l=1}^{h_{S,m}}\sum_{n\neq m}^{N} \mathbf{\Pi}^{R}_{ilmn}\left(\int_0^t \mu_{i0m}\, p_{mn}\beta_{ln}X_{im}(s)ds\right)}_{\text{Routing transitions from phase } i \text{ to phase } l \text{ (station } m \text{ to } n)} \\
& + \underbrace{\sum_{l=1}^{h_{S,m}}\sum_{n\neq m}^{N} \mathbf{\Pi}^{R}_{linm}\left(\int_0^t \mu_{l0n}\, p_{nm}\beta_{im}X_{in}(s)ds\right)}_{\text{Routing transitions from phase } l \text{ to phase } i \text{ (station } n \text{ to } m)}
\end{aligned}$$

$$\text{for } 1 \le i \le h_{S,m} \tag{3.2}$$

where $U_{jm}(t)$ represents phase j of the MAP in station m of the network and $X_{im}(t)$ is the phase i of the phase type distribution in station m of the network at time t.

We assume that each of the Poisson processes in the representation of the $(MAP_t/Ph_t/\infty)^N$ queueing network is of unit rate and pairwise independent. First we begin by describing the Poisson processes that help generate the arrival process. Poisson process, $\mathbf{\Pi}^{A0}_{kjm}(\cdot)$, counts the number of transitions that a token will make from phase k to phase j of the non-arrival producing part of the MAP in station m with transition rate d^{A0}_{kjm}. Poisson process, $\mathbf{\Pi}^{A1}_{kjim}(\cdot)$, counts the number of transitions that a token will make from phase k to phase j of the arrival producing part of the MAP in station m with transition rate d^{A1}_{kjm} and the arrival is heading to phase i of the service process according to the initial distribution of station m given by $\boldsymbol{\beta}_m$. Poisson process, $\mathbf{\Pi}^{S}_{lim}(\cdot)$, counts the internal transitions within each station from phase l to phase i in station m of the service process. Poisson process, $\mathbf{\Pi}^{D}_{im}(\cdot)$, counts the number of departures from phase i in station m of the service process completely out of the network. Lastly, Poisson process, $\mathbf{\Pi}^{R}_{ilmn}(\cdot)$, counts the routing transitions from phase i in station m to phase l in station n of the service process. For the remainder of the paper, we will use the following notation for the stochastic queue length process and its fluid version.

$$\begin{aligned}
\mathbf{Q}(t) = \mathbf{Q}(0) & + \sum_{m=1}^{N}\sum_{j=1}^{h_{A,m}}\sum_{k\neq j}^{h_{A,m}} \mathbf{l}^{A0}_{jkm}\mathbf{\Pi}^{A0}_{kjm}\left(\int_0^t f^{A0}_{jkm}(s,\mathbf{Q}(s))ds\right) \\
& + \sum_{m=1}^{N}\sum_{j=1}^{h_{A,m}}\sum_{k=1}^{h_{A,m}}\sum_{i=1}^{h_{S,m}} \mathbf{l}^{A1}_{jkim}\mathbf{\Pi}^{A1}_{jkim}\left(\int_0^t f^{A1}_{jkim}(s,\mathbf{Q}(s))ds\right) \\
& + \sum_{m=1}^{N}\sum_{i=1}^{h_{S,m}}\sum_{l\neq i}^{h_{S,m}} \mathbf{l}^{S}_{ilm}\mathbf{\Pi}^{S}_{ilm}\left(\int_0^t f^{S}_{ilm}(s,\mathbf{Q}(s))ds\right) \\
& + \sum_{m=1}^{N}\sum_{i=1}^{h_{S,m}} \mathbf{l}^{D}_{im}\mathbf{\Pi}^{D}_{im}\left(\int_0^t f^{D}_{im}(s,\mathbf{Q}(s))ds\right) \\
& + \sum_{m=1}^{N}\sum_{n=1}^{N}\sum_{i=1}^{h_{S,m}}\sum_{l=1}^{h_{S,m}} \mathbf{l}^{R}_{ilmn}\mathbf{\Pi}^{R}_{ilmn}\left(\int_0^t f^{R}_{ilmn}(s,\mathbf{Q}(s))ds\right),
\end{aligned}$$

where we define

$$\mathbf{Q}_m(t) = (U_{1m}(t),\ldots,U_{h_{A,m}m}(t),X_{1m}(t),\ldots,X_{h_{S,m}m}(t))',$$
$$\mathbf{Q}(t) = (\mathbf{Q}_1(t)',\ldots,\mathbf{Q}_N(t)')',$$

$$\mathbf{l}^{A0}_{jkm} : \sum_{m=1}^{N}(h_{A,m}+h_{S,m}) \times 1 \text{ vector},$$
$$\left(\sum_{r=1}^{m-1}(h_{A,r}+h_{S,r})+j\right)\text{th element is } -1,$$
$$\left(\sum_{r=1}^{m-1}(h_{A,r}+h_{S,r})+k\right)\text{th element is } 1,$$

and other elements are 0,

$\mathbf{l}^{A1}_{jkim} : \sum_{m=1}^{N}(h_{A,m}+h_{S,m}) \times 1$ vector,

$\left(\sum_{r=1}^{m-1}(h_{A,r}+h_{S,r})+j\right)$th element is -1,

$\left(\sum_{r=1}^{m-1}(h_{A,r}+h_{S,r})+k\right)$th element is 1,

$\left(\sum_{r=1}^{m-1}(h_{A,r}+h_{S,r})+h_{A,m}+i\right)$th element is 1,

and other elements are 0,

$\mathbf{l}^{S}_{ilm} : \sum_{m=1}^{N}(h_{A,m}+h_{S,m}) \times 1$ vector,

$\left(\sum_{r=1}^{m-1}(h_{A,r}+h_{S,r})+h_{A,m}+i\right)$th element is -1,

$\left(\sum_{r=1}^{m-1}(h_{A,r}+h_{S,r})+h_{A,m}+l\right)$th element is 1, and other elements are 0,

$\mathbf{l}^{D}_{im} : \sum_{m=1}^{N}(h_{A,m}+h_{S,m}) \times 1$ vector,

$\left(\sum_{r=1}^{m-1}(h_{A,r}+h_{S,r})+h_{A,m}+i\right)$th element is -1,

and other elements are 0,

$\mathbf{l}^{R}_{ilmn} : \sum_{m=1}^{N}(h_{A,m}+h_{S,m}) \times 1$ vector,

$\left(\sum_{r=1}^{m-1}(h_{A,r}+h_{S,r})+h_{A,m}+i\right)$th element is -1,

$\left(\sum_{r=1}^{n-1}(h_{A,r}+h_{S,r})+h_{A,n}+l\right)$th element is 1,

and other elements are 0.

Moreover, we use the following notation for the rate functions (integrands of the Poisson processes) and the jump vectors (a value determining whether a jump is added or subtracted).

$f^{A0}_{jkm}(t,\mathbf{q})$: rate function of the (integrand) in $\mathbf{\Pi}^{A0}_{jkm}(\cdot)$,

$f^{A1}_{jkim}(t,\mathbf{q})$: rate function of the (integrand) in $\mathbf{\Pi}^{A1}_{jkim}(\cdot)$,

$f^{S}_{ilm}(t,\mathbf{q})$: rate function of the (integrand) in $\mathbf{\Pi}^{S}_{ilm}(\cdot)$,

$f^{D}_{im}(t,\mathbf{q})$: rate function of the (integrand) in $\mathbf{\Pi}^{D}_{im}(\cdot)$,

$f^{R}_{ilmn}(t,\mathbf{q})$: rate function of the (integrand) in $\mathbf{\Pi}^{R}_{ilmn}(\cdot)$,

where

$\mathbf{q}_m = (u_{1m},\ldots,u_{h_{A,m}m},x_{1m},\ldots,x_{h_{S,m}m})' \in \mathcal{R}_+^{(h_{A,m}+h_{S,m})}$

for $m \in \{1,\ldots,N\}$ and

$\mathbf{q} = (\mathbf{q}_1',\ldots,\mathbf{q}_N')' \in \mathcal{R}_+^{\sum_{m=1}^{N}(h_{A,m}+h_{S,m})}$.

3.1. Fluid limit

With the Poisson process representation for the $(MAP_t/Ph_t/\infty)^N$, we now show how we can use the Poisson representation and strong approximations in order to prove fluid limits for the queue length process. First, we define a sequence of stochastic processes $\{\mathbf{Q}^\eta(t),\, \eta\in\mathcal{N},\, t\in\mathcal{R}_+\}$:

$$\begin{aligned}\mathbf{Q}^\eta(t) &= \mathbf{Q}^\eta(0) + \sum_{m=1}^{N}\sum_{j=1}^{h_{A,m}}\sum_{k\neq j}^{h_{A,m}} \mathbf{l}^{A0}_{jkm}\mathbf{\Pi}^{A0}_{kjm}\left(\eta\int_0^t f^{A0}_{jkm}(s,\bar{\mathbf{Q}}^\eta(s))ds\right)\\ &\quad+\sum_{m=1}^{N}\sum_{j=1}^{h_{A,m}}\sum_{k=1}^{h_{A,m}}\sum_{i=1}^{h_{S,m}} \mathbf{l}^{A1}_{jkim}\mathbf{\Pi}^{A1}_{jkim}\left(\eta\int_0^t f^{A1}_{jkim}(s,\bar{\mathbf{Q}}^\eta(s))ds\right)\\ &\quad+\sum_{m=1}^{N}\sum_{i=1}^{h_{S,m}}\sum_{l\neq i}^{h_{S,m}} \mathbf{l}^{S}_{ilm}\mathbf{\Pi}^{S}_{ilm}\left(\eta\int_0^t f^{S}_{ilm}(s,\bar{\mathbf{Q}}^\eta(s))ds\right)\\ &\quad+\sum_{m=1}^{N}\sum_{i=1}^{h_{S,m}} \mathbf{l}^{D}_{im}\mathbf{\Pi}^{D}_{im}\left(\eta\int_0^t f^{D}_{im}(s,\bar{\mathbf{Q}}^\eta(s))ds\right)\\ &\quad+\sum_{m=1}^{N}\sum_{n=1}^{N}\sum_{i=1}^{h_{S,m}}\sum_{l=1}^{h_{S,m}} \mathbf{l}^{R}_{ilmn}\mathbf{\Pi}^{R}_{ilmn}\left(\eta\int_0^t f^{R}_{ilmn}(s,\bar{\mathbf{Q}}^\eta(s))ds\right)\end{aligned}$$

where we define

$$\bar{\mathbf{Q}}^\eta(t) = \frac{1}{\eta}\mathbf{Q}^\eta(t).$$

Note that we accelerate the arrival rate by setting $\sum_{j=1}^{h_{A,m}} U^\eta_{jm}(t) = \eta$ and $\sum_{m=1}^{N}\sum_{j=1}^{h_{A,m}} U^\eta_{jm}(t) = \eta N$ for $t \geq 0$. Then, the following proposition describes the fluid limits for the $(MAP_t/Ph_t/\infty)^N$ queueing network.

Theorem 3.1 (*Fluid Limit*). *Suppose $\mathbf{Q}^\eta(0)/\eta \to \mathbf{q}(0)$ almost surely as $\eta\to\infty$, then*

$$\lim_{\eta\to\infty}\frac{1}{\eta}\mathbf{Q}^\eta(t) = \mathbf{q}(t) \quad \textit{almost surely},$$

where we define

$\mathbf{q}_m(t) = (u_{1m}(t),\ldots,u_{h_{A,m}m}(t),x_{1m}(t),\ldots,x_{h_{S,m}m}(t))'$

for $m\in\{1,\ldots,N\}$ and

$\mathbf{q}(t) = (\mathbf{q}_1(t)',\ldots,\mathbf{q}_N(t)')'$,

and $\mathbf{q}(t)$ is the solution to the following system of ordinary differential equations:

$$\frac{d}{dt}\mathbf{q}(t) = \mathbf{F}(t,\mathbf{q}(t)),$$

where

$$\begin{aligned}\mathbf{F}(t,\mathbf{q}(t)) &= \sum_{m=1}^{N}\sum_{j=1}^{h_{A,m}}\sum_{k\neq j}^{h_{A,m}} \mathbf{l}^{A0}_{jkm}f^{A0}_{jkm}(t,\mathbf{q}(t))\\ &\quad+\sum_{m=1}^{N}\sum_{j=1}^{h_{A,m}}\sum_{k=1}^{h_{A,m}}\sum_{i=1}^{h_{S,m}} \mathbf{l}^{A1}_{jkim}f^{A1}_{jkim}(t,\mathbf{q}(t))\\ &\quad+\sum_{m=1}^{N}\sum_{i=1}^{h_{S,m}}\sum_{l\neq i}^{h_{S,m}} \mathbf{l}^{S}_{ilm}f^{S}_{ilm}(t,\mathbf{q}(t))\\ &\quad+\sum_{m=1}^{N}\sum_{i=1}^{h_{S,m}} \mathbf{l}^{D}_{im}f^{D}_{im}(t,\mathbf{q}(t))\\ &\quad+\sum_{m=1}^{N}\sum_{n=1}^{N}\sum_{i=1}^{h_{S,m}}\sum_{l=1}^{h_{S,m}} \mathbf{l}^{R}_{ilmn}f^{R}_{ilmn}(t,\mathbf{q}(t)).\end{aligned}$$

Proof. By adding and subtracting the integrand of each Poisson process, we now have the following bound of the difference of the scaled queue length and the fluid limit,

$$\left|\frac{1}{\eta}\mathbf{Q}^\eta(t) - \mathbf{q}(t)\right| \le \int_0^t |\mathbf{F}(s, \bar{\mathbf{Q}}^\eta(s)) - \mathbf{F}(s, \mathbf{q}(s))|ds + |\mathbf{V}^\eta(t)|$$

where $|\cdot|$ denotes the absolute value function and is equal to the sum of the absolute value of each component of the vector. The remainder of the proof follows from a standard argument using Gronwall's lemma. □

3.2. Diffusion limit

Now that we have the fluid limit, $\mathbf{q}(t)$, we derive the diffusion limit with the following proposition.

Let $\mathbf{D}^\eta(t) = \sqrt{\eta}\left(\frac{1}{\eta}\mathbf{Q}^\eta(t) - \mathbf{q}(t)\right)$ and suppose that $\sqrt{\eta}\left(\frac{1}{\eta}\mathbf{Q}^\eta(0) - \mathbf{q}(0)\right)$ converges to $\mathbf{D}(0)$ in distribution as $\eta \to \infty$.

Proposition 3.2. *Suppose that we define $\tilde{\mathbf{D}}^\eta(t)$ as the solution of the following integral equation:*

$$\begin{aligned} \tilde{\mathbf{D}}^\eta(t) &= \int_0^t \partial\mathbf{F}(s, \mathbf{q}(s))\tilde{\mathbf{D}}^\eta(s)ds + \sqrt{\eta}\cdot\mathbf{V}^\eta(t), \\ \tilde{\mathbf{D}}^\eta(0) &= \mathbf{D}^\eta(0), \end{aligned} \tag{3.3}$$

then

$$\lim_{\eta\to\infty}\sup_{0\le t\le T}|\mathbf{D}^\eta(t) - \tilde{\mathbf{D}}^\eta(t)| = 0 \quad \textit{in probability}. \tag{3.4}$$

Proof. To prove this, we define the difference between the two processes as

$$\begin{aligned} \mathbf{E}^\eta(t) &= \mathbf{D}^\eta(t) - \tilde{\mathbf{D}}^\eta(t) \\ &= \int_0^t \left(\partial\mathbf{F}(s, \zeta^\eta(s)) - \partial\mathbf{F}(s, \mathbf{q}(s))\right)\mathbf{D}^\eta(s)ds \\ &\quad + \int_0^t \partial\mathbf{F}(s, \mathbf{q}(s))\mathbf{E}^\eta(s)ds, \end{aligned}$$

where $\zeta^\eta(s)$ is defined as a multivariate stochastic process that lies between $\bar{\mathbf{Q}}^\eta(s)$ and $\mathbf{q}(s)$ like in the standard mean value theorem. Thus, by the mean value theorem, the fact that the rate functions in the Poisson representations are continuously differentiable, stochastic boundedness of $\mathbf{D}^\eta(t)$, and the fluid limit convergence, we obtain our diffusion limit result by applying Gronwall's lemma. □

Theorem 3.3 (*Diffusion Limit*).

$$\lim_{\eta\to\infty}\mathbf{D}^\eta(t) = \mathbf{D}(t) \quad \textit{in distribution on } t \in [0, T],$$

where $T < \infty$ and $\mathbf{D}(t)$ is the solution to the following stochastic differential equation:

$$d\mathbf{D}(t) = d\mathbf{H}(t, \mathbf{q}(t)) + \partial\mathbf{F}(t, \mathbf{q}(t))\mathbf{D}(t)dt,$$

and $\partial\mathbf{F}(t, \mathbf{q}(t))$ is the gradient matrix of $\mathbf{F}(t, \mathbf{q}(t))$ with respect to $\mathbf{q}(t)$. Moreover,

$$\begin{aligned} & d\mathbf{H}(t, \mathbf{q}(t)) \\ &= \sum_{m=1}^{N}\sum_{j=1}^{h_{A,m}}\sum_{k\neq j}^{h_{A,m}} \mathbf{I}^{A0}_{jkm}\sqrt{f^{A0}_{jkm}(t, \mathbf{q}(t))}dW^{A0}_{jkm}(t) \\ &\quad + \sum_{m=1}^{N}\sum_{j=1}^{h_{A,m}}\sum_{k=1}^{h_{A,m}}\sum_{i=1}^{h_{S,m}} \mathbf{I}^{A1}_{jkim}\sqrt{f^{A1}_{jkim}(t, \mathbf{q}(t))}dW^{A1}_{jkim}(t) \\ &\quad + \sum_{m=1}^{N}\sum_{i=1}^{h_{S,m}}\sum_{l\neq i}^{h_{S,m}} \mathbf{I}^{S}_{ilm}\sqrt{f^{S}_{ilm}(t, \mathbf{q}(t))}dW^{S}_{ilm}(t) \\ &\quad + \sum_{m=1}^{N}\sum_{i=1}^{h_{S,m}} \sqrt{f^{D}_{im}(t, \mathbf{q}(t))}dW^{D}_{im}(t) \\ &\quad + \sum_{m=1}^{N}\sum_{n=1}^{N}\sum_{i=1}^{h_{S,m}}\sum_{l=1}^{h_{S,m}} \mathbf{I}^{R}_{ilmn}\sqrt{f^{R}_{ilmn}(t, \mathbf{q}(t))}dW^{R}_{ilmn}(t) \end{aligned}$$

where $W^{A0}_{jkm}(t)$, $W^{A1}_{jkim}(t)$, $W^{S}_{ilm}(t)$, $W^{D}_{im}(t)$, $W^{R}_{ilmn}(t)$ are mutually independent standard Brownian motions.

Proof. In order to construct the diffusion limit, we need to subtract the fluid limit and multiply by $\sqrt{\eta}$. This yields the following expression for $\mathbf{D}^\eta(t)$

$$\begin{aligned} \mathbf{D}^\eta(t) &= \sqrt{\eta}\left(\frac{1}{\eta}\mathbf{Q}^\eta(t) - \mathbf{q}(t)\right) \\ &= \sqrt{\eta}\cdot\left(\int_0^t \mathbf{F}(s, \bar{\mathbf{Q}}^\eta(s)) - \mathbf{F}(s, \mathbf{q}(s))ds\right) + \sqrt{\eta}\cdot\mathbf{V}^\eta(t). \end{aligned}$$

We know by the continuous mapping theorem and the fact that Brownian motion is Hölder continuous, which shows that $\sqrt{\eta}\cdot\mathbf{V}^\eta(t)$ converges to time changed brownian motions. Moreover, we know that $\tilde{\mathbf{D}}^\eta(t)$ converges to $\mathbf{D}(t)$ in distribution on $t \in [0, T]$. Thus, from the result of Proposition 3.2, $\mathbf{D}^\eta(t)$ converges weakly to $\mathbf{D}(t)$ on $t \in [0, T]$. □

Remark (*Steady State Behavior*)**.** Although the weak-convergence is proved on a compact set $[0, T]$, we can think of the stationary distribution for the diffusion process when $t \to \infty$. Considering a $(MAP/Ph/\infty)^N$ network (i.e., all rate functions do not have a parameter t), the fluid limit, $\bar{\mathbf{q}} = \mathbf{q}(\infty)$, should satisfy

$$\frac{d}{dt}\mathbf{q}(t)|_{t=\infty} = \mathbf{F}(\bar{\mathbf{q}}) = 0.$$

Then, the diffusion limit in steady state would be a multi-dimensional Ornstein–Uhlenbeck process given by

$$d\mathbf{D}(t) = d\mathbf{H}(\bar{\mathbf{q}}) + \partial\mathbf{F}(\bar{\mathbf{q}})\mathbf{D}(t)dt,$$

which we know has a steady state distribution that is normally distributed.

4. Conclusion and future research

In this paper, we analyze the $(MAP_t/Ph_t/\infty)^N$ queueing network and prove fluid and diffusion limits for the queue length processes. It is critical to fully understand the behavior of the $(MAP_t/Ph_t/\infty)^N$ queueing network before we can begin to understand networks such as $(MAP_t/MAP_t/n_t + MAP_t)^N$ networks since the infinite-server case represents the offered load of the system with unlimited resources. As future work in the context of finite-server and abandonment settings, it is interesting to apply the methods of [10,12,13,6,22,18,17,20,21,19] to these non-Markovian networks since we should consider the *critically-loaded regime* under which some conditions for deriving diffusion limits are violated.

Acknowledgment

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1005019).

References

[1] Søren Asmussen, Applied Probability and Queues, Vol. 51, Springer Science & Business Media, 2008.

[2] Søren Asmussen, Olle Nerman, Marita Olsson, Fitting phase-type distributions via the EM algorithm, Scand. J. Statist. 23 (4) (1996) 419–441.

[3] Peter Buchholz, Peter Kemper, Jan Kriege, Multi-class Markovian arrival processes and their parameter fitting, Perform. Eval. 67 (11) (2010) 1092–1106.

[4] Srinivas R. Chakravarthy, Markovian arrival processes, in: Wiley Encyclopedia of Operations Research and Management Science, 2010.

[5] Stefan Creemers, Mieke Defraeye, Inneke Van Nieuwenhuyse, G-RAND: A phase-type approximation for the nonstationary $G(t)/G(t)/s(t)+G(t)$ queue, Perform. Eval. 80 (2014) 102–123.

[6] Stefan Engblom, Jamol Pender, Approximations for the moments of nonstationary and state dependent birth-death queues, 2014.

[7] Anja Feldmann, Ward Whitt, Fitting mixtures of exponentials to long-tail distributions to analyze network performance models, Perform. Eval. 31 (3–4) (1998) 245–279.

[8] Ira Gerhardt, Barry L. Nelson, On capturing dependence in point processes: Matching moments and other techniques. Technical Report, Technical Report, Northwestern Univ., 2009.

[9] Sang H. Kang, Yong Han Kim, Dan K. Sung, Bong D. Choi, An application of Markovian arrival process (map) to modeling superposed atm cell streams, IEEE Trans. Commun. 50 (4) (2002) 633–642.

[10] Young Myoung Ko, Natarajan Gautam, Critically loaded time-varying multiserver queues: Computational challenges and approximations, INFORMS J. Comput. 25 (2) (2013) 285–301.

[11] Young Myoung Ko, Jamol Pender, Strong approximations for time varying infinite server queues with non-renewal arrival and service processes, 2016, working paper.

[12] William A. Massey, Jamol Pender, Poster: skewness variance approximation for dynamic rate multiserver queues with abandonment, ACM SIGMETRICS Perform. Eval. Rev. 39 (2) (2011) 74–74.

[13] William A. Massey, Jamol Pender, Gaussian skewness approximation for dynamic rate multi-server queues with abandonment, Queueing Syst. 75 (2–4) (2013) 243–277.

[14] Barry L. Nelson, Michael R. Taaffe, The $Ph_t/Ph_t/\infty$ queueing system: Part I-The single node, INFORMS J. Comput. 16 (3) (2004) 266–274.

[15] Barry L. Nelson, Michael R. Taaffe, The $[Ph_t/Ph_t/\infty]^K$ queueing system: Part II-The multiclass network, INFORMS J. Comput. 16 (3) (2004) 275–283.

[16] Marcel F. Neuts, Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach, Dover Publication, Inc., 1981.

[17] Jamol Pender, A Poisson–charlier approximation for nonstationary queues, Oper. Res. Lett. 42 (4) (2014) 293–298.

[18] Jamol Pender, Gram charlier expansion for time varying multiserver queues with abandonment, SIAM J. Appl. Math. 74 (4) (2014) 1238–1265.

[19] Jamol Pender, An analysis of nonstationary coupled queues, Telecommun. Syst. (2015) 1–16.

[20] Jamol Pender, Nonstationary loss queues via cumulant moment approximations, Probab. Engrg. Inform. Sci. 29 (01) (2015) 27–49.

[21] Jamol Pender, The truncated normal distribution: Applications to queues with impatient customers, Oper. Res. Lett. 43 (1) (2015) 40–45.

[22] Jamol Pender, Sampling the functional Kolmogorov forward equations for nonstationary queueing networks, INFORMS J. Comput. 29 (1) (2016) 1–17.

[23] Jamol Pender, Young Myoung Ko, Approximations for the queue length distributions of time-varying many-server queues, INFORMS J. Comput. (2017) in press.