Open Mathematics

Research Article

Yuejiao Wang, Zaiming Liu, Yuqing Chu*, and Yingqiu Li

An asymptotic property of branching-type overloaded polling networks

https://doi.org/10.1515/math-2019-0116
Received January 2, 2019; accepted September 6, 2019

Abstract: Remerova et al. [Random fluid limit of an overloaded polling model, Adv. Appl. Probab., 2014, 46, 76–101] studied the fluid asymptotics of the joint queue length process for an overloaded cyclic polling system with multigated service discipline by exploiting the connection with multi-type branching processes. In contrast to the heavy traffic behaviors, the cycle time of the overloaded polling system increases by a deterministic times over times under passage to the fluid dynamics and the fluid limit preserves some randomness. The present paper aims to extend the overloaded asymptotics in Remerova et al. [Random fluid limit of an overloaded polling model, Adv. Appl. Probab., 2014, 46, 76–101] to the corresponding polling system with general branching-type service disciplines and customer re-routing policy. A unifying overloaded asymptotic property is derived. Due to the exhaustiveness, the property is a natural extension of the classical polling model with multigated service discipline in Remerova et al. [Random fluid limit of an overloaded polling model, Adv. Appl. Probab., 2014, 46, 76–101] and provides new exact results that have not been observed before for rerouting policy. Additionally, a stochastic simulation is undertaken for the validation of the fluid limit and the optimization of the gating indexes to minimize the total population is considered as an example to demonstrate the usefulness of the random fluid limit.

Keywords: polling networks, overloaded, general branching-type service policies, multi-type branching process, exhaustiveness

MSC 2010: 60J80, 90Bxx

1 Introduction

In this paper, we consider a cyclic N-queue $(Q_1, \cdots, Q_N, N \geq 2)$ polling system with general branching-type service discipline within each queue and customer re-routing policies: after completing service at Q_i, a customer is either routed to Q_j with probability $p_{i,j}$ or leaves the system with probability $p_{i,0}$. The possibility for re-routing of customers further enhances the already-extensive modeling capabilities of polling models, since in many applications, customers require service at more than one facility of the system. Actually, the models of customer re-routing arise naturally in various models of computer, communication and robotic

Yuejiao Wang: College of Mathematics and Computational Science, Hunan First Normal University, Changsha 410205, China, E-mail: yuejiao.wang@csu.edu.cn
Zaiming Liu: School of Mathematics and Statistics, Central South University, 410083, Changsha, China, E-mail: math_lzm@csu.edu.cn
***Corresponding Author: Yuqing Chu:** School of Science, Wuhan University of Technology, Wuhan 430070, Hubei, PR China, E-mail: chuyuqing@whut.edu.cn
Yingqiu Li: School of Mathematics and Statistic, Changsha University of Science and Technology, 410004, Changsha, China, E-mail: liyq-2001@163.com

systems (see [1-4] and references therein). One obvious example is a local area network in which terminals are interconnected in either a physical or logical structure (see [5]).

In the vast majority of papers that have appeared on polling models, it is almost invariably assumed that the system is stable and the stable performance measures are then concerned. With the advent of the era of Internet+, the study of critically or strictly super-critically loaded polling systems is vigorously pioneered due to the overloaded Internet channel or online shopping orders.

The heavy traffic ($\rho \to 1$, ρ is the load of the system) behaviors have gained an ascending attention in the last two decades pioneered by Coffman et al. [6, 7]. By utilizing the connection with multi-type branching process, van der Mei [8] considered a unifying theory on branching-type polling models under heavy-traffic assumptions. In the similar way, Boon et al. [1] discussed the heavy-traffic asymptotic behaviors of a gated polling system with customer re-route policy. Furthermore, Liu et al. [9] extended the results in [1] to the analogous system with a general branching-type service policy in the same form. As an example for non-branching type polling systems, Liu et al. [10] investigated the heavy-traffic behavior of a priority polling system consisting of three M/M/1 queues with threshold policy and proved that the scaled queue-length of the critically loaded queue is exponentially distributed, independent of that of the stable queues.

The study of overloaded ($\rho > 1$) service system is important to control or predict how fast it blows up over time. However, hardly any attention has been given to the overloaded polling system. The few literature refers to [11-14]. By using measure-valued state descriptor, Puha et al. [11] proved that the overloaded $GI/GI/1$ processor sharing queues converge in distribution to supercritical fluid models and a fluid limit result is proved as first order approximations to overloaded processor sharing queues. Using both fluid and diffusion limits, Jennings et al. [13] showed that the virtual waiting time process of an overloaded Multi-class FIFO(first-in-first-out) queue with abandonments converges to a limiting deterministic fluid process. Instead Remerova et al. [14] showed the fluid asymptotic process for the joint queue length process on an overloaded branching-type cyclic polling system by using the asymptotic properties of multi-type branching processes.

The motivation of the present paper is twofold. First, we dedicate to the investigation of a unified overloaded asymptotic fluid process for a very general class of branching-type polling models. Resing [15] observed that for a large class of polling models, including for example cyclic polling models with exhaustive and gated service, the evolution of the system at successive polling instants at a fixed queue can be described as a multi-type branching process with immigration. Models that satisfy this MTBP-structure allow for an exact analysis, whereas models that violate the MTBP-structure are often more intricate. Moreover, van der Mei [8] has given a unifying and insightful theory on branching-type polling models under heavy-traffic assumptions, which shows particularly attractive features. Second, by considering the polling system with re-routing policy, we aim to study how the strikingly simple overloaded fluid limit depends on the system parameters and in particular, on the routing probabilities $p_{i,j}$. Actually, the work carried out here is a natural progression from [9] and a natural extension of [14] due to the exhaustiveness. Our resulting expressions are very insightful, simple to implement and suitable for optimization purposes. Numerical results are presented to assess the accuracy of the results.

The rest of the paper is organized as follows. In Section 2, we describe precisely the polling model. In Sections 3, we introduce the multi-type branching process-structure of polling system. Section 4 provides the main results. Section 5 provides the proofs of Theorems 4.1 and 4.2. Section 6 discusses some numerical issues including the stochastic simulation to test the validity of the overloaded asymptotic behaviors and the optimization of gating indexes to minimize the average growth rate of the total population as an example. Section 7 concludes and provides an outlook on potential further research of our paper.

2 Model description

Consider an asymmetric polling network that consists of $N \geq 2$ queues, $Q_1, \cdots, Q_N$, and attended by a single server that visits the queues in a cyclic order. Customers arrive at Q_i according to a Poisson process $E_i(\cdot)$ with rate λ_i. The service time of each customer at Q_i is a random variable B_i with finite mean value $\mathbb{E}B_i = 1/\mu_i$.

Indices throughout the paper are modulo N, so Q_{N+1} actually refers to Q_1. The interarrival times and the service times (for different queues and for different visits) are assumed to be mutually independent.

In addition, we will consider the impact of customer rerouting policy. Upon completion of service at $Q_i, i = 1, \cdots, N$, a customer is either routed to $Q_j, j = 1, \cdots, N$ with probability $p_{i,j}$ or leaves the system with probability $p_{i,0}$, where

$$\sum_{i=1}^{N} p_{i,0} > 0 \quad \text{and} \quad \sum_{j=0}^{N} p_{i,j} = 1.$$

We assume that all the switches of customers or servers between queues are instantaneous and when the system becomes empty, the server travels a full cycle and subsequently stops right before Q_1 until a new arrival occurs and then cycles along the queues to serve that customer.

Denote the total arrival rate at Q_i by γ_i, which is the unique solution of the following set of linear equation [2]:

$$\gamma_i = \lambda_i + \sum_{j=1}^{N} \gamma_j p_{j,i} \qquad i = 1, \cdots, N.$$

For all i, we assume $\gamma_i/\mu_i < 1$ and for overload we assume $\sum_{i=1}^{N} \rho_i > 1$.

In this paper, we will focus on branching-type service disciplines in a general parameter setting which satisfy the following property (see [15]).

Property 1. *(Branching property ([15], Property 1)) If the server arrives at Q_i to find k_i customers there, then during the course of the server's visit, each of these k_i customers will effectively be replaced in an i.i.d. manner by a random population having probability generating function (p.g.f) $h_i(z_1, z_2, \ldots, z_N)$, which can be any N-dimensional p.g.f..*

According to this property, many classical service disciplines belong to the branching-type service discipline including exhaustive service discipline(per visit the server continues to serve all customers at a queue until it empties), gated service discipline (per visit the server serves only those customers at a queue which are found there upon his visit), binomial-gated [16] and binomial-exhaustive policies [15].

Furthermore, the multigated (X_i-gated) service discipline discussed in [14] is just a special case of Property 1. In Section 4, we can see that Theorem 4.2 in Section 4 is also an extention of [14] (for the special case of multigated (X_i-gated) service discipline and without customer rerouting policy (i.e. $p_{i,j} = 0, i, j = 1, \cdots, N$)). Moreover, we can know how the re-routing policy effect the queue length process from Figures 1 and 2 in Section 4.

Define $t^{(n)}$ as the time point that the server reaches right before Q_1 for the nth time and $t_i^{(n)}$ as the time point that the server reaches Q_i for the nth time ($n \in \mathbb{N} = \{0, 1, 2, \cdots\}, i = 1, 2, \ldots, N$). If the system is empty at $t^{(n)}$, then the interval $[t^{(n)}, t_1^{(n)})$ is the period of waiting until the first arrival, otherwise $t^{(n)} = t_1^{(n)}$.

Let $\mathbf{X}(t^{(n)}) = (X_1(t^{(n)}), \cdots, X_N(t^{(n)}))$, $n \in \mathbb{N}$ be the queue length process at time $t^{(n)}$, where $X_i(t^{(n)})$ is the number of customers at Q_i at time $t^{(n)}$. By Resing [15], branching property implies that the queue length sequence $\{\mathbf{X}(t^{(n)})\}_{n\in\mathbb{N}}$ forms a multi-type branching process with immigration in state $\mathbf{0}$.

In this paper, we assume that all vectors are N–dimensional row vectors, all vectors are typeset in bold italic. The vector with all coordinates equal to 0 is denoted by $\mathbf{0}$ and the vector with coordinate i equals to 1 and the other coordinates equal to 0 by $\mathbf{e}_i$.

3 The MTBP-structure of polling system

In the above section, we have known the queue length process $\{\mathbf{X}(t^{(n)})\}_{n\in\mathbb{N}}$ forms a multi-type branching process with immigration in state $\mathbf{0}$. Let $\mathbf{M} = \{m_{i,j}\}_{i,j=1}^{N}$ be the mean offspring matrix. Also let the vectors

$\mathbf{u} = (u_1, \cdots, u_N)$ and $\mathbf{v} = (v_1, \cdots, v_N)$ be the right and left eigenvectors corresponding to the maximal real-valued, positive eigenvalue θ of $\mathbf{M}$, commonly referred to as the maximum eigenvalue ([17]), normalized such that $\mathbf{vu}^\top = 1$. In this section, we will give the mean offspring matrix associated with the branching process. Moreover, Theorem 5.6.1(supercritical limit theorem) in [17] leads to our main results.

To start with, we give some notations associated with the branching-type polling system.

- Define $\check{\mathbf{L}}_i = (\check{L}_{i,1}, \cdots, \check{L}_{i,N})$ as the visit offspring of a customer at Q_i, which equals in distribution to $\mathbf{X}(t_{i+1}^{(n)})$ given that $\mathbf{X}(t_i^{(n)}) = \mathbf{e}_i$ (its distribution does not depend on n) with $\check{\mathbf{m}}_i = (\check{m}_{i,1}, \cdots, \check{m}_{i,N}) = \mathbb{E}\check{\mathbf{L}}_i$.
- Define $\mathbf{L}_i := (L_{i,1}, \cdots, L_{i,N})$ as the session offspring of a customer at Q_i, which equals in distribution to $\mathbf{X}(t^{(n+1)})$ given that $\mathbf{X}(t^{(n)}) = \mathbf{e}_i$ (its distribution does not depend on n) with $\mathbf{m}_i = (m_{i,1}, \cdots, m_{i,N}) = \mathbb{E}\mathbf{L}_i$. In order to ensure the non-degenerate, we assume that $\mathbb{E}L_{i,j} \log L_{i,j} < \infty$ for all $1 \le i, j \le N$.

To proceed, we need further to define the exhaustiveness f_i of the service descipline at Q_i by (see [18], (55), (56))

$$f_i = 1 - \frac{\partial}{\partial z_i} h_i(z_1, z_2, \cdots, z_N)|_{\mathbf{z}=\mathbf{1}} = 1 - \mathbb{E}\check{L}_{i,i}.$$

It has an appealing interpretation: during the course of the server's visit at Q_i, each customer present at the start of the visit to Q_i will be replaced by a number of customers with mean $1 - f_i$ at the end of the visit to Q_i.

Remark 3.1. *In particular, the exhaustiveness of the multigated (X_i-gated) service discipline at Q_i in [14] (where $X_i = 1$and ∞ of gating index corresponding to conventional gated and exhaustive, respectively) equals $f_i = 1 - \mathbb{E}(\frac{\lambda_i}{\mu_i} + p_{i,i})^{X_i}$.*

Let B_i^E be the total service time of a customer in Q_i before he is either routed to Q_j, $j \neq i$ or leaves the system. Then $b_i^E := \mathbb{E}B_i^E = 1/(\mu_i(1 - p_{i,i}))$.

Define T_i as the busy period in Q_i. This busy period consists of the service of its first customer at Q_i, the services of the customers arriving at Q_i during the service of the first customer (i.e., the children), the services of the customers arriving at Q_i during the service of the children (i.e., the grandchildren), and so forth. Then, we have

$$t_i := \mathbb{E}T_i = f_i \frac{b_i^E}{1 - \lambda_i b_i^E}. \tag{3.1}$$

By Lemma 1 in [9], the mean offspring matrix $\mathbf{M}$ is given in the following proposition.

Proposition 3.1. *For the cyclic branching-type polling system, the mean matrix $\mathbf{M}$ is given by*

$$\mathbf{M} = \mathbf{M}_1 \dots \mathbf{M}_N,$$

where $\mathbf{M}_k = \left(m_{i,j}^{(k)}\right)$ and

$$m_{i,j}^{(k)} = \begin{cases} \delta_{\{i=j\}}, & i \neq k, \\ 1 - f_i, & i = k = j, \\ t_i(\mu_i p_{i,j} + \lambda_j), & i = k \neq j, \end{cases} \tag{3.2}$$

where δ_F denotes the indicator function on F.

Actually, $\mathbf{M}_k$ is the mean session offspring during the visit time on Q_k. Hence, for all i,

$$m_{i,j}^{(i)} = \check{m}_{i,j} = \begin{cases} 1 - f_i, & i = j, \\ t_i(\mu_i p_{i,j} + \lambda_j), & i \neq j. \end{cases}$$

Proof. In the Multi-type branching process, by the definition of $M = \{m_{i,j}\}_{i,j=1}^N$, we have that $m_{i,j} = \frac{\partial f^{(i)}(\mathbf{z})}{\partial z_j}\Big|_{\mathbf{z}=\mathbf{1}}$, $i, j = 1, \cdots, N$, where $f^{(i)}(\mathbf{z})$ is the ith generating function of the distribution of the number of offspring of various types to be produced by a type i particle, and

$$f^{(i)}(\mathbf{z}) = \sum_{j_1, \cdots, j_N \ge 0} p^{(i)}(j_1, \cdots, j_N) z_1^{j_1} \cdots z_n^{j_N}, |z_k| \le 1, i, k = 1, \cdots, N,$$

where $\mathbf{z} = (z_1, \cdots, z_N)$ and $p^{(i)}(j_1, \cdots, j_N)$ = the probability that a type i parent produces j_1 particles of type 1, j_2 of type 2, $\cdots$, j_N of type N.

Let $\mathbf{X}(t^{(n)}) = (X_1(t^{(n)}), \cdots, X_N(t^{(n)}))$, $n \in \mathbb{N}$ be the queue length process at time $t^{(n)}$, where $X_i(t^{(n)})$ is the number of customers at Q_i at time $t^{(n)}$. By Resing [15], branching property implies that the queue length sequence $\{\mathbf{X}(t^{(n)})\}_{n\in\mathbb{N}}$ forms a multi-type branching process with immigration in state $\mathbf{0}$. Therefore, we have

$$f^{(i)}(\mathbf{z}) = h_i(z_1, z_2, \cdots, z_i, f^{(i+1)}(\mathbf{z}), \cdots, f^{(N)}(\mathbf{z})), |z_k| \le 1, i = 1, \cdots, N,$$

where

$$h_i(\mathbf{z}) = \psi_i\Big(\sum_{j\neq i}\lambda_j(1-z_j), z_i, \frac{p_{i,0}}{1-p_{i,i}} + \sum_{j\neq i}\frac{p_{i,j}}{1-p_{i,i}}z_j\Big), i = 1, \cdots, N,$$

where $\psi_i(u, z_1, z_2) = \mathbb{E}(e^{-uT_i}, z_1^{L_i}, Z_2^{M_i})$, L_i is the so-called busy period residue, i.e., the number of type-i children of the original customer that generates this busy period and M_i is the number of customers leave Q_i in the busy period T_i. Then, by the definition of M, we obtain,

$$m_{i,j} = h_{i,j}I_{\{j\le i\}} + \sum_{k=i+1}^{N} h_{i,k}m_{k,j}, i, j = 1, \cdots, N, \tag{3.3}$$

$$h_{i,j} = \frac{\partial h_i(\mathbf{z})}{\partial z_j}\Big|_{\mathbf{z}=\mathbf{1}} = \begin{cases} 1-f_i, i = j, \\ t_k(\lambda_j + p_{i,j}\mu_i), i \neq j, \end{cases} \tag{3.4}$$

Therefore, by (3.3) and (3.4), we obtain (3.2). □

It follows that the auxiliary process $\{\mathbf{X}(t^{(n)})\}_{n\in\mathbb{N}}$ has the following asymptotics (see [17], Theorem 5.6.1), which will be important for proving the main results in the next section.

Proposition 3.2. *If the first arriving customer arrives at Q_i after $t = 0$, then*

$$\frac{\mathbf{X}(t^{(n)})}{\theta^n} \to \zeta_i\mathbf{v} \qquad \text{almost surely (a.s.)} \qquad \text{as } n \to \infty,$$

where the distribution of the random variable ζ_i has a jump of magnitude $q_i = \mathbb{P}(\mathbf{X}(t^{(n)}) = \mathbf{0}$ for some $n|\mathbf{X}(t^{(0)}) = \mathbf{e}_i) < 1$ at 0 and a continuous density function on $(0, \infty)$ and $\mathbb{E}\zeta_i = u_i$.

4 Main results

To give the main results, three more notations are needed.

- Let $\bar{B}_i$ be the total service time of a customer arriving at Q_i from outside, $\bar{c}_i = \mathbb{E}\bar{B}_i$, we have $\bar{c}_i = 1/\mu_i + \sum_{j=1}^{N} p_{i,j}\bar{c}_j$. It is also easy to deduce that $\rho = \sum_{i=1}^{N}\lambda_i\bar{c}_i$.
- For $n \in \mathbb{N}$, let $\eta_n := \begin{cases} \min\{k : t^{(k)} \ge \theta^n\}, & \text{if } n \ge 0; \\ 0, & \text{if } n < 0. \end{cases}$
- For $n \in \mathbb{N}$, define the scaled queue length process $\bar{\mathbf{X}}^{(n)}(t) := \frac{\mathbf{X}(\theta^n t)}{\theta^n}$, $t \in [0, \infty)$.

Theorem 4.1. *There exist constants $\bar{b}_i \in (0, \infty)$ and $\bar{\mathbf{a}}_i = (\bar{a}_{i,1}, \cdots, \bar{a}_{i,N}) \in [0, \infty)^N$, $i = 1, \cdots, N+1$, and a random variable ξ with values in $[1, \theta)$ such that, for all $k \in \mathbb{N}$ and $i = 1, 2, \cdots, N$,*

$$\frac{t_i^{(\eta_n+k)}}{\theta^n} \to \theta^k\bar{b}_i\xi \qquad \text{and} \qquad \frac{\mathbf{X}(t_i^{(\eta_n+k)})}{\theta^n} \to \xi\theta^k\bar{\mathbf{a}}_i \quad a.s. \quad \text{as} \quad n \to \infty.$$

The $\bar{b}_i$ and $\bar{\mathbf{a}}_i$ are given by

$$\bar{b}_1 = 1, \qquad \bar{b}_{i+1} = \bar{b}_i + \left[\frac{\nu_i}{\alpha} + \lambda_i(\bar{b}_i - \bar{b}_1) + \sum_{j=1}^{i-1} p_{j,i}\mu_j(\bar{b}_{j+1} - \bar{b}_j)\right] t_i$$

and for $j = 1, \cdots, N$,

$$\bar{\mathbf{a}}_1 = \frac{\mathbf{v}}{\alpha}, \quad \bar{a}_{i+1,j} = \begin{cases} \bar{a}_{i,j} + [\lambda_j + \mu_i p_{i,j}](\bar{b}_{i+1} - \bar{b}_i) & j \neq i, \\ \bar{a}_{i,i} + [\lambda_i - \mu_i(1 - p_{i,i})](\bar{b}_{i+1} - \bar{b}_i), & j = i. \end{cases}$$

where

$$\alpha = \frac{\sum_{i=1}^{N} \nu_i \bar{c}_i}{\rho - 1}.$$

The distribution of ξ see [14].

Proof. See Section 5. □

Remark 4.1. *Specially, for a polling model with multigated (X_i-gated) service discipline and without customer rerouting policy (i.e. $p_{i,j} = 0, i, j = 1, \cdots, N$), the asymptotics in Theorem 4.1 remains true while the $\bar{b}_i$ and $\bar{\mathbf{a}}_i$ turn to be*

$$\bar{b}_1 = 1, \qquad \bar{b}_{i+1} = \bar{b}_i + \left[\frac{\nu_i}{\alpha} + \lambda_i(\bar{b}_i - \bar{b}_1)\right] t_i$$

and for $j = 1, \cdots, N$,

$$\bar{\mathbf{a}}_1 = \frac{\mathbf{v}}{\alpha}, \quad \bar{a}_{i+1,j} = \begin{cases} \bar{a}_{i,j} + \lambda_j(\bar{b}_{i+1} - \bar{b}_i) & j \neq i, \\ \bar{a}_{i,i} + [\lambda_i - \mu_i](\bar{b}_{i+1} - \bar{b}_i) & j = i. \end{cases}$$

where $t_i = \mathbb{E}T_i = \frac{1 - \mathbb{E}(\frac{\lambda_i}{\mu_i})^{X_i}}{\mu_i - \lambda_i}$ and $\alpha = \frac{\sum_{i=1}^{N} \nu_i/\mu_i}{\rho - 1}$, which is in accord with Theorem 1 in [14].

Theorem 4.2. *There exists a deterministic function $\bar{\mathbf{X}}(\cdot) = (\bar{X}_1, \cdots, \bar{X}_N)(\cdot) \in [0, \infty)^N$ such that,*

$$\bar{\mathbf{X}}^{(n)}(\cdot) \to \xi \bar{\mathbf{X}}(\frac{\cdot}{\xi}) \quad a.s. \qquad as \quad n \to \infty, \tag{4.1}$$

uniformly on compact sets. For all $i = 1, \cdots, N$, the function $\bar{\mathbf{X}}(\cdot)$ is continuous and piecewise linear as depicted in Figure 1 and specified by

$$\bar{\mathbf{X}}(t) = \begin{cases} 0, & \text{if } t = 0; \\ \theta^k \bar{\mathbf{a}}_i + (t - \theta^k \bar{b}_i)\boldsymbol{\lambda} + (t - \theta^k \bar{b}_i)\mu_i \mathbf{p}_i, & \text{if } t \in [\theta^k \bar{b}_i, \theta^k \bar{b}_{i+1}), \end{cases} \tag{4.2}$$

where $\boldsymbol{\lambda} = [\lambda_1, \cdots, \lambda_N]$ and $\mathbf{p}_i = [p_{i,1}, \cdots, p_{i,i-1}, p_{i,i} - 1, p_{i,i+1}, \cdots, p_{i,N}]$.

Proof. See Section 5. □

Remark 4.2. *Theorem 4.2 is also an extention of [14] (for the special case of multigated (X_i-gated) service discipline and without customer rerouting policy (i.e. $p_{i,j} = 0, i, j = 1, \cdots, N$)).*

Corollary 4.1. *Under passage to the fluid dynamics, the fluid total population $(\bar{X}_1 + \bar{X}_2 + \cdots + \bar{X}_N)(\cdot)$ grows at the rate*

$$(\lambda_1 + \cdots + \lambda_N) - p_{i,0}\mu_i$$

when $t \in [\theta^k \bar{b}_i, \theta^k \bar{b}_{i+1})$ for all $k \in \mathbb{N}$ and $i = 1, \cdots, N$. (see Figure 2)

Remark 4.3. *In [9], the fluid asymptotics of the queue length process in the heavy traffic for the same polling model have been discussed. In the heavy traffic, the total scaled workload is effectively constant while the*

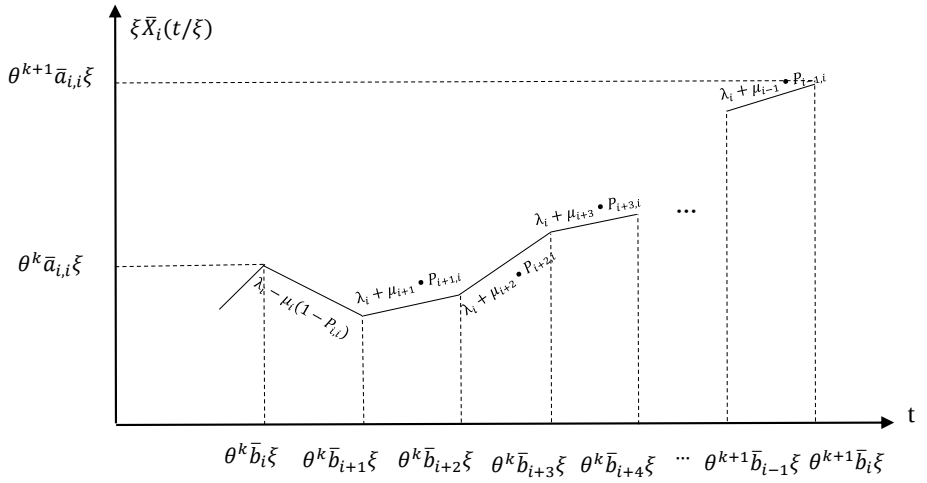

Figure 1: Fluid limit of Q_i.

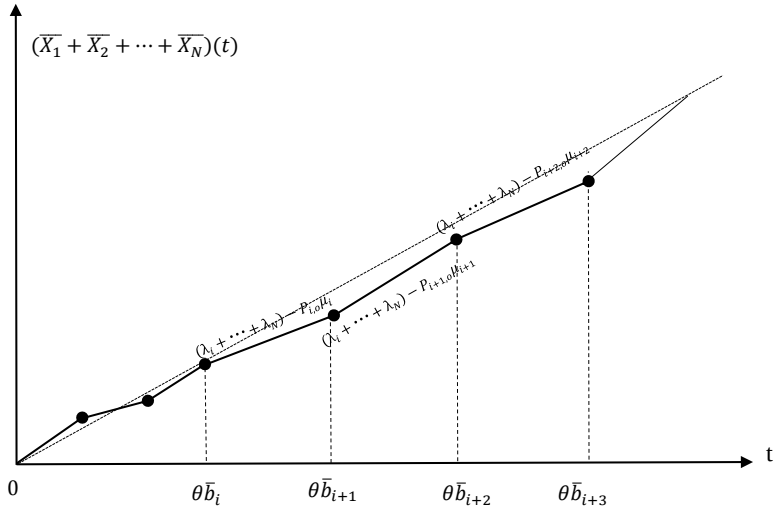

Figure 2: Fluid limit of total population $\bar{X}_1 + \cdots + \bar{X}_N$.

individual queue workload is emptied and refilled at a rate during the course of a cycle. In contrast to the heavy traffic asymptotics, the total overloaded asymptotic workload is always increasing as shown in Corollary 4.1 during the course of a cycle. In addition, the overloaded fluid limit always contains a random variable ξ. However, the individual queue workload is emptied and refilled at the same rate as in the heavy traffic like a fluid model. Hence, our result is a further progress of [9].

Remark 4.4. *For different branching-type service discipline, our main results have shown that the overloaded fluid asymptotics just depend on the exhaustiveness of each service discipline, which also applies to the heavy traffic asymptotics in [9] and the asymptotics with the large-switchover times in [18]. This can be easily interpreted by the fluid approximation. It also proves that the branching-type polling system deserves much more attention.*

Remark 4.5. *The rerouting policy only affects the flow rate both in the heavy traffic asymptotics (see [9]) and in the overloaded asymptotics. In Theorem 4.2, we can see that the fluid limit depends on the re-routing probability $p_{i,j}$. Upon completion of service at Q_i, $i = 1, \cdots, N$, a customer is either routed to Q_i with probability $p_{i,i}$, which*

leads the decreasing rate of the length of Q_i to be $\lambda_i - \mu_i(1-p_{ii})$, or routed to Q_j, $j \neq i$ with probability $p_{i,j}$, which leads the increasing rate of the length of Q_j to be $\lambda_j + \mu_i p_{i,j}$ at time $[\theta^k \bar{b}_i, \theta^k \bar{b}_{i+1})$.

According to Theorem 4.2 and Corrollary 4.1, the fluid limit processes both demonstrate an oscillation waveform with increasing amplitude and cycle time over time. To be more specific, the amplitude and cycle time both increase by $\theta - 1$ times each cycle. Hence, the average growth rate of the scaled total population, denoted by β, equals to the average growth rate in each cycle, as shown in Figure 2. Therefore, we have

$$\sum_{i=1}^{N} \frac{\sum_{j=1}^{N} \bar{a}_{i,j} + \sum_{j=1}^{N} \bar{a}_{i+1,j}}{2} (\bar{b}_{i+1} - \bar{b}_i) = \int_{\bar{b}_1}^{\theta \bar{b}_1} \beta t dt,$$

which yields

$$\beta = \frac{1}{\theta^2 - 1} \left[\sum_{i=1}^{N} \left(\sum_{j=1}^{N} \bar{a}_{i,j} + \sum_{j=1}^{N} \bar{a}_{i+1,j} \right) (\bar{b}_{i+1} - \bar{b}_i) \right]. \tag{4.3}$$

By the definition of the scaled queue length process, the fluid limit could approximate the original queue length process in steady state. Furthermore, the average growth rate in (4.3) allows us to study the optimization problem of how to choose the gating indexes of each queue to minimize the total queue length. Since each of the queues adheres to a branching-type service discipline, we study how to choose the exhaustiveness f_i with the same objective in mind. We would provide an optimization example by utilizing the genetic algorithm in Section 6.

5 Proof of Theorems 4.1 and 4.2

5.1 Proof of Theorem 4.1

Proof. By the tool of Lemma 8 in [14], if we can prove

$$\frac{t_i^{(n)}}{\theta^n} \to b_i \xi \quad \text{and} \quad \frac{\mathbf{X}(t_i^{(n)})}{\theta^n} \to \xi \mathbf{a}_i \quad a.s. \quad \text{as} \quad n \to \infty, \tag{5.1}$$

where $b_i = \alpha \bar{b}_i$, $\mathbf{a}_i = \alpha \bar{\mathbf{a}}_i$, then Theorem 4.1 is concluded. Hence, we focus on the proof of (5.1).

(1) Limit of $t_1^{(n)}/\theta^n$. Define index ν by

$$\nu = \max \left\{ n \in \mathbb{N}^+, \text{such that } \mathbf{X}(t^{(n)}) = \mathbf{0} \text{ and } \mathbf{X}(t^{(m)}) \neq \mathbf{0} \text{ for all } m > n \right\}.$$

By the total workload process, we have, for $n > \nu$,

$$t_1^{(n)} = t^{(n)} = W + t^{(n)} A_1^{(n)} - \theta^n A_2^{(n)}, \tag{5.2}$$

where $\bar{B}_i^{(k)}$ are i.i.d. copies of $\bar{B}_i$. By definition of ν, we know that it is a.s. finite, so that we obtain $W = \sum_{l=0}^{\nu} (t_1^{(l)} - t^{(l)}) < \infty$ a.s..

$$A_1^{(n)} = \sum_{i=1}^{N} \frac{\sum_{k=1}^{E_i(t^{(n)})} \bar{B}_i^{(k)}}{E_i(t^{(n)})} \frac{E_i(t^{(n)})}{t^{(n)}}, A_2^{(n)} = \sum_{i=1}^{N} \frac{\sum_{k=1}^{X_i(t^{(n)})} \bar{B}_i^{(k)}}{X_i(t^{(n)})} \frac{X_i(t^{(n)})}{\theta^n},$$

Since, we know that $\bar{B}_i^{(k)}$ are i.i.d. copies of $\bar{B}_i$, and $\mathbb{E}\bar{B}_i = \bar{c}_i$, then, by the SLLN and Proposition 3.2, we obtain, as $n \to \infty$,

$$A_1^{(n)} \to \sum_{i=1}^{N} \lambda_i \mathbb{E}\bar{B}_i^{(k)} = \sum_{i=1}^{N} \lambda_i \bar{c}_i = \rho, \quad A_2^{(n)} \to \sum_{i=1}^{N} \nu_i \xi \mathbb{E}\bar{B}_i^{(k)} = \sum_{i=1}^{N} \nu_i \bar{c}_i \xi \quad a.s..$$

Therefore, by (5.2), we have, as $n \to \infty$,

$$\frac{t^{(n)}}{\theta^n} \to b_1\xi \quad \text{and} \quad \frac{t_1^{(n)}}{\theta^n} \to b_1\xi,$$

where $b_1 = \frac{\sum_{i=1}^N v_i\bar{c}_i}{\rho-1}$.

(2) Limit of $t_i^{(n)}/\theta^n$. In (1), by utilizing the index v and equation $t_1^{(n)} = t^{(n)}$, we proved $\lim_{n\to\infty} t_1^{(n)}/\theta^n = b_1\xi$. By the symmetry, there also exist positive numbers b_i such that

$$\lim_{n\to\infty} \frac{t_i^{(n)}}{\theta^n} = b_i\xi, \quad i = 1, \cdots, N.$$

It remains to prove the iteration of b_i, which refers to (4) below.

(3)Limit of $X_j(t_i^{(n)})/\theta^n$. Define the renewal processes

$$Y_i(t) = \max\left\{ n \in \mathbb{N}^+, \text{such that} \sum_{i=1}^n B_i^{(k)} \le t \right\},$$

where $B_i^{(k)}$ are i.i.d. copies of B_i. Also let $I_i(t)$ be the whole time that the server has spent at Q_i before time t, i.e.,

$$I_i(t) = \int_0^t I(\text{ queue } i \text{ is in service in time } s)ds \qquad t \in (0, \infty).$$

Let A_i be the position of a customer after completion of service at Q_i, for $i = 1, \cdots, N$, i.e.,

$$A_i = \begin{cases} j, & \text{after receiving service at } Q_i, \text{ a customer is routed to } Q_j; \\ 0, & \text{after receiving service at } Q_i, \text{ a customer leaves the system.} \end{cases}$$

Then $\mathbb{P}(A_i = j) = p_{i,j}$, $j = 0, 1, \cdots, N$. Hence, we have

$$X_j(t_{i+1}^{(n)}) = \begin{cases} X_j(t_i^{(n)}) + E_j(t_{i+1}^{(n)}) - E_j(t_i^{(n)}) + \sum_{k=1}^{Y_i(I_i(t^{n+1}))-Y_i(I_i(t^n))} \delta_{\{A_i^{(k)}=j\}}, & j \ne i; \\ X_i(t_i^{(n)}) + E_i(t_{i+1}^n) - E_i(t_i^{(n)}) - \sum_{k=1}^{Y_i(I_i(t^{n+1}))-Y_i(I_i(t^n))} \delta_{\{A_i^{(k)}\ne i\}}, & j = i, \end{cases} \tag{5.3}$$

where $A_i^{(k)}$ are i.i.d. copies of A_i. Since $\mathbb{E}\delta_{\{A_i^{(k)}=j\}} = \mathbb{P}(A_i^{(k)} = j) = p_{i,j}$ is finite, so that, by SLLN, we have

$$\frac{X_j(t_{i+1}^{(n)})}{\theta^n} \to a_{i+1,j}\xi,$$

where

$$a_{i+1,j} = \begin{cases} a_{i,j} + \lambda_j(b_{i+1} - b_i) + p_{i,j}\mu_j(b_{i+1} - b_i), & j \ne i; \\ a_{i,i} + \lambda_i(b_{i+1} - b_i) - \mu_i(1 - p_{i,i})(b_{i+1} - b_i), & j = i. \end{cases}$$

(4) The iteration of b_i. Recall that

$$t_{i+1}^{(n)} = t_i^{(n)} + \sum_{k=1}^{X_i(t_i^{(n)})} T_i^{(k)}, \tag{5.4}$$

$$X_i(t_i^{(n)}) = X_i(t_1^{(n)}) + E_i(t_i^{(n)}) - E_i(t_1^{(n)}) + \sum_{j=1}^{i-1} \sum_{k=1}^{Y_j(I_j(t^{(n)}))-Y_j(I_j(t^{(n-1)}))} \delta_{\{A_j^{(k)}=i\}}, \tag{5.5}$$

where $T_i^{(k)}$ are i.i.d. copies of T_i. As $n \to \infty$, if $X_i(t_1^{(n)}) = c$, where c is a positive constant, then $\sum_{k=1}^{X_i(t_i^{(n)})} T_i^{(k)}/\theta^n \to 0$, so that $b_{i+1} = b_i$, this case is obviously to us, so we study the case that $X_i(t_1^{(n)}) \to \infty$. θ^n corresponds to T_n in the Proposition 2 of [14], $T_i^{(k)}$ corresponds to $Y_n^{(k)}$ in Proposition 2 of [14], $a_{i,i}$ corresponds

to τ in the Proposition 2 of [14], t_i corresponds to $\mathbb{E}Y$ in Proposition 2 of [14]. Therefore, by Proposition 2 in [14] and (5.4), we get

$$b_{i+1} - b_i = a_{i,i}t_i. \tag{5.6}$$

By (5.5) and the SLLN, we obtain

$$a_{i,i} = \nu_i + \lambda_i(b_i - b_1) + \sum_{j=1}^{i-1} \mu_j(b_{j+1} - b_j)p_{j,i}. \tag{5.7}$$

Then the iteration of b_i can be proved by substituting (5.7) into (5.6).

(5)The equivalence of b_{N+1} and θb_1. By $t_{N+1}^{(n)} = t_1^{(n+1)}$, we obtain $b_{N+1}\xi = \lim_{n\to\infty} t_{N+1}^{(n)}/\theta^n = \lim_{n\to\infty} \theta t_1^{(n+1)}/\theta^{n+1} = \theta b_1 \xi$, i.e. $b_{N+1} = \theta b_1$. This can be proved as follows. By the definition of $\mathbf{M}_i$ in Lemma 3.1, it is easy to give

$$(\mathbf{M}_i - \mathbf{I})\bar{\mathbf{c}}^T = (\rho - 1)t_i\mathbf{e}_i,$$

where $\mathbf{I}$ is the identity matrix, $\cdot^\top$ denotes the operation of transposition and $\bar{\mathbf{c}} = (\bar{c}_1, \ldots, \bar{c}_N)$. Substituting the above equation into (5.6) yields

$$b_{i+1} - b_i = a_{i,i}t_i = \mathbf{a}_i\mathbf{e}_i t_i = \frac{1}{\rho - 1}\mathbf{a}_i(\mathbf{M}_i - \mathbf{I})\bar{\mathbf{c}}^T = \frac{1}{\rho - 1}\left(\mathbf{a}_{i+1} - \mathbf{a}_i\right)\bar{\mathbf{c}}^T.$$

where $\mathbf{a}_i = (a_{i,1}, \cdots, a_{i,N})$, which gives immediately

$$\begin{aligned} b_{N+1} =& \sum_{i=1}^{N}(b_{i+1} - b_i) + b_1 = \frac{1}{\rho - 1}\left(\mathbf{a}_{N+1} - \mathbf{a}_1\right)\bar{\mathbf{c}}^T + b_1 \\ =& \frac{1}{\rho - 1}\left(\mathbf{v}\mathbf{M} - \mathbf{v}\right)\bar{\mathbf{c}}^T + b_1 = (\theta - 1)\frac{\mathbf{v}\bar{\mathbf{c}}^T}{\rho - 1} + b_1 = (\theta - 1)b_1 + b_1 = \theta b_1. \end{aligned}$$

□

5.2 Proof of Theorem 4.2

Proof. For each i, by (4.2), we know that the function $\bar{X}_i(\cdot)$ might have discontinuities only at $t = 0$ and $t = \theta^k\bar{b}_i$ for each $k \in \mathbb{N}$. Since the function $\bar{X}_i(\cdot)$ is càdlàg, the continuity of $\bar{X}(\cdot)$ is evident in combination with the definition of $\mathbf{a_i}$. Additionally, the uniform convergence on compact sets can be proved in the same way as in the proof of Theorem 2 in [14]. Hence, it suffices to prove the point-wise convergence (4.2) for each $i = 1, 2, \ldots, N$.

For $t = 0$, the convergence of (4.2) holds since the system starts empty. For each $i = 1, \ldots, N$, if $t \in [\theta^k\bar{b}_i, \theta^k\bar{b}_{i+1})$, it remains to prove

$$\bar{X}_j(t) = \begin{cases} \theta^k\bar{a}_{i,i} + [\lambda_i - \mu_i(1 - p_{i,i})](t - \theta^k\bar{b}_i), & j = i; \\ \theta^k\bar{a}_{i,j} + [\lambda_j + \mu_i p_{i,j}](t - \theta^k\bar{b}_i), & j \neq i. \end{cases}$$

For all n big enough, $\frac{t_i^{(\eta_n+k)}}{\theta^n} < t < \frac{t_{i+1}^{(\eta_n+k)}}{\theta^n}$ implying that Q_i is in service during $[t_i^{(\eta_n+k)}, \theta^n t)$. Hence,

$$X_j(\theta^n t) = \begin{cases} X_i(t_i^{(\eta_n+k)}) + E_i(\theta^n t) - E_i(t_i^{(\eta_n+k)}) - \sum_{k=1}^{Y_i(I_i(\theta^n t)) - Y_i(I_i(t^{(\eta_n+k)}))} \delta_{\{A_i^{(k)} \neq i\}}, & j = i; \\ X_j(t_i^{(\eta_n+k)}) + E_j(\theta^n t) - E_j(t_i^{(\eta_n+k)}) + \sum_{l=1}^{Y_i(I_i(\theta^n t)) - Y_i(I_i(t^{(\eta_n+k)}))} \delta_{\{A_i^{(l)} = j\}}, & j \neq i. \end{cases}$$

Combining the above equation with Theorem 4.1, we have

$$\bar{X}_j^{(n)}(t) = \frac{X_j(\theta^n t)}{\theta^n} \to \begin{cases} \xi\theta^k\bar{a}_{i,i} + [\lambda_i - \mu_i(1 - p_{i,i})](t - \theta^k\bar{b}_i\xi), & j = i; \\ \xi\theta^k\bar{a}_{i,j} + (\lambda_j + \mu_i p_{i,j})(t - \theta^k\bar{b}_i\xi), & j \neq i, \end{cases}$$

where the right hand-side actually equals $\xi\bar{X}_j(\frac{t}{\xi})$. Therefore, we proved (4.2). □

6 Numerical validation and optimization of gating indexes

6.1 Numerical validation

This subsection is devoted to test the validity of the fluid limits of the scaled queue length process. For simplicity, we consider a 3-queue polling system described in Table 1 with exponentially distributed service time. For this model, it is readily to obtain $\rho_1 = 0.4749$, $\rho_2 = 0.5194$, $\rho_3 = 0.8625$ and $\rho = 1.8568$, which belongs to the overloaded traffic case studied in this paper.

We utilize the SimEvents toolbox of Matlab to undertake the simulations of the polling networks. The exhaustive and gated service policies are taken for example and some vital variables are given in Table 2. In order to illustrate the convergence of the scaled queue length process, we take $n = 1, 5, 8, 10$ in polling network with exhaustive service policies and $n = 1, 10, 18, 20$ in the gated counterpart. The corresponding scaled queue length process of Q_2 and the scaled total queue length process are depicted in Figure 3 and Figure 4, respectively. Apparently, the scaled queue length sample paths get closer and closer as n increases.

Table 1: Parameter values in 3-queue polling network.

Parameter	Considered parameter values
Arriving rate	$\lambda_1 = \lambda_2 = \lambda_3 = 1$
Service rate	$\mu_1 = 8, \mu_2 = 5, \mu_3 = 2$
Transition probability	$P = (p_{i,j})_{3\times 3} = \begin{pmatrix} 0.1 & 0.25 & 0.2 \\ 0.2 & 0.1 & 0.2 \\ 0.2 & 0.1 & 0.25 \end{pmatrix}$

Table 2: Essential variable values in 3-queue polling network.

Variable	Values	
	Exhaustive	Gate
Gating index	∞	1
Exhaustiveness	1	$1 - (\frac{\lambda_i}{\mu_i} + p_{i,i})$
Maximum eigenvalue	$\theta = 3.7497$	$\theta = 1.6394$
Left eigenvector	$v = [0.9731, 0.683, 0]$	$v = [0.7454, 0.5301, 0.4774]$

Moreover, as shown in Figure 3 and Figure 4, the fluid limit processes both demonstrate an oscillation waveform with increasing amplitude and cycle time forward and oscillate at an infinite rate when approaching zero. According to Theorem 4.2, the amplitude and cycle time increase by $\theta - 1$ times within each cycle, which has been easily illustrated by the sample paths.

6.2 Optimization of gating indexes

Subsequently, we consider the optimization of the gating indexes by numerical method. We assume that the gating indexes are integers. Virtually, the fluid limits only depend on the exhaustiveness of the service discipline at each queue (moment of gating index), which allows us to minimize the total queue length through the accommodation of the integer gating indexes.

By (4.3), the average growth rate of the total queue length process β with exhaustive and gated service policies equals 1.5025 and 1.2416 respectively (see Figure 5). This can be intuitively interpreted from the growth

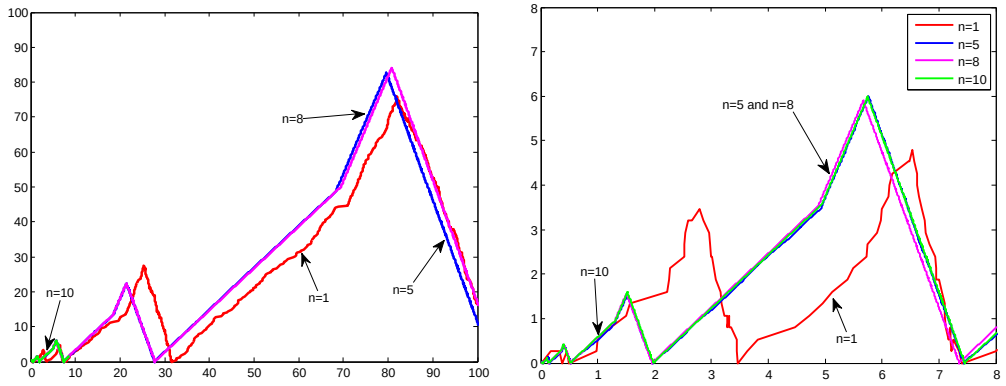

Figure 3: The scaled queue length process of Q_2 for different n (left:$t \in [0, 100]$, right:$t \in [0, 8]$) with exhaustive service policy.

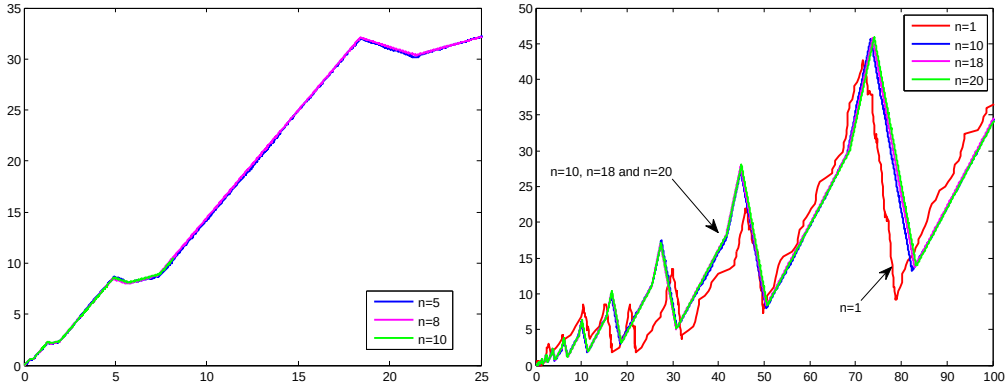

Figure 4: Left: the scaled total queue length process for different n with exhaustive service policy.
Right: the scaled queue length process of Q_2 for different n with gated service policy.

rate during each visiting period on different queues. The visiting period at Q_3 with the maximal growth rate (minimal service rate) takes 4 times as much time as others in exhaustive service policy. Instead, it takes less than 2 times as much time as others in gated service policy. Therefore, to minimize the average growth rate, we need to increase the visiting time at Q_1 and Q_2 and decrease the visiting time at Q_3.

To optimize the gating indexes turns to be an integer programming with three variables. The GA toolbox of Matlab is undertaken here to search for the optimal gating indexes. For our model, it just takes 51 iterations to find the optimal solution: Q_1 and Q_2 both take exhaustive service policy while Q_3 takes gated service policy. The minimal average growth rate equals to 1.19262 and the corresponding exhaustiveness is $f_1 = f_2 = 1$, $f_3 = 0.25$. Figure 5 depicts the process of the optimal average growth rate in each generation. Apparently, the convergence process turns to be very effective. Hence, the average growth rate of the fluid limit provides a simple and transparent method to optimize the gating index.

7 Conclusions and further Research

Inspired by [14], we present the fluid limit of an overloaded polling system with general branching-type service discipline and customer re-routing policies. These results provide new fundamental insight in the impact of exhaustiveness. As a by-product, we propose an optimization problem of gating indexes to minimize the total queue length.

This work gives rise to a variety of directions for further research. A logical follow-up step would be to study the case with non-zero switch-over time. In addition, the asysmptotic behaviors of discrete-time polling

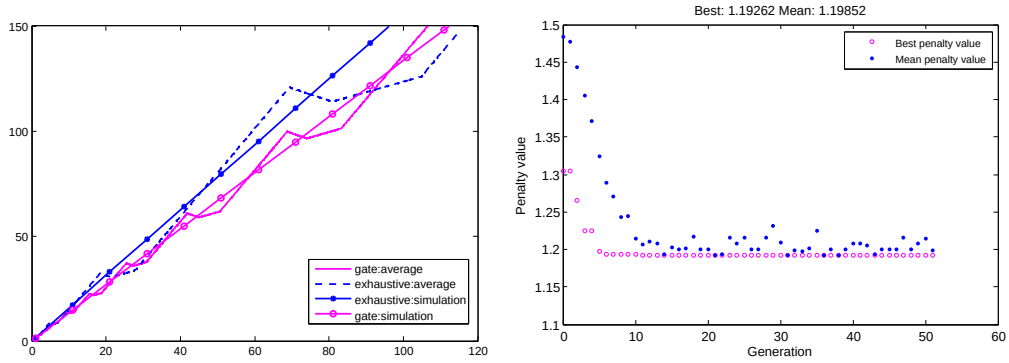

Figure 5: Left: the fluid limit of the scaled total queue length process (exhaustive and gated).
Right: the convergence process of the optimal average growth rate by GA solver.

systems are also direct extensions to this study. Furthermore, the fluid limit allows us to propose control strategies of the growth depression, which requires substantially more effort.

Acknowledgement: This research is partially supported by the National Natural Science Foundation of China (11901186, 11671404) and by the Fundamental Research Funds for the Central Universities (WUT: 2017 IVA 069, WUT: 2018IB018). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

[1] Boon M.A., Van der Mei R.D., Winands E.M., Queueing networks with a single shared server: light and heavy traffic, ACM SIGMETRICS Perform. Eval. Rev., 2011, 39(2), 44–46.
[2] Boon M.A., Van der Mei R.D., Winands E.M., Waiting times in queueing networks with a single shared server, Queueing Syst., 2013, 74(4), 403–429.
[3] Boon M.A., van der Mei R.D., Winands E.M., Heavy traffic analysis of roving server networks, Stochastic Models., 2017, 33(2), 171–209.
[4] Sidi M., Levy H., Customer routing on polling systems, In: King P.J.B. , Mitrani I., Pooley R.J. (Eds.), Performance, North-Holland, Amsterdam, 1990, 319–331.
[5] Sidi M., Levy H., Fuhrmann S.W., A queueing network with a single cyclically roving server, Queueing Syst., 1992, 11(1), 121–144.
[6] Coffman Jr E., Puhalskii A., Reiman M., Polling systems with zero switchover times: a heavy-traffic averaging principle, Ann. Appl. Probab., 1995, 681–719.
[7] Coffman Jr E., Puhalskii A., Reiman M., Polling systems in heavy traffic: A bessel process limit, Math. Oper. Res., 1998, 23(2), 257–304.
[8] Van der Mei R.D., Towards a unifying theory on branching-type polling systems in heavy traffic, Queueing Syst., 2007, 57(1), 29–46.
[9] Liu Z., Chu Y., Wu J., The asymptotic behavior of a branching-type polling network in heavy traffic, Sci. China-Math. (Chinese Series), 2015, 45(5), 515–526.
[10] Liu Z., Chu Y., Wu J., Heavy-traffic asymptotics of priority polling system with threshold service policy, Comput. Oper. Res., 2016, 65, 19–28.
[11] Puha A.L., Stolyar A.L., Williams R.J., The fluid limit of an overloaded processor sharing queue, Math. Oper. Res., 2006, 31(31), 316–350.
[12] Talreja R., Whitt W., Fluid models for overloaded multiclass many-server queueing systems with first-come, first-served routing, Manage. Sci., 2008, 54(8), 1513–1527.
[13] Jennings O.B., Reed J.E., An overloaded multiclass FIFO queue with abandonments, Oper. Res., 2012, 60(5), 1282–1295.
[14] Remerova M., Foss S., Zwart B., Random fluid limit of an overloaded polling model, Adv. in Appl. Probab., 2014, 46(1), 76–101.
[15] Resing J., Polling systems and multitype branching processes, Queueing Syst., 1993, 13(4), 409–426.

[16] Levy H., Analysis of cyclic polling systems with binomial-gated service, In: Hasegawa T., Takagi H., Takahashi Y. (Eds.), Performance of Distributed and Parallel Systems, North-Holland, Amsterdam, 1989, 127-139.
[17] Athreya K.B., Ney P.E., Branching Processes, Springer: Berlin Heidelberg, 1972.
[18] Van der Mei R.D., Polling systems with switch-over times under heavy load: moments of the delay, Queueing Syst., 2000, 36(4), 381–404.