Applied Probability Trust

PERFECT SAMPLING FOR INFINITE SERVER AND LOSS SYSTEMS

JOSE BLANCHET [*] [**] AND

JING DONG,[*] [***] *Columbia University*

Abstract

We present the first class of perfect sampling (also known as exact simulation) algorithms for the steady-state distribution of non-Markovian loss systems. We use a variation of Dominated Coupling From The Past. We first simulate a stationary infinite server system backwards in time and analyze the running time in heavy traffic. In particular, we are able to simulate stationary renewal marked point processes in unbounded regions. We then use the infinite server system as an upper bound process to simulate the loss system. The running time analysis of our perfect sampling algorithm for loss systems is performed in the Quality-Driven (QD) and the Quality-and-Efficiency-Driven regimes. In both cases, we show that our algorithm achieves sub-exponential complexity as both the number of servers and the arrival rate increase. Moreover, in the QD regime, our algorithm achieves a nearly optimal rate of complexity.

Keywords: perfect sampling, dominated coupling from the past, infinite server queues, loss queues, renewal point processes, many-server asymptotics

2010 Mathematics Subject Classification: Primary 65C05, 68U20
Secondary 60K25

[*] Postal address: Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA
[**] Email address: jose.blanchet@columbia.edu
[***] Email address: jd2736@columbia.edu

1. Introduction

We present the first class of exact simulation algorithms for the steady-state distribution of non-Markovian loss systems. The running time of our algorithms is analyzed in the context of many server queues in heavy-traffic; corresponding both to the so-called Quality-Driven (QD) regime, and the Quality-and-Efficiency-Driven (QED, also known as Halfin-Whitt) regime. In both cases, we show that our algorithm achieves sub-exponential complexity as the number of servers and the arrival rate increase. Moreover, in the QD regime, our algorithm achieves a nearly optimal rate of convergence. Our contributions are the first to provide exact simulation methodology with satisfactory running time analysis in heavy-traffic.

Exact simulation (or perfect sampling) consists in sampling without any bias from the steady-state distribution of a given ergodic process. Coupling From The Past (CFTP), introduced in the ground breaking paper, [22], is the most common exact simulation protocol. In its canonical form it applies only to uniformly ergodic Markov chains [15]. A variation of CFTP, called Dominated CFTP (DCFTP)[17], allows one to apply CFTP-type ideas to obtain unbiased samples from the steady-state distribution of ergodic processes without requiring uniform ergodicity. A standard application of DCFTP involves constructing two stationary processes which serve as the upper and lower bounds for the process of interest and can be simulated backwards in time from time zero. When the bounds coincide at some instant in the past, we say that coalescence occurs. The process of interest is then reconstructed forward in time from the coalescence position up to time zero, using the same input sequence that drives the upper and lower bounds. The state of the process of interest at time zero must then follow the corresponding steady-state distribution. More generally, a coalescence time is understood as an instant in the past from which reconstruction up to time zero guarantees a stationary sample. A coalescence time might be detected without the need to simulate upper and lower bound processes or have them coincide. In fact, this is the type of strategy that we follow in this paper.

Generic DCFTP algorithms have been studied for suitably ergodic Harris chains (see for example, [11], [18], and [10]). None of these algorithms apply directly to our setting as one requires information that is not available in the models we consider (see

for example p.788 in [10]). The papers [17], [19] and [14] are close in spirit to the main ideas of our paper as we take a point process approach to the problem. However, their approach requires the use of spatial birth and death processes (generally of Poisson type) as the dominating processes and as pointed out in Section 8 of [3], the complexity of the algorithms appear to increase significantly as the arrival rate increases.

In connection to queueing models, [20] applies the CFTP idea to simulate several queueing models assuming either exponential or bounded service times. In [9], the authors develop a class of DCFTP algorithms for Jackson networks (Poisson arrival and exponential service time). Sigman [23], [24] constructs exact sampling algorithms for multi-server queues assuming Poisson arrivals.

We provide a practical simulation procedure that works only under the assumption of renewal arrivals with finite mean and service time distribution with finite mean (in our running time analysis in heavy traffic we impose mild additional conditions for service times). We test our procedure numerically in Section 4.2.3, see also [5].

In order to implement our strategy for loss systems, we simulate a coupled stationary infinite server system backwards in time. We detect coalescence by observing a time interval on which all customers initially present in the infinite server system leave and no loss of customers occurs. This is an unconventional application of DCFTP in the sense that we use only the upper bound process to detect coalescence.

We summarize our contributions as follows:

1) The design and analysis of the first exact sampling algorithm for the infinite server systems whose running time is shown to be basically linear in the arrival rate and, thus, optimal, as the steady state of the infinite server systems, encoding the remaining service time of each customer, requires on average a vector which grows linearly in the arrival rate (see Theorem 1).
2) The design and analysis of the first exact sampling algorithm for many server loss systems under non-Markovian arrivals and in a heavy-traffic environment. In the QD regime, where service utilization is strictly less than 100%, we show that our algorithm has sub-linear coalescence time. In the QED regime, when the service utilization converges to 100% at a square root speed as a function of the arrival rate, we show that our algorithm has sub-exponential coalescence

time (see Theorems 2 & 3).

We point out that our algorithms allow to simulate stationary renewal processes with independent and identically distributed (i.i.d.) marks on unbounded but stable regions (having finitely many points almost surely). This connection has been noted in [5] for a fixed region (as apposed to a moving frame going backwards in time as we show here). Extensions of (ii) to loss networks with non-Markovian arrivals can also be easily obtained (see Chapter 3 of [12]).

In Section 2 we introduce our notations, describe the general strategy to simulate the infinite server system, and explain how to use it to detect coalescence for the loss system. In Section 3 we give algorithmic details for the strategy explained in Section 2. In Section 4 we study the running time of our algorithms.

2. Basic strategy and main results

In this section we introduce the basic strategy to simulate the systems. We also present some results about the efficiency of our algorithms. We leave the details of the algorithms and proofs of the results to subsequent sections.

To facilitate our explanation, we start with a formal description of the infinite server (GI/GI/∞) system.

2.1. Description of the GI/GI/∞ system

We first introduce some notation and assumptions. Let $N = \{N(t) : t \in (-\infty, 0]\}$ be a one sided time stationary renewal point process. We write $\{A_n : n \geq 1\}$ for the times at which the process N jumps counting backwards in time from time zero with $A_{n+1} < A_n < 0$. Furthermore, we define $X_n = |A_{n+1} - A_n|$. Now let $\{V_n : n \geq 1\}$ be a sequence of i.i.d. random variables (r.v.'s) which are independent of the process N. Define $Z_n = (A_n, V_n)$ and consider the marked point process $\mathcal{M} = \{Z_n : n \geq 1\} \in \mathbb{R}^2$ which we call the "arriving customer stream". More specifically, we consider customers arriving to the system according to a renewal process with i.i.d. interarrival times X_n's. Independent of the arrival process, their service requirements V_n's are also i.i.d..

Figure 1 elaborates on the point process description of the infinite server system and is important for describing our simulation strategy. In Figure 1, the point $Z_n =$

(A_n, V_n) denotes the n-th customer (counting backwards in time), whose arrival time is A_n and service requirement is V_n, $n = 1, \ldots, 4$. One important feature of the infinite server system is that every customer starts service immediately upon arrival (there is no queue). If we project Z_n to the horizontal axis by drawing a -45^o line, then the intersection of this diagonal line with the horizontal axis is the departure time of such n-th customer. We follow the technical tradition that an arrival at time t is counted in the system (closed circle) while a departure at time t is not counted (open circle). We can also draw a vertical line at any $t \in \mathbb{R}$. The height of the intersection of the -45^o lines emanating from the points Z_n with $A_n \leq t$ and such vertical line, if positive, represents the corresponding remaining service time of that customer at time t.

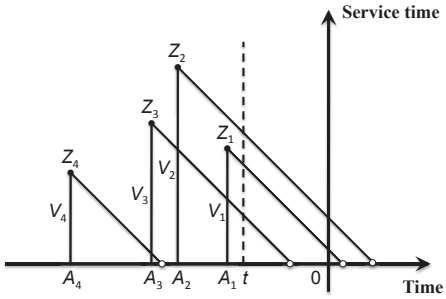

FIGURE 1: Point process description of an infinite server system

We write $G(\cdot) = P(X_n \leq \cdot)$ for the cumulative distribution function (CDF) of X_n, where P is the nominal probability measure, and write $\bar{G}(\cdot) = 1 - G(\cdot)$ for its tail CDF. Similarly, we write $F(\cdot) = P(V_n \leq \cdot)$ as the CDF of V_n and $\bar{F}(\cdot) = 1 - F(\cdot)$ as its tail CDF.

The following assumption is imposed throughout our discussion:

Assumption 1. *$EX_n < \infty$ and $EV_n < \infty$.where E is the expectation.*

We next introduce a Markovian description of the infinite server system. Let $Q(t, y)$ denote the number of people in the infinite server system at time t with residual service time strictly greater than y. Notice that for fixed t, $Q(t, \cdot)$ is a piecewise constant step function. If we denote $(r_{(1)}(t), \ldots, r_{(m)}(t))$ as the ordered (positive) remaining service times of customers in the system at time t, then $Q(t, 0) = m$ and $Q(t, y) =$

$\sum_{i=1}^{m} I(r_{(i)}(t) > y)$. We also let $E(t)$ denote the time elapsed since the previous arrival at time t (i.e. $E(t) = t - \max\{A_n : A_n \leq t\}$) and $W(t) = (E(t), Q(t,\cdot)) \in \mathbb{R}^+ \times \mathcal{D}[0,\infty)$. Then $\{W(t) : t \in R\}$ forms a Markov process which describes the infinite server system.

Similarly, we denote $W^L(t) = (E^L(t), Q^L(t,\cdot)) \in \mathbb{R}^+ \times \mathcal{D}[0,\infty)$ as the state of the loss system with C servers at time t, where $E^L(t) = t - \max\{A_n : A_n \leq t\}$ denotes the time elapsed since the previous arrival, and $Q^L(t,y)$ counts the number of people in the loss system at time t with residual service time strictly greater than y. Only costumers who see less than C servers busy at arrival are admitted to the system and all admitted customers start service immediately upon arrival. If we let $(r^L_{(1)}(t), \ldots, r^L_{(m^L)}(t))$ denote the ordered (positive) remaining service times of customers in the system at time t, then $Q^L(t,0) = m^L$ and $Q^L(t,y) = \sum_{i=1}^{m^L} I(r_{(i)}(t) > y)$.

We now provide a coupling between $W(\cdot)$ and $W^L(\cdot)$ such that $E^L(t) = E(t)$ and $Q^L(t,y) \leq Q(t,y)$ for all $y \geq 0$. In this sense, we say that $W^L(t) \leq W(t)$. The coupling proceeds as follows: we use same stream of customers, $\mathcal{M}$ (same arrival times and service requirements), to update both systems. One can label the servers in the infinite server system, assign customers to the empty server with the smallest label, and by tracking only the state of the first C servers in the infinite server system one automatically tracks the state of the loss system. Based on this coupling, we have that if $W^L(s) = W(s)$, then $W^L(t) \leq W(t)$ for $t \geq s$.

Definition 1. A coalescence time is a time $T < 0$ at which *the state of the loss system is identified* from the coupled infinite server system, i.e. $W^L(T) = W(T)$.

2.2. Coalescence time with an $GI/GI/C/C$ system

As discussed earlier, the infinite server system imposes an upper bound on the coupled loss system. A natural way to construct the coalescence time would be to define it as the first time (going backwards in time) the infinite server system empties (assuming, say, unbounded interarrival time distribution, this will occur). However, this coalescence time generally grows exponentially with the arrival rate [16]. So, instead we consider the following construction. Let $R(t)$ denote the maximum remaining service time among all customers in the system at time t. And consider a random time $\tau < 0$ satisfying

1) $R(\tau) < |\tau| < \infty$;

2) $\inf_{\tau \le t \le \tau + R(\tau)} \{C - Q(t, 0)\} \ge 0$, where C is the number of servers in the loss system.

As we will show in Section 4.2, τ can be identified, and our coalescence time is $T := \tau + R(\tau)$. In simple words, everyone who was present at time τ in the infinite server system will have left at time $\tau + R(\tau)$. And since the infinite server system has less than C customers on $[\tau, \tau + R(\tau)]$, the loss system is also operating below capacity C on that interval. Thus the infinite server system and the loss system must have the same set of customers present in the system at $\tau + R(\tau)$. From then on we can recover the state of the loss system at time zero using the same stream of customers as for the infinite server system on $[\tau + R(\tau), 0]$.

2.3. Basic strategy and main results for the GI/GI/∞ system

Simulating the infinite server system in stationarity and backwards in time is not trivial, so we first need to explain how to do this task. There are two cases to be considered.

Case 1 The interarrival time has finite exponential moments in a neighborhood of the origin. More specifically, define $\psi(\theta) = \log E \exp(\theta X_n)$. There exists $\theta > 0$ such that $\psi(\theta) < \infty$.

Case 2 The interarrival time does not have a finite exponential moment, i.e. $\psi(\theta) = \infty$ for any $\theta > 0$.

As we shall explain, we can always reduce the second case to the first one by defining yet another coupled upper bound process via truncation. Specifically, denote $X_n \wedge b = \min\{X_n, b\}$. We then fix a suitably large constant b and define a coupled infinite server system with truncated interarrival times: $\{X_n \wedge b : n \ge 1\}$. The truncation essentially speeds up the arrival process, thus creating more congestion. By coupling we mean that we use the same stream of customers to update both systems, i.e., we use (X_n, V_n) to update the original system and $(X_n \wedge b, V_n)$ to update the truncated one. We also define the event times as the arrival times and the departure times of the customers. Then the infinite server system with truncated interarrival times imposes an upper bound, in terms of the number of customers in the system, on the original

infinite server system at the corresponding event times. Precisely, the event times are defined as $A_n = \sum_{i=1}^n X_i$ and $A_n + V_n$, $n \geq 1$, for the infinite server system, and $A_n(b) := \sum_{i=1}^n (X_i \wedge b)$ and $A_n(b) + V_n$ for the truncated infinite server system.

In what follows, we shall first concentrate our discussion on Case 1 which also includes the infinite server system with truncated interarrival times. We then explain how to carry out the simulation for Case 2.

2.3.1. *Simulating the stationary $GI/GI/\infty$ system at time zero.* We first introduce the procedure to simulate the state of the stationary infinite server system at time zero. We notice from Figure 2 that customer $Z_n = \{A_n, V_n\}$, with $V_n \leq |A_n|$ (outside the gray triangle region) will have left the system by time 0. Thus if we can find a random number κ such that $V_n \leq |A_n|$ for all $n \geq \kappa$, then we can simulate the customer stream backwards in time up to κ (i.e. $\{Z_n : 1 \leq n \leq \kappa\}$) to recover the state of the system at time zero. The challenge here is that κ defined above depends on an infinite amount of information, i.e. $\{Z_n : n > \kappa\}$, and simulating this information takes infinite amount of time. We overcome this difficulty by defining a sequence of "record breakers". Then, instead of simulating the whole sequence of $\{Z_n : n \geq 1\}$, we only ask a yes/no question defined as "are there any more record breakers". In simulation, answering this yes/no question is equivalent to sampling a Bernoulli random variable with probability of success p, which equals to the probability that there are no more record breakers. If the Bernoulli trial is a failure, we find the next record breaker, say at index n_0, and ask again whether there will be any other record breakers at indices larger than n_0. We repeat the above process until we obtain a successful Bernoulli trial. Then, we know that there are no more record breakers in the remaining infinite sequence. We also locate the position (index) of all the record breakers. In what follows, we shall explain how to use this "record breaker" idea to simulate κ.

We start by separating the simulation of the arrival and service time processes. We write $\mu = EX_n$ and fix an $\epsilon \in (0, \mu)$. Consider any random number κ finite with probability one but large enough such that

$$A_{n+1} \geq n(\mu - \epsilon) \text{ and } V_{n+1} \leq n(\mu - \epsilon) \text{ for all } n \geq \kappa.$$

Let $\kappa(A)$ be a random time satisfying that $A_{n+1} \geq n(\mu - \epsilon)$ for $n \geq \kappa(A)$, and $\kappa(V)$ be a random time satisfying that $V_{n+1} \leq n(\mu - \epsilon)$ for $n \geq \kappa(V)$. Then we can set

$\kappa = \max\{\kappa(A), \kappa(V)\}$. The following proposition states that $\kappa < \infty$ almost surely (a.s.). The proof is given in Appendix A.

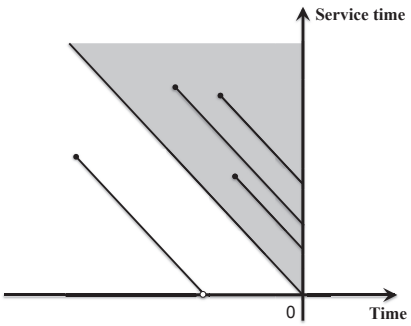

FIGURE 2: Coupling time of the infinite server system

Proposition 1. *Under Assumption 1, the random number κ defined above is finite with probability one.*

As $\{A_n : n \geq 1\}$ and $\{V_n : n \geq 1\}$ are independent of each other, the above construction allows us to sample $\{V_n : n \geq 1\}$ with $\kappa(V)$, and $\{A_n : n \geq 1\}$ with $\kappa(A)$ separately. We next explain the basic sampling strategies for the two processes.

For $\{V_n\}$, we say a record is broken at n, for $n \geq 1$ if $V_{n+1} > n(\mu - \epsilon)$. We then define $J(0) := 0$ and $J(l) = \inf\{n > J(l-1) : V_{n+1} > n(\mu - \epsilon)\}$ for $l = 1, 2, \cdots$. We also write $\gamma := \inf\{l \geq 1 : J(l) = \infty\}$. The $J(l)$'s, for $1 \leq l \leq \gamma - 1$, record the position of the record breakers. We can set $\kappa(V) = J(\gamma - 1) + 1$. We first simulate $J(l)$'s for $l = 1, 2, ..., \gamma - 1$, and then simulate the V_n's conditional on $J(l)$'s (see Section 3.1 for details).

For $\{A_n\}$, we first translate the process into a negative-drifted random walk. Specifically, we define $\tilde{S}_n := n(\mu - \epsilon) - (A_{n+1} - A_1) = \sum_{i=1}^{n} Y_i$, where $Y_i = (\mu - \epsilon) - X_{i+1}$. Note that Y_i's are i.i.d. with $EY_i = -\epsilon$. $A_{n+1} = A_1 - \tilde{S}_n + n(\mu - \epsilon)$. If we can simulate some random time κ^* such that $\tilde{S}_n \leq 0$ for $n \geq \kappa^*$, then $|A_{n+1} - A_1| \leq n(\mu - \epsilon)$ for $n \geq \kappa^*$. Fix any $m > 0$. We ask the yes/no question whenever $\tilde{S}_n < -m$ and we say that a record is broken at index n beyond $k \geq 0$, if for $n > k$, $S_n - S_k > m$. In particular, we define $\Gamma(0) := 0$ and $\Delta(l) := \inf\{n \geq \Gamma(l-1) : \tilde{S}_n \leq -m\}$,

$\Gamma(l) := \inf\{n \geq \Delta(l) : \tilde{S}_n - \tilde{S}_{\Delta(l)} \geq m\}$. Let $\alpha := \inf\{l \geq 1 : \Gamma(l) = \infty\}$. We notice that $\tilde{S}_n$ will never go above 0 from $\Delta(\alpha)$ on; which implies that we can set $\kappa(A) = \Delta(\alpha)$. As we assume the moment generating function of X_n is finite in a neighborhood of the origin, the moment generating function of Y_n is also finite around zero. We simulate $\tilde{S}_n$'s jointly with $\Delta(l)$'s and $\Gamma(l)$'s until α using exponential tilting and the acceptance rejection method (see Section 3.2 for details).

For the heavy-tailed case (Case 2), we can choose the truncation parameter b such that $E[X_n \wedge b] = \int_0^b \bar{G}(x)dx = \mu - 1/2\epsilon$. This is doable because we assume $EX_n = \int_0^\infty \bar{G}(x)dx < \infty$. Set $\epsilon' = 1/2\epsilon$. Then $E[X_n \wedge b] - \epsilon' = \mu - \epsilon$. Let $\kappa(A(b))$ be a random time satisfying that $|A_{n+1}(b)| \geq n(E[X_n \wedge b] - \epsilon')$ for $n \geq \kappa(A(b))$. Then we have $|A_{n+1}| \geq |A_{n+1}(b)| \geq n(\mu - \epsilon)$ for $n \geq \kappa(A(b))$. Thus, we can set $\kappa(A) = \kappa(A(b))$.

While our algorithm works under Assumption 1 only, we impose additional mild conditions on the service time distribution to rigorously show good algorithmic performance, especially in heavy traffic (i.e. as the arrival rate increases).

We consider a sequence of systems indexed by $s \in \mathbb{N}^+$. We shall say that *s is the scale of the system.* We speed up the arrival rate of the s-th system by scale s. That is, the interarrival times of the s-th system are given by $X_n^{(s)} = X_n/s$. We keep the service time distribution fixed for all systems, i.e. the service times do not scale with s. The following theorem summarizes the performance of the procedure we proposed for simulating a stationary infinite server system.

Theorem 1. *Assume $E[X_n] < \infty$, and*

(1) if $EV_n^q < \infty$ for some $q > 2$, then $E_\pi^s \kappa = O(s^{q/(q-1)})$;

(2) if we further assume $E[\exp(\theta V_n)] < \infty$ for some $\theta > 0$, then $E_\pi^s \kappa = O(s \log s)$.

We prove Theorem 1 by establishing two bounds for $\kappa(A)$ and $\kappa(V)$, respectively. The details are given in Section 4.1.

2.3.2. *Simulating the stationary $GI/GI/\infty$ system backwards in time.* We next extend the procedure to simulate states of the stationary infinite server system backwards in time for time intervals of any specified length. The construction is very similar to the single time point (i.e. time zero) case explained above.

Define $\kappa_0 := 1$. We consider a sequence of random times κ_j, $j = 1, 2, \cdots$, finite with

probability one but large enough such that

$$|A_n - A_{\kappa_{j-1}}| \geq (n - \kappa_{j-1})(\mu - \epsilon) \text{ and } V_n \leq (n - \kappa_{j-1})(\mu - \epsilon) \text{ for all } n \geq \kappa_j. \quad (1)$$

Notice that $V_n \leq |A_n - A_{\kappa_{j-1}}|$ for $n \geq \kappa_j$. This implies that a customer who arrives before A_{κ_j} will not be in the system at time $A_{\kappa_{j-1}}$. Thus, using $\{Z_n : 1 \leq n \leq \kappa_j\}$, we can recover the system descriptor $W(t)$ for $t \in [A_{\kappa_{j-1}}, 0]$.

Figure 3 gives more details about the construction. Every point Z_n, with $n > \kappa_j$, will not land in the upper triangle defined by the vertical line at $A_{\kappa_{j-1}}$ and the -45^o line intersecting it at the time axis (x axis).

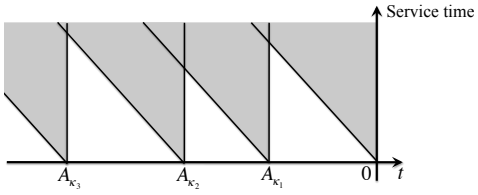

FIGURE 3: Coupling times of the infinite server system

The κ_j's give us some flexibility to separate the simulation of the two processes. We first simulate the service times and then conditional on the sample path of the service time process we simulate the arrival process jointly with κ_j's.

Define $J_1(0) := 1$ and for $k = 1, 2, \cdots$ and $l = 1, 2, \cdots, \gamma_k$ let

$$J_k(l) := \inf\{n > J_k(l-1) : V_n > (n - J_k(0))(\mu - \epsilon)\},$$
$$\gamma_k := \inf\{l \geq 0 : J_k(l) = \infty\} \text{ , and } J_{k+1}(0) := J_k(\gamma_k - 1)$$

for $k = 1, 2, \cdots$ and $l = 1, 2, \cdots, \gamma_k$. We first simulate the random time: $J_k(l)$'s for $k = 1, 2, \cdots$ and $l = 1, 2, \cdots, \gamma_k$, and then simulate $\{V_n : n \geq 1\}$ conditional on $J_k(l)$'s. See Algorithm I in Section 3.1 for details.

Given the sample path of $\{V_n : n \geq 1\}$ and $J_k(l)$'s, we next simulate $\{A_n : n \geq 1\}$ and κ_j's. This is done by simulating the negative-drift random walk $\tilde{S}_n$ jointly with

its running time maximum. Define $\Delta_1(0) := 0$ and $\Gamma_1(0) := 0$. Fix $m > 0$ and let

$$\Delta_j(l) := \inf\{n \geq \Gamma_j(l-1) : \tilde{S}_n - \tilde{S}_{\Delta_j(0)} \leq -m\} \ ,$$
$$\Gamma_j(l) := \inf\{n \geq \Delta_j(l) : \tilde{S}_n - \tilde{S}_{\Delta_j(l)} \geq m\},$$
$$\alpha_j := \inf\{l \geq 1 : \Gamma_j(l) = \infty\} \ , \ \kappa_j := \min\{J_k(0) : J_k(0) \geq \Delta_j(\alpha_j) + 1\},$$
$$\Delta_{j+1}(0) := \kappa_j - 1 \ , \ \Gamma_{j+1}(0) := \Delta_{j+1}(0)$$

for $j = 1, 2, ...$ and $l = 1, 2, \cdots, \alpha_j$. Notice that the process $\tilde{S}_n$ will never go above $\tilde{S}_{\Delta_j(0)}$ from $\Delta_j(\alpha_j)$ on. This implies that $|A_n - A_{\kappa_{j-1}}| \geq (n - \kappa_{j-1})(\mu - \epsilon)$ for $n \geq \kappa_j$. Under the light-tail assumption (Case 1), we simulate the random times $\Delta_j(l)$ and $\Gamma_j(l)$ for $j = 1, 2, ...$, $l = 1, 2, ..., \alpha_j$ and $\{\tilde{S}_n : n \geq 0\}$ by the exponential tilting and acceptance-rejection method. The details are explained in Algorithm II in Section 3.2.

For the heavy-tailed case (Case 2), we again simulate the infinite server system with truncated interarrival times first. We carefully choose the truncation parameter b such that $E[X_n \wedge b] - \epsilon'$, where $\epsilon' = \epsilon/2$, coincides with $\mu - \epsilon$. Then the $\kappa_j(b)$'s we constructed for the truncated system must automatically satisfy the conditions characterizing κ_j's in (1) for the original system as well.

2.4. Basic strategy and main results for the $GI/GI/C/C$ system

Once we simulate the customer stream backwards in time and construct the dominating stationary infinite server system accordingly, we can check and find the coalescence time $T = \tau + R(\tau)$ where τ is defined in Section 2.2 backwards in time. We then use the state of the infinite server system at time $T < 0$ as the state of the many-server loss system at the same time, and go forwards in time using the same stream of customers to construct the state of the loss system up to time 0.

As in the infinite server system case, we again consider a sequence of systems indexed by $s \in \mathbb{N}^+$ where the arrival rate of the s-th system is scaled by s and the service rate is kept fixed. Let $\rho = E[V_n]/E[X_n]$ (the ratio of the mean service time and mean interarrival time of the base system). We analyze the system in two heavy-traffic asymptotic regimes. One is the quality driven (QD) regime where $\rho < 1$ and the number of servers in the s-th system, C_s, satisfies $C_s = s$. The other regime is the quality and efficiency driven (QED) regime where $\rho = 1$ and the number of servers in the s-th system, C_s, satisfies $s + \beta\sqrt{s}$ for some $\beta > 0$.

Theorem 2 summarizes the performance of the coalescence time in the QD regime.

Theorem 2. *Assume $EX_n < \infty$ and X_n's are non-lattice and strictly positive. We also assume that $EV_n^q < \infty$ for any $q > 0$ and the cumulative distribution function (CDF) of V_n is continuous. Then $E_\pi^s \tau = o(s^\delta)$, for any $\delta > 0$.*

Remark 1. The assumption about the existence of all moments on the service time distribution covers a range of heavy tailed distributions, such as Weibull and log-normal, which are known to fit well data in applications [8].

Theorem 3 analyzes the performance of the coalescence time in the QED regime.

Theorem 3. *Assume $EX_n^2 < \infty$. We also assume $EV_n^q < \infty$ for any $q > 0$ and the CDF of V_n is continuous. Then for β large enough, we have $\log E_\pi^s \tau = o(s^\delta)$ for any $\delta > 0$.*

The main difficulty in the proof of Theorem 2 and Theorem 3 is that it involves tracking the state of the system on a time interval rather than a single time point. In Section 4.2, we prove the results using a geometric trial construction. To control the variation of the path on a time interval, for Theorem 2, we use the sample path large deviation results [4] for infinite server queues; for Theorem 3, we apply Borell-TIS inequality [1] to the diffusion limit processes of infinite server queues [21].

3. Detailed simulation algorithms

In order to provide the details of our simulation algorithms outlined in Section 2.3, we shall first work under the light-tailed case (Case 1) where we assume there exists $\theta > 0$ such that $\psi(\theta) < \infty$. The extension to the heavy-tailed case (Case 2) was introduced in Section 2 and we shall provide more details in Section 3.3.

We further impose the following assumptions on our ability to simulate the service times and interarrival times.

Assumption 2. *We assume that for each $x \geq 0$, $F(x)$ is easily computable, either in closed form or via efficient numerical procedures. Moreover, we can simulate V_n conditional on $V_n \in (a, b]$ with $P(V_n \in (a, b]) > 0$. The sampling time of V_n conditional on $V_n \in (a, b]$ is assumed to be independent of a and b.*

Assumption 3. *Suppose that $G(\cdot)$ is known and that it is possible to simulate from $G_{eq}(\cdot) := \mu^{-1}\int_{\cdot}^{\infty}\overline{G}(t)\,dt$. Moreover, let $G_\theta(\cdot) = E\exp(\theta X_n - \psi(\theta))I(X_n \le \cdot)$ be the associated exponentially tilted distribution with parameter θ for $\psi(\theta) < \infty$. We assume that we can simulate from $G_\theta(\cdot)$.*

Remark 2. Assumption 2 can be applied to virtually any model used in practice, including distributions such as Gamma, phase-type, Pareto, Weibull, Lognormal, and mixtures of them. Knowledge of the underlying distribution is required in Procedure A below. Note that the required simulation procedure is not restricted to the inversion method. One can use, for example, the acceptance/rejection method, but a good proposal distribution for the conditional distribution given $V_n \in (a,b]$ might have to be constructed based on knowledge of the density function to increase efficiency. Assumption 3 is applicable to models for which the moment generating function is finite, these include distributions such as Gamma, phase type, hyperexponential, and other mixtures of them.

We next introduce our algorithm to simulate $\{V_n : n \ge 1\}$. Conditional on the sample path of $\{V_n : n \ge 1\}$, we then explain how to to simulate $\{A_n : n \ge 1\}$ together with κ_j's.

3.1. Simulation of $\{V_n : n \ge 1\}$ and $J_k(l)$'s for $k = 1, 2, \cdots$, $l = 1, 2, \cdots, \gamma_k$

We will first introduce the procedure to simulate $J_1(l)$ for $l = 1, 2, \cdots, \gamma_1$. Recall that $J_1(l)$'s keep track of all the record breakers, $\{n : V_n > n(\mu - \epsilon)\}$. Let $p(n) = P(V_1 > n(\mu-\epsilon))$. Then $P(J_1(l) = \infty | J_1(l-1) = k) = \prod_{n=k+1}^{\infty}(1-p(n))$, which is the probability that after k there are no more record breakers (i.e. a success of the Bernoulli trial occurs), and it is the product of infinitely many terms. We do not know how to evaluate the infinite product exactly. However, we can find a sequence of upper bounds and lower bounds of $P(J_1(l) = \infty | J_1(l-1) = k)$ denoted by $g(h)$'s and $f(h)$'s for $h > k$ respectively, such that

$$f(h) < f(h+1) < \cdots < P(J_1(l) = \infty | J_1(l-1) = k) < \cdots < g(h+1) < g(h),$$

and $\lim_{h\to\infty} f(h) = \lim_{h\to\infty} g(h) = P(J_1(l) = \infty | J_1(l-1) = k)$. For $U \sim$Uniform$[0,1]$, we can then determine whether the Bernoulli trial is a success ($U \le P(J_1(l) = \infty | J_1(l-$

$1) = k)$) or a failure ($U > P(J_1(l) = \infty | J_1(l-1) = k)$) by checking if $U < f(h)$ or $U > g(h)$ for some $h > k$. As

$$\prod_{n=k+1}^{h}(1-p(n)) \geq P(J_1(l) = \infty | J_1(l-1) = k) \geq \prod_{n=k+1}^{h}(1-p(n)) \times \exp(-\frac{2\int_h^\infty P(V_1 > \nu)d\nu}{\mu - \epsilon}), \tag{2}$$

the upper bound is easily obtained by truncating the infinite product up to finitely many terms, i.e. $g(h) = \prod_{n=k+1}^{h}(1-p(n))$. For the lower bound, let $u(h) := \exp(-2\int_h^\infty P(V_1 > \nu)/(\mu - \epsilon)d\nu)$, then we have $f(h) = g(h)u(h)$. Moreover, it is easy to check that $g(h) - g(h-1) = p(h)\prod_{i=k}^{h-1}(1-p(i)) = P(J_1(l) = h | J_1(l-1) = k)$. Thus, if $g(h+1) < U < g(h)$, we can also claim that $J_1(l) = h+1$. The "sandwiching" idea just described is the key in Procedure A introduced below.

Procedure A (Simulate $J_1(l)$ given $J_1(l-1) = k$)

1. Initialize $h = k+1$, $g = 1 - p(h)$ and $f = gu(h)$. Simulate $U \sim \text{Unif}[0,1]$.
2. While $f < U < g$, set $h = h+1$, $g = g(1-p(h))$ and $f = gu(h)$.
3. If $U \leq f$, then $J_1(l) = \infty$. Otherwise, $J_1(l) = h$.

The following lemma guarantees the finite termination of Procedure A.

Lemma 1. *If $EV_1 < \infty$, then*

$$P(J_1(1) = \infty) = \prod_{n=1}^{\infty}(1-p(n)) \geq \exp(-\frac{cEV_1}{\mu - \epsilon}) > 0 \tag{3}$$

for some $c > 0$ depending on the value of $p(1)$, thus, $E\gamma_1 \leq \exp(cEV_1/(\mu - \epsilon)) < \infty$.

Proof.

$$P(J_1(1) = \infty) = \prod_{n=1}^{\infty}(1-p(n)) \geq \prod_{n=1}^{\infty}\exp(-cp(n))$$
$$\geq \exp(-\frac{c}{\mu - \epsilon}\int_0^\infty P(V_1 > \nu)d\nu) = \exp(-\frac{cEV_1}{\mu - \epsilon}).$$

For $l = 2, 3, \cdots$, conditional on $J_1(l-1) = k$:

$$P(J_1(l) = \infty | J_1(l-1) = k) = \prod_{n=k+1}^{\infty}(1-p(n))$$
$$\geq \exp(-\frac{c\int_k^\infty P(V_1 > \nu)d\nu}{\mu - \epsilon}) \geq \exp(-\frac{cEV_1}{\mu - \epsilon}).$$

Thus γ_1 is stochastically dominated by a geometric random variable with parameter $p = \exp(-cEV_1/(\mu-\epsilon))$. The result then follows. □

The simulation of $J_k(l)$ for $k = 1, 2, \ldots$, $l = 1, 2, \ldots, \gamma_k$ is very similar to Procedure A. Let $p_k(n) = P(V_1 > n(\mu-\epsilon)|V_1 \leq (n + J_k(0) - J_{k-1}(0))(\mu-\epsilon))$. Then following the same argument leading to (3) and (2), we have $P(J_k(1) = \infty) > 0$, and for $h > n$,

$$\begin{aligned}
&\prod_{i=n+1}^{h} (1 - p_k(i)) \\
\geq &P(J_k(l) - J_k(0) = \infty | J_k(l-1) - J_k(0) = n) \\
\geq &\prod_{i=n+1}^{h} (1 - p_k(i)) \times \exp\left(- \frac{2\int_h^\infty P(V_1 > \nu | V_1 \leq \nu + (J_k(0) - J_{k-1}(0))(\mu-\epsilon)) d\nu}{\mu - \epsilon} \right).
\end{aligned}$$

Let $u_k(h) := \exp\left(-2\int_h^\infty P(V_1 > \nu|V_1 \leq \nu + (J_k(0) - J_{k-1}(0))(\mu-\epsilon))\, d\nu/(\mu-\epsilon)\right)$. We now propose a modification of Procedure A that allows us to simulate $J_k(l)$ conditional on $J_k(l-1) - J_k(0) = n$.

Procedure A1 (Simulate $J_k(l)$ given $J_k(l-1) - J_k(0) = n$)

1. Initialize $h = n+1$, $g = 1 - p_k(h)$ and $f = gu_k(h)$. Simulate $U \sim \text{Unif}[0,1]$.
2. While $f < U < g$, set $h = h+1$, $g = g(1-p_k(h))$ and $f = gu_k(h)$.
3. If $U \leq f$, then $J_k(l) = \infty$. Otherwise, $J_k(l) = J_k(0) + h$.

Based on Procedure A1 and our previous analysis we have:

Algorithm I (Sample V_n's jointly with $J_k(l)$'s)

Step 0. Set $J_0(0) = -\infty$, $J_1(0) = 1$, $k = 1$, $l = 1$. Simulate V_1 according to its nominal distribution.

Step 1. Simulate $J_k(l)$ conditional on the value of $J_k(l-1)$ using Procedure A1.

Step 2. If $J_k(l) = \infty$, set $\gamma_k = l$, $J_{k+1}(0) = J_k(\gamma_k - 1)$, $k = k+1$, $l = 1$ and go back to Step 1. Otherwise, go to Step 3.

Step 3. Simulate V_n for $J_k(l-1) < n < J_k(l)$ by conditioning on $V_n \leq (n - J_k(0))(\mu-\epsilon)$ and simulate $V_{J_k(l)}$ by conditioning on $(J_k(l) - J_k(0))(\mu-\epsilon) < V_{J_k(l)} \leq (J_k(l) - J_{k-1}(0))(\mu-\epsilon)$. Set $l = l+1$ and go back to Step 1.

When running the above algorithm, we specify K as the number of intervals ($[J_k(0), J_k(\gamma_k - 1)]$) we want to simulate. We then run Algorithm I from $k = 1$ until $k = K$. The program will give us $\{V_n : 1 \leq n \leq J_K(\gamma_K - 1)\}$ and $J_k(l)$'s for $k = 1, 2, \cdots, K$, and $l = 1, 2, \cdots, \gamma_k$.

3.2. Simulation of $\{A_n : n \geq 1\}$ and $\Delta_j(l)$'s, $\Gamma_j(l)$'s for $j = 1, 2, ...,$ $l = 1, 2, ..., \alpha_j$

Given the sample path of $\{V_n : n \geq 1\}$, we will first explain how to simulate the $\Delta_j(l)$'s and $\Gamma_j(l)$'s sequentially and jointly with the underlying random walk $\{\tilde{S}_n : n \geq 1\}$. We then simulate A_1 according to $G_{eq}(\cdot)$ and set $A_{n+1} = A_1 + n(\epsilon - \mu) - \tilde{S}_n$. The methodology in this subsection follows closely those in [13] and [7]. The same procedure can be used to simulate a negative drifted random walk, $\tilde{S}_n$, together with its running time maximum defined as $\max_{k \geq n}\{\tilde{S}_k\}$.

Let $\mathcal{F}_n = \sigma\{Y_1, Y_2, \cdots, Y_n\}$, denote the σ-field generated by the Y_j's up to time n. Define $T_\xi := \inf\{n \geq 0 : \tilde{S}_n > \xi\}$ for $\xi \geq 0$. Then by the strong Markov property, we have that for $1 \leq l \leq \alpha_j$, $P(\Gamma_j(l) = \infty | \mathcal{F}_{\Delta_j(l)}) = P(\Gamma_j(l) = \infty | \tilde{S}_{\Delta_j(l)}) = P(T_m = \infty) > 0$. It is important then to notice that $P(\alpha_j = k) = P(T_m < \infty)^{k-1} P(T_m = \infty)$, for $k \geq 1$. In other words, α_j is geometrically distributed. The procedure that we have in mind is to simulate each stage $\Delta_j(\alpha_j)$ in time intervals, and the number of time intervals is precisely α_j.

Let $\psi_Y(\theta) = \log E \exp(\theta Y_i)$ be the log moment generating function of Y_i. As we assume $\psi_X(\theta)$ is finite in a neighborhood of the origin, $\psi_Y(\cdot)$ is also finite in a neighborhood of the origin. Moreover $EY_i = \psi_Y'(0) = -\epsilon$ and $\mathrm{Var}(Y_i) = \psi_Y''(0) > 0$. Then by the convexity of $\psi_Y(\cdot)$, one can always select $\epsilon > 0$ sufficiently small so that there exists $\eta > 0$ with $\psi_Y(\eta) = 0$ and $\psi_Y'(\eta) \in (0, \infty)$. The root η allows us to define a new measure P_η based on exponential tilting so that

$$\frac{dP_\eta}{dP}(Y_i) = \exp(\eta Y_i).$$

Moreover, under P_η, $\tilde{S}_n$ is random walk with positive drift equal to $\psi_Y'(\eta)$ [2]. Therefore, $P_\eta(T_\xi < \infty) = 1$ and $q(\xi) := P(T_\xi < \infty) = E_\eta \exp(-\eta \tilde{S}_{T_\xi})$ for each $\xi \geq 0$.

Based on the above analysis we now introduce a convenient representation to sim-

ulate a Bernoulli random variable $J(\xi)$ with probability of success $q(\xi)$, namely,

$$J(\xi) = I(U \leq \exp(-\eta \tilde{S}_{T_\xi})), \tag{4}$$

where U is a uniform random variable independent of everything else under P_η. Sampling $\{\tilde{S}_1, \cdots, \tilde{S}_{T_\xi}\}$ conditional on $T_\xi < \infty$, as we shall explain now, corresponds to basically the same procedure. First, let us write $P^*(\cdot) := P(\cdot | T_\xi < \infty)$. The following result provides an expression for the likelihood ratio between P^* and P_η.

Lemma 2. *We have that*

$$\frac{dP^*}{dP_\eta}(\tilde{S}_1, ..., \tilde{S}_{T_\xi}) = \frac{\exp(-\eta \tilde{S}_{T_\xi})}{P(T_\xi < \infty)} \leq \frac{1}{P(T_\xi < \infty)}.$$

Proof.

$$\begin{aligned} P(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_\xi} \in H_{T_\xi} | T_\xi < \infty) &= \frac{P(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_\xi} \in H_{T_\xi}, T_\xi < \infty)}{P(T_\xi < \infty)} \\ &= \frac{E_\eta[\exp(-\eta \tilde{S}_{T_\xi}) I(\tilde{S}_1 \in H_0, ..., \tilde{S}_{T_\xi} \in H_{T_\xi})]}{P(T_\xi < \infty)}. \ \square \end{aligned}$$

The previous lemma provides the basis for a simple acceptance / rejection procedure to simulate $\{\tilde{S}_1, ..., \tilde{S}_{T_\xi}\}$ conditional on $T_\xi < \infty$. More precisely, we generate a proposal $\{\tilde{S}_1, ..., \tilde{S}_{T_\xi}\}$ from $P_\eta(\cdot)$. Then one generates a uniform random variable U independent of everything else and accept the proposal if

$$U \leq P(T_\xi < \infty) \times \frac{dP^*}{dP_\eta}(\tilde{S}_1, ..., \tilde{S}_{T_\xi}) = \exp(-\eta \tilde{S}_{T_\xi}).$$

This criterion coincides with $J(\xi)$ according to (4). So, the procedure above simultaneously obtains both a Bernoulli random variable $J(\xi)$ with parameter $q(\xi)$, and the corresponding path $\{\tilde{S}_1, ..., \tilde{S}_{T_\xi}\}$ conditional on $T_\xi < \infty$ under $P(\cdot)$ if $J(\xi) = 1$.

As $E[Y_i] = -\epsilon < 0$, by strong law of large numbers we have $\Delta_j(l) < \infty$ almost surely for $j = 1, 2, ...$ and $l = 1, 2, ..., \alpha_j$. We next define $\bar{q}(\xi) := 1 - q(\xi) = P(T_\xi = \infty)$ and $P'(\cdot) := P(\cdot | T_\xi = \infty)$. The following result provides an expression for the likelihood ratio between P' and P.

Lemma 3. *We have that*

$$\frac{dP'}{dP}(\tilde{S}_1, ..., \tilde{S}_n) = \frac{I(T_\xi > l)\bar{q}(\xi - \tilde{S}_n)}{P(T_\xi = \infty)} \leq \frac{1}{P(T_\xi = \infty)}.$$

Proof.

$$\begin{aligned}&P(\tilde{S}_1 \in H_1, ..., \tilde{S}_n \in H_n | T_\xi = \infty)\\ =&\frac{P(\tilde{S}_1 \in H_1, ...\tilde{S}_n \in H_n, T_\xi = \infty)}{P(T_\xi = \infty)}\\ =&\frac{E[I(\tilde{S}_1 \in H_1, ..., \tilde{S}_n \in H_n) I(T_\xi > n) P(T_\xi = \infty | \tilde{S}_1, ..., \tilde{S}_n)]}{P(T_\xi = \infty)}.\end{aligned}$$

The result then follows from the strong Markov property and homogeneity of the random walk. □

We shall apply acceptance / rejection to sample from P'. According to Lemma 3, to sample $\{\tilde{S}_1, ..., \tilde{S}_n\}$ given $T_\xi = \infty$, we propose from the original (nominal) distribution and accept with probability $\bar{q}(\xi - \tilde{S}_n)$ as long as $\tilde{S}_j \leq \xi$ for all $0 \leq j \leq n$. In order to perform the acceptance/rejection step we need to sample a Bernoulli with parameter $\bar{q}(\xi - \tilde{S}_n)$, but this can be easily done using identity (4).

Consider $0 \leq \xi_1 < \xi_2$, we define $P^o(\cdot) := P(\cdot | T_{\xi_1} < \infty, T_{\xi_2} = \infty)$. The following result provides an expression for the likelihood ratio between P^o and P_η.

Lemma 4. *We have that*

$$\frac{dP^o}{dP_\eta}(\tilde{S}_1, ...\tilde{S}_{T_{\xi_1}}) = \frac{\exp(-\eta \tilde{S}_{T_{\xi_1}}) \bar{q}(\xi_2 - \tilde{S}_{T_{\xi_1}})}{P(T_{\xi_1} < \infty, T_{\xi_2} = \infty)} \leq \frac{1}{P(T_{\xi_1} < \infty, T_{\xi_2} = \infty)}.$$

Proof.

$$\begin{aligned}&P(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_{\xi_1}} \in H_{T_{\xi_1}} | T_{\xi_1} < \infty, T_{\xi_2} = \infty)\\ =&\frac{E_\eta[I(\tilde{S}_1 \in H_1, ..., \tilde{S}_{T_{\xi_1}} \in H_{T_{\xi_1}}) \exp(-\eta \tilde{S}_{T_{\xi_1}}) P(T_{\xi_2} = \infty | \tilde{S}_1, ..., \tilde{S}_{T_{\xi_1}})]}{P(T_{\xi_1} < \infty, T_{\xi_2} = \infty)}.\ \square\end{aligned}$$

We again use acceptance/rejection to sample $\{\tilde{S}_1, ..., \tilde{S}_{T_{\xi_1}}\}$ given $T_{\xi_1} < \infty$ and $T_{\xi_2} = \infty$. We propose $\{\tilde{S}_1, ..., \tilde{S}_{T_{\xi_1}}\}$ from $P_\eta(\cdot)$. Then we simulate a uniform random variable U independent of all else and accept the proposal if

$$U \leq \frac{P(T_{\xi_1} < \infty, T_{\xi_2} = \infty)}{\exp(-\eta \xi_1)} \times \frac{dP^o}{dP_\eta}(\tilde{S}_1, ..., \tilde{S}_{T_{\xi_1}}) = \exp(-\eta \tilde{S}_{T_{\xi_1}}) q(\xi_2 - \tilde{S}_{T_{\xi_1}}).$$

Based on the above analysis we propose the following algorithm.

Algorithm II (Given V_n's and $J_k(l)$'s, sample $\tilde{S}_n$'s together with $\Delta_j(l)$'s, $\Gamma_j(l)$'s and κ_j's)

Step 0. Set $\Delta_1(0) = \Gamma_1(0) = 0$, $\tilde{S}_0 = 0$, $j = 1$, $l = 1$, $\xi = \infty$, $\gamma = -m$. Sample A_1 according to $G_{eq}(\cdot)$.

Step 1. Simulate $S_1, ..., S_{T_\gamma}$ from the original (nominal) distribution.

Step 2. If $S_i \leq \xi$ for all $1 \leq i \leq T_\gamma$ then sample a Bernoulli $J(\xi - S_{T_\gamma})$ with parameter $q(\xi - S_{T_\gamma})$ using (4) and continue to step 3. Otherwise (i.e. $S_i > \xi$ for some $1 \leq i \leq T_\gamma$) go back to step 1.

Step 3. If $J(\xi - S_{T_\gamma}) = 1$, go back to step 1. Otherwise $J(\xi - S_{T_\gamma}) = 0$, let $\Delta_j(l) = \Gamma_j(l-1) + T_\gamma$ and $\tilde{S}_{\Gamma_j(l-1)+i} = \tilde{S}_{\Gamma_j(l-1)} + S_i$ for $i = 1, ..., T_\gamma$. If $j \geq 2$, set $\xi = \tilde{S}_{\Delta_{j-1}(\alpha_{j-1})} + m - \tilde{S}_{\Delta_j(l)}$.

Step 4. Simulate $S_1, ..., S_{T_m}$ from $P_\eta(\cdot)$. Sample a Bernoulli $J(\xi - S_{T_m})$ with parameter $q(\xi - S_{T_m})$ using (4) and $U \sim \text{Unif}[0,1]$. Let $J^* = I(U \leq \exp(-\eta S_{T_m})) \times (1 - J(\xi - S_{T_m}))$.

Step 5. If $J^* = 1$, let $\Gamma_j(l) = \Delta_j(l) + T_m$ and $\tilde{S}_{\Delta_j(l)+i} = \tilde{S}_{\Delta_j(l)} + S_i$ for $1 \leq i \leq T_m$. Set $\gamma = \min\{0, \tilde{S}_{\Delta_j(0)} - m - \tilde{S}_{\Gamma_j(l)}\}$. If $j \geq 2$, set $\xi = \tilde{S}_{\Delta_{j-1}(\alpha_{j-1})} + m - S_{\Gamma_j(l)}$. Set $l = l+1$ and go back to step 1. Otherwise $J^* = 0$, set $\alpha_j = l$, $\kappa_j = \inf\{J_k(0) : J_k(0) \geq \Delta_j(\alpha_j) + 1\}$, $\Delta_{j+1}(0) = \kappa_j - 1$, $\xi = m$ and continue to step 6.

Step 6. Let $h = \Delta_{j+1}(0) - \Delta_j(\alpha_j)$. Sample $S_1, ..., S_h$ from the original distribution.

Step 7. If $S_i \leq \xi$ for all $1 \leq i \leq h$ then sample a Bernoulli $J(\xi - S_h)$ with parameter $q(\xi - S_h)$ using (4) and continue to step 8. Otherwise (i.e. $S_i > \xi$ for some $1 \leq i \leq h$), go back to step 6.

Step 8. If $J(\xi - S_h) = 1$, go back to step 6. Otherwise $J(\xi - S_h) = 0$, let $\tilde{S}_{\Delta_j(\alpha_j)+i} = \tilde{S}_{\Delta_j(\alpha_j)} + S_i$ for $i = 1, ..., h$. Set $A_{n+1} = A_1 + n(\epsilon - \mu) - \tilde{S}_n$ for $n = \Delta_j(0) + 1, ..., \Delta_{j+1}(0)$. Set $j = j + 1$, $l = 1$, $\xi = \tilde{S}_{\Delta_{j-1}(\alpha_{j-1})} + m - \tilde{S}_{\Delta_j(0)}$, $\gamma = -m$ and go back to step 1.

When applying Algorithm II, we must specify K as the number of intervals ($[\kappa_{j-1}, \kappa_j]$) we want to simulate. We then run Algorithm II from $j = 1$ until $j = K$, and get $\{A_n : 1 \leq n \leq \kappa_K\}$ and $\{\kappa_j : 1 \leq j \leq K\}$.

3.3. Coupled infinite server system with truncated interarrival times

In this subsection, we provide some additional details for simulating the coupled truncated infinite server system with the original infinite server system.

We first explain how to simulate A_1 jointly with $A_1(b)$. The equilibrium distribution of X_n is $G_{eq}(x) = \int_0^x \bar{G}(u)du/EX_n$ and the equilibrium distribution of $X_n \wedge b$ is $G^b_{eq}(x) = \int_0^x \bar{G}(u)duI\{x \le b\}/E[X_n \wedge b]$. Thus we simulate A_1 with CDF $G_{eq}(x)$, if $A_1 \le b$, we set $A_1(b) = A_1$. Otherwise if $A_1 > b$, we keep simulating X_e with CDF $G_{eq}(x)$ until $X_e \le b$ and set $A_1(b) = X_e$. In particular we have $A_1(b) \le A_1$.

When simulating $X_n \wedge b$'s from the nominal distribution, we first simulate X_n with CDF $G(\cdot)$ and set $X_n \wedge b = \min\{X_n, b\}$. Denote $Y_n(b) = (E[X_n \wedge b] - \epsilon') - X_n \wedge b$ and let $\eta_b > 0$ be chosen such that $\log E\exp(\eta_b Y_n(b)) = 0$. When simulating $X_n \wedge b$'s under exponential tilting $P_{\eta_b}(\cdot)$, we first simulate $Y_n(b)$ under $P_{\eta_b}(\cdot)$ and set $X_n \wedge b = (E[X_n \wedge b] - \epsilon') - Y_n(b)$. If $X_n \wedge b < b$, set $X_n = X_n \wedge b$, otherwise $(X_n \wedge b = b)$, sample X_n conditional on $X_n \ge b$ under the nominal distribution $P(\cdot)$.

4. Performance analysis

In this section, we analyze the running time of our algorithms. We start with the infinite server system and then analyze the coalescence time of the many-server loss system.

4.1. Termination time for the infinite server system (Proof of Theorem 1)

Theorem 1 provides the relationship between the moment of the service times and $E^s_\pi\kappa$. We next give a proof of it. We shall omit the subscription π and s when there is no confusion for notational convenience. We first give a proof of the light tailed case. Recall that $\kappa = \max\{\kappa(V), \kappa(A)\}$, where $\kappa(V) = \inf\{k > 1 : V_{n+1} \le n(\mu-\epsilon)/s \text{ for all } n \ge k\}$ and $\kappa(A) = \inf\{k > 1 : A_{n+1} \ge n(\mu-\epsilon)/s \text{ for all } n \ge k\}$. We prove the theorem by establishing the bounds for $\kappa(V)$ (Lemma 5) and $\kappa(A)$ (Lemma 6) separately.

Lemma 5. *If $EV_n^q < \infty$ for some $q > 2$, then $E\kappa(V) = O(s^{q/(q-1)})$.*

Proof. Let $p(n) = P(V_1 > n(\mu-\epsilon)/s)$. For k sufficiently large, we have

$$P(\kappa(V) > k) = 1 - \prod_{n=k+1}^{\infty}(1-p(n)) \leq 1 - \exp(-\frac{2s}{\mu-\epsilon}\int_{k(\mu-\epsilon)/s}^{\infty} P(V>\nu)d\nu).$$

By Chebyshev's inequality $P(V_n > \nu) \leq EV_n^q/\nu^q$. Let $\delta = 1/(q-1)$, then for s sufficiently large, we have

$$\begin{aligned}\sum_{k=s^{1+\delta}}^{\infty} P(\kappa(V)>k) &\leq \sum_{k=s^{1+\delta}}^{\infty}\frac{2s}{\mu-\epsilon}\int_{k(\mu-\epsilon)/s}^{\infty} P(V>\nu)d\nu \\ &\leq \frac{2EV_n^q s^q}{(q-1)(q-2)(\mu-\epsilon)^q}\sum_{k=s^{1+\delta}}^{\infty}\frac{1}{k^{q-1}} = O(s^{q-(1+\delta)(\delta-2)}).\end{aligned}$$

As $q-(1+\delta)(q-2) = 1+\delta$,

$$\begin{aligned}E\kappa(V) &= \sum_{k=0}^{\infty} P(\kappa(V)>k) \\ &= \sum_{k=0}^{s^{1+\delta}-1} P(\kappa(V)>k) + \sum_{k=s^{1+\delta}}^{\infty} P(\kappa(V)>k) \leq s^{1+\delta} + O(s^{1+\delta}).\end{aligned}$$

Notice that when $E\exp(\theta V_n) < \infty$ for some $\theta > 0$, $P(V_n > \nu) \leq E\exp(\theta(V_n-\nu)) = E\exp(\theta V_n)\exp(-\theta\nu)$. Similarly as above, for s sufficiently large we have

$$\sum_{k=\lceil \frac{2}{\theta(\mu-\epsilon)} s\log s\rceil}^{\infty} P(\kappa(V)>k) \leq \frac{2E\exp(\theta V_n)}{(\mu-\epsilon)^2\theta^2}$$

and

$$E\kappa(V) = \sum_{k=0}^{s\log s-1} P(\kappa(V)>k) + \sum_{k=s\log s}^{\infty} P(\kappa(V)>k) \leq s\log s + O(1).$$

Thus if $E\exp(\theta V) < \infty$ for some $\theta > 0$, then $E\kappa(V) = O(s\log s)$. □

Lemma 6. *Assume there exist $\theta > 0$, such that $\psi(\theta) < \infty$, then $E\kappa(A) = O(s)$.*

Proof. Based on the algorithm proposed in Section 3.2, we divide the proof into two parts. We first prove that the expected number of iterations is $O(1)$. We then prove that the expected number of steps to pass $-m$ or m from 0 is $O(s)$.

Let $T_\xi = \inf\{n \geq 0 : \tilde{S}_n > \xi\}$. Recall that for the base system there exist $\eta > 0$ with $\psi_Y(\eta) = 0$ and $\psi_Y'(\eta) > 0$. And the number of iterations is distributed as a geometric random variable with probability of success $P(T_m = \infty) = 1 - E_\eta \exp(-\eta\tilde{S}_{T_m})$.

Then for the s-th system with $Y_i^s = Y_i/s$ we have $\tilde{S}_n/s > m$ is equivalent to $\tilde{S}_n > sm$. Thus the number of iterations is a Geometric random variable with probability of success $P(T_{sm} = \infty) = 1 - E_\eta \exp(-\eta \tilde{S}_{T_{sm}}) \geq 1 - \exp(-\eta sm)$.

Similarly, let $T'_\xi = \inf\{n \geq 0 : \tilde{S}_n < \xi\}$. Define $M_n = \tilde{S}_n + n\epsilon$, then M_n is a martingale with respect to the filtration generated by $\{Y_1, Y_2, ..., Y_n\}$. As $EY_i = -\epsilon < 0$, $P(T'_{-m} < \infty) = 1$. By the Optional Sampling Theorem, $EM_{T'_{-m}} = E\tilde{S}_{T'_{-m}} + \epsilon ET'_{-m} = 0$. Thus $ET'_{-m} = m/\epsilon - E[m - S_{T'_{-m}}]/\epsilon$. Then for the s-th system we have $ET'_{-sm} = sm/\epsilon - E[sm - S_{T'_{-sm}}]/\epsilon$. As $(sm - S_{T'_{-sm}})$ converges to the ladder hight distribution as $s \to \infty$[2] and $\sup_m E[(sm - S_{T'_{-sm}})^p] < \infty$ for $p > 1$, $ET'_{-sm} = O(s)$. □

For the heavy-tailed case, we select the truncation parameter b such that $E[X_n \wedge b] = \mu - 1/2\epsilon$. Then we set $\epsilon' = 1/2\epsilon$ and define $\kappa(A(b))$ as a random time satisfying that $|A_{n+1}| \geq n(E[X_n \wedge b] - \epsilon') = n(\mu - \epsilon)$ for $n \geq \kappa(A(b))$. As $|A_{n+1}| \geq |A_{n+1}(b)|$ under our coupling scheme, we can set $\kappa(A) = \kappa(A(b))$. By Lemma 6, we have $E\kappa(A) = E\kappa(A(b)) = O(s)$.

As $\kappa = \max\{\kappa(V), \kappa(A(b)\}$, we have $E\kappa = O(s \log s)$. This concludes the proof of Theorem 1.

4.2. Coalescence time for the many-server loss system (Proof of Theorem 2 and Theorem 3)

As we are simulating the process backwards in time, it is natural to define the filtration $\overleftarrow{\mathcal{H}}_t = \sigma\{W(-u) : 0 \leq u \leq t\}$, for which $\overleftarrow{\mathcal{H}}_u \subset \overleftarrow{\mathcal{H}}_t$ for $0 \leq u \leq t$. τ is a stopping time with respect to $\overleftarrow{\mathcal{H}}_t$. We next try to draw connections between the backward process and some forward process. Define $\tau^* := \inf\{t + R(t) : \sup_{t \leq u \leq t+R(t)}\{Q(u, 0)\} < s, t \geq 0\}$. τ^* is a stopping time with respect to $\mathcal{H}_t$ where $\mathcal{H}_t = \sigma\{M(u) : 0 \leq u \leq t\}$. The stochastic process $\{Q(t, 0) : t \in \mathbb{R}\}$ has a piecewise constant sample path with finitely many points of discontinuity on any finite length intervals almost surely. Thus

for any fixed $T > 0$, we have

$$\begin{aligned}
P_\pi(\tau > T) &= P_\pi(\bigcap_{-T\le t\le 0}(\{R(t) > -t\}\bigcup(\bigcup_{t\le u\le (t+R(t))\wedge 0}(\{Q(u,0) > s\})))\\
&= P_\pi(\bigcap_{-T\le t\le 0}(\{R(T+t) > -t\}\bigcup(\bigcup_{T+t\le u\le (T+t+R(T+t))\wedge T}\{Q(u,0) > s\})))\\
&= P_\pi(\bigcap_{0\le w\le T}(\{R(w) > T-w\}\bigcup(\bigcup_{w\le u\le (w+R(w))\wedge T}\{Q(u,0) > s\})))\\
&= P_\pi(\tau^* > T).
\end{aligned}$$

The second equality holds by stationarity. This gives us $E_\pi\tau = E_\pi\tau^*$.

Next, we use a special construction similar to that in Section 4 of [6] to prove the results for $E^s_\pi\tau^*$. The idea is to use a geometric trial argument. We divide the time frame into blocks that are roughly independent. And if the process is well-behaved (staying around its measure-valued fluid limit) on one block, then τ^* is reached before the end of that block.

Let $\bar{Q}(t,y)$ denote the number of customers in the infinite server system that starts empty at time zero with remaining service time greater than y at time $t \ge 0$. For convenience, we also define $\bar{Q}_u(t,y) = \bar{Q}(u+t,y) - \bar{Q}(u,t+y)$ as the number of customers who arrive after u with remaining service time larger than y at time $u+t$.

4.2.1. *Proof of Theorem 2.* We first prove the theorem for the light-tailed case. The heavy-tail case proceeds by selecting the truncation parameter b sufficiently large.

For the QD regime, by "well-behaved", we mean that the process does not deviate δs, for some $\delta > 0$, from its fluid limit. The following lemma states that the probability of not being "well-behaved" decays exponentially fast with the system scale.

Lemma 7. *Assume $\psi(\theta) < \infty$ for some $\theta > 0$ and X_n's are non-lattice and strictly positive. We also assume the CDF of V_n is continuous. Then for any $\delta > 0$, there exist $I^*(\delta) > 0$, such that*

$$P(\bar{Q}(t,y) > (1+\delta)\lambda s\int_y^{t+y}\bar{F}(u)du \text{ for some } t\in[0,1], y\in[0,\infty)) = \exp(-sI^*(\delta)+o(s)).$$

The proof of Lemma 7 follows form the tow-parameter sample path large deviation result for infinite server queues in [4]. We shall omit it here.

We next introduce our construction of "blocks". Let $l(s) = \inf\{y : (1+\delta)s \int_y^\infty \bar{F}(u)du \leq \frac{1}{2}\}$, we define the following sequence of random times Ξ_i's: $\Xi_0 := 0$. Given Ξ_{i-1} for $i = 1, 2, \cdots$, define

$$r_i := \inf\{k : k \geq R(\Xi_{i-1}), k = 1, 2, \cdots) \ , \ z := \inf\{k : k \geq l(s), k = 1, 2, \cdots\},$$
$$\Xi_i := \Xi_{i-1} + r_i + z.$$

We define a Bernoulli random variable ξ_i, with $\xi_i = 1$ if and only if $\bar{Q}_{\Xi_{i-1}+(k-1)t_0}(t, y) \leq (1+\delta)\lambda s \int_y^{t+y} \bar{F}(u)du$, for all $t \in [0, 1]$, $y \in [0, \infty)$ and every $k = 1, 2, \cdots, r_i + z$.

Choose $\delta < 1/\rho - 1$. We first check that $\xi_i = 1$ implies that τ^* is reached before Ξ_i. Since $r_i \geq R(\Xi_{i-1})$, all the customers in the system at time $\Xi_{i-1} + r_i$ will be those who arrive after Ξ_i. Then $\xi_i = 1$ implies that

$$\begin{aligned} Q(\Xi_{i-1} + r_i, y) &\leq \sum_{k=1}^{r_i/t_0} \int_{(k-1)t_0+y}^{kt_0+y} \bar{F}(u)du \\ &= (1+\delta)\lambda s \int_y^{r_i+y} \bar{F}(u)du \leq (1+\delta)\lambda s \int_y^\infty \bar{F}(u)du, \end{aligned}$$

thus, $R(\Xi_{i-1} + r_i) \leq l(s)$.

And for every $t \in (k-1, k]$, $k = 1, 2, ..., z$

$$Q(\Xi_{i-1} + r_i + t, y) \leq (1+\delta)\lambda s \int_y^{r_i+t+y} \bar{F}(u)du \leq (1+\delta)\lambda s \int_y^\infty \bar{F}(u)du.$$

Thus, $Q(\Xi_{i-1} + r_i + t, 0) \leq (1+\delta)\rho s \leq s$ for $t \in [0, R(\Xi_{i-1} + r_i)]$. Now let $N = \inf\{i \geq 1 : \xi_i = 1\}$, then $E\tau^* \leq E\sum_{i=1}^N (r_i + z)$.

We now give a bound for $E\sum_{i=1}^N (r_i + z)$. The proof is given in the Appendix B.

Lemma 8. *Assume $\psi(\theta) < \infty$ for some $\theta > 0$ and $\psi_N(\theta)$ is continuously differentiable throughout $\mathbb{R}$. We also assume the CDF of V_n is continuous and $EV_n^q < \infty$ for any $q > 0$. Then $E[\sum_{i=1}^N (r_i + z)] = o(s^\delta)$, for any $\delta > 0$.*

This concludes the proof of the light tailed case. We next extend the theorem to the heavy-tailed case. We prove it by drawing connection to the truncated system. Here we delicately choose the truncation parameter b so that the truncated system still operating the QD regime. More specifically, we choose b such that $\int_b^\infty \bar{G}(x)dx < 1/\rho - 1$. This can be achieved since $EX_n = \int_0^\infty \bar{G}(x)dx < \infty$. Then for fixed such b we have

$$\rho_b = \frac{E[V_n]}{E[X_n \wedge b]} = \frac{EV_n}{EX_n - \int_b^\infty \bar{G}(x)dx} < 1$$

and $E^s_\pi \tau(b) = o(s^\delta)$, for any $\delta > 0$, where $\tau(b)$ denote the coalescence time of the truncated system.

We next prove by contradiction that *the coalescence in the truncated system implies the coalescence in the original system with the same amount of information simulated.* Recall that $\tau(b)$ is a random time satisfying that the system has less than s customers at $\tau(b)$. The maximum remaining service time among all customers in the system at time τ is denoted as $R(\tau(b))$. $R(\tau(b)) \leq |\tau(b)|$ and during $R(\tau(b))$ unites of time from $\tau(b)$ on the system always has less than s customers. We can look for $\tau(b)$ at departure times of customers. We assume the process $Q(t, y)$ is right continuous with left limit, so customers departure at time t will not counted in $Q(t, 0)$. Suppose $\tau(b)$ equals to the departure time of the n-th customer. Then every customer arriving between $\tau(b)$ and $\tau(b) + R(\tau(b))$ sees strictly less than s customers (excluding himself) when he enters the system. We set τ equal to the departure time of the n-th customer in the original system and $R(\tau)$ by definition equals to the maximum remaining service time among all customers in the system at time τ. We have $R(\tau) \leq R(\tau(b))$. We claim that every customer arriving between τ and $\tau + R(\tau)$ must see less than s customers (excluding himself) when he enters the system. Suppose this is not the case. Then there exist a customer m, $1 \leq m \leq n$ who arrives between τ and $\tau + R(\tau)$ and finds at least s customers in the system already. The customer with the same index m must have arrived between $\tau(b)$ and $\tau(b) + R(\tau(b))$ in the truncated system and $Q(A_m(b)-) \geq Q(A_m-) \geq s$. We get a contradiction. Therefore, we must have seen the coalescence in the original system as well with the same amount of information simulated.

4.2.2. *Proof of Theorem 3.* For QED regime, by "well-behaved", we mean that the process does not deviate $C\sqrt{s}$, for some $C > 0$, from its fluid limit. The following lemma states that the probability of both being "well-behaved" and not "well-behaved" are bounded away from zero.

Lemma 9. *Fix any $\eta > 0$. Let $\nu(y) = (\int_y^\infty \bar{F}(u)du)^{1/(2+\eta)}$. Assume $EX_n^2 < \infty$ and $EV_n^q < \infty$ for any $q > 0$. Then for any large enough C, there exists $\zeta_1(C) > 0$ and*

$\zeta_2(C) > 0$, *such that*

$$P(\bar{Q}(t,y) \le \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) \ge \zeta_1(C) \quad (5)$$

and

$$P(\bar{Q}(t,y) > \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \text{ for some } t \in [0,1], y \in [0,\infty)) \ge \zeta_2(C). \quad (6)$$

The proof of Lemma 9 follows form the proof of Lemma 9 in [6]. Our case is actually simpler, as we are dealing with a one sided bound (upper bound) as apposed to the two sided bound in [6]. This simplification allows us to remove the light-tail assumption on interarrival time distribution required in [6]. We shall only give an outline of the procedure here.

For Inequality (5), the idea is to consider the diffusion limit of $Q(t,y)$ as a two dimensional Gaussian random field [21], and then invoke Borell-TIS inequality [1].

Assume $EX_n^2 < \infty$, $EV_n < \infty$ and the CDF of V_n is continuous. Pang and Whitt [21] has proved that for $GI/GI/\infty$ queues with any given initial age $E(0)$,

$$\frac{\bar{Q}(t,y) - \lambda s \int_t^{t+y} \bar{F}(u)du}{\sqrt{s}} \Rightarrow R(t,y) \text{ in } D_{D[0,\infty)}[0,\infty),$$

where $R(t,y) = R_1(t,y) + R_2(t,y)$ is a Gaussian random field with

$$R_1(t,y) = \lambda \int_0^t \int_0^\infty I(u+x > t+y)dK(u,x) \text{ and } R_2(t,y) = \lambda c_a^2 \int_0^t \bar{F}(t+y-u)dB(u),$$

where $K(u,x) = W(\lambda u, F(x)) - F(x)W(\lambda u, 1)$ in which $W(\cdot,\cdot)$ is a standard Brownian sheet on $[0,\infty) \times [0,1]$ and $B(\cdot)$ is a standard Brownian motion independent of $W(\cdot,\cdot)$. The constant c_a is coefficient of variation of the interarrival times, i.e. $c_a = \sqrt{\text{Var}(X_n)}/EX_n$. We denote $\tilde{R}_i(t,y) := R_i(t,y)/v(y)$ and define the d-metric (a pseudo-metric) for $i = 1, 2$

$$d_i((t,y),(t',y')) := E[(\tilde{R}_1(t,y) - \tilde{R}_2(t',y'))^2]$$

We then invoke the Borell-TIS inequality. We shall skip the verification of the conditions for such invocation here as it is tedious and detailedly proved in [6]. Let $S = [0,1] \times [0,\infty)$. it is shown in [6] that, there exist constants $M_{i,1} > 0$ and $M_{i,2} > 0$, such that $E[\sup_S \tilde{R}_i(t,y)] \le M_{i,1} < \infty$ and $\sup_S E[\tilde{R}_i(t,y)^2] \le M_{i,2} < \infty$. And for $C_i \ge E[\sup_S \tilde{R}_i(t,y)]$, for $i = 1, 2$,

$$P(\sup_S \tilde{R}_i(t,y) \ge C_i) \le \exp\left\{-\frac{1}{2\sup_S E[\tilde{R}_i(t,y)^2]}(C_i - E[\sup_S \tilde{R}_i(t,y)])^2\right\}.$$

Let $C \geq 2\max\{E[\sup_S \tilde{R}_1(t,y)], E[\sup_S \tilde{R}_2(t,y)]\}$. Then

$$\begin{aligned}
&P(R(t,y) \leq C\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) \\
\geq &P(\sup_S \tilde{R}_1(t,y) + \sup_S \tilde{R}_2(t,y) \leq C) \\
\geq &P(\sup_S \tilde{R}_1(t,y) \leq \frac{C}{2}) P(\sup_S \tilde{R}_2(t,y) \leq \frac{C}{2}) > 0.
\end{aligned}$$

Let X_0 denote the interarrival time of the first customer and V_0 denote its service time. We also denote $\bar{Q}^0(t,y)$ as an independent infinite server process starting empty and with $E(0) = 0$. Then for s large enough, we have

$$\begin{aligned}
&P(\bar{Q}(t,y) \leq \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) \\
=&P(\bar{Q}^0(t-X_0,y) + 1\{V_0 > t+y\} \leq \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \\
&\text{for all } t \in [X_0,1], y \in [0,\infty)) \\
\geq &P(\bar{Q}^0(t,y) + 1\{V_0 > t + X_0 + y\} \leq \lambda s \int_y^{t+X_0+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \\
&\text{for all } t \in [0, 1-X_0], y \in [0,\infty)) \\
\geq &P(\bar{Q}^0(t,y) \leq \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) \\
=&P(\frac{\bar{Q}^0(t,y) - \lambda s \int_y^{t+y} \bar{F}(u)du}{\sqrt{s}} \leq C\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)).
\end{aligned}$$

It is easy to check that the set $\{f : |f(t,y)| \leq C\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)\}$ is a continuity set, thus by the Functional Central Limit Theorem result in [21], we have

$$\begin{aligned}
&P(\frac{\bar{Q}^0(t,y) - \lambda s \int_y^{t+y} \bar{F}(u)du}{\sqrt{s}} \leq C\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) \\
\to &P(R(t,y) \leq C\nu(y) \text{ for all } t \in [0,1], y \in [0,\infty)) > 0.
\end{aligned}$$

Inequality (6) is easy to prove as we can always isolate a point (t^*, y^*) inside S. The projection of the process on that point posses Gaussian distribution. More specifically,

$$\begin{aligned}
&P(\bar{Q}(t,y) > \lambda s \int_y^{t+y} \bar{F}(u)du + C\sqrt{s}\nu(y) \text{ for some } t \in [0,1], y \in [0,\infty)) \\
\geq &P(\bar{Q}(t^*,y^*) > \lambda s \int_{y^*}^{t^*+y^*} \bar{F}(u)du + C\sqrt{s}\nu(y^*)) \\
=&P(\frac{\bar{Q}(t^*,y^*) - \lambda s \int_{y^*}^{t^*+y^*} \bar{F}(u)du}{\sqrt{s}} > C\nu(y^*)),
\end{aligned}$$

and by Fatou's lemma

$$\liminf_{s\to\infty} P(\frac{\bar{Q}(t^*,y^*)-\lambda s\int_{y^*}^{t^*+y^*}\bar{F}(u)du}{\sqrt{s}} > C\nu(y^*)) \geq P(R(t^*,y^*) > C\nu(y^*)) > 0.$$

Let $m(s) = \inf\{y : C\sqrt{s}(v(y) + \int_y^\infty v(s)ds) \leq \frac{1}{2}\}$. Following the same construction as for the QD regime, we define the sequence of random times Ξ_i's as follows: $\Xi_0 := 0$. Given Ξ_{i-1} for $i = 1, 2, \cdots$,

$$r_i := \inf\{k : k \geq R(\Xi_{i-1}), k = 1, 2, ...) \ , \ z := \inf\{k : k \geq m(s), k = 1, 2, ...\},$$
$$\Xi_i := \Xi_{i-1} + r_i + z.$$

We introduce a Bernoulli random variable ξ_i with $\xi_i = 1$ if and only if $\bar{Q}_{\Xi_{i-1}+(k-1)t_0}(t,y) \leq \lambda s\int_y^{t+y}\bar{F}(u)du + C\sqrt{s}\nu(y)$, for all $t \in [0,1]$, $y \in [0,\infty)$ and every $k = 1, 2, ..., r_i + z$.

We next show that $\xi_i = 1$ implies that τ^* is reached before Ξ_i. Since $r_i \geq R(\Xi_{i-1})$, all the customers at time $\Xi_{i-1} + r_i$ will be those arrive after Ξ_i. Thus we have $\xi_i = 1$ implies that

$$\begin{aligned} Q(\Xi_{i-1}+r_i, y) &\leq \sum_{k=1}^{r_i}\{\lambda s\int_{(k-1)t_0+y}^{kt_0+y}\bar{F}(u)du + C\sqrt{s}\nu((k-1)+y))\} \\ &\leq \lambda s\int_y^\infty \bar{F}(u)du + C\sqrt{s}(\nu(y) + \int_y^\infty \nu(u)du). \end{aligned}$$

As $\int_y^\infty \bar{F}(u)du$ decays faster than $\nu(y)$ as y grows large, for s large enough, we have $R(\Xi_{i-1} + r_i) < m(s)$.

Likewise for every $t \in (k-1, k]$ and $k = 1, 2, \cdots, z$,

$$Q(\Xi_{i-1}+r_i+t, y) \leq \lambda s\int_y^\infty \bar{F}(u)du + C\sqrt{s}(\nu(y) + \int_y^\infty \nu(u)du).$$

Thus when $\beta > C(\nu(0) + \int_0^\infty \nu(u)du)$, for $t \in [0, R(\Xi_{i-1} + r_i)]$, we have

$$Q(\Xi_{i-1}+r_i+t, 0) \leq s + C(\nu(0) + \int_0^\infty \nu(u)du)\sqrt{s} \leq s + \beta\sqrt{s}.$$

Now let $N = \inf\{i \geq 1 : \xi_i = 1\}$. Then $E\tau^* \leq E[\sum_{i=1}^N (r_i + z)]$.

We next show a bound for $E\sum_{i=1}^N (r_i + z)$. The proof is given in the Appendix B.

Lemma 10. *Assume $EX_n^2 < \infty$ and $EV_n^q < \infty$ for any $q > 0$. Then $\log E[\sum_{i=1}^N (r_i + z)] = o(s^\delta)$, for any $\delta > 0$.*

Notice that our proof of Theorem 3 only requires the existence of the second moment of the interarrival time distribution. We thus conclude the proof of Theorem 3.

TABLE 1: Simulation results for τ (QD: $\lambda = s, C_s = 1.2s$)

s	mean	95% confidence interval
100	22.6297	$[21.3381, 23.9213]$
500	15.6162	$[15.1791, 16.0533]$
1000	15.8816	$[15.4559, 16.3073]$

TABLE 2: Simulation results for τ (QED: $\lambda = s, C_s = s + 2\sqrt{s}$)

s	mean	95% confidence interval
100	22.6297	$[21.3381, 23.9213]$
500	37.0449	$[32.7770, 41.3128]$
1000	42.0704	$[37.9622, 46.1786]$

4.2.3. *Simulation experiment* In this subsection, we run some numerical experiments aimed at verifying the running time of our algorithm measured by $E^s_\pi[\tau]$ for different values of s. The algorithms appear to have substantially better performance in practice. In the QD regime, our numerical experiments suggest that $E^s_\pi[\tau]$ is almost bounded as apposed to grow sub-linearly with s indicated by Theorem 2. This is because in the QD regime, the stationary probability that the queue length process is above C_s decays exponentially with the system scale s. In the QED regime, our numerical experiments suggest a growth rate of $O(\sqrt{s})$ as apposed to the sub-exponentially growth rate in Theorem 3. This empirical bound is intuitive, as in the QED regime, the situation when coalescence occurs is similar to the case when a mean zero random walk spends s units of time below 0. If the increments of the random walk have finite variance, this situation occurs with probability $O(1/\sqrt{s})$.

The performance was tested using a wide range of distributions and the overall conclusions are similar. The numbers displayed (Table 1 and Table 2) are obtained assuming that a generic base interarrival time, X_n, follows a Gamma distribution with shape parameter 2 and rate parameter 2 ($\Gamma(2,2)$). For the s-th system, the interarrival is distributed as X_n/s, and a generic service time, V_n, follows lognormal distribution, where $\log V_n \sim N(-1/2, 1/2)$. We use 10^3 replications for each value of s.

We tested our codes in the case of of Poisson arrivals and exponential service time

distributions. In this case, $E[Q^L(\infty,0)]$ can be computed analytically. We tried: a) $\lambda = 100$, $\mu = 1$ and $C = 105$, in this case, $E[Q^L(\infty,0)] = 95.1739$; b) $\lambda = 100$, $\mu = 1$ and $C = 105$, in this case, $E[Q^L(\infty,0)] = 97.2537$; c) $\lambda = 500$, $\mu = 1$ and $C = 511$, in this case, $E[Q^L(\infty,0)] = 488.7970$; d) $\lambda = 500$, $\mu = 1$ and $C = 550$, in this case, $E[Q^L(\infty,0)] = 499.2344$. In all of theses cases, we simulated 10^3 replications. The corresponding 95% confidence interval contains the true value in each case.

Appendix A. Proof of Proposition 1

By Chebyshev's inequality,

$$P(A_{n+1} < n(\mu-\epsilon)) \le E[\exp(\theta(n(\mu-\epsilon) - A_{n+1}))] \le \exp(-n(-\theta(\mu-\epsilon) - \psi(-\theta)))$$

for any $\theta \ge 0$.

Let $I(-\epsilon) := \max_{\theta\ge 0}\{-\theta(\mu-\epsilon) - \psi(-\theta)\}$. As $\psi(0) = 0$, $\psi'(0) = \mu$ and $\psi''(0) = Var(X) > 0$, $I(-\epsilon) > 0$. Then $P(A_{n+1} < n(\mu-\epsilon)) \le \exp(-nI(-\epsilon))$, and

$$\sum_{n=1}^{\infty} P(A_{n+1} < n(\mu-\epsilon)) \le \frac{\exp(-I(-\epsilon))}{1-\exp(-I(-\epsilon))} < \infty.$$

By Borel-Cantelli lemma, $\{A_{n+1} \ge n(\mu-\epsilon)\}$ eventually almost surely.

Similarly and independently we have

$$\begin{aligned}\sum_{n=1}^{\infty} P(|V_{n+1}| > (n(\mu-\epsilon))^{\alpha}) &= \sum_{n=1}^{\infty} P(|V_1|^{1/\alpha} > n(\mu-\epsilon)) \\ &\le \frac{1}{\mu-\epsilon}\int_0^{\infty} P(|V_1|^{1/\alpha} > \nu)d\nu < \infty.\end{aligned}$$

Thus, again by Borel-Cantelli lemma, $\{|V_{n+1}| \le (n(\mu-\epsilon))^{\alpha}\}$ eventually almost surely. Therefore, $P(\kappa < \infty) = 1$.

Appendix B. Proof of Lemma 8 and Lemma 10

We first prove the following two lemmas (Lemma 11 and Lemma 12) as a preparation.

Lemma 11. *If $EV_n^q < \infty$ for any $q > 0$, then for any fixed $p > 0$, $E[(\max_{k=1,2,\ldots,n} V_k)^p] = o(n^{\delta})$ for any $\delta > 0$.*

Proof. For any fixed $\delta > 0$ we can find $\delta' \in (0, \delta)$. Let $q = 1/\delta' + p$. By Chebyshev's inequality we have $\bar{F}(u) \le E[V^q]/u^q$. Let $\bar{F}_n(u) = P(\max_{k=1,2,\ldots,n} V_k > u)$ then

$$\begin{aligned} E[(\max_{k=1,2,\ldots n} V_k)^p] &= p\int_0^\infty u^{p-1}\bar{F}_n(u)\,du \\ &\le n^{1/(q-p)} + np\int_{n^{1/(q-p)}}^\infty u^{p-1}\bar{F}(u)\,du \\ &\le n^{1/(q-p)} + np\int_{n^{1/(q-p)}}^\infty \frac{EV^q}{u^{q-p+1}}\,du = n^{\delta'} + \frac{p}{q-p}EV^q. \ \square \end{aligned}$$

Now we turn to Lemma 12. First notice that by Holder's inequality, we have

$$E[\sum_{i=1}^N (r_i + z)] = E[\sum_{i=1}^\infty (r_i + z) I\{N \ge i\}] \le \sum_{i=1}^\infty E[(r_i+z)^2]^{1/2} P(N \ge i)^{1/2}.$$

Lemma 12. *If $EX_n < \infty$ and $EV_n^q < \infty$ for any $q > 0$, then for any $p \ge 1$ we have $E[(r_i + z)^p]^{1/p} = o(s^\delta)$, for any $\delta > 0$.*

Proof. By Minkowski inequality, $E[(r_i+z)^p]^{1/p} \le E[r_i^p]^{1/p} + z$. Using similar argument as in the proof of Lemma 11, we can show that $l(s) = o(s^\delta)$ for any $\delta > 0$, thus $z = o(s^\delta)$ for any $\delta > 0$.

For fixed $\delta > 0$, we can find $\delta' \in (0, p\delta/(1+p\delta))$, such that

$$\begin{aligned} E[r_i^p] &\le E[E[(\max_{k=1,\ldots,N_s(\Xi_{i-1})-N_s(\Xi_{i-2})} V_k)^p | N_s(\Xi_{i-1}) - N_s(\Xi_{i-2})]] \\ &\le CE[(N_s(\Xi_{i-1}) - N_s(\Xi_{i-2}))^{\delta'}] \text{ Lemma 11} \\ &\le C(E[N_s(\Xi_{i-1}) - N_s(\Xi_{i-2})])^{\delta'} \text{ Jensen's inequality for concave function} \\ &\le C\tilde{\lambda}^{\delta'} s^{\delta'} E[r_{i-1} + z]^{\delta'} \text{ Key Renewal Theorem.} \end{aligned}$$

Let $w_i = r_i + z$ for $i = 1, 2, \cdots$. As z it is a constant that only depends on s and $z = o(s^{\delta'})$, then $Ew_i \ge z \ge 1$ and $Ew_i = Er_i + z \le C\tilde{\lambda}^{\delta'} s^{\delta'} (Ew_{i-1})^{\delta'} + z \le \tilde{C} s^{\delta'} (Ew_{i-1})^{\delta'}$, where $\tilde{C} = C\tilde{\lambda}^{\delta'} + 1$. As $E[r_1^p] = E_\pi[R(0)^p] = o(s^{\delta'})$. By iteration we have $Ew_i \le \tilde{C}^{1/(1-\delta')} s^{\delta'/(1-\delta')}$ for $i = 1, 2, \cdots$. Thus $Er_i^p = o(s^{p\delta})$ and $E[(r_i+z)^p]^{1/p} = o(s^\delta)$. $\square$

Proof of Lemma 8. We first notice that $P(\xi_i = 0) \le E[w_1]\exp(-sI^*(\delta) + o(s))$ by Lemma 7. $P(N \ge 1) = 1$ and $P(N \ge 2) = P(\xi_1 = 0) \le E[w_1]\exp(-sI^*(\delta) + o(s))$. Recall that $w_i = r_i + z$ for $i = 1, 2, \cdots$.

$$\begin{aligned} P(N \ge 3) = P(N \ge 1)P(N \ge 3 | N \ge 2) &= P(\xi_1 = 0)P(\xi_2 = 0 | \xi_1 = 0) \\ &\le P(\xi_1 = 0)E[w_2 | \xi_1 = 0]\exp(-sI^*(\delta) + o(s)) \\ &\le E[w_1]E[w_2|\xi_1 = 0]\exp(-2sI^*(\delta) + o(s)). \end{aligned}$$

We next prove that $E[w_2|\xi_1 = 0] = \exp(o(s))$. Notice that $P(\xi_i = 0) \geq \exp(-sI^*(\delta) + o(s))$ by Lemma 7. Then for any $p > 0$, $q > 0$ and $1/p + 1/q = 1$,

$$\begin{aligned} E[w_2|\xi_1 = 0] &= \frac{E[w_2 I\{\xi_1 = 0\}]}{P(\xi_1 = 0)} \\ &\leq \frac{E[w_2^p]^{1/p} P(\xi_1 = 0)^{1/q}}{P(\xi_1 = 0)} \text{ Holder's inequality} \\ &\leq E[w_2^p]^{1/p} E[w_1]^{1/q} \exp(\frac{1}{p} sI^*(\delta) + o(s)). \end{aligned}$$

Thus, $1/s \log E[w_2|\xi_1 = 0] \leq 1/s\,(1/p \log E[w_2^p] + 1/q \log E[w_1] + o(s)) + 1/pI^*(\delta)$. By sending p to infinity, we have $E[w_2|\xi_1 = 0] = \exp(o(s))$.
Similarly by iteration, $P(N \geq k) = \exp(-ksI^*(\delta) + o(s))$ for $k = 4, 5, \cdots$. Then $\sum_{i=1}^{\infty} P(N \geq i)^{1/2} = O(1)$. As $E[\sum_{i=1}^{N}(r_i + z)] \leq \sum_{i=1}^{\infty} E[(r_i + z)^2]^{1/2} P(N \geq i)^{1/2}$ and $E[(r_i + z)^2]^{1/2} = o(s^\delta)$ for any $\delta > 0$, we have $E[\sum_{i=1}^{N}(r_i + z)] = o(s^\delta)$. □

Proof of Lemma 10. $P(N \geq 1) = 1$.

$$\begin{aligned} P(N \geq 2) = P(\xi_1 = 0) &\leq 1 - E[\zeta_1(C)^{w_1}] \text{ Lemma 9} \\ &\leq 1 - \zeta_1(C)^{E[w_1]} \text{ Jensen's inequality} \\ &= 1 - b \exp(-o(s^\delta)). \end{aligned}$$

Moreover,

$$\begin{aligned} P(N \geq 3) = P(N > 2|N > 1)P(N > 1) &= P(\xi_2 = 0|\xi_1 = 0)P(\xi_1 = 0) \\ &\leq E[1 - \zeta_1(C)^{w_2}|\xi_1 = 0]P(\xi_1 = 0) \\ &\leq (1 - \zeta_1(C)^{E[w_2|\xi_1=0]})P(\xi_1 = 0). \end{aligned}$$

We next show that $E[w_2|\xi_1 = 0] = o(s^\delta)$ for any $\delta > 0$. Notice that $P(\xi_i = 0) \geq \zeta_2(C)$ by Lemma 9, then $E[w_2|\xi_1 = 0] = E[w_2 I\{\xi_1 = 0\}]/P(\xi_1 = 0) \leq E[w_2]/\zeta_2(C)$.
Similarly by iteration we have $P(N \geq k) \leq (1 - b \exp(-o(s^\delta)))^k$ for any $\delta > 0$ and $k = 4, 5, \ldots$ Then for any $\delta > 0$, $\log \sum_{i=1}^{\infty} P(N \geq i)^{1/2} = o(s^\delta)$. As $E[\sum_{i=1}^{N}(r_i + z)] \leq \sum_{i=1}^{\infty} E[(r_i + z)^2]^{1/2} P(N \geq i)^{1/2}$ and $E[(r_i + z)^2]^{1/2} = o(s^\delta)$ for any $\delta > 0$, we have $\log E[\sum_{i=1}^{N}(r_i + z)] = o(s^\delta)$. □

Acknowledgement

NSF support from grants DMS 1320550 and CMMI 1069064 is gratefully acknowledged. The authors would also like to thank the referee for her/his careful reading of our manuscript and insightful comments.

References

[1] Adler, R. J. (1990). An introduction to continuity, extrema, and related topics for general Gaussian processes. *IMS Lecture Notes: Monograph Series* **12,**.

[2] Asmussen, S. (2003). *Applied Probability and Queues* 2 ed. Spinger, New York.

[3] Berthelsen, K. and Møller, J. (2002). A primer on perfect simulation for spatial point process. *Bull Braz Math Soc* **33(3),** 351–367.

[4] Blanchet, J., Chen, X. and Lam, H. (2012). Two-parameter sample path large deviation for infinite server queues. *working paper*.

[5] Blanchet, J. and Dong, J. (2012). Sampling point processes on stable unbounded regions and exact simulation of queues. *Proc. of 2012 Winter Simulation Conference*.

[6] Blanchet, J. and Lam, H. (2012). Rare-event simulation for many-sever queues. *working paper*.

[7] Blanchet, J. and Sigman, K. (2011). On exact sampling of stochastic perpetuities. *J. Appl. Probab.* **48A,** 165–182.

[8] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2002). Statistical analysis of a telephone call center: a queueing-science perspective. *Preprint*.

[9] Busic, A., Gaujal, B. and Perronnin, F. (2012). Perfect sampling of networks with finite and infinite capacity queues. In *ASMTA*. vol. 7314. Springer pp. 136–149.

[10] Connor, S. and Kendall, W. (2007). Perfect simulation for a class of positive recurrent Markov chains. *Ann. Appl. Probab.* **3,** 781–808.

[11] Corcoran, J. and Tweedie, R. (2001). Perfect sampling of ergodic Harris chains. *Ann. of Appl. Probab.* **11,** 438–451.

[12] Dong, J. (2014). Studies in stochastoc networks: Efficient Monte Carlo methods, modeling and asymptotic analysis. Ph.D. Thesis, Columbia University.

[13] Ensor, K. and Glynn, P. (2000). Simulating the maximum of a random walk. *Journal of Statistical Planning and Inference* **85,** 127–135.

[14] Fernandez, R., Ferrari, P. and Garcia, N. (2002). Perfect simulation for interacting point processes, loss networks and ising models. *Stoch. Process. Appl.* **102(1),** 63–88.

[15] Foss, S. and Tweedie, R. (1998). Perfect simulation and backward coupling. *Stochastic Models* **14,** 187–203.

[16] Kelly, F. (1991). Loss networks. *Annals of Applied Probability* 319–378.

[17] Kendall, W. (1998). Perfect simulation for area-interaction point processes. In *Probability Towards 2000.* ed. L. Accardi and C. Heyde. Spinger, New York pp. 218–234.

[18] Kendall, W. (2004). Geometric ergodicity and perfect simulation. *Electron. Comm. Probab.* **9,** 140–151.

[19] Kendall, W. and Møller, J. (2000). Perfect simulation using dominating processes on ordered spaces, with application to locally stable point pocesses. *Adv. Appl. Prob.* **32,** 844–865.

[20] Murdoch, D. and Takahara, G. (2006). Perfect sampling for queues and network models. *ACM Transactions of Modeling and Computer Simulation* **16,** 76–92.

[21] Pang, G. and Whitt, W. (2010). Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems: Theory and Applications* 325–264.

[22] Propp, J. and Wilson, D. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9,** 223–252.

[23] Sigman, K. (2011). Exact simulation of the stationary distribution of the FIFO M/G/c queue. *Journal of Applied Probability* **48A,** 209–216.

[24] Sigman, K. (2012). Exact simulation of the stationary distribution of the FIFO M/G/c queue: The general case of $\rho < c$. *Queueing Systems: Theory and Applications* **70,** 37–43.