A Law of Large Numbers for M/M/c/Delayoff-Setup Queues with Nonstationary Arrivals

Jamol Pender[1] and Tuan Phung-Duc[2(✉)]

[1] School of Operations Research and Information Engineering, Cornell University, Ithaca, USA
jjp274@cornell.edu
[2] Division of Policy and Planning Sciences, University of Tsukuba, Tsukuba, Japan
tuan@sk.tsukuba.ac.jp

Abstract. Cloud computing is a new paradigm where a company makes money by selling computing resources including both software and hardware. The core part or infrastructure of cloud computing is the data center where a large number of servers are available for processing incoming data traffic. These servers not only consume a large amount of energy to process data, but also need a large amount of energy to keep cool. Therefore, a reduction of a few percent of the power consumption means saving a substantial amount of money for the company as well as reduce our impact on the environment. As it currently stands, an idle server still consumes about 60 % of its peak energy usage. Thus, a natural suggestion to reduce energy consumption is to turn off servers which are not processing data. However, turning off servers can affect the customer experience. Customers trying to access computing power will experience delays if their data cannot be processed quickly enough. Moreover, servers require setup times in order to move from the off state to the on state. In the setup phase, servers consume energy, but cannot process data. Therefore, there exists a trade-off between power consumption and delay performance. In [7,9], the authors analyze this tradeoff using an M/M/c queue with setup time for which they present a decomposition property by solving difference equations. In this paper, we complement recent stationary analysis of these types of models by studying the sample path behavior of the queueing model. In this regard, we prove a weak law of large numbers or fluid limit theorem for the queue length and server processes as the number of arrivals and number of servers tends to infinity. This methodology allows us to consider the impact of nonstationary arrivals and abandonment, which have not been considered in the literature so far.

Keywords: Setup time · Abandonment · Power-saving

S. Wittevrongel and T. Phung-Duc (Eds.): ASMTA 2016, LNCS 9845, pp. 253–268, 2016.
DOI: 10.1007/978-3-319-43904-4_18

1 Introduction

1.1 Motivation

The core part of cloud computing is the data center where a large number of servers are available to serve the demand generated by the arrival of data traffic. These servers consume a large amount of energy, which translates into a large cost for many cloud computing companies. It is reported that data centers worldwide consume as much as about 20–30 GW of electricity [11]. However, a large part of this energy is consumed by idle servers which do not process any jobs. In fact, it is reported that an idle server still consumes about 60 % of its peak energy usage when processing jobs [3]. Thus, an important issue for the management of these data centers is to minimize the power consumption while maintaining a high quality of service for their customers. A simple way to minimize the power consumption in data center is to turn off idle servers. However, servers that are off eventually need to be turned on in order to process waiting jobs, which causes more delays. In fact, servers require some setup time in order to gain the ability to start processing jobs. Moreover, during this setup time, servers also consume a substantial amount of energy but cannot process waiting jobs. Thus, there exists a trade-off between saving power and the quality of service provided by the company. This motivates our study of multiserver queues with setup times.

In practice, the amount of requests that arrive at a data center varies time to time. It is natural that traffic in daytime is different from that in nighttime. The amount of traffic is also different on weekdays and weekends. This motivates us to consider time-non-homogeneous arrival processes. Furthermore, today data centers are partially operated by renewable energy such that wind energy or solar energy [11]. These energy sources depend on the weather and often vary on time. Thus, the number of available servers is time-dependent. This calls for the need of studying a queueing system with time-dependent number of servers. As mentioned above since setup time not only incurs in extra waiting time it may also incur in increasing energy consumption because a server consumes a large amount of energy during setup. Therefore, it is not a good strategy to turn off a server immediately upon idle. In our model, we allow an idle time before shutdown. A job arriving during the idle time is served immediately while if there is no arriving customer during the idle time, the server is switched off. Requests to data centers have time limiting nature and thus they will abandon after some waiting time. This may cause by impatient user or by the timeout of a web browser. We incorporate customer abandonment in our model. We allow all the setup rate and abandonment rate to be time-dependent. To the best of our knowledge, this paper is the first to consider a time-dependent queueing model for power-saving data centers.

1.2 Literature Review

Artalejo et al. [2] present a thorough analysis for multiserver queues with setup times where the authors consider the case in which at most one server can be in

the setup mode at a time. This policy is later referred to as staggered setup in the literature [9]. Artalejo et al. [2] show an analytical solution by solving the set of balance equations for the joint stationary distribution of the number of active servers and that of jobs in the system using a difference equation approach. The solution of the staggered setup model is significantly simplified by Gandhi et al. [9] who also present a decomposition property for the queue length and the waiting time.

Recently, motivated by applications in data centers, multiserver queues with setup times have been extensively investigated in the literature. In particular, Gandhi et al. [9] present a stationary analysis for multiserver queues with setup times. They obtain some closed form approximations for the ON-OFF policy where any number of servers can be in the setup mode. As is pointed out in Gandhi et al. [9], from an analytical point of view the most challenging model is the ON-OFF policy where the number of servers in setup mode is not limited. Recently, Gandhi et al. [7,8] analyze the M/M/c/Setup model with ON-OFF policy using a recursive renewal reward approach. Gandhi et al. [7,8] claim that the model is difficult to be solved using conventional methods such as generating function or matrix analytic methods. As a result, the recursive renewal reward approach is presented as a new mathematical tool to resolve the problem. Phung-Duc [28] analyzes the same model via generating function and matrix analytic methods. It should be noted that in all the work above, arrival, service and setup processes are time-homogeneous and abandonment of customers is not taken into account.

However, as is mentioned above, in reality, traffic to data center has time-inhomogeneous nature because it is generated by human users whose activities clearly depend on time. Furthermore, nowadays, many data centers partially operate using renewable resources such as wind or solar energies [1,10]. As a result, the number of available servers also depends on time. On the other hand, ON-OFF control of servers may also incur in extra delays which cause abandonment of customers. Therefore, there is a need to develop and analyze a model taking into account all of these factors and that is the aim of the current paper.

1.3 Main Contributions of Paper

In this work, we make the following contributions to the literature on queueing theory:

1. We develop a new queueing model that incorporates a stochastic number of servers with the Delay-off feature, abandonment of jobs, and nonstationary arrival times of jobs.
2. We propose a heuristic mean field limit for the queue length and non-idle server processes.
3. We prove that the mean field heuristic is asymptotically true when the arrival rate and number of servers tend to infinity.

1.4 Organization of Paper

The rest of this paper is organized as follows. Section 2 presents the model in detail while Sect. 3 is devoted to the analysis where we present a mean field approximation and fluid limit. Section 4 shows some numerical examples showing insights into the performance of the system. Concluding remarks are presented in Sect. 5.

1.5 Notation

The paper will use the following notation:

- $\lambda(t)$ is the external arrival rate of jobs to the data center at time t
- $\mu(t)$ is the service rate of all of the servers at time t
- $\theta(t)$ is the abandonment rate of jobs at time t
- $\beta(t)$ is the rate at which needed servers transition from the OFF state to the ACTIVE (BUSY) state at time t
- $\gamma(t)$ is the rate at which unneeded servers transition from the IDLE state to the OFF state at time t
- C_{max} is the bound on the number of servers in the data center facility
- $x \wedge y = \min(x, y)$
- $(x - y)^+ = \max(0, x - y)$

2 $M_t/M/c$/Delayoff-Setup+M Queueing Model

We consider $M_t/M/C_{max}(t)$/Setup+M queueing systems with ON-OFF policy. Jobs arrive at the system according to a time-dependent Poisson process with rate $\lambda(t)$. In this system, after a service completion, if there is a waiting job, the server pickups this job to process immediately. Otherwise, the server stays IDLE for a while and then is switched off. We assume that the switch-off time is instantaneous. The service rate of a server is $\mu(t)$. The rate at which the server changes to OFF state is $\gamma(t)$. However, if there is some waiting customer, an OFF server is switched to the ON state with rate $\beta(t)$. We call $\beta(t)$ the setup rate. Because jobs have time-limiting nature, we assume that each waiting job abandons with rate $\theta(t)$. For this system, let $Q(t)$ denote the number of jobs in the system at time t and $C(t)$ denote the total number of BUSY and IDLE servers at time t. In our system, a server can take one of the following states: OFF, IDLE (not serving a job), BUSY (serving a job), SETUP. In the OFF state, the server does not consume energy but also does not process a job. In the IDLE state, the server consumes energy but does not process any job. In this the IDLE state, the server can process an arriving job immediately. If a job arrives at the system and there are not idle servers, the job is queued and an OFF server is activated and that server changes to the SETUP state. After the setup time, the server processes the waiting job.

Under our setting, the number of servers in setup at time t is given by $S(t) = ((Q(t) - C(t))^+ \wedge (C_{max}(t) - C(t))$ and the number of IDLE servers at time t is given by $(C(t) - Q(t))^+$, respectively.

We will model our version of the setup queue model with abandonment with a two dimensional Markov process. In fact, it is possible to derive a sample path representation of the queueing model via the work of [17] or [18] that is given by the following stochastic integral equation

$$\begin{aligned} Q(t) &= Q(0) + \Pi_1\left(\int_0^t \lambda(s)ds\right) - \Pi_2\left(\int_0^t \mu(s)\cdot(Q(s)\wedge C(s))ds\right) \\ &\quad -\Pi_3\left(\int_0^t \theta(s)\cdot(Q(s)-C(s))^+ds\right) \qquad (1) \\ C(t) &= C(0) + \Pi_4\left(\int_0^t \beta(s)\cdot S(s)ds\right) \\ &\quad -\Pi_5\left(\int_0^t \gamma(s)\cdot(C(s)-Q(s))^+ds\right). \qquad (2) \end{aligned}$$

The first stochastic process $Q(t)$ is for the queue length and the second stochastic process $C(t)$ is to keep track of the total number of busy servers and idle servers. With this construction, we need to define the Poisson processes Π_i that are used. For the first Poisson process Π_1, it counts the number of arrivals of jobs to be processed at the data center during the interval $(0,t]$. For Π_2, we have that it counts the number of service completions from the data center in the interval $(0,t]$. Similarly, Π_3, we have that it counts the number of jobs that have abandoned or timed out from the data center in the interval $(0,t]$. For the Poisson process Π_4 we have that it counts the number of servers that have been turned on when there is sufficient number of jobs that need to be processed. Lastly, Π_5 represents the number of servers that have been turned off because the idle times expire and jobs do not arrive.

With our stochastic model representation for a data center, there are several important observations to make under certain parameter settings. When we let the delay-off parameter $\gamma = 0$, we construct a situation where none of the servers can be turned off. Thus, the number of servers will increase until it reaches its maximum and when the maximum is reached, the queue will behave as a nonstationary multiserver or $M_t/M/C_{max}$ queue. When $\gamma = \infty$ server is turned off immediately when they are considered to be idle. In this case, the number of servers mimicks the number of jobs in the system as the number of jobs decreases. Moreover, in this setting, the least amount of energy is used since servers are immediately turned off when they become idle. However, turning a server off immediately can cause unnecessary delays for future jobs.

Since the joint process $(Q(t), C(t))$ is clearly Markovian with time-dependent infinitesimal generator A_t defined on continuous and bounded functions $h : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ which has the following representation

$$\begin{aligned} \mathcal{A}_t h(x,y) &\equiv \lim_{\Delta\to 0}\frac{E[g(t,\Delta)|(Q(t),C(t))=(x,y)]-h(x,y)}{\Delta} \\ &= \sum_{c\in\mathcal{C}} r_c(x,y,t)\cdot[h(x+\delta_{(x,c)},y+\delta_{(y,c)})-h(x,y)], \end{aligned}$$

where $g(t, \Delta) = h(Q(t+\Delta), C(t+\Delta))$. It thus follows by Dynkin's formula, see for example Lemma 17.21 of [15], that for $t \in \mathbb{R}_+$

$$E[h(Q(t), C(t))] = h(x_0, y_0) + \int_0^t E[\mathcal{A}_s h(Q(s), C(s))]ds$$

Using the Dynkin's formula for Markov processes and the Poisson process representation of the stochastic queueing model, we can subsequently derive the functional Kolmogorov forward equations for the two dimensional Markov process as

$$\begin{aligned}
&\frac{d}{dt}E[h(Q(t), C(t))] \equiv \dot{E}[h(Q(t), C(t))] \\
&\equiv \dot{E}[h(Q(t), C(t)) \mid Q(0) = Q_0, C(0) = C_0\ ] \\
&= E[\lambda(t) \cdot (h(Q+1, C) - h(Q, C))] \\
&+ E[\mu(t) \cdot (Q \wedge C) \cdot (h(Q-1, C) - h(Q, C))] \\
&+ E[\theta(t) \cdot (Q - C)^+ \cdot (h(Q-1, C) - h(Q, C))] \\
&+ E[\beta(t) \cdot (S \cdot (h(Q, C+1) - h(Q, C))] \\
&+ E[\gamma(t) \cdot (C - Q)^+ \cdot (h(Q, C-1) - h(Q, C))],
\end{aligned}$$

where we omit (t) of $Q(t), C(t)$ and $S(t)$ in the right hand side for simplicity. When we let $h(x, y) := x$ or $h(x, y) := y$, we have the following equations for the mean queue length and mean number of non-idle servers

$$\begin{aligned}
\dot{E}[Q(t)] &= \lambda(t) - \mu(t) \cdot E[(Q \wedge C)] - \theta(t) \cdot E[(Q - C)^+] \\
\dot{E}[C(t)] &= \beta(t) \cdot E[((Q - C)^+ \wedge (C_{max} - C)] - \gamma(t) \cdot E[(C - Q)^+].
\end{aligned}$$

Equations for second-order moments can be obtained by choosing $h(x, y) := (x \cdot y, x^2, y^2)$. In fact, monomial functions of any order can be used to obtain equations for moments of arbitrary orders by letting $h(x, y) := x^i \cdot y^j$. However, if the rate functions, which define the time changed Poisson processes, are non-linear (as is usually the case and is the case here), the term $E[\mathcal{A}_s h(Q(s), C(s))]$ will involve expectations of non-linear functions of the stochastic processes and will thus need to be simplified by applying some form of moment-closure approximation. One type of moment closure technique is the mean field approximation.

2.1 Mean Field Approximation

Using the functional Kolmogorov forward equations as outlined in [5,12,19,20], we have the following system of differential equations for the mean queue length and the mean number of non-idle servers

$$\begin{aligned}
\dot{E}[Q(t)] &= \lambda(t) - \mu(t) \cdot E[(Q \wedge C)] - \theta(t) \cdot E[(Q - C)^+] \\
\dot{E}[C(t)] &= \beta(t) \cdot E[((Q - C)^+ \wedge (C_{max} - C)] - \gamma(t) \cdot E[(C - Q)^+]
\end{aligned}$$

Now if we use a mean field approximation i.e.

$$E[f(X)] = f(E[X]) \tag{3}$$

we have that

$$\dot{E}[Q(t)] \approx \lambda(t) - \mu(t) \cdot (E[Q] \wedge E[C]) - \theta(t) \cdot (E[Q] - E[C])^+$$
$$\dot{E}[C(t)] \approx \beta(t) \cdot ((E[Q] - E[C])^+ \wedge (C_{max} - E[C]) - \gamma(t) \cdot (E[C] - E[Q])^+$$

Unlike the exact equations for the mean queue length and the mean number of non-idle servers, the system of equations for the mean field approximation is an autonomous dynamical system and can be solved numerically quite easily. However, this approximation is a heuristic and is not rigorous. We will show in the sequel that the mean field approximation can be made rigorous by an appropriate scaling limit of our queueing model.

3 A Weak Law of Large Numbers Limit

In order to prove a fluid limit for the queue length process and the number of servers, we need to scale our system appropriately. We define $Q^\eta(t)$ and $C^\eta(t)$ as the following stochastic processes in terms of time changed Poisson processes.

$$Q^\eta(t) = Q^\eta(0) + \Pi_1\left(\int_0^t \eta\lambda(s)ds\right) - \Pi_2\left(\eta\int_0^t \mu(s)\cdot(\bar{Q}^\eta(s) \wedge \bar{C}^\eta(s))ds\right)$$
$$-\Pi_3\left(\int_0^t \eta\theta(s)\cdot(\bar{Q}^\eta(s) - \bar{C}^\eta(s))^+ds\right)$$
$$C^\eta(t) = C^\eta(0) - \Pi_5\left(\int_0^t \eta\ \gamma(s)\cdot(\bar{C}^\eta(s) - \bar{Q}^\eta(s))^+ds\right) + \Pi_4\left(\eta\int_0^t \beta(s)\cdot\bar{S}^\eta(s)ds\right)$$

where

$$\bar{Q}^\eta(t) = \frac{1}{\eta}Q^\eta(t), \quad \bar{C}^\eta(t) = \frac{1}{\eta}C^\eta(t), \quad \bar{S}^\eta(t) = \frac{1}{\eta}S^\eta(t).$$

Let $\mathcal{D}\left([0,\infty), \mathbb{R}^2\right)$ be the space of right continuous functions with left limits in $\mathbb{R}^2$ having the domain $[0,\infty)$. We give the space $\mathcal{D}\left([0,\infty), \mathbb{R}^2\right)$ the standard Skorokhod J_1 topology. Suppose $\{X^\eta\}_{\eta=1}^\infty$ is a sequence of stochastic processes, then $X^\eta \Rightarrow x$ means that X^η converges weakly to the stochastic process x.

Definition 1. *If there exists a limit in distribution for the scaled processes* $\{\bar{Q}^\eta\}_{\eta=1}^\infty$ *and* $\{\bar{C}^\eta\}_{\eta=1}^\infty$ *i.e.* $\bar{Q}^\eta(t) \Rightarrow q(t)$ *and* $\bar{C}^\eta(t) \Rightarrow c(t)$*, then* $(q(t), c(t))$ *is called the fluid limit for the original stochastic model.*

Proposition 1. *The sequence of scaled stochastic processes* $(\bar{Q}^\eta, \bar{C}^\eta)$ *are relatively compact and all weak limits are almost surely continuous.*

Proof. In order to show that $(\bar{Q}^\eta, \bar{C}^\eta)$ *is relatively compact with continuous limits, it is sufficient by Theorem 10.2 of* [6] *to show that the stochastic processes satisfy the following two conditions.*

1. *Compact Containment: for any $T \geq 0$, $\epsilon > 0$, there exists a compact set $\Gamma_T \subset \mathbb{R}^2$ such that*

$$\lim_{\eta \to \infty} \mathbb{P}\left(\left(\bar{Q}^\eta, \bar{C}^\eta\right) \in \Gamma_T, t \in [0,T]\right) \to 1, \tag{4}$$

2. *Oscillation Bound: for any $\epsilon > 0$, and $T \geq 0$ there exists a $\delta > 0$ such that*

$$\limsup_{\eta \to \infty} \mathbb{P}\left(\omega\left(\left(\bar{Q}^\eta, \bar{C}^\eta\right), \delta, T\right) \geq \epsilon\right) \leq \epsilon, \tag{5}$$

where

$$\omega(\boldsymbol{x}, \delta, T) := \sup_{s,t \in [0,T], |s-t| < \delta} \max_j |x_j(s) - x_j(t)|, \tag{6}$$

The proof of compact containment can be shown easily since there are no initial customers in the queue. Even if there were initial customers in the system, we can still bound the initial customers by a constant. In the case where there are no initial customers in the system, we can bound the queue length process by the arrival process. By defining the following quantity

$$\bar{\lambda} = \sup_{t \in [0,T]} \lambda(t) \tag{7}$$

it is trivial to show using the Law of Large numbers for Poisson processes that

$$\Gamma_T = \left\{(q,c) \middle| \; q + c \leq q(0) + \bar{\lambda} \cdot T + C_{max}\right\} \tag{8}$$

that the compact containment condition holds. Now it remains to prove the oscillation bound for the queueing process. First we bound the difference of the queue length process

$$\begin{aligned} Q^\eta(t) - Q^\eta(u) \leq \Pi_1\left(\eta \cdot \int_u^t \lambda(s)ds\right) + \Pi_2\left(\int_u^t \mu(s) \cdot (Q^\eta(s) \wedge C^\eta(s))ds\right) \\ + \Pi_3\left(\int_u^t \theta(s) \cdot (Q^\eta(s) - C^\eta(s))^+ ds\right) \end{aligned}$$

Now we bound the difference of the process that keeps track of the number of servers that are not idling.

$$C^\eta(t) - C^\eta(u) \leq \Pi_4\left(\int_u^t \beta(s) \cdot S^\eta(s)ds\right) + \Pi_5\left(\int_u^t \gamma(s) \cdot (C^\eta(s) - Q^\eta(s))^+ ds\right)$$

From the compact containment property, we know that there exists a finite constant K^* such that

$$\mathbb{P}\left(\bar{Q}^\eta(s) + \bar{C}^\eta(s) \leq K^*, s \in [0,T]\right) \to 1, \text{as } \eta \to \infty. \tag{9}$$

Thus on the event $\Omega_\eta = \{\bar{Q}^\eta(s) + \bar{C}^\eta(s) \leq K^*, s \in [0,T]\}$, then we have the subsequent inequalities for the rate functions for all $u, t \in [0,T]$ where $|t-u| \leq \delta$.

$$\int_u^t \lambda(s)ds \le \bar{\lambda} \cdot \delta =: c_1(\delta)$$

$$\int_u^t \mu(s) \cdot (\bar{Q}^\eta(s) \wedge \bar{C}^\eta(s))ds \le (\mu K^*) \cdot \delta =: c_2(\delta)$$

$$\int_u^t \theta(s) \cdot (\bar{Q}^\eta(s) - \bar{C}^\eta(s))^+ \le (\theta K^*) \cdot \delta =: c_3(\delta)$$

$$\int_u^t \beta(s) \cdot ((\bar{Q}^\eta(s) - \bar{C}^\eta(s))^+ \wedge (\bar{C}^\eta_{max}(s) - \bar{C}^\eta(s))ds \le (\beta K^*) \cdot \delta =: c_4(\delta)$$

$$\int_u^t \gamma(s) \cdot (\bar{C}^\eta(s) - \bar{Q}^\eta(s))^+ ds \le (\gamma K^*) \cdot \delta =: c_5(\delta)$$

Now by using the above inequalities, the Law of Large numbers for Poisson processes, and the continuity of the moduli of continuity function, the oscillation bound holds with

$$\delta = \frac{\epsilon}{\bar{\lambda} + (\mu + \theta + \beta + \gamma) \cdot K^*}. \tag{10}$$

Now the proof is complete.

Theorem 1. *If we are given determinsitic values* $(q(0), c(0))$ *and we assume that* $(\bar{Q}^\eta(0), \bar{C}^\eta(0)) \Rightarrow (q(0), c(0))$ *as* $\eta \to \infty$*, then the fluid limit*

$$\lim_{\eta\to\infty} \frac{1}{\eta} Q^\eta(t) \Rightarrow q(t) \quad \text{and} \quad \lim_{\eta\to\infty} \frac{1}{\eta} C^\eta(t) \Rightarrow c(t)$$

of the original stochastic queueing model is the unique solution to the following system of ordinary differential equations

$$\begin{aligned} \frac{d}{dt} q(t) &= \lambda(t) - \mu(t) \cdot (q(t) \wedge c(t)) - \theta(t) \cdot (q(t) - c(t))^+ \\ \frac{d}{dt} c(t) &= \beta(t) \cdot (q(t) - c(t))^+ \wedge (c_{max}(t) - c(t)) - \gamma(t) \cdot (c(t) - q(t))^+. \end{aligned} \tag{11}$$

Proof. Now that we know that the queueing and server non-idle processes are relatively compact, we can now use this result to prove the fluid limit theorem. Since $\left(\bar{Q}^\eta(\cdot), \bar{C}^\eta(\cdot)\right)$ *is relatively compact, we know that that given any subsequence* $\left(\bar{Q}^{\eta_m}(\cdot), \bar{C}^{\eta_m}(\cdot)\right)$ *we can construct another subsequence* $\left(\bar{Q}^{\eta_{m_l}}(\cdot), \bar{C}^{\eta_{m_l}}(\cdot)\right)$ *that converges weakly in* $\mathbb{D}([0,\infty), \mathbb{R}^2)$*, to a continuous process* $(q^*(\cdot), c^*(\cdot))$*. Thus, we know that* $v^*(\cdot)$ *is at least one limit of the original stochastic process sequence* $\left(\bar{Q}^\eta(\cdot), \bar{C}^\eta(\cdot)\right)$*. Therefore, if we can prove that* $(q^*(\cdot), c^*(\cdot))$ *satisfies the fluid limit Eq.* (11) *and the fluid limit equations have a unique solution, then by the arbitrariness of the limit* $v^*(\cdot)$*, there exists unique fluid limit that is given by the equations of* (11)*. From the representation of* $\left(\bar{Q}^\eta(t), \bar{C}^\eta(t)\right)$ *we have that*

$$\bar{Q}^\eta(t) = \bar{Q}^\eta(0) + M_Q^\eta\left(\bar{Q}^\eta(t), \bar{C}^\eta(t)\right) + \int_0^t A_Q^\eta\left(\bar{Q}^\eta(u), \bar{C}^\eta(u)\right) du$$

$$\bar{C}^\eta(t) = \bar{C}^\eta(0) + M_C^\eta\left(\bar{Q}^\eta(t), \bar{C}^\eta(t)\right) + \int_0^t A_C^\eta\left(\bar{Q}^\eta(u), \bar{C}^\eta(u)\right) du$$

where

$$
\begin{aligned}
&M_Q^\eta\left(\bar{Q}^\eta(t),\bar{C}^\eta(t)\right) = \\
&\left(\frac{1}{\eta}\cdot \Pi_1\left(\eta\cdot\int_0^t \lambda(s)ds\right) - \int_0^t \lambda(s)ds\right) \\
&-\frac{1}{\eta}\cdot\Pi_2\left(\int_0^t \mu(s)\cdot(Q^\eta(s)\wedge C^\eta(s))ds\right) + \int_0^t \mu(s)\cdot(\bar{Q}^\eta(s)\wedge \bar{C}^\eta(s))ds \\
&-\frac{1}{\eta}\cdot\Pi_3\left(\int_0^t \theta(s)\cdot(Q^\eta(s)-C^\eta(s))^+ds\right) + \int_0^t \theta(s)\cdot(\bar{Q}^\eta(s)-\bar{C}^\eta(s))^+ds
\end{aligned}
$$

$$
\begin{aligned}
&M_C^\eta\left(\bar{Q}^\eta(t),\bar{C}^\eta(t)\right) = \\
&\frac{1}{\eta}\cdot\Pi_4\left(\int_0^t \beta(s)\cdot S^\eta(s)ds\right) - \int_0^t \beta(s)\cdot S^\eta(s)ds \\
&-\frac{1}{\eta}\cdot\Pi_5\left(\int_0^t \gamma(s)\cdot(C^\eta(s)-Q^\eta(s))^+ds\right) + \int_0^t \gamma(s)\cdot(C^\eta(s)-Q^\eta(s))^+ds
\end{aligned}
$$

and

$$
\begin{aligned}
\int_0^t A_Q^\eta\left(\bar{Q}^\eta(u),\bar{C}^\eta(u)\right)du &= \int_0^t \lambda(u)du - \int_0^t \mu(u)\cdot(\bar{Q}^\eta(u)\wedge\bar{C}^\eta(u))du \\
&\quad - \int_0^t \theta(u)\cdot(\bar{Q}^\eta(u)-\bar{C}^\eta(u))^+du \\
\int_0^t A_C^\eta\left(\bar{Q}^\eta(u),\bar{C}^\eta(u)\right)du &= \int_0^t \beta(u)\cdot\bar{S}^\eta(u)du - \int_0^t \gamma(u)\cdot(\bar{C}^\eta(u)-\bar{Q}^\eta(u))^+du
\end{aligned}
$$

Since we know that $\bar{\mathbf{V}}^{\eta_m}(\cdot) = (\bar{Q}^{\eta_m}(\cdot),\bar{C}^{\eta_m}(\cdot)) \overset{d}{\Rightarrow} v^*(\cdot) = (q^*(\cdot),c^*(\cdot))$ *and that* $v^*(\cdot)$ *is continuous, then we have that*

$$
\begin{aligned}
\bar{Q}^{\eta_m}(\cdot) - \bar{Q}^{\eta_m}(0) - \int_0^\cdot A_Q(\bar{Q}^{\eta_m})ds &\overset{d}{\Rightarrow} q^*(\cdot) - q(0) - \int_0^\cdot A_Q(q^*(s))ds \\
\bar{C}^{\eta_m}(\cdot) - \bar{C}^{\eta_m}(0) - \int_0^\cdot A_C(\bar{C}^{\eta_m})ds &\overset{d}{\Rightarrow} c^*(\cdot) - c(0) - \int_0^\cdot A_C(c^*(s))ds.
\end{aligned}
$$

Thus, if we can show that

$$
\lim_{m\to\infty} \mathbf{M}^{\eta_m}(\cdot) \equiv \lim_{m\to\infty} (M_Q^{\eta_m}(\cdot), M_C^{\eta_m}(\cdot)) = 0, \tag{12}
$$

then we have that all of the limits satisfy the fluid limit Eq. (11) *and since the functional* $A(\cdot)$ *is Lipschitz continuous, the fluid equations have a unique solution. This implies that all of the fluid limits are the same and are all equal to the solution of the fluid limit Eq.* (11)*. Now it remains to prove that*

$$
\lim_{m\to\infty} \mathbf{M}^{\eta_m}(\cdot) = 0. \tag{13}
$$

Using the law of large numbers for Poisson processes, we know that

$$\lim_{\eta\to\infty} \mathbf{Y}(\eta\,\cdot)/\eta - \cdot \stackrel{d}{\Rightarrow} 0 \quad \text{in } \mathcal{D}([0,\infty),\mathbb{R}). \tag{14}$$

Moreover, since we have that $\bar{\mathbf{V}}^{\eta_m}(\cdot) \stackrel{d}{\Rightarrow} v^*(\cdot)$ *as* $m \to \infty$ *and we know that the limit* $v^*(\cdot)$ *is continuous, then we have that*

$$\begin{aligned}
\lim_{\eta\to\infty} \int_0^{\cdot} \mu(s)\cdot(\bar{Q}^\eta(s)\wedge\bar{C}^\eta(s))ds &\stackrel{d}{\Rightarrow} \int_0^{\cdot} \mu(s)\cdot(q^*(s)\wedge c^*(s))ds\\
\lim_{\eta\to\infty} \int_0^{\cdot} \theta(s)\cdot(\bar{Q}^\eta(s)-\bar{C}^\eta(s))^+ds &\stackrel{d}{\Rightarrow} \int_0^{\cdot} \theta(s)\cdot(q^*(s)-c^*(s))^+ds\\
\lim_{\eta\to\infty} \int_0^{\cdot} \beta(s)\cdot\bar{S}^\eta(s)ds &\stackrel{d}{\Rightarrow} \int_0^{\cdot} \beta(s)\cdot(q^*(s)-c^*(s))^+\wedge(c_{max}(s)-c^*(s)))ds\\
\lim_{\eta\to\infty} \int_0^{\cdot} \gamma(s)\cdot(\bar{C}^\eta(s)-\bar{Q}^\eta(s))^+ds &\stackrel{d}{\Rightarrow} \int_0^{\cdot} \gamma(s)\cdot(c^*(s)-q^*(s))^+ds.
\end{aligned}$$

Now by the random time change Theorem of [4]*, we have that*

$$\lim_{\eta\to\infty} Y_{il}^S\left(\eta_m\int_0^{\cdot} f(s,\bar{\mathbf{V}}^{\eta_m}(s))ds\right)/\eta - \int_0^{\cdot} f(s,\bar{\mathbf{V}}^{\eta_m}(s)) \stackrel{d}{\Rightarrow} 0$$

and this completes the proof for the fluid limit since the other terms of $\mathbf{M}^\eta(\cdot)$ *can also be shown to converge to 0.*

4 Performance Measures and Numerics

In this section, we compare our limit theorems with a discrete event simulation of the delay-off queueing process. We show that the fluid limit is quite accurate at approximating the mean dynamics of the queueing process.

4.1 Mean Queue Length and Mean Non-Idle Servers

Our first comparison between simulation and our fluid limits is given on the left of Fig. 1. In this example, the turn off rate is of moderate size meaning that it is not too high or too low. On the left of Fig. 1, we see that the simulated mean queue length is well approximated by the fluid limit and we also see similar accuracy for the mean number of idle servers. Our second comparison is given on the right of Fig. 1 and in this example the turn off rate is high. This situation is closest to when the servers are immediately shut off when they become idle. Once again on the right of Fig. 1, we see that the simulated mean queue length and the mean number of idle servers are well approximated by the mean field approximation or the fluid limit.

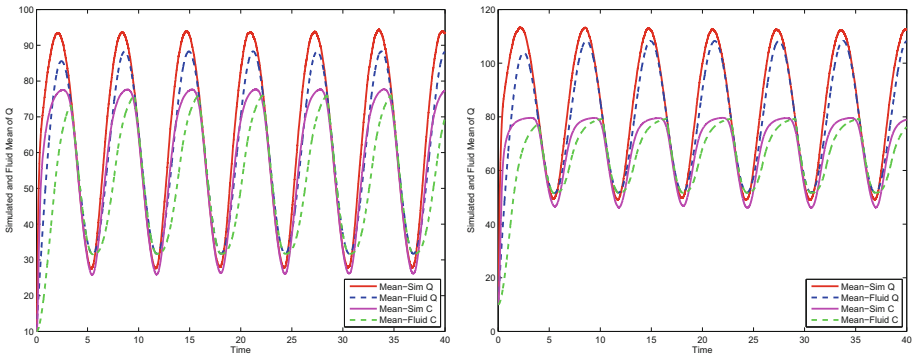

Fig. 1. $\lambda(t) = 80 + 20 \cdot \sin t$, $\mu = 1$, $\theta = 1$, $\beta = 1$, $\gamma = 1$ $(Left)$, $\gamma = 1000$ $(Right)$, $C_{max} = 80$. (Mean Queue Length and Mean Non-Idle Servers).

4.2 Energy Consumption

In addition to understanding how well our limit theorem approximates the actual stochastic system, it is important to analyze the power consumption of the system in a variety of parameter settings.

The mean energy consumption in the nonstationary setting is now given by

$$E[ActEne(t)] = \int_0^t E[C(u)] \times c_1(u)du,$$

where $c_1(t)$ is the energy cost for an active or idle server at time t. For simplicity we may consider the simple case where $c_1(t) = c_1$.

Furthermore, let $S(t)$ denote the number of servers in setup at time t and let $c_2(t)$ denote the energy cost for a server in setup mode at time t, then the energy consumption by servers in setup mode is given by

$$E[SetEne(t)] = \int_0^t E[S(u)] \times c_2(u)du.$$

By considering a simple case where $c_2(t) = c_2$ and since also in practice, it is empirically seen that $c_2 = c_1$. Thus, in the numerical examples, we consider the case $c_1 = c_2 = 1$. The overall energy consumption in the time interval $[0, t]$ is given by

$$E[TotalEne(t)] = E[ActEne(t)] + E[SetEne(t)].$$

We would like to minimize the above total energy consumption. On the other hand, we also would like to minimize the mean waiting cost which is calculated based on the queue length, i.e., $\int_0^t Q(u)du$. Thus, we need to consider a cost function which is a combination of the power consumption and the waiting cost.

In Figs. 2 and 3, we plot a convex combination of the power used and the queue length or delay of the system integrated over time. These plots represent the trade-off between delays experienced by customers and the power cost of the

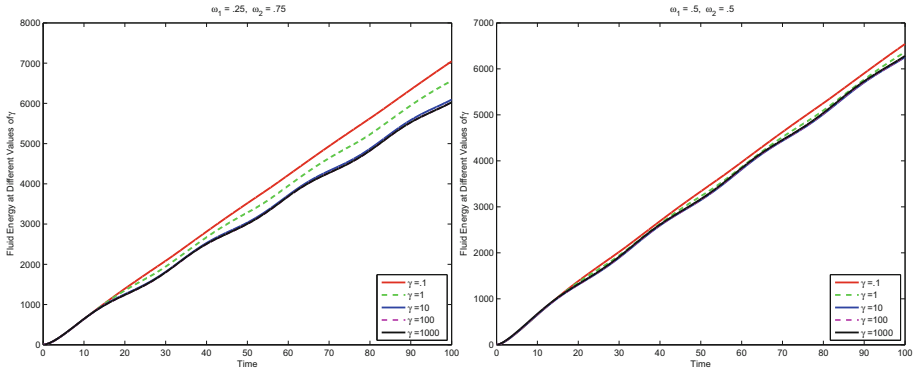

Fig. 2. $\lambda(t) = 60 + 20 \cdot \sin t$, $\mu = 1$, $\theta = 1$, $\beta = 1$, $C_{max} = 100$. Fluid Energy as γ varies and $\omega_1 = .25, \omega_2 = .75$ (Left). Fluid Energy as γ varies and $\omega_1 = .5, \omega_2 = .5$ (Right).

data center. On the left of Fig. 2, we weight the delay by $\omega_1 = .25$ and weight the power by $\omega_2 = .75$. We see that as we increase γ the total power and delay cost decreases. This is partially because we are weighting the power as more costly in this example. On the right of Fig. 2, we weight the delay by $\omega_1 = .5$ and weight the power by $\omega_2 = .5$. We see that as we increase γ the total power and delay cost decreases, but only slightly since the weighting is equal. However, we see that the power is a bit more influential on the cost, but very slight. On the left of Fig. 3, we weight the delay by $\omega_1 = .75$ and weight the power by $\omega_2 = .25$. We see that as we increase γ the total power and delay cost increases. This is partially because we are weighting the delay as more costly in this example.

On the right of Fig. 3, we plot the power consumption as we vary the parameter γ. We see that we increase γ, the power consumption goes down, especially

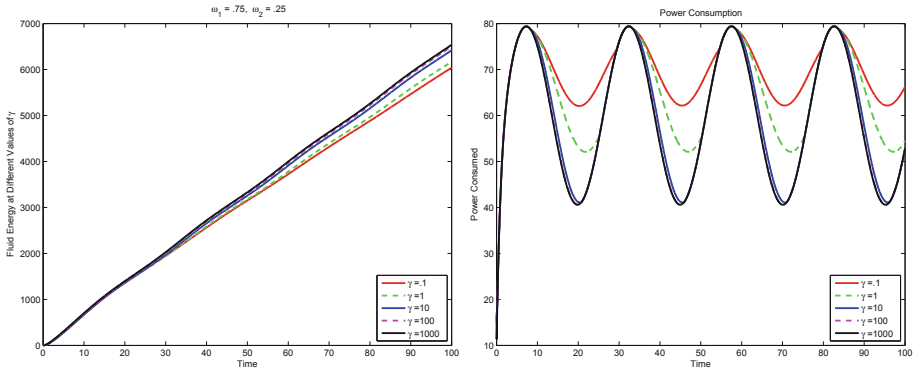

Fig. 3. $\lambda(t) = 60 + 20 \cdot \sin t$, $\mu = 1$, $\theta = 1$, $\beta = 1$, $C_{max} = 100$. Fluid Energy as γ varies and $\omega_1 = .75, \omega_2 = .25$. (Left) $\lambda(t) = 60 + 20 \cdot \sin t$, $\mu = 1$, $\theta = 1$, $\beta = 1$, $C_{max} = 100$. Power Consumption as γ varies. (Right)

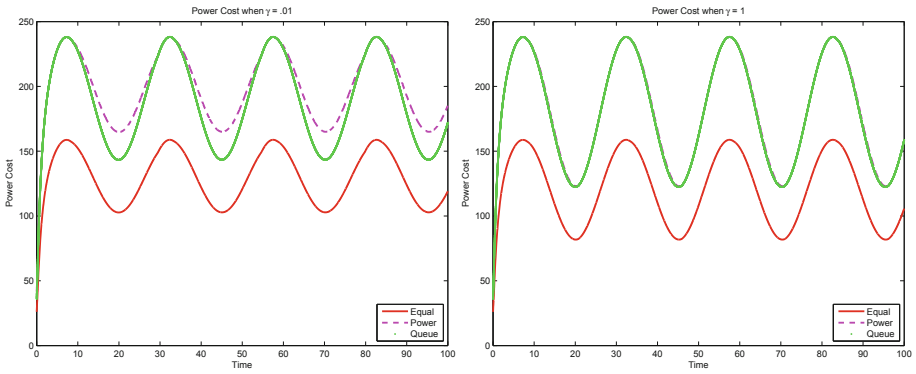

Fig. 4. $\lambda(t) = 60 + 20 \cdot \sin t$, $\mu = 1$, $\theta = 1$, $\beta = 1$, $C_{max} = 100$. Power Cost when $\gamma = .01$ (Left) Power Cost when $\gamma = 1$ (Right)

when the queue moves from the overloaded to underloaded regime. However, in the overloaded setting, the parameter γ does not do much in terms of saving power consumption.

In Fig. 4, we plot the trade-off between delays experienced by customers (in terms of $Q(t)$) and the power cost of the data center over time. In these figures, we have three different scenarios. The first corresponds to 'Equal' where the power and queue length are weighted equally. The second corresponds to power where the power cost is multiplied by a factor of 2. Lastly, the third scenario corresponds to the queue length being multiplied by 2. On the left of Fig. 4, we see that power is very important and therefore the plot that weights power more is higher than the plot that weights delay more. However, as γ gets larger on the right of Fig. 4, this difference between the two plots disappears and is negligible.

5 Conclusion and Final Remarks

In this paper, we analyze multi-server setup queues with non-stationary arrivals and abandonment. We show that a heuristic mean field limit can be made rigorous by scaling the number of arrivals and servers to infinity. This is an appropriate regime since the amount of data traffic is large and the number of servers in most data centers is also large. We show that we are able to capture the salient features of the queueing model with our weak law of large numbers limit.

There are many extensions of this work that are worth pursuing. One extension that is important is to generalize the arrival and service times of the data. Currently there is no empirical evidence to support that the inter-arrival and service times are exponential random variables. One way to generalize these results would be to use Markovian Arrival Processes like in the work of [16,26]. Moreover, refining the approximations using orthogonal polynomial methods like in the work of [22–25,27] is also an important area of study. We hope to consider these generalizations in future work. Moreover, we would like to incorporate the

energy impact of using renewable energy such as wind and solar. This would involve additional stochastic models for understanding the mix and cost of the energy being provided to the data center. In the case when wind and solar energy are used, it may be cost effective to keep the servers on even though servers are not needed since the energy used is cheaper. We plan to pursue this extension as well. Lastly, we are interested in optimal control methods for these delay-off systems. In this context, we can use the work of [13,14,21] to find optimal turn off or on policies for our delay-off model. We also plan to complete this work in a follow up paper.

Acknowledgments. Tuan Phung-Duc was supported in part by JSPS KAKENHI Grant Number 2673001. The authors would like to thank the four referees for their constructive comments which improve the presentation of the paper.

References

1. Akoush, S., Sohan, R., Rice, A., Moore, A.W., Hopper, A.: Free lunch: exploiting renewable energy for computing. In: Proceedings of HotOS, p. 17 (2011)
2. Artalejo, J.R., Economou, A., Lopez-Herrero, M.J.: Analysis of a multiserver queue with setup times. Queueing Syst. **51**(1–2), 53–76 (2005)
3. Barroso, L.A., Hölzle, U.: The case for energy-proportional computing. Computer **12**, 33–37 (2007)
4. Billingsley, P.: Convergence of Probability Measures, vol. 493. Wiley, London (2009)
5. Engblom, S., Pender, J.: Approximations for the moments of nonstationary and state dependent birth-death queues. Submitted to Stochastic Systems (2014)
6. Ethier, S.N., Kurtz, T.G.: Markov Processes: Characterization and Convergence, vol. 282. Wiley, London (2009)
7. Gandhi, A., Doroudi, S., Harchol-Balter, M., Scheller-Wolf, A.: Exact analysis of the m/m/k/setup class of Markov chains via recursive renewal reward. SIGMETRICS Perform. Eval. Rev. **41**(1), 153–166 (2013)
8. Gandhi, A., Doroudi, S., Harchol-Balter, M., Scheller-Wolf, A.: Exact analysis of the m/m/k/setup class of Markov chains via recursive renewal reward. Queueing Syst. **77**(2), 177–209 (2014)
9. Gandhi, A., Harchol-Balter, M., Adan, I.: Server farms with setup costs. Perform. Eval. **67**(11), 1123–1138 (2010)
10. Gao, V., Zeng, Z., Liu, X., Kumar, P.R.: The answer is blowing in the wind: analysis of powering internet datacenters with wind energy. In: IEEE 2013 Proceedings of INFOCOM, pp. 520–524. IEEE (2013)
11. Goiri, Í., Haque, M.E., Le, K., Beauchea, R., Nguyen, T.D., Guitart, J., Torres, J., Bianchini, R.: Matching renewable energy supply and demand in green datacenters. Ad Hoc Netw. **25**, 520–534 (2015)
12. Grier, N., Massey, W.A., McKoy, T., Whitt, W.: The time-dependent erlang loss model with retrials. Telecommun. Syst. **7**(1–3), 253–265 (1997)
13. Hampshire, R.C., Massey, W.A.: Variational optimization for call center staffing. In: Proceedings of the 2005 Conference on Diversity in Computing, pp. 4–6. ACM (2005)

14. Hampshire, R.C., Massey, W.A.: Dynamic optimization with applications to dynamic rate queues. TUTORIALS in Operations Research, INFORMS Society, pp. 210–247 (2010)
15. Kallenberg, O.: Foundations of Modern Probability. Springer Science & Business Media, New York (2006)
16. Ko, Y.M., Pender, J.: Strong approximations for time varying infinite-server queues with non-renewal arrival and service processes
17. Kurtz, T.G.: Strong approximation theorems for density dependent Markov chains. Stoch. Process. Appl. **6**(3), 223–240 (1978)
18. Mandelbaum, A., Massey, W.A., Reiman, M.I.: Strong approximations for Markovian service networks. Queueing Syst. **30**(1–2), 149–201 (1998)
19. Massey, W., Pender, J.: Skewness variance approximation for dynamic rate multiserver queues with abandonment. Perform. Eval. Rev. **39**, 74 (2011)
20. Massey, W., Pender, J.: Gaussian skewness approximation for dynamic rate multiserver queues with abandonment. Queueing Syst. **75**(2), 243–277 (2013)
21. Niyirora, J., Pender, J.: Optimal staffing of clinical revenue centers in health care organizations (2015, under review)
22. Pender, J.: Gram Charlier expansion for time varying multiserver queues with abandonment. SIAM J. Appl. Math. **74**(4), 1238–1265 (2014)
23. Pender, J.: A Poisson-Charlier approximation for nonstationary queues. Oper. Res. Lett. **42**(4), 293–298 (2014)
24. Pender, J.: An analysis of nonstationary coupled queues. Telecommun. Syst. **61**, 823–838 (2016)
25. Pender, J.: Nonstationary loss queues via cumulant moment approximations. Probab. Eng. Inf. Sci. **29**(01), 27–49 (2015)
26. Pender, J., Ko, Y.M.: Approximations for the queue length distributions of time-varying many-server queues
27. Pender, J.: Laguerre polynomial expansions for time varying multiserver queueswith abandonment. Working paper (2014). http://www.columbia.edu/~jp3404/LSA.html
28. Phung-Duc, T.: Exact solutions for m/m/c/setup queues. Telecommunication Systems (2016). doi:10.1007/s11235-016-0177-z