Avoiding Overages by Deferred Aggregate Demand for PEV Charging on the Smart Grid

G. Pang and G. Kesidis
IE, CSE and EE Depts
Penn State University
University Park, PA
{gup3, gik2}@psu.edu

T. Konstantopoulos
Dept of Maths
Uppsala University
Uppsala, Sweden
takiskonst@gmail.com

Abstract—In this paper, we model the aggregate overnight demand for electricity by a large community of (possibly hybrid) plug-in electric vehicles (PEVs) each of whose power demand follows a prescribed profile and is interruptible. The community is serviced by a regional electrical utility which is assumed to purchase electricity from a state/national distribution grid according to a flat-rate $\phi$ per kilowatt-unit-time up to a threshold $\pi$, and thereafter overage charges $\pi > \phi$ are levied per kilowatt-unit-time. Rather than a spot-price system for household consumers (which would necessarily need to be operated by automated means overnight when most people sleep), the “grid” (regional utility) is “smart” in that it monitors its total load and, when overages threaten, can reduce load by signaling certain consumers to defer their load by one unit of time. In this paper, we model undeferred load by a Gaussian process which we justify by means of a functional central limit theorem. This limiting Gaussian process is the arrival process of a discrete-time queue which is used to model the deferred load over a finite time-horizon. We can then compute the mean amount of overage at the end of this time horizon (say at 6 AM when charging is to be completed ahead of the morning commute).

I. INTRODUCTION

Much activity has recently focused on highly dynamic economic models for electricity distribution systems based on precise and ubiquitous load and network metering [12], [16]. Metering is intended to inform a deregulated spot-price system which is in turn intended to incentivize the efficiencies found in free markets (e.g., [23], [11], [4]), in particular through demand-side management. Such consumer load management is even more crucial as plug-in electric vehicles (PEVs), including plug-in hybrid electric vehicles (PHEVs), are coming to the market [18], [13]. With battery capacities varying from 15 to 50 kWh, these vehicles are expected to double the average household load during charging time [9].

The design of appropriate incentives and efficient energy-consumption scheduling algorithms (e.g., [15], [1], [13]) is a primary goal for the smart grid, i.e., incentivizing consumers to offload demand to off-peak periods, also known as “peak shaving,” see, e.g., [25].

Basically, there are two frameworks, one in which the grid signals the consumers with prices and lets them decide, and another where the grid controls at least some instrumented outlets in the consumers’ houses and can cycle power to them to shave peak demand. As we are interested in automated decisions by both the grid and the consumers (as would be the case in the evening while electric vehicles are charging and consumers are sleeping), we herein assume that the grid is capable of deferring load. We assume that users who both submit to control by the grid of their primary evening power load (plug-in electric vehicle) and choose a random start time of their load (so as to desynchronize demand and reduce the “load” on the deferment mechanism) will be charged discounted overages if they do occur. By simply monitoring the start time of a user’s evening power consumption over several evenings, the grid can perform a straightforward statistical test to reckon the latter.

Several authors have considered “smart” users of a spot-price system, e.g., [13], based on the Nash certainty equivalence principle for the LQG framework [6], or see, e.g., [14] for smart users of a queueing system as in Section IV below. We instead assume that the regional utility is “smart” in its ability to dynamically defer demand to avoid overages, and to steady and reduce (shave) peak total power demand.

The focus of this paper is on automated management by a regional utility of its collective PEV load, a load which will likely be borne in the earlier morning hours while most consumers sleep. We assume that the utility’s ability to defer load would be overwhelmed if all users simply started charging at, say, midnight. Thus, to avoid spikes in demand due to possibly synchronized users, particularly at the start of charging period, we assume that, night after night, the users independently and uniformly at random choose initial start times in the evening charging intervals, either voluntarily or by request of the utility, cf. Section II-B. Given this, there may still be moderate overage periods which the utility will need to manage by deferring demand. By using smart meters, the regional utility can obtain an empirical start-time distribution for each user to police their compliance with a randomly chosen start-time; users who are deemed in compliance can be assigned a reduced share of any overage charges incurred.

Given a characterization of aggregate load by a functional central limit theorem enabled by this assumption of asynchronous users, a regional utility can seek to minimize its own objective function which is a weighted combination of the probability of overage (social welfare) and the contractual cost of their supply from a broader electrical distribution system.

The rest of this paper is organized as follows. We motivate our model of consumer demand without deferment in
In Section III, we use a functional central limit theorem (FCLT) to derive a simple diffusion model of the unattended load on the utility. In Section IV, we summarize our discrete-time load-deferment framework, potentially resulting in overages at the end of the deferment interval, and describe a numerical study under idealized load deferment. After a discussion on how to choose which users to defer in Section V, we conclude in Section VI with a discussion of future work. The proofs of the theorems given below can be found in the appendices.

II. PROBLEM SET-UP

We consider a $T$-hour time period, e.g., $T = 6$ hours from midnight to 6AM, during which $n$ customers need to automatically schedule their electrical demand. Note that our time horizon is finite since users do not want their jobs to be indefinitely delayed.

A. Demand

We consider two broad types of power-consumption profiles for each user: constant and unimodal.

1) Constant-power demand: Constant power-consumption (charging rate) profiles may be approximately true for appliances such as clothes or dish washers that use pre-heated water or clothes dryers. The consumption profile for the $i^{th}$ such appliance/user is parameterized by $(\eta_i, H_i, \tau_i)$, where $\tau_i$ is the starting time of service request, $\eta_i$ is the service request duration and $H_i$ is the power charging rate during the service request duration. So, the power consumption profile from the $i^{th}$ user can be written simply as

$$g(t; \tau_i, \eta_i, H_i) = H_i 1(\tau_i \leq t \leq \tau_i + \eta_i).$$

The total energy demand $\beta_i$ for user $i$ is simply

$$\beta_i = H_i \eta_i. \tag{2}$$

b) Unimodal power demand: A unimodal power consumption profile is more typical of battery charging [20], [5]: when charging of an empty Lithium-Ion (Li-Ion) battery commences (say at time $0$), the current is constant and the voltage grows until a peak-power threshold at time $\zeta > 0$ is reached, whereafter the voltage is constant and the current diminishes. We will approximate this power charging rate by a triangle-pulse function, $h(t)$, starting at time $0$ with support $\eta > \zeta$ and peak value $H$ at time $\zeta$; see top graph of Figure 1. The $h(t)$ function has the following explicit expression

$$h(t; \eta, H, \zeta) = \begin{cases} \frac{\eta t}{\eta - \zeta}, & 0 \leq t < \zeta, \\ \frac{H}{\eta - \zeta} t + \frac{H \eta - \zeta}{\eta - \zeta}, & \zeta \leq t \leq \eta. \end{cases} \tag{3}$$

A given smart grid PEV customer/user at a given night may have a residual charge on their battery$^1$. So, their power demand profile will look like, e.g., the front-clipped (advanced) triangle pulse at the bottom of Figure 1 i.e.,

$$h(t + \xi; \eta, H, \zeta)u(t + \xi),$$

where $u$ is Heaviside’s unit-step function and $\xi > 0$ is the time corresponding to the amount of residual charge $\alpha = h(\xi)$. So, the PEV power consumption profile of the $i^{th}$ residential smart grid customer/user is here parameterized by $(\tau_i, \xi_i, \eta_i, H_i, \zeta_i)$, where again $\tau_i$ is the starting time of service request, i.e.,

$$g(t; \tau_i, \xi_i, \eta_i, H_i, \zeta_i) = h(t + \xi_i - \tau_i; \eta_i, H_i, \zeta_i)u(t + \xi_i - \tau_i).$$

Note that the factor $(\xi, \alpha := h(\xi))$ may alternatively (or additionally) reflect the presence and initial use of a limited local energy supply generated by, e.g., solar radiation or wind during the day and stored by the consumer for later use.

c) Random demand parameters: Additionally, assume that the parameters, $\eta_i, \zeta_i$ and $H_i$ are mutually independent random variables over customer index $i$. Also, assume that $\xi_i \in [0, \eta_i]$ is chosen independently given $\eta_i$. Let $m_H = EH$, $\sigma_H^2 = \text{var}(H)$, etc. For the case of unimodal power profile, note from the distribution of these parameters we can obtain the moments of the initial charge: $E\alpha = Eh(\xi) =: m_\alpha$ and $\text{var}(\alpha) = \text{var}(h(\xi)) =: \sigma_\alpha^2$.

Variations in the parameters $\eta_i, \zeta_i$ and $H_i$ may be naturally due to different types/models of batteries used in different hybrid electric and different purely electric vehicles, as well as variations due to manufacture and due to performance degradation with age, cf. the end of Section III-A. Assume that

$$EH^{2+\delta}, E\eta^{2+\delta} < \infty \text{ for some } \delta > 0.$$

So, the power consumption profiles for different customers are independent. Assume that the joint distributions of $(\xi_i, \eta_i, H_i, \zeta_i)$ are such that there are $K$ classes of profiles/flags with the $k^{th}$ class $(\xi^{(k)}_i, \eta^{(k)}_i, H^{(k)}_i, \zeta^{(k)}_i)$ occurring with probability $p_k$. The total energy demand for a user of class $k$ is simply

$$\beta^{(k)} = \int_0^\infty g(t; \tau, \xi^{(k)}_i, \eta^{(k)}_i, H^{(k)}_i, \zeta^{(k)}_i)dt, \tag{4}$$

1 Or users may have auxiliary on-site power generation (e.g., p. 3 of [21]) which they could use to charge their battery before engaging the grid.
which does not depend on the start time \( \tau \in [0,T] \), of course. So,

\[
\beta^{(k)} = \begin{cases} \\
\frac{1}{\pi} H^{(k)}(\eta^{(k)}) - \frac{1}{\pi} h^{(k)}(\xi^{(k)}) \xi^{(k)} & \text{if } \xi^{(k)} < \zeta^{(k)} \\
\frac{1}{\pi} \frac{(\eta^{(k)} - \xi^{(k)})^2 h^{(k)}(\eta^{(k)})}{\pi^2 - \zeta^{(k)}} & \text{else.}
\end{cases}
\]

That is, the total energy demand for any user is \( \beta = \beta^{(k)} \) with probability \( p_k \).

### B. Supply from the regional utility

First suppose that the utility employs a dynamic system wherein the spot prices are periodically communicated to the consumers’ computer (during the evening) so that a pre-programmed policy of “engagement” in the power grid can be employed, e.g., if the spot-price is below a threshold, then charge the battery during the corresponding period of time. If all consumers follow such policies, then potentially dramatic swings in demand may ensue. To mitigate such swings, the utility could also signal the consumers of an overload situation which will result in overage charges notwithstanding the low spot price just announced. Consumers may randomly delay engagement times to charge their batteries on with the grid in order to avoid synchronization of their demand, particularly if an overload is signaled.

Again, in the following, a highly dynamic spot-price system is assumed not in play. Rather, we assume that the utility has a good estimate of the statistics of aggregate consumer demand and this knowledge has informed a contract with the distribution system: a flat rate \( \phi \) per kilowatt-unit-time up to a threshold \( L \), after which an overcharge \( \pi > \phi \) per kilowatt-unit-time is levied. Also, we assume that consumers select a start-time for their demand chosen uniformly at random from \( \{0, T_o, 2T_o, ..., T_o \} \) where \( T_o \leq T - \eta_{\max} \), \( T_o \ll \eta_{\max} \), and \( \eta_{\max} \) is the maximum service request duration. The goal of this strategy is to “even out” demand, cf. (next) Section III.

### III. UNDEFERRED LOAD

As motivated by the previous subsection, we now assume that the start time of demand for each user \( i \), \( \tau_i \), is chosen independently and uniformly at random from the continuous interval of time \([0, T_o]\) (for simplicity in this section). Assuming the number of consumers \( n \gg 1 \) is given, conditional uniformity, the arrival epochs \( \{ \tau_i : i \geq 1 \} \) of service requests can be regarded as the arrival times of a Poisson process \( \{ N(t) : t \geq 0 \} \). Let \( \lambda = n / T_o \) be its arrival rate.

The case of a constant consumption profile is covered in Appendix A.

---

2The risk of overload as a function of announced price may be learned and factored into the price that is advertised to the users.
3This tactic is not new: An intention of the proposed random early discard (RED) mechanism operating at an Internet packet-queue is to desynchronize the distributed TCP sessions potentially congesting it; see, e.g., [17].
4A good strategy for a “no information” scenario too [1].

---

### A. Unimodal consumption profile

Let \( \lambda_k = p_k \lambda \) be the arrival rate of the class \( k \) users and \( \{ N^{(k)}(t) : t \geq 0 \} \) as the corresponding Poisson process with arrival epochs \( \{ \tau^{(k)}_i : i \geq 1 \} \). So, \( N(t) = \sum_{k=1}^{K} N^{(k)}(t) \) is a Poisson process with rate \( \lambda \).

When the consumption profiles are unimodal, the total instantaneous power charging rate for class \( k \) users is given by the flag process

\[
D^{(k)}(t) = \sum_{i=1}^{N^{(k)}(t)} g(t; \tau^{(k)}_i, \xi^{(k)}_i, \eta^{(k)}_i, H^{(k)}_i, \zeta^{(k)}_i)
\]

and the total instantaneous power charging rate for all users is given by \( D(t) = \sum_{k=1}^{K} D^{(k)}(t) \) for \( t \geq 0 \). The energy consumption for all class \( k \) users is given by the integrated (average) flag process \( \{ C^{(k)}(t) : t \geq 0 \} \)

\[
C^{(k)}(t) = \int_0^t D^{(k)}(s) ds
\]

and the energy consumption for all users is given by \( \{ C(t) : t \geq 0 \} \), where \( C(t) = \sum_{k=1}^{K} C^{(k)}(t) \).

We take an asymptotic approach by scaling the processes \( \{ D^{(k)}(t) : t \geq 0 \} \) and \( \{ C^{(k)}(t) : t \geq 0 \} \) by the total arrival rate \( \lambda \) and letting \( \lambda \rightarrow \infty \). Assume that there exist \( \lambda_k \in (0,1) \) such that \( \sum_{k=1}^{K} \lambda_k = 1 \) and \( \lambda_k / \lambda \rightarrow \lambda_k \) as \( \lambda \rightarrow \infty \) for each \( k \in \{1, ..., K\} \). We index by \( \lambda \) the relevant processes: \( N^{(k)}(t), N^{(k)}(t), D^{(k)}(t), D^{(k)}(t), C^{(k)}(t) \) and \( C^{(k)}(t) \).

Define the diffusion-scaled processes \( \{ \hat{C}^{(k)}(t) : t \geq 0 \} \), for \( k \in \{1, ..., K\} \), and \( \{ \hat{C}^{(k)}(t) : t \geq 0 \} \) by

\[
\hat{C}^{(k)}(t) = \frac{1}{\sqrt{\lambda}} \left( C^{(k)}(t) - \mu^{(k)} t \right),
\]

\[
\hat{C}^{(k)}(t) = \sum_{k=1}^{K} \hat{C}^{(k)}(t) = \frac{1}{\sqrt{\lambda}} \left( C^{(k)}(t) - \mu^{(k)} t \right),
\]

where

\[
\mu^{(k)} = \lambda_k E[\beta^{(k)}_i], \quad \mu = \sum_{k=1}^{K} \mu^{(k)} = E[\beta^{(k)}_i].
\]

**Theorem 1**: Under the assumption of unimodal consumption profiles, as \( \lambda \rightarrow \infty \)

\[
(\hat{C}^{(1)}_{\lambda}, ..., \hat{C}^{(K)}_{\lambda}, \hat{C}_{\lambda}) \Rightarrow (\sigma_1 B^{(1)}, ..., \sigma_K B^{(K)}, \sigma B)
\]

in \( D([0,T], \mathbb{R}^{K+1}) \) (endowed with the Skorohod \( J_1 \) product topology), where \( (B^{(1)}, ..., B^{(K)}) \) is a \( K \)-dimensional standard BM, \( B \) is a standard BM such that \( \sigma B = \sigma_1 B^{(1)} + ... + \sigma_K B^{(K)} \) (equal in distribution) with \( \sigma^2 = \sigma_1^2 + \cdots + \sigma_K^2 \) and

\[
\sigma_k^2 = \text{var}[\beta^{(k)}_i] + \lambda_k^2 E[\beta^{(k)}_i]^2, \quad k = 1, ..., K.
\]
B. Optimal power overage level from diffusion approximation

From Theorem 1 we can approximate the energy consumption up to time \( t \), \( C(t) \), for large \( \lambda \) by

\[
C(t) \approx X(t) := \lambda mt + \sqrt{\lambda \sigma} B(t), \quad t \geq 0. \tag{13}
\]

Given the threshold total charging rate \( L \) \( (L > \lambda \mu) \) at each time and a fixed time window \( t_o \), we define the energy consumption overages above \( L \) over \([0, T_o]\) by

\[
C(t_o, L, T_o) := \sum_{i=1}^{[T_o/t_o]} [C(i t_o) - C((i-1)t_o) - L t_o]^+ \tag{14}
\]

**Theorem 2:** When the demand is large, the expected energy consumption overages \( \mathbb{E}C(t_o, L, T_o) \) over any time interval \([0, T_o]\),

\[
\mathbb{E}C(t_o, L, T_o) \approx \frac{(\lambda \mu - L) T_o}{2} \Phi \left( \frac{(\lambda \mu - L)}{\sigma} \sqrt{t_o/\lambda} \right) + T_o \sigma \sqrt{\frac{\lambda}{2\pi \lambda t_o}} \exp \left( -\frac{(\lambda \mu - L)^2}{2\lambda \sigma^2} t_o \right), \tag{15}
\]

where \( \Phi(\cdot) \) is the cdf of the standard normal distribution.

**Note:** \( \lambda \mu < L \) so that the first term on the right hand side of (15) is negative, but the second term will dominate and the sum is nonnegative.

IV. SMART GRID SCHEDULING BY LOAD DEFERRAL

A. Summary framework

The assumption herein is that the smart grid is aware of aggregate demand \( X \) at each epoch (of duration \( t_o \)). If the demand exceeds a threshold \( L \), we assume that the smart grid is capable of deferring the overage \((X - L)\) to the next epoch. To do so, the grid needs to be able to ascertain which users to defer (cf. Section V) and needs to be able to temporarily cut supply to those users (e.g., car battery charging can be paused).

B. Overage

**Theorem 1** gives an expression for the limiting total energy consumption \( (13) \) over \([0, T_o]\), where \( T_o \leq T - \eta_{\max} \). Discretizing time by \( t_o \) seconds, we can obtain an approximate i.i.d. Gaussian process for the power \( P_j \), where at every time \( j t_o, j \in \mathbb{Z}^+ \):

\[
P_j \overset{d}{=} N(\lambda \mu t_o, \lambda \sigma^2 t_o)
\]

We assume idealized deferment of the excess aggregate load, where the carry-over load is the backlog of a discrete-time GI/D/1 queue \( X \) with arrivals \( P_j \) and deterministic service times \( L \),

\[
X_j = (X_{j-1} + P_j - L t_o)^+, \quad j \in \mathbb{Z}^+
\]

with \( X_0 = 0 \). Let \( J = T_o/t_o \) and note that the residual demand over time \([T_o, T]\) is

\[
X_J \overset{d}{=} \max\{Y_0, Y_1, \ldots, Y_J\}
\]

where \( Y_j = \sum_{i=1}^j (P_i - L t_o) \) for \( j \geq 1 \), and \( Y_0 = 0 \). Thus, by the reflection principle [3],

\[
P(X_j > x) = 2P(Y_j > x) = 2P(N(\lambda \mu - L T_o, \lambda \sigma^2 T_o) > x). \tag{16}
\]

So, we assume the average total overages levied by the utility will be

\[
\pi \Omega := \pi \mathbb{E}(X_j - L(T - T_o))^+ \tag{16}
\]

where \( \pi \) k$/\text{kWh}$ is the overage billing rate. One may expect that the *contract* between the electrical distribution system and the utility will depend on the \( \pi, L \), and \( \phi \leq \pi \) (the cost per kWh consumed below \( L \) will itself cost the utility in a manner increasing in \( L \); also, one would expect that \( \phi \) would be an increasing function of \( L \), but \( \pi \) may be decreasing in \( L \).

V. DISCUSSION: CHOOSING WHICH USER TO DEFER

Simply, users could be chosen uniformly at random by the grid for load deferment. Alternatively, deferment could be based on a score \( S \) of each user’s residual power-demand profile and the remaining time \( T - t \). For example, for a user that started at time \( \tau \) with initial charge zero, a large \( S \) would correspond to a larger probability of deferment so that \( S(h(t - \tau), \int \cdots \cdots h(r) dr, (\eta - (t - \tau))/(T - t)) \) would be an increasing function of its first argument (current/instantaneous power demand) and a decreasing function of the last two arguments (residual energy demand, and residual charge time as a fraction of remaining time), e.g., [1].

Such deferment policy may need to rely on trusting the residual demand reported by the user, unless that somehow securely metered by the grid. Again note that smart metering could be used to detect if a user misstates their load, does not obey a request to defer, or simply does not disclose their demand to the grid for purposes of “smart” deferment. Given simple applied cryptographic techniques (which may be in play in any case to ensure privacy), the utility can accept load attestations from the consumers in a non-repudiatable fashion (i.e., using accompanying digital signatures).

Recall that in our model, total overages are nominally shared by all significant overnight electricity consumers in proportion to their energy consumption, peak power demand, or some combination. Alternatively, uncooperative users might be penalized with an additional overage share, particularly if significant overages are experienced prior to the end-time of the charging period, \( T \), notwithstanding deferment actions by the grid.

Finally, we reiterate how the grid can easily police whether the users start at random times based on sampling their starting times night after night and capturing the distribution and comparing to a uniform via, e.g., Kolmogorov-Smirnov [3].

3And likewise honest consumers can confirm their load attestations were received by the grid through a signed acknowledgement.
VI. Future Work

As future work, we are developing experimental results for different user deferrament strategies. Also, rather than a central limit theorem, we plan to employ a large deviations principle (LDP) to model aggregate demand, see [2].

Acknowledgements: We thank Heath Hofmann and Jeff Rowe for helpful discussions.

References


Appendix A: Constant Consumption Profile

By the consumption profile in [1], the total instantaneous power consumption (charging rate) for all users is given by the flag process \( \{D(t) : t \geq 0\} \):

\[
D(t) = \sum_{i=1}^{N(t)} H_i 1(\tau_i \leq t \leq \tau_i + \eta_i).
\]

The energy (cumulative power) consumption for all users is given by the integrated (average) flag process \( \{C(t) : t \geq 0\} \), where

\[
C(t) = \int_0^t D(s) ds.
\]

We take an asymptotic approach by scaling the processes \( \{D(t) : t \geq 0\} \) and \( \{C(t) : t \geq 0\} \) with the arrival rate \( \lambda \) and letting \( \lambda \to \infty \). (This is justified because we consider a large number of consumers and \( \lambda = \eta / T_p \).) We index the the relevant processes \( N_\lambda(t) \), \( D_\lambda(t) \) and \( C_\lambda(t) \) by \( \lambda \). Define the diffusion-scaled processes \( \{C_\lambda(t) : t \geq 0\} \),

\[
\hat{C}_\lambda(t) = \frac{1}{\lambda} (C_\lambda(t) - \mu_\lambda t),
\]

where

\[
\mu_\lambda = m_\lambda + m_2 \eta.
\]

Thus, we can prove the following FCLT for \( \hat{C}_\lambda \) by applying Theorem 1 in [8]. The proof is omitted since it is similar to but simpler than the proof of Theorem 1 below. Let \( D([0,T], \mathbb{R}) \) be the space of 1-dimensional real-valued functions defined on \([0, T]\) that are right continuous on \([0, T]\) and has left limits on \((0, T]\).

Theorem 3: Under the assumption of constant consumption profiles in [1],

\[
C_\lambda \Rightarrow \sigma_r B \text{ in } D([0,T], \mathbb{R}) \text{ as } \lambda \to \infty,
\]

where \( D([0,T], \mathbb{R}) \) is endowed with the Skorohod \( J_1 \) topology, \( B \) is a standard Brownian motion (BM), and

\[
\sigma_r^2 = m_2 \eta + \frac{\eta^2}{2} \lambda^2 + \frac{\eta^2}{2} \lambda^2 + \frac{\eta^2}{2} \lambda^2 + \frac{\eta^2}{2} \lambda^2.
\]

We remark that when \( H_i = 1 \) a.s. for all users, the instantaneous charging rate process \( D(t) \) in fact corresponds to the queue-length process in an M/GI/\( \infty \) queue with Poisson arrival rate \( \lambda \) and service time \( \eta_i \). At each time \( t \), \( D(t) \) is a Poisson random variable with rate \( \lambda \int_0^t (1 - F(s)) ds \), where \( F(\cdot) \) is the cdf of \( \eta_i \), see e.g., p. 237 of [24]. In heavy traffic (\( \lambda \) is large while service times \( \eta_i \) are fixed), when \( \eta_i \) are exponential, Iglehart [7] first proved that the process \( D(t) \) can be approximated by an Ornstein-Uhlenbeck (OU) diffusion process,

\[
D(t) \approx \lambda m_H + \sqrt{\lambda m_H} OU(t;m_H, 2m_H), \text{ } t \geq 0,
\]

and

\[
D(\infty) \overset{d}{=} N(\lambda m_H, \lambda m_H)
\]

where \( OU(t; \nu, \sigma^2) \) is an OU process with drift \( -\nu x \) and diffusion coefficient \( \sigma^2 \), which has normal marginal distribution, and \( \overset{d}{=} \) means “approximately distributed as.”
APPENDIX B: PROOF OF THEOREM 1

Proof: We will follow Theorem 1 in [8] of a FCLT for general integrated (averaged) flag processes. First of all, for the independent Poisson processes $N^{(k)}$, we have

\[
\left(\tilde{N}^{(1)}_\lambda, \ldots, \tilde{N}^{(K)}_\lambda\right) \Rightarrow (\tilde{X}_1, \ldots, \tilde{X}_K)
\]

in $D([0, T], \mathbb{R}^K)$ as $\lambda \to \infty$, where $\tilde{N}^{(k)}_\lambda := N^{(k)}_\lambda/\lambda$, $e(t) = t$ for $0 \leq t \leq T$. Moreover, define the diffusion-scaled processes

\[
\tilde{N}^{(k)}_\lambda := \frac{1}{\sqrt{\lambda}} (N^{(k)}_\lambda - \lambda_k e),
\]

for $k = 1, \ldots, K$. Then,

\[
\left(\tilde{N}^{(1)}_\lambda, \ldots, \tilde{N}^{(K)}_\lambda\right) \Rightarrow (\tilde{B}_{a,1}, \ldots, \tilde{B}_{a,K})
\]

in $D([0, T], \mathbb{R}^K)$ as $\lambda \to \infty$, where $(\tilde{B}_{a,1}, \ldots, \tilde{B}_{a,K})$ is a $K$-dimensional standard BM.

Next, we consider the partial sum process $\{S^{(k)}_i : i \geq 1\}$ defined by $S^{(k)}_i = \beta^{(k)}_1 + \cdots + \beta^{(k)}_i$, where $\beta^{(k)}_i$ is defined as in (5). We can easily compute the mean and variance of $\beta^{(k)}_i$,

\[
\mu_{s,k} = \mathbb{E}[\beta^{(k)}_i], \quad \sigma^2_{s,k} = \text{var}[\beta^{(k)}_i].
\]

Define the diffusion-scaled partial sum processes $\{\tilde{S}^{(k)}_\lambda(t) : t \geq 0\}$ for $k = 1, \ldots, K$ by

\[
\tilde{S}^{(k)}_\lambda(t) = \frac{1}{\sqrt{\lambda}} \left(\frac{S^{(k)}_{\lambda t}}{\lambda t} - \lambda_k \mu_{s,k} t\right).
\]

We have the following FCLT for $\tilde{S}^{(k)}_\lambda(t)$,

\[
(\tilde{S}^{(1)}_\lambda, \ldots, \tilde{S}^{(K)}_\lambda) \Rightarrow (\sigma_{s,1} B_{s,1}, \ldots, \sigma_{s,K} B_{s,K})
\]

where $(B_{s,1}, \ldots, B_{s,K})$ is a $K$-dimensional standard BM, independent of $(B_{a,1}, \ldots, B_{a,K})$.

We then define the diffusion-scaled compound processes $\{\tilde{\gamma}^{(k)}_\lambda(t) : t \geq 0\}$ for $k = 1, \ldots, K$

\[
\tilde{\gamma}^{(k)}_\lambda(t) = \tilde{S}^{(k)}_\lambda \circ \tilde{N}^{(k)}_\lambda(t).
\]

and obtain the following FCLT for $\tilde{\gamma}^{(k)}_\lambda(t)$,

\[
(\tilde{\gamma}^{(1)}_\lambda, \ldots, \tilde{\gamma}^{(K)}_\lambda) \Rightarrow (\mu^{(1)} B_{a,1}, \ldots, \mu^{(K)} B_{a,K}).
\]

It is easy to show as Theorem 1 [8] that

\[
\sup_{0 \leq t \leq T} |\tilde{\gamma}^{(k)}_\lambda(t) - \tilde{\gamma}^{(k)}_\lambda(t)| = o(\lambda^{-1-\delta})
\]

as $\lambda \to \infty$ for each $k = 1, \ldots, K$. We need to check that the following condition holds,

\[
\int_a^\infty h(s + \xi_i; H, \eta_i, \zeta_i) ds = o(a^{-1-\delta})
\]

as $a \to \infty$ for some $\delta > 0$ and for all $i$. This is guaranteed by the assumption $\mathbb{E}[H_i^{2+\delta}] < \infty$ and $\mathbb{E}[\eta_i^{2+\delta}] < \infty$, since the integral in (23) is bounded above by the consumption profile with $h(t)$ equal to constant $H$ and the upper bound is of order $o(a^{-1-\delta})$ by the same argument in the first example in §4 of [8].

Thus, by Theorem 1 in [8], we have

\[
\sigma_k B^{(k)}(t) \overset{d}{=} \sigma_{s,k} B_{s,k} + \mu^{(k)} B_{a,k}, \quad k = 1, \ldots, K.
\]

and this implies that $\sigma_k^2 = \sigma_{s,k}^2 + (\mu^{(k)})^2$, which gives (12).

For the convergence of $\tilde{C}_\lambda$, we apply the continuous mapping theorem to the addition map.

APPENDIX C: PROOF OF THEOREM 2

Proof: From the approximation of $C(t)$ by the BM $X(t)$ in (13), we can approximate $C(t_o, L, T_o)$ in (14) by

\[
C(t_o, L, T_o) \approx \sum_{i=1} \left[ X(it_o) - X((i - 1)t_o) - L t_o \right]^+
\]

\[
= \sum_{i=1} \left[ (\lambda \mu - L)t_o + \sqrt{\lambda \sigma}(B(it_o) - B((i - 1)t_o)) \right]^+
\]

\[
\overset{d}{=} \sum_{i=1} \left[ (\lambda \mu - L)t_o + \sqrt{\lambda \sigma} B_i(t_o) \right]^+
\]

where $B_i$’s are mutually independent standard BMs, and the last equation is equal in distribution (d) by the property of independent increments for BMs. Recall that

\[
\mathbb{E}[N(\mu, \sigma^2)^+] = \mu \Phi(\mu/\sigma) + (\sigma/\sqrt{2\pi}) e^{-\mu^2/2\sigma^2} \geq (25)
\]

So,

\[
\mathbb{E}[N(\mu, \sigma^2)^+] = \mu \Phi(\mu/\sigma) + (\sigma/\sqrt{2\pi}) e^{-\mu^2/2\sigma^2} \geq (25)
\]

and thus, we can approximate $\mathbb{E} C(t_o, T_o)$ by (15).