Centralized Wireless Data Networks With User Arrivals and Departures

Rajat Prakash and Venugopal V. Veeravalli, *Fellow, IEEE*

***Abstract*—A dynamic-user model for centralized wireless networks is studied, where users arrive with a certain file size and depart when the file is served by a central server. Although the exact analysis of dynamic-user systems can be complicated, it is shown that an approximate analysis can be performed in a time-scale separation regime where the file size is much larger than the time scale of service process fluctuation. A first-order approximation result is derived that shows that when file sizes are large, a complicated service process can be replaced by a simple constant-rate service process. The accuracy of the approximation is further improved through a second order approximation result that incorporates the effect of service variability. Variability in the service process is shown to reduce the effective service rate, leading to a quantification of the conventional heuristic that service variability degrades system performance.**

***Index Terms*—Blocking probability, cellular networks, fading channel, mean delay, offered load, queueing systems, steady-state distribution.**

I. Introduction

In a centralized wireless data network many users share the services of a central unit or base station. Depending on whether the number of users in the system varies or remains fixed, centralized wireless systems can be divided into two broad categories. The first category consists of systems with a fixed number of users, where each user has an infinite backlog of information to transmit. For such *fixed-user systems*, the most important performance metric is the throughput attainable by each user. There is a vast body of work about the computing and optimizing the throughput for fixed-user systems. See, for example, [1], [2] for an information-theoretic overview, and [3] for a signal processing overview.

Our focus is on a second category of centralized wireless data systems, where the number of users varies with time. In such systems, new users arrive according to a stochastic arrival process, and each user has a finite sized file for transmission. A user leaves the system when the entire file is transmitted. Due to the dynamic evolution with time of the number of users, we call such systems *dynamic-user systems*. The operation of dynamic-user centralized wireless systems can be divided into two stages: initial access and data transfer. In the initial access stage, a new user informs the base station that he has data to transmit. Once the base station is aware of the user's request, it allocates a communication resource (say a frequency slot) to the user. After this allocation is complete, the new user enters the data transfer stage, and transmits data to the base station using the allocated resource.

Initial access is usually implemented through some version of Aloha, and the initial access problem has been studied extensively in the literature [4]. In contrast, there has not been much work on the data transfer aspect of dynamic-user centralized wireless systems, and most of the analyses rely on a fixed-user model. In this paper, we develop a framework for the analysis of data transfer schemes in dynamic-user systems.

The two most important performance metrics for a dynamic-user system are the mean delay and the blocking probability. The mean delay is the mean time between a user's arrival and completion of data transfer. Computation of the mean delay allows us to plot the tradeoff between offered load (users per second) and delay, and to determine the maximum allowable load that can be sustained while maintaining a tolerable delay. Using Little's law, the mean delay can be calculated from the steady-state distribution of the number of users, $\pi(u)$. The distribution π can also be used to compute the blocking probability, that is the probability that the number of users in the system is more than a certain threshold $u_{\max}$. Thus, the distribution $\pi(u)$ characterizes the performance of a dynamic-user system, and the goal of this paper is to evaluate this distribution.

The problem of evaluating $\pi(u)$ for dynamic-user systems was first studied by Telatar and Gallager [5], under the assumption that the service process has a constant rate $\phi(u)$ bits per second, that depends only on the number of users u in the system (a symmetric queue model). However, the constant service rate assumption in [5] is restrictive. In practice, the service process in wireless systems is often stochastic, i.e., the service rate varies with time. Reasons for the stochastic nature of the service process include channel fading and inherent randomness in the communication protocol (e.g., retransmission timers for automatic repeat request (ARQ)). Due to these complexities of the service process, the simple model of [5] is not applicable to many systems of interest. The exact analysis of realistic dynamic-user systems is often complicated, and can involve finding the steady-state distribution of a Markov chain with a large state space (Raychaudhuri [6], Zhang *et al.* [7]). The main contribution of this paper is to develop an approximation

Manuscript received January 31, 2004: revised November 11, 2006. This work was supported by the National Science Foundation CAREER/PECASE award CCF 0049089, through the University of Illinois. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Lausanne, Switzerland, June/July 2002, and at the IEEE International Symposium on Information Theory, Yokohama, Japan, June/July 2003.

R. Prakash is with Qualcomm, Inc., San Diego, CA 92122 USA (e-mail: rprakash@qualcomm.com).

V. V. Veeravalli is with the Department of Electrical and Computer Engineering, and the Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801 USA (e-mail: vvv@uiuc.edu).

Communicated by G. Sasaki, Associate Editor for Communication Networks.

Digital Object Identifier 10.1109/TIT.2006.889461

technique that simplifies the analysis of dynamic-user systems with complicated Markovian service models.

To analyze a given dynamic-user system, we first construct a set of fixed-user systems by blocking the arrival and departure of users in the given dynamic-user system. For this set of fixed-user systems, we evaluate the first- and second-order statistics of the service process, this evaluation being considerably simpler than the analysis of the original dynamic-user system. Then, we approximate the performance of the dynamic-user system in terms of statistics of the service in the set of fixed-user systems. This technique allows us to break the complex problem of dynamic-user system analysis into two simpler problems: fixed-user system analysis, followed by symmetric queue analysis.

Our first result is a first-order approximation, where we approximate the distribution π in terms of the mean $\phi(u)$ of the service of a fixed-user system with u users. In an asymptotic regime where the file sizes are much larger than the time scale of variation of the service process, we show that the given dynamic-user system can be approximated by a symmetric queue with service rate $\phi(u)$. We used this approximation earlier in [8] and [9], and recently, such an approximation has also been used in [10] and [11], albeit without a rigorous justification. In this paper, we give a rigorous justification for the first-order approximation.

Our second result is a second-order approximation, where, for a certain class of dynamic-user systems, we refine the first-order approximation in terms of the variance time constant $\sigma^2(u)$ of the service process. We show that when the file size $\bar{S}$ is large, the dynamic-user system is well approximated by a symmetric queue with service rate

$$\phi_{\text{eff}}(u) = \phi(u) - \frac{\sigma^2(u)}{2\bar{S}} \tag{1}$$

leading to a second-order approximation for the distribution π. This approximation quantifies the effect service process variability has on the mean delay. In particular, it shows that service process variability reduces the effective service rate, and that the reduction is proportional to the variance of the service process.

The second-order approximation result can be contrasted with two other results from queueing that relate variability with service time. For the M/G/1/FCFS queue with constant service rate, it is well known that the mean delay increases linearly with file size variance. On the other hand, for the M/G/1/PS queue with constant service rate, it is known that the mean delay is not a function of file size variance. The second-order approximation implies that for an M/G/1/PS queue with nonconstant service rates, the mean delay increases with service rate variance. Note that the second-order approximation provides a reduced effective service rate for each u values in the M/G/1/PS, and is not equivalent to considering an increased effective file size.

Our first- and second-order approximations say that as viewed from a higher layer perspective, a complex physical layer can be modeled by a constant-rate data pipe with appropriately selected rates. This constant-rate data pipe model is related to [12], where large deviations and the theory of effective bandwidths are used. However, the technique in [12] is applicable only to single-user systems, and their main objective is to bound the probability of buffer overflow at the queue of this single user.

Although not directly related with our work, we also mention the following stream of work that seeks to reconcile queueing and physical layer considerations. For systems with a fixed number of users, a random packet arrival process, and a randomly varying channel, the issue of deciding which user transmits in a given slot (the scheduling problem) has been considered in [13]–[18]. Our analysis is different from these works because we allow the number of users in the system to vary.

The rest of the paper is organized as follows. An overview of the time-scale separation results, without the formal definitions, is presented in Section II. A detailed system model and statement of results is given in Section III. Section V contains the proof of time-scale separation, and uses two lemmas for the proof. These lemmas, in turn, are proved in Section VI. Finally, conclusions are given in Section VIII.

II. Overview of Results

In this section, we present an overview of our results on the asymptotic analysis of dynamic-user systems. Our approximation results take the form of approximating the given dynamic-user system by a symmetric queue, which is described next.

A. Symmetric Queue Model

The symmetric queue model was used for the analysis of centralized wireless systems by Telatar and Gallager [5]. In a symmetric queue, users arrive into the system according to a continuous time Poisson process with rate λ users per second. An arriving user is blocked if the number of users reaches an admission threshold u_0. The file size of each user is independent and identically distributed with a typical file size S bits. When there are u users in the system, data is transmitted at a constant rate of $\phi(u)$ bits per second. This data rate is divided equally among all users, giving rate $\frac{1}{u}\phi(u)$ bits per second to each user. A user leaves the system when the entire file has been transmitted.

For the symmetric queue, a key quantity of interest is the steady-state distribution of the number of users in the system, and is denoted by $\mu(u) = \mathsf{Pr}\{u \text{ users in system}\}$. The steady-state distribution can be used to compute the mean delay D using Little's law

$$D = \frac{1}{\lambda}\sum_{u=1}^{\infty} u\mu(u). \tag{2}$$

Also, using the Poisson arrivals see time averages (PASTA) property [19], the blocking probability can be computed as $\mu(u_0)$.

For the symmetric queue, $\mu(u)$ can be computed in closed form, and it can be shown that μ depends on λ and $\bar{S}$ only through the offered load $\Omega \stackrel{\text{def}}{=} \lambda\bar{S}$.

$$\mu(u) = \frac{\Omega^u}{K\prod_{j=1}^{u}\phi(j)}\mathbb{1}_{\{u\le u_0\}} \tag{3}$$

where K is a normalizing constant and $\mathbb{1}_{\{.\}}$ is the indicator function. Although (3) is valid irrespective the distribution of the file size S (see [20, Sec. 3.3] for details), in the special case of exponentially distributed file sizes,(3) can be interpreted as arising from the Markov chain in Fig. 1.

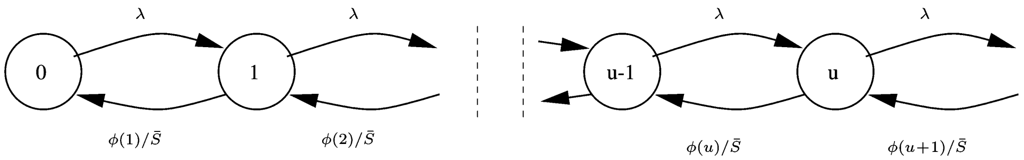

Fig. 1. Markov chain model for number of users u for a symmetric queue with exponential file sizes.

B. *First-Order Approximation*

The symmetric queue model of the previous section requires a constant service rate assumption. For more general Markov modulated service models, such as may be found in realistic dynamic-user systems, the complexity of the Markov state space makes it difficult to analytically evaluate the steady-state distribution of the number of users. We denote this steady-state distribution by $\pi(u)$ (to distinguish it from the symmetric queue distribution $\mu(u)$). We will show that when the average file size is much larger than the time scale of variation of the service process, a *time-varying* service process can be effectively replaced by a *constant-rate* service process with the same mean. Although we consider general Markov modulated models for the service process, we continue to make two symmetry assumptions inherent in the model of [5].

- *Symmetric Users:* All users have statistically identical channels and file size distributions. In the context of cellular systems, this means that all users are subjected to the same slow fading (that does not change over the time period of interest), with the fast fading being statistically identical for all users. If the slow-fading terms are not the same (due to varying distances from the base station), we can satisfy the symmetric users condition by using power control to compensate for the slow fading.
- *Symmetric Resource Allocation:* Service is provided to each user with equal priority. Symmetric resource allocation rules out the use of residual file size based resource allocation. It is known that file size based prioritization of users can reduce the average delay seen by a user. For example, smallest remaining processing time (SRPT) [21] can result in a reduction in the mean delay. Some of these issues are addressed by Telatar in [22].

We call a dynamic-user system to be *separable* if it satisfies both of the preceding symmetry properties. For nonseparable systems, the analysis of dynamic-user systems is difficult even for a constant service rate. In this paper, our focus is on extending the analysis of [5] to stochastic or discrete service models, and we do not consider the problem of relaxing the symmetry assumption (a discussion of the asymmetric users case is given in [23])

The first-order approximation will be applicable to separable systems. To model systems with large file sizes, consider a sequence of dynamic-user systems indexed by α, where each system has the same service process statistics. For system-α, let the average file size be $\frac{1}{\alpha}$ bits, and the arrival rate be $\lambda = \alpha\Omega$ users per second (giving the same offered load Ω bits per second for each system in the sequence). Further, let the steady-state distribution of the number of users for system-α be $\pi^{\alpha}(u)$.

To characterize the statistics of the general service process, we construct a fixed-user system by disallowing arrivals and setting the file size of each user to infinity in the given dynamic-user system. For such a fixed-user system with u users, let $\tilde{I}_j(t)$ be the service given to user j in time t, and let the average service rate be

$$\frac{\phi(u)}{u} = \lim_{t\to\infty} \frac{\tilde{I}_j(t)}{t}, \qquad j = 1, \ldots, u. \tag{4}$$

Note that the symmetric users assumption is reflected in the above equation because each user has the same average service rate.

When the file size is large (α is small), each user will be in the system for a long time. Then, we can use the limit in (4) to argue that each user sees a nearly constant service rate $\phi(u)/u$. Thus, for large file sizes, we should be able to approximate the given system by a symmetric queue with service rate $\phi(u)$ and steady-state distribution $\mu(u)$. This heuristic is stated in the following first-order approximation result.

Result 1: The steady-state distribution π^{α} of a separable dynamic-user system can be approximated by the steady-state distribution μ of an equivalent symmetric queue, i.e.,

$$\pi^{\alpha}(u) \approx \mu(u). \tag{5}$$

A formal statement of this result is given in Section III. This result allows us to simplify the complexities of the physical layer by replacing the service process by a simple constant-rate service process that depends only on the number of users in the system.

C. *Second-Order Approximation*

The second-order approximation improves the first-order approximation by using the service process variance time-constant $\sigma^2(u)$, that is defined as

$$\sigma^2(u) \overset{def}{=} \lim_{t\to\infty} \frac{1}{t} E\left[\tilde{I}(t) - \phi(u)t\right]^2. \tag{6}$$

Here, $\tilde{I}(t)$ is the sum of the service provided to all users in a fixed-user system with u users. Such a time constant will exist for a large variety of service processes, including the class of service processes that are governed by a Markov chain.

The second-order approximation is applicable to a class of systems that satisfy the following conditions: independent service, stationary arrival state, exponential file sizes, and negligible discreteness. The exact definitions of these conditions are provided in Section III-D, and the conditions are informally described as follows.

- *Independent Service:* The independent service condition is a technical condition that is described in precise form in Section III-D. This condition depends on the physical layer, and is satisfied for most physical layer scenarios when the channel fading is uncorrelated across time slots. When the channel is correlated in time, the independent service condition is satisfied by time-division multiple access (TDMA) and frequency-division multiple access (FDMA), but not by channel condition based scheduling schemes.

- *Stationary Arrival State:* The stationary arrival state condition is a technical condition that is described in precise form in Section III-D. The stationary arrival state condition requires that the fading state of an arriving user is distributed according to the marginal distribution of fading. This is a reasonable assumption for wireless systems because the arrival process (governed by the application layer) can be modeled as being independent of the channel state.

  The exact statement of the independent service and independent arrival state conditions is given in Section III-D.
- *Exponential File Sizes:* This condition means that the file size of arriving users is exponentially distributed.
- *Negligible Discreteness:* The difference between the dynamic-user system and the symmetric queue is caused by the deviation of the dynamic-user service process from a constant rate. This deviation is $\tilde{I}(t) - \phi(u)t$, and it can be looked upon as having two distinct components, one discrete and the other stochastic. The discrete component of the service process arises due to the fact that service is provided in chunks proportional to the slot length δ, rather than as a continuous flow. The stochastic component of the service process, on the other hand, arises due to fading induced variations in the service process, and is represented by the variance time constant $\sigma^2(u)$. The negligible discreteness condition is said to hold when δ is significantly smaller than a $\sum_u \sigma^2(u)/\phi(u)$ (a motivation for this condition is provided when the condition is discussed formally).

  Unlike the independent service and exponential file size conditions, the negligible discreteness condition is not a fundamental property of the system, but rather depends on the numerical parameters of the system. The exact nature of this dependence on numerical parameters is considered when the second-order approximation is discussed formally (Section III-D). Numerical examples demonstrating the role of negligible discreteness are given in [23].

To approximate π^α for a dynamic-user system that satisfies the above three conditions, we consider a symmetric queue with modified service rates

$$\phi^\alpha(u) = \phi(u) - \frac{1}{2}\alpha\sigma^2(u) \tag{7}$$

and corresponding steady-state distribution μ^α. The following result links the distribution of interest π^α with the distribution μ^α.

TABLE I
GLOSSARY

Symbol	**Interpretation**
E_x	Expectation, with initial state fixed at x
u	Number of users in the system
Ω	Offered load in bits per second
δ	Slot duration, in seconds
$\tilde{I}_j(t)$	Service process for user-j in a fixed-user system (bits)
$\tilde{I}(t)$	Sum of service processes of all users in a fixed-user system
$\phi(u)$	Mean rate of $\tilde{I}(t)$ for a fixed-user system with u users (bits/sec)
$\sigma^2(u)$	Variance time-constant of $\tilde{I}(t)$ for fixed-user system with u users (bits^2/sec)
α	Index of sequence of dynamic-user systems with mean filesize $1/\alpha$
$X^\alpha[k]$	Markov chain for dynamic-user system
$\mathcal{X}$	Space in which $X^\alpha[k]$ evolves
$\tilde{X}[k]$	Markov chain for fixed-user system
$\mathcal{X}_u$	Space in which $\tilde{X}[k]$ evolves for fixed-user system with u users
f_j	function that maps state $x \in \mathcal{X}$ to service received by user-j
f	sum of f_j over all users in the system
$\pi^\alpha(x)$	Steady state distribution of $X^\alpha[n]$
$\mu(u)$	Steady state distribution for symmetric queue
μ^α	Steady state distribution for modified symmetric queue
η_u	Steady state distribution of fixed-user system with u users
$\bar{O}(\alpha)$	An $O(\alpha)$ term where the constant depends only on the symmetric queue parameters
A_k	Number of arrivals in slot k
$D_{j,k}$	Boolean variable indicating departure of user j in slot k
G^α	The first slot with an arrival or departure
$H^\alpha(x)$	Mean of G^α when the initial condition is x
X_n^α	Jump chain, with jump instants defined as arrival and departure instants
$\gamma^\alpha(x)$	Distribution of jump chain X_n^α
$\bar{U}(t)$, $\bar{U}^\alpha(t)$	Markov process for symmetric queue and modified symm. queue
$\bar{U}_n$, $\bar{U}_n^\alpha$	Jump chain for $\bar{U}(t)$, $\bar{U}^\alpha(t)$
$\bar{H}(u)$, $\bar{H}^\alpha(u)$	Mean holding time in state u for $\bar{U}(t)$, $\bar{U}^\alpha(t)$
$\bar{P}(j\|i)$, $\bar{P}^\alpha(j\|i)$	Transition probabilities for $\bar{U}_n$, $\bar{U}_n^\alpha$
$\bar{\gamma}(u)$, $\bar{\gamma}^\alpha(u)$	Distribution of jump chain $\bar{U}_n$, $\bar{U}_n^\alpha$

Result 2: For a separable dynamic-user system, that satisfies the independent service, stationary arrival state, negligible discreteness, and exponential file size conditions, an improved approximation for π^α is given by the steady-state distribution μ^α of a modified symmetric queue

$$\pi^\alpha(u) \approx \mu^\alpha(u). \tag{8}$$

For a formal statement of the second-order approximation result, see Section III-D. The second-order approximation result means that variance in the service process results in a reduction in the effective service rate according to (7). From Little's law and the structure of the symmetric queue, we can compute the increase in delay caused by service variability. Further, from the form of (7), it follows that the increase in delay is larger for smaller file sizes. This is so because users with smaller file sizes experience less averaging of the service variability, and their delay can be adversely effected by small periods of poor service.

III. Detailed System Model and Results

In this section, we give a formal definition of the system model and formally state the first- and second-order approximation results of Section II. A glossary of notation is given in Table I.

A. A General Dynamic User Model

In this subsection, we describe a system model that is required to prove the first- and second-order approximation results. Our system model is general in the sense that it can be used to describe various service schemes at the physical layer. The model uses slotted time, and we assume that the users are provided service only at boundaries of slots with length δ seconds. To simplify the notation, we assume that the file sizes are exponentially distributed. The model can be extended to accommodate nonexponential files (Section VII), but in that case, as stated in Result 2, only the first-order approximation result holds.

The details of the dynamic-user model are as follows. The number of users arriving in any given slot k is $A^{\delta}[k]$, and has a Poisson distribution with mean $\alpha\Omega\delta$ (giving an arrival rate of $\alpha\Omega$ users per second, as motivated in Section II-B). An arriving user is blocked if there are u_0 users in the system. [1] The file sizes of arriving users are independent and identically distributed with exponential random variable S denoting the typical file size. The mean of the random variable S is $1/\alpha$. This gives the effective arrival rate as Ω bits per second. The service provided to user j from time 0 to time t is $I_j(t)$, and a user departs when the service exceeds the file size.

We define general dynamic-user systems through a general Markov chain $X^{\alpha}[k]$. The state vector for this Markov chain contains information about the number of users in the system, the channel state of each user, and possibly protocol specific information, such as the active user index for a TDMA system.

Dynamic-User System: A dynamic-user system is defined by a Markov chain $X^{\alpha}[k]$ with state vectors in a space $\mathcal{X}$ such that the following properties hold.

1) There is a partition $\mathcal{X} = \cup_{u=0}^{u_0}\mathcal{X}_u$ such that $X^{\alpha}[k] \in \mathcal{X}_u$ implies that there are u users in the system. For analytical ease, assume that the $\mathcal{X}_u$ is finite for each u, and that $\mathcal{X}_0$ has only one element.
2) The service for user j in slot k is given by

$$I_{j,k} = f_j(X^{\alpha}[k]).$$

Here, interpret the state X^{α} as containing the channel state information of the system, and function f_j as a map from channel state to the service level.
3) A Bernoulli random variable $D_{j,k}$ denotes the departure of user j in slot k, with

$$D_{j,k} = \begin{cases} 1, & \text{w.p. } 1-\exp(-\alpha f_j(X^{\alpha}[k])) \\ 0, & \text{w.p. } \exp(-\alpha f_j(X^{\alpha}[k])). \end{cases} \tag{9}$$

The random variables $D_{j,k}$ are independent across j for a fixed k. This form of $D_{j,k}$ follows from the exponential file size distribution and the symmetric service assumption.
4) The number of new users arriving in slot k is A_k, and has a Poisson distribution with mean $\delta\Omega\alpha$. The arrivals are independent across slots.
5) The distribution of the new state is given by a kernel K

$$\mathsf{K}(X^{\alpha}[k+1]|X^{\alpha}[k], D_{1,k}, \ldots, D_{u,k}, A_k).$$

This kernel does not depend on α or k. Further, the kernel is such that the number of users never exceeds u_0. The kernel K describes the stochastic evolution of the channel states and protocol specific information contained in the state vector.

The preceding description forces the dynamic-user system to obey the symmetric resource allocation and exponential file size conditions. This follows from the definition of departures in part 3, which states that, irrespective of the system state, the residual file size is exponential with mean $1/\alpha$. Thus, the system state is not allowed to contain the residual file size of the users in the system, and service prioritization based on residual file sizes is not allowed. Apart from these two restrictions (exponential file sizes and symmetric resource allocation), the description of dynamic-user systems we use in this paper is quite general. An extension to nonexponential file sizes is described in Section VII. Asymmetric resource allocation schemes are considered in [23].

B. The Fixed-User System

To develop the time-scale separation approximation, we define a fixed-user system that is obtained by blocking arrivals and departures in the dynamic-user system. This fixed-user system is relatively easy to analyze, and its analysis, coupled with the time-scale separation result, yields a good approximation to the performance of the given dynamic-user system.

Definition 1: A fixed-user system with u users is a Markov chain $\tilde{X}[k]$ (in state space $\mathcal{X}_u$) that is obtained by disallowing arrivals and departures in the probability transition kernel to give a new kernel

$$\tilde{K}_u(\tilde{X}[k+1]|\tilde{X}[k]) = \mathsf{K}(\tilde{X}[k+1]|\tilde{X}[k], 0, \ldots, 0, 0).$$

The service for user j in slot k is given by $\tilde{I}_{j,k} = f_j(\tilde{X}[k])$, with cumulative service for user j being denoted by

$$\tilde{I}_j[n] = \sum_{k=0}^{n-1} \tilde{I}_{j,k}$$

and cumulative service for all users being denoted by

$$\tilde{I}[n] = \sum_{j=1}^{u} \tilde{I}_j[n].$$

C. First-Order Approximation

The first-order approximation of Section II-B required that the time average of the service process converge to a constant $\phi(u)$ for each user in the system. When the fixed-user Markov chain $\tilde{X}[k]$ is ergodic, it follows from the theory of Markov chains ([19, Sec. 3.5] or [24, Theorem 1.1]) that the time average service for each user converges to a constant. User symmetry requires that the averaged service rate be the same for each user, i.e., in each slot, each user is served an average of $\delta\phi(u)/u$ bits. This requirement, and some technical conditions are stated in the following definition of *separability*.

[1] The case where $u_0 = \infty$ is discussed in [23].

Definition 2: Separable System: A dynamic-user system is said to be separable if the corresponding fixed-user system with u users is ergodic with stationary distribution η_u such that

$$\sum_{x \in \mathcal{X}_u} \eta_u(x) f_j(x) = \delta \frac{\phi(u)}{u}, \qquad j = 1, \ldots, u.$$

The first-order approximation aims to compute the steady-state distribution of the number of users in the dynamic-user system, and we argue next that a steady-state distribution exists for separable dynamic-user systems because the Markov chain $X^\alpha[k]$ is ergodic when the fixed-user system is separable.

Ergodicity: First, we note that ergodicity of the fixed-user system is a reasonable assumption because in the examples we will consider, the state vector $\tilde{X}[k]$ will represent the channel state, and ergodicity of the channel is a reasonable assumption.

We argue now that when the fixed-user system is ergodic, the dynamic-user system Markov chain $X^\alpha[k]$ is also ergodic. First, we argue that $\mathcal{X}_0$ can be reached from any $x \in \mathcal{X}_u$. This will be demonstrated if we can show a path from $x \in \mathcal{X}_u$ to $\mathcal{X}_{u-1}$ for all $u > 0$. Consider sample paths where there are no arrivals. Then, by ergodicity of the fixed-user system, the service process is nonzero for at least some sample path. For this sample path, the probability of departure is nonzero, giving a path to $\mathcal{X}_{u-1}$. Second, we argue that every $y \in \mathcal{X}_u$ can be reached from every $x \in \mathcal{X}_u$. This follows by considering sample paths with no arrivals and departures, and using the fact that the fixed-user system is ergodic. Third, by definition of the arrival process, it follows that for every $y \in \mathcal{X}_u$ it is possible to reach some state in $\mathcal{X}_{u+1}$. The above three arguments together imply that $\mathcal{X}_u$ is irreducible. Further, the time spent in $\mathcal{X}_0$ (which is a singleton) can take any value, making the chain aperiodic. This concludes the argument for the ergodicity of the dynamic-user system.

If $X^\alpha[k]$ is ergodic, there is a steady-state distribution π^α on $\mathcal{X}$, such that for all initializations $x \in \mathcal{X}$, the distribution of $X^\alpha[k]$ converges to π^α for large k. With this definition of π^α, the steady-state probability of u users being present in the system is $\pi^\alpha(\mathcal{X}_u)$. The following theorem states that $\pi^\alpha(\mathcal{X}_u)$ can be approximated by the steady-state distribution of an equivalent symmetric queue.

Theorem 1 (First-Order Approximation) : Consider a separable dynamic-user system with limiting service rate $\phi(u)$. Let $\mu(u)$ be the steady-state distribution of a symmetric queue with service rate $\phi(u)$ and offered load Ω. Then

$$\pi^\alpha(\mathcal{X}_u) - \mu(u) = O(\alpha). \tag{10}$$

Proof: See Section V. □

The proof uses multiplicative ergodicity, and exploits the link between the dynamic-user system and the symmetric queue. The theorem can also be proved by using the theory of time-scale decomposition of Markov chains [25], and a heuristic argument using this theory is as follows. Under our model, transitions within the sets $\mathcal{X}_u$ take place much faster than transitions across the boundaries of the sets $\mathcal{X}_u$. Thus, we can assume that the state distribution within $\mathcal{X}_u$ is η_u (the steady-state distribution of the fixed-user system). Using the definition of a separable system, the mean service rate under the distribution η_u is $\phi(u)$. Thus, downward transitions ($\mathcal{X}_u$ to $\mathcal{X}_{u-1}$) have a rate $\alpha\phi(u)$. Further, by definition of the arrival process, upward transitions have a rate $\alpha\Omega$. Thus, the sets $\mathcal{X}_u$ can be replaced by the elements u of the symmetric queue, giving a justification for the first order approximation.

D. Second-Order Approximation

The second-order approximation of Result 2 in Section II-C required the existence of a variance time constant in the limit of (6). From the theory of Markov chains [26, Theorem 17.5.3], we know that this limit exists, i.e., the process $\tilde{I}(t)$ has a time-average variance constant that satisfies for all initial conditions $x \in \mathcal{X}_u$

$$\lim_{n \to \infty} \frac{1}{n} \mathsf{E}_x \left[(\tilde{I}[n] - n\delta\phi(u))^2 \right] = \sigma^2(u)\delta. \tag{11}$$

In the preceding equation, the variance time constant is denoted by $\sigma^2(u)\delta$ to ensure consistency with the variance time constant $\sigma^2(u)$ of the continuous-time service process in (6).

In Section II-C, we stated that the second-order approximation requires the independent service and independent arrival state condition. Next, we define these conditions precisely in terms of the way transitions occur from $\mathcal{X}_u$ to $\mathcal{X}_{u-1}$ or $\mathcal{X}_{u+1}$. The independent service condition says that if X^α is initialized with the distribution η_u, and a departure occurs in slot 0, then the resulting state in $\mathcal{X}_{u-1}$ is distributed according to η_{u-1}. Similarly, the stationary arrival state condition says that in case of an arrival, the resulting state in $\mathcal{X}_{u+1}$ is distributed according to η_{u+1}. We define systems that satisfy both these conditions to be nicely separable.

Definition 3: Nice Separability: A separable dynamic-user system is said to be nicely separable if it has independent service and stationary arrival state. When X^α is initialized in distribution η_u, i.e., $\mathsf{Pr}(X^\alpha = x) = \eta_u(x)$, the independent service condition is that the following is satisfied:

$$\mathsf{Pr}_{\eta_u}(X^\alpha[1] = x | A_0 = 0, D_0 = 1) = \eta_{u-1}(x) + O(\alpha) \tag{12}$$

and the stationary arrival state condition is that the following is satisfied:

$$\mathsf{Pr}_{\eta_u}(X^\alpha[1] = x | A_0 = 1, D_0 = 0) = \eta_{u+1}(x) + O(\alpha) \tag{13}$$

where A_0 and D_0 are the number of arrivals and departures, respectively, in slot 0.

A discussion of physical layer schemes that satisfy the nice separability condition is given in Section IV.

To proceed with the second-order approximation for a nicely separable system, we define the service discreteness parameter $\delta_d = \Omega\delta + \delta_f$, where

$$\begin{aligned} \delta_f &= \max_{u \le u_0} \frac{1}{\delta(\Omega + \phi(u))} \sum_{x \in \mathcal{X}_u} \eta_u(x) f(x)^2 \qquad &(14)\\ &\le \max_{u \le u_0} \frac{\phi(u)}{\Omega + \phi(u)} \max_{x \in \mathcal{X}_u} f(x) \\ &\le \max_{x \in \mathcal{X}} f(x) &(15) \end{aligned}$$

where we have used Definition 2 in the second step.

The following theorem says that nicely separable dynamic-user systems are approximated well by a symmetric queue with effective service rates given by (7), with the error in the approximation being small when the discreteness parameter defined above is small.

Theorem 2 (Second-Order Approximation): For a nicely separable dynamic-user system with exponential file sizes, and a symmetric queue with effective service rates ϕ^α

$$\pi^\alpha(\mathcal{X}_u) - \mu^\alpha(u) = o(\alpha) + \delta_d \bar{O}(\alpha) \tag{16}$$

where the constant in the $\bar{O}(\alpha)$ term is a particular $O(\alpha)$ term that depends only on the parameters of the symmetric queue (and not on δ and the structure of the Markov chain $\tilde{X}$).

Proof: See Section V. □

Relation with Result 2: We argue next that Theorem 2 leads to Result 2. In order to show that the second-order approximation provides an improvement over the first-order approximation, we will show that $\pi^\alpha(\mathcal{X}_u) - \mu^\alpha(u)$ is much smaller than $\pi^\alpha(\mathcal{X}_u) - \mu(u)$

Consider the difference $\mu^\alpha(u) - \mu(u)$. By definition of μ^α in (7) and (3), it can be shown that this term is of the order $\alpha \sum_u \sigma^2(u)/\phi(u)$.[2]

Next, consider $\pi^\alpha(\mathcal{X}_u) - \mu^\alpha(u)$ that is described in Theorem 2. The first term on the right-hand side is $o(\alpha)$ and for small enough α, is much smaller than $\alpha \sum_u \sigma^2(u)/\phi(u)$. For the second term, we will show that $\delta_d \bar{O}(\alpha)$ is much smaller than $\alpha \sum_u \sigma^2(u)/\phi(u)$. Assuming that the service provided in one time slot of duration δ (the quantity $f(x)$) is related to the slot duration δ, we have that δ_d defined in (14) satisfies

$$\delta_d < K_0 \delta$$

where K_0 is the maximum instantaneous service rate possible in the system. Also, from Theorem 2, we know that $\bar{O}(\alpha)$ is an $O(\alpha)$ term that depends only on the symmetric queue, giving for small enough α

$$\bar{O}(\alpha) < K_1 \alpha$$

where K_1 depends only on the parameters of the symmetric queue. Putting the two equations above together

$$\delta_d \bar{O}(\alpha) < K_1 K_2 \delta \alpha.$$

From the negligible discreteness condition, we have that δ is suitably smaller than $\sum_u \sigma^2(u)/\phi(u)$. The suitably small requirement should be met to the extent that

$$\delta \ll \frac{1}{K_0 K_1} \sum_u \sigma^2(u)/\phi(u).$$

This immediately gives $\delta_d \bar{O}(\alpha) \ll \alpha \sum_u \sigma^2(u)/\phi(u)$.

This completes the demonstration of result 2.

IV. The Nice-Separability Condition

The nice-separability condition of the previous section had a technical form. In this section, we show that the nice-separability condition is satisfied by several physical layer schemes of interest.

A. Model Overview

To model a physical layer scheme, let the state vector $X^\alpha[k]$ contain the number of users in the system, and the channel states of each user in the system, i.e.,

$$X^\alpha[k] = (U[k], h_{1,k}, \ldots, h_{u,k}) \tag{17}$$

where $h_{j,k}$ is the fading level of user j at time k. For simplicity of notation, we will assume that $h_{j,k}$ takes values 0 through $L-1$, with marginal distribution $q(h_{j,k})$. Also, define $\boldsymbol{h}_k^{1,u} = (h_{1,k}, \ldots, h_{u,k})$, and the product distribution

$$q^u\left(\boldsymbol{h}_k^{1,u}\right) = \prod_{j=1}^{u} q(h_{j,k}).$$

Let the fading for each user evolve according to a Markov process with transition from state $h_{j,k}$ to state $h_{j,k+1}$ occurring with probability $Q(h_{j,k+1}|h_{j,k})$. Further, let the fading process evolve independently across users, i.e.,

$$\Pr\left(\boldsymbol{h}_{k+1}^{1,u} | \boldsymbol{h}_k^{1,u}\right) = \prod_{j=1}^{u} Q(h_{j,k+1}|h_{j,k}). \tag{18}$$

To complete the description of the physical layer, let the service received by user j in slot k depend on the channel state through a function $f_j(X^\alpha[k])$.

Under these conditions on the channel, $X^\alpha[k]$ itself forms a Markov chain, with the state transition described by the definition of the dynamic-user system in Section III-A. The state transition rules of Section III-A can be summarized as follows. Consider an initial state $X^\alpha[k] = x \in \mathcal{X}_u$. Then, the new state X^α is determined as follows.

First, determine the service received by each of the users using f_j. Then, determine if user j departs in slot 0 by generating a random variable $D_{j,k}$ that is distributed according to (9). Repeat the process for all users to generate $D_{1,k}$ through $D_{u,k}$. Finally, for the users that remain in the system, determine the new channel state according to the Markov model for channel evolution. Then, determine the new arrivals in the system by generating a Poisson random variable A_0. If there is a new arrival, generate its channel state according to the marginal distribution of fading. This process generates the new channel state $X^\alpha[k+1]$.

To further define the state evolution, assume that if u users are present and user j departs, then users $j+1$ through u have their user index reduced by one. Similarly, if a user arrives, assume that this arriving user is inserted at position j, with j uniformly distributed between 1 and $u+1$. Users after position j have their user indices increased by one.

For the state evolution described above, it can be seen that the stationary arrival state condition (13) is satisfied due to the assumption that the channel state of new users is distributed according to the marginal distribution.

The independent service condition (12) requires a more elaborate treatment, and the rest of this section is devoted to

[2] It may be possible study the effect of $\sigma^2(u)$ through a tighter function, but we select this summation for simplicity.

establishing that the following three interesting physical layer schemes satisfy the independent service condition.

(i) The function f is arbitrary, and fading is independent across time, i.e., $Q(\ell_1|\ell_0) = q(\ell_1)$.

(ii) The service received by user j depends only on the channel state of user j, and not on the channel states of other users. Thus (with some abuse of notation)

$$f_j(X^\alpha[k]) = f_j(h_{j,k}).$$

Further, the channel is correlated across time slots. This includes the case of FDM sharing of resources by users.

(iii) Users share the channel using TDMA, with channel correlation across time slots. In case of TDMA, the state of (17) needs to be enlarged to accommodate the round-robin nature of TDMA.

For cases (i) and (ii), the steady-state distribution η_u for the fixed-user system with u users is given by

$$\eta\left(u, \boldsymbol{h}_1^{1,u}\right) = q^u\left(\boldsymbol{h}_1^{1,u}\right). \tag{19}$$

This is because the fixed-user system has no arrivals and departures, and has independent fading across users, with marginal fading distribution $q(\ell)$.

B. Case (i)

The independent service condition is easy to satisfy when fading is independent across time slots. Instead of initializing $X^\alpha[0]$ by the distribution η_u, as required in the definition of independent service, we will initialize $X^\alpha[0]$ at $x_0 \in \mathcal{X}_u$, and show independent service

$$\mathsf{Pr}_{x_0}(X^\alpha[1] = x|A_0 = 0, D_0 = 1) = \eta_{u-1}(x)$$

and stationary arrival state

$$\mathsf{Pr}_{x_0}(X^\alpha[1] = x|A_0 = 1, D_0 = 0) = \eta_{u+1}(x).$$

When there is no arrival, and one departure, from the independence of fading across time, it follows that the new state is drawn independently of the past, following the distribution q^{u-1}. Since η_{u-1} and q^{u-1} are identical, the independent service condition is established. Similarly, using the fact that an arriving user has fading levels distributed according to the marginal distribution q, the stationary arrival state condition can be verified. Note that in case (i), the requirements of nice separability are met exactly, with no $O(\alpha)$ error term.

C. Case (ii)

Initialize the state $X^\alpha[0]$ according to the steady-state distribution η_u. Then, the special structure of case (ii) gives us the following independent and identically distributed Markov chains (in the notation of [1]):

$$\begin{array}{ccc} h_{1,0} & \rightarrow & (D_{1,0}, h_{1,1}) \\ h_{2,0} & \rightarrow & (D_{2,0}, h_{2,1}) \\ \cdots & & \cdots \\ h_{u,0} & \rightarrow & (D_{u,0}, h_{u,1}). \end{array} \tag{20}$$

We will now evaluate the distribution of $h_{j,0}$ conditioned on the event that user j does not depart from the system at time 0

$$\mathsf{Pr}_{\eta_u}(h_{j,0} = \ell|D_{j,0} = 0).$$

This probability can be simplified, using Bayes rule, to give

$$\frac{\mathsf{Pr}_{\eta_u}(D_{j,0} = 0|h_{j,0} = \ell)\mathsf{Pr}_{\eta_u}(h_{j,0} = \ell)}{\sum_m \mathsf{Pr}(D_{j,0} = 0|h_{j,0} = m)\mathsf{Pr}_{\eta_u}(h_{j,0} = m)}.$$

The distribution of $D_{j,0}$ is given by the definition of dynamic-user systems, and we know, from the structure of the function f in case (ii), that

$$\mathsf{Pr}(D_{j,0} = 0|h_{j,0} = \ell) = 1 - f_j(\ell)\alpha + o(\alpha).$$

This, together with $\mathsf{Pr}_{\eta_u}(h_{j,0} = \ell) = q(\ell)$ reduces the probability of interest to

$$\mathsf{Pr}_{\eta_u}(h_{j,0} = \ell|D_{j,0} = 0) = \frac{[1 - f_j(\ell)\alpha + O(\alpha^2)]q(\ell)}{\sum_m [1 - f_j(m)\alpha + o(\alpha)]q(m)}.$$

Since $q(m)$ in the denominator sums to one, we get

$$\mathsf{Pr}_{\eta_u}(h_{j,0} = \ell|D_{j,0} = 0) = q(\ell) + O(\alpha). \tag{21}$$

With the aid of the preceding equation, we are ready to show that (12) and (13) are satisfied in case (ii).

To check (12), first consider the case when the departing user has index u. Then, we wish to compute the distribution of the channels of the remaining $u - 1$ users

$$\mathsf{Pr}_{\eta_u}\left(\boldsymbol{h}_1^{1,u-1}|A_1 = 0, D_{j,0} = \delta_{ju}\right)$$

where δ_{ij} equals one when $i = j$ and zero otherwise. From the independence structure in (20), it follows that the above expression is equal to

$$\prod_{j=1}^{u-1} \mathsf{Pr}_{\eta_u}(h_{j,1}|A_1 = 0, D_{j,0} = 0).$$

From the Markov evolution of the channel states, this is equal to

$$\prod_{j=1}^{u-1}\left[\sum_m Q(h_{j,1}|m)\mathsf{Pr}_{\eta_u}(h_{j,0} = m|A_1 = 0, D_{j,0} = 0)\right].$$

However, we know from (21) that

$$\mathsf{Pr}_{\eta_u}(h_{j,0} = m|A_1 = 0, D_{j,0} = 0)$$

is close to the steady-state distribution $q(m)$ of the transition matrix $Q(\cdot|\cdot)$. Thus

$$\sum_m Q(h_{j,1}|m)\mathsf{Pr}_{\eta_u}(h_{j,0} = m|A_1 = 0, D_{j,0} = 0) = q(h_{j,1}) + O(\alpha).$$

This gives the quantity of interest in the independent service condition as

$$\mathsf{Pr}_{\eta_u}\left(\boldsymbol{h}_1^{1,u-1}|A_1 = 0, D_{j,0} = \delta_{ju}\right) = q^{u-1}\left(\boldsymbol{h}_1^{1,u-1}\right) + O(\alpha).$$

Also, by symmetry, we get a similar result when the departing user has index i

$$\mathsf{Pr}_{\eta_u}\left(\boldsymbol{h}_1^{1,u-1}|A_1=0,D_{j,0}=\delta_{ji}\right)=q^{u-1}\left(\boldsymbol{h}_1^{1,u-1}\right)+O(\alpha). \tag{22}$$

To prove (12), we need to average the above expression over the index of the departing user i, and evaluate

$$\mathsf{Pr}_{\eta_u}\left(\boldsymbol{h}_1^{1,u-1}|A_1=0,D_1=1\right).$$

Now, the event $D_1=0$ can be written as a union of disjoint events as $\cup_{i=1}^{u}\{D_{j,0}=\delta_{ji}\}$. Further, from (22), we know that conditioned on each of these constituent events, the probability of interest is the same (up to $O(\alpha)$). Thus, we get

$$\mathsf{Pr}_{\eta_u}\left(\boldsymbol{h}_1^{1,u-1}|A_1=0,D_1=1\right)=q^{u-1}\left(\boldsymbol{h}_1^{1,u-1}\right)+O(\alpha).$$

To understand the step from (22) to the above equation more closely, observe that we have used simple conditional probability definitions, as illustrated below for dummy events F and $E=\cup_i E_i$

$$\mathsf{Pr}(F|E)=\frac{1}{\mathsf{Pr}(E)}\sum_i\mathsf{Pr}(F|E_i)\mathsf{Pr}(E_i).$$

When the terms $\mathsf{Pr}(F|E_i)$ are the same for all i, we get

$$\mathsf{Pr}(F|E)=\frac{1}{\mathsf{Pr}(E)}\mathsf{Pr}(F|E_1)\mathsf{Pr}(\cup_i E_i)=\mathsf{Pr}(F|E_1).$$

This completes verification of the independent service condition for case (ii).

To verify the stationary arrival state condition, first modify the above proof to show that when there are no departures, the channel states of the existing users are distributed according to q^u, i.e.,

$$\mathsf{Pr}_{\eta_u}(\boldsymbol{h}_1^{1,u}|D_0=0)=q^u(\boldsymbol{h}_1^{1,u})+O(\alpha).$$

Then, use the fact that the arriving user has state independent of the other users to show

$$\mathsf{Pr}_{\eta_u}\left(\boldsymbol{h}_1^{1,u}|D_0=0,A_0=1\right)=q^{u+1}\left(\boldsymbol{h}_1^{1,u+1}\right)+O(\alpha).$$

D. Case (iii)

In the case of TDMA, the state vector needs to be modified to

$$X^\alpha[k]=(u[k],\mathrm{TDM}[k],h_{1,k},\ldots,h_{u,k}) \tag{23}$$

where $\mathrm{TDM}[k]$ is the index of the user to be served in slot k. The evolution follows $\mathrm{TDM}[k+1]=\mathrm{TDM}[k]+1$ when $\mathrm{TDM}[k]<u[k+1]$, and $\mathrm{TDM}[k+1]=1$ otherwise.

The steady-state distribution η_u for TDMA is the distribution for the channel states q^u, multiplied by a uniform distribution between 1 and u for $\mathrm{TDM}[k]$

$$\eta_u(X[0])=q^u\left(\boldsymbol{h}_k^{1,u}\right)\cdot\frac{1}{u}.$$

For the independent service condition, we are required to show

$$\mathsf{Pr}_{\eta_u}(X^\alpha[0]=(u[k],\mathrm{TDM}[1],h_{1,1},\ldots,h_{u-1,1})|A_0=0,D_0=1)=\frac{q^{u-1}(\boldsymbol{h}_k^{1,u-1})}{u-1}. \tag{24}$$

We will evaluate this probability by first conditioning on $\mathrm{TDM}=u$, and showing

$$\mathsf{Pr}_{\eta_u}(h_{1,1},\ldots,h_{u-1,1}|A_0=0,D_0=1,\mathrm{TDM}[0]=j)=q^{u-1}\left(\boldsymbol{h}_k^{1,u-1}\right). \tag{25}$$

To prove the above, observe that for the TDMA case, only user TDM receives service. Thus, the event $\{A_0=0,D_0=1,\mathrm{TDM}[0]=j\}$ is independent of the channel states of the remaining users at time 0, i.e.,

$$\mathsf{Pr}_{\eta_u}(h_{i,0},i\neq j|A_0=0,D_0=1,\mathrm{TDM}[0]=j)=q^{u-1}(h_{i,0},i\neq j).$$

Also, we know from the channel-state transition rule (18) that when the channel is initialized in the steady state q, the channel after one time step is also in steady state. This proves (25). Also, under steady state, departure is equally likely to occur for each user, giving

$$\mathsf{Pr}_{\eta_u}(\mathrm{TDM}[1]|A_0=0,D_0=1,\mathrm{TDM}[0]=j)=\frac{\mathbb{1}_{\{0<\mathrm{TDM}[1]\leq u\}}}{u-1}.$$

Using the above equation, and (25), we can prove the desired result (24).

The stationary arrival state condition can be verified exactly as for case (ii). The state of the new arrival is distributed independent of other users, and thus the stationary arrival state condition (13) is satisfied for the channel state part of the state vector. The condition is also satisfied for TDM because the arriving user is inserted at a random position at arrival. This verifies the stationary arrival state condition for the entire state vector (23).

V. Proving Time-Scale Separation Results

The first-order approximation result (Theorem 1) can be proven by using the theory of time-scale decomposition (for an overview of time-scale decomposition, see [25]). Results similar to our second-order approximation (involving an $o(\alpha)$ approximation error) have been considered in [25], [27]–[31]. However, these references are devoted either to very specific problems, or to general numerical techniques, and are not readily applicable to our problem. For this reason, we do not use the standard time-scale decomposition theorems, but rather, develop a unified proof for the first- and second-order approximation results. Our proof technique focuses on the structure of the service process $\tilde{I}[k]$ and gives insight about the role of service variability.

Our proof relies on relating the dynamic-user system with the symmetric queue. First, we give a jump chain representation of the symmetric queue, and determine the holding times and transition probabilities for the jump chain. Then, we give two lemmas that link the holding times and transition probabilities of the symmetric queue with corresponding quantities for the dynamic-user system. Finally, we use the similarity of the holding times and transition probabilities of the two systems to prove the first- and second-order approximation theorems.

A. Jump Chain for the Symmetric Queue

As we argued in Section II-A, in a symmetric queue the number of users at time t forms a Markov process $U(t)$. The steady-state distribution of this Markov process is $\mu(u)$, and can be computed using (3). To prove the time-scale separation theorems, we give an alternative characterization of $\mu(u)$ in terms of a jump chain. Our treatment effectively involves treating the given Markov process as a semi-Markov process (which is defined in [19, Sec. 4.8]).

To distinguish the symmetric queue from the dynamic-user system, we will use the "bar" notation (for example, $\bar{U}(t)$) to denote quantities associated with the symmetric queue. Consider a symmetric queue with offered load Ω, average file size 1, and service rates $\phi(u)$. Let the jump instants (corresponding to arrivals or departures) be $\bar{t}_n$, and the jump chain be $\bar{U}_n$, i.e., $\bar{U}_n$ is the state immediately after the jump. Further, conditioned on $U(0) = u$, let the holding time in state u be denoted by the random variable $\bar{\tau}_u$. Then, $\bar{U}_n$ forms a jump Markov chain with transition probabilities given by matrix $\bar{P}$

$$\bar{P}(j|i) = \begin{cases} \frac{\Omega}{\Omega+\phi(u)}, & j = i+1, 0 < i < u_0 \\ \frac{\phi(u)}{\Omega+\phi(u)}, & j = i-1, 0 < i < u_0 \\ 1, & j = u_0 - 1,\ i = u_0 \\ 1, & j = 1,\ i = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

Further, the random variables $\bar{\tau}_u$ have an exponential distribution with mean $\bar{H}(u) = (\Omega + \phi(u))^{-1}$ for $0 < u < u_0$, and the means at the extreme points given by $\bar{H}_0 = \Omega^{-1}$ and $\bar{H}_{u_0} = \phi(u_0)^{-1}$. Let the steady-state distribution of $\bar{U}_n$ be $\bar{\gamma}$. Then, $\bar{\gamma}$ is the solution to

$$\bar{\gamma}(j) = \sum_i \bar{\gamma}(i)\bar{P}(j|i),$$
$$\sum_i \bar{\gamma}(i) = 1. \tag{27}$$

The theory of semi-Markov processes tells us that the steady-state distribution of the continuous-time Markov process $\bar{U}(t)$ is

$$\mu(i) = \frac{\bar{\gamma}(i)\bar{H}(i)}{\sum_j \bar{\gamma}(j)\bar{H}(j)}. \tag{28}$$

Further, let the holding times and transition probabilities for the modified symmetric queue be $\bar{H}_i^\alpha$ and $\bar{P}^\alpha(i|j)$, respectively, and let the distribution of the jump chain be $\bar{\gamma}^\alpha(u)$. Then, $\bar{\gamma}^\alpha$ satisfies

$$\bar{\gamma}^\alpha(j) = \sum_i \bar{\gamma}^\alpha(i)\bar{P}^\alpha(i|j)$$
$$\sum_i \bar{\gamma}^\alpha(i) = 1. \tag{29}$$

Also, μ^α satisfies

$$\mu^\alpha(u) = \frac{\bar{\gamma}^\alpha(u)\bar{H}^\alpha(u)}{\sum_j \bar{\gamma}^\alpha(j)\bar{H}^\alpha(j)}. \tag{30}$$

B. Holding Time and Transition Probability Lemmas

In this subsection, we give two lemmas that show that for separable dynamic-user systems, the sets $\mathcal{X}_u$ are analogous to the states u in the symmetric queue. In particular, the transition probabilities and the holding times of the sets $\mathcal{X}_u$ are similar to those for the symmetric queue.

Define the jump instants k_n to be the slots where an arrival or departure occurs, and define the jump chain $X_n^\alpha = X^\alpha[k_n + 1]$. Define the steady-state distribution of the jump chain to be $\gamma^\alpha(x)$, i.e.,

$$\gamma^\alpha(x) = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{\{X_n^\alpha = x\}}. \tag{31}$$

Next, we link the holding times in the symmetric queue and the given dynamic-user system. When the system is initialized at x, define random variable G^α to be the first time slot when an arrival or departure occurs. Then, G^α is the holding time in set $\mathcal{X}_u$, with initialization at x, and the mean holding time can be defined as $H^\alpha(x) \overset{\text{def}}{=} \mathrm{E}_x[\delta G^\alpha]$.

Now, consider the mean holding time $H^\alpha(\mathcal{X}_u)$ in set $\mathcal{X}_u$, when initialization is according to the jump chain distribution. By the following argument, $H^\alpha(\mathcal{X}_u)$ can be computed from $H^\alpha(x)$ and $\gamma^\alpha(x)$. Transitions into $\mathcal{X}_u$ take place according to the distribution of the jump chain γ^α. In particular, the probability of a jump into $\mathcal{X}_u$ reaching state x is $\gamma^\alpha(x)/\gamma^\alpha(\mathcal{X}_u)$. Thus, the mean holding time in the set $\mathcal{X}_u$ is the average of $H^\alpha(x)$ over the distribution $\gamma^\alpha(x)/\gamma^\alpha(\mathcal{X}_u)$, giving

$$H^\alpha(\mathcal{X}_u) = \frac{1}{\gamma^\alpha(\mathcal{X}_u)}\sum_{x\in\mathcal{X}_u}\gamma^\alpha(x)H^\alpha(x).$$

We show that after scaling by the file size, the mean holding time in each $x \in \mathcal{X}_u$ is roughly the same as the mean holding time in state u of the symmetric queue, i.e.,

$$\alpha H^\alpha(x) = \bar{H}_u + O(\alpha).$$

By definition of $H^\alpha(\mathcal{X}_u)$, this immediately implies

$$\alpha H^\alpha(\mathcal{X}_u) = \bar{H}_u + O(\alpha).$$

For the second-order approximation, we require a finer characterization of $H^\alpha(x)$. In particular, we are interested in a result that has $o(\alpha)$ accuracy. Unfortunately, the difference $\alpha H^\alpha(x) - \bar{H}_u$ is strictly $O(\alpha)$. However, we are able to show that for nicely separable systems, averaging across γ^α cancels the $O(\alpha)$ terms in the difference $\alpha H^\alpha(\mathcal{X}_u) - \bar{H}_u$. This result is stated in the following lemma.

Lemma 1: (Holding Times) For a separable system, with $x \in \mathcal{X}_u$, the mean holding time $H^\alpha(x)$ satisfies

$$\alpha H^\alpha(x) = \bar{H}(u) + O(\alpha) \tag{32}$$

and for a nicely separable system it satisfies

$$\alpha \sum_{x\in\mathcal{X}_u} \gamma^\alpha(x)H^\alpha(x) = \gamma^\alpha(\mathcal{X}_u)\bar{H}^\alpha(u) + \delta\bar{O}(\alpha). \tag{33}$$

Proof: See Section VI-C. □

Next, we are interested in the transition probability for the jump chain X_n^α. The transition probability is $\mathsf{Pr}_x\,(X_1^\alpha \in \mathcal{X}_j)$. As we argued for the holding time lemma, the quantity of interest is the mean transition probability from $\mathcal{X}_i$ into $\mathcal{X}_j$ when the initial state in $\mathcal{X}_i$ is distributed according to γ^α. We want this mean transition probability to be close to the transition probability $\bar{P}(j|i)$ of the separable system (26). The following lemma gives an $O(\alpha)$ accurate result for the mean transition probability for separable systems, and an $o(\alpha)$ result for the transition probability of nicely separable systems.

Lemma 2 (Transition Probability): For a separable system, with $x \in \mathcal{X}_i$, the transition probability $\mathsf{Pr}_x\,(X_1^\alpha \in \mathcal{X}_j)$ satisfies

$$\mathsf{Pr}_x\,(X_1^\alpha \in \mathcal{X}_j) = \bar{P}(j|i) + O(\alpha) \tag{34}$$

and for a nicely separable system it satisfies

$$\sum_{x\in\mathcal{X}_i} \gamma^\alpha(x)\mathsf{Pr}_x\,(X_1^\alpha \in \mathcal{X}_j) = \gamma^\alpha(\mathcal{X}_i)\bar{P}^\alpha(j|i)+\delta_d\bar{O}(\alpha). \tag{35}$$

Proof: See Section VI-D. □

With these two lemmas, we are ready to prove the first- and second-order approximation results.

C. Proving Time-Scale Separation

The proof of time-scale separation relies on the following proposition about the jump chain X_n^α:

Proposition 1: For any set $A \subset \mathcal{X}$, the steady-state distribution γ^α and the transition probabilities $\mathsf{Pr}_x\,(X_1^\alpha \in A)$ are linked by the flow balance relation

$$\gamma^\alpha(A) = \sum_i \sum_{x\in\mathcal{X}_i} \gamma^\alpha(x)\mathsf{Pr}_x(X_1^\alpha \in A). \tag{36}$$

Further, the steady-state distribution π^α of $X^\alpha[k]$ is linked with the mean holding times $H^\alpha(x)$ by

$$\pi^\alpha(\mathcal{X}_j) = \frac{\sum_{x\in\mathcal{X}_j}\gamma^\alpha(x)H^\alpha(x)}{\sum_i\sum_{x\in\mathcal{X}_i}\gamma^\alpha(x)H^\alpha(x)}. \tag{37}$$

Proof: Initialize $X^\alpha[0]$ according to the jump chain's stationary distribution γ^α. The flow balance relation (36) follows immediately from the stationarity of the jump chain distribution γ^α.

To see (37), let $X^\alpha[0]$ be initialized at some x, and let the jump chain evolve over N jumps. Further, let x be visited $N(x)$ times. Then

$$\lim_{N\to\infty} \frac{N(x)}{N} = \gamma^\alpha(x).$$

Further, the jump chain is renewed at each visit to x, and thus the holding times following each visit to x are independent. Let the total holding time after visits to state x be $T(x)$. Then, by the law of large numbers, $T(x)$ satisfies

$$\lim_{N\to\infty} \frac{T(x)}{N(x)} = H^\alpha(x)$$

or equivalently

$$\lim_{N\to\infty} \frac{T(x)}{N} = \gamma^\alpha(x)H^\alpha(x).$$

Further, the total time spent in $\mathcal{X}_u$ is $T(\mathcal{X}_u) = \sum_{x\in\mathcal{X}_u} T(x)$ and satisfies

$$\lim_{N\to\infty} \frac{T(\mathcal{X}_u)}{N} = \sum_{x\in\mathcal{X}_u} \gamma^\alpha(x)H^\alpha(x).$$

The quantity of interest in (37) is $\pi^\alpha(\mathcal{X}_u)$, which is the fraction of time $X^\alpha[k]$ spends in $\mathcal{X}_u$ in the limit of large N. This fraction of time can be computed from the above equation, proving (37). □

With the above proposition, and the holding time and transition probability lemmas, we are ready to prove the first-order approximation result.

Proof of Theorem 1: Consider the flow balance relation in (36), with $A = \mathcal{X}_j$. The probability $\mathsf{Pr}_x\,(X_1^\alpha \in \mathcal{X}_j)$ in this equation is given by the transition probability lemma. Using the fact that the sum of a finite number of $O(\alpha)$ terms remains $O(\alpha)$, (36) simplifies to

$$\gamma^\alpha(\mathcal{X}_j) = \sum_i \gamma^\alpha(\mathcal{X}_i)\bar{P}(j|i) + O(\alpha).$$

We also know that

$$\sum_i \gamma^\alpha(\mathcal{X}_i) = 1.$$

Thus, $\gamma^\alpha(\mathcal{X}_i)$ satisfies the same set of linear equations as $\bar{\gamma}(i)$ in (27), with the exception of the $O(\alpha)$ term. This leads to

$$\gamma^\alpha(\mathcal{X}_i) = \bar{\gamma}(i) + O(\alpha). \tag{38}$$

Next, consider property (37) of the jump chain. Using the holding time lemma (Lemma 1), and the fact that the linear combination of a finite number of $O(\alpha)$ remains $O(\alpha)$, we can simplify (37) to

$$\pi^\alpha(\mathcal{X}_j) = \frac{\gamma^\alpha(\mathcal{X}_j)\bar{H}(j) + O(\alpha)}{\sum\limits_i \gamma^\alpha(\mathcal{X}_i)\bar{H}(i) + O(\alpha)}. \tag{39}$$

Combining the above equation with (38) and the property (28) of $\mu(u)$ gives the desired result

$$\pi^\alpha(\mathcal{X}_j) = \mu(j) + O(\alpha).$$ □

Proof of Theorem 2: The method of proof is similar to the proof for the first-order approximation. Combining the flow balance relation (36) with the transition probability lemma for the nicely separable system gives

$$\gamma^\alpha(\mathcal{X}_j) = \sum_i \gamma^\alpha(\mathcal{X}_i)\bar{P}^\alpha(i|j) + \delta_d\bar{O}(\alpha). \tag{40}$$

In addition, γ^α sums to one, showing that $\gamma^\alpha(\mathcal{X}_i)$ satisfies a system of linear equations similar to (29) for the modified symmetric queue, giving

$$\gamma^\alpha(\mathcal{X}_j) = \bar{\gamma}^\alpha(j) + \delta_d \bar{O}(\alpha). \tag{41}$$

The following comment about the $\bar{O}(\alpha)$ term is in order here. Recall that $\bar{O}(\alpha)$ was defined earlier as an $O(\alpha)$ term where the constants depend only on the symmetric queue. In solving the system of linear (40), the error term $\delta_d\bar{O}(\alpha)$ is multiplied by the inverse of the matrix $(I-P)$. However, this matrix depends only on the symmetric user system, and thus, the result of the multiplication remains $\delta\bar{O}(\alpha)$.

Continuing with the proof of Theorem 2, the quantity of interest, $\pi^\alpha(\mathcal{X}_j)$, is given by (37). The holding time lemma reduces (37) to

$$\pi^\alpha(\mathcal{X}_j) = \frac{\gamma^\alpha(\mathcal{X}_j)\bar{H}^\alpha(j) + \delta_d\bar{O}(\alpha)}{\sum_i \gamma^\alpha(\mathcal{X}_i)\bar{H}^\alpha(i) + \delta_d\bar{O}(\alpha)}.$$

Using (41) simplifies this to

$$\pi^\alpha(\mathcal{X}_j) = \frac{\bar{\gamma}^\alpha(j)\bar{H}^\alpha(j) + \delta_d\bar{O}(\alpha)}{\sum_i \bar{\gamma}^\alpha(j)(\bar{H}^\alpha(i) + \delta_d\bar{O}(\alpha))}.$$

Comparing this with property (30) of the symmetric queue gives the desired result

$$\pi^\alpha(\mathcal{X}_j) = \mu^\alpha(j) + \delta_d\bar{O}(\alpha). \qquad \square$$

VI. Proving Holding Time and Transition Probability Lemmas

This section is dedicated to proving the holding time and transition probability lemmas. The proofs rely on a link between dynamic-user and fixed-user systems, which is outlined in Section VI-A, and multiplicative ergodic theory, which is summarized in Section VI-B. The proofs of the lemmas themselves are given in Section VI-C and VI-D.

A. Link Between Dynamic-User and Fixed-User Systems

The holding time lemma deals with G^α, the holding time. In this subection, we show that the distribution of G^α is linked to the service process in the fixed-user system. First, consider a simple model where the service process is constant rate, with service $\delta\phi(u)/u$ for each user in each slot when the number of users is u. Then, the probability of $G^\alpha \geq k$ turns out to be simply the probability that there are no arrivals and departures in slots 0 through k. This probability can be written as

$$\mathsf{Pr}(G^\alpha \geq k) = e^{-\alpha\phi(u)\delta k} e^{-\alpha\Omega\delta k}.$$

However, for the general service model we are interested in, the preceding expression will not be valid. The probability of interest will depend on the service statistics and on the initial state. This dependence takes the following form.

Proposition 2 (Departure Probability): For a given dynamic-user system, the distribution of the holding time G^α is given by

$$\mathsf{Pr}_{x_0}(G^\alpha \geq k) = \mathsf{E}_{x_0}\left[e^{-\alpha\tilde{I}[k]}\right] \cdot e^{-\alpha\Omega\delta k}.$$

Proof: We use the following notation: $A_{[0,k-1]}$ and $D_{[0,k-1]}$, respectively, are the number of arrivals and departures in slots 0 to $k-1$. The event $\mathcal{A}_{[0,k-1]}$ denotes $A_{[0,k-1]} = 0$, and $\mathcal{D}_{[0,k-1]}$ denotes $D_{[0,k-1]} = 0$. Further, for $x \in \mathcal{X}_u$, we define $f(x) = \sum_{j\leq u} f_j(x)$, making $f(X^\alpha[k])$ the sum of the service provided to all users in slot k.

The probability of interest can be simplified, using the definition of G^α, as

$$\mathsf{Pr}_{x_0}(G^\alpha \geq k) = \mathsf{Pr}_{x_0}(\mathcal{D}_{[0,k]}|\mathcal{A}_{[0,k-1]})\mathsf{Pr}_{x_0}(\mathcal{A}_{[0,k-1]}).$$

The second probability can be evaluated using the Poisson arrivals property, and to prove the proposition, it remains to be shown that

$$\mathsf{Pr}_{x_0}(\mathcal{D}_{[0,k-1]}|\mathcal{A}_{[0,k-1]}) = \mathsf{E}_{x_0}\left[e^{-\alpha\tilde{I}[k]}\right]. \tag{42}$$

Let the initial state x_0 belong to $\mathcal{X}_u$. Then, given $\mathcal{A}_{[0,k-1]}$, the event $\mathcal{D}_{[0,k-1]}$ is equivalent to the event that $X^\alpha[n]$ are in $\mathcal{X}_u$ for all $n = 1,\ldots,k$, giving

$$\mathsf{Pr}_{x_0}(\mathcal{D}_{[0,k-1]}|\mathcal{A}_{[0,k-1]}) = \mathsf{Pr}_{x_0}(X^\alpha[n] \in \mathcal{X}_u \forall\, n \leq k|\mathcal{A}_{[0,k-1]}).$$

Let $\boldsymbol{x} = (x_1,\ldots,x_k)$ denote a k-dimensional vector in $\mathcal{X}^k$, and let $\mathcal{B}$ denote the k-fold product of $\mathcal{X}_u$ with itself. Further, let F_k denote the distribution of $(X^\alpha[1],\ldots X^\alpha[k])$ conditioned on $X^\alpha[0] = x_0$ and $\mathcal{A}_{[0,k-1]}$

$$F_k(\boldsymbol{x}) = \mathsf{Pr}_{x_0}(\boldsymbol{X}^\alpha = \boldsymbol{x}|\mathcal{A}_{[0,k-1]}).$$

Then, the probability of interest is

$$\mathsf{Pr}_{x_0}(\mathcal{D}_{[0,k-1]}|\mathcal{A}_{[0,k-1]}) = \sum_{\boldsymbol{x}\in\mathcal{B}} F_k(\boldsymbol{x}). \tag{43}$$

Recall that $\mathcal{D}_{[0,k-1]}$ is the probability of no departures occurring during times 0 through $k-1$, while $\mathcal{A}_{[0,k-1]}$ is the probability of no arrivals occurring during times 0 through $k-1$. The distribution F_k can be constructed from the description of state transition probabilities for the Markov chain $X^\alpha[k]$ (see (9)). Consider the special case of transition from $x_{n-1} \in \mathcal{X}_u$ to $x_n \in \mathcal{X}_u$, conditioned on there being no arrivals. The probability of this transition is given by part 5 of the description of dynamic-user systems. When there are no arrivals, both x_{n-1} and x_n can belong to $\mathcal{X}_u$ only when there is no departure in slot $n-1$. This gives

$$\begin{aligned}
&\mathsf{Pr}(X^\alpha[n] = x_n|X^\alpha[n-1] = x_{n-1}, \mathcal{A}_{[0,k-1]})\\
&\quad = \mathsf{Pr}(X^\alpha[n] = x_n, \mathcal{D}_0|X^\alpha[n-1] = x_{n-1}, \mathcal{A}_{[0,k-1]})\\
&\quad = \mathsf{K}(x_n|x_{n-1},0,\ldots,0)\\
&\qquad \cdot \mathsf{Pr}(\mathcal{D}_0|\mathcal{A}_{[0,k-1]}, X^\alpha[n-1] = x_{n-1}).
\end{aligned}$$

The probability of no departure in slot $n-1$ is

$$\mathsf{Pr}(D_{n-1} = 0|\mathcal{A}_{[0,k-1]}, X^\alpha[n-1] = x_{n-1}) = e^{-\alpha f(x_{n-1})}.$$

This gives the transition probability

$$\begin{aligned}
&\mathsf{Pr}(X^\alpha[n] = x_n|X^\alpha[n-1] = x_{n-1}, \mathcal{A}_{[0,k-1]})\\
&\qquad\qquad = \mathsf{K}(x_n|x_{n-1},0,0,\ldots,0)e^{-\alpha f(x_{n-1})}.
\end{aligned}$$

Now, observe that the transition probability above is the same as the transition probability for a fixed-user system (Definition 1) multiplied by a correction term

$$\begin{aligned}&\mathsf{Pr}(X^\alpha[n] = x_n | X^\alpha[n-1] = x_{n-1}, A_{n-1} = 0)\\ &\qquad = \mathsf{Pr}(\tilde{X}[n] = x_n | \tilde{X}[n-1] = x_{n-1}) e^{-\alpha f(x_{n-1})}.\end{aligned}$$

For a fixed-user system initialized at x_0, let $\tilde{F}_k$ denote the distribution of $(\tilde{X}[1], \ldots \tilde{X}[k])$. For $\boldsymbol{x} \in \mathcal{B}$, the above equation gives

$$F_k(\boldsymbol{x}) = \tilde{F}_k(\boldsymbol{x}) \prod_{n=0}^{k-1} e^{-\alpha f(x_n)}. \tag{44}$$

This simplifies (43) to

$$\mathsf{Pr}_{x_0}(\mathcal{D}_{[0,k-1]} | \mathcal{A}_{[0,k-1]}) = \sum_{\boldsymbol{x} \in \mathcal{B}} \tilde{F}_k(\boldsymbol{x}) \prod_{n=0}^{k-1} e^{-\alpha f(x_n)}.$$

Writing the summation as an expectation, and using the definition of the cumulative service process $\tilde{I}[k]$ (Definition 1) verifies (42). This completes the proof of Proposition 2. □

B. Multiplicative Ergodicity Basics

In Proposition 2 in the previous section, we argued that departures in the dynamic-user system are characterized by the fixed-user system service process $\tilde{I}[k]$, through the quantity

$$\mathsf{E}_x \left[e^{-\alpha \tilde{I}[k]} \right]. \tag{45}$$

The service process $\tilde{I}[k]$ is determined by the fixed-user Markov chain $\tilde{X}$ with state transition probability matrix $\check{\mathsf{K}}$ with $\tilde{I}[k] = \sum_{n=0}^{k-1} f(\tilde{X}[k])$.

The theory of multiplicative ergodicity deals with the behavior of the quantity (45). Results on multiplicative ergodicity are given by Balaji and Meyn [24], and Kontoyiannis and Meyn [32]. For our application, the technical conditions required in [24], [32] will be simplified considerably due to our assumption that $\mathcal{X}_u$ is finite. The results required for proving the holding time and transition probability lemmas are given below.

The log moment-generating function of $\tilde{I}_j$ is defined by

$$\Lambda(-\alpha) = \lim_{k \to \infty} \frac{1}{k} \log \mathsf{E} \left[e^{-\alpha \tilde{I}[k]} \right].$$

This result suggests that we can approximate the quantity of interest as

$$\mathsf{E}_x \left[e^{-\alpha I[k]} \right] \approx e^{\Lambda(-\alpha)k}.$$

A more accurate form of the above approximation is given by [32, Theorem 4.1], which says that there are constants B_0 and $b_0 > 0$, and a function $\check{f}_\alpha(x)$ such that

$$\left| \mathsf{E}_x \left[e^{-\alpha \tilde{I}[k-1]} \right] \cdot e^{-k\Lambda(-\alpha)} - \check{f}_\alpha(x) \right| \le B_0 |\alpha| e^{-b_0 k}. \tag{46}$$

The function $\check{f}_\alpha(x)$ precisely determines the dependence of the quantity of interest on the initial state x.

The time-scale separation argument deals with the limit of a small α. Next, we describe $\Lambda(-\alpha)$ and $\check{f}_\alpha(x)$ for small values of α. From [24, Theorem 6.2], the derivatives of $\Lambda(\alpha)$ at 0 are related to the service process through the mean service rate by (see Definition 2)

$$\frac{d}{d\alpha} \Lambda(\alpha)|_{\alpha=0} = \sum_{x \in \mathcal{X}_u} \eta_u(x) f(x) = \delta \phi(u)$$

and to the variance by (see (11))

$$\frac{d^2}{d\alpha^2} \Lambda(\alpha)|_{\alpha=0} = \mathrm{Var}_{\eta_u}[f] = \delta \sigma^2(u).$$

These properties of $\Lambda(-\alpha)$ can be combined to get

$$\Lambda(-\alpha) = -\phi(u)\delta\alpha + \delta\sigma^2(u)\alpha^2/2 + O(\alpha^3). \tag{47}$$

It can also be verified [32, Proposition 4.9] that the derivative of $\check{f}_\alpha(x)$ with respect to α at $\alpha = 0$ is well defined. Let this derivative be denoted by $\hat{F}(x)$. Also, it can be verified easily from the definition of $\check{f}_\alpha(x)$ in [32, Theorem 4.1] that $\check{f}_0(x) = 1$, giving

$$\check{f}_\alpha(x) = 1 + \alpha \hat{F}(x) + O(\alpha^2). \tag{48}$$

This completes the description of $\check{f}_\alpha$ and $\Lambda(-\alpha)$ for small values of α.

The function $\hat{F}(x)$ is a solution to Poisson's equation

$$\sum_{y \in \mathcal{X}_u} \tilde{K}(y|x) \hat{F}(y) = \hat{F}(x) - f(x) + \sum_{y \in \mathcal{X}_u} \eta_u(y) f(y) \tag{49}$$

and is known as the relative value function. Further, from [32, Proposition 4.9], $\hat{F}$ has the property

$$\sum_{x \in \mathcal{X}_u} \eta(x) \hat{F}(x) = 0 \tag{50}$$

and

$$\lim_{k \to \infty} \mathsf{E}_x \left[\tilde{I}[k] - k\delta\phi(u) \right] = \hat{F}(x). \tag{51}$$

Next, we describe a result that allows us to generalize (46). From [32, Theorem 4.1], for a given function $g : \mathcal{X} \to \Re^+$

$$\begin{aligned}&\left| \mathsf{E}_x \left[e^{-\alpha \tilde{I}[k-1]} g(X^\alpha[k]) \right] \cdot e^{-k\Lambda(-\alpha)} - \check{\mu}_\alpha(g) \check{f}_\alpha(x) \right|\\ &\qquad\qquad \le B_0 \sup_x g(x) e^{-b_0 k}\end{aligned} \tag{52}$$

where $\check{\mu}_\alpha$ is a distribution on $\mathcal{X}_u$ that satisfies ([32, Proposition 4.9])

$$\check{\mu}_\alpha(x) = \eta_u(x) + \alpha \mu'(x) + O(\alpha^2).$$

With these results on multiplicative ergodicity, we are ready to prove the holding time and transition probability lemmas.

C. Holding Times

The holding time lemma for nicely separable systems involves the steady-state distribution γ^α of the jump chain. In proving the the second part of the lemma (relating to nicely separable systems), we will use the following result about the structure of $\gamma^\alpha(x)$.

Lemma 3 (Jump Chain Distribution): For a nicely separable system, and a given $x \in \mathcal{X}_u$

$$\gamma^\alpha(x) = \eta_u(x)\gamma^\alpha(\mathcal{X}_u) + O(\alpha).$$

Proof: See Section VI-E. □

Proof of Lemma 1: The quantity of interest in the holding time lemma is $\alpha H^\alpha(x)$, which is defined to be $\alpha\delta \mathsf{E}_x[G^\alpha]$. The expected value of G^α is given by $\sum_{k\geq 1} \mathsf{Pr}(G^\alpha \geq k)$. Thus, to prove the lemma, it will be enough to show the following two results:

$$\alpha\delta \sum_{k=1}^{\infty} \mathsf{Pr}_{x_0}(G^\alpha \geq k) = \frac{1}{\Omega + \phi(u)} + O(\alpha). \qquad (53)$$

Further, for a nicely separable system, it will be enough to prove

$$\sum_{x\in\mathcal{X}_u} \gamma^\alpha(x) \left[\alpha\delta \sum_{k=1}^{\infty} \mathsf{Pr}_x(G^\alpha \geq k)\right] = \frac{\gamma(\mathcal{X}_u)}{\Omega + \phi^\alpha(u)} + \delta\bar{O}(\alpha). \qquad (54)$$

To prove the above two equations, observe that Proposition 2 shows that

$$\mathsf{Pr}_x(G^\alpha \geq k) = \mathsf{E}_x\left[e^{-\alpha\tilde{I}[k-1]}\right] e^{-\alpha\Omega\delta k}. \qquad (55)$$

Multiplying both sides of the multiplicative ergodicity result (46) by $e^{k\Lambda(-\alpha)}$ and $e^{-k\Omega\alpha\delta}$ gives

$$\left|\mathsf{E}_x\left[e^{-\alpha\tilde{I}[k-1]-\alpha\Omega\delta k}\right] - \check{f}_\alpha(x)e^{k\Lambda(-\alpha)-\alpha\Omega\delta k}\right| \leq B_0|\alpha|e^{-b_0 k}e^{k\Lambda(-\alpha)-\alpha\Omega\delta k}. \qquad (56)$$

Consider the right-hand side of the above inequality. Since $b_0 > 0$ is a constant, and from (47), $\Lambda(-\alpha)$ is negative for small enough α, we have

$$\sum_{k=0}^{\infty} B_0|\alpha|e^{-b_0 k}e^{k\Lambda(-\alpha)-\alpha\Omega\delta k} \leq \alpha B_0(1-e^{-b_0})^{-1} = O(\alpha).$$

Further, the second term on the left of (56) can be summed to give

$$\sum_{k=0}^{\infty} \check{f}_\alpha(x)e^{k\Lambda(-\alpha)-\alpha\Omega\delta k} = \frac{\check{f}_\alpha(x)}{1-e^{\Lambda(-\alpha)-\Omega\alpha\delta}}.$$

By combining the above two equations, the quantity of interest in (53) can be written as

$$\alpha\delta \sum_{k=1}^{\infty} \mathsf{E}_x\left[e^{-\alpha\tilde{I}[k-1]}e^{-\alpha\Omega\delta k}\right] = \frac{\check{f}_\alpha(x)\alpha\delta}{1-e^{\Lambda(-\alpha)-\Omega\alpha\delta}} + \delta O(\alpha^2).$$

Using the expansion of $\Lambda(-\alpha)$ in (47) and the expansion of $\check{f}_\alpha$ in (48), this can be simplified to

$$\alpha\delta \sum_{k=1}^{\infty} \mathrm{E}_x\left[e^{-\alpha\tilde{I}[k-1]}e^{-\alpha\Omega\delta k}\right] = \frac{1}{\phi(u) - u\sigma^2\alpha/2 + \Omega} - \frac{\alpha\hat{F}(x)}{\phi(u)+\Omega} + O(\alpha^2) + \frac{1}{2}\delta\alpha. \qquad (57)$$

Equation (53) immediately follows from the above.

To verify (54), we only need to consider the first two terms in the above equation (the last two terms are negligible). From the definition of $\gamma^\alpha(x)$ it immediately follows that

$$\sum_{x\in\mathcal{X}_u} \frac{\gamma^\alpha(x)}{\phi(u) - u\sigma^2\alpha/2 + \Omega} = \frac{\gamma(\mathcal{X}_u)}{\phi^\alpha(u) + \Omega}.$$

Also, from the property of γ^α in the jump chain distribution lemma, and the property of the relative value function $\hat{F}$ in (50), it follows that

$$\sum_{x\in\mathcal{X}_u} \frac{\gamma^\alpha(x)\alpha\hat{F}(x)}{\phi(u)+\Omega} = O(\alpha^2).$$

This completes the proof of (54). □

D. Transition Probabilities

We are interested in the probability $\mathsf{Pr}_x\left\{X_1^\alpha \in \mathcal{X}_j\right\}$ where $x \in \mathcal{X}_i$. For $|i-j| = 0$ and $|j-i| \geq 2$, we know that $\bar{P}^\alpha(i|j) = 0$, and proving the transition probability lemma requires showing that $\mathsf{Pr}_x\left\{X_1^\alpha \in \mathcal{X}_j\right\}$ is small. This is established through the following lemma, which says that the probability of two or more events in a slot is negligible.

Lemma 4 (Negligible Events): For any initial state x, the probability of two or more arrival and departure events happening in slot G^α is negligible

$$\mathsf{Pr}_x(A_{G^\alpha} + D_{G^\alpha} \geq 2) = \delta_d\bar{O}(\alpha).$$

Proof: See Appendix II □

From the preceding lemma, it follows that

$$\sum_{j:|j-i|\geq 2, j=i} \mathsf{Pr}_x\{X_1^\alpha \in \mathcal{X}_j\} = \delta_d\bar{O}(\alpha). \qquad (58)$$

To prove the transition probability lemma, it remains to prove (34) and (35) for $j = i+1$ and $j = i-1$.

The case of $j = i+1$: We are interested in the probability $\mathsf{Pr}_{x_0}\left\{X_1^\alpha \in \mathcal{X}_j\right\}$. To simplify the calculation, we consider instead, the probability $\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1)$. We make use of the fact that the probability of two or more events in a slot is negligible, to get

$$\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1) = \mathsf{Pr}_{x_0}(X_1^\alpha \in \mathcal{X}_j) + \delta_d\bar{O}(\alpha).$$

Then, proving (34) requires us to show

$$\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1) = \bar{P}(i+1|i) + O(\alpha) \qquad (59)$$

and proving (35) requires us to show

$$\sum_{x\in\mathcal{X}_i} \gamma(x)\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1) = \gamma^\alpha(\mathcal{X}_i)\bar{P}(i+1|i) + \delta_d\bar{O}(\alpha). \qquad (60)$$

Next, we simplify $\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1)$. By definition of G^α,

$$\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1) = \mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1, \mathcal{D}_{[0,G^\alpha-1]}, \mathcal{A}_{[0,G^\alpha-1]}).$$

Expanding this out by conditioning over all possible values of G^α gives

$$\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1) = \sum_{k=0}^{\infty} \mathsf{Pr}_{x_0}(A_k \geq 1 | \mathcal{D}_{[0,k-1]}, \mathcal{A}_{[0,k-1]}) \cdot \mathsf{Pr}_{x_0}(\mathcal{D}_{[0,k-1]}, \mathcal{A}_{[0,k-1]}). \quad (61)$$

However, the arrival process is independent of the past, giving

$$\mathsf{Pr}_{x_0}(A_k \geq 1 | \mathcal{D}_{[0,k-1]}, \mathcal{A}_{[0,k-1]}) = (1 - \exp(-\Omega\delta\alpha)).$$

Substituting back in (61), and noting that the term above can be reduced to $\Omega\delta\alpha + \delta^2\bar{O}(\alpha^2)$ gives

$$\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1) = \left(\Omega\delta\alpha + \delta^2\bar{O}(\alpha^2)\right) \cdot \sum_{k=0}^{\infty} \mathsf{Pr}_{x_0}(\mathcal{D}_{[0,k-1]}, \mathcal{A}_{[0,k-1]}).$$

Combining this with (53) gives

$$\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1) = \frac{\Omega}{\Omega + \phi(u)} + O(\alpha)$$

which proves (59).

Further, applying (54) gives

$$\sum_{x \in \mathcal{X}_i} \gamma^\alpha(x) \mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1) = \gamma^\alpha(\mathcal{X}_u) \frac{\Omega}{\Omega + \phi^\alpha(u)} + \Omega\delta_d\bar{O}(\alpha)$$

which proves (60). □

The case of $j = i - 1$*:* We have already established (34) and (35) for all $j \neq i - 1$. The first-order result for $j = i - 1$ follows immediately by combining result (58) with the fact that the probabilities must sum to one (over j). The probability of interest for the second-order result can be written as

$$\sum_{x \in \mathcal{X}_i} \gamma^\alpha(x) \mathsf{Pr}_x(X_1^\alpha \in \mathcal{X}_{i-1}) = \sum_{x \in \mathcal{X}_i} \gamma^\alpha(x) \left(1 - \sum_{j \neq i-1} \mathsf{Pr}_x(X_1^\alpha \in \mathcal{X}_j)\right).$$

Using (58) and the result for the case $j = i + 1$ (which we have already proved) gives the desired result (35) for $j = i - 1$. This completes the proof of the transition probability lemma. □

E. Finer Structure of the Jump Chain Distribution

In this subsection, we prove the jump chain distribution lemma. This lemma is applicable exclusively to nicely separable systems, and is not used in the proof of the first-order result. For nicely separable systems, we wish to show for $x \in \mathcal{X}_u$, that

$$\gamma^\alpha(x) = \eta_u(x)\gamma^\alpha(\mathcal{X}_u) + O(\alpha).$$

The proof of the lemma involves three steps. The first step is to find the distribution of the state $X^\alpha[k]$ for $k < G^\alpha$, i.e., the states before transition. The second step is the use the nicely separable condition to find the distribution of the state after transition. The third step is to tie the above two results together to prove the jump chain distribution lemma.

For the first step, we show that for $k < G^\alpha$, the probability $\mathsf{Pr}_{x_0}(X^\alpha[k])$ is nearly the same as the fixed-user system's steady-state distribution η_u. This is stated in the following proposition.

Proposition 3: Conditioned on there being no arrivals and departures for $n \leq k$, the distribution of $X^\alpha[k]$ is close to η_u, i.e.,

$$\mathsf{Pr}_{x_0}(X^\alpha[k] = x | \mathcal{D}_{[0,k-1]}, \mathcal{A}_{[0,k-1]}) = \eta_u(x) + O(\alpha) + B_0 e^{-b_0 k}. \quad (62)$$

Proof: The result follows from (52). In the notation of the theory of multiplicative ergodicity, the result follows from the fact that the "twisted transition probability" or the taboo probability is nearly the same as the original transition probability for small values of α. See Appendix I for details of the proof. □

Proposition 3 says that the state before a transition is nearly distributed according to $\eta_u(x)$. For a nicely separable system, we know that if the state before a transition event is distributed according to $\eta_u(x)$, then the state after transition is distributed according to $\eta_{u+1}(x)$ or $\eta_{u-1}(x)$ depending on whether the transition event was an arrival or departure. Thus, we argue that irrespective of the initial state x_0, the state after transition is distributed according to $\eta(x)$. This is formalized in the following proposition.

Proposition 4: For $x \in \mathcal{X}_u$ and x_0 in either $\mathcal{X}_{u-1}$ or $\mathcal{X}_{u+1}$

$$\mathsf{Pr}_{x_0}(X^\alpha[G^\alpha + 1] = x | X^\alpha[G^\alpha + 1] \in \mathcal{X}_u) = \eta_u(x) + O(\alpha). \quad (63)$$

Proof: First, consider the case when $x_0 \in \mathcal{X}_{u+1}$, and the event of interest is a downward transition. The method of proof is identical for the other case, when $x_0 \in \mathcal{X}_{u-1}$, and the event of interest is an upward transition. From the Markov structure of $X^\alpha[k]$, and the definition of the jump times G^α, the above probability can be reduced to

$$\begin{aligned}&\mathsf{Pr}_{x_0}\left(X^\alpha[G^\alpha + 1] = x, G^\alpha < \frac{1}{\sqrt{\alpha}} | X^\alpha[G^\alpha + 1] \in \mathcal{X}_u\right) \\ &\cdot \mathsf{Pr}(G^\alpha < \frac{1}{\sqrt{\alpha}} | X^\alpha[G^\alpha + 1] \in \mathcal{X}_u) \\ &+ \sum_{k \geq 1/\sqrt{\alpha}} \Bigg(\mathsf{Pr}_{x_0}(G^\alpha = k | X^\alpha[k+1] \in \mathcal{X}_u) \\ &\cdot \mathsf{Pr}_{x_0}(X^\alpha[k+1] = x | X^\alpha[k+1] \in \mathcal{X}_u, G^\alpha = k)\Bigg). \quad (64)\end{aligned}$$

Since the mean file size is $1/\alpha$, and the arrival rate is $\alpha\Omega$, it can be argued that

$$\mathsf{Pr}_{x_0}\left(G^\alpha < \frac{1}{\sqrt{\alpha}} | X^\alpha[k+1] \in \mathcal{X}_u\right) = O(\alpha). \quad (65)$$

For a more detailed argument, see the proof of the first part of the holding time lemma. Further, we will prove for $k \geq 1/\sqrt{\alpha}$ that

$$\mathsf{Pr}_{x_0}(X^\alpha[k+1] = x | X^\alpha[k+1] \in \mathcal{X}_u, G^\alpha = k) = \eta_u(x) + O(\alpha). \quad (66)$$

This equation, together with (65) will complete the proof of (63). To prove (66), observe that from the negligible discreteness condition, it will be enough to show

$$\mathsf{Pr}_{x_0}(X^\alpha[k+1] = x | \mathcal{A}_k, D_k = 1, G^\alpha = k) = \eta_u(x) + O(\alpha). \quad (67)$$

Further, by definition G^α, we have

$$\begin{aligned}&\mathsf{Pr}_{x_0}(X^\alpha[k+1]=x|\mathcal{A}_k, D_k=1, G^\alpha=k)\\ &\quad=\mathsf{Pr}_{x_0}(X^\alpha[k+1]=x|\mathcal{A}_k, D_k=1,\\ &\qquad \mathcal{A}_{[0,k-1]}, \mathcal{D}_{[0,k-1]}).\end{aligned} \tag{68}$$

This can be simplified to

$$\frac{\mathsf{Pr}_{x_0}(X^\alpha[k+1]=x, \mathcal{A}_k, D_k=1|\mathcal{A}_{[0,k-1]}, \mathcal{D}_{[0,k-1]})}{\mathsf{Pr}_{x_0}(\mathcal{A}_k, D_k=1|\mathcal{A}_{[0,k-1]}, \mathcal{D}_{[0,k-1]})}. \tag{69}$$

The numerator can be simplified further as

$$\begin{aligned}&\sum_{x_1}(\mathsf{Pr}(X^\alpha[k+1]=x, \mathcal{A}_k, D_k=1|X^\alpha[k]=x_1)\\ &\qquad\cdot\mathsf{Pr}_{x_0}(X^\alpha[k]=x_1|\mathcal{A}_{[0,k-1]}, \mathcal{D}_{[0,k-1]}).\end{aligned}$$

However, from Proposition 3, we know that the second term is $O(\alpha)$ away from η_{u+1} (because the initial state x_0 is in $\mathcal{X}_{u+1}$, and $k \geq \frac{1}{\sqrt{\alpha}}$). This allows us to further simplify the numerator of (69) to

$$\begin{aligned}&\sum_{x_1}\mathsf{Pr}(X^\alpha[k+1]=x, \mathcal{A}_k, D_k=1|X^\alpha[k]=x_1)\eta_{u+1}(x_1)\\ &\qquad+O(\alpha).\end{aligned} \tag{70}$$

Now, we use the nice-separability condition of Definition 3 to simplify the above summation. Apply the independent service condition for a departure from $\mathcal{X}_{u+1}$ to $\mathcal{X}_u$

$$\frac{\mathsf{Pr}_{\eta_{u+1}}(X^\alpha[1]=x|\mathcal{A}_0, D_0=1)}{\mathsf{Pr}_{\eta_{u+1}}(\mathcal{A}_0, D_0=1)}=\eta_u(x)+O(\alpha). \tag{71}$$

The numerator in the above expression can be written as

$$\sum_{x_1}\mathsf{Pr}(X^\alpha[1]=x, \mathcal{A}_0, D_0=1|X^\alpha[0]=x_1)\eta_{u+1}(x_1).$$

We will now establish that the expressions in (69) and (71) are close, in the sense that both the numerators and the denominators differ by $O(\alpha)$.

The numerator of (71) differs from (70) only in the time index, and an $O(\alpha)$ term. Thus, by time homogeneity of the system, we can see that the numerator of (71) differs from (70) only by $O(\alpha)$.

Further, the denominator of (69) is nothing but the sum of the numerator for different values of x, and thus, the denominator of (69) differs from the denominator of (71) only by $O(\alpha)$. This proves that the expression in (69) differs from the expression in (71) only by $O(\alpha)$. This proves (66), and completes the proof of Proposition 4. □

With the aid of Proposition 4, we are ready to prove the jump chain distribution lemma.

Proof of Lemma 3: The important part of the proof is contained in Proposition 4, which says that when we initialize the system in $\mathcal{X}_u$, and the state after transition is in $\mathcal{X}_{u-1}$ or $\mathcal{X}_{u+1}$, then the state is distributed according to the desired distribution η_{u-1} or η_{u+1}, respectively. To prove Lemma 3, we will rely on the fact that transitions to other sets have low probability, and thus, the state after a transition is always distributed according to η

Consider the jump chain X_n^α, initialized at its steady-state distribution $\gamma^\alpha(x)$, and recall that for all $A \subset \mathcal{X}$, the distribution γ^α satisfies (36). Further, when $A \subseteq \mathcal{X}_u$, we know from the negligible events lemma that only $i=u-1$ and $i=u+1$ make a significant contribution to the summation in (36), giving

$$\gamma^\alpha(A)=\sum_{\substack{i=u-1,\\ u+1}}\sum_{x_0\in\mathcal{X}_i}\gamma^\alpha(x_0)\mathsf{Pr}_{x_0}(X_1^\alpha\in A)+\delta\bar{O}(\alpha).$$

Now, substitute $A=\mathcal{X}_u$ to get

$$\gamma^\alpha(\mathcal{X}_u)=\sum_{\substack{i=u-1,\\ u+1}}\sum_{x_0\in\mathcal{X}_i}\gamma^\alpha(x_0)\mathsf{Pr}_{x_0}(X_1^\alpha\in\mathcal{X}_u)+\delta\bar{O}(\alpha). \tag{72}$$

Also, substitute $A=x\in\mathcal{X}_u$, and use the fact that $X_1^\alpha=x$ implies $X_1^\alpha\in\mathcal{X}_u$ to get

$$\begin{aligned}\gamma^\alpha(x)=&\sum_{\substack{i=u-1,\\ u+1}}\sum_{x_0\in\mathcal{X}_i}(\gamma^\alpha(x_0)\mathsf{Pr}_{x_0}(X_1^\alpha\in\mathcal{X}_u)\\ &\qquad\cdot\mathsf{Pr}_{x_0}(X_1^\alpha=x|X_1^\alpha\in\mathcal{X}_u)+\delta\bar{O}(\alpha)).\end{aligned} \tag{73}$$

Use Proposition 4 to rewrite the above equation as

$$\gamma^\alpha(\mathcal{X}_u)=\eta_u(x)\sum_{\substack{i=u-1,\\ u+1}}\sum_{x_0\in\mathcal{X}_i}\gamma^\alpha(x_0)\mathsf{Pr}_{x_0}(X_1^\alpha\in\mathcal{X}_u)+\delta\bar{O}(\alpha).$$

The term inside the parenthesis is given by (72), completing the proof of Lemma 3. □

This completes the proof of our main time-scale separation results.

VII. First-Order Approximation for Nonexponential File Sizes and Other Extensions

Until now, we have considered dynamic-user systems with exponentially distributed file size S. In this section, we briefly describe a technique to extend the first-order approximation to a more general class of file size distributions. Numerical examples and additional details about these extensions can be found in [23].

Kelly [20] gives a technique for the analysis of symmetric queues with general file size distributions. The analysis makes use of the fact that any distribution can be approximated by a mixture of a finite number of gamma distributions. Let the file size distribution be a mixture of M gamma distributions, with gamma distribution m corresponding to the sum of n_m exponential random variables, each with mean $\bar{S}_m$. Let the probability associated with gamma distribution m be p_m, with $m=1,\ldots,M$. Then, the mean file size is related to the means of the constituent exponential random variables by

$$\bar{S}=\sum_{m=1}^{M}p_m n_m \bar{S}_m. \tag{74}$$

This model for the file sizes can be interpreted as dividing users into M types, where a type m user wishes to sequentially transfer n_m files, each exponentially distributed with mean $\bar{S}_m$.

In this case, the number of users in the symmetric queue $\bar{U}(t)$ does not form a Markov process, and the state needs to be enlarged to a multidimensional Markov process $\bar{V}(t)$ in a finite state-space $\mathcal{V}$. Details about the structure of $\mathcal{V}$ be found in [20].

Returning to the dynamic-user system, earlier we considered exponentially distributed S. The method of proof for the first-order approximation result relied on splitting the state space $\mathcal{X}$ into a partition $\mathcal{X}_u$, with $u = 0, 1, \ldots$, where $\mathcal{X}_u$ corresponds to state $\bar{U}(t) = u$ for the symmetric queue.

With a general distribution for S, the definition of dynamic-user systems in Section III-A needs to change. In particular, for a nonexponential dynamic-user system, the state space $\mathcal{X}$ should be split into a partition $\mathcal{X}_v$, where v lies in the set $\mathcal{V}$. For this partition of the state space, a new version of the holding time and transition probability lemmas can be proved to show that transitions probabilities and mean holding times for the sets $\mathcal{X}_v$ are well approximated by the transition probabilities and mean holding times of states v in the symmetric queue. These new lemmas can be used to modify the proof of the first-order approximation theorem to prove that the steady-state distribution of the sets $\mathcal{X}_v$ is the same as the steady-state distribution of the Markov process $\bar{V}(t)$. In this manner, the first-order approximation can be extended to nonexponential dynamic-user systems.

A. *Other Extensions*

Several extensions to the results presented in this paper are given in [23], but are not included in this paper for brevity. These are itemized below.

- Numerical examples for the validity of the first- and second-order approximations.
- Extension of the second-order approximation to systems with dependent service. This extension is considerably more complicated than the second-order approximation for independent service systems, and has a detailed dependence on the structure of the Markov chain $X^\alpha[k]$.
- Extension of the second-order approximation to nonexponential file sizes. For this case, a heuristic is provided and numerical results are presented.
- Use of the second-order approximation to construct a new power control scheme for a simple on–off channel. Some details may also be found in [33].

VIII. Conclusion

We presented a time-scale separation technique for analyzing dynamic-user centralized wireless systems. We showed that the physical layer design affects the queueing level performance only through an effective service rate. Furthermore, under some reasonable conditions, the effective service rate depends only on the mean throughput $\phi(u)$ and the variance σ^2 of the service process. Thus, our technique simplifies the analysis of dynamic-user wireless systems by decoupling the analysis of the physical layer and user dynamics into two separate problems.

From the broader perspective of interlayer interaction in communication systems, our analysis shows that the data rate offered by the physical layer to the higher layers is not necessarily the same as the mean data rate of the physical layer. Rather, the effective data rate depends on the service variability (a physical layer parameter), as well as the file size (an application layer parameter). Thus, time-scale separation offers a way to study interlayer interaction.

Appendix I
Proving Proposition 3

Proof: The proof will rely on the multiplicative ergodicity result (52) and the notation developed in Section VI-A. The probability of interest is, after some manipulation

$$\begin{aligned}\mathsf{Pr}_{x_0}(X^\alpha[k] = x | G^\alpha \geq k) \\ = \frac{\mathsf{Pr}_{x_0}(X^\alpha[k] = x, \mathcal{D}_{[0,k-1]} | \mathcal{A}_{[0,k-1]})}{\mathsf{Pr}_{x_0}(\mathcal{D}_{[0,k-1]} | \mathcal{A}_{[0,k-1]})}. \quad (75)\end{aligned}$$

The denominator is evaluated in Proposition 2. Applying (46) and (48) shows that the denominator satisfies

$$\begin{aligned}\mathsf{Pr}_{x_0}\left(\mathcal{D}_{[0,k-1]} | \mathcal{A}_{[0,k-1]})\right) e^{-\alpha k \Lambda(-\alpha)} \\ = 1 + O(\alpha) + B_0 |\alpha| e^{-b_0 k}. \quad (76)\end{aligned}$$

To evaluate the numerator, we rely on (44). We can show that

$$\begin{aligned}\mathsf{Pr}_{x_0}(X^\alpha[k] = x, \mathcal{D}_{[0,k-1]} | \mathcal{A}_{[0,k-1]}) \\ = \sum_{\boldsymbol{x} \in \mathcal{B}} \mathbb{1}_{\{x_k = x\}} \prod_{n=1}^{k-1} e^{-\alpha f(x_n)} \tilde{F}_k(\boldsymbol{x}).\end{aligned}$$

This can be simplified to

$$\begin{aligned}\mathsf{Pr}_{x_0}(X^\alpha[k] = x, \mathcal{D}_{[0,k-1]} | \mathcal{A}_{[0,k-1]}) \\ = \mathsf{E}_{x_0}\left[\mathbb{1}_{\{\tilde{X}[k] = x\}} e^{-\alpha \tilde{I}[k-1]}\right].\end{aligned}$$

Applying (52) with $g(\cdot) = \mathbb{1}_{\{\cdot = x\}}$ gives

$$\begin{aligned}\mathsf{Pr}_{x_0}(X^\alpha[k] = x, \mathcal{D}_{[0,k-1]} | \mathcal{A}_{[0,k-1]}) e^{-\alpha k \Lambda(-\alpha)} \\ = \eta_u(x) + O(\alpha) + B_0 e^{-b_0 k}.\end{aligned}$$

The probability of interest in (75) is then given by dividing the above equation by (76), giving the desired result (62).

Appendix II
Probability of Two Events in a Slot

We wish to show that the probability of $D_{G^\alpha} + A_{G^\alpha} \geq 2$ is $\delta_d \bar{O}(\alpha)$ for all initial conditions x. We evaluate the probability by considering three events: $\{A_{G^\alpha} \geq 2\}$, $\{A_{G^\alpha} \geq 1, D_{G^\alpha} \geq 1\}$, and $\{D_{G^\alpha} \geq 2\}$. We show that the probability of the first two events is $\Omega \delta \bar{O}(\alpha)$, while the probability of the third event is $\delta_f \bar{O}(\alpha)$. Thus, using (14) we conclude that the probability of more than one event in a slot is $\delta_d \bar{O}(\alpha)$.

Two Arrivals: Let G_A^α be the first slot after slot 0 when an arrival occurs. The probability that more than one arrival event occurs at time G_A^α is (if Y is Poisson with mean a, consider $\mathsf{Pr}(Y \geq 2 | Y \geq 1)$)

$$\mathsf{Pr}_{x_0}(A_{G_A^\alpha} \geq 2) = \Omega \alpha \delta + \Omega \delta O(\alpha^2).$$

We are interested in the probability of $\{A_{G^\alpha} \geq 2\}$. Since $A_{G^\alpha} \geq 2$ implies that $G_A^\alpha = G^\alpha$, we have

$$\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 2) = \mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 2, G_A^\alpha = G^\alpha)$$

giving $\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 2) = \delta \Omega \bar{O}(\alpha)$

Arrival and Departure: Next, consider the event $\{A_{G^\alpha} \geq 1, D_{G^\alpha} = 1\}$. By expanding this probability over different values of G^α, (and keeping in mind the definition of the events $A_{[0,k-1]}$ and $D_{[0,k-1]}$), we get

$$\begin{aligned}
&\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1, D_{G^\alpha} \geq 1) \\
&\quad = \sum_k \mathsf{Pr}_{x_0}(A_k \geq 1, D_k \geq 1, G^\alpha = k) \\
&\quad = \sum_k \mathsf{Pr}_{x_0}(A_k \geq 1, D_k \geq 1, G^\alpha \geq k) \\
&\quad = \sum_k \mathsf{Pr}_{x_0}(D_k \geq 1 | G^\alpha \geq k) \\
&\qquad \cdot \mathsf{Pr}_{x_0}(G^\alpha \geq k) \mathsf{Pr}_{x_0}(A_k \geq 1).
\end{aligned}$$

The last term in the above summation is simply the probability of one or more arrivals in a slot, and can be written as

$$\mathsf{Pr}_{x_0}(A_k \geq 1) = \left(1 - e^{-\Omega\alpha\delta}\right).$$

This gives (using the identity $1 - e^{-a} < a$)

$$\begin{aligned}
&\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1, D_{G^\alpha} \geq 1) \\
&\quad \leq (\Omega\alpha\delta) \sum_k \mathsf{Pr}_{x_0}(D_k \geq 1 | G^\alpha \geq k) \mathsf{Pr}_{x_0}(G^\alpha \geq k). \quad (77)
\end{aligned}$$

Consider the first term in the summation. We know that

$$\mathsf{Pr}(D_k \geq 1 | X^\alpha[k] = x) = 1 - \exp(-\alpha f(x))$$

giving $\mathsf{Pr}(D_k \geq 1 | X^\alpha[k] = x) \leq \alpha f(x)$. The distribution of $X^\alpha[k]$ is given in Proposition 3, and can be used to obtain the following bound:

$$\begin{aligned}
\mathsf{Pr}_{x_0}(D_k = 1 | G^\alpha \geq k) &\leq \sum_{x \in \mathcal{X}_u} \eta_u(x) \alpha f(x) \\
&\quad + \left(\alpha \max_{x \in \mathcal{X}_u} f(x))(O(\alpha) + B_0 e^{-b_0 k}\right). \quad (78)
\end{aligned}$$

In evaluating the summation in (77), first consider the error term (the last term in the above equation). Using Proposition 2, the error term is

$$(\Omega\alpha\delta) \sum_k (\alpha \max_{x \in \mathcal{X}_u} f(x))(O(\alpha) + B_0 e^{-b_0 k}) \mathsf{Pr}_{x_0}(G^\alpha \geq k).$$

The summation can be computed using (53) to show that the error term is $O(\alpha^2)$.

Next, consider the first term on the right in (78). We have (by Definition 2)

$$\sum_{x \in \mathcal{X}_u} \eta_u(x) \alpha f(x) = \alpha\delta\phi(u).$$

The quantity of interest in (77) can then be written as

$$\begin{aligned}
&\mathsf{Pr}_{x_0}(A_{G^\alpha} \geq 1, D_{G^\alpha} = 1) \\
&\qquad \leq \alpha^2 \delta^2 \phi(u) \Omega \sum_k \mathsf{Pr}_{x_0}(G^\alpha \geq k) + O(\alpha^2).
\end{aligned}$$

Equation (53) can be used to show that this term is $\delta\Omega\bar{O}(\alpha)$.

Two Departures: The probability that two departures occur in slot G^α can be simplified, following the first equation in the previous section, to give

$$\mathsf{Pr}_{x_0}(D_{G^\alpha} \geq 2) = \sum_{k=0}^{\infty} \mathsf{Pr}_{x_0}(D_k \geq 2 | G^\alpha \geq k) \mathsf{Pr}_{x_0}(G^\alpha \geq k). \quad (79)$$

Our approach to evaluating the above summation will be similar to the one for the summation in (77). First, we use (62) to bound the first term in the summation. By definition of the departure events in (9), we get

$$\mathsf{Pr}(D_k \geq 2 | X^\alpha[k] = x) = \sum_{i < j \leq u} (1 - e^{-\alpha f_i(x)})(1 - e^{-\alpha f_j(x)}).$$

Some manipulation of the above summation, using $f(x) = \sum_j f_j(x)$, gives

$$\mathsf{Pr}(D_k \geq 2 | X^\alpha[k] = x) \leq \alpha^2 f(x)^2.$$

The probability of two or more departures in slot k, conditioned on no arrivals or departures till time k can then be bound using (62). This gives

$$\begin{aligned}
&\mathsf{Pr}_{x_0}(D_k \geq 2 | G^\alpha \geq k) \\
&\qquad \leq \sum_x \eta_u(x) \alpha^2 f(x)^2 + O(\alpha^3) + O(\alpha^2) e^{-b_0 k}.
\end{aligned}$$

From the above equation, and (53), it can be shown that

$$\mathsf{Pr}_{x_0}(D_{G^\alpha} \geq 2) \leq \frac{1}{\delta(\Omega + \phi(u))} \mathsf{E}_{\eta_u}[f(x)^2] \bar{O}(\alpha) + O(\alpha^2).$$

This shows $\mathsf{Pr}_{x_0}(D_{G^\alpha} \geq 2) \leq \delta_f \bar{O}(\alpha)$.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[2] S. Verdu and S. Shamai (Shitz), "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 622–640, Mar. 1999.

[3] S. Verdú, *Multiuser Detection*. Cambridge, U.K.: Cambridge Uni. Press, 1998.

[4] J. M. Massy, Ed., "Special Issue on Random Access Communication," *IEEE Trans. Inf. Theory*, vol. 31, no. 2, Mar. 1985.

[5] İ. E. Telatar and R. G. Gallager, "Combining queueing theory with information thoery for multiaccess," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 963–969, Aug. 1995.

[6] D. Raychaudhari, "Performance analysis of random access packet-switched code division multiple access systems," *IEEE Trans. Commun.*, vol. COM-29, no. 6, pp. 895–901, Jun. 1981.

[7] Q. Zhang, T. F. Wong, and J. S. Lehnert, "Performance of a type-II hybrid ARQ protocol in slotted DS-CDMA packet radio systems," *IEEE Trans. Commun.*, vol. 47, no. 2, pp. 281–290, Feb. 1999.

[8] R. Prakash and V. V. Veeravalli, "Wireless packet data systems with incremental redundancy—uplink analysis," in *Proc. Conf. Information Sciences and Systems*, Princeton, NJ, Mar. 2002, pp. 790–795.

[9] ——, "Traffic load based reverse link power allocation for cellular packet data systems," in *Proc. IEEE Vehicular Technology Conf.*, Vancouver, BC, Canada, Sep. 2002, pp. 2332–2336.

[10] B. Lu, X. Wang, and J. Zhang, Throughput of CDMA Data Networks With Multiuser Detection, ARQ, and Packet Combining, submitted for publication.

[11] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proc. IEEE INFOCOM*, San Francisco, CA, Apr. 2003, vol. 1, pp. 321–331.

[12] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[13] A. Eryilmaz, R. Srikant, and J. Perkins, Stable Scheduling Policies for Fading Wireless Channels, 2002 [Online]. Available: http://comm.csl.uiuc.edu/~srikant

[14] E. M. Yeh, "An inter-layer view of multiaccess communications," in *Proc. IEEE Intl. Symp. Information Theory*, Laussane, Switzerland, Jun./Jul. 2002, p. 112.

[15] ——, "Throughput and delay optimal resource allocation in multiaccess fading channels," in *Proc. IEEE Intl. Symp. Information Theory*, Yokohoma, Japan, Jun. 2003, p. 245.

[16] S. Shakkottai, R. Srikant, and A. L. Stolyar, "Pathwise optimality and state space collapse for the exponential rule," in *Proc. IEEE Intl. Symp. Information Theory*, Laussane, Switzerland, Jun./Jul. 2002, p. 379.

[17] X. Liu, E. Chong, and N. Shroff, "A framework for opportunistic scheduling in wireless networks," *Comput. Netw.*, vol. 41, pp. 451–474, Mar. 2003.

[18] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.

[19] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Engelwood Cliffs, NJ: Prentice-Hall, 1989.

[20] F. P. Kelly, *Reversibility and Stochastic Networks*. Chichester, U.K.: Wiley, 1979.

[21] D. R. Smith, "A new proof of the optimality of the shortest remaining processing time discipline," *Operations Res.*, vol. 26, no. 1, pp. 197–199, 1978.

[22] İ. E. Telatar, "Job scheduling and multiple access," in *DIMACS Workshop on Network Information Theory*, Rutgers Univ., Piscataway, NJ, Mar. 2003 [Online]. Available: http://cm.bell-labs.com/cm/ms/events/NIT03/abstract.html

[23] R. Prakash, "Centralized wireless networks with user arrivals and departures," Ph.D. dissertation, Univ. Illinois, Urbana-Champaign, 2003.

[24] S. Balaji and S. P. Meyn, "Multiplicative ergodicity and large deviations for an irreducible markov chain," *Stochastic Processes and Their Applications*, vol. 1, pp. 123–144, 2000.

[25] W. J. Stewart, *An Introduction to the Numerical Solution of Markov Chains*. Princeton, NJ: Princeton Univ. Press, 1994.

[26] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. London, U.K.: Springer-Verlag, 1993.

[27] H. Vantilborgh, "Exact aggregation in exponential queueing networks," *J. Assoc. Comp. Mach.*, vol. 25, no. 4, pp. 620–629, 1978.

[28] ——, "Aggregation with an error of $O(\epsilon^2)$," *J. Assoc. Comp. Mach.*, vol. 32, no. 1, pp. 162–190, 1985.

[29] D. Kafeety, C. Meyer, and W. Stewart, "A general framework for iterative aggregation /disaggregation methods," in *Proc. 4th Copper Mountain Conf. Iterative Methods*, Copper Mountain, CO, Apr. 1992.

[30] P. J. Courtois, "Error analysis in nearly-completely decomposable stochastic systems," *Econometrica*, vol. 43, no. 4, pp. 691–709, Jul. 1975.

[31] P. J. Schweitzer, "Perturbation series expansions for nearly completely-decomposable markov chains," in *Teletraffic Analysis and Computer Performance Evaluation*, O. J. Boxma, J. W. Cohen, and H. C. Tijms, Eds. Amsterdam, The Netherlands: Elsevier Science (North-Holland), 1986, vol. 7, pp. 319–328.

[32] I. Kontoyiannis and S. P. Meyn, "Spectral theory and limit theorems for geometrically ergodic markov processes," *Ann. Probab.*, vol. 13, pp. 304–362, 2003.

[33] S. Prakash and V. V. Veeravalli, "The impact of service rate fluctuations in wireless packet data systems," in *Proc. IEEE Int. Symp. Information Theory*, Yokohama, Japan, Jun./Jul. 2003, p. 131.