The impact of dependent service times in large-scale service systems

Guodong Pang* and Ward Whitt**
*Pennsylvania State University, University Park, PA, 16802
Email: gup3@psu.edu
** Columbia University, New York, NY, 10027
Email: ww2040@columbia.edu

April 16, 2012

1 Introduction

In service systems there can be dependence among successive service times. For example, in a telephone call center responding to service calls, a product defect can lead to many calls concerning that same product over a relatively short interval of time, which may require similar handling times. For another example, in a hospital emergency room, there may be multiple patients associated with the same medical incident. Several people may be victims of a single highway accident or food poisoning at the same restaurant. There may be rapid spread of a contagious disease. The common causes may lead to dependent service times. Thus it is interesting to understand the impact of dependence among the successive service times upon the ability of a service system to respond to service requests. There are well developed methods to study the impact of average service times, but the impact of the dependence, for given mean, has evidently not been studied before.

2 The models

We consider the many-server queueing models with general arrival processes possibly having time-varying arrival rates and multiple parallel statistically identical servers. Instead of assuming i.i.d. service time distributions, we allow the successive service times to be dependent; specifically, we assume that the successive service times form a sequence of weakly dependent random variables satisfying certain mixing conditions. For example, the successive service times can form a sequence of discrete autoregressive process of order one, which is a sequence of batches of independent service times or identical service times. For another example, customers arrive in batches, and service times of customers within each batch are correlated while those across different batches are mutually independent. Our goal is to find the impact of such dependence among successive service times upon certain performance measures, both in steady-state and in transient states.

3 Heavy-traffic stochastic-process limits

We establish heavy-traffic fluid limit and its refined stochastic limit for the number of customers in the system for the above models with infinite number of servers under the assumption that the arrival processes satisfy a functional central limit theorem (FCLT). We find that the fluid limit is the same as that for the models without dependence among successive service times. However, its refined stochastic limit is very different, capturing the impact of the dependence. In order to prove the convergence to the limits, we represent the number of customers in the system at each time by an integral against a sequential empirical process driven by the successive service times. This insightful perspective allows us to view service times sequentially for all customers, and thus, to incorporate dependence among successive service times. An FCLT are established for diffusion-scaled sequential empirical processes in a proper function-valued functional space, and its limit is a generalized Kiefer process (Gaussian random field) when certain mixing conditions are satisfied for the sequence of dependent service times. The FCLT for the number of customers in the system is a functional of such a generalized Kiefer process, which is a Gaussian process (random field) when the arrival limit process is a Brownian motion.

4 Steady-state approximations

When the arrival rate is constant and the arrival limit process is a Brownian motion (BM), we derive the stationary distribution of the number of customers in the system, which is Gaussian with explicit mean and variance formulas. We find that the mean is the same as that without dependence, and the variance has an additional term besides those without dependence, which completely captures the dependence effect itself. A very useful performance measure is the peakedness, the ratio of the variance to the mean of the number of customers in the infinite-server queue. Since the dependence in the service times in our models does not affect the mean number of customers in system at all, our heavy-traffic approximation for the variance of the number of customers in system translates directly into an associated approximation for the peakedness. In fact, the approximate peakedness measure has an additional term besides those without dependence, which can be written as integrals of a sum of the deviation of bivariate joint tail distributions of two dependent service times of any distance apart from those in the independent case. This additional term can also be represented as a sum of the differences of expected minimum of two dependent service times of any distance apart from that of two independent ones. If we use extremal bivariate distributions with correlations as approximations, the approximate peakedness measure can be written as a linear functional of the aggregate correlation parameter. For example, when the service times form a discrete autoregressive process of order one (or called randomly repeated service times), that is, each successive service time be a mixture of the previous service time with probability p or a new independent service time having the same distribution with probability $1-p$, then the aggregate correlation parameter is simply $p/(1-p)$. For another example, in the batch arrival model, if service times within each batch have the same correlation, then this aggregate correlation parameter is a linear in this correlation. Moreover, we show that in the batch model, when the arrival process of batches is Poisson

and the service times within each batch is multivariate Marshall-Olkin exponential distribution, such a linearity relationship is exact. We conduct extensive simulation experiments to verify that the approximations are very accurate for general interarrival and service time distributions.

5 Approximations with time-varying arrival rates

When arrival rates are time-varying and the arrival limit process is a BM, the Gaussian approximation of the number of customers in the system has both time-varying mean and variance functions. As in the stationary scenario, the mean function is the same as without dependence and has been well studied in the literature as offered load function for general many-server queues (with or without abandonment). However, the variance function is affected by the dependence among service times, and has not been paid any attention. One can regard the variance function as the variance of offered load in general many-server queues. We derive an exact approximations of the variance function by representing it as a pointwise stationary approximation, in which the arrival rate function is shifted by the stationary excess random variables of a service time, minimum of two independent service times and minimum of two dependent service times of any distance apart. When the arrival rate function is sinusoidal, we are able to write this approximation explicitly in terms of the expectations of these random variables and their sinusoidal functionals. When the arrival rate function is general, we use two approximation methods to approximate the variance function: Taylor approximations of second order and approximations based on a recent average arrival rate. We derive explicit expressions for these exact and approximation expressions when the service times are generated in the randomly repeated scheme and in the batch model with multivariate Marshall-Olkin exponential and hyperexponential (newly introduced) distributions, and conduct simulations to compare with the exact and approximation expressions. We find that in all simulation experiments, the Taylor approximations of second order are very accurate. We also observe that the variance function is larger for all time when the probability of repeating its previous service time is larger in the models with randomly repeated service times and when the correlation among service time within each batch is larger in the batch models.

6 Delay probability approximation

It is well known that the so-called Halfin-Whitt delay function is a very good approximation of the delay probability for Markovian many-server queues. For general many-server queueing models with dependent service times and finite number of servers, we modify the Halfin-Whitt delay function by using the peakedness measure to approximate the delay probability. Our simulation experiments show that such an approximation is also very effective. In the models with randomly repeated service times, the delay probability increases as the probability of repeating its previous service time increases. In the batch models, the delay probability increases as the correlation among service times within each batch increases.