Heavy-Traffic Limits for Nearly Deterministic Queues

Karl Sigman, Ward Whitt
Department of Industrial Engineering and Operations Research
Columbia University
500 West 120th Street
New York, NY 10027-6699
{ks20,ww2040}@columbia.edu

ABSTRACT

We establish heavy-traffic limits for "nearly deterministic" queues, such as the $G/D/n$ many-server queue. Waiting times before starting service in the $G/D/n$ queue are equivalent to waiting times in an associated $G_n/D/1$ model, where the G_n denotes "cyclic thinning" of order n, indicating that the original (possibly general) point process of arrivals is thinned to contain only every n^{th} point. We thus focus on the $G_n/D/1$ model and the generalization to $G_n/G_n/1$, where "cyclic thinning" is applied to both the arrival and service processes. As $n \to \infty$, the $G_n/G_n/1$ models approach the deterministic $D/D/1$ model. The classical example is the Erlang $E_n/E_n/1$ queue, where cyclic thinning of order n is applied to both the interarrival times and the service times, starting from a "base" $M/M/1$ model. We establish different limits in two cases: (i) when $(1-\rho_n)\sqrt{n} \to \beta$ as $n \to \infty$ and (ii) $(1-\rho_n)n \to \beta$ as $n \to \infty$, where ρ_n is the traffic intensity in model n, and $0 < \beta < \infty$. The nearly deterministic feature leads to interesting nonstandard scaling. We also establish revealing heavy-traffic limits for the stationary waiting times and other performance measures in the $G_n/G_n/1$ queues, by letting $\rho_n \uparrow 1$ as $n \to \infty$.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Queueing Theory

General Terms

Performance, Theory

Keywords

heavy traffic; nearly deterministic queues; cyclic thinning, point processes, Erlang queues; many-server queues; deterministic service times; Gaussian random walk; functional central limit theorems; invariance principles.

1. INTRODUCTION

A primary cause of congestion in a queueing system is stochastic fluctuations in the arrival times and the service times. We say that a queueing system is *nearly deterministic* if these stochastic fluctuations are low. At customary loads, the congestion in a nearly deterministic queueing system will be negligible. However, if the system is nearly deterministic, then it is natural to operate the system at higher loads. In this research we explore the interplay between low variability and high loads. In particular, we establish heavy-traffic (HT) limits for some nearly deterministic queueing models. We consider single-server models (indexed by n) which approach the $D/D/1$ queue as $n \to \infty$.

A classic example is the FIFO $GI/D/n$ multiserver queue when n is large. It is well known that waiting times (before starting service) in this model can be identified with waiting times in the corresponding FIFO $GI_n/D/1$ model, where the GI_n means that the arrival process is the renewal process whose interarrival times are distributed as the sum of n interarrival times from the original renewal arrival process; a "cyclic thinning" of the original arrival process; e.g., see Theorem 4.6.1 of [9]. (Due to the deterministic service, customers to a FIFO $GI/D/n$ queue depart in the same order they arrive, and so without loss of generality, the customers can be assigned to the servers in a cyclic order; each of the servers thus becomes a FIFO $GI_n/D/1$ queue.) The GI_n interarrival times become nearly deterministic as n increases by virtue of the law of large numbers. In applications, a Poisson arrival process is often a realistic assumption. The reduction of $M/D/n$ to $E_n/D/1$ for the waiting times is often mentioned in textbooks. Otherwise, the renewal process assumption is not so realistic. Thus, it is important that both the reduction of the $GI/D/n$ model to the $GI_n/D/1$ model and our HT limits hold for more general "G" arrival processes (e.g., such as a general point process that satisfies a functional central limit theorem (FCLT)). There are no algorithms available to compute the steady-state waiting time distribution or even only its mean in the new $G_n/D/1$ model. Thus, the simple approximations stemming from the HT limits we establish can be very useful.

Motivated by the $GI_n/D/1$ example, we consider the waiting time process in FIFO single-server queues with a $G_n/G_n/1$ structure, where cyclic thinning is applied to *both* the interarrival times and the service times. Our results cover the two models $D/G_n/1$ and $G_n/D/1$ as special cases because of the way we do the cyclic thinning: When working with partial sums of interarrival times or service times, we let the new sequence of G_n partial sums $\{S_{n,k} : k \geq 1\}$ be defined in terms of the sequence of original G partial sums $\{S_{1,k} : k \geq 1\}$ by letting $S_{n,k} \equiv S_{1,kn}/n$, $k \geq 1$; i.e., we scale the index k in the original partial sums $S_{1,k}$ by n because we add n consecutive times, but we also divide by n in order to keep the mean fixed in the identically distributed case. It is easy to see that $D_n = D$ with this construction.

If the traffic intensity ρ_n in the $G_n/G_n/1$ model (assumed well defined) is held fixed at a stable value or, more generally, satisfies $\rho_n \to \rho < 1$ as $n \to \infty$, then the $G_n/G_n/1$ model approaches the purely deterministic $D/D/1$ model, and the stationary waiting time becomes asymptotically negligible.

However, we let $\rho_n \uparrow 1$ as $n \to \infty$. We thus obtain an interesting *double limit*, in which the models approach $D/D/1$, while the traffic intensity increases. On the one hand, congestion should decrease, because the models are becoming less variable, approaching $D/D/1$. On the other hand, the congestion should increase because we let $\rho_n \uparrow 1$. We let ρ_n approach 1 at an appropriate rate so that we get revealing nondegenerate limits.

For the multiserver $G/D/n$ model mentioned at the outset, the double limit coincides with the familiar many-server HT limit, in which we let the traffic intensities ρ_n approach 1 as the number of servers, n, increases [2]. In fact, HT limits were already established for stationary performance measures in [3] in the so-called Halfin-Whitt or quality-and-efficiency-driven (QED) regime, in which

$$(1-\rho_n)\sqrt{n} \to \beta \quad \text{as} \quad n \to \infty, \quad 0 < \beta < \infty. \tag{1}$$

We establish stochastic-process HT limits in this case, but we also establish stochastic-process limits when

$$(1-\rho_n)n \to \beta \quad \text{as} \quad n \to \infty, \quad 0 < \beta < \infty. \tag{2}$$

as $n \to \infty$. Let $W^c_{n,k}$ be the waiting time of arrival k in the $G_n/G_n/1$ queue, where the superscript c indicates that cyclic thinning of order n is applied to a base $G/G/1$ model. In case (1) we obtain a limit for the *spatially-scaled* waiting times $\sqrt{n}W^c_{n,k}$; in case (2) we obtain a stochastic-process limit for the *temporally-scaled* waiting times $W^c_{n,nk}$ [6].

The asymptotically-deterministic feature is critical for these limits. For example, the HT limits for the $G_n/GI/1$ model as $n \to \infty$ with fixed service-time distribution are significantly different in the two cases: (i) when the GI service-time distribution is D and (ii) when the service-time distribution is not D (and we do not perform the cyclic thinning on the service times, replacing GI by GI_n). When the service-time distribution is not deterministic, the $G_n/GI/1$ model is not asymptotically deterministic as $n \to \infty$. As a consequence, the HT limit agrees with the conventional one for the corresponding $D/GI/1$ model, with the usual scaling, obtained by simply replacing the interarrival-time distribution in the G_n process by a deterministic interarrival times with the same mean. In contrast, that is *not* the case with the nearly deterministic $G_n/D/1$ model.

In many ways, the HT behavior of the $G_n/G_n/1$ models as $n \to \infty$ is different from the conventional HT behavior of the $GI/GI/1$ model, discussed in Chapters 5 and 9 of [11]. Unlike the conventional HT theory for the $GI/GI/1$ model, for the $G_n/G_n/1$ models there need not be any spatial scaling. In the conventional HT limit, the queue-length and waiting time processes have the same asymptotic behavior; both processes behave like reflected Brownian motion, after the same scaling. In contrast, for the $G_n/G_n/1$ models, the waiting-time and queue-length processes look very different.

To illustrate these differences, we now plot sample paths of the waiting times (before starting service) of successive arrivals and the continuous-time queue-length process from one simulation run of the $E_{100}/D/1$ queue with traffic intensity $\rho = 0.99$ and unit service times. Figure 1 shows the waiting times at arrival epochs and the continuous-time queue length process, starting empty, in the final subinterval of length 200 ending at $t = 5 \times 10^4$ from a single run over the time interval $[0, 5 \times 10^4]$.

First, all values of both processes over the full time interval of length $50,000$ fall in the interval $[0, 5]$ without any

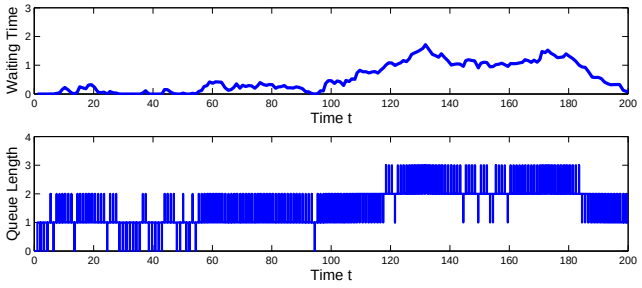

Figure 1: Simulation plots of the waiting times at arrival epochs and queue lengths at arbitrary times in the $E_{100}/D/1$ model with $\rho = 0.99$, starting empty, for a time interval of length 200 ending at $t = 50,000$.

spatial scaling. The waiting times are comparable to the unit service times, e.g., the average waiting time is about 0.5. Second the waiting-time plots look continuous, like a plot of reflected Brownian motion, which we will show is indeed its HT limit. In contrast, the queue-length process is integer-valued, making frequent jumps of size 1. Evidently its limiting behavior is more complicated. We explain these plots in this work [6].

We also consider heavy-traffic limits for the stationary distributions of waiting times under the 2 previous conditions, (1)and (2) [7]. We first address a foundational issue. We show that stationary waiting times are well defined by placing the $G_n/G_n/1$ model in a stationary framework [5]. To do so, we show that stationarity and ergodicity assumed for a point process are inherited by the new point process created by cyclic thinning. We establish limits for stationary point processes modified by cyclic thinning. We show that counting processes created by cyclic thinning do not have the same relatively simple asymptotic behavior as the associated partial sums. (The continuous mapping theorem with the inverse map discussed in §13 of [11] does not apply in the usual way.)

Within the stationary framework for the $G_n/G_n/1$ model, the waiting times $W^c_{n,k}$ converge to stationary waiting times $W^c_{n,\infty}$ as $k \to \infty$ for each n. We can apply the stochastic-process HT limits to generate approximations for those stationary waiting times by considering the iterated limit in which first $n \to \infty$ and then $k \to \infty$. We provide conditions under which the limit interchange is valid. In particular, we provide conditions under which $\sqrt{n}W^c_{n,\infty}$ and $W^c_{n,\infty}$ converge in distribution to proper limits as $n \to \infty$ in cases (1) and (2), respectively, and identify the limits with the iterated limits already established.

Special cases of the two limits for stationary waiting times were established previously. First, in Example 3.1 of [1], exploiting the known Laplace transform of the stationary waiting time, the authors showed in case (2) that $W^c_{n,\infty} \Rightarrow W^c_\infty$ as $n \to \infty$, where $\Rightarrow$ denotes convergence in distribution and W^c_∞ is an exponential random variable, when the $G_n/G_n/1$ model is $E_n/E_n/1$, i.e., when the base model is $M/M/1$. Second, in [3] the authors showed in case (1) that $\sqrt{n}W^c_{n,\infty} \Rightarrow \tilde{W}^c_\infty$ as $n \to \infty$, where $\tilde{W}^c_\infty$ is the maximum of a Gaussian random walk with negative drift, when the $G_n/G_n/1$ model is $GI_n/D/1$. In [3] the authors actually

$E_k/E_k/1$ queue with mean service time 1 and $\rho_k \equiv 1-(1/k)$

	$k=10$	$k=10^2$	$k=10^3$	$k=10^4$
$P(W>0)$ exact	0.7102	0.9036	0.9688	0.9900
approx case (1)	0.6279	0.8666	0.9561	0.9843
approx case (2)	1.0000	1.0000	1.0000	1.0000
$E[W\|W>0]$ exact	1.054	1.0175	1.0055	1.0018
approx case (1)	1.216	1.0265	1.0189	1.0076
approx case (2)	1.000	1.0000	1.0000	1.0000
$E[W]$ exact	0.7484	0.9195	0.97417	0.99018
approx case (1)	0.7635	0.9201	0.9742	0.99018
approx case (2)	1.0000	1.0000	1.0000	1.00000

Table 1: A comparison of the approximations for three steady-state performance measures in the two cases of heavy-traffic scaling with exact numerical values computed using the numerical algorithm, for Erlang models as the Erlang order increases.

considered the $G/D/n$ model, but they analyzed it by exploiting the fact that the waiting times are the same as in the associated $GI_n/D/1$ model. Our results are extensions of those two results. We obtain some results for stationary waiting times for general $G_n/G_n/1$ models, but most of our results are for the special case $GI_n/GI_n/1$ in which the base model is $GI/GI/1$.

The two different scalings in (1) and (2) indicate that high-order cyclic thinning produces some interesting behavior. To a large extent, this phenomenon can be explained by the fact that two parts of the distribution of $W^c_{n,\infty}$ tend to have different asymptotic behavior. Paralleling the relatively well understood many-server queue; e.g., [10], the delay probability $P(W^c_{n,\infty}>0)$ and the conditional delay distribution $P(W^c_{n,\infty}>t|W^c_{n,\infty}>0)$ behave differently. In case (1), the delay probability $P(W^c_{n,\infty}>0)$ has a nondegenerate limit α (with $0<\alpha<1$) as $n\to\infty$, without scaling, while $W^c_{n,\infty}\Rightarrow 0$, i.e., $P(W^c_{n,\infty}>t|W^c_{n,\infty}>0)\to 0$ as $n\to\infty$ for each $t\geq 0$. On the other hand, in case (2), $P(W^c_{n,\infty}>0)\to 1$, a degenerate limit, while $P(W^c_{n,\infty}>t|W^c_{n,\infty}>0)$ has a nondegenerate limit as $n\to\infty$ for each t. This is a unifying theme throughout this research.

Since we have two candidate approximations for the stationary waiting time distribution provided by the HT limits in the two cases (1) and (2), it is interesting to see how they compare to exact values. Hence we made comparisons with exact numerical results [1] and simulation estimates for $E_k/E_l/1$ high-order Erlang models.

For example, Table 1 compares the approximations for the $E_k/E_k/1$ model with $\rho_k=1-(1/k)$ to the exact numerical results in Table 1 of [1], for $k=10^j$ for four values of j. The scaling here is naturally in case (2), because $(1-\rho_k)k=1$ for all k, but we considered both cases (1) and (2). In case (1) we evaluated approximations for the delay probability and the mean using matlab. In case (2) the approximations for the delay probability and the mean wait are both 1.

Even though the scaling puts these examples naturally in the domain of case (2), we find that the approximations based on case (1) consistently perform better than the approximations based on case (2). Case (2) becomes competitive and even preferred to case (1) when we focus on the expected conditional delay, given that the wait is positive $E[W|W>0]$. When we look at the conditional mean $E[W|W>0]$, we find that each approximation works better when we expect it to. Overall, these approximations are remarkably effective, given the huge error in the mean using a simple $M/M/1$ approximation. For example, for $k=100$, the $M/M/1$ approximation yields $P(W>0)=0.99$ and $E[W]=99$. The $M/M/1$ approximation for the mean is off by a factor of 99.

We close this abstract by briefly mentioning two other connections. First, on account of the D service, the associated $G/D/n+GI$ queueing model with customer abandonment (the $+GI$) has interesting periodic behavior when it is overloaded ($\rho>1$) [4].

Second, another source of motivation came from [8], wherein the authors studied a basic (r,q) inventory model, in which the demand forms a Poisson process at rate λ and the lead times are i.i.d. distributed as L. Every q^{th} demand from the Poisson process triggers an order requiring time L to arrive. Thus there is a $E_q/GI/\infty$ queue in the background. They were interested in the joint effect upon performance of $Var(L)$ and the lot size q. Because the model is intractable, they use a HT approximation, first considering deterministic lead times. In [6] we show that our analysis adds insight.

This research was partially supported by NSF grant CMMI 0948190.

2. REFERENCES

[1] J. Abate, G. L. Choudhury, and W. Whitt. Calculation of the gi/gi/1 steady-state waiting-time distribution and its cumulants from pollaczek's formulas. *International Journal of Electronics and Communications (AËU)*, 47:311–321, 1993.

[2] S. Halfin and W. Whitt. Heavy traffic limits for queues with many exponential servers. *Operations Research*, 29:567–587, 1981.

[3] P. Jelenkovic, A. Mandelbaum, and P. Momcilovic. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems*, 47:53–69, 2004.

[4] Y. Liu and W. Whitt. The heavily loaded many-server queue with abandonment and deterministic service times. *Columbia University, NY, NY*, 2010. http://www.columbia.edu/∼ww2040/allpapers.html.

[5] K. Sigman. *Stationary Marked Point Processes: An Intuitive Approach*. Chapman Hall, NY, NY, 1995.

[6] K. Sigman and W. Whitt. Heavy-traffic limits for nearly deterministic queues. *Columbia University, NY, NY*, 2010. http://www.columbia.edu/∼ww2040/allpapers.html.

[7] K. Sigman and W. Whitt. Heavy-traffic limits for nearly deterministic queues: stationary distributions. *Columbia University, NY, NY*, 2010. http://www.columbia.edu/∼ww2040/allpapers.html.

[8] J.-S. Song and P. H. Zipkin. The joint effect of leadtime variance and lot size in a parallel processing environment. *Management Science*, 42:1352–1363, 1996.

[9] H. C. Tijms. *Stochastic Models: An Algorithmic Approach*. John Wiley, NY, NY, 1994.

[10] W. Whitt. Understanding the efficiency of multi-server service systems. *Management Science*, 38:708–723, 1992.

[11] W. Whitt. *Stochastic-Process Limits*. Springer, NY, NY, 2002.