On the Many-Server Fluid Limit for a Service System with Routing based on Delayed Information

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA

Abstract

Pender, Rand and Wesson [1] established a delay differential equation limit for a parallel service system with routing based on delayed information. We provide an interpretation of their scaling under which their limit can be regarded as an instance of a law of large numbers in the familiar many-server heavy-traffic scaling. It requires scaling in the probabilistic routing function. We also obtain related limits for more general models.

Keywords: queues, delayed information, delay differential equation, fluid limit, many-server heavy-traffic scaling

1. Introduction

Pender, Rand and Wesson [1] established a delay differential equation (deterministic fluid) limit for a parallel service system with routing based on delayed information. That limit is useful because it helps quantify and understand the impact of the delay. We introduce an interpretation of the scaling used in [1], which shows that their limit can be regarded as being consistent with a law of large numbers in the familiar many-server heavy-traffic scaling in [2] and many other papers. With many-server scaling, an additional scaling of the probabilistic routing function is required. (See (2)-(5) for the definition of the routing and (7)-(8) for the proposed scaling.) Lemma 3.1 shows that the many-server scaling here with the scaling of the routing in (8) produces the same scaling used in [1]. Hence, the limit obtained by the scaling here is equivalent to the limit in [1], so we are primarily just translating into the many-server heavy-traffic framework. The paper [1] nicely shows the implications of the limit by analyzing the limiting delay differential equation.

Related early work involving delayed information can be found in [3, 4] on the study of rate control in communication networks and in citations to these papers. More recently, information delay has played an important role in high-speed financial trading; e.g., [5, 6].

In §2 we introduce our model, which is a modification of the model in [1], covering the model with finitely many servers in each group and customer abandonment (analog of the Erlang A model). In §3 we introduce the scaling. In §4 we present the limit. In §5 we extend the results to the more general time-varying non-Markov $G_t/GI/s_t + GI$ model in [7, 8]. In §6 we conclude with additional discussion.

2. The Model

The queueing model has Poisson arrivals at rate λ, routing to one of N multi-server service groups based on the queue lengths (numbers in the service group, either waiting or being served) in these N service groups in the past. For simplicity, suppose that the system starts empty at time 0 and was empty in the past before time 0. A somewhat more general initial condition is considered in [1]; it is not difficult to treat that extension. Let the total number of arrivals over the interval $[0, t]$ be denoted by $A(t)$. Thus $A \equiv \{A(t) : t \geq 0\}$ is a Poisson process with rate λ (with $\equiv$ denoting equality by definition).

Let service group i have s_i servers, $1 \leq i \leq N$, working independently in parallel. The special case $s_i = \infty$ for all i is considered In [1], which is an important special case. Let $Q(t) \equiv (Q_1(t), \ldots, Q_N(t))$ be the vector of queue lengths at time t. Let customers have i.i.d exponential service times in each service group. Let $1/\mu_i$ denote the mean service time of each customer served in service group i. Then μ_i is the service rate of individual customers in service group i. (In [1] it is assumed that $\mu_i = \mu$ for all i.) We assume that the arrival process and service times are mutually independent. Let customers waiting in queue have i.i.d. exponential patience times with mean $1/\alpha_i$. If the customer has not entered service by their patience time, they abandon the system without otherwise influencing the system.

We now turn to the routing, which is a form of probabilistic routing. let $X_k \equiv (X_{k,1}, \ldots X_{k,N})$ represent the routing of arrival number k, $k \geq 1$. Hence X_k is a vector with all components equal to 0 except a single component equal to 1, which is the index of the service group to which the arrival is routed. Each arrival is routed to the designated service group immediately upon arrival, after which it immediately enters service if possible; otherwise the customer goes to the end of the queue at server group i. Then we have the arrival process to each queue i resulting from the routing defined as

$$A_i(t) = \sum_{k=1}^{A(t)} 1_{\{X_{k,i}=1\}}, \quad t \geq 0, \quad 1 \leq i \leq N. \tag{1}$$

We assume that the routing at time t depends on the system history up to time t only through the history of the queue-length vector $\{Q(u) : 0 \leq u \leq t\}$ at time t. In particular, we assume that there is probabilistic routing depending upon that history. In general, the way that the arriving customers could obtain past state information can be quite complex. For example, each service group might make periodic state estimates and broadcast them to all potential customers. That is like the emergency department delays discussed by [9]. In that case the delays are periodically 0 and then increase linearly over the interval between updates. These updates might or might not be coordinated and synchronized. See [10] for a study of such systems.

However, following [1], we avoid that complexity and, assume a simplified approximate representation of the probabilistic routing based on delays. For $t \geq \Delta$, where Δ is a constant time, an arrival at time t is routed to service

group i with probability

$$p_i(t, \{Q(u) : t - \Delta \le u \le t\}) \equiv \frac{\tilde{r}_i(t, \{Q(u) : t - \Delta \le u \le t\})}{\sum_{j=1}^{N} \tilde{r}_j(t, \{Q(u) : t - \Delta \le u \le t\})}, \quad (2)$$

where

$$\tilde{r}_i(t, \{Q(u) : t - \Delta \le u \le t\}) \equiv r_i(t, \theta_i \int_0^{\Delta} Q_i(t-u)\, dG_i(u)) \quad (3)$$

for a specified cdf G_i, with r_i being a continuous positive real-valued function over its domain satisfying

$$r_i(t, 0) \equiv 1 \quad \text{for all} \quad t \quad \text{and} \quad i. \quad (4)$$

We assume that $r_i(t, x)$ is strictly decreasing in x for each t as well. Hence, the routing to a particular service group gets less likely as the queue length there increases. Thus, the routing is a probabilistic analog of the join-the-shortest-queue routing. Condition (4) makes the routing to each queue be with probability $1/N$ when all N service groups are empty. By our assumption that the system starts out empty, this routing prevails immediately after time 0.

The routing formulation above is more general than in [1]. First, in [1] the model is a symmetric model in which (μ_i, θ_i, G_i) is independent of the server group i. Second, in [1] the cdf G is the unit point mass on Δ, but a randomized delay is studied in [11] and an asymmetric model is studied in [12]. In [1, 11, 12] r_i is assumed to take the special form

$$\begin{aligned} r_i(t, \theta_i \int_0^{\Delta} Q_i(t-u)\, dG_i(u)) &\equiv \exp\{-\theta_i \int_0^{\Delta} Q_i(t-u)\, dG_i(u)\} \\ &\equiv \exp\{-\theta_i Q_i(t-\Delta)\}; \end{aligned} \quad (5)$$

i.e., there is an explicit exponential form, motivated by the multinomial logit customer choice model. It is easy to see that (5) is consistent with our assumptions.

3. Many-Server Scaling

We now introduce what we regard as natural many-server scaling. We construct a sequence of parallel-server models indexed by positive integers η,

as in [1]. For each η, there is a parallel-server model consisting of N infinite-server queues. Consistent with [2], we let the arrival rate and number of servers in each service group grow, but we hold the service-time distribution and the information delay fixed. In particular, we let the parameters in model η be

$$\begin{aligned}(\lambda^{(\eta)}, s_i^{(\eta)}) &\equiv (\eta\lambda, \eta s_i), \\ (\mu_i^{(\eta)}, \alpha_i^{\eta}, \Delta_i^{(\eta)}, G_i^{(\eta)} &\equiv (\mu_i, \alpha_i, \Delta_i, G_i) \quad 1 \le i \le N, \quad \text{for all} \quad \eta \ge 1. \end{aligned} \tag{6}$$

In other words, the arrival rate and number of servers in each service group is scaled up by the constant factor η, while the service-time, patience and delay distributions are independent of η. The idea is that large scale is directly determined by the increasing arrival rate and system capacity (growing together), while the service, patience and delay times should be relatively insensitive to scale.

Because we are interested in the law of large numbers (LLN) yielding the deterministic limit, we focus on the scaled queueing process

$$\bar{Q}^{(\eta)}(t) \equiv \eta^{-1} Q^{(\eta)}(t), \quad t \ge 0, \tag{7}$$

where $Q^{(\eta)}(t)$ is the vector queue-length process in model η. (This is consistent with [2].)

Turning to the routing, we assume that the routing depends on $\bar{Q}^{(\eta)}$ instead of $Q^{(n)}$. Equivalently, we assume that the routing constants $\theta_i^{(\eta)}$ appearing in (3) and (5) scale by

$$\theta_i^{(\eta)} \equiv \theta_i/\eta \quad \text{for all} \quad i \quad \text{and} \quad \eta. \tag{8}$$

This seems to be a reasonable initial assumption in face of increasing scale. (Refined scaling might be considered for refined diffusion limits.)

With this scaling, we again get a generalization of equation (5) of [1], but with a different interpretation. The quantity $Q_i^{\eta}(t)$ in [1] is replaced by $\bar{Q}_i^{(\eta)}(t)$ in (7) above and in the unscaled system θ_i is replaced by $\theta_i^{(\eta)} \equiv \theta_i/\eta$ in (8). Assuming that equation (5) here applies to the unscaled number in system, we have

$$\theta_i^{(\eta)} Q_i^{(\eta)}(t) = \theta_i \bar{Q}_i^{(\eta)}(t) \quad \text{for all} \quad \eta \ge 1, \quad t \quad \text{and} \quad i. \tag{9}$$

With the interpretation/definitions above, we can apply the results in [1] to obtain the natural generalization of their stated limit to the case of finitely many servers.

To state the expression explicitly, let Π_i^a, Π_i^s and Π_i^l be $3N$ independent rate-1 Poisson processes. Then, in our notation under the simplifying assumption of starting empty, we have the following explicit construction. Let $x \wedge y \equiv \min\{x, y\}$ and $(x)^+ \equiv \max\{x, 0\}$ for real numbers x.

Lemma 3.1. (*explicit construction of the scaled queue-length process*) *Under the assumptions above, for each $\eta \geq 1$ and for each i, $1 \leq i \leq N$, the scaled queue-length process at service group i, $\bar{Q}_i^{(\eta)}(t)$, has the explicit representation*

$$\begin{aligned}
&\bar{Q}_i^{(\eta)}(t) \equiv \eta^{-1} Q_i^{(\eta)}(t) \\
= \;& \eta^{-1}\Pi_i^a\left(\int_0^t \left(\frac{\lambda^{(\eta)} r_i(u, \theta_i^{(\eta)} \int_0^\Delta Q_i^{(\eta)}(u-v)\, dG_i(v))}{\sum_{j=1}^N r_j(u, \theta_j^{(\eta)} \int_0^\Delta Q_j^{(\eta)}(u-v)\, dG_j(v))}\right) du\right) \\
& -\eta^{-1}\Pi_i^s\left(\int_0^t \mu_i(Q_i^{(\eta)}(u) \bigwedge s_i^\eta) du\right) - \eta^{-1}\Pi_i^l\left(\int_0^t \alpha_i (Q_i^{(\eta)}(u) - s_i^\eta)^+ du\right) \\
= \;& \eta^{-1}\Pi_i^a\left(\eta \int_0^t \left(\frac{\lambda r_i(u, \theta_i \int_0^\Delta \bar{Q}_i^{(\eta)}(u-v)\, dG_i(v))}{\sum_{j=1}^N r_j(u, \theta_j \int_0^\Delta \bar{Q}_j^{(\eta)}(u-v)\, dG_j(v))}\right) du\right) \qquad (10) \\
& -\eta^{-1}\Pi_i^s\left(\eta \int_0^t \mu_i(\bar{Q}_i^{(\eta)}(u) \bigwedge s_i) du\right) - \eta^{-1}\Pi_i^l\left(\eta \int_0^t \alpha_i (\bar{Q}_i^{(\eta)}(u) - s_i)^+ du\right).
\end{aligned}$$

Proof. This construction is justified in §2.1 of [2]. The final display follows by simple algebra. □

As can be seen from the first line of (10), the many-server scaling involves scaling $\theta_i^{(n)}$ as in (8). The scaling in [1] corresponds to the second line of (10) in the special case $s_i = \infty$ for all i. (That seems unclear from [1].) In that context, there is scaling inside both Poisson processes, but no scaling of θ. It is clear that these representations are equivalent. We are thus only elaborating on the interpretation.

4. Convergence as $\eta \to \infty$

As noted in [1], the functional strong law of large numbers for a Poisson process can be applied to obtain the desired limit. The first step has

$$\eta^{-1}\Pi(\eta t) \to t \quad \text{as} \quad \eta \to \infty, \qquad (11)$$

uniformly in t over bounded intervals with probability 1, where Π is a unit-rate Poisson process, as in (10).

That yields the limit in [1], extended to the multi-server model. We use function space notation as in [13].

Theorem 4.1. (*law of large numbers, following [1]*) *Given the representation in Lemma* 3.1,

$$\bar{Q}^{(\eta)} \to q \quad as \quad \eta \to \infty \quad w.p.1 \quad in \quad D([0,\infty), \mathbb{R}^N), \tag{12}$$

where q *is the deterministic vector-valued function with*

$$\begin{aligned} q_i(t) &= \int_0^t \left(\frac{\lambda r_i(u, \theta_i \int_0^\Delta q_i(u-v)\, dG_i(v))}{\sum_{j=1}^N r_j(s, \theta_j \int_0^\Delta q_j(u-v)\, dG_j(v))} \right) du - \int_0^t \mu_i (q_i(u) \wedge s_i) du \\ &\quad - \int_0^t \alpha_i (q_i(u) - s_i)^+ du, \quad 1 \le i \le N. \end{aligned} \tag{13}$$

The expression (13) is equivalent to the N-dimensional delay differential equation (system of delay differential equations)

$$\dot{q}_i(t) = \left(\frac{\lambda r_i(t, \theta_i \int_0^\Delta q_i(t-u)\, dG_i(u))}{\sum_{j=1}^N r_j(t, \theta_j \int_0^\Delta q_j(t-u)\, dG_j(u))} \right) - \mu_i (q_i(t) \wedge s_i) - \alpha_i (q_i(t) - s_i)^+ \tag{14}$$

for $1 \le i \le N$ with initial condition $q_i(0) = 0$, $1 \le i \le N$ (under our assumptions).

The expression for the limit in (14) coincides with the delay differential equation in equation (7) of [1] when the model is symmetric, $s_i = \infty$ for all i and we use the special form for the functions r_i in (5) for the special case $(\theta_i, \mu_i) = (\theta, \mu)$ for all i. Note that our representation leads to the asymptotic parameter vector $(\lambda, s_i, \mu_i, \theta_i, \Delta, G_i)$, $1 \le i \le N$. We obtain λ and θ_i from $\lambda^{(n)}$ and $\theta_i^{(n)}$ from (6) and (8).

In summary, we have given a careful specification of the scaling and the resulting interpretation of equation (5) in [1], which serves to justify or explain the limit there. In particular, we have identified $Q_i^\eta(t)$ in [1] with $\bar{Q}^{(\eta)}(t)$ in (7) here. Then the routing probabilities require the scaling in (8). Moreover, our interpretation makes clear that similar many-server limits can be established for generalizations of the model considered in [1], e.g., such as in [14, 15, 16].

5. Extension to Time-Varying Non-Markov Models

In this section we briefly indicate how to obtain corresponding results for the time-varying non-Markov parallel network of N $\cdot/GI/s_t + GI$ models with a single G_t arrival process by modifying the results in [7, 8]. We first consider how to treat the extension of the fluid model in [7].

Initially assume that each arrival is routed to each of the N queues with probability $1/N$. Then each of the N queues is a $G_t/GI/s_t + GI$ model studied in [7]. For each of these models, the performance of the fluid model is characterized by a pair of two-parameter functions $(B(t,y), Q(t,y))$: $B(t,y)$ is the quantity of fluid in service at time t that has been so for a time less than or equal to y; $Q(t,y)$ is the quantity of fluid waiting in queue at time t that has been so for a time less than or equal to y. These key quantities admit integrable representations as in equation (2) of [7]. The paper imposes several regularity conditions, which we assume here as well. Under those regularity conditions, a full algorithm is developed for the performance descriptors of this model. The full algorithm is outlined in §8 of [7].

In order to treat this more general model, we make an additional assumption about the routing function r_i in (3). In particular, we assume that the randomization of the delay takes place over the interval $[\delta, \Delta]$, where $0 < \delta \leq \Delta$; i.e., we have

$$r_i(t, \theta_i \int_{\delta}^{\Delta} Q_i(t-u)\, dG_i(u). \tag{15}$$

Given definition (15), we can apply the algorithm in [7] with the initially specified arrival rate functions $\{\lambda_i(t) : 1 \leq i \leq N\}$ to compute the performance descriptors for each of the N queues. The key observation is that these performance descriptors are valid for the associated fluid model with delayed information over the initial interval $[0, \delta]$. The algorithm from [7] will thus yield the initial candidate N fluid content functions $\{Q_i(t) : 1 \leq i \leq N\}$, which are valid over the time interval $[0, \delta]$. We use these numbers in system to compute new routing probabilities, which will produce new arrival rates that are now valid over the interval $[0, \delta]$, because they incorporate the delayed information.

Then we repeat this calculation to produce new arrival rates and associated new performance descriptors that are valid over the interval $[0, 2\delta]$. We iteratively repeat this process to obtain valid performance descriptors over the interval $[0, k\delta]$ for any desired k.

To establish the associated functional law of large numbers, it suffices to use the same iteration. For the initial model, that is a direct application of [8]. Then we apply that same limit theorem iteratively to obtain the limit over the interval $[0, k\delta]$ for any desired k.

6. Discussion

In this final section we discuss the applied relevance.

6.1. The Role of Infinite-Server Service Groups

At first glance, the formulated routing problem in [1] with infinitely many servers in each service group may not seem very interesting or useful. With the present formulation, the congestion experienced by each customer is just their own service-time distribution. If $\mu_i = \mu$ for all i as assumed in [1], then all service-time distributions are i.i.d. Hence, all routing schemes should be equally good for customers. With infinitely many servers, there should be no concern about how many other customers are already in service. Of course, there might well be other associated measures of congestion associated with a large queue length, but that has not yet been discussed.

Nevertheless, we think that the model with infinite-server groups developed in [1] can be useful. It can be justified through a series of approximations. First, suppose that we actually have service groups with large finite numbers of servers. In that context, it seems natural to regard the customer goal as being to minimize the expected response time. By using Little's law, we see that the expected response time is the expected number in system divided by the arrival rate. So we can translate the goal to minimizing weighted mean queue lengths. Finally, we can approximate the number in system in a large finite server group by the number in system in an associated infinite group. Thus, that is one way that we can employ the infinite-server model to get useful results for finite-server models.

Alternatively, as shown in §4 above, we can directly treat the Erlang-A model, but that produces a new delay differential equation.

6.2. Interpreting the Numerical Examples

In order to understand the numerical examples in [1] and interpret the applied relevance, it is necessary to understand what parameters we should look at. The relevant parameters to interpret the conclusions about the deterministic limiting model in [1] are the asymptotic parameters $(\lambda, \mu_i, \theta_i, \Delta)$.

It is helpful to think of the process of interest being the scaled queue-length process $\bar{Q}^{(\eta)}(t) \equiv \eta^{-1} Q^{(\eta)}(t)$. With that understanding, equation (5) of [1] reduces to the last line in the display in (10) here when $s_i = \infty$ for all i. With that interpretation, the model parameters $(\lambda, \mu, \Delta, \theta)$ do not change as we increase n (η in [1]), while the behavior of $\bar{Q}^{(n)}(t)$ approaches the deterministic limit $q(t)$. Thus, the conclusions for the fluid model should apply directly (as an approximation) to $\bar{Q}^{(n)}(t)$ in the stochastic model. As η increases from 10 to 100 in the numerical examples, we see the improved fit of the limiting deterministic model. (According to Pender (private communication), the numerical examples in §2.3 of [1] all have parameter $\theta = 1$.) In that setting, the derived critical delay Δ_{cr} in the fluid model applies directly to the stochastic model as an approximation. This shows an advantage of the scaling in the second line of (10), which is used by [1]. Lemma 3.1 shows how to translate the conclusions to the unscaled system. That view forces us to scale theta in the routing function by (8).

6.3. Differential Delays

The paper [1] provides a nice study of the symmetric model. It remains to carefully investigate the impact of differential delay information when the delay cdf G_i varies with i. Progress in that direction has been made by [12].

Acknowledgement. I thank Jamol Pender for helpful comments and suggestions.

References

[1] J. Pender, R. Rand, E. Wesson, A stochastic analysis of queues with customer choice and delayed information, Mathematics of Operations Research, published in Articles in Advance, April 10, 2020 (2020).

[2] G. Pang, R. Talreja, W. Whitt, Martingale proofs of many-server heavy-traffic limits for Markovian queues, Probability Surveys 4 (2007) 193–267.

[3] K. W. Fendick, M. A. Rodrigues, A. Weiss, Analysis of a rate-based control strategy with delayed feedback, Performance Evaluation 16 (1992) 67–84.

[4] K. W. Fendick, M. A. Rodrigues, Asymptotic analysis of adaptive rate control for diverse sources with delayed feedback, IEEE Transactions on Information Theory 40 (6) (1994) 2008–2025.

[5] E. Budish, P. Cramton, J. Shim, The high-frequence trading arms race: frequent batch auctions as a market design response, The Quarterly Journal of Economics 130 (4) (2015) 1547–1621.

[6] M. E. Lewis, Flash Boys, A Wall Street Revolt, Norton, 2014.

[7] Y. Liu, W. Whitt, The $G_t/GI/s_t + GI$ many-server fluid queue, Queueing Systems 71 (4) (2012) 405–444.

[8] Y. Liu, W. Whitt, A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading, Operations Research Letters 40 (2012) 307–312.

[9] J. Dong, E. Yom-Tov, G. B. Yom-Tov, The impact of delay announcements on hospital network coordination and waiting times, Management Science 65 (5) (2018) 1969–1994.

[10] S. Novitzky, J. Pender, To update or not to update: queues with information updates, Cornell University, Ithaca NY 14853 (2020).

[11] S. Novitzky, J. Pender, Queues with delayed information: a probabilistic perspective, Cornell University, Ithaca NY 14853 (2020).

[12] P. Doldo, J. Pender, R. Rand, Breaking the symmetry in queues with delayed information, Cornell University, Ithaca NY 14853 (2020).

[13] W. Whitt, Stochastic-Process Limits, Springer, 2002.

[14] I. Gurvich, W. Whitt, Scheduling flexible servers with convex delay costs in many-server service systems, Manufacturing and Service Operations Management 11 (2) (2009) 237–253.

[15] G. Pang, W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues, Queueing Systems 65 (2010) 325–364.

[16] X. Sun, W. Whitt, Delay-based service differentiation with many servers and time-varying arrival rates, Stochastic Systems 8 (3) (2018) 230–263.