Submitted to *Management Science*
manuscript MS-00251-2007

Electronic Companion – Fluid Models for Overloaded Multi-Class Many-Server Queueing Systems with FCFS Routing

Rishi Talreja, Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027
{rt2146, ww2040}@columbia.edu

In this electronic companion, we first discuss transient dynamics for the fluid model discussed in the main paper. Next, we give simulation results for the X model with batch arrivals and the N model. Finally, we discuss open questions related to this work.

1. Transient Dynamics of the Fluid Model

We discuss the transient dynamics of the fluid model, thus elaborating on Conjecture 1 of the main paper. We assume we are given transient routing flow rates for the model and specify dynamics in terms of these flow rates. This section closely follows the discussion of transient dynamics of the single-class single-pool fluid model given in Whitt (2006). We can think of $B_{i,j}(t,y)$ as the amount of class i *fluid* in service in service pool j at time t that has been in service for time less than or equal to y, and $Q_i(t,y)$ as the amount of class i *fluid* in queue at time t that has been in queue for time less than or equal to y. As stated in Conjecture 1 in the main paper, we assume that these functions are continuous with respect to y. We further stipulate that they should have densities b and q:

$$B_{i,j}(t,y) = \int_0^y b_{i,j}(t,u)du \quad \text{and} \quad Q_i(t,y) = \int_0^y q_i(t,u)du, \quad \text{for all} \quad i \in \mathcal{C},\, j \in \mathcal{S}, \quad \text{and} \quad y \geq 0. \tag{1}$$

Also, let $B_{i,j}(t) \equiv B_{i,j}(t,\infty)$ be the total amount of class i fluid in service in service pool j at time t and let $Q_i(t) \equiv Q_i(t,\infty)$ be the total amount of class i fluid in queue at time t, for $i \in \mathcal{C}$ and $j \in \mathcal{S}$. A clear relationship between Q and B is if the class i queue is not empty then all service pools that can serve class i fluid must be full:

$$Q_i(t) > 0 \quad \Longrightarrow \quad \sum_{j \in S(i)} B_{i,j}(t) = 1 \quad \text{for all} \quad t \geq 0, \quad \text{for all} \quad i \in \mathcal{C}. \tag{2}$$

In other words, the model is work-conserving.

Similarly, $R_{i,j}(t)$ can be thought of as the amount of class i fluid that has been routed to pool j by time t. We stipulate that this function also has a density $r_{i,j}$ so that

$$R_{i,j}(t) = \int_0^t r_{i,j}(u)du, \quad \text{for all} \quad t \geq 0, \quad \text{for all} \quad i \in \mathcal{C}, j \in \mathcal{S}. \tag{3}$$

We assume here that the functions $R_{i,j}$, $i \in \mathcal{C}$, $j \in \mathcal{S}$, are given. We do not describe their dynamics here. Our goal here is to specify the dynamics of Q_i and $B_{i,j}$, $i \in \mathcal{C}$, $j \in \mathcal{S}$, in terms of $R_{i,j}$, $i \in \mathcal{C}$, $j \in \mathcal{S}$. In the main paper we discuss how to determine $R_{i,j}$, $i \in \mathcal{C}$, $j \in \mathcal{S}$, when the system is in stationarity.

To describe service and abandonment, we work with hazard rates of the service-time and abandonment-time distributions, which are well defined by the assumption that F and G are absolutely continuous:

$$h_{s,j}(x) \equiv \frac{g_i(x)}{G_i^c(x)} \quad \text{and} \quad h_{a,i}(x) \equiv \frac{f_i(x)}{F_i^c(x)} \quad \text{for } x \geq 0, \quad \text{for all} \quad i \in \mathcal{C} \quad \text{and} \quad j \in \mathcal{S}.$$

(In this context, $0/0$ is naturally to be interpreted as 0.) Clearly, $h_{s,j}(x)$ is the conditional rate of service for a customer in service pool $j \in \mathcal{S}$ that has been in service for x amount of time, conditional on that customer not having been served previously. Similarly, $h_{a,i}(x)$ is the conditional rate of abandonment of a class $i \in \mathcal{C}$ customer that has been in queue for x amount of time, conditional on that customer not having abandoned previously. The total service rate for class-i customers in service pool j at time t is then

$$\sigma_{i,j}(t) \equiv \int_0^\infty b_{i,j}(t,x) h_{s,j}(x) dx, \quad \text{for all} \quad t \geq 0, \tag{4}$$

and the total class-i abandonment rate is

$$\alpha_i(t) \equiv \int_0^\infty q_i(t,x) h_{a,i}(x) dx, \quad \text{for all} \quad t \geq 0. \tag{5}$$

Define the vector versions of the above functions in the obvious way.

For each class $i \in \mathcal{C}$ and pool $j \in \mathcal{S}$, fluid in service at time t that is not served in the next u time units remains in service, giving us the equation

$$b_{i,j}(t+u, x+u) = b_{i,j}(t,x) \frac{G_j^c(x+u)}{G_j^c(x)}, \quad \text{for all} \quad x \geq 0, t \geq 0, \quad \text{and} \quad u > 0. \tag{6}$$

Similarly, for each class $i \in \mathcal{C}$, fluid waiting in queue at time t that does not abandon or go into service in the next u time units, remains in the queue, giving us

$$q_i(t+u, x+u) = q_i(t,x) \frac{F_i^c(x+u)}{F_i^c(x)}, \quad \text{for all} \quad x \geq 0, t \geq 0, \quad \text{and} \quad u > 0, \tag{7}$$

for content that has not moved into service.

If the queue is not empty, class-i fluid moves into service at time t at the rate

$$\nu_i(t) \equiv \sum_{j \in S(i)} r_{i,j}(t), \quad \text{for all} \quad t \geq 0. \tag{8}$$

The fluid that moves into service is always the fluid that has been waiting the longest. Therefore, for each class i and time t, there exists some queue boundary $q_i^b(t)$ such that

$$q_i(t,x) = 0, \quad \text{for all} \quad t \geq 0 \quad \text{and} \quad x > q_i^b(t). \tag{9}$$

We note here that if $W_i(t)$ is the amount of time class-i fluid waits before entering service, then for each $t \geq 0$, $x \geq 0$, $W_i(t)$ satisfies the relationsip $W_i(t - q_i^b(t)) = q_i^b(t)$.

Also, for each class $i \in \mathcal{C}$, the corresponding queue will receive new input at the rate λ_i as long as class i input to the system can not go directly into service. Therefore,

$$q_i(t,0) = \lambda_i, \quad \text{for all} \quad t \geq 0 \text{ such that } q_i^b(t) > 0. \tag{10}$$

Furthermore, for the case where the class $i \in \mathcal{C}$ queue is empty but all eligible service pools are busy we have

$$q_i(t,0) = \lambda_i - \nu_i(t), \quad \text{for all} \quad t \geq 0 \text{ such that } \sum_{l \in C(j)} B_{l,j}(t) = 1 \quad \text{for all} \quad j \in S(i) \quad \text{and} \quad Q_i(t) = 0. \tag{11}$$

In this case λ_i units of class $i \in \mathcal{C}$ fluid enter the system in one unit of time but in this unit of time $\nu_i(t)$ units of fluid enter service directly. The rest of the fluid joins the class i queue.

We also have directly from the definitions of $b(t,x)$ and $r(t)$,

$$b_{i,j}(t,0) = r_{i,j}(t), \quad \text{for all} \quad t \geq 0, \quad \text{for all} \quad i \in \mathcal{C}, j \in \mathcal{S}. \tag{12}$$

We now conjecture that, given R, the equations above fully specify the dynamics of B and Q.

CONJECTURE 1. *Given R such that $R_{i,j}(0) = 0$ for all $i \in \mathcal{C}$, $j \in \mathcal{S}$, there exist unique continuous functions B and Q satisfying the description in (1)-(12) and $B_{i,j}(0,y) = Q_i(0,y) = 0$ for all $y \geq 0$, $i \in \mathcal{C}$, and $j \in \mathcal{S}$.*

We show in §7 of the main paper that it is in general not possible to determine the rates R without taking into account stochastic properties of the pre-limit queueing model. This is why we assume above that R is given.

2. Additional Simulation Results

In this section we present additional simulation results. We consider the X model with batch arrivals and the N model. As in the main paper, all results in this section were computed by simulating 100,000 arrivals to the queueing system, disregarding the initial transient 20%. Also, each of the simulated values are given as 95% confidence intervals.

2.1. X Model with Batch Arrivals

The parameters we used for our X model simulations are:

$$\theta_1 = \theta_2 = \frac{1}{2}, \quad \lambda_1 = 2000,\ \lambda_2 = 3000, \quad \eta_1 = 1,\ \eta_2 = 2, \quad s_1 = s_2 = 1000.$$

Table 1 X **model simulation results with batch arrivals for various combinations of arrival processes and service time distributions.**

Interarrival / Service	Param.	*EXP*	*GAMMA*	*HYPEREXP*	*UNIFORM*	*TWOPOINT*	*CONSTANT*
EXP	p_{11}	0.3331 ± 0.0032	0.3336 ± 0.0028	0.3327 ± 0.0027	0.3335 ± 0.0040	0.3345 ± 0.0030	0.3326 ± 0.0035
	p_{21}	0.3337 ± 0.0021	0.3336 ± 0.0028	0.3338 ± 0.0026	0.3330 ± 0.0035	0.3319 ± 0.0024	0.3340 ± 0.0020
GAMMA	p_{11}	0.3332 ± 0.0023	0.3332 ± 0.0033	0.3328 ± 0.0016	0.3337 ± 0.0029	0.3322 ± 0.0035	0.3337 ± 0.0023
	p_{21}	0.3331 ± 0.0026	0.3331 ± 0.0018	0.3330 ± 0.0032	0.3330 ± 0.0019	0.3331 ± 0.0021	0.3335 ± 0.0020
HYPEREXP	p_{11}	0.3337 ± 0.0047	0.3335 ± 0.0043	0.3329 ± 0.0024	0.3348 ± 0.0044	0.3324 ± 0.0048	0.3322 ± 0.0038
	p_{21}	0.3341 ± 0.0025	0.3339 ± 0.0032	0.3324 ± 0.0027	0.3331 ± 0.0031	0.3327 ± 0.0057	0.3324 ± 0.0027
UNIFORM	p_{11}	0.3331 ± 0.0038	0.3327 ± 0.0025	0.3330 ± 0.0038	0.3331 ± 0.0032	0.3323 ± 0.0015	0.3341 ± 0.0032
	p_{21}	0.3334 ± 0.0024	0.3333 ± 0.0028	0.3337 ± 0.0022	0.3331 ± 0.0021	0.3335 ± 0.0023	0.3332 ± 0.0019
TWOPOINT	p_{11}	0.3324 ± 0.0026	0.3342 ± 0.0042	0.3344 ± 0.0044	0.3338 ± 0.0030	0.3334 ± 0.0031	0.3340 ± 0.0027
	p_{21}	0.3330 ± 0.0012	0.3335 ± 0.0020	0.3328 ± 0.0029	0.3337 ± 0.0015	0.3333 ± 0.0027	0.3330 ± 0.0021
CONSTANT	p_{11}	0.3340 ± 0.0030	0.3334 ± 0.0028	0.3346 ± 0.0036	0.3324 ± 0.0034	0.3327 ± 0.0045	0.3343 ± 0.0035
	p_{21}	0.3339 ± 0.0024	0.3341 ± 0.0025	0.3331 ± 0.0028	0.3340 ± 0.0024	0.3338 ± 0.0027	0.3327 ± 0.0024

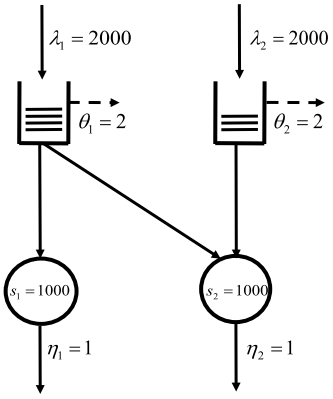
These parameters are the same as in §9.2 of the main paper, but now arrivals occur in batches of size 2. The fluid approximation gives $p_{1,1} = p_{2,1} = 1 - p_{1,2} = 1 - p_{2,2} = 1/3$. Table 1 gives our simulation results.

The rows and columns are labelled as in Table 2 of the main paper. Notice that in all cases we get results consistent with our fluid approximation. This gives us further evidence that the arrival distributions do not play a role in the fluid approximations.

2.2. N Model

In this section we simulate an N model with parameters given in Figure 1, and with exponential interarrival, abandonment, and service distributions. Since the N model is sparsely connected,

Figure 1 N **model simulation parameters.**



we compute our fluid approximation as in §5 of of the main paper. We give numerical results for

Table 2 Simulation results for the N model.

	$p_{1,1}$	$p_{1,2}$	$p_{2,2}$	w_1	w_2	Q_1	Q_2
Fluid Approx.	1	0	1	0.346	0.346	499.4	499.4
Simulation	0.975 ± .007	0.025 ± .007	1	0.335 ± .004	0.359 ± .005	490.7 ± 7.9	512.9 ± 7.0

both the fluid approximation and the simulation of the queueing model in Table 2. We find this N model interesting because our fluid approximation gives $p_{1,2} = 0$. The simulation also gives us $p_{1,2} \approx 0$, but the error is larger than in our previous examples. We attribute the small deviation from 0 here to not being close enough to the fluid limit (r may not be large enough). We believe that a stochastic refinement to our fluid model may explain the deviation and provide an improved approximation.

For recent work on the N model, but under a different routing policy, see Tezcan and Dai (2006).

3. Future Work

There are a number of interesting questions we would like to answer about multi-class many-server queueing systems with FCFS routing and corresponding fluid models. Here we list some of them.

- **When global FCFS fails.** We have yet to describe the stationary behavior in the sparsely connected and hybrid cases when global FCFS fails. We have shown that this occurs when the characterizing linear system yields negative flows. Evidently, the system behavior can be described by setting some of these negative flows to zero and analyzing the separate components of the resulting decomposed network, but we have yet to determine precisely what happens. We conjecture that system behavior can be accurately described by decomposing the network by eliminating a subset of the arcs for which the corresponding flows $r_{i,j}$ the are initially not strictly positive. It remains to develop a systematic procedure to determine the proper decomposition. This problem occurs when there are initially two negative flows. We have given examples with two intial negative flows for which it is appropriate to delete (i) one of the two arcs corresponding to the two negative flows, and (ii) both of these arcs.

- **More general routing graphs.** We have only described how to compute the flow rates for sparsely connected and fully connected routing graphs, and combinations of these. We have yet to

analyze other routing graphs.

- **Proof of fluid limit.** It remains to prove that the steady-state fluid equations are asymptotically correct. In particular, it remains to prove the conjectures here and in the main paper. Perhaps this can be done first in discrete-time as in Whitt (2006) and then extended to continuous-time. Alternatively, perhaps it can be done directly for the fully Markovian model, with Poisson arrivals, exponential service, and exponential abandonment. This has been done in the single-class single-pool case in Whitt (2004).

- **Class-dependent service.** We have restricted our analysis to systems where service-time distributions depend only on the service pool. We would like to be able to extend the analysis to systems where service times depend on both service pool and customer class. Toward this end, the recent QED limits by Gurvich and Whitt (2006) may provide guidance.

- **Stochastic refinement.** It remains to develop stochastic refinements to the fluid approximations. Stochastic refinements would give us a better second-order understanding of the queueing systems we have considered.

- **State-space collapse**. A more tractable first step would be to determine when the queueing systems considered here admit state-space collapse. Positive results could potentially be established using results in Dai and Tezcan (2006), which rely on hydrodynamic scaling as described in Bramson (1998). State-space collapse results would facilitate arguments towards stochastic refinements. We have reason to believe that state-space collapse does occur here because of the fact that in the fluid limit virtual wait times at all queues are the same. See Gurvich and Whitt (2006) for related work.

References

Bramson, M. 1998. State space collapse with applications to heavy traffic limits for multiclass queueing networks. *Queueing Systems* **30** 89–148.

Dai, J. G., T. Tezcan. 2006. State space collapse in many-server diffusion limits of parallel server systems. Working Paper, Georgia Institute of Technology, `http://www.isye.gatech.edu/~dai/publications/preprints/daiTezcanSSC.pdf`.

Gurvich, I., W. Whitt. 2006. Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. Working Paper, Columbia University, `http://www.columbia.edu/~ww2040`.

Tezcan, T., J. G. Dai. 2006. Dynamic control of n-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. Working Paper, Georgia Institute of Technology, `http://www.isye.gatech.edu/~dai/publications/preprints/NModel821.pdf`.

Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Sci.* **50**(10) 1449–1461.

Whitt, W. 2006. Fluid models with multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.