Join-Idle-Queue system with general service times: Large-scale limit of stationary distributions

Sergey Foss
Heriot-Watt University, Edinburgh, UK
and Novosibirsk State University
s.foss@hw.ac.uk

Alexander L. Stolyar
University of Illinois at Urbana-Champaign
Urbana, IL
stolyar@illinois.edu

ABSTRACT

A parallel server system with n identical servers is considered. The service time distribution has a finite mean $1/\mu$, but otherwise is arbitrary. Arriving customers are to be routed to one of the servers immediately upon arrival. Join-Idle-Queue routing algorithm is studied, under which an arriving customer is sent to an idle server, if such is available, and to a randomly uniformly chosen server, otherwise. We consider the asymptotic regime where $n \to \infty$ and the customer input flow rate is λn. Under the condition $\lambda/\mu < 1/2$, we prove that, as $n \to \infty$, the sequence of (appropriately scaled) stationary distributions concentrates at the natural equilibrium point, with the fraction of occupied servers being constant equal λ/μ. In particular, this implies that the steady-state probability of an arriving customer having to wait for service vanishes.

1. INTRODUCTION

We consider a parallel server system consisting of n servers, processing a single input flow of customers. The service time of any customer by any server has the same distribution with finite mean $1/\mu$. Each customer has to be assigned (routed) to one of the servers immediately upon arrival. (This model is sometimes referred to as "supermarket" model.) We study a Join-Idle-Queue routing algorithm, under which an arriving customer is sent to an idle server, if such is available; if there are no idle servers, a customer is sent to one of the servers chosen uniformly at random.

We consider an asymptotic regime such that $n \to \infty$ and the input rate is λn, where the system load $\lambda/\mu < 1$. Thus, the system remains subcritically loaded. Under the additional assumption that the service time distribution has *decreasing hazard rate* (DHR), it is shown in [10] that the following property holds.

Asymptotic optimality: As $n \to \infty$, the sequence of the system stationary distributions is such that the fraction of occupied servers converges to constant λ/μ; consequently, the steady-state probability of an arriving customer being routed to a non-idle server vanishes.

The results of [10] apply to far more general systems, where servers may be non-identical. However, the analysis in [10] does rely in essential way on the DHR assumption on the service times; under this assumption the system process has *monotonicity* property, which is a powerful tool for analysis. Informally speaking, monotonicity means that two versions of the process, such that the initial state of the first one is dominated (in the sense of some natural partial order) by that of the second one, can be coupled so that this dominance persists at all times.

When the service time distribution is general, the monotonicity under JIQ no longer holds, which requires a different approach to the analysis. In the present work we prove the following fact.

Main result (Theorem 2 in Section 2): *The asymptotic optimality holds for an arbitrary service time distribution, if the system load $\lambda/\mu < 1/2$.*

We believe that condition $\lambda/\mu < 1/2$ is purely technical (required for the proof in this paper) and that our main result in fact holds for $\lambda/\mu < 1$, i.e. as long as the system is stable. This will be discussed in more detail in Section 3.

The key feature of the JIQ algorithm (as well as more general *pull-based* algorithms [1, 6, 10, 11]), is that it does not utilize any information about the current state of the servers besides them being idle or not. This allows for a very efficient practical implementation, requiring very small communication overhead between the servers and the router(s) [6, 10, 11]. In fact, in the asymptotic regime that we consider, JIQ is much superior to the celebrated "power-of-d-choices" (or Join-Shortest-Queue(d), or JSQ(d)) algorithm [12, 7, 2, 3], in terms of both performance and communication overhead (see [10, 11] for a detailed comparison). The JSQ(d) algorithm routes a customer to the shortest queues among the $d \geq 1$ servers picked uniformly at random.

We note that when the service time distribution is general, there is no monotonicity under JSQ(d) (just like under JIQ in our case), and this also makes the analysis far more difficult. Specifically, the result for JSQ(d), which is a counterpart of our main result for JIQ, is Theorem 2.3 in [2], which shows the asymptotic independence of individual server states. (Our main result also implies asymptotic independence of server states; see formal statement in Corollary 3.) Theorem 2.3 in [2] imposes even stronger assumptions than ours, namely a finite second moment of the service time and load $\lambda/\mu < 1/4$ (for non-trivial values of d, which are $d \geq 2$); our Theorem 2 only requires a finite first moment of the service time and load $\lambda/\mu < 1/2$.

In a different asymptotic regime, so called Halfin-Whitt regime (when the system capacity exceeds its load by $O(\sqrt{n})$, as opposed to $O(n)$), and Markov assumptions (Poisson input flows and exponentially distributed service times), JIQ has been recently analyzed in [4, 8]. These papers study diffusion limits of the system transient behavior; Markov assumptions appear to be essential for the analysis. Finally, we mention a recent paper [9], which proposes and studies a version of JIQ for systems with *packing constraints* at the servers.

Basic notation. We say that a function is RCLL if it is *right-continuous with left-limits*. Symbol $\Rightarrow$ signifies convergence of random elements in distribution. Indicator of event or condition B is denoted by $\mathbf{I}(B)$.

2. MODEL AND MAIN RESULT

We consider a service system, consisting of n parallel servers. The system is homogeneous in that all servers are identical, with the same customer service time distribution, given by the cdf $F(\xi), \xi \geq 0$. This distribution has finite mean, which WLOG can be assumed to be 1:

$$\int_0^\infty F^c(\xi) = 1, \quad \text{where } F^c(\xi) \doteq 1 - F(\xi).$$

Otherwise, the cdf $F(\cdot)$ is arbitrary. The service/queueing discipline at each server is arbitrary, as long as it is work-conserving and non-idling.

Customers arrive as a Poisson process. (This assumption can be relaxed to a renewal arrival process; see Section 4.) The arrival rate is λn, where $\lambda < 1$, so that the system load is strictly subcritical.

The routing algorithm is Join-Idle-Queue (JIQ), which is defined as follows. (The JIQ algorithm can be viewed, in particular, as a specialization of the PULL algorithm [10, 11] to a homogeneous system with "single router.")

Definition 1 (JIQ). *An arriving customer is routed to an idle server, if there is one available. Otherwise, it is routed to server chosen uniformly at random.*

We consider the sequence of systems with $n \to \infty$. From now on, the upper index n of a variable/quantity will indicate that it pertains to the system with n servers, or n-th system. Let $W_i^n(t)$ denote the workload, i.e. unfinished work, in queue i at time t in the n-th system. Consider the following *fluid-scaled* quantities:

$$x_w^n(t) \doteq (1/n) \sum_i \mathbf{I}\{W_i^n(t) > w\}, \quad w \geq 0. \tag{1}$$

That is, $x_w^n(t)$ is the fraction of servers i with $W_i^n(t) > w$. Then $x^n(t) = (x_w^n(t),\ w \geq 0)$ is the system state at time t; $\rho^n(t) \equiv x_0^n(t)$ is the fraction of busy servers (the instantaneous system load).

For any n, the state space of the process $(x^n(t),\ t \geq 0)$ is a subset of a common (for all n) state space $\mathcal{X}$, whose elements $x = (x_w,\ w \geq 0)$ are non-increasing RCLL functions of w, with values $x_w \in [0,1]$. This state space $\mathcal{X}$ is equipped with Skorohod metric, topology and corresponding Borel σ-algebra.

Then, for any n, process $x^n(t),\ t \geq 0$ is Markov with state space $\mathcal{X}$, and sample paths being RCLL functions (with values in $\mathcal{X}$), which are in turn elements of (another) Skorohod space.

Stability (positive Harris recurrence) of the process $(x^n(t),\ t \geq 0)$, for any n, is straightforward to verify. Indeed, as long as a server remains busy, it receives each new arrival with probability at most $1/n$, and therefore receives the new work at the average rate at most λ. Thus, the process has unique stationary distribution. Let $x^n(\infty)$ be a random element whose distribution is the stationary distribution of the process; in other words, this is a random system state in stationary regime.

The system *equilibrium point* $x^* \in \mathcal{X}$ is defined as follows. Let $\Phi^c(w)$ denote the complementary (or, tail) distribution function of the steady-state residual service time; the latter is the steady-state residual time of a renewal process with renewal time distribution function $F(\cdot)$. We have

$$\Phi^c(w) = \int_w^\infty F^c(\xi) d\xi, \quad w \geq 0.$$

Then,

$$x^* = (x_w^* = \lambda \Phi^c(w),\ w \geq 0) \in \mathcal{X}.$$

In particular, the equilibrium point is such that "the fraction of occupied servers" $x_0^* = \lambda$. Our main result is the following

Theorem 2. *If $\lambda < 1/2$, then $x^n(\infty) \Rightarrow x^*$ as $n \to \infty$.*

The detailed proof of Theorem 2 can be found in [5]. In this short paper we will only give (in Section 3) a high level description of the proof approach and the intuition for condition $\lambda/\mu < 1/2$.

Theorem 2 shows, in particular, that if $\lambda < 1/2$, then as $n \to \infty$ the steady-state probability of an arriving customer waiting for service (or sharing a server with other customers) vanishes. It also implies the following

Corollary 3. *Assume $\lambda < 1/2$. Suppose that JIQ is completely symmetric with respect to the servers. Specifically, if at the time of a customer arrival there are idle servers, the customer is routed to one of them chosen uniformly at random. Then the states of individual servers in stationary regime are asymptotically independent. Moreover, for any fixed m, the stationary distribution of $(W_1^n, \ldots, W_m^n)$ converges to that of $(\widetilde{W}_1, \ldots, \widetilde{W}_m)$, with i.i.d. components such that $\mathbb{P}\{\widetilde{W}_1 > w\} = x_w^* = \lambda \Phi^c(w),\ w \geq 0$.*

Indeed, by symmetry with respect to servers, the stationary distribution of $(W_1^n, \ldots, W_m^n)$, i.e. of the residual work on the fixed set of servers $1, \ldots, m$, is same as that on a set of m servers, *chosen uniformly at random.* But, $x^n(\infty)$, which describes the overall distribution of server workloads in the system, converges in distribution to the non-random point x^*. This implies Corollary 3.

3. PROOF APPROACH AND DISCUSSION OF CONDITION $\lambda < 1/2$

The approach we use to establish the convergence of stationary distributions in Theorem 2 is as follows. (The detailed proof is in [5].) We find a set $A \in \mathcal{X}$ and a fixed finite time T, such that, with high probability, for all large n, (a) $x^n(\infty) \in A$ and (b) $x^n(0) \in A$ implies that $x^n(T)$ is close to x^*. Property (b) is key. When n is large, the trajectory $x^n(t)$ is "almost deterministic." (In fact, the problem reduces to the analysis of "fluid limit" trajectories, which are the limits of $x^n(t)$ as $n \to \infty$.) Then, informally speaking, property (b) above reduces to the property (b'): trajectories $x^n(t)$ converge to x^* as $t \to \infty$. The absence of process monotonicity (described in Section 1) makes proving (b') difficult. We now describe – very informally – the key idea, which we use in our proof of convergence (b'), and which relies on the condition $\lambda < 1/2$.

Suppose n is large. Consider an initial state $x^n(0)$, such that *the total amount of (fluid-scaled, i.e. multiplied by $1/n$) unfinished work is upper bounded by $C < \infty$*. Pick α such that $\alpha > \lambda$ and $\alpha + \lambda < 1$; this can be done if and only if $\lambda < 1/2$. Then, at some finite time τ, the system must reach a state with αn servers being idle. (Otherwise, if at least $(1-\alpha)n$ servers would continue to be busy as time goes to infinity, the unfinished work would become negative, since $1-\alpha > \lambda$.) Denote by S_α the set of those αn servers, which are idle at time τ. Starting time τ, WLOG, assume that all new arriving customers go to an idle server in S_α, as long as there is one available. Consider the subsystem, consisting only of the servers in S_α; starting time τ and until the (random) time when *all* servers in S_α become busy, the behavior of this subsystem is obviously equivalent to that of the infinite-server system, $M/GI/\infty$, with idle initial state. If n is large, the behavior of $x^n(t)$ *for such $M/GI/\infty$ system* is "almost deterministic" and such that the (scaled) number of occupied servers $x_0^n(t)$ in it is "almost monotone increasing, converging to $\lambda < \alpha$" and, moreover, $x^n(t)$ "converges" to x^*. But this means that after time τ the subsystem S_α will "always" have idle servers, which in turn means that *its* state will "converge" to x^* as $t \to \infty$. Also, after time τ, the subsystem consisting of the servers outside S_α will "never" receive any new arrivals and will "eventually" empty. Thus, $x^n(t)$ for our entire system "converges" to x^*.

Turning the key intuition, described above informally, into a formal proof is the subject of paper [5]. Set $A \in \mathcal{X}$ is picked by using a constructed uniform in n upper bound on the stationary distribution of the workload of an individual server. The states in A are such that the total (scaled) workload is not necessarily upper bounded by a constant C (in fact, if the second moment of the service time is infinite, the steady-state total workload in the system is infinite with probability 1); however, for states in A the (scaled) workload is bounded by C on a close-to-1 fraction of servers – this suffices for the proofs. The property (b') is proved uniformly for fluid limits starting from A – from here we obtain that (b) holds for the pre-limit processes with high probability, uniformly for all large n.

As explained above, the proof of Theorem 2 relies in essential way on condition $\lambda < 1/2$. However, we believe that this condition is purely technical, and Theorem 2 in fact holds for any $\lambda < 1$. Establishing this fact will most likely require a different proof approach, although some elements of the analysis in [5] may turn out to be useful for the proof of a more general result.

4. GENERALIZATIONS

The following generalizations of Theorem 2 hold. Here we only state them – see [5] for a detailed discussion.

Renewal arrival process. Theorem 2 and its proof easily generalize to the case when the arrival process is renewal; namely, when in the n-th system the interarrival times are i.i.d., equal in distribution to A/n, where A is a positive random variable, $\mathbb{E}A = 1/\lambda$. (Mild assumptions on the interarrival time distribution are needed. For example, it suffices that this distribution has an absolutely continuous component.) The common process state space contains an additional scalar variable u, which is the residual interarrival time; clearly $u^n(\infty) \Rightarrow 0$ as $n \to \infty$. The more general form of Theorem 2 is as follows:
If $\lambda < 1/2$, then $(u^n, x^n)(\infty) \Rightarrow (0, x^*)$.

Biased routing when all servers are busy. The specific rule – uniform at random – for routing arriving customers when all servers are busy, can be relaxed as follows. It suffices that the arrival rate into a server *when it is busy* is upper bounded by some $\bar{\lambda} < 1$; this holds, for example, when the probability that any busy server receives an arrival does not exceed $(1/n)(\bar{\lambda}/\lambda)$ for some $\bar{\lambda} < 1$. When this condition holds – in addition to $\lambda/\mu < 1/2$ – Theorem 2 holds as is.

Finite buffers. Suppose, we allow some or all servers to have finite buffers (of same or different sizes). If a server has finite buffer of size $B \geq 1$, and already has B customers, any new customer routed to to this server is blocked and leaves the system. For such more general system Theorem 2 holds as is.

5. REFERENCES

[1] Badonnel, R. and Burgess, M. (2008). Dynamic pull-based load balancing for autonomic servers. *Network Operations and Management Symposium, NOMS 2008*, 751–754.

[2] Bramson, M., Lu, Y., and Prabhakar, B. (2012). Asymptotic independence of queues under randomized load balancing. *Queueing Systems* **71**, 247–292.

[3] Bramson, M., Lu, Y., and Prabhakar, B. (2013). Decay of tails at equlibrium for fifo join the shortest queue networks. *The Annals of Applied Probability* **23**, 1841–1878.

[4] Eschenfeldt, P. and Gamarnik, D. (2015). Join the shortest queue with many servers. the heavy traffic asymptotics. arXiv:1502.00999.

[5] Foss, S. and Stolyar, A. L. (2016). Large-scale Join-Idle-Queue system with general service times. http://arxiv.org/abs/1605.05968

[6] Lu, Y., Xie, Q., Kliot, G., Geller, A., Larus, J., and Greenberg, A. (2011). Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation* **68**, 1057–1071.

[7] Mitzenmacher, M. (2001). The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems* **12**, 10, 1094–1104.

[8] Mukherjee, D., Borst, S., van Leeuwaarden, J., and Whiting, P. (2015). Universality of load balancing schemes on diffusion scale. arXiv:1510.02657.

[9] Stolyar, A. L. (2017). Large-scale heterogeneous service systems with general packing constraints. *Advances in Applied Probability* **49**, 1. arXiv:1508.07512.

[10] Stolyar, A. L. (2015). Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems* **80**, 4, 341–361.

[11] Stolyar, A. L. (2016). Pull-based load distribution among heterogeneous parallel servers: the case of multiple routers. *Queueing Systems*. DOI 10.1007/s11134-016-9508-8. arXiv:1512.07873.

[12] Vvedenskaya, N., Dobrushin, R., and Karpelevich, F. (1996). Queueing system with selection of the shortest of two queues: an asymptotic approach. *Problems of Information Transmission* **32**, 1, 20–34.