Explicit Stationary Comparisons for Join-the-Shortest-Queue Approximations

Alper Ferudun*

28 September 2026

Abstract

We compare the stationary occupancy law of a finite join-the-shortest-queue system with its finite-buffer and sampled-routing approximations. For N unit-rate exponential servers, Poisson arrival rate $0 < \lambda < N$, and $\rho = \lambda/N$, regeneration gives the explicit constant $H = ((1-\rho)^{-N}-1)/\lambda$. If every queue has capacity b and the blocking probability is B_b, the total variation error is at most $\lambda H(bN+1)B_b$. An explicit bound on B_b and a matching-order lower bound show that this error has order λ^{bN+1} as λ tends to zero with N, b fixed. For shortest-of-k routing we obtain an explicit total variation bound for both sampling conventions. These estimates answer the two comparison requests in an AIM problem for all bounded occupancy observables. The constants can be loose or trivial at high load or large N; no sharp many-server approximation is asserted.

1 Model, statements, and prior work

There are $N \geq 1$ identical queues, with independent unit-rate exponential service times and a Poisson arrival stream of total rate $\lambda \in (0, N)$. Full join-the-shortest-queue routing (JSQ) chooses a shortest queue, uniformly among ties. Its labelled state is $x = (x_1, \ldots, x_N) \in \mathbb{Z}^N_{\geq 0}$ and its occupancy state is

$$q_i(x) = \#\{j : x_j \geq i\} \quad (i \geq 1), \qquad q_0 = N, \qquad J(x) = \sum_{j=1}^{N} x_j = \sum_{i \geq 1} q_i(x).$$

Write π for its stationary occupancy law. Let π_b be the stationary law when every queue has capacity $b \geq 1$, an integer: an arrival is blocked only when all queues have length b. Let $\pi^{(k)}$ denote the stationary law under shortest-of-k routing, where the k queue labels are sampled uniformly, either independently with replacement or as a uniform subset without replacement. These are different models. Ties among the sampled shortest queues are resolved symmetrically; the occupancy transition law does not depend on that choice.

All laws are embedded in the same countable occupancy space. We use $d_{\mathrm{TV}}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$ and put

$$\rho = \frac{\lambda}{N}, \qquad H = \frac{(1-\rho)^{-N} - 1}{\lambda}, \qquad D_b = \min\left\{ \frac{\rho}{1+\rho}, e^{-bN(\log(1/\rho) - 1 + \rho)} \right\}. \tag{1}$$

*Mercury Software GmbH. Email: alper@mercurycodelab.com. GitHub: AlperTheKing.

Theorem 1 (Finite buffers)**.** *For every $N \geq 1$, $b \geq 1$ and $0 < \lambda < N$, let B_b be the stationary blocking probability of the capacity-b chain. Then*

$$B_b \leq D_b, \tag{2}$$

$$d_{\mathrm{TV}}(\pi, \pi_b) \leq \min\{1, \lambda H(bN+1) B_b\} \leq \min\{1, \lambda H(bN+1) D_b\}, \tag{3}$$

$$d_{\mathrm{TV}}(\pi, \pi_b) \geq (1-\rho)^N \left(\frac{\lambda}{\lambda + N}\right)^{bN+1}. \tag{4}$$

In particular, with N, b fixed, $d_{\mathrm{TV}}(\pi, \pi_b) = \Theta_{N,b}(\lambda^{bN+1})$ as $\lambda \downarrow 0$.

Theorem 2 (Sampled routing)**.** *Define*

$$\epsilon_k = \begin{cases} ((N-1)/N)^k, & \textit{with replacement, } k \geq 1, \\ (N-k)/N, & \textit{without replacement, } 1 \leq k \leq N. \end{cases}$$

Then

$$d_{\mathrm{TV}}(\pi, \pi^{(k)}) \leq \min\left\{1, \lambda H\left(\frac{\lambda}{1-\rho} + 2\right) \epsilon_k\right\}. \tag{5}$$

For any bounded real occupancy function f, each upper bound also bounds the corresponding expectation difference after multiplication by $\mathrm{osc}(f) = \sup f - \inf f$. Conversely, indicator functions recover total variation. Without replacement, (5) is zero at $k = N$, as it must be. With replacement, even $k = N$ can miss all shortest queues.

Problem 1.35, attributed to Han Gan in the AIM list for the 2018 workshop on Stein's method [1], asks for explicit expectation comparisons with capacity two and with shortest-of-k routing. Theorems 1–2 give a finite-system answer for every bounded occupancy observable, and extend capacity two to arbitrary b. They do not address every possible test class or limiting regime.

The methods are standard: weak-majorization comparison, regeneration, and a Poisson-equation generator identity. Weak-majorization optimality of JSQ is classical; see, for example, [2]. Ma–Maguluri [3] use hitting times to obtain a two-server transient convergence rate. We extend their one-extra-customer induction to arbitrary fixed N and use it for stationary perturbation bounds. Generator comparison and coupling-based Stein factors for JSQ appear in Braverman [4]; that work treats a diffusion limit, not the present finite-system comparison. The many-server universality results of [5] are likewise in a different regime and are not improved by our constants. We give elementary proofs of the comparison and hitting-time estimates needed here.

2 Recurrence and moments

Let A be the full-JSQ generator and $A^{(k)}$ a sampled-routing generator on labelled states. If L is the selected queue length, then for either rule

$$\mathbb{E}[L \mid x] \leq J(x)/N, \qquad \mathbb{E}[a^L \mid x] \leq N^{-1} \sum_j a^{x_j} \quad (a > 1).$$

For sampled routing, compare to the first uniform sample; for full JSQ, compare the minimum to the average. Choose $1 < a < 1/\rho$ and set $W(x) = \sum_j a^{x_j}$. With $n_0(x)$ empty queues,

$$\begin{aligned} AW(x) &\leq \rho(a-1)W(x) - (1 - 1/a)(W(x) - n_0(x)) \\ &\leq -cW(x) + d, \\ c &= (a-1)(1/a - \rho) > 0, \qquad d = N(1 - 1/a). \end{aligned} \tag{6}$$

The same inequality holds for $A^{(k)}$. Both chains are irreducible: any state can drain to zero, and balanced states can be reached by successive shortest-queue arrivals; from a sufficiently high balanced state, suitable services reach any prescribed state. Each required finite transition sequence has positive probability. The jump rates are bounded by $\lambda + N$, so the chains are nonexplosive. The Foster drift (6), with finite sublevel sets, gives positive recurrence.

The stationary moment claim need not assume integrability of W. For $W_M = W \wedge M$, concavity gives

$$AW_M \le \mathbf{1}_{\{W<M\}}(-cW + d).$$

Its stationary mean is zero because W_M is bounded. Hence $c\mathbb{E}_\pi[W\mathbf{1}_{\{W<M\}}] \le d$; monotone convergence gives $\mathbb{E}_\pi W \le d/c$. The same argument applies to $\pi^{(k)}$. In particular, all polynomial moments are finite.

Since $AJ = \lambda - q_1$, stationarity yields $\mathbb{E}_\pi q_1 = \lambda$. For $V(x) = \sum_j x_j^2$ we also have

$$AV(x) \le -2(1-\rho)J(x) + \lambda + q_1(x).$$

These identities hold for sampled routing as well. Taking stationary means gives

$$\mathbb{E}_\pi J \le \frac{\lambda}{1-\rho}, \qquad \mathbb{E}_{\pi^{(k)}} J \le \frac{\lambda}{1-\rho}. \tag{7}$$

All uses of an unbounded stationary generator identity here and below are justified by the established exponential moments: both the function and its absolute jump increments are integrable. With bounded rates, absolute summation of the stationary balance equations then gives $\nu Ah = 0$. The finite capacity chain is irreducible and has a unique stationary distribution without any moment issue.

3 Empty-state comparison and hitting times

Lemma 3. *The full-JSQ empty-state probability satisfies $\pi(0) \ge (1-\rho)^N$.*

Proof. Sort queue lengths in descending order. Put $S_l(x) = \sum_{i=1}^{l} x_i$ and write $x \preceq y$ if $S_l(x) \le S_l(y)$ for all $1 \le l \le N$. Two elementary transition maps preserve this order.

First, add one job to a shortest coordinate of x and to any coordinate of y, then sort. It suffices to consider a shortest coordinate of y, since it minimizes every resulting upper partial sum. Write $r_x = \#\{i : x_i > \min x\}$. The partial sum $S_l(x)$ increases exactly when $l > r_x$. A violation therefore requires $S_l(x) = S_l(y)$, an increase on the x side, and none on the y side. Set $m = \min x$. Then $x_l = \cdots = x_N = m$, and the inequality at $l-1$ gives $y_l \le m$. The inequality at N, together with equality at l, forces $y_{l+1} = \cdots = y_N = m$. Thus $y_l = \min y = m$, a contradiction. For $l = N$ both arrival maps increase the sum, so failure is impossible.

Second, apply a potential service to the same rank j in both vectors, decreasing that coordinate if positive, and sort. For a positive coordinate, S_l decreases precisely when its equal-value block ends at or before l. A violation requires equality at l and a decrease only on the y side. Necessarily $j \le l$ and $y_j > 0$. If $x_j = 0$, equality at l contradicts the inequality at $j-1$. Otherwise the plateau in x extends beyond l; write $x_j = \cdots = x_{l+1} = u$. The inequality at $l+1$ gives $y_{l+1} \ge u$, while equality at l and the inequality at $j-1$ give $\sum_{i=j}^{l} y_i \le (l-j+1)u$. Descending order forces all these entries and y_{l+1} to equal u. The y plateau also extends past l, a contradiction. At $l = N$ the plateau case cannot occur.

Now couple full JSQ and uniform random routing, starting empty, using common arrivals and independent rate-one potential service clocks at each rank. These have the correct sorted marginals: the departure rate from a length class is its number of queues. The two transition facts give $J_{\mathrm{JSQ}}(t) \le J_{\mathrm{random}}(t)$ pathwise. The random-routing stationary law is that of N independent $M/M/1$ queues of load ρ. Passing to stationarity gives the claim. This is weak-majorization comparison, not coordinatewise domination of labelled JSQ queues by labelled randomly routed queues. $\square$

Lemma 4. *For full JSQ, the hitting time τ_0 of the empty state satisfies $\mathbb{E}_x\tau_0 \le HJ(x)$.*

Proof. Let h_1 be the mean hitting time from a state with one job. A regenerative cycle at zero has an idle interval of mean $1/\lambda$ followed by a busy excursion of mean h_1. Positive recurrence and the renewal reward identity give

$$h_1 = \frac{\pi(0)^{-1} - 1}{\lambda} \le H.$$

Two labelled JSQ chains preserve coordinatewise order using common labelled service clocks and a common uniform random priority permutation for tie breaking at each arrival. Indeed, an arrival could violate $x \le y$ only at a coordinate j with $x_j = y_j$, chosen by the lower chain but not the upper, which chooses i. Shortest-queue selection gives $x_j \le x_i \le y_i \le y_j = x_j$. Thus both i, j are minima in both chains, contradicting their common priority order. Services preserve the order directly.

The excess total population cannot increase in this coupling. If $y = x + e_j$, then when the lower chain empties the upper contains at most one job. The strong Markov property gives $\mathbb{E}_y\tau_0 \le \mathbb{E}_x\tau_0 + h_1$. Induction on $J(y)$ proves the lemma. This is the one-job induction from [3], used here for general N. $\square$

4 An explicit Poisson-equation estimate

For a bounded occupancy function f, also regard f as a symmetric function of labelled states. Normalize $\operatorname{osc}(f) \le 1$, and let P_t be the full-JSQ semigroup. Couple chains from x and from $Y \sim \pi$ below the chain from their coordinatewise maximum, using the coupling of Lemma 4. The lower chains coalesce by the time the upper one empties. Hence

$$\int_0^\infty |P_t f(x) - \pi f|\, dt \le H(J(x) + \mathbb{E}_\pi J) < \infty. \tag{8}$$

Here a stationary labelled realization is used for Y; its occupancy law is π. Define

$$g(x) = -\int_0^\infty (P_t f(x) - \pi f)\, dt$$

and subtract its value at zero. Finite-neighbor summation, absolute convergence in (8), and the backward equation imply

$$Ag = f - \pi f. \tag{9}$$

The endpoint at infinity is πf, by positive recurrence (or the same coupling). Comparing x with zero gives $|g(x)| \le HJ(x)$.

The same argument for neighboring states gives

$$|g(x + e_i) - g(x)| \le H(J(x) + 1), \tag{10}$$

$$|g(x + e_i) - g(x + e_j)| \le H(J(x) + 2). \tag{11}$$

For (10), the upper initial state is $x + e_i$. For (11), dominate both states by $x + e_i + e_j$; if $i = j$ the difference is zero. Lemma 4 bounds the corresponding coupling times. Symmetry ensures that these estimates also hold for the associated admissible occupancy increments. The linear growth of g and (7) justify all stationary generator comparisons below.

5 Finite-buffer comparison and light traffic

Proof of Theorem 1. Let $\mathbf{b} = (b, \dots, b)$ and let A_b be the finite-chain generator. It differs from full JSQ only at an arrival in state $\mathbf{b}$. Thus (9) and $\pi_b A_b g = 0$ give the exact identity

$$\pi_b f - \pi f = \lambda B_b\big[g(\mathbf{b} + e_1) - g(\mathbf{b})\big]. \tag{12}$$

Apply (10) with $J(\mathbf{b}) = bN$ and take the supremum over indicator functions. This proves the first bound in (3).

In sorted or occupancy coordinates, let η be the state with one queue of length $b-1$ and all others of length b. Balance at the all-full state gives $NB_b = \lambda\pi_b(\eta)$. Since $B_b + \pi_b(\eta) \le 1$, we obtain $B_b \le \rho/(1+\rho)$. Also, the finite-buffer population has arrival intensity at most λ and departure intensity $q_1 \ge J/b$. It is dominated by an immigration-death chain with birth rate λ and death rate z/b in state z. This coupling is valid even though J alone need not be Markov: at equality of the populations, pair each death of the dominating chain with a finite-buffer departure; extra finite-buffer departures and blocked births only improve the order. The dominating stationary law is Poisson with mean $b\lambda$. Exponential Markov inequality, optimized at $e^\theta = 1/\rho$, gives

$$B_b \le \mathbb{P}\{\text{Poisson}(b\lambda) \ge bN\} \le \exp[-bN(\log(1/\rho) - 1 + \rho)].$$

This proves (2) and the remaining upper bound.

For the lower bound, let z_m be the balanced occupancy state reached from zero by m successive arrivals without service, for $0 \le m \le bN+1$. The transition $z_{m-1} \to z_m$ has rate λ. Stationary point balance and the total exit-rate bound $\lambda + N$ therefore give

$$\pi(z_m) \ge \frac{\lambda}{\lambda + N}\pi(z_{m-1}).$$

Lemma 3 bounds $\pi(z_0)$ from below. The state z_{bN+1} has a queue of length $b+1$ and lies outside the support of π_b, so its probability proves (4). Finally, $H \to 1$ as $\lambda \downarrow 0$, and

$$e^{-bN(\log(1/\rho)-1+\rho)} = e^{bN - b\lambda}(\lambda/N)^{bN}.$$

The upper and lower bounds have the same power λ^{bN+1} for fixed N, b. □

6 Sampled-routing comparison

Proof of Theorem 2. For an occupancy state q, put $s = \min\{i \ge 1 : q_i < N\}$. Full JSQ increments q_s. The sampled arrival increments q_i with probability

$$p_i(q) = \begin{cases} (q_{i-1}/N)^k - (q_i/N)^k, & \text{with replacement,} \\ \dfrac{\binom{q_{i-1}}{k} - \binom{q_i}{k}}{\binom{N}{k}}, & \text{without replacement.} \end{cases}$$

Use $\binom{u}{k} = 0$ when $u < k$. In particular, $p_i = 0$ for $i < s$. The probability of missing every globally shortest queue is

$$a_k(q) = \sum_{i>s} p_i(q) = \begin{cases} (q_s/N)^k, & \text{with replacement,} \\ \binom{q_s}{k}/\binom{N}{k}, & \text{without replacement,} \end{cases} \qquad a_k(q) \le \epsilon_k,$$

because $q_s \le N-1$. All service terms cancel in the generator comparison, and (9) gives

$$\pi^{(k)} f - \pi f = \lambda \mathbb{E}_{\pi^{(k)}} \sum_{i>s} p_i(q)[g(q+e_s) - g(q+e_i)]. \tag{13}$$

Each pair of successors in the sum arises by adding one job to two queues of the same labelled starting state. Consequently, (11) and (7) yield

$$|\pi^{(k)} f - \pi f| \le \lambda H \mathbb{E}_{\pi^{(k)}}[(J+2)a_k(q)] \le \lambda H \epsilon_k \left(\frac{\lambda}{1-\rho} + 2\right).$$

Taking the supremum over indicators and capping by one proves the theorem. □

7 Scope and reproducibility

The finite-buffer blocking probability can be found by solving a finite stationary linear system. For example, $N = 2$, $b = 2$, $\rho = 1/10$ gives $B_2 = 1/5861$. The fully explicit bound using D_2 is less than 0.0042923856; using B_2 in (3) is sharper. These are error bounds, not measured values of the infinite-system distance. For $N = 1$, the exact finite-buffer distance is ρ^{b+1}, consistent with the asserted light-traffic order.

The accompanying checker performs 126,180 finite transition, drift, generator, and stationary-distribution checks with rational arithmetic. Two Python runtimes produce the same report. Floating-point displays of the Chernoff expression are diagnostics, not interval certificates. Finite capacity-four versus capacity-two computations do not certify an infinite-state error. The proofs above, rather than the finite tests, establish the infinite-state conclusions.

The constants in (1) are not uniform in N or near $\rho = 1$. The capped upper bounds may equal one; no sharp Halfin–Whitt rate, unbounded-observable estimate, or exact light-traffic leading coefficient is claimed. A bounded literature search did not locate these paired explicit bounds and the general-capacity light-traffic statement; folklore or an earlier formulation is not excluded. No absolute priority claim is made.

This unrefereed note was prepared with AI assistance in derivation, literature search, exact-arithmetic checking, and exposition. The supporting record includes an originating-researcher self-audit; no independent peer review or proof-assistant verification is claimed.

References

[1] American Institute of Mathematics, *Stein's method and applications in high-dimensional statistics*, 2018 workshop problem list, Problem 1.35 (Han Gan). Archived AIM Problem Lists, accessed 28 September 2026.

[2] O. T. Akgun, R. Righter, and R. Wolff, Multiple-server system with flexible arrivals, *Advances in Applied Probability* **43** (2011), 985–1004. doi:10.1239/aap/1324045695.

[3] Y. Ma and S. T. Maguluri, Convergence rate of the join-the-shortest-queue system, *Journal of Applied Probability*, first published online 9 September 2026. arXiv:2503.15736.

[4] A. Braverman, The Join-the-Shortest-Queue System in the Halfin–Whitt Regime: Rates of Convergence to the Diffusion Limit, *Stochastic Systems* **13** (2023), 1–39. doi:10.1287/stsy.2022.0102.

[5] D. Mukherjee, S. C. Borst, J. S. H. van Leeuwaarden, and P. A. Whiting, Universality of Power-of-d Load Balancing in Many-Server Systems, *Stochastic Systems* **8** (2018), 265–292. doi:10.1287/stsy.2018.0016.