

Steady-State Limit Interchange for Join-the-Shortest-Queue in the Nondegenerate Slowdown Regime

Wei You

July 21, 2026

Abstract

We establish steady-state validity of the nondegenerate slowdown diffusion approximation for the stationary marginal of normalized total population in a homogeneous Markovian parallel-server system operating under join-the-shortest-queue, with one queue at each server. Without abandonment, the finite-horizon diffusion limit and its invariant law were known, but stationary convergence remained open because irrevocable routing allows idle servers and waiting jobs to coexist. A Palm analysis of marked excursions controls stationary idleness in high-population states; combined with a state-space-collapse estimate, this yields population tail control. We prove convergence of the stationary normalized-population marginal, interchange of the two iterated limits, and convergence of stationary moments. We also study customer abandonment with mean patience proportional to system size. From balanced initial queue lengths we prove the normalized-population diffusion limit, and separately prove convergence of the stationary normalized-population marginal. Abandonment adds a linear mean-reverting drift and gives Gaussian-type decay in the limiting invariant density without altering the lower-boundary effect of idleness.

Keywords. join-the-shortest-queue, nondegenerate slowdown, interchange of limits, steady-state diffusion approximation, customer abandonment.

1 Introduction

Join-the-shortest-queue (JSQ) is a canonical load-balancing policy for parallel-server systems with irrevocable dispatching and no job migration. We study a homogeneous Markovian system in which each of a growing number of unit-rate servers has its own queue and every arrival joins a shortest queue. The offered load grows with the number of servers while aggregate spare capacity remains fixed. Utilization therefore approaches one without making congestion negligible: waiting times remain comparable to service times, and customers experience a nontrivial slowdown. This is the *nondegenerate slowdown* (NDS) regime. It differs from the Halfin–Whitt regime, where spare capacity grows on the square-root scale and waiting becomes negligible relative to service (Halfin and Whitt, 1981; Eschenfeldt and Gamarnik, 2018; Braverman, 2020).

Atar (2012) introduced NDS diffusion models for pooled many-server systems. For the JSQ system studied here, Gupta and Walton (2019) proved that the normalized total population, observed on the NDS time scale and started from balanced queue lengths, converges to a one-dimensional diffusion. Its drift contains an idleness term that becomes singular near the lower boundary and vanishes once the normalized population reaches two. They also identified the diffusion’s invariant law.

That finite-horizon theorem does not determine the equilibrium behavior of the finite systems. Balanced initial states form only a small part of the irreducible state space, and process convergence supplies neither uniform control of the prelimit stationary laws nor averaging of idleness under those laws. The obstruction is structural. If several servers are idle, aggregate service capacity is reduced by exactly that amount, so even a large total population need not have negative drift. Irrevocable routing permits waiting jobs and idle servers to coexist. In a pooled many-server queue, work conservation makes the total customer count a one-dimensional Markov process; with one queue at each server, the total population is not Markov because its departure rate also depends on idleness. A steady-state proof must therefore control the joint occurrence of a large population and substantial idleness.

We prove that the NDS diffusion approximation is asymptotically exact for the stationary marginal of normalized total population. That marginal converges to the diffusion invariant law, the two iterated limits agree, and stationary polynomial moments and a range of exponential moments converge. We also analyze a model with exponential patience times of rate γ/N . From balanced initial queue lengths we establish its normalized-population diffusion limit. Furthermore, we prove convergence and moment convergence for the stationary normalized populations; together, these results yield interchange of the two iterated limits. Abandonment adds a linear mean-reverting drift but leaves unchanged the averaged contribution of idleness responsible for the lower-boundary behavior.

Methodologically, the proof separates stationary occupation control from stochastic averaging. A Palm analysis of marked deficit excursions, based on Campbell’s formula, controls integrated idleness in high-population states. Together with the bound for the perpendicular component, these occupation estimates yield a uniform exponential-moment bound for the stationary normalized population. A finite- N Poisson equation for the frozen idleness chain then removes the centered fast variable from the stationary generator. The resulting basic adjoint relationship (BAR) identifies the limit in the interior, pooled comparison excludes boundary mass, and zero probability flux determines the invariant density.

1.1 Related literature

NDS diffusion models and the direct predecessor. Atar (2012) introduced the NDS regime for pooled many-server systems with heterogeneous exponential servers. Huang et al. (2016) subsequently developed diffusion limits for pooled many-server queues with general patience-time distributions in both the Halfin–Whitt and NDS regimes. The direct predecessor to the present

paper is Gupta and Walton (2019), who established the NDS diffusion limit and its invariant law for the same JSQ system without abandonment. Their result is a finite-horizon limit from balanced initial states; convergence of the finite-system stationary laws was left open.

Many-server JSQ limits. For JSQ in the Halfin–Whitt regime, Eschenfeldt and Gamarnik (2018) proved the diffusion limit, while Braverman (2020, 2023) established steady-state convergence and rates. Zhao et al. (2025) obtained process and stationary limits in the super-Halfin–Whitt window, and Bhambay et al. (2025) proved a diffusion limit and interchange result for heterogeneous speed-aware JSQ in the Halfin–Whitt regime. These papers establish steady-state validity under other many-server scalings, but they do not address the NDS equilibrium problem with fixed aggregate spare capacity.

Customer abandonment. For pooled many-server queues, Dai et al. (2010) established diffusion limits with general patience, Dai et al. (2014) proved steady-state convergence and interchange in the Halfin–Whitt regime, and Gurvich et al. (2014) developed steady-state approximations for the Erlang–A queue across loading regimes, including NDS. For systems with one queue at each server, Cao et al. (2021) compared pooled and separate-queue architectures through a fluid model in a moderately overloaded regime with abandonment. Jhunjhunwala et al. (2025) characterized steady-state queue-vector limits for a fixed number of discrete-time JSQ queues as the abandonment probability vanishes. Wang and Maguluri (2026) obtained stationary Gaussian approximations and tail bounds for a fixed-dimensional heterogeneous JSQ system in a heavily overloaded, vanishing-abandonment regime. The closest JSQ-abandonment results are stationary asymptotics with a fixed number of queues, while the closest many-server abandonment results concern pooled or fluid models. To the best of our knowledge, no prior work proves both the scalar NDS diffusion limit and steady-state limit interchange for normalized total population in many-server JSQ with abandonment.

Steady-state methods. Jhunjhunwala et al. (2024) proved the uniform exponential-moment bound for the component perpendicular to the state-space-collapse manifold used here, but their stronger total-queue estimate requires a scale condition that fails when normalized slack is of order one over the number of servers. More broadly, steady-state interchange has been proved for generalized Jackson networks through Lyapunov and strong-approximation methods (Gamarnik and Zeevi, 2006; Budhiraja and Lee, 2009) and through the BAR approach (Braverman et al., 2017). Poisson equations have also been used to average fast modulation in JSQ systems (Hurtado-Lange and Grosof, 2026). These methods do not directly resolve the present problem because the population drift depends on fast idleness dynamics before averaging and becomes singular near the lower boundary. Our proof combines Palm/Campbell occupation control with a finite-system Poisson equation to address these two obstacles.

1.2 Main contributions

The paper has three limit results and two methodological contributions.

1. **Steady-state validity without abandonment.** Theorem 4 and Corollaries 1 and 2 show that the stationary marginal of normalized total population converges to the explicit invariant law of the NDS diffusion, that the two iterated limits agree, and that all polynomial moments and a nontrivial range of exponential moments converge. Thus the diffusion obtained from balanced initial states remains exact for this scalar equilibrium marginal, although the finite-system invariant law is supported on all queue configurations.
2. **Steady-state validity with abandonment.** For exponential patience times of rate γ/N , Proposition 4, Theorem 6, and Corollary 3 identify the invariant density, prove convergence of the stationary normalized-population marginal, and establish moment convergence and interchange. The abandonment term changes the upper-tail behavior of the limiting invariant density from exponential to Gaussian-type decay.
3. **NDS diffusion limit with abandonment.** From balanced initial queue lengths, Theorem 5 proves the functional normalized-population limit. Its proof requires localized occupancy-profile estimates, finite-system averaging, joint Brownian identification, and removal of localization.
4. **Stationary occupation-time control.** A Palm analysis of marked excursions of the deficit process converts subcritical excursion dynamics into bounds on idleness in high-population states without requiring a regenerative decomposition of the full system. Combined with a uniform exponential-moment bound for the perpendicular component, these estimates yield the uniform exponential-moment and tail bounds in Theorem 2.
5. **Finite-system stochastic averaging.** Theorem 3 replaces the idleness process, under the prelimit invariant law, by its state-dependent equilibrium mean in the population drift. The finite- N Poisson equation reduces the generator comparison to a weighted approximation of the actual fast transition rate and bounds on the increments of its solution.

Organization. Section 2 states the baseline and abandonment models, the principal results, and the proof architecture. Sections 3 to 6 develop the common stationary argument in proof order: state-space collapse and excursion estimates, exponential-moment control, stochastic averaging, and limit identification. Section 7 proves the diffusion limit with abandonment. The appendices contain fixed-system recurrence and moment arguments, Palm-construction details, localization proofs, and auxiliary estimates.

Notation conventions. For $x \in \mathbb{R}$, $x^+ = \max\{x, 0\}$. We use $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$, $a \wedge b = \min\{a, b\}$, and $a \vee b = \max\{a, b\}$. The vector e_i is the i th coordinate vector, $\mathbf{1}_N = (1, \dots, 1)$, and $\mathbb{1}_A$ is the indicator of A . For open $U \subset \mathbb{R}$, $C_c^k(U)$ denotes the real-valued C^k functions with compact support in U ; $C_c(U) = C_c^0(U)$, and $C_c^\infty(U)$ is the corresponding smooth test-function space. The notation

$\text{Law}(Y)$ is the law of Y , $\Rightarrow$ denotes weak convergence, and $U \leq_{\text{st}} V$ means $\mathbb{P}(U > x) \leq \mathbb{P}(V > x)$ for all x . Unless stated otherwise, $O(\cdot)$, $o(\cdot)$, and $\asymp$ refer to $N \rightarrow \infty$; their constants may depend on fixed model parameters but not on N , and $a_N \asymp b_N$ means that a_N/b_N is bounded above and below by positive constants.

2 Model and main results

2.1 Finite-system dynamics and stationary identities

Fix $\beta > 0$. For each integer $N > \beta$, consider N identical single-server queues. Arrivals form a Poisson process of rate

$$\lambda_N = N - \beta.$$

Service requirements are independent exponential random variables of mean one. Each nonempty queue is served at rate one, buffers are infinite, and an arrival joins a shortest queue, with ties broken uniformly. For every process-level result, the initial queue vector is independent of the future arrival, service, tie-breaking, and abandonment primitives. We use the completed right-continuous filtration generated by the initial state and those primitives.

Let $Q^N(t) = (Q_1^N(t), \dots, Q_N^N(t)) \in \mathbb{Z}_+^N$ be the queue-length vector. The generator is

$$\mathcal{L}_N F(q) = \lambda_N \frac{1}{|A(q)|} \sum_{i \in A(q)} [F(q + e_i) - F(q)] + \sum_{i: q_i > 0} [F(q - e_i) - F(q)],$$

where $m(q) = \min_{1 \leq i \leq N} q_i$ and $A(q) = \{i : q_i = m(q)\}$ for $q \in \mathbb{Z}_+^N$. The total jump rate is at most $2N$, so the chain is nonexplosive.

For $j \geq 0$, introduce the exact-level counts $M_j^N(q) = \#\{i : q_i = j\}$, and, for $j \geq 1$, the occupancy counts $Z_j^N(q) = \#\{i : q_i \geq j\}$. Along the process, use the same symbols for the corresponding evaluations: $M_j^N(t) = M_j^N(Q^N(t))$ and $Z_j^N(t) = Z_j^N(Q^N(t))$. For a deterministic state $q \in \mathbb{Z}_+^N$, write

$$I(q) = M_0^N(q) = \#\{i : q_i = 0\}, \quad S(q) = \sum_{i=1}^N q_i = \sum_{j \geq 1} Z_j^N(q).$$

Along the process, set $I^N(t) = I(Q^N(t))$ and $S^N(t) = S(Q^N(t))$. A deterministic argument q denotes evaluation of a coordinate function at state q , as in $M_j^N(q)$, $Z_j^N(q)$, $I(q)$, and $S(q)$; a time argument denotes evaluation along the stochastic process. When no confusion is possible, the superscript N is suppressed.

The following results introduce basic properties of the queue-length process. The proofs are collected in Section A.1.

Proposition 1 (Positive recurrence). *For every $N > \beta$, the chain Q^N is irreducible and positive recurrent and therefore has a unique stationary distribution π_N .*

Let $Q^N(\infty)$ have law π_N , with $I^N(\infty) = I(Q^N(\infty))$ and $S^N(\infty) = S(Q^N(\infty))$. In stationary calculations we suppress “ ∞ ” and write Q^N , I^N , S^N , and $X_N = S^N/N$. The symbol “ ∞ ” denotes

a stationary copy, not an almost-sure path limit. Unless another law is specified, unqualified probabilities and expectations of finite- N variables without time arguments are taken under π_N .

Lemma 1 (Exponential moments for fixed N). *For every $N > \beta$ and every z satisfying $1 < z < N/(N - \beta)$, one has*

$$\mathbb{E}_{\pi_N} \sum_{i=1}^N z^{Q_i^N} < \infty. \quad (1)$$

Consequently, every polynomial moment of S^N is finite.

Lemma 2 (Stationary generator identity). *If $F : \mathbb{Z}_+^N \rightarrow \mathbb{R}$ satisfies $|F(q)| + |\mathcal{L}_N F(q)| \leq C(1 + S(q)^p)$ for some $C, p < \infty$, then F and $\mathcal{L}_N F$ are π_N -integrable and*

$$\mathbb{E}_{\pi_N}(\mathcal{L}_N F)(Q^N) = 0.$$

Applying Lemma 2 to $F = S$ gives the exact flow balance

$$\mathbb{E}_{\pi_N} I^N = \beta, \quad (2)$$

because the birth and death rates of S^N are $N - \beta$ and $N - I^N$, respectively. This fixed first moment of idleness is used throughout the tightness and averaging arguments.

2.2 NDS scaling and the limiting diffusion

Define the accelerated, normalized total population

$$\hat{X}^N(t) = \frac{S^N(Nt)}{N}.$$

A martingale representation is

$$\hat{X}^N(t) = \hat{X}^N(0) + \mathcal{M}^N(t) - \beta t + \frac{1}{N} \int_0^{Nt} I^N(s) \, ds,$$

where $\mathcal{M}^N$ is the martingale term with respect to the natural filtration. The only nontrivial drift term is the integrated idleness.

Set

$$r(x) = (2 - x)^+, \quad g(x) = \frac{r(x)}{1 - r(x)} = \begin{cases} \frac{2 - x}{x - 1}, & 1 < x < 2, \\ 0, & x \geq 2. \end{cases}.$$

Whenever $h \in C_c((1, \infty))$ is multiplied by g , we extend h by zero to $\mathbb{R}$ and use the convention $h(x)g(x) = 0$ for any $x \leq 1$.

For $1 < x < 2$, a balanced occupancy profile contains approximately $N(2 - x)$ one-job queues and $N(x - 1)$ two-job queues. Service completions from one-job queues create idleness at rate approximately $N(2 - x)$, while arrivals remove idleness at rate approximately N . The corresponding fast $M/M/1$ chain has mean $(2 - x)/(x - 1) = g(x)$.

The following is the diffusion limit of Gupta and Walton (2019), written in the present normalization.

Theorem 1 (Gupta–Walton NDS limit). *Suppose that the initial queues are balanced,*

$$\mathbb{P}\left(\max_i Q_i^N(0) - \min_i Q_i^N(0) \leq 1\right) = 1,$$

and that $\widehat{X}^N(0) \Rightarrow X(0)$ with $\mathbb{P}(X(0) > 1) = 1$. Then, in the topology of uniform convergence on compact time intervals,

$$\widehat{X}^N \Rightarrow X,$$

where X is the pathwise unique solution of

$$dX(t) = [g(X(t)) - \beta] dt + \sqrt{2} dB(t), \quad X(t) > 1, \quad (3)$$

and B is a standard one-dimensional Brownian motion. The boundary 1 is inaccessible.

The next proposition verifies positive recurrence and identifies the invariant law through the scale and speed measures; it also records the ergodic convergence used in the interchange argument.

Proposition 2 (Invariant density of the NDS diffusion). *The diffusion (3) is nonexplosive and positive recurrent. Its unique invariant probability density is*

$$p_\beta(x) = C_\beta \begin{cases} (x-1)e^{-(\beta+1)(x-2)}, & 1 < x \leq 2, \\ e^{-\beta(x-2)}, & x \geq 2, \end{cases} \quad \text{where } C_\beta = \frac{\beta(1+\beta)^2}{1+\beta e^{1+\beta}}. \quad (4)$$

From every initial law supported on $(1, \infty)$, the law of $X(t)$ converges to this invariant law as $t \rightarrow \infty$.

The proof is deferred to Section C.1. The two pieces of (4) correspond to the two drift regimes: above two the density has exponential rate β , as in the pooled $M/M/N$ queue under NDS scaling, while near one it vanishes linearly because $g(x) \sim (x-1)^{-1}$. It remains to show that the finite-system stationary laws converge to this invariant law.

2.3 Steady-state results without abandonment

Let ν_N be the stationary law of $X_N = S^N/N$, and let ν_β have density (4). The proof separates compactness from identification. The first theorem supplies both tightness and the uniform integrability needed for moment convergence.

Theorem 2 (Uniform exponential-moment and tail bounds). *Fix $\beta > 0$. For every*

$$0 < a < \min\{\beta, 1/288\}, \quad (5)$$

there exist $N_0 < \infty$ and $C_a < \infty$ such that

$$\sup_{N \geq N_0} \mathbb{E}_{\pi_N} e^{aS^N/N} \leq C_a. \quad (6)$$

Consequently,

$$\sup_{N \geq N_0} \mathbb{P}_{\pi_N} \left(\frac{S^N}{N} > x \right) \leq C_a e^{-ax}, \quad x \geq 0, \quad (7)$$

and the sequence $\{\nu_N\}$ is tight.

Tightness alone is insufficient because the stationary generator still contains the fast variable I^N . The next theorem replaces it, against test functions supported away from the lower boundary, by the averaged value determined by X_N .

Theorem 3 (Stationary stochastic averaging). *For every $h \in C_c^1((1, \infty))$,*

$$\mathbb{E}_{\pi_N} [h(X_N)I^N - h(X_N)g(X_N)] \longrightarrow 0. \quad (8)$$

Together, the two theorems reduce steady-state convergence to identification: subsequential limits satisfy the diffusion BAR on $(1, \infty)$, a pooled comparison excludes mass at one, and zero-flux uniqueness determines the law.

Theorem 4 (Steady-state NDS limit). *For every fixed $\beta > 0$,*

$$\frac{S^N(\infty)}{N} \Rightarrow X_\beta,$$

where X_β has density (4). Equivalently, we have $\nu_N \Rightarrow \nu_\beta$.

Together with fixed-system ergodicity, the diffusion limit in Theorem 1 and the diffusion ergodicity in Proposition 2 then give the interchange statement.

Corollary 1 (Interchange of limits). *Let $\{Q^N(0)\}$ be an initial sequence satisfying the hypotheses of Theorem 1. Then*

$$\lim_{N \rightarrow \infty} \lim_{t \rightarrow \infty} \text{Law} \left(\frac{S^N(Nt)}{N} \right) = \nu_\beta = \lim_{t \rightarrow \infty} \lim_{N \rightarrow \infty} \text{Law} \left(\frac{S^N(Nt)}{N} \right).$$

Moreover, for an arbitrary sequence of finite-system initial laws on $\mathbb{Z}_+^N$, the steady-state-then-many-server limit on the left still equals ν_β ; the balanced-initial-state condition is needed only for the many-server-then-long-time limit on the right.

Corollary 2 (Convergence of stationary moments). *For every integer $k \geq 1$,*

$$\mathbb{E}_{\pi_N} \left(\frac{S^N}{N} \right)^k \longrightarrow \mathbb{E} X_\beta^k.$$

More generally, if $0 < \eta < a$ for an a satisfying (5), then

$$\mathbb{E}_{\pi_N} e^{\eta S^N/N} \longrightarrow \mathbb{E} e^{\eta X_\beta}.$$

2.4 Abandonment model: diffusion and steady-state results

Fix $\gamma > 0$. For $q \in \mathbb{Z}_+^N$, let

$$W(q) = \sum_{i=1}^N (q_i - 1)^+ = S(q) - N + I(q)$$

be the number of waiting customers. We attach the superscript (N, γ) to the processes from Section 2.1; the superscript identifies the abandonment model but does not change the definitions of M_j , Z_j , I , S , or W . For $N > \beta$, let π_N^γ denote the stationary law whose existence is established in Proposition 3, and set

$$X_N^\gamma = \frac{S^{N,\gamma}(\infty)}{N}, \quad \widehat{X}^{N,\gamma}(t) = \frac{S^{N,\gamma}(Nt)}{N}.$$

In the abandonment model, each waiting customer abandons independently at rate γ/N ; customers in service do not abandon, and all other primitives are unchanged. In physical time, service times are exponential with rate one and patience times are exponential with rate γ/N . Under the accelerated clock $t \mapsto Nt$, these rates become N and γ , respectively. Operationally, this scaling represents increasingly patient customers: the mean patience time is N/γ in physical time. Abandonment is therefore rare at the individual-customer level. Nevertheless, when the waiting population is of order N , the aggregate abandonment rate $(\gamma/N)W^{N,\gamma}$ is of order one. This flow is on the same scale as the fixed spare-capacity parameter β , so it has a first-order effect on the normalized population and its stationary distribution.

The generator of $Q^{N,\gamma}$ is

$$\mathcal{L}_N^\gamma F(q) = \mathcal{L}_N F(q) + \frac{\gamma}{N} \sum_{i=1}^N (q_i - 1)^+ [F(q - e_i) - F(q)]. \quad (9)$$

Consequently, the population drift is

$$(\mathcal{L}_N^\gamma S)(q) = I(q) - \beta - \frac{\gamma}{N} W(q). \quad (10)$$

This normalization corresponds to the $\alpha = 1/2$ specialization of the pooled NDS formulation in Huang et al. (2016, Section 3.3 and Theorem 3.3). Set their sequence index to $n = N^2$ and their base service rate to $\mu = 1$. Their system then has $n^\alpha = N$ servers, per-server service rate $\mu^n = n^{1-\alpha}\mu = N$, and diffusion scale $\sqrt{n} = N$, exactly as on our accelerated clock. The accelerated arrival rate is $\lambda^n = N(N - \beta) = n - \beta\sqrt{n}$, so their traffic parameter $\sqrt{n}(\lambda^n/(n\mu) - 1)$ equals $-\beta$. With the fixed patience distribution $F(x) = 1 - e^{-\gamma x}$, their no-scaling condition gives $f(x) = \gamma x$ and hence the abandonment drift $-\mu f(y/\mu) = -\gamma y$ for their scaled waiting-population variable y . In the present model, on the accelerated clock,

$$y^N(t) = \frac{W^{N,\gamma}(Nt)}{N} = \widehat{X}^{N,\gamma}(t) - 1 + \frac{I^{N,\gamma}(Nt)}{N}.$$

The localized idleness estimate proved below shows that the final term vanishes, yielding the limiting abandonment drift $-\gamma(X^\gamma(t) - 1)$.

An abandonment changes the normalized population by $-1/N$. Its contribution to predictable quadratic variation is therefore of order N^{-1} and vanishes in the limit; abandonment changes the limiting drift but does not create an additional Brownian term. Although the physical-time arrival rate is $O(N)$, the aggregate abandonment rate is $O(1)$, so only an $O(N^{-1})$ fraction of arrivals

abandon. This order-one flow nevertheless competes with the fixed spare-capacity parameter β and changes the normalized-population limit.

Under stationarity, the population drift identity gives

$$\mathbb{E}_{\pi_N^\gamma} I^{N,\gamma} = \beta + \frac{\gamma}{N} \mathbb{E}_{\pi_N^\gamma} W^{N,\gamma}.$$

Thus stationary idleness accounts for both the fixed spare capacity and the loss of waiting customers through abandonment.

The extension begins with the pathwise order that supplies stability and transfers upper bounds from the model without abandonment. For $q \in \mathbb{Z}_+^N$, let $q^\uparrow$ be its nondecreasing rearrangement and write $q \preceq q'$ when $q_i^\uparrow \leq (q'_i)^\uparrow$ for every i . Thus $q \preceq q'$ precisely when the sorted queue-length vectors are ordered componentwise.

Proposition 3 (Finite-system stability and monotone coupling). *For every $N > \beta$, the chain $Q^{N,\gamma}$ is irreducible, nonexplosive, and positive recurrent and therefore has a unique stationary law π_N^γ . If $q \preceq q'$, there is a coupling of the abandonment system started from q and the system without abandonment started from q' such that*

$$Q^{N,\gamma}(t) \preceq Q^{N,0}(t), \quad t \geq 0, \quad (11)$$

almost surely. In particular, π_N^γ is stochastically dominated by π_N with respect to $\preceq$.

The baseline steady-state theorem is proved independently. Under abandonment, the diffusion and stationary limits require distinct compactness arguments. For the diffusion limit, we localize the normalized population on compact subintervals of $(1, \infty)$, establish the required occupancy-profile estimates, average the fast idleness variable through the finite- N Poisson equation, identify the Brownian driver in the joint limiting filtration, and then remove localization. For the stationary limit, Proposition 3 transfers the baseline exponential-moment control, while abandonment-modified marked-excursion bounds and the idleness-weighted approximation of the one-job queue count yield stationary averaging. A comparison with the pooled $M/M/N + M$ queue then excludes boundary mass and completes the zero-flux identification.

Theorem 5 (NDS diffusion limit with abandonment). *Suppose that the initial queues are balanced in the sense of Theorem 1 and that*

$$\widehat{X}^{N,\gamma}(0) \Rightarrow X(0), \quad \mathbb{P}(X(0) > 1) = 1.$$

Then, in the topology of uniform convergence on compact time intervals,

$$\widehat{X}^{N,\gamma} \Rightarrow X^\gamma, \quad (12)$$

where $X^\gamma(0)$ has law $X(0)$, B is a standard Brownian motion independent of $X^\gamma(0)$, and X^γ is the pathwise unique nonexplosive solution of

$$dX^\gamma(t) = [g(X^\gamma(t)) - \beta - \gamma(X^\gamma(t) - 1)] dt + \sqrt{2} dB(t), \quad X^\gamma(t) > 1. \quad (13)$$

The boundary 1 is inaccessible.

Remark 1 (Role of balanced initialization). *Exact balance supplies the four initial-layer controls used in the proof: it places the initial occupancy profile on the state-space-collapse manifold, identifies the initial fraction of one-job queues, bounds the initial maximum queue length by the total population, and ensures that time zero contributes no uncontrolled marked deficit excursion with large $Z_3^{N,\gamma}$. Extending the process theorem to asymptotically collapsed or stationary initial states requires a separate initial-layer analysis.*

Proposition 4 (Invariant law with abandonment). *The diffusion (13) is positive recurrent and has the unique invariant density*

$$p_{\beta,\gamma}(x) = C_{\beta,\gamma} \begin{cases} (x-1) \exp\left\{-(1+\beta)(x-2) - \frac{\gamma}{2}((x-1)^2 - 1)\right\}, & 1 < x \leq 2, \\ \exp\left\{-\beta(x-2) - \frac{\gamma}{2}((x-1)^2 - 1)\right\}, & x \geq 2, \end{cases} \quad (14)$$

where $C_{\beta,\gamma}$ is the normalizing constant. From every initial law supported on $(1, \infty)$, the law of $X^\gamma(t)$ converges to this invariant law as $t \rightarrow \infty$.

Theorem 6 (Steady-state NDS limit with abandonment). *For every fixed $\beta, \gamma > 0$,*

$$X_N^\gamma \Rightarrow X_{\beta,\gamma}, \quad (15)$$

where $X_{\beta,\gamma}$ has density (14).

Corollary 3 (Interchange and stationary moments with abandonment). *For every initial sequence satisfying the hypotheses of Theorem 5,*

$$\lim_{N \rightarrow \infty} \lim_{t \rightarrow \infty} \text{Law}\left(\frac{S^{N,\gamma}(Nt)}{N}\right) = \text{Law}(X_{\beta,\gamma}) = \lim_{t \rightarrow \infty} \lim_{N \rightarrow \infty} \text{Law}\left(\frac{S^{N,\gamma}(Nt)}{N}\right).$$

The steady-state-then-many-server limit holds from arbitrary finite-system initial laws. For every $k \geq 1$,

$$\mathbb{E}_{\pi_N^\gamma} (X_N^\gamma)^k \longrightarrow \mathbb{E} X_{\beta,\gamma}^k.$$

If $0 < \eta < a$ for an a satisfying (5), then

$$\mathbb{E}_{\pi_N^\gamma} e^{\eta X_N^\gamma} \longrightarrow \mathbb{E} e^{\eta X_{\beta,\gamma}}.$$

The invariant density is the baseline density multiplied by an explicit abandonment tilt:

$$p_{\beta,\gamma}(x) = \frac{C_{\beta,\gamma}}{C_\beta} p_\beta(x) \exp\left\{-\frac{\gamma}{2}((x-1)^2 - 1)\right\}.$$

Because this tilt is finite and positive at one, the density still vanishes linearly at the lower boundary: the singular upward drift generated by idleness dominates, while the linear abandonment drift vanishes. In the upper tail,

$$\log p_{\beta,\gamma}(x) = -\frac{\gamma}{2}x^2 + O(x).$$

Thus the limiting invariant density changes from exponential upper-tail decay to Gaussian-type decay. This statement concerns the limit density; the finite- N stationary systems are controlled

by the inherited uniform exponential-moment bound. Both the baseline diffusion and its invariant density are recovered as $\gamma \downarrow 0$. The extension also shows that the proof architecture is robust to a perturbation of the slow drift once the modified rate bounds are verified.

The interchange statements in this section concern iterated weak limits of the laws of normalized total population. They do not assert convergence of the stationary occupancy profile, a joint limit in N and time, stationary process-level convergence, or a uniform mixing estimate. Arbitrary finite-system initial laws are allowed only in the fixed- N , long-time-first order; the many-server-first order uses the balanced initialization in Theorems 1 and 5.

2.5 Proof architecture

The stationary proofs for both models follow one chain of implications. Marked excursions control integrated idleness when many queues contain at least three jobs. Combined with the perpendicular-component estimate, these bounds yield a uniform exponential-moment bound for the stationary normalized population. The finite- N Poisson equation then replaces idleness in the stationary generator by its state-dependent equilibrium mean. The resulting BAR identifies the law on the interior, pooled comparison excludes boundary mass, and zero probability flux determines the invariant density.

The diffusion-limit proof with abandonment uses a separate localized branch. The monotone coupling supplies the upper-tail bound needed for compact containment; modified deficit excursions give the required occupancy-profile and idleness estimates; and the same Poisson equation yields stopped averaging. The martingale functional central limit theorem identifies the marginal noise law, and a joint martingale argument shows that the limiting noise is Brownian in the filtration containing the initial state and stopped population. Boundary estimates then remove localization.

The rest of the paper is dedicated to the proofs of the main results.

3 State-space collapse and excursion estimates

The first part of the stationary argument controls two features hidden by the scalar population: imbalance across queues and aggregate shortfall below level two. The former is handled through a uniform exponential-moment bound for the perpendicular component and the latter through the deficit process and its marked excursions. The resulting occupation estimates are the common input to tightness and averaging; stationary generator calculations use Lemma 2 and the flow balance (2).

3.1 State-space collapse and the deficit process

Uniform exponential-moment bound for the perpendicular component. For a deterministic $q \in \mathbb{Z}_+^N$, define

$$\bar{q} = \frac{1}{N} \sum_{i=1}^N q_i, \quad q^\perp = q - \bar{q} \mathbf{1}_N.$$

For the stationary random state, set

$$Q^{N,\perp} = (Q^N)^\perp, \quad Q_i^{N,\perp} = Q_i^N - X_N.$$

The following result is Proposition 3 of Jhunjhunwala et al. (2024). Their homogeneous FCFS JSQ model matches ours under $n = N$ and service rate $\mu = 1$, and their load-slack parameter becomes $\varepsilon_n = \beta/N$.

Theorem 7 (Uniform exponential-moment bound for the perpendicular component). *There exist universal constants $\kappa_\perp = 128$ and $\theta_\perp = 1/96$ with the following property: if $\beta/N \leq 1/2$ and $0 < \theta < \theta_\perp$, then*

$$\mathbb{E}_{\pi_N} \left[\sum_{i=1}^N e^{\theta Q_i^{N,\perp}} + \sum_{i=1}^N e^{-\theta Q_i^{N,\perp}} \right] \leq \kappa_\perp N. \quad (16)$$

Let

$$Q_{\max}^N = \max_{1 \leq i \leq N} Q_i^N, \quad D_N = \max_{1 \leq i \leq N} |Q_i^N - X_N| = \max_{1 \leq i \leq N} |Q_i^{N,\perp}|.$$

By Markov's inequality,

$$\mathbb{P}_{\pi_N}(D_N > u) \leq \kappa_\perp N e^{-\theta u}, \quad u \geq 0. \quad (17)$$

A second immediate consequence is an exponential-moment estimate weighted by idleness. On $\{I^N > 0\}$, every idle coordinate has $Q_i^{N,\perp} = -X_N$, so

$$I^N e^{\theta X_N} \leq \sum_{i=1}^N e^{-\theta Q_i^{N,\perp}}.$$

Consequently,

$$\mathbb{E}_{\pi_N} [I^N e^{\theta X_N}] \leq \kappa_\perp N. \quad (18)$$

The factor N prevents (18) from yielding a uniform exponential-moment bound on its own.

The deficit process. Recall that $Z_3^N = \#\{i : Q_i^N \geq 3\}$ is the number of queues with at least three jobs. Following the aggregate coordinate used by Gupta and Walton (2019) in their finite-horizon averaging argument, define the aggregate shortfall below level two, for a deterministic state q , by

$$V(q) = 2I(q) + M_1^N(q),$$

and along the process set $V^N(t) = V(Q^N(t))$. We refer to V^N as the *deficit process*, and use $V^N = V(Q^N)$ for its stationary version. Equivalently, $V(q) = \sum_{i=1}^N (2 - q_i)^+$: an idle queue contributes two units of shortfall, a one-job queue contributes one, and a queue with at least two jobs contributes none. Thus V^N measures the aggregate shortfall below level two without crediting excess jobs above level two. Its exact dynamics are particularly simple.

Lemma 3 (Exact deficit dynamics). *At every state,*

$$V^N \longrightarrow V^N + 1 \quad \text{at rate} \quad M_1^N + M_2^N = N - I^N - Z_3^N, \quad (19)$$

while

$$V^N \longrightarrow V^N - 1 \quad \text{at rate} \quad N - \beta, \quad V^N > 0. \quad (20)$$

Moreover, Z_3^N can increase only when $V^N = 0$.

Proof. A service completion from a one-job queue changes (I, M_1) to $(I + 1, M_1 - 1)$ and therefore increases V by one. A completion from a two-job queue creates a one-job queue and also increases V by one. No other completion changes V , which proves (19).

If $I > 0$, an arrival changes (I, M_1) to $(I - 1, M_1 + 1)$, decreasing V by one. If $I = 0$ and $M_1 > 0$, it joins a one-job queue and again decreases V by one. These are exactly the states $V > 0$, proving (20). Finally, Z_3 increases only when an arrival joins a two-job queue. Under JSQ this requires every queue to have at least two jobs, equivalently $I = M_1 = 0$ and $V = 0$. $\square$

The lemma exposes the useful asymmetry. On $\{Z_3^N \geq K\}$ with $K > \beta$, the upward rate of V^N is at most $N - K$, while its downward rate is $N - \beta$; hence the deficit is uniformly subcritical. Because Z_3^N cannot increase while $V^N > 0$, this comparison persists until the excursion returns to zero or Z_3^N falls below K . The next subsection turns this excursion-level control into stationary occupation bounds.

3.2 Marked-excursion decomposition

Fix integers K, L satisfying

$$\beta < K < N, \quad 1 \leq L < N - \beta. \quad (21)$$

A deficit excursion is *marked* when its initial jump from zero occurs with $Z_3^N \geq K$. The decomposition below splits its integrated idleness at the first hit of level L : idleness before that hit is indexed by the individual idleness excursions, while the remaining idleness is indexed by the containing deficit excursion. The half-open endpoint convention assigns the hitting time to the post-hit component and excludes the terminal return time. Figure 1 illustrates the allocation before the formal statement.

For $0 \leq b < d$ and $i \geq 0$, define

$$F_{b,d}(i) = \frac{i^2}{2(d-b)} + \frac{(b+d)i}{2(d-b)^2}.$$

We use a two-sided stationary version of the chain and write $\mathbb{E}_V^0$ and $\mathbb{E}_I^0$ for Palm expectation with respect to the marked deficit- and idleness-excursion point processes.

Proposition 5 (Palm excursion decomposition). *Suppress the superscript N . Mark the starts*

$$\mathcal{E}_V = \{\sigma : V(\sigma-) = 0, V(\sigma) = 1, Z_3(\sigma) \geq K\}, \quad \Xi_V = \sum_{\sigma \in \mathcal{E}_V} \delta_\sigma.$$

For $\sigma \in \mathcal{E}_V$, let

$$\tau(\sigma) = \inf\{t > \sigma : V(t) = 0 \text{ or } Z_3(t) < K\}, \quad \chi(\sigma) = \inf\{t \in [\sigma, \tau(\sigma)) : V(t) \geq L\},$$

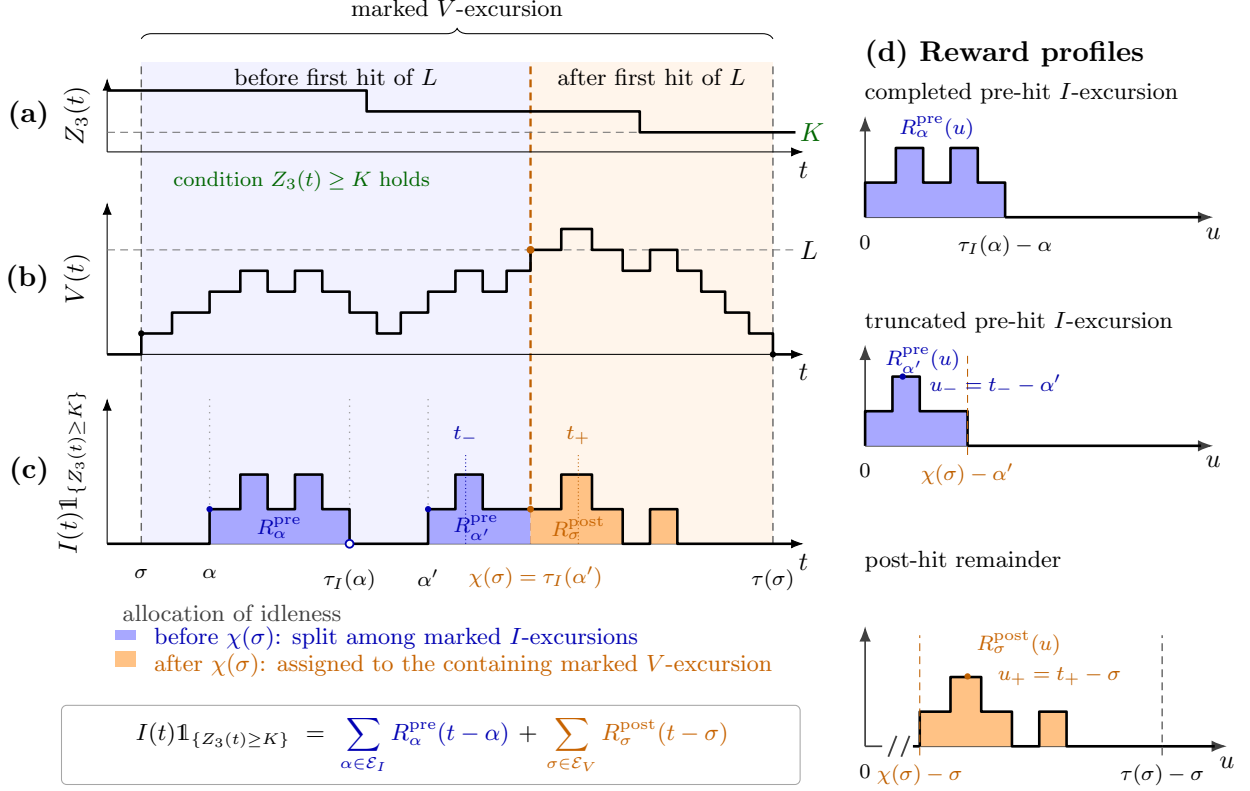


Figure 1: Panel (a) shows that $Z_3(t)$ remains at least K during the marked excursion. Panel (b) shows the associated deficit path, which first reaches level L at $\chi(\sigma)$. Panel (c) allocates the idleness restricted to $\{Z_3(t) \geq K\}$ between its pre-hit and post-hit components, and panel (d) re-indexes the corresponding reward profiles by elapsed time since the marked event. Shifting and superposing these profiles recovers $I(t)\mathbb{1}_{\{Z_3(t) \geq K\}}$.

with $\chi(\sigma) = \infty$ if the second set is empty. Let $\mathcal{E}_I$ consist of jumps α with $I(\alpha-) = 0$ and $I(\alpha) = 1$ occurring in $[\sigma, \chi(\sigma) \wedge \tau(\sigma))$ for a marked excursion σ , and write $\Xi_I = \sum_{\alpha \in \mathcal{E}_I} \delta_\alpha$. If $\sigma(\alpha)$ is the containing marked excursion, set

$$\tau_I(\alpha) = \inf\{t > \alpha : I(t) = 0\} \wedge \chi(\sigma(\alpha)) \wedge \tau(\sigma(\alpha)),$$

and define

$$R_\alpha^{\text{pre}}(u) = I(\alpha + u)\mathbb{1}_{\{0 \leq u < \tau_I(\alpha) - \alpha\}}, \quad R_\sigma^{\text{post}}(u) = I(\sigma + u)\mathbb{1}_{\{\chi(\sigma) < \tau(\sigma), \chi(\sigma) - \sigma \leq u < \tau(\sigma) - \sigma\}}.$$

Then the following assertions hold.

- (a) Ξ_V and Ξ_I are stationary, simple, locally finite point processes with measurable shift-covariant marks. If

$$\Gamma_{K,L}(t) = \sum_{\sigma \in \mathcal{E}_V} \mathbb{1}_{\{\sigma < t \leq \chi(\sigma) \wedge \tau(\sigma)\}},$$

then, whenever $V(t-) > 0$ and σ_t is the most recent jump $V : 0 \rightarrow 1$ before t ,

$$\Gamma_{K,L}(t) = \mathbb{1}_{\{Z_3(t-) \geq K, \sup_{\sigma_t \leq s < t} V(s) < L\}}, \quad (22)$$

with $\Gamma_{K,L}(t) = 0$ when $V(t-) = 0$. In particular $\Gamma_{K,L}$ is predictable. Define

$$\Theta_V(t) = \mathbb{1}_{\{V(t-)=0, Z_3(t-) \geq K\}}, \quad \Theta_I(t) = \Gamma_{K,L}(t) \mathbb{1}_{\{I(t-)=0, V(t-) \leq L-2\}},$$

Then Θ_V and Θ_I are predictable, and the exact compensator measures are

$$\Lambda_V(dt) = \Theta_V(t) M_2(t-) dt, \quad \Lambda_I(dt) = \Theta_I(t) M_1(t-) dt. \quad (23)$$

Consequently, the corresponding intensities satisfy

$$\iota_V \leq N - K, \quad \iota_I \leq L. \quad (24)$$

(b) Away from the locally finite jump set,

$$I(t) \mathbb{1}_{\{Z_3(t) \geq K\}} = \sum_{\alpha \in \mathcal{E}_I} R_\alpha^{\text{pre}}(t - \alpha) + \sum_{\sigma \in \mathcal{E}_V} R_\sigma^{\text{post}}(t - \sigma). \quad (25)$$

Campbell's formula therefore gives

$$\mathbb{E}_{\pi_N} [I^N \mathbb{1}_{\{Z_3^N \geq K\}}] = \iota_I \mathbb{E}_I^0 \int_0^\infty R_0^{\text{pre}}(u) du + \iota_V \mathbb{E}_V^0 \int_0^\infty R_0^{\text{post}}(u) du. \quad (26)$$

(c) Conditional on the post-jump state at a Palm point, the future segment has the original deterministic-state Markov law. Uniformly over the possible post-jump states,

$$\mathbb{E}_I^0 \int_0^\infty R_0^{\text{pre}}(u) du \leq F_{L,N-\beta}(1) = \frac{N - \beta}{(N - \beta - L)^2}, \quad (27)$$

$$\mathbb{P}_q(\chi < \tau) \leq \left(\frac{N - K}{N - \beta} \right)^{L-1}, \quad (28)$$

$$\sup_{\substack{q': V(q')=L \\ Z_3(q') \geq K}} \mathbb{E}_{q'} \int_0^{\tau'} I(t) dt \leq \frac{1}{2} F_{N-K,N-\beta}(L), \quad (29)$$

where q ranges over marked post-jump states and τ' is the first time, from q' , that $V = 0$ or $Z_3 < K$.

The proof, including the two-sided construction, predictable thinning, endpoint verification, and the Palm disintegration argument, is given in Section B.1.

3.3 Occupation bounds for level-three occupancy and the deficit process

We now apply Proposition 5 to bound the stationary idleness accumulated while $Z_3^N \geq K$, leading to the general occupation estimate in Proposition 6. The proposition splits this idleness into a pre-hit part, accumulated before the deficit process reaches level L , and a post-hit part, accumulated

after that threshold is reached. Campbell's formula turns this pathwise decomposition into two intensity-times-reward terms, which we bound separately below.

For the pre-hit term, marked I -excursions start at intensity at most L and have Palm mean total reward bounded by $F_{L,N-\beta}(1)$. For the post-hit term, a marked V -excursion reaches level L with probability at most

$$\left(\frac{N-K}{N-\beta}\right)^{L-1},$$

after which $I \leq V/2$ controls the remaining integrated idleness.

Proposition 6 (Idleness on states with $Z_3^N \geq K$). *For K, L satisfying (21),*

$$\mathbb{E}_{\pi_N}[I^N \mathbf{1}_{\{Z_3^N \geq K\}}] \leq \frac{L(N-\beta)}{(N-\beta-L)^2} + \frac{N-K}{2} \left(\frac{N-K}{N-\beta}\right)^{L-1} \left[\frac{L^2}{2(K-\beta)} + \frac{(2N-K-\beta)L}{2(K-\beta)^2} \right].$$

Proof. By (23)–(24) and (26), the pre-hit contribution is at most $LF_{L,N-\beta}(1)$. For the post-hit contribution, combine the marked deficit-excursion intensity, the threshold-hitting probability (28), and the reward estimate (29). Substitution of $F_{b,d}$ gives the displayed bound. $\square$

Set

$$K_N = \lceil N^{3/4} \rceil, \quad L_N = \lceil N^{1/2} \rceil.$$

For these choices, the pre-hit term in Proposition 6 is $O(N^{-1/2})$. Moreover, $K_N L_N / N \asymp N^{1/4}$, so the threshold-hitting probability is at most $e^{-cN^{1/4}}$ for some $c > 0$. Including the polynomial intensity and reward factors, the post-hit term is $O(N^{5/4}e^{-cN^{1/4}})$, and therefore decays faster than every inverse power of N . More generally, the bound in Proposition 6 vanishes for any sequences satisfying the range conditions in (21), together with $K_N \rightarrow \infty$, $\frac{L_N}{N} \rightarrow 0$, and $\frac{K_N L_N}{N \log N} \rightarrow \infty$.

Corollary 4 (Idleness bound above the level-three threshold). *There are constants $c, C, N_0 > 0$, depending only on β , such that*

$$\mathbb{E}_{\pi_N}[I^N \mathbf{1}_{\{Z_3^N \geq K_N\}}] \leq CN^{-1/2} + CN^{5/4}e^{-cN^{1/4}}, \quad N \geq N_0. \quad (30)$$

Proof. Apply Proposition 6. With $K = K_N$ and $L = L_N$, its pre-hit term is $O(N^{-1/2})$, while

$$\left(\frac{N-K_N}{N-\beta}\right)^{L_N-1} \leq \exp\left\{-\frac{(L_N-1)(K_N-\beta)}{N-\beta}\right\} \leq e^{-cN^{1/4}}.$$

The post-hit integrated-idleness factor is $O(N^{1/4})$ and the marked-excursion start intensity is $O(N)$, giving the second term in (30). $\square$

The averaging branch also requires a tail bound for the deficit while Z_3^N remains large.

Proposition 7 (Occupation above a deficit threshold). *For $\beta < K < N$ and every integer $v \geq 1$,*

$$\mathbb{P}_{\pi_N}(V^N \geq v, Z_3^N \geq K) \leq \frac{Nv}{K-\beta} \left(\frac{N-K}{N-\beta}\right)^{v-1}. \quad (31)$$

The proof is deferred to Section B.2. Its two estimates are a birth–death hitting bound for reaching v within a marked excursion and a localized Dynkin bound $v/(K-\beta)$ for the remaining duration; Campbell's formula, together with the coarser bound $\iota_V \leq N$ in place of $\iota_V \leq N-K$, then gives (31).

4 Uniform exponential-moment bound and tightness

We now convert the occupation estimates into a uniform exponential-moment bound for the stationary normalized population. Because $(\mathcal{L}_N S)(q) = I(q) - \beta$, only high-population states that retain idleness can have positive population drift. The argument first shows that their expected idleness vanishes, then strengthens this to an exponential-moment bound weighted by idleness, and finally uses an exact level-crossing identity to control the exponential transform of the normalized population.

4.1 Vanishing idleness in high-population states

For $x \geq 0$, define the expected idleness above population level x by

$$J_N(x) = \mathbb{E}_{\pi_N} [I^N \mathbf{1}_{\{X_N \geq x\}}].$$

The estimate splits according to the number of queues with at least three jobs. Large Z_3^N is controlled by the excursion bound, whereas small Z_3^N together with $X_N > 2$ forces the excess population onto a few long queues and therefore creates a large perpendicular component.

Proposition 8 (Vanishing idleness in high-population states). *For every fixed $x_0 > 2$,*

$$J_N(x_0) \longrightarrow 0.$$

More precisely, for any $0 < \theta < \theta_\perp$, there are constants $c, C, N_0 > 0$ such that

$$J_N(x_0) \leq CN^{-1/2} + CN^{5/4} e^{-cN^{1/4}} + \kappa_\perp N^2 \exp \left\{ -\theta(x_0 - 2) \left(\frac{N}{K_N} - 1 \right) \right\}, \quad N \geq N_0, \quad (32)$$

where $K_N = \lceil N^{3/4} \rceil$.

Proof. The contribution from $\{Z_3^N \geq K_N\}$ is bounded by Corollary 4. On $\{X_N \geq x_0, Z_3^N = 0\}$ one would have $S^N \leq 2N$, so this event is empty because $x_0 > 2$. It remains to consider $Z_3^N = j$ with $1 \leq j < K_N$. The $N - j$ queues below level three contain at most two jobs each, so $NX_N \leq 2(N - j) + jQ_{\max}^N$. On $\{X_N \geq x_0, Z_3^N = j\}$ this implies

$$Q_{\max}^N - X_N \geq (x_0 - 2) \left(\frac{N}{j} - 1 \right) \geq (x_0 - 2) \left(\frac{N}{K_N} - 1 \right).$$

By (17) and $I^N \leq N$,

$$\mathbb{E}[I^N \mathbf{1}_{\{X_N \geq x_0, Z_3^N < K_N\}}] \leq \kappa_\perp N^2 \exp \left\{ -\theta(x_0 - 2) \left(\frac{N}{K_N} - 1 \right) \right\}.$$

Combining the two regions proves (32). Since $N/K_N - 1 \asymp N^{1/4}$, both stretched-exponential terms there are $o(N^{-1/2})$; in particular, $J_N(x_0) \leq C_0 N^{-1/2}$ for all sufficiently large N . $\square$

4.2 Exponential-moment control weighted by idleness

The fixed-level estimate is not yet strong enough for the transform argument. We interpolate it with the global bound (18) to obtain a uniform exponential-moment bound under idleness weighting.

Proposition 9 (Uniform exponential-moment bound weighted by idleness). *Fix $0 < \theta < \theta_\perp$ and*

$$0 < a < \min \left\{ \beta, \frac{\theta}{3} \right\}.$$

Then there are $N_0 < \infty$ and $C < \infty$ such that

$$\sup_{N \geq N_0} \mathbb{E}_{\pi_N} [I^N e^{aX_N}] \leq C.$$

The proof in Section C.2 uses a layer-cake decomposition, splitting at $c_* \log N$. The interval $1/(\theta - a) < c_* < 1/(2a)$ is nonempty precisely when $3a < \theta$.

4.3 Level crossings and the exponential transform

Let

$$p_m^N = \mathbb{P}_{\pi_N}(S^N = m), \quad j_m^N = \mathbb{E}_{\pi_N} [I^N \mathbf{1}_{\{S^N = m\}}].$$

For $m \geq 0$, apply Lemma 2 to the bounded state function

$$F_m(q) = \mathbf{1}_{\{S(q) \leq m\}}.$$

Only two kinds of transitions cross the cut $\{S \leq m\}$. An arrival leaves the cut when $S = m$, while a service completion enters it when $S = m + 1$; the latter occurs at total rate $N - I$. Therefore

$$0 = \mathbb{E}_{\pi_N} (\mathcal{L}_N F_m)(Q^N) = -(N - \beta)p_m^N + \mathbb{E}_{\pi_N} [(N - I^N) \mathbf{1}_{\{S^N = m+1\}}].$$

Equivalently, the exact level-crossing identity is

$$(N - \beta)p_m^N = Np_{m+1}^N - j_{m+1}^N. \tag{33}$$

This identity yields the transform bound: after multiplication by z^{m+1} and summation over m , the exponential transform of the population distribution is bounded by the corresponding transform weighted by idleness, while the restriction $a < \beta$ keeps the resulting coefficient positive.

Proof of Theorem 2. Fix a satisfying (5). Choose θ with $3a < \theta < \theta_\perp$, then Proposition 9 applies. Set $z = e^{a/N}$ and define the finite sums

$$F_M = \sum_{m=0}^M z^m p_m^N, \quad G_M = \sum_{m=0}^M z^m j_m^N.$$

Multiply (33) by z^{m+1} and sum from $m = 0$ to M . Since $j_0^N = Np_0^N$, the boundary terms at zero cancel, and

$$(N - \beta)zF_M = NF_{M+1} - G_{M+1}.$$

Therefore

$$[N - (N - \beta)z]F_M = G_{M+1} - Nz^{M+1}p_{M+1}^N \leq G_{M+1}. \quad (34)$$

Now $N - (N - \beta)e^{a/N} \rightarrow \beta - a > 0$. For all sufficiently large N , the coefficient in (34) is bounded below by $(\beta - a)/2$, while $G_{M+1} \leq \mathbb{E}[I^N e^{aX_N}]$ which is uniformly bounded by Proposition 9. Letting $M \rightarrow \infty$ and applying monotone convergence yields (6). Markov's inequality gives (7). $\square$

5 Stationary averaging of the idleness contribution

The second part of the stationary argument removes the idleness term from the population generator. We first show that the true birth rate M_1^N of the idleness process is close to $Nr(X_N)$ in the $(I^N + 1)$ -weighted sense needed below. We then solve the Poisson equation for the idleness chain with X_N frozen and use that solution to remove $I^N - g(X_N)$ from the stationary generator. We call this finite- N solution the corrector.

5.1 Stationary tails of idleness

Let $p_j^{I,N} = \mathbb{P}_{\pi_N}(I^N = j)$. The idleness process jumps upward through a completion at a one-job queue and downward through an arrival whenever it is positive. For $0 \leq j < N$, apply Lemma 2 to

$$G_j(q) = \mathbb{1}_{\{I(q) \leq j\}}.$$

A completion at a one-job queue leaves this cut exactly when $I = j$, at rate M_1^N , and an arrival enters it exactly when $I = j + 1$, at rate $N - \beta$. No other transition crosses the cut. Hence

$$0 = \mathbb{E}_{\pi_N}(\mathcal{L}_N G_j)(Q^N) = -\mathbb{E}_{\pi_N}[M_1^N \mathbb{1}_{\{I^N=j\}}] + (N - \beta)p_{j+1}^{I,N},$$

and therefore

$$(N - \beta)p_{j+1}^{I,N} = \mathbb{E}_{\pi_N}[M_1^N \mathbb{1}_{\{I^N=j\}}].$$

Since $M_1^N \leq N - I^N$, this identity yields the recursive bound

$$p_{j+1}^{I,N} \leq \frac{N - j}{N - \beta} p_j^{I,N}. \quad (35)$$

Lemma 4 (Gaussian-type tail bound for idleness and an $o(N)$ second-moment bound). *There are constants $c, C, N_0 > 0$, depending only on β , such that*

$$\mathbb{P}_{\pi_N}(I^N \geq L) \leq CNe^{-cL^2/N}, \quad 0 \leq L \leq N, \quad N \geq N_0. \quad (36)$$

Consequently,

$$\frac{1}{N} \mathbb{E}_{\pi_N}(I^N)^2 \rightarrow 0.$$

The proof is deferred to Section C.3. It follows by iterating (35) above a fixed level and then splitting the second moment at $N^{3/4}$. The prefactor N in (36) does not affect the conclusion because the estimate is used at $L = N^{3/4}$, where $Ne^{-cL^2/N} = Ne^{-cN^{1/2}}$.

5.2 Idleness-weighted approximation of the one-job queue count

The corrector comparison multiplies the mismatch in the birth rate of the idleness process by $I^N + 1$, so unweighted convergence of M_1^N/N is not sufficient. The required approximation is

$$\frac{M_1^N}{N} \approx r(X_N) = (2 - X_N)^+,$$

in an $(I^N + 1)$ -weighted sense. The following identity isolates the two geometric error terms.

Define the aggregate number of jobs beyond level two by

$$H^N = \sum_{i=1}^N (Q_i^N - 2)^+.$$

For each queue, we have $2 - Q_i^N = 2\mathbb{1}_{\{Q_i^N=0\}} + \mathbb{1}_{\{Q_i^N=1\}} - (Q_i^N - 2)^+$. Summing gives $2N - S^N = V^N - H^N$. Therefore $Nr(X_N) = (V^N - H^N)^+$. Because $M_1^N = V^N - 2I^N$,

$$|M_1^N - Nr(X_N)| \leq H^N + 2I^N. \quad (37)$$

Thus the discrepancy from the one-job queue count associated with the balanced profile is controlled by two explicit terms: the aggregate number of jobs beyond level two and twice the number of idle servers.

Proposition 10 (Idleness-weighted approximation of the one-job queue count). *As $N \rightarrow \infty$,*

$$\mathbb{E}_{\pi_N} \left[(I^N + 1) \left| \frac{M_1^N}{N} - r(X_N) \right| \right] \rightarrow 0. \quad (38)$$

The proof is deferred to Section C.3. The argument splits according to whether $Z_3^N < \lceil N^{1/2} \rceil$. In the first region, (37) and $H^N \leq 2\lceil N^{1/2} \rceil D_N$ reduce the error to a large perpendicular component or atypically large idleness. In the complementary region, the error is at most V^N/N and Proposition 7 applies. The resulting stretched-exponential probability estimate remains negligible after multiplication by $I^N + 1$.

5.3 The frozen-chain Poisson equation

With the weighted birth-rate approximation in place, we solve the Poisson equation for the idleness chain with the slow coordinate frozen. The frozen family is used only for $x > 1$; the zero extension below is a bookkeeping convention and does not define a positive recurrent chain for $x \leq 1$.

Fix $x > 1$. Consider the birth-death generator

$$\mathcal{A}_x^N \varphi(i) = Nr(x) [\varphi(i+1) - \varphi(i)] + N\mathbb{1}_{\{i>0\}} [\varphi(i-1) - \varphi(i)]. \quad (39)$$

Its birth-to-death ratio is $r(x) < 1$, and its stationary mean is $g(x) = r(x)/(1 - r(x))$.

Let $h \in C_c^1((1, \infty))$ and use its zero extension fixed above. Define the global function

$$a_h(x) = \begin{cases} 0, & x \leq 1, \\ \frac{h(x)}{x-1}, & 1 < x < 2, \\ h(x), & x \geq 2. \end{cases}$$

The function h vanishes on a neighborhood of one, so the denominator in the middle expression is bounded away from zero wherever that expression is nonzero. Each piece has bounded derivative, and the two nonzero pieces agree at 2 because $a_h(2-) = h(2) = a_h(2+)$. Hence a_h is bounded and globally Lipschitz on $\mathbb{R}$, although its derivative need not be continuous at 2. Define

$$\Psi_N(x, i) = -\frac{a_h(x)}{2N}i(i+1).$$

Its same- x increments are

$$\Psi_N(x, i+1) - \Psi_N(x, i) = -\frac{a_h(x)}{N}(i+1), \quad (40)$$

$$\Psi_N(x, i-1) - \Psi_N(x, i) = \frac{a_h(x)}{N}i. \quad (41)$$

For $x > 1$, substitution in (39) yields the exact Poisson equation

$$\mathcal{A}_x^N \Psi_N(x, i) = h(x)(i - g(x)). \quad (42)$$

For $x \leq 1$, define both sides of (42) to be zero according to our convention.

Remark 2 (Why the corrector is quadratic). *For a constant-rate birth–death chain, the generator maps the quadratic $i(i+1)$ to an affine function of i . The factor $a_h(x)/(2N)$ normalizes that affine function so that its centered form is exactly $h(x)(i - g(x))$. Solving this equation at finite N cancels the centered idleness term before passage to the stationary limit; the remaining errors come only from the actual birth-rate mismatch and the $1/N$ motion of the scaled population.*

5.4 Generator comparison and averaging

Lift the two-coordinate corrector to the queue-state space by defining

$$x_N(q) = \frac{S(q)}{N}, \quad \psi_N(q) = \Psi_N(x_N(q), I(q)).$$

At a deterministic state q , write

$$x = x_N(q), \quad i = I(q), \quad m = M_1^N(q).$$

The pair of coordinate values (x, i) has the following relevant transitions:

$$(x, i) \longrightarrow \left(x + \frac{1}{N}, i - 1\right) \quad \text{at rate } N - \beta, \quad i > 0, \quad (43)$$

$$(x, i) \longrightarrow \left(x - \frac{1}{N}, i + 1\right) \quad \text{at rate } m,$$

$$(x, i) \longrightarrow \left(x - \frac{1}{N}, i\right) \quad \text{at rate } N - i - m. \quad (44)$$

If $i = 0$, an arrival changes x but not i ; it contributes zero to Ψ_N because $\Psi_N(x, 0) = 0$.

Relative to the frozen generator, the actual chain has birth rate m rather than $Nr(x)$, death rate $N - \beta$ rather than N , and slow-coordinate shifts of size $1/N$. The corrector estimate separates the resulting rate-mismatch term from the lower-order error caused by motion of x .

Lemma 5 (Corrector error). *There is a constant $C_{h,\beta} < \infty$, depending only on β and $\|a_h\|_{\text{BL}} := \|a_h\|_\infty + \text{Lip}(a_h)$, where $\text{Lip}(a_h) := \sup_{x \neq y} |a_h(x) - a_h(y)|/|x - y|$, and in particular independent of N , x , i , and q , such that every $q \in \mathbb{Z}_+^N$ satisfies*

$$|(\mathcal{L}_N \psi_N)(q) - h(x)(i - g(x))| \leq C_{h,\beta} \left[(i+1) \left| \frac{m}{N} - r(x) \right| + \frac{(i+1)^2}{N} \right], \quad (45)$$

where $x = x_N(q)$, $i = I(q)$, and $m = M_1^N(q)$.

The proof is deferred to Section C.3. The same- x increments (40)–(41) yield the birth-rate and death-rate mismatch term, while the global Lipschitz bound on a_h controls all $1/N$ -shifts of x by $C_{h,\beta}(i+1)^2/N$.

Proof of Theorem 3. For each fixed N , $0 \leq I(q) \leq N$ for every q and a_h is bounded, so

$$\|\psi_N\|_\infty \leq \frac{N+1}{2} \|a_h\|_\infty < \infty.$$

The total transition rate is at most $2N$, and therefore

$$\|\mathcal{L}_N \psi_N\|_\infty \leq 4N \|\psi_N\|_\infty < \infty.$$

Thus the hypotheses of Lemma 2 hold and

$$\mathbb{E}_{\pi_N} [(\mathcal{L}_N \psi_N)(Q^N)] = 0.$$

Evaluate (45) at the stationary state Q^N and take expectations. By Proposition 10, the expectation of the first term on its right-hand side tends to zero. By Lemma 4 and $\mathbb{E}I^N = \beta$,

$$\frac{1}{N} \mathbb{E}(I^N + 1)^2 \rightarrow 0.$$

The left-hand side therefore yields (8). □

6 Identification of subsequential stationary limits

It remains to identify the tight subsequential limits. Stationary averaging yields the BAR for test functions supported in $(1, \infty)$. We then locate the support, exclude an atom at the lower boundary, and use zero probability flux to identify the unique solution.

6.1 The BAR for subsequential limits

Proposition 11 (BAR for subsequential limits). *The family $\{\nu_N\}$ is tight. Every weak subsequential limit ν satisfies*

$$\int_{(1,\infty)} [f''(x) + (g(x) - \beta)f'(x)] \nu(dx) = 0, \quad f \in C_c^\infty((1, \infty)). \quad (46)$$

Proof. Tightness follows from Theorem 2. Let $f \in C_c^3((1, \infty))$ and extend f by zero to $\mathbb{R}$; the extension is C^3 because the support is a positive distance from one. Define the deterministic state function $\varphi_N(q) = f(S(q)/N)$. Applying Lemma 2 to φ_N gives

$$0 = \mathbb{E} \left[(N - \beta) \left(f \left(X_N + \frac{1}{N} \right) - f(X_N) \right) \right] + \mathbb{E} \left[(N - I^N) \left(f \left(X_N - \frac{1}{N} \right) - f(X_N) \right) \right].$$

Taylor's theorem gives

$$f \left(x \pm \frac{1}{N} \right) - f(x) = \pm \frac{1}{N} f'(x) + \frac{1}{2N^2} f''(x) + R_{N,\pm}(x), \quad |R_{N,\pm}(x)| \leq \frac{\|f'''\|_\infty}{6N^3}.$$

Multiplying the generator identity by N therefore gives

$$0 = \mathbb{E} \left[(I^N - \beta) f'(X_N) + \left(1 - \frac{\beta + I^N}{2N} \right) f''(X_N) \right] + O_f(N^{-1}).$$

Indeed, since $I^N \leq N$, the absolute contribution of the two third-order remainders is at most

$$\frac{(N - \beta) + (N - I^N)}{6N^2} \|f'''\|_\infty \leq \frac{\|f'''\|_\infty}{3N}.$$

Thus $O_f(N^{-1})$ has absolute value at most $C_{f,\beta}/N$, where $C_{f,\beta}$ depends only on β and the sup norms of the first three derivatives of f . Because $\mathbb{E} I^N = \beta$,

$$\mathbb{E} [(I^N - \beta) f'(X_N) + f''(X_N)] \longrightarrow 0.$$

Apply Theorem 3 with $h = f'$ to obtain

$$\mathbb{E} [f''(X_N) + (g(X_N) - \beta) f'(X_N)] \longrightarrow 0. \quad (47)$$

If $\nu_{N_k} \Rightarrow \nu$, the integrand in (47) is bounded and continuous because f' is supported away from one. Passing to the limit proves (46) first for C_c^3 test functions and hence, in particular, for C_c^∞ test functions. $\square$

6.2 Support and exclusion of boundary mass

The BAR in Proposition 11 uses test functions compactly supported in $(1, \infty)$ and therefore cannot detect mass at the boundary 1. We first locate the support of any subsequential limit. Since the number of busy servers is at most the total number of jobs,

$$(N - S^N)^+ \leq I^N.$$

Together with the flow balance (2), this gives

$$\mathbb{E}(1 - X_N)^+ \leq \frac{\beta}{N}. \quad (48)$$

If $X_{N_k} \Rightarrow X_\star$, then $(1 - x)^+$ is bounded and continuous on $[0, \infty)$, so (48) gives $\mathbb{E}(1 - X_\star)^+ = 0$. Every weak subsequential limit is therefore supported on $[1, \infty)$. It remains to exclude an atom at

one. A pooled $M/M/N$ queue is the natural lower comparison: at equal total population, complete pooling keeps at least as many servers busy as the JSQ system with one queue at each server. Its stationary NDS limit lies strictly above one, so stochastic domination rules out boundary mass for JSQ. Let $C^N(t)$ have arrival rate $N - \beta$ and N unit-rate servers, and let $C^N(\infty)$ denote its stationary population.

Proposition 12 (Pooled $M/M/N$ comparison and boundary exclusion). *The pooled system C^N and the JSQ total population S^N can be coupled so that*

$$C^N(0) \leq S^N(0) \implies C^N(t) \leq S^N(t) \quad \text{for all } t \geq 0.$$

Consequently, their stationary populations satisfy

$$C^N(\infty) \leq_{\text{st}} S^N(\infty). \quad (49)$$

Moreover,

$$\frac{C^N(\infty)}{N} \Rightarrow 1 + E_\beta, \quad (50)$$

where E_β is exponential with rate β . Therefore, every weak subsequential limit ν of ν_N has no atom at the lower boundary:

$$\nu(\{1\}) = 0. \quad (51)$$

The proof in Section C.4 couples common arrivals with a rate- N stream of potential service events. At equal populations the pooled system has at least the JSQ service rate. Its stationary excess above N is geometric with parameter β/N , while its mass below N is $O(N^{-1/2})$; weak-limit preservation of stochastic order then excludes an atom at one.

6.3 Probability flux and uniqueness

With the support and boundary mass settled, it remains to identify the law on $(1, \infty)$. We use the following one-dimensional regularity fact; its distributional proof is deferred to Section C.5.

Lemma 6 (Distributional probability flux and density). *Let $U \subset \mathbb{R}$ be an open interval, let b be a locally bounded Borel function, and let μ be a locally finite nonnegative measure on U . If*

$$\int_U [f''(x) + b(x)f'(x)] \mu(dx) = 0, \quad f \in C_c^\infty(U), \quad (52)$$

then μ has a locally absolutely continuous density p and there is $c \in \mathbb{R}$ such that

$$p'(x) - b(x)p(x) = c \quad \text{for a.e. } x \in U. \quad (53)$$

Equivalently, the probability flux $\mathcal{J}_p(x) = b(x)p(x) - p'(x)$ is constant almost everywhere on U and equals $-c$.

Proposition 13 (Uniqueness of the BAR under the boundary condition). *Let ν be a probability measure supported on $[1, \infty)$ such that $\nu(\{1\}) = 0$ and*

$$\int_{(1, \infty)} [f''(x) + (g(x) - \beta)f'(x)] \nu(dx) = 0, \quad f \in C_c^\infty((1, \infty)).$$

Then $\nu = \nu_\beta$.

Proof. Set $b(x) = g(x) - \beta$. The restriction of ν to $(1, \infty)$ has total mass one. Applying Lemma 6 on $U = (1, \infty)$ gives a locally absolutely continuous density p and a constant c such that

$$p'(x) - b(x)p(x) = c \quad \text{for a.e. } x > 1.$$

For $x > 2$, $b(x) = -\beta$, so

$$p(x) = Ae^{-\beta(x-2)} + \frac{c}{\beta}.$$

Integrability and nonnegativity on $(2, \infty)$ force $c = 0$. Thus the stationary probability flux $bp - p'$ vanishes and $p' = bp$. The integrable normalized solution of this zero-flux equation was computed in Proposition 2; hence $p = p_\beta$ and $\nu = \nu_\beta$. $\square$

6.4 Completion of the baseline stationary results

Proof of Theorem 4. By Proposition 11, $\{\nu_N\}$ is tight and every subsequential limit satisfies (46). Such a limit is supported on $[1, \infty)$ by (48) and has no atom at one by (51). Proposition 13 therefore identifies every subsequential limit as ν_β . Tightness then implies convergence of the full sequence. $\square$

Proof of Corollary 1. For each fixed N , Lemma 14 gives convergence of $Q^N(t)$ to π_N in total variation from every initial law. Hence the steady-state-then-many-server limit is Theorem 4. Under the balanced initialization of Theorem 1, the many-server limit at each fixed time is the diffusion (3); Proposition 2 then sends that diffusion to ν_β as time tends to infinity. $\square$

Proof of Corollary 2. Apply Lemma 13 to the weak convergence in Theorem 4 and the uniform exponential-moment bound in Theorem 2. $\square$

6.5 Stationary identification with abandonment

The stationary abandonment proof retains the baseline BAR and zero-flux identification, with abandonment-specific compactness and averaging estimates. The stochastic domination in Proposition 3 yields the inherited exponential-moment and maximum-queue bounds in Lemma 17, while Lemma 21 provides the required first- and second-moment control of idleness. For marked deficit excursions, Proposition 18 replaces the baseline upward rate by $N - K + \gamma$ against downward rate $N - \beta$, so subcriticality requires $K > \beta + \gamma$ and has margin $K - \beta - \gamma$. These estimates yield the idleness-weighted approximation of the one-job queue count in Proposition 19. The finite- N Poisson

equation is unchanged; on the relevant compact population sets, its additional abandonment contribution is bounded by a constant times $(I^{N,\gamma} + 1)^2/N^2$ and is negligible by Lemma 21. Consequently, Proposition 20 establishes stationary stochastic averaging, and Proposition 21 gives the BAR for every subsequential stationary limit. Finally, Proposition 22 supplies the pooled $M/M/N + M$ comparison and its stationary limit, while Proposition 23 excludes boundary mass and identifies the unique zero-flux limit.

Proof of Theorem 6. Tightness follows from Lemma 17. By Proposition 23, every weakly convergent subsequence has the law of $X_{\beta,\gamma}$ as its limit. Therefore the full sequence converges, proving (15). $\square$

7 Diffusion limit with abandonment

The goal of this section is to prove the NDS diffusion limit with abandonment in Theorem 5. The proof has two parts. We first localize the normalized population to a compact subinterval of $(1, \infty)$ and establish the finite-horizon occupancy-profile estimates needed to average the idleness process; this part culminates in Proposition 16. We then combine that averaging result with the population martingale decomposition to identify the stopped diffusion in Proposition 17. Finally, Lemma 10 removes the localization boundaries and completes the proof of Theorem 5.

Fix $T < \infty$, $0 < \eta < 1$, and $R > 2$, and define

$$\tau_N^{\eta,R} = \inf\{t \in [0, T] : \widehat{X}^{N,\gamma}(t) \notin (1 + \eta, R)\} \wedge T.$$

Unless stated otherwise, the process estimates below are stopped at $\tau_N^{\eta,R}$.

7.1 Localized occupancy-profile estimates and averaging

The deficit process provides the excursion coordinate used to control the fast idleness dynamics. Lemma 7 gives its exact rates under abandonment, while Lemma 8 bounds the number of relevant excursion starts on the accelerated horizon. Together, these results allow geometric excursion-height probabilities to be union-bounded over the entire original-time interval $[0, NT]$.

Lemma 7 (Deficit dynamics with abandonment). *Let*

$$V^{N,\gamma} = 2I^{N,\gamma} + M_1^{N,\gamma}.$$

At every state,

$$V^{N,\gamma} \longrightarrow V^{N,\gamma} + 1 \quad \text{at rate } M_1^{N,\gamma} + \left(1 + \frac{\gamma}{N}\right) M_2^{N,\gamma}, \quad (54)$$

$$V^{N,\gamma} \longrightarrow V^{N,\gamma} - 1 \quad \text{at rate } N - \beta, \quad V^{N,\gamma} > 0. \quad (55)$$

The process $Z_3^{N,\gamma}$ can increase only when $V^{N,\gamma} = 0$. On the set $\{Z_3^{N,\gamma} \geq K\}$, the upward rate in (54) is at most

$$b_{N,K}^\gamma = N - K + \gamma. \quad (56)$$

Lemma 8 (Excursion-start compensators). *Let $C_I^N(t)$ count transitions $I^{N,\gamma} : 0 \rightarrow 1$ by original time t . For an integer K , let $C_{V,K}^N(t)$ count transitions $V^{N,\gamma} : 0 \rightarrow 1$ whose post-jump state has $Z_3^{N,\gamma} \geq K$. Their predictable intensities are, respectively,*

$$\mathbb{1}_{\{I^{N,\gamma}(t-)=0\}} M_1^{N,\gamma}(t-), \quad (57)$$

$$\mathbb{1}_{\{V^{N,\gamma}(t-)=0, Z_3^{N,\gamma}(t-)\geq K\}} \left(1 + \frac{\gamma}{N}\right) M_2^{N,\gamma}(t-). \quad (58)$$

Consequently,

$$C_I^N(NT) = O_{\mathbb{P}}(N^2), \quad C_{V,K}^N(NT) = O_{\mathbb{P}}(N^2), \quad (59)$$

uniformly over K . More explicitly, the two counts are stochastically dominated by Poisson random variables with means N^2T and $N(N + \gamma)T$, respectively.

On $\{Z_3^{N,\gamma} \geq K\}$, the deficit process has upward rate at most $N - K + \gamma$ against downward rate $N - \beta$, and hence subcriticality margin $K - \beta - \gamma$. Since only $O_{\mathbb{P}}(N^2)$ relevant excursions can begin before original time NT , the detailed excursion estimates control the probability that any such excursion reaches a large threshold. The first consequence is the following localized approximation of the fraction of queues containing exactly one job.

Proposition 14 (Stopped state-space-collapse estimate for one-job queues). *For every fixed T, R, η as above,*

$$\sup_{t \leq \tau_N^{\eta,R}} \left| \frac{M_1^{N,\gamma}(Nt)}{N} - r(\widehat{X}^{N,\gamma}(t)) \right| \xrightarrow{\mathbb{P}} 0. \quad (60)$$

The unweighted approximation in Proposition 14 is not sufficient for the Poisson-equation argument, because the corrector increments grow with the number of idle servers. The next result provides a logarithmic bound on idleness and the corresponding idleness-weighted approximation.

Proposition 15 (Logarithmic idleness bound under localization). *For every fixed T, R, η , there is a constant $c_\eta < \infty$ such that*

$$\mathbb{P} \left(\sup_{t \leq \tau_N^{\eta,R}} I^{N,\gamma}(Nt) > c_\eta \log N \right) \longrightarrow 0. \quad (61)$$

In particular, the stopped supremum is $O_{\mathbb{P}}(\log N)$. Moreover,

$$\sup_{t \leq \tau_N^{\eta,R}} (I^{N,\gamma}(Nt) + 1) \left| \frac{M_1^{N,\gamma}(Nt)}{N} - r(\widehat{X}^{N,\gamma}(t)) \right| \xrightarrow{\mathbb{P}} 0. \quad (62)$$

The weighted error in (62) is the leading term in the comparison between the true queueing generator and the frozen idleness generator. The finite- N Poisson equation therefore gives the process-level averaging estimate needed to identify the population drift.

Proposition 16 (Stopped averaging via the Poisson equation). *For every fixed T, R, η ,*

$$\sup_{t \leq T} \left| \int_0^{t \wedge \tau_N^{\eta,R}} [I^{N,\gamma}(Ns) - g(\widehat{X}^{N,\gamma}(s))] ds \right| \xrightarrow{\mathbb{P}} 0. \quad (63)$$

The proofs of Lemmas 7 and 8 and Propositions 14 to 16 are given in Section D.2. In the idleness argument, the process is stopped at the first time the subcritical birth-rate bound ceases to hold, so the birth–death comparison remains adapted. In the Poisson-equation argument, a decomposition by jump type shows that the corrector martingale has predictable quadratic variation $o_{\mathbb{P}}(N^2)$.

7.2 Localized diffusion identification and removal of localization

Define

$$b_{\gamma}(x) = g(x) - \beta - \gamma(x - 1), \quad x > 1.$$

By Propositions 15 and 16, the stopped population drift is asymptotically b_{γ} . It remains to identify the population martingale jointly with the initial state and stopping time, identify the limiting first-exit time, and show that the localization boundaries are asymptotically irrelevant.

The next lemma gives the first-exit regularity required for convergence of the stopping times.

Lemma 9 (First-exit crossing and exit-time continuity). *Fix $T > 0$ and $1 < \ell < u < \infty$, and let Y be a continuous adapted solution of $dY(t) = b_{\gamma}(Y(t)) dt + \sqrt{2} dB(t)$ up to its first exit from an open neighborhood of $[\ell, u]$, where B is Brownian with respect to a filtration containing $Y(0)$. For $z \in C[0, T]$, set*

$$\tau_{\ell, u}(z) = \inf\{t \in [0, T] : z(t) \notin (\ell, u)\} \wedge T, \quad \inf \emptyset = \infty.$$

If $Y(0) \in (\ell, u)$ and $\tau_{\ell, u}(Y) < T$, then, almost surely, for every $\varepsilon > 0$ there is $s \in (\tau_{\ell, u}(Y), \tau_{\ell, u}(Y) + \varepsilon) \cap [0, T]$ with $Y(s) \notin [\ell, u]$; if $Y(0) \in \{\ell, u\}$, the same statement holds with $s \in (0, \varepsilon) \cap [0, T]$. At every continuous path with the corresponding first-exit property, the capped exit-time functional is continuous under uniform convergence on $[0, T]$.

With the first-exit map controlled, joint tightness and the martingale problem identify the stopped limiting process.

Proposition 17 (Localized C -tightness and identification). *For every fixed $T < \infty$, $0 < \eta < 1$, and $R > 2$, the stopped processes*

$$\widehat{X}_{\eta, R}^{N, \gamma}(t) = \widehat{X}^{N, \gamma}(t \wedge \tau_N^{\eta, R}), \quad 0 \leq t \leq T,$$

are C -tight. Every subsequential limit is the absorbed solution

$$X_{\eta, R}^{\gamma}(t) = X(0) + \sqrt{2}B(t \wedge \tau^{\eta, R}) + \int_0^{t \wedge \tau^{\eta, R}} b_{\gamma}(X_{\eta, R}^{\gamma}(s)) ds, \quad (64)$$

where B is Brownian with respect to the joint limiting filtration and is independent of $X(0)$. Here $\tau^{\eta, R}$ is the first exit from $(1 + \eta, R)$, capped at T ; if $X(0) \notin (1 + \eta, R)$, then $\tau^{\eta, R} = 0$ and the solution is absorbed at its initial value. In particular, the stopped limit is unique in law and the entire stopped sequence converges.

The localized result must next be extended to the full state space. The following lemma controls escape through both localization boundaries and yields C -tightness of the unstopped processes.

Lemma 10 (Compact containment and removal of localization). *For every $T < \infty$,*

$$\lim_{\eta \downarrow 0} \lim_{R \uparrow \infty} \limsup_{N \rightarrow \infty} \mathbb{P} \left(\inf_{0 \leq t \leq T} \hat{X}^{N,\gamma}(t) \leq 1 + 2\eta \text{ or } \sup_{0 \leq t \leq T} \hat{X}^{N,\gamma}(t) \geq R \right) = 0. \quad (65)$$

Consequently, the full sequence $\{\hat{X}^{N,\gamma}\}$ is C -tight on every compact time interval.

The detailed proof first establishes joint tightness of the initial state, stopped population, extended martingale, and exit time; see Lemma 19. A martingale-problem argument in the joint limiting filtration then identifies the limiting driver as a Brownian motion independent of the initial state; see Lemma 20. Lemma 9 identifies the limiting exit time, and Lemma 10 removes localization in the order $N \rightarrow \infty$, $R \uparrow \infty$, and $\eta \downarrow 0$. These results complete the proof of Theorem 5.

Proof of Theorem 5. By the C -tightness conclusion of Lemma 10, every subsequence has a further subsequence converging in $C[0, T]$. It remains to identify an arbitrary subsequential limit.

Fix $0 < \eta < 1$ and $R > 2$, and apply Proposition 17 with upper boundary $2R$. On the event

$$\left\{ \inf_{0 \leq t \leq T} \hat{X}^{N,\gamma}(t) > 1 + 2\eta, \quad \sup_{0 \leq t \leq T} \hat{X}^{N,\gamma}(t) < R \right\},$$

the full process and the process stopped on leaving $(1 + \eta, 2R)$ agree on $[0, T]$. By Lemma 10, the probability of the complementary event can be made arbitrarily small by taking $N \rightarrow \infty$, then $R \uparrow \infty$, and finally $\eta \downarrow 0$.

By Proposition 17, the stopped processes converge to the absorbed solution of (64). The boundary classification established in the proof of Proposition 4 shows that the candidate diffusion is nonexplosive and does not reach the lower endpoint. Hence its paths stopped on leaving $(1 + \eta, 2R)$ converge uniformly on $[0, T]$ to the global solution as $R \uparrow \infty$ and $\eta \downarrow 0$. Hence, every subsequential limit solves (13) on $[0, T]$.

The drift b_γ is locally Lipschitz on $(1, \infty)$. Local pathwise uniqueness on an increasing sequence of compact subintervals, together with the boundary classification and nonexplosion, therefore gives global pathwise uniqueness. The limiting law is unique, and (12) follows on $[0, T]$. Since T is arbitrary, the convergence holds uniformly on compact time intervals. $\square$

8 Conclusion and outlook

We have established steady-state validity of the NDS diffusion approximation for the stationary marginal of normalized total population in the JSQ system with one queue at each server. In the baseline model, this marginal converges to the invariant law of the Gupta–Walton diffusion, the two iterated limits agree, and stationary moments converge. For the abandonment model, we proved the normalized-population diffusion limit from balanced initial queue lengths and, separately, convergence of the finite-system stationary normalized populations. Marked excursions provide the stationary occupation bounds, while a finite- N Poisson equation removes the centered fast variable from the stationary generator.

The explicit invariant laws give a scalar equilibrium description of congestion in the NDS regime. Since stationary idleness is of order one, the normalized number of waiting customers has the same limit as the normalized population minus one. With abandonment, the stationary mean aggregate abandonment rate therefore converges to $\gamma\mathbb{E}(X_{\beta,\gamma} - 1)$. The abandonment parameter leaves the singular lower-boundary behavior and diffusion coefficient unchanged, but adds a linear mean-reverting drift; the resulting limiting invariant density has Gaussian-type rather than exponential upper-tail decay.

The analysis assumes Poisson arrivals, exponential service and patience, and homogeneous servers. It does not establish convergence of the stationary occupancy profile or waiting-time distributions, rates of convergence, a joint N -and-time limit, uniform mixing, or stationary process-level convergence. Extending the framework requires a suitable auxiliary excursion coordinate, control of the frozen fast process, and, for more general process initializations, explicit initial-layer estimates. General service or patience distributions require augmented, generally measure-valued state descriptors; server heterogeneity and limited-information routing require new state-space-collapse and occupation-time estimates.

References

- Rami Atar. A Diffusion Regime with Nondegenerate Slowdown. *Operations Research*, 60(2):490–500, 2012. doi: 10.1287/opre.1110.1030.
- François Baccelli and Pierre Brémaud. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*, volume 26 of *Stochastic Modelling and Applied Probability*. Springer, Berlin, Heidelberg, second edition, 2003. ISBN 978-3-540-66088-0. doi: 10.1007/978-3-662-11657-9.
- Sanidhay Bhambay, Burak Büke, and Arpan Mukhopadhyay. Asymptotic Optimality of the Speed-Aware Join-the-Shortest-Queue in the Halfin–Whitt Regime for Heterogeneous Systems. *Stochastic Systems*, 15(2):147–193, 2025. doi: 10.1287/stsy.2023.0054.
- Anton Braverman. Steady-State Analysis of the Join-the-Shortest-Queue Model in the Halfin–Whitt Regime. *Mathematics of Operations Research*, 45(3):1069–1103, 2020. doi: 10.1287/moor.2019.1023.
- Anton Braverman. The Join-the-Shortest-Queue System in the Halfin-Whitt Regime: Rates of Convergence to the Diffusion Limit. *Stochastic Systems*, 13(1):1–39, 2023. doi: 10.1287/stsy.2022.0102.
- Anton Braverman, J. G. Dai, and Masakiyo Miyazawa. Heavy Traffic Approximation for the Stationary Distribution of a Generalized Jackson Network: The BAR Approach. *Stochastic Systems*, 7(1):143–196, 2017. doi: 10.1287/15-SSY199.
- Amarjit Budhiraja and Chihoon Lee. Stationary Distribution Convergence for Generalized Jackson

- Networks in Heavy Traffic. *Mathematics of Operations Research*, 34(1):45–56, 2009. doi: 10.1287/moor.1080.0353.
- Ping Cao, Shuangchi He, Junfei Huang, and Yunan Liu. To Pool or Not to Pool: Queueing Design for Large-Scale Service Systems. *Operations Research*, 69(6):1866–1885, 2021. doi: 10.1287/opre.2019.1976.
- J. G. Dai, Shuangchi He, and Tolga Tezcan. Many-Server Diffusion Limits for $G/Ph/n + GI$ Queues. *The Annals of Applied Probability*, 20(5):1854–1890, 2010. doi: 10.1214/09-AAP674.
- J. G. Dai, A. B. Dieker, and Xuefeng Gao. Validity of Heavy-Traffic Steady-State Approximations in Many-Server Queues with Abandonment. *Queueing Systems*, 78(1):1–29, 2014. doi: 10.1007/s1134-014-9394-x.
- Patrick Eschenfeldt and David Gamarnik. Join the Shortest Queue with Many Servers. The Heavy-Traffic Asymptotics. *Mathematics of Operations Research*, 43(3):867–886, 2018. doi: 10.1287/moor.2017.0887.
- David Gamarnik and Assaf Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *The Annals of Applied Probability*, 16(1):56–90, 2006. doi: 10.1214/105051605000000638.
- Varun Gupta and Neil Walton. Load Balancing in the Nondegenerate Slowdown Regime. *Operations Research*, 67(1):281–294, 2019. doi: 10.1287/opre.2018.1768.
- Itai Gurvich, Junfei Huang, and Avishai Mandelbaum. Excursion-Based Universal Approximations for the Erlang-A Queue in Steady-State. *Mathematics of Operations Research*, 39(2):325–373, 2014. doi: 10.1287/moor.2013.0606.
- Shlomo Halfin and Ward Whitt. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*, 29(3):567–588, 1981. doi: 10.1287/opre.29.3.567.
- Junfei Huang, Hanqin Zhang, and Jiheng Zhang. A Unified Approach to Diffusion Analysis of Queues with General Patience-Time Distributions. *Mathematics of Operations Research*, 41(3):1135–1160, 2016. doi: 10.1287/moor.2015.0772.
- Daniela Hurtado-Lange and Izzy Grosof. Markov modulated JSQ in heavy traffic via the poisson equation, 2026. arXiv:2603.00387v1, submitted February 28, 2026.
- Prakirt R. Jhunjunwala, Martin Zubeldia, and Siva Theja Maguluri. Join-the-Shortest Queue with Abandonment: Critically Loaded and Heavily Overloaded Regimes. *Mathematics of Operations Research*, 51(1):398–439, 2025. doi: 10.1287/moor.2023.0098.
- Prakirt Raj Jhunjunwala, Daniela Hurtado-Lange, and Siva Theja Maguluri. Exponential tail bounds on queues: A confluence of non-asymptotic heavy traffic and large deviations. *ACM*

- SIGMETRICS Performance Evaluation Review*, 51(4):18–19, 2024. doi: 10.1145/3649477.3649488. Full version: arXiv:2306.10187.
- Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Communications and Control Engineering. Springer-Verlag, London, first edition, 1993. ISBN 978-3-540-19832-1. doi: 10.1007/978-1-4471-3267-7.
- Zedong Wang and Siva Theja Maguluri. Tail Bounds for Queues with Abandonment: Constant, Moderate, Large Deviations, and Efficient Concentration, 2026. arXiv:2603.14163v2, revised March 19, 2026.
- Ward Whitt. Proofs of the Martingale FCLT. *Probability Surveys*, 4:268–302, 2007. doi: 10.1214/07-PS122.
- Zhisheng Zhao, Sayan Banerjee, and Debankur Mukherjee. Many-server asymptotics for join-the-shortest-queue in the super-halfin–whitt scaling window. *Mathematics of Operations Research*, 2025. doi: 10.1287/moor.2021.0133. Articles in Advance.

Appendix

The appendix contains the auxiliary arguments and detailed proofs deferred from the main text. Appendix A records common lemmas and fixed-system foundations, Appendix B supports the marked-excursion analysis in Section 3, Appendix C supplies the remaining technical proofs for the baseline stationary analysis, and Appendix D contains the technical proofs for the abandonment results.

A Common auxiliary lemmas and fixed-system foundations

Lemma 11 (Ergodicity through a skeleton chain). *Let Y be a time-homogeneous Markov process with invariant probability measure π . Suppose that, for some $h > 0$, the h -skeleton is ψ -irreducible, aperiodic, and positive Harris recurrent. Then*

$$\|\text{Law}_y(Y(t)) - \pi\|_{\text{TV}} \longrightarrow 0, \quad t \rightarrow \infty,$$

for every initial state y , and hence from every initial law.

Proof. The aperiodic ergodic theorem for the h -skeleton gives total-variation convergence at times nh . Write $t = nh + s$ with $s \in [0, h)$. Invariance of π and contraction of total variation under a Markov kernel give

$$\|P_t(y, \cdot) - \pi\|_{\text{TV}} \leq \|P_{nh}(y, \cdot) - \pi\|_{\text{TV}}.$$

Integration with respect to an arbitrary initial law proves the final assertion. $\square$

Lemma 12 (Scale–speed classification and ergodicity). *Let $D = (\ell, r)$ and let b be locally Lipschitz on D . Consider*

$$dY(t) = b(Y(t)) dt + \sqrt{2} dB(t),$$

up to its first exit from D . Fix $x_0 \in D$ and define

$$s'(x) = \exp \left\{ - \int_{x_0}^x b(y) dy \right\}, \quad m(x) = \frac{1}{s'(x)}.$$

If

$$\int_{\ell}^{x_0} s'(x) dx = \int_{x_0}^r s'(x) dx = \infty, \quad \int_{\ell}^r m(x) dx < \infty,$$

then the solution is nonexplosive and positive recurrent, with unique invariant density $m / \int_D m$. If its transition kernel has a strictly positive density on D , then its law converges to the invariant law in total variation from every initial law supported on D .

Proof. The scale-function hitting formula and the two divergent scale integrals make both endpoints inaccessible. Finite total speed measure then gives positive recurrence and invariant measure proportional to $m(x) dx$ by the standard scale–speed classification for regular one-dimensional diffusions. A strictly positive transition density makes the time- h skeleton, for any $h > 0$, irreducible and aperiodic. The invariant probability measure makes that skeleton positive Harris recurrent, so Lemma 11 gives total-variation convergence and uniqueness. $\square$

Lemma 13 (Moment convergence from exponential integrability). *Let $Y_n \geq 0$ and suppose that $Y_n \Rightarrow Y$ and, for some $a > 0$ and n_0 ,*

$$\sup_{n \geq n_0} \mathbb{E} e^{aY_n} < \infty.$$

Then, for every integer $k \geq 1$ and every $0 < \eta < a$,

$$\mathbb{E} Y_n^k \longrightarrow \mathbb{E} Y^k, \quad \mathbb{E} e^{\eta Y_n} \longrightarrow \mathbb{E} e^{\eta Y}.$$

Proof. For each k , $x^k \leq C_{k,a} e^{ax/2}$ on $[0, \infty)$, so $\{Y_n^k : n \geq n_0\}$ is uniformly integrable. For $0 < \eta < a$, choose $p > 1$ with $p\eta \leq a$; then $\{e^{\eta Y_n} : n \geq n_0\}$ is bounded in L^p and hence uniformly integrable. Weak convergence and the continuous mapping theorem complete both limits. $\square$

Lemma 14 (Fixed-system ergodicity). *For every fixed $N > \beta$, both Q^N and $Q^{N,\gamma}$ converge in total variation to their respective stationary laws from every initial law.*

Proof. The two chains are irreducible, nonexplosive, and positive recurrent by Propositions 1 and 3. For either model, every prescribed finite transition path has positive probability of occurring during one unit of time with no other effective mark, so the time-one skeleton is irreducible. The probability of no effective transition during one unit of time is positive at every state, so the skeleton is aperiodic. Its invariant probability measure makes it positive Harris recurrent, and Lemma 11 applies. $\square$

A.1 Fixed-system recurrence, moments, and stationary identities

Proof of Proposition 1. From any state there is positive probability of enough service completions and no arrivals to reach zero. Conversely, from zero, $N \max_i q_i$ arrivals with no service fill all queues to the common level $\max_i q_i$. A prescribed sequence of completions can then reach any target q , so the chain is irreducible.

For positive recurrence, set $\mathcal{W}_2(q) = \sum_i q_i^2$. An arrival changes $\mathcal{W}_2$ by $2m(q) + 1$ and a completion at queue i by $-2q_i + 1$, so

$$\mathcal{L}_N \mathcal{W}_2(q) = \lambda_N (2m(q) + 1) - 2S(q) + N - I(q) \leq -\frac{2\beta}{N} S(q) + 2N - \beta.$$

The norm-like function $\mathcal{W}_2$ therefore has strictly negative drift outside a finite set. The continuous-time Foster criterion gives positive recurrence (Meyn and Tweedie, 1993, Chapter 4). $\square$

Proof of Lemma 1. Fix any $1 < z < N/(N - \beta)$ and put $\Phi_z(q) = \sum_i z^{q_i}$. Since an arrival contributes $(z - 1)z^{m(q)}$ and completions contribute $-(1 - z^{-1})(\Phi_z - I)$,

$$\mathcal{L}_N \Phi_z(q) \leq (z - 1) \left(\frac{N - \beta}{N} - \frac{1}{z} \right) \Phi_z(q) + N(1 - z^{-1}) = -c_z \Phi_z(q) + N(1 - z^{-1})$$

for some $c_z > 0$. Geometric Foster–Lyapunov theory gives (1); for fixed N , every polynomial in S^N is bounded by a constant times $1 + \Phi_z$. $\square$

Proof of Lemma 2. Integrability follows from Lemma 1. Start in stationarity and stop at $\tau_m = \inf\{t : S^N(t) \geq m\}$. Before τ_m the state space is finite, so Dynkin's formula applies. If $N_A^N(t)$ is the arrival count, then

$$\sup_{s \leq t} S^N(s \wedge \tau_m) \leq S^N(0) + N_A^N(t),$$

whose moments of every order are finite independently of m . The growth bound permits dominated convergence as $m \rightarrow \infty$, giving

$$\mathbb{E}F(Q^N(t)) - \mathbb{E}F(Q^N(0)) = \mathbb{E} \int_0^t (\mathcal{L}_N F)(Q^N(s)) ds.$$

Stationarity makes the left side zero and the right side $t\mathbb{E}(\mathcal{L}_N F)(Q^N)$. □

B Marked-excursion construction and auxiliary occupation estimates

B.1 Marked excursions and compensators

We realize the process on the canonical two-sided path space $D(\mathbb{R}, \mathbb{Z}_+^N)$, equipped with the Borel σ -field generated by the coordinate maps and the stationary law of the coordinate process Q^N . For $s \in \mathbb{R}$, the time shift is $\vartheta_s \omega(t) = \omega(t + s)$. We use the completed right-continuous natural filtration and suppress the superscript N throughout this subsection. Since the total jump rate is at most $2N$, every sample path has only finitely many jumps on compact time intervals.

The irreducible positive recurrent chain is ergodic under its stationary law, and $\pi_N(\{(2, \dots, 2)\}) > 0$. Applying the ergodic theorem to the indicator of this state on the positive and negative half-lines shows that $\{V = 0\}$ is visited at arbitrarily large positive and negative times almost surely. Together with local finiteness of the jump times, this implies that every time with $V > 0$ belongs to a unique finite excursion away from zero.

For a deterministic state q , $\mathbb{P}_q$ and $\mathbb{E}_q$ refer to the original chain started from q . The notations $\mathbb{E}_V^0$ and $\mathbb{E}_I^0$ denote Palm expectation for Ξ_V and Ξ_I , respectively; the generic notation $\mathbb{E}^0$ in Lemma 15 applies to whichever stationary point process is under discussion. Palm expectation is used only to average post-jump marks, while all Markov estimates below are uniform deterministic-state estimates under the original laws $\mathbb{P}_q$. Here δ_t is the unit point mass at t , and under Palm indexing the distinguished point is at time zero.

The following lemma packages the general Palm and birth–death facts used in the excursion estimates.

Lemma 15 (Palm identity and birth–death estimates). *(a) Let $\Xi = \sum_k \delta_{T_k}$ be a stationary simple point process of finite intensity ι , and attach nonnegative shift-covariant profiles $Y_k(u)$ that vanish for $u < 0$. If $G(t) = \sum_k Y_k(t - T_k)$, then*

$$\mathbb{E}G(0) = \iota \mathbb{E}^0 \int_0^\infty Y_0(u) du,$$

with the value 0 when $\iota = 0$.

(b) Let Y be an adapted nonnegative integer-valued unit-jump process with $Y(0) = i \geq 0$, let $\tau \leq \tau_0 := \inf\{t : Y(t) = 0\}$, and suppose on $[0, \tau)$ that upward jumps occur at predictable rate at most b , downward jumps at rate d , and other jumps do not change Y . If $0 \leq b < d$, then

$$\mathbb{E} \int_0^\tau Y(t) dt \leq F_{b,d}(i), \quad F_{b,d}(i) = \frac{i^2}{2(d-b)} + \frac{(b+d)i}{2(d-b)^2}.$$

In particular, $F_{b,d}(1) = d/(d-b)^2$.

(c) If, from $Y(0) = 1$, the process is stopped on first hitting 0, L , or an additional killing set, with upward rate at most $b < d$ and downward rate d , then

$$\mathbb{P}(Y \text{ reaches } L \text{ before being stopped}) \leq (b/d)^{L-1}.$$

The proof is given in Section B.2. The Palm excursion proposition in the main text states the point processes, exact compensators, reward decomposition, and reward estimates. Figure 1 records the endpoint conventions used in that statement.

Proof of Proposition 5. All assertions are on the canonical two-sided realization above. Local finiteness of the jump times makes Ξ_V and Ξ_I simple and locally finite. Their definitions commute with every time shift ϑ_s , so stationarity of the coordinate process gives stationarity of both point processes.

For measurability, the events $\{\tau(\sigma) - \sigma > u\}$ and $\{\chi(\sigma) - \sigma > u\}$ are determined by the path on $[\sigma, \sigma + u]$ and may be checked at rational times together with the endpoint. Thus τ and χ are measurable hitting-time functionals. The containing start $\sigma(\alpha)$ is the last locally finite marked start preceding α whose terminal time has not occurred, so it and $\tau_I(\alpha)$ are measurable. Evaluation of the càdlàg coordinate path at these times gives joint measurability of the profiles. The identities

$$\begin{aligned} \mathcal{E}_V(\vartheta_s \omega) &= \mathcal{E}_V(\omega) - s, & \tau_{\vartheta_s \omega}(\sigma - s) &= \tau_\omega(\sigma) - s, & \chi_{\vartheta_s \omega}(\sigma - s) &= \chi_\omega(\sigma) - s, \\ \mathcal{E}_I(\vartheta_s \omega) &= \mathcal{E}_I(\omega) - s, & \tau_{I, \vartheta_s \omega}(\alpha - s) &= \tau_{I, \omega}(\alpha) - s \end{aligned}$$

prove shift covariance of the point processes and marks.

Formula (22) depends only on the strict past. Its endpoint conventions make it left-continuous: the indicator is zero at the starting jump, remains one through a jump that first reaches L or terminates the marked excursion, and becomes zero immediately afterward. Hence $\Gamma_{K,L}$, Θ_V , and Θ_I are predictable. When $V(t-) = 0$, one has $I(t-) = M_1(t-) = 0$, and a jump $V : 0 \rightarrow 1$ is exactly a service completion from a two-job queue. Its rate is $M_2(t-)$ and it leaves Z_3 unchanged, giving Λ_V . On $\{\Theta_I = 1\}$, a service completion from a one-job queue is exactly a transition $I : 0 \rightarrow 1$ whose post-jump deficit remains below L ; its rate is $M_1(t-)$. Conversely, every point of Ξ_I is such a completion, giving Λ_I . On $\{\Theta_V = 1\}$, $M_2 = N - Z_3 \leq N - K$, while on $\{\Theta_I = 1\}$, $M_1 \leq V \leq L - 2 < L$. The exact compensators in (23) therefore imply finite intensity and (24).

To prove (25), fix a nonjump time t . If $I(t) = 0$ or $Z_3(t) < K$, both sides vanish. Otherwise t belongs to a unique finite V -excursion with start σ . Because Z_3 cannot increase while $V > 0$,

one has $Z_3(\sigma) \geq Z_3(t) \geq K$, so $\sigma \in \mathcal{E}_V$ and $t < \tau(\sigma)$. If $t < \chi(\sigma)$, the current positive I -excursion has a unique start $\alpha \in \mathcal{E}_I$, and exactly $R_\alpha^{\text{pre}}(t - \alpha)$ equals $I(t)$. If $t > \chi(\sigma)$, all pre-hit profiles have ended and exactly $R_\sigma^{\text{post}}(t - \sigma)$ equals $I(t)$. The half-open definitions allocate the first hitting time to the post-hit profile and exclude the terminal time. Termination through $V = 0$ gives $I = 0$, while termination through $Z_3 < K$ makes the indicator on the left zero. This proves the pathwise identity away from the locally finite jump set. The Palm identity in Lemma 15(a) then gives (26).

Both point processes are predictable thinnings of the Markov-chain jump measure. Their successive points after a deterministic time are stopping times. Conditional on the post-jump state at such a point, the strong Markov property gives the original future law $\mathbb{P}_q$; disintegration of the Palm law changes only the distribution of q . For a pre-hit I -excursion, the post-jump state has $I = 1$, $V < L$, and $Z_3 \geq K$. Until its stopping time, the upward rate of I is $M_1 \leq V < L$ and its downward rate is $N - \beta$. The deterministic-state estimate in Lemma 15(b), uniform in the post-jump state, proves (27).

For a marked V -excursion started from any post-jump state q with $V(q) = 1$ and $Z_3(q) \geq K$, the upward rate of V before τ is at most $N - K$ and the downward rate is $N - \beta$. Part (c) of Lemma 15 gives (28). On $\{\chi < \tau\}$, unit jumps imply $V(\chi) = L$. From every possible hitting state q' with $Z_3(q') \geq K$, the inequality $I \leq V/2$ and part (b) give (29). The strong Markov property at χ yields

$$\mathbb{E}_q \left[\mathbb{1}_{\{\chi < \tau\}} \int_\chi^\tau I(t) dt \right] \leq \mathbb{P}_q(\chi < \tau) \sup_{\substack{q': V(q')=L \\ Z_3(q') \geq K}} \mathbb{E}_{q'} \int_0^{\tau'} I(t) dt.$$

Conditioning under $\mathbb{E}_V^0$ on the post-jump state proves the claimed post-hit integrability. No independence under the Palm law is used. $\square$

B.2 Proofs for the excursion estimates

Proof of Lemma 15. For part (a), if $\iota = 0$ the point process is empty almost surely. Otherwise, stationarity, Tonelli's theorem, and the refined Campbell formula (Baccelli and Brémaud, 2003) give

$$\mathbb{E}G(0) = \mathbb{E} \int_0^1 G(t) dt = \iota \int_{\mathbb{R}} \mathbb{E}^0 \left[\int_0^1 Y_0(t-s) dt \right] ds = \iota \mathbb{E}^0 \int_0^\infty Y_0(u) du.$$

For part (b), a direct calculation shows, for $i \geq 1$,

$$b(F_{b,d}(i+1) - F_{b,d}(i)) + d(F_{b,d}(i-1) - F_{b,d}(i)) = -i.$$

The function $F_{b,d}$ is increasing, so replacing b by a smaller predictable upward rate only decreases its drift. Apply Dynkin's formula after stopping additionally at a deterministic time T and at $\tau_m^Y = \inf\{t : Y(t) \geq m\}$. This gives

$$\mathbb{E} \int_0^{\tau \wedge T \wedge \tau_m^Y} Y(t) dt \leq F_{b,d}(i).$$

Letting first $m \rightarrow \infty$ and then $T \rightarrow \infty$, monotone convergence proves the claim.

For part (c), the case $b = 0$ is immediate. Otherwise put $z = d/b > 1$. Before stopping, the predictable drift of $z^{Y(t)}$ is

$$b_t z^{Y(t-)}(z - 1) + dz^{Y(t-)}(z^{-1} - 1) = z^{Y(t-)}(z - 1)(b_t - b) \leq 0.$$

Thus the stopped process is a bounded supermartingale. If p_L is the probability of reaching L , optional stopping yields $z \geq p_L z^L + (1 - p_L)$ and hence $p_L \leq z^{-(L-1)}$. $\square$

Proof of Proposition 7. The marked V -excursions have intensity $\iota_V \leq N - K \leq N$; we use the latter bound to retain the simpler expression in (31). Start the original chain from any possible post-jump state q with $V(q) = 1$ and $Z_3(q) \geq K$, and let

$$T_v = \inf\{t : V(t) \geq v\}, \quad T_K = \inf\{t : V(t) = 0 \text{ or } Z_3(t) < K\}.$$

By Lemma 15(c),

$$\mathbb{P}_q(T_v < T_K) \leq \left(\frac{N - K}{N - \beta} \right)^{v-1}.$$

On $\{T_v < T_K\}$, unit jumps give $V(T_v) = v$. To bound the remaining duration uniformly, start instead from any state q' with $V(q') = v$ and $Z_3(q') \geq K$, and set $\tau_m = \inf\{t : V(t) \geq m\}$ for $m > v$. Before T_K ,

$$(\mathcal{L}_N V)(Q(t)) = \beta - I(t) - Z_3(t) \leq -(K - \beta).$$

Dynkin's formula at $T_K \wedge T \wedge \tau_m$ gives

$$0 \leq \mathbb{E}_{q'} V(T_K \wedge T \wedge \tau_m) \leq v - (K - \beta) \mathbb{E}_{q'} [T_K \wedge T \wedge \tau_m].$$

Letting first $m \rightarrow \infty$ and then $T \rightarrow \infty$ yields

$$\sup_{\substack{q': V(q')=v \\ Z_3(q') \geq K}} \mathbb{E}_{q'} T_K \leq \frac{v}{K - \beta}.$$

The strong Markov property at T_v , used conditionally rather than as an independence assertion, now gives

$$\begin{aligned} \mathbb{E}_q [\mathbf{1}_{\{T_v < T_K\}}(T_K - T_v)] &= \mathbb{E}_q [\mathbf{1}_{\{T_v < T_K\}} \mathbb{E}_{Q(T_v)} T_K] \\ &\leq \mathbb{P}_q(T_v < T_K) \sup_{\substack{q': V(q')=v \\ Z_3(q') \geq K}} \mathbb{E}_{q'} T_K \\ &\leq \frac{v}{K - \beta} \left(\frac{N - K}{N - \beta} \right)^{v-1}, \end{aligned}$$

where the inner T_K is the exit time for the restarted chain. Attach to each marked excursion the indicator profile of the interval from its first hit of v until its exit time T_K . This profile dominates the occupation of $\{V \geq v, Z_3 \geq K\}$ within that excursion. Campbell's formula, conditioning under $\mathbb{E}_V^0$ on the post-jump state and using the uniform deterministic-state bound above, together with $\iota_V \leq N$, proves (31). $\square$

C Technical proofs for the baseline stationary analysis

C.1 Invariant law and ergodicity of the limiting diffusion

Proof of Proposition 2. Write $b(x) = g(x) - \beta$. With normalization at two, the scale and speed densities are

$$s'(x) = \begin{cases} \frac{e^{(1+\beta)(x-2)}}{x-1}, & 1 < x \leq 2, \\ e^{\beta(x-2)}, & x \geq 2, \end{cases} \quad m(x) = \frac{1}{s'(x)}.$$

The scale integrals diverge at both endpoints, while

$$\int_1^\infty m(x) dx = \int_1^2 (x-1)e^{-(1+\beta)(x-2)} dx + \int_2^\infty e^{-\beta(x-2)} dx < \infty.$$

The nondegenerate constant diffusion coefficient gives a strictly positive transition density on the interior. Thus Lemma 12 gives nonexplosion, positive recurrence, uniqueness, and total-variation ergodicity. The invariant density is proportional to m , and

$$\left(\int_1^\infty m(x) dx \right)^{-1} = \frac{\beta(1+\beta)^2}{1+\beta e^{1+\beta}} = C_\beta,$$

which yields (4). □

C.2 Auxiliary proof of the uniform exponential-moment bound

Proof of Proposition 9. Fix $x_* > 2$. By Proposition 8, after increasing N_0 ,

$$J_N(x_*) \leq C_0 N^{-1/2}, \quad N \geq N_0. \quad (66)$$

Choose c_* so that

$$\frac{1}{\theta - a} < c_* < \frac{1}{2a}, \quad (67)$$

which is possible because $3a < \theta$, and set $\ell_N = c_* \log N$. Increase N_0 so that $\ell_N > x_*$ for every $N \geq N_0$.

For every nonnegative random variable X ,

$$e^{aX} = 1 + a \int_0^X e^{ax} dx.$$

Multiplying by I^N , taking expectations, and using Tonelli's theorem gives

$$\mathbb{E}[I^N e^{aX_N}] = \beta + a \int_0^\infty e^{ax} J_N(x) dx. \quad (68)$$

Split the integral at x_* and ℓ_N . On $[0, x_*]$, use $J_N(x) \leq \mathbb{E}I^N = \beta$. On $[x_*, \ell_N]$, monotonicity and (66) give a bound of order

$$N^{-1/2} e^{a\ell_N} = N^{-1/2+ac_*},$$

which is uniformly bounded by the upper inequality in (67). Finally, (18) gives

$$J_N(x) \leq \kappa_\perp N e^{-\theta x}.$$

Hence the integral over $[\ell_N, \infty)$ is bounded by a constant times

$$N e^{-(\theta-a)\ell_N} = N^{1-c_*(\theta-a)},$$

which is uniformly bounded by the lower inequality in (67). Substitution in (68) proves the proposition. $\square$

C.3 Auxiliary proofs for stationary averaging

Proof of Lemma 4. Let $j_0 = \lceil 2\beta \rceil \vee 1$. For $j_0 \leq j < N$ and all sufficiently large N ,

$$\frac{N-j}{N-\beta} = 1 - \frac{j-\beta}{N-\beta} \leq e^{-j/(2N)}.$$

Hence, for $j > j_0$, iterating (35) yields

$$p_j^{I,N} \leq p_{j_0}^{I,N} \exp \left\{ -\frac{1}{2N} \sum_{k=j_0}^{j-1} k \right\} \leq C e^{-cj^2/N},$$

for constants $c, C > 0$ independent of j and N ; the finitely many levels below j_0 are absorbed by increasing C . Summing the bound over $j = L, \dots, N$ and decreasing c if needed proves (36). Taking $L = N^{3/4}$ in that estimate, and using $I^N \leq N$ and $\mathbb{E}I^N = \beta$, gives

$$\mathbb{E}(I^N)^2 \leq N^{3/4} \mathbb{E}I^N + N^2 \mathbb{P}(I^N > N^{3/4}) \leq \beta N^{3/4} + C N^3 e^{-cN^{1/2}}.$$

Dividing by N proves the result. $\square$

Proof of Proposition 10. Set

$$\tilde{K}_N = \lceil N^{1/2} \rceil, \quad \delta_N = N^{-1/4}, \quad \Delta_N = \left| \frac{M_1^N}{N} - r(X_N) \right|.$$

We first prove that

$$\mathbb{P}(\Delta_N > \delta_N) \leq C N^c e^{-c' N^{1/4}} \tag{69}$$

for suitable constants $c, c', C > 0$.

On $\{Z_3^N < \tilde{K}_N\}$, at least one queue has length at most two. Hence $X_N \leq 2 + D_N$. For every queue with length at least three,

$$Q_i^N - 2 \leq (Q_i^N - X_N) + (X_N - 2) \leq 2D_N.$$

Thus

$$H^N \leq 2\tilde{K}_N D_N \quad \text{on } \{Z_3^N < \tilde{K}_N\}.$$

By (37), on this event

$$\Delta_N \leq \frac{2\tilde{K}_N D_N + 2I^N}{N}.$$

If $\Delta_N > \delta_N$, then either

$$D_N > \frac{\delta_N N}{4\tilde{K}_N} \asymp N^{1/4} \quad \text{or} \quad I^N > \frac{\delta_N N}{4} \asymp N^{3/4}.$$

The probabilities are stretched-exponentially small by (17) and (36).

On $\{Z_3^N \geq \tilde{K}_N\}$, both M_1^N and $(V^N - H^N)^+$ lie in $[0, V^N]$. Therefore

$$\Delta_N \leq \frac{V^N}{N}.$$

Apply Proposition 7 with $K = \tilde{K}_N$ and $v = \lceil \delta_N N \rceil$. The right side of (31) is bounded by a polynomial in N times $e^{-cN^{1/4}}$. This proves (69).

Finally, $0 \leq M_1^N/N \leq 1$ and $0 \leq r(X_N) \leq 2$, so $0 \leq \Delta_N \leq 2$; also $I^N \leq N$. Splitting according to $\{\Delta_N \leq \delta_N\}$ gives

$$\mathbb{E}[(I^N + 1)\Delta_N] \leq \delta_N(\mathbb{E}I^N + 1) + 2(N + 1)\mathbb{P}(\Delta_N > \delta_N) \leq (\beta + 1)N^{-1/4} + o(1),$$

which proves (38). $\square$

Proof of Lemma 5. Fix $q \in \mathbb{Z}_+^N$, retain the deterministic values $x = x_N(q)$, $i = I(q)$, and $m = M_1^N(q)$, and put $\delta = 1/N$. Using (43)–(44), split each transition into a same- x increment and an x -shift. The same- x part of the actual generator is

$$(N - \beta)\mathbb{1}_{\{i > 0\}}[\Psi_N(x, i - 1) - \Psi_N(x, i)] + m[\Psi_N(x, i + 1) - \Psi_N(x, i)],$$

where the two increments are given by (40)–(41). Subtracting the frozen same- x generator gives

$$- \beta\mathbb{1}_{\{i > 0\}} \frac{a_h(x)i}{N} - (m - Nr(x)) \frac{a_h(x)(i + 1)}{N}.$$

Its absolute value is bounded by

$$C_{h,\beta} \left[(i + 1) \left| \frac{m}{N} - r(x) \right| + \frac{(i + 1)^2}{N} \right]. \quad (70)$$

It remains to control the x -shifts. Since a_h is Lipschitz,

$$|\Psi_N(x \pm \delta, j) - \Psi_N(x, j)| \leq \frac{C_{h,\beta}}{N^2} j(j + 1).$$

The total rate of all transitions is at most $2N$, and the shifted indices are among $i - 1, i, i + 1$. Hence all x -shift terms are bounded by $C_{h,\beta}(i + 1)^2/N$. Combining this with (42) and (70) proves the lemma. $\square$

C.4 Comparison with the pooled many-server queue and boundary exclusion

Proof of Proposition 12. Couple the two systems using common arrivals and a rate- N Poisson process of potential service events. At each potential service event, draw a uniform mark U on $[0, N]$. The pooled system completes a service when $U < \min(C^N, N)$, whereas the JSQ system completes a service when $U < N - I^N$. Conditional on a JSQ departure, use a second independent mark to select a busy queue uniformly. Each busy JSQ queue then receives service at rate one.

Suppose that $C^N = S^N$ immediately before a potential service event. Since the number of busy JSQ queues cannot exceed either S^N or N , we have $N - I^N \leq \min(S^N, N) = \min(C^N, N)$. Thus, whenever the JSQ system departs at equality, the pooled system also departs. When $C^N < S^N$, the two populations differ by at least one, and their unit jumps cannot reverse the order. Common arrivals also preserve the order. Hence

$$C^N(0) \leq S^N(0) \implies C^N(t) \leq S^N(t) \quad \text{for all } t \geq 0.$$

Starting both systems empty and letting $t \rightarrow \infty$ gives (49).

It remains to establish (50). Set $\rho_N = (N - \beta)/N = 1 - \beta/N$. The stationary birth-death probabilities of the pooled $M/M/N$ system satisfy

$$\mathbb{P}(C^N(\infty) = N + j) = \mathbb{P}(C^N(\infty) = N) \rho_N^j, \quad j \geq 0.$$

Therefore, conditional on $C^N(\infty) \geq N$, the excess $C^N(\infty) - N$ is geometric on $\mathbb{Z}_+$ with success probability β/N . For every $x \geq 0$,

$$\mathbb{P}\left(\frac{C^N(\infty) - N}{N} > x \mid C^N(\infty) \geq N\right) = \rho_N^{\lfloor Nx \rfloor + 1} \longrightarrow e^{-\beta x}.$$

Below level N , the stationary probabilities are proportional to those of a Poisson random variable with mean $N - \beta$. Hence, relative to the stationary mass at N , the total mass below N is

$$\frac{\mathbb{P}(\text{Pois}(N - \beta) \leq N - 1)}{\mathbb{P}(\text{Pois}(N - \beta) = N)} = O(\sqrt{N}),$$

where the estimate follows from Stirling's formula. By contrast, the stationary mass at and above N , again relative to the mass at N , is $\sum_{j \geq 0} \rho_N^j = 1/(1 - \rho_N) = N/\beta$. It follows that

$$\mathbb{P}(C^N(\infty) < N) = O(N^{-1/2}) \longrightarrow 0.$$

Combining this with the conditional geometric limit gives $(C^N(\infty) - N)/N \Rightarrow E_\beta$, and hence (50).

Now let $\nu_{N_k} \Rightarrow \nu$, and let $X_\star$ have law ν . For every bounded increasing continuous function φ , (49) gives

$$\mathbb{E}\varphi\left(\frac{C^{N_k}(\infty)}{N_k}\right) \leq \mathbb{E}\varphi\left(\frac{S^{N_k}(\infty)}{N_k}\right).$$

Using (50) and $\nu_{N_k} \Rightarrow \nu$ on the two sides yields

$$\mathbb{E}\varphi(1 + E_\beta) \leq \mathbb{E}\varphi(X_\star).$$

Bounded increasing continuous functions characterize stochastic order, so $1 + E_\beta \leq_{\text{st}} X_\star$. Consequently, for every $\eta > 0$,

$$\mathbb{P}(X_\star = 1) \leq \mathbb{P}(X_\star < 1 + \eta) \leq \mathbb{P}(1 + E_\beta < 1 + \eta) = 1 - e^{-\beta\eta}.$$

Letting $\eta \downarrow 0$ yields $\nu(\{1\}) = \mathbb{P}(X_\star = 1) = 0$. $\square$

C.5 Distributional probability flux and density

For a distribution T , $\langle T, \varphi \rangle$ denotes its action on a test function and DT its distributional derivative. For an interval U , $W^{1,1}(U)$ is the Sobolev space of $L^1(U)$ functions with weak derivative in $L^1(U)$. We use “a.e.” for Lebesgue-almost everywhere, $K \Subset U$ for compact containment, $|K|$ for Lebesgue length, and $\|b\|_{\infty, K}$ for the essential supremum of $|b|$ on K . Convolution with functions or locally finite measures is denoted by $*$.

Proof of Lemma 6. Regard μ as a distribution and set $T = D\mu - b\mu$. Equation (52) gives $\langle T, f' \rangle = 0$. Because U is an interval, every $\varphi \in C_c^\infty(U)$ with zero integral has a compactly supported smooth primitive in U . Thus T annihilates every such φ . Choose $\psi \in C_c^\infty(U)$ with $\int_U \psi = 1$. For arbitrary $\varphi \in C_c^\infty(U)$, the function $\varphi - (\int_U \varphi)\psi$ has zero integral, and hence

$$\langle T, \varphi \rangle = c \int_U \varphi(x) dx, \quad c := \langle T, \psi \rangle.$$

Therefore $T = c dx$, or equivalently

$$D\mu = b\mu + c dx. \tag{71}$$

To identify a density, fix nondegenerate compact intervals $K \Subset K' \Subset U$. Choose a nonnegative $\rho \in C_c^\infty((-1, 1))$ with $\int_{\mathbb{R}} \rho = 1$, set $\rho_\varepsilon(x) = \varepsilon^{-1} \rho(x/\varepsilon)$, and take $0 < \varepsilon < \text{dist}(K, \mathbb{R} \setminus K')$. In the following convolutions, restrict μ and $b\mu$ to K' and extend those finite measures by zero to $\mathbb{R}$. For $x \in K$, the support of $\rho_\varepsilon(x - \cdot)$ lies in K' , so the local distributional identity (71) gives, with $p_\varepsilon = \rho_\varepsilon * \mu$,

$$p'_\varepsilon = \rho_\varepsilon * (b\mu) + c \quad \text{on } K.$$

Consequently,

$$\|p_\varepsilon\|_{L^1(K)} \leq \mu(K'), \quad \|p'_\varepsilon\|_{L^1(K)} \leq \|b\|_{\infty, K'} \mu(K') + |c| |K|.$$

Thus $\{p_\varepsilon\}$ is bounded in $W^{1,1}(K^\circ)$, and hence in $BV(K^\circ)$. The one-dimensional compactness theorem for bounded sets in BV gives relative compactness in $L^1(K)$. Along a sequence $\varepsilon_n \downarrow 0$, let $p_{\varepsilon_n} \rightarrow p_K$ in $L^1(K)$. On K° , the measures $p_{\varepsilon_n}(x) dx$ converge vaguely both to $p_K(x) dx$ and, by the approximate-identity property, to μ . Hence $\mu = p_K dx$ on K° .

Applying the same argument along an exhaustion of U by compact intervals gives local densities. On overlaps they agree almost everywhere because they represent the same measure, so they patch to a single $p \in L^1_{\text{loc}}(U)$ with $\mu = p dx$. Substitution in (71) gives

$$Dp = (bp + c) dx$$

locally on U . Since the right-hand side is locally integrable, p has a locally absolutely continuous representative and satisfies (53) almost everywhere. $\square$

D Technical proofs for the abandonment results

The proof is organized around five structural steps: a finite-system order construction, finite-horizon control of the occupancy profile, stopped diffusion convergence and localization, stationary stochastic averaging, and boundary identification. Each rate affected by abandonment is retained explicitly, while results from the preceding analysis are invoked only after their hypotheses have been verified under the abandonment dynamics.

D.1 Order-preserving construction and finite-system consequences

Lemma 16 (Order preservation by the transition maps). *Let $x, y \in \mathbb{Z}_+^N$ be nondecreasing and satisfy $x_i \leq y_i$ for every i . Define*

$$A(x) = (x + e_1)^\uparrow, \quad D_k(x) = (x - \mathbb{1}_{\{x_k > 0\}} e_k)^\uparrow, \quad 1 \leq k \leq N.$$

Then

$$A(x) \preceq A(y), \quad D_k(x) \preceq D_k(y), \quad 1 \leq k \leq N.$$

Moreover, decreasing any single coordinate of x by one and resorting produces a vector dominated by y .

Proof. If $u_i \leq v_i$ for every i , then $u_i^\uparrow \leq v_i^\uparrow$ for every i ; this follows, for example, by comparing the number of coordinates exceeding each integer threshold. Since $x + e_1 \leq y + e_1$ coordinatewise, the assertion for A follows after sorting.

For the service map, put

$$\tilde{x} = x - \mathbb{1}_{\{x_k > 0\}} e_k, \quad \tilde{y} = y - \mathbb{1}_{\{y_k > 0\}} e_k.$$

For coordinates other than k , the original order is unchanged. At coordinate k , if $x_k = 0$ then $\tilde{x}_k = 0 \leq \tilde{y}_k$; if $x_k > 0$, then $y_k > 0$ and $x_k - 1 \leq y_k - 1$. Thus $\tilde{x} \leq \tilde{y}$ coordinatewise, and sorting proves the claim. Finally, decreasing a coordinate of x gives a vector coordinatewise dominated by x , before sorting, and hence dominated by y after sorting. $\square$

Proof of Proposition 3. We first construct the two sorted processes on one probability space. Arrivals are generated by a common Poisson process of rate $N - \beta$ and apply the map A to both systems. Potential-service marks form a common Poisson process of rate N ; each mark carries an independent rank, uniform on $\{1, \dots, N\}$, and applies D_k to both systems. This is the correct sorted-state representation because choosing a uniform rank is equivalent to choosing a uniform labeled server, and, when several coordinates are tied, the sorted post-transition vector does not depend on which tied label is selected.

The abandonment system has additional state-dependent marks. At a sorted state q , rank k is decremented and the vector is resorted at rate

$$\frac{\gamma}{N} (q_k - 1)^+.$$

These rates can be realized by independent Poisson random measures with state-dependent thinning. Their sum is $\gamma W(q)/N$, exactly the additional rate in (9). An abandonment acts only on the abandonment system and lowers one coordinate. By Lemma 16, the order $\preceq$ is preserved at every arrival, service, and abandonment mark. Independent Poisson constructions have no simultaneous effective marks almost surely; imposing any fixed order on the null event of a tie gives the same conclusion. This proves (11).

For stability, use the quadratic Lyapunov function $\mathcal{W}_2(q) = \sum_i q_i^2$ from the proof of Proposition 1. The abandonment contribution to its drift is

$$\frac{\gamma}{N} \sum_{i=1}^N (q_i - 1)^+ (-2q_i + 1) \leq 0.$$

Consequently,

$$\mathcal{L}_N^\gamma \mathcal{W}_2(q) \leq -\frac{2\beta}{N} S(q) + 2N - \beta,$$

so the same continuous-time Foster criterion gives positive recurrence. Irreducibility follows from the finite paths used in the proof of Proposition 1: each such path has positive probability when no abandonment mark occurs. To prove nonexplosion, let $N_A^N(t)$ be the arrival count. Pathwise,

$$S^{N,\gamma}(t) \leq S^{N,\gamma}(0) + N_A^N(t).$$

On every bounded time interval the right-hand side is finite, and the total effective jump rate is then bounded by $2N + \gamma S^{N,\gamma}/N$. Thus only finitely many jumps occur.

Start both systems from zero. The pathwise order and convergence of each positive recurrent irreducible chain to its unique stationary law imply, for every bounded increasing function of the sorted vector,

$$\mathbb{E}_{\pi_N^\gamma} F(Q) \leq \mathbb{E}_{\pi_N} F(Q).$$

A monotone-class argument gives stationary stochastic domination. $\square$

The localization argument in Lemma 2 applies to $\mathcal{L}_N^\gamma$. Indeed, stationary domination by π_N gives all fixed- N polynomial moments needed for dominated convergence. We will use the resulting identity

$$\mathbb{E}_{\pi_N^\gamma} (\mathcal{L}_N^\gamma F)(Q^{N,\gamma}) = 0 \tag{72}$$

for every function F satisfying the polynomial-growth condition of Lemma 2.

Lemma 17 (Inherited stationary upper bounds). *For every a satisfying (5), there are N_0 and C_a such that*

$$\sup_{N \geq N_0} \mathbb{E}_{\pi_N^\gamma} e^{aX_N^\gamma} \leq C_a. \tag{73}$$

Moreover, for every fixed $0 < \theta_0 < \theta_\perp$, there is $C < \infty$ such that

$$\mathbb{P}_{\pi_N^\gamma}(Q_{\max}^{N,\gamma} > u) \leq C e^{-au/2} + \kappa_\perp N e^{-\theta_0 u/2}, \quad u \geq 0, \tag{74}$$

and hence

$$\mathbb{E}_{\pi_N^\gamma} (Q_{\max}^{N,\gamma})^2 = O((\log N)^2). \tag{75}$$

Proof. The maps $q \mapsto S(q)$ and $q \mapsto Q_{\max}(q)$ are increasing with respect to $\preceq$. Stationary domination and Theorem 2 give (73). They also give

$$\mathbb{P}(Q_{\max}^{N,\gamma} > u) \leq \mathbb{P}(Q_{\max}^N > u).$$

Since $Q_{\max}^N \leq X_N + D_N$, (7) and (17) yield (74). Integrating the tail, split at a constant multiple of $\log N$; below the split use the bound one and above it use (74). This proves (75). $\square$

D.2 Localized occupancy-profile estimates

Proof of Lemma 7. Service completions from one- and two-job queues increase V by one, at total rate $M_1 + M_2$. An abandonment changes V only when it occurs from a two-job queue; it then changes that queue to level one and increases V by one. Such abandonments occur at rate $\gamma M_2/N$. Arrivals decrease V by one whenever $V > 0$, at rate $N - \beta$. These observations prove (54)–(55). An increase of Z_3 requires an arrival to a two-job queue, which under JSQ is possible only when every queue has at least two jobs, equivalently $V = 0$. Finally,

$$M_1 + M_2 = N - I - Z_3 \leq N - K, \quad \frac{\gamma}{N} M_2 \leq \gamma,$$

on $\{Z_3 \geq K\}$, proving (56). $\square$

Proof of Lemma 8. A transition $I : 0 \rightarrow 1$ is exactly a service completion from a one-job queue, giving (57). When $V = 0$, one has $I = M_1 = 0$. A transition $V : 0 \rightarrow 1$ is then either a service completion from a two-job queue or an abandonment from such a queue; both leave Z_3 unchanged. This gives (58). The first intensity is at most N and the second at most $N + \gamma$. Standard thinning of homogeneous Poisson processes gives the asserted dominations and hence (59). In particular, for a constant $c_T < \infty$,

$$\mathbb{P}(C_I^N(NT) + C_{V,K}^N(NT) > c_T N^2) \leq 2e^{-c'_T N^2}$$

for all sufficiently large N and a suitable $c'_T > 0$. $\square$

Lemma 18 (Coarse finite-horizon idleness bound). *For every $\varepsilon \in (0, 1/2)$ there are constants $c, C > 0$, depending on T, β , such that*

$$\mathbb{P}\left(\sup_{0 \leq s \leq NT} I^{N,\gamma}(s) \geq N^{1/2+\varepsilon}\right) \leq CN^2 e^{-cN^{2\varepsilon}} + o(1).$$

Proof. At level i , the upward rate of $I^{N,\gamma}$ is $M_1^{N,\gamma} \leq N - i$, while the downward rate is $N - \beta$ whenever $i > 0$. Thus every excursion of $I^{N,\gamma}$ away from zero is dominated by the birth–death chain with birth rate $N - i$ at level i and death rate $N - \beta$.

Let $L < N$ and start that comparison chain from one. The standard birth–death hitting formula, bounded by its last denominator term, gives

$$\mathbb{P}_1(T_L < T_0) \leq \prod_{j=1}^{L-1} \frac{N-j}{N-\beta}.$$

Let $j_0 = \lceil 2\beta \rceil + 1$. The finitely many factors below j_0 are bounded by a constant depending only on β , and for $j \geq j_0$,

$$\log \frac{N-j}{N-\beta} = \log \left(1 - \frac{j-\beta}{N-\beta} \right) \leq -\frac{j-\beta}{N-\beta}.$$

Therefore, for $L \geq 2j_0$ and all large N ,

$$\mathbb{P}_1(T_L < T_0) \leq C_\beta \exp \left\{ -c_\beta \frac{L^2}{N} \right\}. \quad (76)$$

By Lemma 8, the number of excursions begun by time NT is dominated by a Poisson random variable of mean N^2T . A union bound using (76), with $L = \lceil N^{1/2+\varepsilon} \rceil$, proves the displayed estimate apart from the initial state. Balanced initialization and $\hat{X}^{N,\gamma}(0) \Rightarrow X(0) > 1$ imply

$$\mathbb{P}(I^{N,\gamma}(0) > 0) \leq \mathbb{P}(\hat{X}^{N,\gamma}(0) < 1) \longrightarrow 0,$$

which is the final $o(1)$ term. □

Proof of Proposition 14. Choose $0 < \varepsilon < 1/8$ and set

$$K_N = \lceil N^{1/2+2\varepsilon} \rceil, \quad L_N = \lceil N^{1/2+\varepsilon} \rceil.$$

Consider a $V^{N,\gamma}$ -excursion that starts from zero while $Z_3^{N,\gamma} \geq K_N$, and stop it when V returns to zero or Z_3 falls below K_N . By Lemma 7, while the excursion is active its upward rate is at most $N - K_N + \gamma$ and its downward rate is $N - \beta$. Hence Lemma 15(c) gives, uniformly over its post-jump starting state,

$$\mathbb{P}(V^{N,\gamma} \text{ reaches } L_N) \leq \left(\frac{N - K_N + \gamma}{N - \beta} \right)^{L_N - 1} \leq e^{-cN^{3\varepsilon}}.$$

The compensator bound in Lemma 8, followed by the strong Markov property at each start, gives

$$\mathbb{P}(\text{some marked excursion reaches } L_N \text{ by } NT) \leq N(N + \gamma)Te^{-cN^{3\varepsilon}}.$$

Under balanced initialization, time zero contributes no uncontrolled high- Z_3 deficit excursion: if a balanced initial state has $Z_3 \geq K_N$, then every queue has at least two jobs and therefore $V = 0$. Combining this observation with Lemma 18 shows that, with probability tending to one,

$$\begin{aligned} I^{N,\gamma}(s) &< L_N, \\ Z_3^{N,\gamma}(s) &\geq K_N \implies V^{N,\gamma}(s) < L_N, \quad 0 \leq s \leq NT. \end{aligned} \quad (77)$$

We next control the aggregate number of jobs beyond level two. Under JSQ, an arrival can increase the current maximum queue length only if all queues have that common maximum. Service completions and abandonments never increase the maximum. Consequently, on every sample path,

$$Q_{\max}^{N,\gamma}(s) \leq \max \left\{ Q_{\max}^{N,\gamma}(0), 1 + \sup_{0 \leq u \leq s} \frac{S^{N,\gamma}(u)}{N} \right\}. \quad (78)$$

On $\{\tau_N^{\eta,R} > 0\}$, balanced initialization gives $Q_{\max}^{N,\gamma}(0) = \lceil \widehat{X}^{N,\gamma}(0) \rceil \leq R + 1$. Together with (78) and the possible $1/N$ overshoot at the exit time, this yields

$$\sup_{t \leq \tau_N^{\eta,R}} Q_{\max}^{N,\gamma}(Nt) \leq R + 2. \quad (79)$$

If $\tau_N^{\eta,R} = 0$ and $\widehat{X}^{N,\gamma}(0) \geq 1$, the balanced initial vector has only two adjacent occupied levels and satisfies $M_1^{N,\gamma}(0) = Nr(\widehat{X}^{N,\gamma}(0))$ exactly. The remaining event $\{\widehat{X}^{N,\gamma}(0) < 1\}$ has probability tending to zero. It therefore remains to work on $\{\tau_N^{\eta,R} > 0\}$.

Let

$$H^{N,\gamma} = \sum_{i=1}^N (Q_i^{N,\gamma} - 2)^+.$$

The deterministic identity (37) is model-independent. On $\{Z_3^{N,\gamma} < K_N\}$ and $\{\tau_N^{\eta,R} > 0\}$, (79) gives

$$H^{N,\gamma} \leq (R + 2)K_N,$$

and hence, on the event (77),

$$\left| \frac{M_1^{N,\gamma}}{N} - r \left(\frac{S^{N,\gamma}}{N} \right) \right| \leq \frac{(R + 2)K_N + 2L_N}{N}.$$

On $\{Z_3^{N,\gamma} \geq K_N\}$, both $M_1^{N,\gamma}$ and

$$Nr(S^{N,\gamma}/N) = (V^{N,\gamma} - H^{N,\gamma})^+$$

lie in $[0, V^{N,\gamma}]$, so their normalized difference is at most L_N/N on the same good event. Since $K_N/N \rightarrow 0$ and $L_N/N \rightarrow 0$, (60) follows. $\square$

Proof of Proposition 15. On $[0, \tau_N^{\eta,R}]$,

$$r(\widehat{X}^{N,\gamma}) \leq 1 - \eta.$$

Define the failure time

$$\zeta_N = \inf \left\{ t \in [0, \tau_N^{\eta,R}] : M_1^{N,\gamma}(Nt) > N(1 - \eta/2) \right\},$$

with the convention $\inf \emptyset = \tau_N^{\eta,R}$. By Proposition 14, $\mathbb{P}(\zeta_N < \tau_N^{\eta,R}) \rightarrow 0$. Before ζ_N , the birth rate of the idleness process is at most $N(1 - \eta/2)$ and the idleness-death rate is $N - \beta$ whenever idleness is positive. Hence every excursion begun before $N\zeta_N$, stopped at $N\zeta_N$ if necessary, is dominated by a birth–death chain whose probability of reaching level L is at most

$$\rho_N^{L-1}, \quad \rho_N = \frac{N(1 - \eta/2)}{N - \beta} \leq 1 - \eta/4$$

for all sufficiently large N . The compensator estimate in Lemma 8 and a union bound therefore show that, for a sufficiently large constant c_η ,

$$\mathbb{P} \left(\sup_{t \leq \tau_N^{\eta,R}} I^{N,\gamma}(Nt) > c_\eta \log N \right) \leq \mathbb{P}(\zeta_N < \tau_N^{\eta,R}) + \mathbb{P}(I^{N,\gamma}(0) > 0) + o(1).$$

The second term charges the entire idleness excursion already active at time zero; excursions beginning later are covered by the compensator union bound. Balanced initialization implies that $I^{N,\gamma}(0) > 0$ only when $\widehat{X}^{N,\gamma}(0) < 1$. The latter event has probability tending to zero, so the right-hand side vanishes. This proves (61).

The proof of Proposition 14 gives the quantitative bound

$$\sup_{t \leq \tau_N^{\eta,R}} \left| \frac{M_1^{N,\gamma}(Nt)}{N} - r(\widehat{X}^{N,\gamma}(t)) \right| = O_{\mathbb{P}}(N^{-1/2+2\varepsilon})$$

for any fixed $\varepsilon \in (0, 1/8)$. Multiplication by the logarithmic bound proves (62). $\square$

Proof of Proposition 16. Choose $h \in C_c^1((1, \infty))$ equal to one on $[1 + \eta, R]$, and let a_h , Ψ_N , and ψ_N be the corrector functions defined in Section 5.3. For $x = S(q)/N$, $i = I(q)$, and $m = M_1^{N,\gamma}(q)$, the only term in $\mathcal{L}_N^\gamma \psi_N$ not present in Lemma 5 is

$$\frac{\gamma W(q)}{N} \left[\Psi_N \left(x - \frac{1}{N}, i \right) - \Psi_N(x, i) \right]. \quad (80)$$

On the stopped set, $W(q) \leq S(q) \leq RN$, and the Lipschitz property of a_h gives

$$\left| \Psi_N \left(x - \frac{1}{N}, i \right) - \Psi_N(x, i) \right| \leq \frac{C_h}{N^2} (i + 1)^2.$$

Thus Lemma 5 becomes, on the stopped set,

$$|(\mathcal{L}_N^\gamma \psi_N)(q) - h(x)(i - g(x))| \leq C_{h,\beta,\gamma,R} \left[(i + 1) \left| \frac{m}{N} - r(x) \right| + \frac{(i + 1)^2}{N} + \frac{(i + 1)^2}{N^2} \right]. \quad (81)$$

The last term is displayed separately to identify the abandonment contribution, although it is dominated by the preceding term.

Let

$$M_\psi^{N,\gamma}(u) = \psi_N(Q^{N,\gamma}(u)) - \psi_N(Q^{N,\gamma}(0)) - \int_0^u (\mathcal{L}_N^\gamma \psi_N)(Q^{N,\gamma}(s)) ds.$$

Dynkin's formula at $N(t \wedge \tau_N^{\eta,R})$, divided by N , expresses the integral in (63) as the sum of a corrector boundary term, $-M_\psi^{N,\gamma}/N$, and the integrated error in (81). By (61),

$$\frac{1}{N} \sup_{t \leq \tau_N^{\eta,R}} |\psi_N(Q^{N,\gamma}(Nt))| \leq \frac{C_h}{N^2} \left(1 + \sup_{t \leq \tau_N^{\eta,R}} I^{N,\gamma}(Nt) \right)^2 \xrightarrow{\mathbb{P}} 0.$$

The integrated error vanishes uniformly by (62) and (61).

It remains to control the corrector martingale. Let c_η be the constant in (61), put $L_N^I = \lceil c_\eta \log N \rceil$, and introduce the genuine accelerated-time stopping time

$$\kappa_N = \inf\{t \in [0, \tau_N^{\eta,R}] : I^{N,\gamma}(Nt) \geq L_N^I\} \wedge \tau_N^{\eta,R}.$$

Up to κ_N , the jump sizes and original-time rates satisfy the deterministic bounds

transition type	total rate	$ \Delta\psi_N $
arrival with $I > 0$ or service from a one-job queue	$\leq 2N$	$\leq CL_N^I/N$
other arrival or service transition	$\leq 2N$	$\leq C(L_N^I)^2/N^2$
abandonment	$\leq \gamma R$	$\leq C(L_N^I)^2/N^2$

Consequently,

$$\begin{aligned} \langle M_\psi^{N,\gamma} \rangle (N(T \wedge \kappa_N)) &\leq C_T \left[(L_N^I)^2 + \frac{(L_N^I)^4}{N^2} + \frac{\gamma R (L_N^I)^4}{N^3} \right], \\ \frac{1}{N^2} \langle M_\psi^{N,\gamma} \rangle (N(T \wedge \kappa_N)) &\longrightarrow 0. \end{aligned}$$

Doob's inequality and the deterministic bracket bound imply

$$\sup_{t \leq T} \frac{1}{N} \left| M_\psi^{N,\gamma} (N(t \wedge \kappa_N)) \right| \xrightarrow{\mathbb{P}} 0.$$

Moreover,

$$\mathbb{P}(\kappa_N < \tau_N^{\eta,R}) \leq \mathbb{P} \left(\sup_{t \leq \tau_N^{\eta,R}} I^{N,\gamma}(Nt) \geq L_N^I \right) \longrightarrow 0$$

by (61). Removing this adapted localization gives

$$\sup_{t \leq T} \frac{1}{N} \left| M_\psi^{N,\gamma} (N(t \wedge \tau_N^{\eta,R})) \right| \xrightarrow{\mathbb{P}} 0,$$

which completes the proof. $\square$

D.3 Stopped process limits, compact containment, and localization

Proof of Lemma 9. Choose $\delta > 0$ so that $[\ell - \delta, u + \delta]$ lies in both $(1, \infty)$ and the neighborhood on which the stopped SDE holds. Let ρ be the first exit of Y from the enlarged interval, and define the globally bounded Lipschitz extension $\bar{b}(x) = b_\gamma(\Pi x)$, where Π is projection onto $[\ell - \delta, u + \delta]$. Let $\bar{Y}$ solve $d\bar{Y} = \bar{b}(\bar{Y}) dt + \sqrt{2} dB$ from $Y(0)$. Local pathwise uniqueness makes Y and $\bar{Y}$ agree until either leaves the enlarged interval; continuity then identifies the two exit times, so $\bar{Y}(\cdot \wedge \rho) = Y(\cdot \wedge \rho)$.

Set $\theta(t) = \bar{b}(\bar{Y}(t))/\sqrt{2}$. Novikov's condition holds because, with $K = \sup_x |\bar{b}(x)|$,

$$\mathbb{E} \exp \left\{ \frac{1}{2} \int_0^T \theta(s)^2 ds \right\} \leq e^{K^2 T/4}.$$

Under the equivalent measure generated by $\mathcal{E}(-\int_0^\cdot \theta(s) dB(s))_T$, $\tilde{B}(t) = B(t) + \int_0^t \theta(s) ds$ is Brownian in the same filtration and $\bar{Y}(t) = Y(0) + \sqrt{2}\tilde{B}(t)$. Thus $Y(\cdot \wedge \rho)$ is Brownian motion stopped on leaving the enlarged interval under an equivalent measure.

If $Y(0) \in (\ell, u)$ and $\tau = \tau_{\ell,u}(Y) < T$, then $\tau < \rho$ is a stopping time and $Y(\tau) \in \{\ell, u\}$. The strong Markov property at τ and regularity of zero for both Brownian half-lines give outward crossings at arbitrarily small positive times after τ . The same argument at time zero applies when

$Y(0)$ is an endpoint. Equivalence of measures transfers these properties to the original law; no assertion is made at other boundary-hit times.

For continuity, let $z_n \rightarrow z$ uniformly and write $\tau_n = \tau_{\ell,u}(z_n)$ and $\tau = \tau_{\ell,u}(z)$. If $z(0) \notin [\ell, u]$, then $\tau_n = 0$ eventually. If $z(0)$ is an endpoint, time-zero crossing gives $\limsup_n \tau_n \leq \varepsilon$ for every $\varepsilon > 0$. If $z(0) \in (\ell, u)$, compactness of $z([0, t])$ for $t < \tau$ gives $\liminf_n \tau_n \geq \tau$; when $\tau < T$, first-exit crossing gives $\limsup_n \tau_n \leq \tau + \varepsilon$ for every $\varepsilon > 0$, while for $\tau = T$ the lower bound and $\tau_n \leq T$ suffice. Hence $\tau_n \rightarrow \tau$. $\square$

Lemma 19 (Joint tightness and marginal martingale limit). *Under the assumptions of Theorem 5, fix $T < \infty$, $0 < \eta < 1$, and $R > 2$, and let $\overline{\mathcal{M}}^{N,\gamma}$ be the extension of the stopped population martingale defined below. Then*

$$(\widehat{X}^{N,\gamma}(0), \widehat{X}_{\eta,R}^{N,\gamma}, \overline{\mathcal{M}}^{N,\gamma}, \tau_N^{\eta,R})$$

is tight in $\mathbb{R} \times D[0, T] \times D[0, T] \times [0, T]$. The two process coordinates are C -tight, and

$$\overline{\mathcal{M}}^{N,\gamma} \Rightarrow \sqrt{2}B \quad \text{in } D[0, T]$$

marginally, where B is standard Brownian motion.

Proof. A random-time-change representation gives

$$\widehat{X}^{N,\gamma}(t) = \widehat{X}^{N,\gamma}(0) + \mathcal{M}^{N,\gamma}(t) + \int_0^t \left[I^{N,\gamma}(Ns) - \beta - \gamma \frac{W^{N,\gamma}(Ns)}{N} \right] ds, \quad (82)$$

where $\mathcal{M}^{N,\gamma}$ is a square-integrable local martingale with jumps of absolute size $1/N$ and predictable quadratic variation

$$\langle \mathcal{M}^{N,\gamma} \rangle(t) = \int_0^t \left[\frac{N - \beta}{N} + \frac{N - I^{N,\gamma}(Ns)}{N} + \gamma \frac{W^{N,\gamma}(Ns)}{N^2} \right] ds. \quad (83)$$

On the stopped interval,

$$\frac{W^{N,\gamma}}{N} = \widehat{X}^{N,\gamma} - 1 + \frac{I^{N,\gamma}}{N},$$

and therefore (82) can be written as

$$\widehat{X}_{\eta,R}^{N,\gamma}(t) = \widehat{X}^{N,\gamma}(0) + \mathcal{M}^{N,\gamma}(t \wedge \tau_N^{\eta,R}) + \int_0^{t \wedge \tau_N^{\eta,R}} b_\gamma(\widehat{X}^{N,\gamma}(s)) ds + \mathcal{E}_N^{\eta,R}(t), \quad (84)$$

where

$$\mathcal{E}_N^{\eta,R}(t) = \int_0^{t \wedge \tau_N^{\eta,R}} [I^{N,\gamma}(Ns) - g(\widehat{X}^{N,\gamma}(s))] ds - \frac{\gamma}{N} \int_0^{t \wedge \tau_N^{\eta,R}} I^{N,\gamma}(Ns) ds.$$

By Propositions 15 and 16,

$$\sup_{t \leq T} |\mathcal{E}_N^{\eta,R}(t)| \xrightarrow{\mathbb{P}} 0. \quad (85)$$

We next verify the martingale functional central limit theorem without assuming convergence of the random exit times. Enlarge the probability space by an independent standard Brownian motion $\tilde{B}^N$ and extend the stopped martingale after $\tau_N^{\eta,R}$ by setting

$$\overline{\mathcal{M}}^{N,\gamma}(t) = \mathcal{M}^{N,\gamma}(t \wedge \tau_N^{\eta,R}) + \sqrt{2} \int_0^t \mathbf{1}_{\{s > \tau_N^{\eta,R}\}} d\tilde{B}^N(s).$$

Its predictable quadratic variation satisfies, because $W^{N,\gamma} \leq RN$ before the stop,

$$\sup_{t \leq T} \left| \left\langle \overline{\mathcal{M}}^{N,\gamma} \right\rangle(t) - 2t \right| \leq T \left[\frac{\beta}{N} + \frac{1}{N} \sup_{s \leq \tau_N^{\eta,R}} I^{N,\gamma}(Ns) + \frac{\gamma R}{N} \right] \xrightarrow{\mathbb{P}} 0.$$

The abandonment term is the final term in brackets and is therefore asymptotically negligible. Moreover,

$$\sup_{t \leq T} |\Delta \overline{\mathcal{M}}^{N,\gamma}(t)| \leq \frac{1}{N} \longrightarrow 0, \quad (86)$$

so the conditional Lindeberg condition holds. The martingale FCLT (Whitt, 2007, Theorem 2.1) therefore gives only the marginal convergence

$$\overline{\mathcal{M}}^{N,\gamma} \Rightarrow \sqrt{2}B \quad \text{in } D[0, T].$$

The stopped population is also C -tight. Indeed, if $\sigma_N \leq T$ are stopping times and $\delta_N \downarrow 0$, optional sampling and (83) give

$$\begin{aligned} \mathbb{P} \left(|\mathcal{M}^{N,\gamma}((\sigma_N + \delta_N) \wedge \tau_N^{\eta,R}) - \mathcal{M}^{N,\gamma}(\sigma_N \wedge \tau_N^{\eta,R})| > \varepsilon \right) \\ \leq \frac{C_{R,\gamma} \delta_N + o(1)}{\varepsilon^2}. \end{aligned}$$

The drift b_γ is bounded on $[1 + \eta, R]$, (85) is uniformly negligible, and (86) excludes discontinuous limits. Aldous' criterion applied to (84) therefore proves C -tightness of $\hat{X}_{\eta,R}^{N,\gamma}$. The initial states are tight by their assumed convergence, and $\tau_N^{\eta,R} \in [0, T]$. Product-space tightness now proves the joint assertion. $\square$

Lemma 20 (Joint Brownian identification). *For every joint subsequential limit supplied by Lemma 19,*

$$(\hat{X}^{N,\gamma}(0), \hat{X}_{\eta,R}^{N,\gamma}, \overline{\mathcal{M}}^{N,\gamma}) \Rightarrow (X(0), X_{\eta,R}^\gamma, \mathcal{M}),$$

the process $\mathcal{M}$ is a continuous martingale with respect to the usual augmentation of the filtration generated jointly by $X(0)$, $X_{\eta,R}^\gamma$, and $\mathcal{M}$. Moreover, $\mathcal{M}/\sqrt{2}$ is a standard Brownian motion in that filtration and is independent of $X(0)$.

Proof. Let $\mathcal{G}_t^N$ be the completed right-continuous filtration generated by the initial queue vector, the queueing primitives through original time Nt , and the independent auxiliary Brownian motion through accelerated time t . Then $\overline{\mathcal{M}}^{N,\gamma}$ is a square-integrable $\mathcal{G}^N$ -martingale. By Lemma 19, the marginal law of every limit $\mathcal{M}$ is that of $\sqrt{2}$ times standard Brownian motion; this fact alone makes no claim about the joint filtration.

Fix $0 \leq s < t \leq T$ and let H be a bounded continuous cylinder function of the initial state and the values of the stopped population and extended martingale at finitely many times not exceeding s . The corresponding prelimit random variable is $\mathcal{G}_s^N$ -measurable, so

$$\mathbb{E} \left[H_N(\overline{\mathcal{M}}^{N,\gamma}(t) - \overline{\mathcal{M}}^{N,\gamma}(s)) \right] = 0.$$

The brackets of the extended martingales are bounded on $[0, T]$ by a deterministic constant depending only on T, R, γ , so the martingale increments are uniformly bounded in L^2 . The preceding joint tightness lemma makes the relevant cylinder evaluations continuity functionals at every subsequential limit. Joint weak convergence and uniform integrability therefore allow passage to the limit. A monotone-class argument shows that $\mathcal{M}$ is a martingale with respect to the raw filtration

$$\mathcal{G}_t^0 = \sigma\{X(0), X_{\eta,R}^\gamma(u), \mathcal{M}(u) : 0 \leq u \leq t\}.$$

Define its usual augmentation by

$$\mathcal{G}_t = \overline{\bigcap_{\varepsilon>0} \mathcal{G}_{(t+\varepsilon) \wedge T}^0}, \quad 0 \leq t \leq T,$$

where the overline denotes completion under the limiting law. The martingale property extends from $\mathcal{G}^0$ to $\mathcal{G}$. Indeed, for $A \in \mathcal{G}_s$, replace A by an event in the uncompleted right-continuous intersection that differs from it only by a null set. For deterministic $s_n \downarrow s$ with $s_n < t$, that event belongs to $\mathcal{G}_{s_n}^0$, so the $\mathcal{G}^0$ -martingale identity applies at s_n ; continuity of $\mathcal{M}$ gives $\mathcal{M}(s_n) \rightarrow \mathcal{M}(s)$ in L^2 .

The marginal Brownian law gives continuity and, along deterministic refining partitions, the pathwise quadratic variation $[\mathcal{M}](t) = 2t$. Lévy's characterization now implies that $\mathcal{M}/\sqrt{2}$ is Brownian with respect to $\mathcal{G}$. Since $X(0)$ is $\mathcal{G}_0$ -measurable, the Brownian increments, and hence the Brownian path, are independent of $X(0)$. $\square$

Proof of Proposition 17. By Lemma 19, take a jointly convergent subsequence of the initial state, stopped population, extended martingale, and exit time. By Lemma 20, write the martingale limit as $\sqrt{2}B$, where B is Brownian in the joint limiting filtration and is independent of $X(0)$. On a Skorokhod representation of this joint law, assume that the path components converge uniformly and that

$$\tau_N^{\eta,R} \longrightarrow \bar{\tau} \in [0, T]$$

almost surely. Continuity of the limiting martingale gives

$$\mathcal{M}^{N,\gamma}(\cdot \wedge \tau_N^{\eta,R}) = \overline{\mathcal{M}}^{N,\gamma}(\cdot \wedge \tau_N^{\eta,R}) \longrightarrow \sqrt{2}B(\cdot \wedge \bar{\tau}) \quad \text{uniformly on } [0, T]. \quad (87)$$

For every $t < \bar{\tau}$, one has $t < \tau_N^{\eta,R}$ for all sufficiently large N . Passing to the limit in (84), using (85) and (87), shows that the limiting path solves the SDE with drift b_γ on $[0, \bar{\tau})$.

Let

$$\tau = \inf\{t \leq T : X_{\eta,R}^\gamma(t) \notin (1 + \eta, R)\} \wedge T,$$

with $\tau = 0$ when $X(0) \notin (1 + \eta, R)$. We identify $\bar{\tau}$ by cases. If $X(0)$ lies strictly outside the closed interval, uniform convergence of the initial states gives $\tau_N^{\eta,R} = 0$ eventually and hence $\bar{\tau} = \tau = 0$. If

$X(0)$ is an endpoint and $\bar{\tau} > 0$, the limiting path solves the SDE immediately after time zero; the time-zero clause of Lemma 9 then gives an exterior point before $\bar{\tau}$, contradicting uniform convergence of prelimit paths that remain in the open interval before their exit. Thus $\bar{\tau} = \tau = 0$ also in this case.

Suppose now that $X(0) \in (1 + \eta, R)$. Then $\tau_N^{\eta, R} > 0$ for all sufficiently large N . If $\bar{\tau} < T$, the prelimit exit value is outside $(1 + \eta, R)$ with overshoot at most $1/N$; uniform convergence gives $X_{\eta, R}^\gamma(\bar{\tau}) \in \{1 + \eta, R\}$ and hence $\tau \leq \bar{\tau}$. If $\tau < \bar{\tau}$, the first-exit clause of Lemma 9 gives a time $t \in (\tau, \bar{\tau})$ with $X_{\eta, R}^\gamma(t) \notin [1 + \eta, R]$. For all sufficiently large N , one has $t < \tau_N^{\eta, R}$ and the prelimit path lies in $(1 + \eta, R)$, contradicting uniform convergence. Therefore $\bar{\tau} = \tau$. This also covers the capped case $\tau = T$.

Every prelimit stopped path is constant after $\tau_N^{\eta, R}$, so the limit is constant after τ . Together with (87), this proves the absorbed SDE (64). Pathwise uniqueness on compact subintervals of $(1, \infty)$ implies uniqueness in law and convergence of the entire stopped sequence. $\square$

Proof of Proposition 4. The drift b_γ is locally Lipschitz on $(1, \infty)$. With normalization at two, its scale density is

$$s'_\gamma(x) = \begin{cases} \frac{1}{x-1} \exp\left\{(1+\beta)(x-2) + \frac{\gamma}{2}((x-1)^2 - 1)\right\}, & 1 < x \leq 2, \\ \exp\left\{\beta(x-2) + \frac{\gamma}{2}((x-1)^2 - 1)\right\}, & x \geq 2. \end{cases}$$

Both scale integrals diverge, while $m_\gamma = 1/s'_\gamma$ is integrable because it is of order $x-1$ near one and has Gaussian-type decay at infinity. The nondegenerate constant diffusion coefficient gives a strictly positive transition density on the interior. Hence Lemma 12 gives nonexplosion, positive recurrence, uniqueness, and total-variation ergodicity. Normalizing m_γ gives exactly (14). $\square$

Proof of Lemma 10. For the upper boundary, couple $Q^{N, \gamma}$ with the system $Q^{N, 0}$ without abandonment, started from the same balanced initial vector. By Proposition 3,

$$\widehat{X}^{N, \gamma}(t) \leq \widehat{X}^{N, 0}(t), \quad t \geq 0,$$

almost surely. The Gupta–Walton theorem Theorem 1 implies tightness of $\sup_{t \leq T} \widehat{X}^{N, 0}(t)$, and therefore

$$\lim_{R \uparrow \infty} \limsup_{N \rightarrow \infty} \mathbb{P} \left(\sup_{t \leq T} \widehat{X}^{N, \gamma}(t) \geq R \right) = 0. \quad (88)$$

Fix $\eta > 0$ and $R > 2$. We use the enlarged localization interval $(1 + \eta, 2R)$ so that, on the event $\sup_{t \leq T} \widehat{X}^{N, \gamma}(t) < R$, its upper boundary cannot trigger the stop; only the lower-boundary event is relevant. On the event

$$\left\{ \inf_{t \leq T} \widehat{X}^{N, \gamma}(t) \leq 1 + 2\eta, \sup_{t \leq T} \widehat{X}^{N, \gamma}(t) < R \right\},$$

the process stopped on leaving $(1 + \eta, 2R)$ has infimum at most $1 + 2\eta$; the jump size $1/N$ makes this implication immediate for all large N . By Proposition 17 and the portmanteau theorem applied to the closed set

$$\left\{ z \in C[0, T] : \inf_{t \leq T} z(t) \leq 1 + 2\eta \right\},$$

we obtain

$$\limsup_{N \rightarrow \infty} \mathbb{P} \left(\inf_{t \leq T} \widehat{X}^{N,\gamma}(t) \leq 1 + 2\eta, \sup_{t \leq T} \widehat{X}^{N,\gamma}(t) < R \right) \leq \mathbb{P} \left(\inf_{t \leq T} X^\gamma(t) \leq 1 + 2\eta \right).$$

Here the stopped limit is identified, by pathwise uniqueness, with the global solution X^γ stopped on leaving $(1 + \eta, 2R)$. By Proposition 4, X^γ has continuous paths and never reaches one. Hence its minimum on $[0, T]$ is strictly larger than one almost surely, and the probability on the right tends to zero as $\eta \downarrow 0$. Combining this fact with (88), in the order $N \rightarrow \infty$, then $R \uparrow \infty$, then $\eta \downarrow 0$, proves (65).

It remains to verify the oscillation condition for the full sequence. For $z \in D[0, T]$, let

$$\omega_T(z, \delta) = \sup\{|z(t) - z(s)| : 0 \leq s, t \leq T, |t - s| \leq \delta\}.$$

Let $\tilde{\tau}_N^{\eta, R}$ be the first exit from $(1 + \eta, 2R)$, capped at T . On the event

$$\left\{ \inf_{t \leq T} \widehat{X}^{N,\gamma}(t) > 1 + 2\eta, \sup_{t \leq T} \widehat{X}^{N,\gamma}(t) < R \right\},$$

the full path agrees on $[0, T]$ with the path stopped at $\tilde{\tau}_N^{\eta, R}$. Hence, for every $\varepsilon > 0$,

$$\begin{aligned} \mathbb{P}(\omega_T(\widehat{X}^{N,\gamma}, \delta) > \varepsilon) &\leq \mathbb{P} \left(\inf_{t \leq T} \widehat{X}^{N,\gamma}(t) \leq 1 + 2\eta \text{ or } \sup_{t \leq T} \widehat{X}^{N,\gamma}(t) \geq R \right) \\ &\quad + \mathbb{P}(\omega_T(\widehat{X}^{N,\gamma}(\cdot \wedge \tilde{\tau}_N^{\eta, R}), \delta) > \varepsilon). \end{aligned}$$

For fixed η and R , the second term vanishes after taking $\lim_{\delta \downarrow 0} \limsup_{N \rightarrow \infty}$ by Proposition 17. The first term is made arbitrarily small by (65), in the order $N \rightarrow \infty$, $R \uparrow \infty$, and $\eta \downarrow 0$. Together with the compact-containment estimate and tightness of the initial conditions, this proves C -tightness of the full sequence. $\square$

D.4 Stationary moment bounds and stochastic averaging

Lemma 21 (Stationary idleness moments with abandonment). *There are constants $c, C, N_0 > 0$, depending on β and γ , such that*

$$\mathbb{E}_{\pi_N^\gamma} I^{N,\gamma} \leq C, \tag{89}$$

$$\mathbb{P}_{\pi_N^\gamma}(I^{N,\gamma} \geq L) \leq C N e^{-cL^2/N}, \quad 0 \leq L \leq N, \tag{90}$$

$$\mathbb{E}_{\pi_N^\gamma}(I^{N,\gamma})^2 = O(N^{3/4}), \quad \frac{1}{N} \mathbb{E}_{\pi_N^\gamma}(I^{N,\gamma})^2 \longrightarrow 0. \tag{91}$$

Proof. Apply (72) to S and use (10). This gives

$$\mathbb{E} I^{N,\gamma} = \beta + \frac{\gamma}{N} \mathbb{E} W^{N,\gamma} \leq \beta + \gamma \mathbb{E} X_N^\gamma \leq C,$$

where the last bound follows from (73).

Abandonment never changes I . Therefore the stationary level-crossing identity across the cut $\{I \leq j\}$ is exactly

$$(N - \beta)\mathbb{P}(I^{N,\gamma} = j + 1) = \mathbb{E}[M_1^{N,\gamma} \mathbb{1}_{\{I^{N,\gamma} = j\}}].$$

Since $M_1^{N,\gamma} \leq N - I^{N,\gamma}$,

$$\mathbb{P}(I^{N,\gamma} = j + 1) \leq \frac{N - j}{N - \beta} \mathbb{P}(I^{N,\gamma} = j).$$

Let $j_0 = \lceil 2\beta \rceil \vee 1$. For $j \geq j_0$ and all large N ,

$$\frac{N - j}{N - \beta} \leq e^{-j/(2N)}.$$

Iteration gives $\mathbb{P}(I = j) \leq Ce^{-cj^2/N}$, and summation over $j \geq L$ proves (90). With $L = N^{3/4}$, use $I^2 \leq LI$ on $\{I \leq L\}$ and $I \leq N$ to obtain

$$\mathbb{E}I^2 \leq L\mathbb{E}I + N^2\mathbb{P}(I > L) \leq CN^{3/4} + CN^3e^{-cN^{1/2}}.$$

This proves (91). □

Proposition 18 (Marked deficit excursions with abandonment). *For integers K, v satisfying*

$$\beta + \gamma < K < N, \quad v \geq 1,$$

one has

$$\mathbb{P}_{\pi_N^\gamma}(V^{N,\gamma} \geq v, Z_3^{N,\gamma} \geq K) \leq \frac{(N + \gamma)v}{K - \beta - \gamma} \left(\frac{N - K + \gamma}{N - \beta} \right)^{v-1}. \quad (92)$$

Proof. Work on the two-sided stationary path space of $Q^{N,\gamma}$. Nonexplosion gives locally finite paths. Mark the transitions σ satisfying

$$V^{N,\gamma}(\sigma-) = 0, \quad V^{N,\gamma}(\sigma) = 1, \quad Z_3^{N,\gamma}(\sigma) \geq K.$$

Predictable thinning of the service and abandonment marks, as formalized in Proposition 5, gives the marked-excursion compensator

$$\mathbb{1}_{\{V(t-)=0, Z_3(t-)\geq K\}} \left(1 + \frac{\gamma}{N}\right) M_2(t-) dt.$$

At a marked start, $V = 0$ implies $I = M_1 = 0$ and $M_2 = N - Z_3 \leq N - K$. Its stationary intensity is therefore at most

$$\iota_{V,\gamma} \leq \left(1 + \frac{\gamma}{N}\right) (N - K) \leq N + \gamma. \quad (93)$$

Start the original chain from any possible post-jump state with $V = 1$ and $Z_3 \geq K$, and let

$$T_v = \inf\{t : V(t) \geq v\}, \quad T_K = \inf\{t : V(t) = 0 \text{ or } Z_3(t) < K\}.$$

Before T_K , Lemma 7 gives upward rate at most

$$b = N - K + \gamma$$

and downward rate

$$d = N - \beta.$$

The condition $K > \beta + \gamma$ is exactly $b < d$. Hence Lemma 15(c) yields

$$\mathbb{P}(T_v < T_K) \leq \left(\frac{N - K + \gamma}{N - \beta} \right)^{v-1}. \quad (94)$$

The exact generator drift of V is

$$\begin{aligned} (\mathcal{L}_N^\gamma V)(q) &= M_1(q) + M_2(q) + \frac{\gamma}{N} M_2(q) - (N - \beta) \\ &= \beta - I(q) - Z_3(q) + \frac{\gamma}{N} M_2(q). \end{aligned}$$

Thus, before T_K ,

$$(\mathcal{L}_N^\gamma V)(Q(t)) \leq -(K - \beta - \gamma).$$

Starting from any state with $V = v$ and $Z_3 \geq K$, Dynkin's formula stopped additionally at $V = m$ and at deterministic time T , followed by $m \rightarrow \infty$ and $T \rightarrow \infty$, gives

$$\sup_{\substack{q: V(q)=v \\ Z_3(q) \geq K}} \mathbb{E}_q T_K \leq \frac{v}{K - \beta - \gamma}. \quad (95)$$

The strong Markov property at T_v combines (94) and (95). Attach to each marked excursion the indicator profile from its first hit of v until T_K . This profile dominates occupation of $\{V \geq v, Z_3 \geq K\}$ within that excursion. Campbell's formula Lemma 15(a), together with (93), proves (92). The bound is governed by the marked-excursion start intensity $N + \gamma$, the subcritical birth rate $N - K + \gamma$, and the drift margin $K - \beta - \gamma$. $\square$

Proposition 19 (Stationary idleness-weighted approximation of the one-job queue count). *As $N \rightarrow \infty$,*

$$\mathbb{E}_{\pi_N^\gamma} \left[(I^{N,\gamma} + 1) \left| \frac{M_1^{N,\gamma}}{N} - r(X_N^\gamma) \right| \right] \rightarrow 0. \quad (96)$$

Proof. Set

$$\tilde{K}_N = \lceil N^{1/2} \rceil, \quad \delta_N = N^{-1/4}, \quad \Delta_N^\gamma = \left| \frac{M_1^{N,\gamma}}{N} - r(X_N^\gamma) \right|.$$

The deterministic bound (37) remains valid. On $\{Z_3^{N,\gamma} < \tilde{K}_N\}$,

$$H^{N,\gamma} = \sum_i (Q_i^{N,\gamma} - 2)^+ \leq \tilde{K}_N Q_{\max}^{N,\gamma}.$$

Therefore, by Cauchy–Schwarz, Lemmas 17 and 21,

$$\begin{aligned} &\mathbb{E}[(I^{N,\gamma} + 1) \Delta_N^\gamma; Z_3^{N,\gamma} < \tilde{K}_N] \\ &\leq \frac{\tilde{K}_N}{N} \mathbb{E}[(I^{N,\gamma} + 1) Q_{\max}^{N,\gamma}] + \frac{2}{N} \mathbb{E}[I^{N,\gamma} (I^{N,\gamma} + 1)] \end{aligned}$$

$$\begin{aligned}
&\leq \frac{\tilde{K}_N}{N} (\mathbb{E}(I^{N,\gamma} + 1)^2)^{1/2} (\mathbb{E}(Q_{\max}^{N,\gamma})^2)^{1/2} + O(N^{-1/4}) \\
&= O(N^{-1/8} \log N) + O(N^{-1/4}) = o(1).
\end{aligned}$$

On $\{Z_3^{N,\gamma} \geq \tilde{K}_N\}$, both $M_1^{N,\gamma}$ and $Nr(X_N^\gamma)$ lie in $[0, V^{N,\gamma}]$, so $\Delta_N^\gamma \leq V^{N,\gamma}/N$. The contribution from $V^{N,\gamma} \leq \delta_N N$ is at most

$$\delta_N \mathbb{E}(I^{N,\gamma} + 1) = o(1)$$

by (89). On the complementary event, use $\Delta_N^\gamma \leq 2$, $I^{N,\gamma} + 1 \leq N + 1$, and Proposition 18 with

$$K = \tilde{K}_N, \quad v = \lceil \delta_N N \rceil = \lceil N^{3/4} \rceil.$$

For all large N , $K > \beta + \gamma$, and

$$\left(\frac{N - K + \gamma}{N - \beta} \right)^{v-1} \leq e^{-cN^{1/4}}.$$

The remaining factor is polynomial in N , so this contribution also tends to zero. Combining the three regions proves (96). $\square$

Proposition 20 (Stationary stochastic averaging with abandonment). *For every $h \in C_c^1((1, \infty))$,*

$$\mathbb{E}_{\pi_N^\gamma} [h(X_N^\gamma) I^{N,\gamma} - h(X_N^\gamma) g(X_N^\gamma)] \longrightarrow 0. \quad (97)$$

Proof. Use the same corrector $\psi_N(q) = \Psi_N(S(q)/N, I(q))$ as in Section 5.4. The abandonment contribution is (80). Because a_h has compact support, that contribution can be nonzero only when $S(q)/N$ lies in a fixed compact enlargement of the support. On this set $W(q) \leq S(q) \leq C_h N$, and therefore

$$\left| \frac{\gamma W(q)}{N} \left[\Psi_N \left(x - \frac{1}{N}, i \right) - \Psi_N(x, i) \right] \right| \leq \frac{C_{h,\gamma}}{N^2} (i+1)^2.$$

Combining this estimate with Lemma 5 gives, for all states,

$$|(\mathcal{L}_N^\gamma \psi_N)(q) - h(x)(i - g(x))| \leq C_{h,\beta,\gamma} \left[(i+1) \left| \frac{m}{N} - r(x) \right| + \frac{(i+1)^2}{N} + \frac{(i+1)^2}{N^2} \right]. \quad (98)$$

For fixed N , ψ_N is bounded. The abandonment part of its generator is also bounded because it vanishes outside the compact enlargement just described. Hence (72) applies and $\mathbb{E} \mathcal{L}_N^\gamma \psi_N = 0$. Taking expectations in (98), the first term vanishes by Proposition 19, while the remaining terms vanish by (89) and (91). This proves (97). $\square$

Proposition 21 (BAR for subsequential stationary limits with abandonment). *The laws of X_N^γ are tight. Every weak subsequential limit ν is supported on $[1, \infty)$ and satisfies*

$$\int_{(1,\infty)} [f''(x) + b_\gamma(x) f'(x)] \nu(dx) = 0, \quad f \in C_c^\infty((1, \infty)). \quad (99)$$

Proof. Tightness follows from (73). Apply (72) to $f(S/N)$ for $f \in C_c^3((1, \infty))$. The upward rate of S is $N - \beta$, and its total downward rate is

$$N - I^{N,\gamma} + \frac{\gamma}{N} W^{N,\gamma}.$$

Multiplying the generator identity by N and applying Taylor's theorem gives

$$0 = \mathbb{E} \left[\left(I^{N,\gamma} - \beta - \gamma \frac{W^{N,\gamma}}{N} \right) f'(X_N^\gamma) \right] + \mathbb{E} \left[\left(1 - \frac{\beta + I^{N,\gamma}}{2N} + \gamma \frac{W^{N,\gamma}}{2N^2} \right) f''(X_N^\gamma) \right] + O_{f,\beta,\gamma}(N^{-1}). \quad (100)$$

The remainder is uniform: a jump can affect $f(S/N)$ only when S/N lies in a fixed compact enlargement of $\text{supp } f$, where $W = O(N)$.

Use the exact identity

$$\frac{W^{N,\gamma}}{N} = X_N^\gamma - 1 + \frac{I^{N,\gamma}}{N}.$$

The I/N terms vanish by Lemma 21, and Proposition 20 replaces $I^{N,\gamma}$ against f' by $g(X_N^\gamma)$. Passing along a weakly convergent subsequence in (100) proves (99).

Finally,

$$(1 - X_N^\gamma)^+ \leq \frac{I^{N,\gamma}}{N},$$

because $N - S = I - W \leq I$. Equation (89) therefore implies $\mathbb{E}(1 - X_N^\gamma)^+ \rightarrow 0$, so every subsequential limit is supported on $[1, \infty)$. $\square$

D.5 Pooled comparison, zero flux, and completion

Proposition 22 (Pooled $M/M/N + M$ comparison). *Let $C^{N,\gamma}$ be the pooled $M/M/N + M$ queue with arrival rate $N - \beta$, unit-rate servers, and waiting-customer abandonment rate γ/N . For all sufficiently large N ,*

$$C^{N,\gamma}(\infty) \leq_{\text{st}} S^{N,\gamma}(\infty). \quad (101)$$

Moreover,

$$\frac{C^{N,\gamma}(\infty)}{N} \Rightarrow Y_{\beta,\gamma}, \quad (102)$$

where $Y_{\beta,\gamma}$ is supported on $(1, \infty)$ and has density

$$y \mapsto \frac{e^{-\beta(y-1) - \gamma(y-1)^2/2}}{\int_0^\infty e^{-\beta z - \gamma z^2/2} dz} \mathbf{1}_{\{y>1\}}. \quad (103)$$

Proof. At population m , the pooled downward rate is

$$d_p(m) = \min(m, N) + \frac{\gamma}{N}(m - N)^+.$$

If the JSQ system has B busy servers and total population m , then its service-plus-abandonment rate is

$$d_d(m, B) = B + \frac{\gamma}{N}(m - B) = \frac{\gamma m}{N} + \left(1 - \frac{\gamma}{N}\right) B.$$

For $N \geq \gamma$, the coefficient of B is nonnegative. Since $B \leq \min(m, N)$,

$$d_d(m, B) \leq d_p(m).$$

The two systems have the same upward rate. A standard skip-free coupling uses common arrivals; when the populations are equal, every downward jump of the JSQ system is paired with a pooled downward jump, while the excess pooled downward rate acts only on the pooled process. This preserves $C^{N,\gamma}(t) \leq S^{N,\gamma}(t)$ from equal initial populations. Starting both from zero and passing to stationarity proves (101).

Let ϖ_N denote the stationary distribution of the pooled birth–death chain and put $\lambda_N = N - \beta$. For $j \geq 0$, detailed balance gives

$$\frac{\varpi_N(N+j)}{\varpi_N(N)} = a_N(j) := \prod_{\ell=1}^j \frac{N-\beta}{N+\gamma\ell/N}.$$

For every fixed $y \geq 0$,

$$\begin{aligned} \log a_N(\lfloor Ny \rfloor) &= \sum_{\ell=1}^{\lfloor Ny \rfloor} \left[\log \left(1 - \frac{\beta}{N} \right) - \log \left(1 + \frac{\gamma\ell}{N^2} \right) \right] \\ &\longrightarrow -\beta y - \frac{\gamma}{2} y^2, \end{aligned}$$

uniformly for y in compact subsets of $[0, \infty)$. Also,

$$a_N(j) \leq \left(1 - \frac{\beta}{N} \right)^j \leq e^{-\beta j/N}.$$

Dominated Riemann-sum convergence therefore yields

$$\frac{1}{N} \sum_{j=0}^{\infty} a_N(j) \longrightarrow \int_0^{\infty} e^{-\beta y - \gamma y^2/2} dy. \quad (104)$$

For the mass below N , detailed balance gives, for $1 \leq k \leq N$,

$$\frac{\varpi_N(N-k)}{\varpi_N(N)} = \prod_{\ell=0}^{k-1} \frac{N-\ell}{N-\beta}.$$

After absorbing the finitely many factors with $\ell \leq 2\beta$ into a constant, the logarithm of this product is bounded above by $-ck^2/N$. Hence

$$\sum_{m=0}^{N-1} \frac{\varpi_N(m)}{\varpi_N(N)} \leq C \sum_{k=1}^N e^{-ck^2/N} = O(\sqrt{N}). \quad (105)$$

Equations (104)–(105) show that stationary mass below N vanishes after normalization. Applying the same Riemann-sum argument with a bounded continuous test function identifies the limit of $(C^{N,\gamma} - N)/N$ as the law on $[0, \infty)$ with density proportional to $e^{-\beta y - \gamma y^2/2}$. This proves (102)–(103). $\square$

Proposition 23 (Boundary exclusion and zero-flux uniqueness). *Every weak subsequential limit of X_N^γ assigns probability one to $(1, \infty)$ and is the law with density (14).*

Proof. Let $X_{N_j}^\gamma \Rightarrow \nu$. By Proposition 21, ν is supported on $[1, \infty)$. The stochastic order (101) is closed under weak convergence. Since the limit in (102) assigns probability one to $(1, \infty)$,

$$1 = \mathbb{P}(Y_{\beta, \gamma} > 1) \leq \nu((1, \infty)),$$

so $\nu(\{1\}) = 0$.

Apply Lemma 6 to the BAR (99) on $(1, \infty)$. The measure ν has a locally absolutely continuous density p satisfying

$$p'(x) - b_\gamma(x)p(x) = c \tag{106}$$

for some constant c . For $x > 2$, the general solution is

$$\begin{aligned} p(x) = & \exp\left\{-\beta(x-2) - \frac{\gamma}{2}((x-1)^2 - 1)\right\} \\ & \times \left[A + c \int_2^x \exp\left\{\beta(y-2) + \frac{\gamma}{2}((y-1)^2 - 1)\right\} dy\right]. \end{aligned} \tag{107}$$

If $c < 0$, the bracket in (107) is eventually negative, contradicting nonnegativity of p . If $c > 0$, l'Hospital's rule gives

$$\begin{aligned} & \exp\left\{-\beta(x-2) - \frac{\gamma}{2}((x-1)^2 - 1)\right\} \int_2^x \exp\left\{\beta(y-2) + \frac{\gamma}{2}((y-1)^2 - 1)\right\} dy \\ & \sim \frac{1}{\beta + \gamma(x-1)}, \quad x \rightarrow \infty, \end{aligned}$$

which is not integrable. Thus $c = 0$. Solving (106) with zero probability flux on $(1, 2)$ and $(2, \infty)$, matching at two, and normalizing gives exactly (14). Hence every subsequential limit is the same law. $\square$

D.6 Interchange and moment convergence

Proof of Corollary 3. By Lemma 14, the fixed- N laws converge to π_N^γ from every initial law, so the steady-state-then-many-server limit is Theorem 6. Under the balanced initialization of Theorem 5, that theorem gives the many-server limit at each fixed time, and Proposition 4 sends the limiting diffusion to its invariant law as time tends to infinity. Finally, apply Lemma 13 to Theorem 6 and (73). $\square$