Critical diffusion limits and a spatial convergence rate in queueing networks with unreliable nodes and two-stage dynamic routing

El'mira Yu. Kalimulina

Lomonosov Moscow State University, Moscow, Russian Federation

elmira.yu.k@gmail.com ORCID 0000-0001-7158-040X

Abstract

We study an open queueing network in which each node is a pool of N parallel queues. After service, a customer who remains in the network first chooses an available destination according to the current routing probabilities, then joins the shortest of d queues sampled there. A failed node holds its customers until repair. The offered load is $1 - 1/N$ per active node; the environmental transition rates scale as $1/N$. We start from the joint stationary law of all queue lengths and the environmental phase specifying node availability and routing. Along subsequences on which the fractional part of $\log_d N$ converges, we prove joint convergence at times Nt of the centered node means and the phase to a switching diffusion and its driving chain. At repair, the first Wasserstein distance between the centered empirical queue-length distribution and the profile indexed by the centered mean at failure vanishes in probability. On growing, locally stabilizing graphs we obtain a stationary infinite-network limit on every fixed set of nodes. For the limiting diffusions, the Wasserstein error in the law of a node's trajectory decays exponentially in the radius of graph agreement. Under either a volume-growth bound or a bound on incoming routing weights, the numbers of nodes and queues per node may grow at arbitrary relative rates.

Keywords. Queueing networks; unreliable nodes; two-stage dynamic routing; join-the-shortest-queue routing with d choices; heavy traffic; switching diffusion.
Mathematics Subject Classification (2020). 60K25, 60F17, 60K37.

1 Introduction

We consider an open queueing network with unreliable nodes. Each node is a pool of parallel queues. When a node fails, its customers wait for repair while traffic continues along the surviving routes. The routing mechanism is motivated by large data centers. For example, Meta's ServiceRouter can change the probabilities of routing to different regions during an outage and select a server within the chosen region by sampling server loads [14, Section 3.4]. In our model a transfer first selects an available node and then compares queue lengths within it. We study the distribution of queue lengths that a node brings back into service at repair and the subsequent fluctuations of its mean queue length.

In the finite network each active node has one unit of spare capacity across its N queues. At physical times Nt, both failures and order-one fluctuations of the mean are visible. Since a failed node retains its queues, its distribution at repair is determined at the random time of failure.

Literature review. Dynamic rerouting around failed stations was studied for Jackson-type networks with one queue per station in [6, 7]. Chen and Whitt [3] obtained heavy-traffic limits

for single-server networks with service interruptions. Lorek and Szekli [10] bounded the rate of convergence to equilibrium for unreliable Jackson networks with reversible routing and repetitive-service blocking. In that model a transition aimed at a failed node causes another service at the source. Krenzler, Daduna and Otten [8] derived product forms under skipping and reflection, and Daduna and Szekli [4] studied temporal correlations in unreliable networks. Li, Li and Zhang (2017, arXiv:1703.08823) studied shortest-of-d routing with individually repairable servers.

Vvedenskaya, Dobrushin and Karpelevich [17] obtained the stationary mean-field queue-length law for shortest-of-two routing. With availability held fixed, our active nodes form the fast Jackson network of Martin and Suhov [12, (2.1), Theorem 4(ii)], who treated fixed subcritical load. Suhov and Vvedenskaya [16] allowed more general load balancing. The author's scalar preprint (2026, arXiv:2611.22569v1) gives the one-pool critical diffusion. You (2026, arXiv:2611.19766v1, Theorems 1 and 4) treats policies that give priority to idle servers; a sample of d misses a sole idle server with probability $(1 - 1/N)^d$. The author's companion manuscript *A stationary heavy-traffic diffusion limit for fast Jackson networks with dynamic routing within pools* gives the fixed-routing stationary reference used here (Theorem 1) and treats growing networks with fixed routes (EC.1–EC.3). The frozen diffusion fits the product-form framework of O'Connell and Ortmann [13, Theorem 3.2].

Atar (2026, arXiv:2606.01209) obtains a reflected McKean–Vlasov process for exchangeable single-server stations with uniform routing. In a pool the critical mean is centered by $\lfloor \log_d N \rfloor$. An empty pool thus corresponds to $-\lfloor \log_d N \rfloor$ in centered coordinates, and the limiting mean takes values on the whole real line. Budhiraja and Liu [1, 2] study reflected switching diffusions in generalized Jackson networks. Sen and Selvaraju [15] treat shared-buffer stations in a rapidly switching environment. Goldsztajn, Borst and van Leeuwaarden [5] compare queue lengths across a rapidly resampled graph and obtain a fluid limit. Lacker, Ramanan and Wu [9] derive marginal dynamics in terms of neighborhood histories on a unimodular Galton–Watson tree.

Contribution. Theorem 1 derives a switching diffusion for the centered means from the joint stationary law of the queues and the environmental phase in a fixed finite network. At all repairs on a bounded interval, the centered empirical distributions approach profiles indexed by the centered means at failure, or at time zero for nodes initially down. The stationary preparation step is needed because an outage changes the traffic seen by other nodes. Theorem 6 extends the limit to any locally finite routing graph with locally stabilizing weights and uniform traffic reserves, first taking the pool-size limit and then the graph limit. It bounds the spatial approximation error for the limiting diffusions and gives two conditions for simultaneous growth in the discrete network. A reciprocal route can leave memory in a node's path even when the environmental phase is observed (Corollary 7).

Section 2 defines the finite network and states the critical limit. Section 3 proves Theorem 1. Section 4 gives the stationary limit on growing networks and its spatial error bound. Section 5 illustrates the routing assumptions and examines what a single node's trajectory reveals.

2 The finite network and its critical limit

2.1 Network model and routing

Each of $K \geq 2$ nodes has $N \geq 2$ infinite-buffer, single-server queues. The directed routing graph may be sparse, asymmetric or disconnected. In physical time s an exogenous irreducible continuous-time Markov chain $E^N(s)$ on a finite set $\mathcal{E}$ has generator Q/N and invariant law π. Phase e has a nonempty set A_e of working nodes; each node works in at least one phase. The instantaneous

failure and repair rates are $\alpha_i(e)/N$ and $\beta_i(e)/N$, respectively, where

$$\alpha_i(e) = \sum_{f:\, i \notin A_f} Q_{ef} \quad (i \in A_e), \qquad \beta_i(e) = \sum_{f:\, i \in A_f} Q_{ef} \quad (i \notin A_e).$$

One transition can change the status of several nodes. For example, if one node remains available and another alternates between working and failed at rates α/N and β/N, the latter is available with stationary probability $\beta/(\alpha+\beta)$.

In phase e the substochastic routing matrix $P^e = (p^e_{ij})$ has zero diagonal and is supported on directed edges between nodes in A_e. A completion at $i \in A_e$ sends the customer to j with probability p^e_{ij} or out of the network with the remaining probability p^e_{i0}. All matrices are specified in advance. At a switch from e to f, P^f takes effect immediately for subsequent completions. One way to specify these matrices is to draw an ordered list of candidate destinations, skip unavailable nodes, and exit if the list is exhausted. A switch moves no customers. A failed node receives no arrivals, suspends service and retains its queues until repair.

For each active i assume

$$a^e_i = 1 - \sum_{j \in A_e} p^e_{ji} \geq a_* > 0, \qquad p^e_{i0} = 1 - \sum_{j \in A_e} p^e_{ij} \geq \eta > 0. \tag{1}$$

Conditional on the phase path, external arrivals at i are Poisson with rate $(N-1)a^e_i$ while active and zero while down. Given that path, the external streams, unit-rate Poisson service clocks at the servers, and fresh routing and sampling marks are mutually independent. A service clock removes one customer exactly when its server is working and occupied. Service requirements are thus exponential of mean one in working time. Every arrival, external or internal, samples a fixed $d \geq 2$ queues at its destination uniformly with replacement and joins a shortest sampled queue, choosing the first minimizer in sampling order. The fixed-phase traffic equation $\lambda^e_i = (N-1)a^e_i + \sum_{j \in A_e} p^e_{ji}\lambda^e_j$ has the unique solution $\lambda^e_i = N-1$ at every active node: row sums of the active routing matrix are at most $1-\eta$. Thus P^e may be reducible; the irreducibility assumption concerns the environment. The matrices and Q are fixed as N grows on a fixed graph.

The finite stationary law. Write L_e for the queue generator with phase e held fixed. An invariant law of the form $\pi_e \nu(dq)$ exists exactly when $\nu L_e = 0$ for every e: in the balance equation for phase e the switching terms cancel by $\pi Q = 0$. If all $P^e = 0$, each phase preserves $\nu_N^{\otimes K}$, where ν_N is the joint invariant law of an isolated working pool. The finite stationary law is then $\pi \otimes \nu_N^{\otimes K}$. Within a pool, exponential clocks alone do not yield independent queues. Consider two queues with $d = 2$ and total input rate one. If their stationary lengths were independent with common probabilities r_k, flow balance would give $r_0 = 1/2$, and balance at the empty state would give $r_1 = r_0/2 = 1/4$. Yet balance of arrivals to an empty tagged queue and departures from length one gives $r_1 = r_0[r_0/2 + 3(1-r_0)/4] = 5/16$. The product density in Theorem 1 describes the limiting centered node means.

2.2 First main result: the critical limit on a fixed network

Write $q_{ib}(s)$ for the full queue length, including a job in service, and $M_i(s) = N^{-1}\sum_b q_{ib}(s)$. Set $m_N = \lfloor \log_d N \rfloor$ and $\theta_N = \log_d N - m_N$. On accelerated time define

$$X^N_i(t) = M_i(Nt) - m_N, \qquad J^N(t) = E^N(Nt), \qquad \mu^N_i(t) = N^{-1}\sum_b \delta_{q_{ib}(Nt)-m_N}.$$

Thus $\mu_i^N(t)$ is the empirical measure of centered queue lengths in pool i, and its mean is exactly $X_i^N(t)$. The process J^N records the environmental phase, which determines the routing matrix. For $w > 0$ let ξ_w on $\mathbb{Z}$ have tails $\xi_w([j,\infty)) = \exp(-wd^j)$. Its mean $r(w) = \sum_{j\geq 1}\exp(-wd^j) - \sum_{j\leq 0}(1-\exp(-wd^j))$ decreases continuously from $+\infty$ to $-\infty$. Put $\omega = r^{-1}$ and $\Pi_x = \xi_{\omega(x)}$. Stochastic ordering gives $W_1(\Pi_x, \Pi_y) = |x-y|$. Write

$$g_\theta(x) = (d-1)d^\theta\omega(x), \qquad p_\theta(x) = Z_\theta^{-1}\exp\left[-x - \int_x^\infty g_\theta(v)\,dv\right].$$

Here Z_θ normalizes the density. For these one-pool laws the identity $r(dw) = r(w) - 1$ implies $\omega(x) \asymp d^{-x}$ and ensures integrability at both ends.

On A_e write $R_e = I - (P^e)^\top$ and $\Gamma_e = R_e + R_e^\top$. Extend both matrices by zero on all other entries. In particular the diagonal of Γ_e is twice the indicator of A_e. Evaluation of g_θ on a vector is componentwise. Write

$$d_{i,N}(t) = W_1\big(\mu_i^N(t), \Pi_{X_i^N(t)}\big), \qquad D_N(t) = \sum_i d_{i,N}(t)$$

for the discrepancy in pool i and its sum over all pools.

Theorem 1. *Under* (1) *the joint finite network and environment have a unique invariant probability at each N. Start in this law and take any subsequence $\theta_N \to \theta \in [0,1]$. For each finite T, (X^N, J^N) converges in distribution in the Skorokhod J_1 topology on $[0,T]$ to the stationary process (X,J), where X is continuous and*

$$dX_t = R_{J(t-)}\big(g_\theta(X_t) - \mathbf{1}\big)\,dt + \big(\Gamma_{J(t-)}\big)^{1/2}\,dB_t. \tag{2}$$

Here $J(t-)$ is the environmental phase immediately before t, the ith component of $g_\theta(X_t) - \mathbf{1}$ is $g_\theta(X_i(t)) - 1$, and $(\Gamma_e)^{1/2}$ is the unique symmetric positive semidefinite square root of Γ_e. The chain J has generator Q. The K-dimensional Brownian motion B can be chosen independent of $(X(0), J)$. An inactive coordinate remains fixed. A transfer between working nodes changes their means in opposite directions, which accounts for the off-diagonal covariances in Γ_e. The unique invariant law is

$$\mathbb{P}\{J = e, X \in dx\} = \pi_e \prod_{i=1}^K p_\theta(x_i)\,dx_i. \tag{3}$$

In particular, the limiting means $X_1(t), \ldots, X_K(t)$ are independent at each fixed time in stationarity. The empirical measures satisfy

$$\int_0^T D_N(t)\,dt \;\longrightarrow\; 0 \quad \textit{in probability.} \tag{4}$$

Also $D_N(t) \to 0$ in probability at each fixed $t \in [0,T]$, including $t = 0$. Jointly with the J_1 convergence of (X^N, J^N), $(\mu_i^N)_{i=1}^K$ converges in distribution to $(\Pi_{X_i(\cdot)})_{i=1}^K$ in the time-integrated metric $\int_0^T \sum_i W_1(\mu_i(t), \nu_i(t))\,dt$ on paths identified up to equality almost everywhere.

For nonstationary starts, suppose $(X^N(0), J^N(0))$ converges in law and $D_N(0) \to 0$ in probability. The same path and tracking limits hold with that joint initial law in (2)*. Future driving marks are fresh conditional on the initial state.*

The queue-length profile at repair. Under either initial condition, the profile approximation holds at all repairs in $(0, T]$. Fix a pool i and $\varepsilon > 0$. Write $\mathcal{R}_{i,T}^N$ for its repair times and extend $\beta_i(e)$ by zero when $i \in A_e$. On accelerated time the repair count has intensity $\beta_i(J^N(s-))$ relative to the joint history of the queues and the environment. Set $\beta_i^* = \max_e \beta_i(e)$ and $Z_{i,N} = \int_0^T d_{i,N}(s)\, ds$.

At a switch the queue configuration is unchanged almost surely, so $d_{i,N}(\tau) = d_{i,N}(\tau-)$ at every repair. We may therefore count repairs with a large error using a predictable indicator. The probability of at least one such repair is bounded by the expected number of such repairs. Compensation gives

$$\begin{aligned}\mathbb{P}\{\exists\, \tau \in \mathcal{R}_{i,T}^N :\ d_{i,N}(\tau) > \varepsilon\} &\le \mathbb{E} \sum_{\tau \in \mathcal{R}_{i,T}^N} \mathbf{1}_{\{d_{i,N}(\tau-) > \varepsilon\}} \\ &= \mathbb{E} \int_0^T \beta_i(J^N(s-)) \mathbf{1}_{\{d_{i,N}(s-) > \varepsilon\}}\, ds \\ &\le \beta_i^*\, \mathbb{E} \min\{T, Z_{i,N}/\varepsilon\}.\end{aligned}$$

The last bound uses the time spent above ε. By (4), $Z_{i,N} \to 0$ in probability. The minimum is bounded by T, so its expectation tends to zero. The same argument for all environmental switches, whose total rate in phase e is $-Q_{ee}$, bounds the probability of a switch with $D_N > \varepsilon$ by $q_*\, \mathbb{E} \min\{T, \varepsilon^{-1} \int_0^T D_N(s)\, ds\}$, where $q_* = \max_e(-Q_{ee})$.

If pool i fails at σ and reopens at τ, its queues remain fixed throughout the outage, including at repair. Thus $\mu_i^N(\tau) = \mu_i^N(\sigma)$ and $X_i^N(\tau) = X_i^N(\sigma)$. Hence the profile at repair is approximated by $\Pi_{X_i^N(\sigma)}$. For a pool already down at time zero, the corresponding profile at its first repair is $\Pi_{X_i^N(0)}$.

Observing the active set. The entire history of the active *sets* can be read from the limiting means in the interior of an observation interval. Indeed (2) gives

$$[X_i]_t = 2 \int_0^t \mathbf{1}_{\{i \in A_{J_s}\}}\, ds, \qquad [X_i, X_j]_t = -\int_0^t (p_{ij}^{J_s} + p_{ji}^{J_s})\, ds \quad (i \ne j), \tag{5}$$

with routing entries zero off the active block. Right derivatives of the first brackets identify A_{J_t} for $t < T$ since there are finitely many jumps on compact intervals. If several phases have the same set, this observation identifies the set itself. Thus instantaneous independence in (3) is compatible with pathwise observation of failures.

Corollary 2. *If a pool is active in some phases and inactive in others, the stationary mean vector X alone is not Markov in its natural filtration.*

Proof. At deterministic $t > 0$ there is almost surely no phase change at t. The left derivative of $[X_i]$ in (5) makes the indicator H of pool i being inactive measurable in the completed past of X. On an interval of length h the probability of a phase change, conditionally on the joint past, is at most $h \max_e(-Q_{ee})$. On an interval without a change, an inactive coordinate stays fixed and an active coordinate has an atomless transition law by ellipticity of its active block. It follows that the conditional probability of $X_i(t+h) = X_i(t)$ given the past of X differs from H by at most $h \max_e(-Q_{ee})$. Under the Markov property these probabilities would be measurable functions of $X(t)$. Their limit H would then have the same property, contradicting its independence of $X(t)$ and its nondegeneracy under (3). □

3 Proof of Theorem 1

We first prove the process limit for initial configurations whose empirical measures are close to the profiles indexed by their means. We follow the network between successive changes of environment and identify the invariant law of the limiting diffusion. We then show that the finite network's own stationary law provides the tightness and the initial profile approximation required by this argument.

3.1 From a frozen environment to the switching diffusion

Fix a phase e. Inactive pools retain their queues, while the active pools form an autonomous network with fixed routing. The fixed-routing theorem supplies a stationary reference on this active set. To pass from one holding interval to the next, we show that a prepared profile remains close to the profile indexed by its current mean.

Lemma 3 (Transfer from a stationary reference)**.** *Fix e and $T < \infty$. Along a subsequence with lattice phase converging to θ, let deterministic initial configurations satisfy $X^N(0) \to x$ and $D_N(0) \to 0$. With the environment held at e, the mean path on $[0,T]$ converges in law to the frozen diffusion started at x. Moreover, for every $\varepsilon > 0$,*

$$\sup_{0\le t\le T} \mathbb{P}\{D_N(t) > \varepsilon\} \longrightarrow 0, \qquad \int_0^T D_N(t)\,dt \longrightarrow 0 \quad \textit{in probability.}$$

Proof. We work on the active coordinates; the other means and empirical measures are constant. Couple two copies of the active network by common external arrivals, potential services at each label, destinations and ordered lists of probes, choosing the first probe of minimum length. Each mark preserves coordinatewise order. Indeed, if an arrival chooses a label in the smaller configuration whose length agrees in both, the same label wins in the larger one. A service performed in just the larger configuration removes a unit of excess at the source before any routing. The total excess between ordered configurations can only decrease under a common mark. Comparison with their coordinatewise minimum therefore shows that the sum of absolute differences between arbitrary labelled configurations is nonincreasing.

Permute labels within each pool to match the two initial configurations in increasing order. This preserves the law of all observables used here. The initial sum of absolute differences divided by N then equals the sum of the poolwise empirical W_1 distances. Call it d_0. At every later time the sum of these distances and the sum of the absolute mean differences are each at most d_0. Since $W_1(\Pi_u, \Pi_v) = |u - v|$, the two profile errors differ by at most $2d_0$ at every time.

Let B_δ be the open sup-norm ball of radius $\delta > 0$ about x_{A_e}. Start a reference copy of the active network in stationarity and condition only on $E_{N,\delta} = \{X^{N,\mathrm{ref}}_{A_e}(0) \in B_\delta\}$. Its transition law is unchanged. Theorem 1, equations (5)–(6), of the companion manuscript gives the stationary mean-path limit and vanishing expected profile error. The limiting mean density $\prod_{i\in A_e} p_\theta(z_i)$ is continuous and positive, so $\mathbb{P}_{\mathrm{stat}}(E_{N,\delta}) \to b_\delta > 0$ and the boundary of B_δ has zero limiting mass. For $D^{\mathrm{ref}}_{A,N}$, the sum of the active-pool errors in the reference copy, stationarity gives

$$\sup_{t\le T} \mathbb{E}_{\mathrm{stat}}[D^{\mathrm{ref}}_{A,N}(t) \mid E_{N,\delta}] \le \frac{\mathbb{E}_{\mathrm{stat}} D^{\mathrm{ref}}_{A,N}(0)}{\mathbb{P}_{\mathrm{stat}}(E_{N,\delta})} \longrightarrow 0.$$

The conditional expected time integral therefore also vanishes.

For fixed δ, the triangle inequality through the two mean-indexed profiles gives $d_0 \le |A_e|\delta + o_{\mathsf{P}}(1)$. The coupling thus proves both profile estimates, first taking $N \to \infty$ and then $\delta \downarrow 0$.

At fixed δ the reference mean path converges to the frozen diffusion with its stationary initial law conditioned to lie in B_δ, by joint convergence with the initial value and the null boundary of B_δ. The frozen diffusion has locally Lipschitz drift, constant covariance and nonexplosive paths. Coupling by common Brownian increments and stopping on compact sets gives continuous dependence in probability on the initial point on $[0,T]$. The limiting reference law therefore tends to the path law from x as $\delta \downarrow 0$. The uniform bound d_0 on the difference of the means proves the claim. □

For random starts, the empirical measures determine the transition law of the observables up to permutations of queue labels. Preparedness and initial convergence give joint convergence to $(x, (\Pi_{x_i})_i)$ with the prescribed random x. On a Skorokhod representation the lemma applies to almost every sequence of initial states. Integrating bounded continuous tests also carries along any jointly convergent past. At $t_N \to t$ the means converge by continuity and the profile estimate holds by its uniformity in deterministic time.

Conditional on the initial phase, the future environment is independent of the initial queues and of the primitive Poisson clocks and uniform marks used to construct the future queue dynamics. Fix the environmental path on $[0,T]$. Its finitely many holding intervals have deterministic lengths under this conditioning. Iterate Lemma 3 on these intervals: each switch preserves the prepared endpoint, and the next interval uses fresh queue marks. Concatenation gives the conditional mean-path limit. Adding the profile integrals and then integrating over the environment gives joint convergence. The isometry of $x \mapsto \Pi_x$ identifies the empirical limit in time-integrated W_1.

Identification in stationarity. The law $\pi(de) \prod_i p_\theta(x_i)\, dx_i$ is invariant for the switching process. Indeed, each frozen generator preserves the product measure: the active block preserves its factor by the fixed-graph theorem, and all other coordinates stay fixed. Conditional on any sequence of holding intervals, the law of the vector therefore remains the same product.

To prove uniqueness, fix $t > 0$ and a terminal phase e_*. From every initial phase there is positive probability of following, before t, a finite itinerary ending in e_* whose active sets together cover all pools. This follows from irreducibility of the environment and the assumption that each pool is active in some phase. On each holding interval of positive length the transition law of the active block is equivalent to Lebesgue measure on that block: the covariance is nonsingular, and on increasing balls stopped Girsanov equivalence applies to its locally bounded drift. Conditional on the holding times, after successive visits the law on the union of the visited active sets is equivalent to Lebesgue measure there. Indeed, the next interval retains the marginal of all previously visited coordinates outside its active set and gives a law equivalent to Lebesgue measure on that active set for every starting point. Disintegration gives the assertion on their union, including when the sets overlap. Thus every invariant probability of the time-t kernel dominates in the sense of absolute continuity the same nonzero measure, Lebesgue measure in phase e_*. Two distinct invariant probabilities would give two mutually singular ones: the positive and negative parts of their difference are separately invariant, since application of a Markov kernel preserves their masses and can only reduce total variation. This contradiction proves uniqueness. Thus any subsequential limit obtained from tight, prepared stationary finite systems is the stationary law and process stated in the theorem.

3.2 Tightness under the finite-network stationary law

We first record the one-pool estimate needed for a pool receiving rerouted traffic. Its total spare capacity can be any fixed positive number. Throughout this subsection $m_N = \lfloor \log_d N \rfloor$.

Lemma 4 (A fixed reserve)**.** *Consider N unit-rate servers with shortest-of-d sampling with replacement and Poisson input $N-\beta$, for fixed $\beta>0$ and $N>\beta$. Let M be their stationary mean length. The laws of $M-m_N$ are tight, as are the laws of their centered empirical measures in $(\mathcal{P}_1(\mathbb{Z}),W_1)$.*

Proof. Write $\rho=1-\beta/N$, $U_l=N^{-1}\#\{Q_b\ge l\}$, $D_l=N(1-U_l)$, $I=D_1$ and $X=M-m_N$. Positive recurrence with finite first moments follows from the quadratic Foster bound for this single pool (scalar preprint, Lemma 3.1). We extend the unit-reserve estimates in Section 3 and Supplement S2 of that preprint to cover the smaller reserve used in the comparison. The total-population balance gives $\mathbb{E}I=\beta$. The drift of $A_l=\sum_{h=1}^{l}D_h$ is $D_{l+1}-I-N\rho(1-U_l^d)$. Consequently $\mathbb{E}D_l\le\beta(d^l-1)/(d-1)$. Define

$$J_l=\rho U_l^d-U_{l+1},\qquad \mathcal{R}=\sum_{l\ge0}d^{-l-1}J_l^2,\qquad L=\log N.$$

For $N\ge2\beta$ the constant-service case of Proposition S2.1, equations (S2.1)–(S2.2), of the scalar preprint applies at the general load ρ, with its $b=1-\rho=\beta/N$. Direct substitution gives $\mathbb{E}\mathcal{R}\le C_\beta L^2/N^2$. To get the second deficit estimate put $z_l=\rho^{(d^l-1)/(d-1)}$. Iterating $|U_{l+1}-z_{l+1}|\le d|U_l-z_l|+|J_l|$ and using weighted Cauchy–Schwarz gives $|U_l-z_l|^2\le Cd^{2l}\mathcal{R}$. Since $N(1-z_l)\le C_\beta d^l$, it follows that $\mathbb{E}D_l^2\le C_\beta d^{2l}L^2$. The stationary spread bound $\mathbb{E}H\le C_d$, $H=N^{-1}\sum_b\exp(t_d|Q_b-M|)$, holds uniformly for $\rho\ge1/2$ by Lemma 3.3 of the scalar preprint.

Set $k=\lfloor\frac14\log_d N\rfloor$, $Y=D_k/d^k$. For any $f\in C_b^2(\mathbb{R})$, stationarity of $A_lf(X)$ gives

$$\mathbb{E}[D_{l+1}f(X)]=d\mathbb{E}[D_lf(X)]+\mathbb{E}[If(X)]+O_{\beta,f}(d^{2l}L^2/N).$$

Indeed $|\mathcal{L}f(X)|\le C_{\beta,f}(1+I)/N$, events changing A_l have expected rate $O_\beta(d^{l+1})$, and the error replacing $N\rho(1-U_l^d)$ by dD_l is bounded by $C_d(D_l^2+\beta D_l)/N$. Iteration from $D_1=I$ gives $\mathbb{E}[If(X)]-(d-1)\mathbb{E}[Yf(X)]\to0$. In particular $\mathbb{E}Y$ stays bounded and $\mathbb{E}Y^2\le C_\beta L^2$.

We detail the weighted cutoff used for the right tail. Put $G=\{L^{-3}\le Y\le L^3,\ H\le L^4,\ \mathcal{R}\le L^6/N^2\}$. The preceding bounds give $\mathbb{E}[Y;G^c]=O_\beta(L^{-1})$. Let $w=\lceil B\log L\rceil$ for a fixed sufficiently large B, and put $\theta_N=\log_d N-m_N$. Uniformly for $|j|\le w$ on G,

$$U_{m_N+j}=\exp(-Yd^{j-\theta_N})+O_{\beta,B}(N^{-1/8}L^{C_B}).$$

Indeed the square of the error in iterating the recursion from k to s is at most $Cd^{2s-k}\mathcal{R}$; the resulting expression is $\rho^{(d^{s-k}-1)/(d-1)}(1-Yd^k/N)^{d^{s-k}}$. Here $|\log\rho|\le2\beta/N$ for $N\ge2\beta$, giving a tilt error $O_{\beta,B}(N^{-1/4}L^{B\log d})$. The quadratic error in the second logarithm is $O_B(N^{-3/4}L^{6+B\log d})$.

The reference law with the displayed exponential tails places at least three quarters of its mass in $[-w_0,w_0]$, for $w_0=(3/\log d)\log L+O(1)$. Hence at least half the empirical mass belongs there, up to enlarging w_0 by one. On G this and the bound on H yield $|X|\le w_0+t_d^{-1}\log(2L^4)$. By increasing B, the empirical mean excess to the right of w is $o(1)$, using $H\exp[-t_d(w-|X|)]/t_d$. The reference mean excess to the left of $-w$ is $O(L^3d^{-w})$. Summing the tail differences in the intervening window therefore gives $X\le r(d^{-\theta_N}Y)+o(1)$ on G, where $r(z)$ is the mean of the integer law with tails $\exp(-zd^j)$. Since $r(dz)=r(z)-1$, the continuous function $r(z)+\log_d z$ is bounded. We obtain $\limsup_N\mathbb{E}[Y;X>R]\le C_\beta d^{-R}$.

Let $h_R(x)=h(x-R)$ be a smooth increasing cutoff from zero to one on $[R,R+1]$, with $0\le h'\le K_0$. Apply stationary balance to an antiderivative of $h_R(X)$, multiply by N, and use the drift numerator $I-\beta$ and scaled bracket rate at most two. The resulting inequality is $\beta\mathbb{E}h_R(X)\le\mathbb{E}[Ih_R(X)]+\mathbb{E}h_R'(X)+O_\beta(N^{-1})$. Thus $H_R^*=\limsup_N\mathbb{P}(X\ge R)$ satisfies $(\beta+K_0)H_{R+1}^*\le K_0H_R^*+C_\beta d^{-R}$ at positive integers; iteration proves upper tightness. The bound on $\mathbb{E}D_l$ implies $\mathbb{E}[N^{-1}\sum_b(m_N-Q_b-R)_+]\le C_\beta d^{-R}$ and gives lower tightness. Finally, on $|X|\le L_0$,

the average centered excess over $R > L_0$ is bounded by $t_d^{-1} H \exp[-t_d(R - L_0)]$. These estimates give tight first-moment tails in probability, which characterize tightness of the random empirical laws in $\mathcal{P}_1(\mathbb{Z})$. □

Lemma 5 (Stationary tightness)**.** *Suppose each pool is active in some phase. The joint finite system has a unique invariant probability law with finite first moments. In that law the vector of centered empirical measures is tight in $\mathcal{P}_1(\mathbb{Z})^K$ uniformly in N. This remains true conditionally on any specified present phase.*

Proof. Fix a pool i. Given an entire environment path, split the independent potential service clocks at each source by the prescribed destination probabilities in the current phase. Include a potential inward transfer even when its source queue is empty. On the active clock $\int_0^s \mathbf{1}_{\{i \in A_{E^N(u)}\}}\, du$, this produces, together with the external stream, an inhomogeneous Poisson process of intensity $(N-1)a_i^e + N\sum_{j \in A_e} p_{ji}^e = N - a_i^e$. Add independent dummy arrivals of intensity $a_i^e - a_*$ on that clock. The resulting rate is constantly $N - a_*$. These candidate arrivals are independent of the pool's own potential service clocks conditional on the environment, since $p_{ii}^e = 0$. Thus the comparison pool accepting every candidate, with the same ordered probes and the pool's own potential service clocks in active time, is precisely the scalar system with reserve a_*. A real internal transfer uses the candidate only if the source is occupied. First-minimum tie breaking on uniformly ordered probes preserves the coordinatewise order: if a lower state selects a coordinate on which the two states agree, that same coordinate is minimal and first among minimizers in the upper state. Potential services preserve order as well. Starting the real system empty and the upper pool in equilibrium gives coordinatewise domination at every physical time; the upper marginal conditional on the entire environment path is stationary. This construction is needed separately for each pool.

For fixed N these scalar bounds make the occupation averages of the real process from empty tight and bound the first moment of each pool. Any subsequential Cesáro limit of these occupation laws is invariant: for a bounded test, a fixed time shift changes its average over $[0, S]$ by $O(S^{-1})$, and tight convergence on the countable state space is in total variation. The first-moment bound passes by truncation. Every state can be emptied by active services with exits along a finite phase itinerary; every prescribed queue can be built by external arrivals probing that queue throughout their sample. Suppressing other marks in this finite construction has positive probability. Irreducibility proves uniqueness and the scalar domination passes to the invariant law, also conditional on the present phase when the environment is started in equilibrium.

Let I_i be the number of empty queues in pool i. Total-population balance in the invariant law gives

$$\mathbb{E} \sum_{i \in A_{E^N}} p_{i0}^{E^N} I_i = \mathbb{E} \sum_{i \in A_{E^N}} p_{i0}^{E^N} \le K,$$

because $\sum_{i \in A_e} a_i^e = \sum_{i \in A_e} p_{i0}^e$. Hence $\mathbb{E}[\mathbf{1}_{\{i \text{ active}\}} I_i] \le K/\eta$. Writing $D_{i,l} = \#\{b : q_{ib} < l\}$, testing $A_{i,l} = \sum_b (l - q_{ib})_+$ and using the true active input bound $\Lambda_i \le N$ gives $\mathbb{E}[\mathbf{1}_{\{i \text{ active}\}} D_{i,l}] \le C d^l$ by the same deficit recursion. To include inactive phases, let $U = \{e : i \in A_e\}$, $D = U^c$ and $v_e = \mathbb{E}[D_{i,l} \mathbf{1}_{\{E^N = e\}}]$. In a phase in D, pool i is frozen. Balance applied separately to each such phase yields $v_D Q_{DD} + v_U Q_{UD} = 0$, with the common factor $1/N$ canceled. If D is nonempty, $-Q_{DD}$ has a nonnegative inverse since the finite irreducible chain hits U. This proves $\mathbb{E} D_{i,l} \le C d^l$, with a constant independent of l, N. In particular the mean left excess beyond $-R$ of the centered empirical measure is bounded in expectation by $C d^{-R}$. Its right excess beyond R is dominated by the scalar stationary right excess. The preceding lemma and finiteness of K prove the claimed

Wasserstein tightness. The phase marginal is the fixed invariant law of Q, positive in every state, so unconditional tightness also yields each conditional assertion. □

3.3 Recovery of the stationary queue-length profile

Completion of the proof of Theorem 1. Start the finite system in its invariant law. Write $I_j(s) = \#\{b : q_{jb}(s) = 0\}$ and, for a configuration of pool i, denote its centered empirical measure by μ_i and its centered mean by x_i. Let $b_N = \sqrt{\log N}$ in physical time. The probability of a phase change before b_N is at most $b_N \max_e(-Q_{ee})/N = o(1)$. Fix an initial phase e and $i \in A_e$. From the same initial configuration construct a scalar comparison system with fresh Poisson input $N-1$ and unit potential service clocks. To couple it up to the first switch τ, aggregate the external input to i and all potential transfers to i as in Lemma 5, holding the mark probabilities at their phase-e values. This Poisson stream has rate $N - a_i^e$. Independently thin it to rate $N-1$ for the scalar system; the network accepts the original external candidates and the internal candidates with busy sources. Use common probes and common service clocks at i. Continue the scalar drivers with phase-e marks regardless of τ. They have the requisite scalar law, independently of the initial configuration conditional on e.

Up to τ, the rate of arrivals accepted in exactly one of the coupled systems is bounded by

$$1 - a_i^e + \sum_{j \in A_e} p_{ji}^e I_j(s-).$$

This bound counts independently thinned candidates and candidates at idle sources. A discrepant arrival creates at most one unit of labelled ℓ^1 distance; the other marks are nonexpansive. By stationarity and the total exit balance in Lemma 5, the expected number of discrepant arrivals before $b_N \wedge \tau$, conditional on the initial phase e, is $O(b_N/\pi_e)$. Indeed one may discard the initial-phase indicator and the stopping restriction in the unconditional upper bound involving the currently active idle counts. Dividing by N makes the discrepancy vanish in probability. The same bound controls both the empirical W_1 distance and the absolute mean difference at all times up to $b_N \wedge \tau$.

The conditional initial centered measures are tight in W_1 by Lemma 5. Their means are tight by continuity. From any subsequence extract a further one on which the conditional mean converges in law. The constant-service part of Theorem 2.2(i), equations (16)–(17), of the scalar preprint applies to the scalar system on $[0, b_N]$. At b_N its empirical measure is close in W_1 to the law indexed by its initial centered mean, and its mean changes by $o_{\mathsf{P}}(1)$ on the whole layer. The isometry of $x \mapsto \Pi_x$, the discrepancy bound and $\mathbb{P}(\tau \le b_N \mid E^N(0) = e) = o(1)$ show that the network pool is prepared at b_N in probability given the initial phase e. The conclusion holds on the original subsequence because the error limit is zero on every such extraction.

By stationarity and the small probability of a change up to b_N, the probability of an unprepared pool at a stationary observation time, jointly with its being *active*, tends to zero. To recover the observations while it is inactive, fix $\epsilon > 0$ and apply the down-phase balance of Lemma 5 to the bounded function $\mathbf{1}_{\{W_1(\mu_i, \Pi_{x_i}) > \epsilon\}}$ of this pool alone. Its values cannot change while the pool is down. The same fixed matrix $Q_{UD}(-Q_{DD})^{-1}$ therefore transfers the vanishing probability from active phases to every down phase. This argument preserves the distribution of the backlog retained through an outage.

All stationary initial vectors are now tight and prepared. Along each convergent subsequence Lemma 3 and the concatenation argument give the process limit from its subsequential initial law. Deterministic-time observations of the joint process converge, since X is continuous and J has no fixed jump times. Stationarity passes to the limit. Uniqueness in (3) identifies the same law along every subsequence and completes the proof. □

4 Second main result: the stationary limit on growing networks

How much can distant routes affect the stationary diffusion observed at one node? When do the discrete queues converge to this local process as both the graph and the pools grow? The finite routing graphs tend locally to a countable locally finite graph: each fixed-radius ball around a finite window in the undirected union of routes eventually agrees, including activity labels and all phase-dependent weights incident to its interior. Fix $\theta \in [0,1]$, the generator Q and d. Assume (1) uniformly, and extend inactive coefficients by zero. Write $H_i = \mathbf{1}_{\{i \in A_{J_t}\}}$, $h = g_\theta - 1$ and $\widehat{p}_{ji} = \max_e p^e_{ji}$.

Theorem 6 (Stationary limit on growing graphs)**.** *On each fixed finite window and horizon, the stationary finite-graph diffusions of Theorem 1 converge, jointly with their environment, to a stationary switching process on the infinite graph. Its continuous coordinates have drifts $H_i h(X_i) - \sum_j p^{J_t}_{ji} h(X_j)$ and brackets (5). Conditional on the environmental path, their joint law at a fixed time is the product of the densities p_θ. For the stationary queues this gives the iterated limit, first $N \to \infty$ with $\theta_N \to \theta$ on each fixed graph, then $K \to \infty$.*

Write $\mathcal{L}^{G_K}_{o,T}$ and $\mathcal{L}^{G}_{o,T}$ for the stationary diffusion path laws at node o on the finite and infinite graphs, respectively, with the same Q and θ. If the radius-L balls agree, including the phase labels and all weights incident to the interior, choose $\delta > 0$ with $c_\delta = e^\delta(1-a_*) < 1$ and put $D_\delta = (1-c_\delta)^{-1}$ and $q_* = \max_e(-Q_{ee})$. Then

$$W_{1,\infty}(\mathcal{L}^{G_K}_{o,T}, \mathcal{L}^{G}_{o,T}) \le 4T\big[D_\delta \exp\{q_* T(D_\delta - 1)\} - 1\big] e^{-\delta L}. \tag{6}$$

Here $W_{1,\infty}$ uses the uniform norm on $C([0,T])$. The bound depends on the common traffic reserve and the maximum switching rate and is uniform in the graphs, the root and the radius.

For the stationary queues the iterated limit also holds simultaneously at arbitrary relative rates $N, K \to \infty$ with $\theta_N \to \theta$, under either of the following uniform bounds:

$$|B(v,l)| \le Ce^{\beta l}, \quad e^\beta(1-\eta) < 1, \qquad \text{or} \qquad q := \sup_{K,i} \sum_j \widehat{p}_{ji} < 1.$$

Here $B(v,l)$ is the radius-l ball in the undirected union of all phase graphs. The first bound holds with common $C, \beta \ge 0$ for all v, l and finite graphs and includes uniformly polynomial growth. In both limits the means and environment converge in J_1; the empirical measures track their mean-indexed profiles in time-integrated W_1 and at all repairs in the window, in probability.

Proof. Fix a root o. For $L \ge 1$ write $B = \{i : \mathrm{dist}(o,i) < L\}$ in the union graph. We first compare finite-graph diffusions. Drive each directed edge $j \to i$ by an independent Brownian motion with coefficient $\sqrt{p^e_{ji}}$ and opposite signs at its ends; add independent vertex noises of variance rate $H_i(a^e_i + p^e_{i0})$. Use common phases and iid initial p_θ coordinates independent of the environment. Share the Brownian drivers at vertices in B and on every incident directed edge, using agreement on the one-step enlargement. The noises cancel throughout B. With $z_i = X_i - X'_i$ and $v_i = H_i|h(X_i) - h(X'_i)|$, monotonicity on a holding interval $[s,u]$ in phase e gives

$$|z_i(t)| + \int_s^t v_i\, dr \le |z_i(s)| + \sum_j p^e_{ji} \int_s^t v_j\, dr \qquad (i \in B,\ s \le t \le u).$$

Set $w_i = e^{-\delta \mathrm{dist}(o,i)}$. By (1), $w_i \sum_j p^e_{ji} y_j \le c_\delta \sup_{j: p^e_{ji} > 0} w_j y_j$ for $y_j \ge 0$, taking the empty supremum as zero. Condition on J. Write $\Delta_1, \dots, \Delta_n$ for its holding interval lengths up to T and $m = \int |h| p_\theta$.

Since $p'_\theta = hp_\theta$ and p_θ vanishes at both ends, $\int g_\theta p_\theta = 1$ and $m \le 2$. On interval r, let S_r be the larger of the two quantities

$$\sup_{i\in B} w_i \mathbb{E}\left[\sup_{t \text{ in } r} |z_i(t)| \mid J\right], \qquad \sup_{i\in B} w_i \mathbb{E}\left[\int_r v_i(t)\,dt \mid J\right].$$

For a neighboring vertex outside B, the weighted conditional integral of v_i is at most $2m\Delta_r e^{-\delta L}$ by conditional stationarity. The holding-interval inequality therefore gives $S_r \le S_{r-1} + c_\delta(S_r + 2m\Delta_r e^{-\delta L})$, with $S_0 = 0$. Iterating and summing, using $c_\delta D_\delta = D_\delta - 1$, yields

$$\sum_{r=1}^{n} S_r \le 2me^{-\delta L} \sum_{j=1}^{n} \Delta_j (D_\delta^{n-j+1} - 1) \le 2mTe^{-\delta L}(D_\delta^n - 1).$$

The left side bounds both the conditional expected path supremum at o and $\mathbb{E}[\int_0^T v_o(t)\,dt \mid J]$. The number of switches $n-1$ is dominated by a Poisson variable of mean q_*T, so $\mathbb{E}D_\delta^n \le D_\delta \exp\{q_*T(D_\delta - 1)\}$.

Identify agreeing balls of increasing radii with the infinite graph and use common marks there. The estimate gives local Cauchy paths in L^1 of the uniform norm and Cauchy drift integrands $H_i h(X_i)$ in L^1 of time and probability. Continuity of h and local finiteness identify the limiting equations. The supremum norm of a stationary coordinate has finite first moment: its drift has uniformly bounded first moment by the one-time product law, and its martingale has bracket at most $2T$. Averaging the coupled path estimate and passing to the infinite graph gives (6). For $L = 0$, couple the root initial values and scalar martingale noises, which have the same variance rate $2H_o$. Monotonicity of h bounds the expected path supremum of the difference by the two integrated absolute incoming drifts, at most $2mT(1 - a_*)$ in total, within (6). The isometry $W_1(\Pi_x, \Pi_y) = |x - y|$ gives the same bound for the laws of the corresponding profile paths with cost $\sup_{t\le T} W_1$. Stationarity and the conditional one-time product law pass to the limit. For pathwise uniqueness with common initial values and drivers in the class $\sup_i \mathbb{E}\int_0^T H_i|h(X_i)|dt < \infty$, fix an itinerary of n phases. Multiply the interval inequalities by its indicator and average over holding times: D_δ^n is fixed and $e^{-\delta L}$ times the two unconditional moment bounds controls the exterior. Let $L \to \infty$ and take the countable union of itineraries. Given the present mean vector and phase, resample the past conditionally and use the same independent future drivers. Pathwise uniqueness makes the futures coincide, which gives the Markov property for the full stationary process. For the iterated queue limit apply Theorem 1 on each fixed graph and use the isometry of $x \mapsto \Pi_x$; the repair estimate there holds for all switches.

For the simultaneous limit we bound the effect of the exterior directly on the stationary queues. For fixed routes this uses the stochastic comparison of Martin [11] and EC.2 of the companion. With switching routes, per-pool balance retains the current phase. Writing $u_i = \mathbb{E}[H_i I_i]$, we have

$$u_i = \mathbb{E}[H_i a_i^{E^N}] + \sum_j \mathbb{E}[p_{ji}^{E^N} H_j I_j].$$

Under the volume bound choose $\alpha > \beta$ with $c = e^\alpha(1 - \eta) < 1$, and put $w_i = e^{-\alpha \mathrm{dist}(o,i)}$ at root o. Summing over spheres bounds $\sum_i w_i$ uniformly and $\sum_{\mathrm{dist}\ge L} w_i \le C_\alpha e^{-(\alpha-\beta)L}$. In every phase $\sum_i w_i p^e_{ji} \le c w_j$ by the exit bound. Multiplying this balance by w_i and summing gives $(1 - c)\sum_i w_i u_i \le \sum_i w_i$. Taking each vertex as root yields $\sup_{i,K,N} u_i < \infty$.

On B, accept every exterior potential transfer, including rings at idle servers, and send outgoing transfers to the exit. Conditional on the phase path, split exterior Poisson clocks by destination. This upper network depends only on the common ball. Couple from empty with common phases,

services and ordered probes. The upper queues dominate the original queues; both have stationary couplings by the active-time scalar domination and time averaging of Lemma 5; the upper marginal is unique by the same irreducibility argument. Its potential input is still $N - a_i^e$ when active. Write $b_i = \mathbb{E}[H_i(I_i - I_i^+)]$, where $+$ denotes the upper queues. This is the mean rate of unmatched services and, by stationary population balance, of unmatched arrivals at i. These occur at rate $\sum_j p_{ji}^{E^N} H_j V_j$, where $V_j = I_j - I_j^+$ inside B and $V_j = I_j$ outside. Under the volume bound the phasewise weight inequality gives $(1-c)\sum_{i\in B} w_i b_i \le c \sum_{j\notin B} w_j u_j$. Thus $b_o \le C_\alpha e^{-(\alpha-\beta)L}$ uniformly.

Alternatively, under the bound on $\widehat{p}$, per-pool balance gives $u_i \le 1 + \sum_j \widehat{p}_{ji} u_j$, so $\sup_i u_i \le (1-q)^{-1}$ on every finite network. The same upper network on B gives $b_i \le \sum_j \widehat{p}_{ji} \tilde{b}_j$, where $\tilde{b}_j = b_j$ inside B and $\tilde{b}_j = u_j$ outside. Since $0 \le b_j \le u_j$ and every path from o to the exterior has at least L steps, iteration gives $b_o \le q^L/(1-q)$. Both bounds are uniform in N, K.

To compare the profiles, for the deficits of Lemma 5 put $F_{i,l} = D_{i,l} - D_{i,l}^+ \ge 0$. Subtract the stationary balances of $\sum_b (l - q_{ib})_+$ in the two systems. Their arrival rates satisfy $\Lambda_i \le \Lambda_i^+ \le N H_i$; the d-Lipschitz bound for u^d gives $\mathbb{E}[H_i F_{i,l+1}] \le b_i + d\,\mathbb{E}[H_i F_{i,l}]$ with first term $\mathbb{E}[H_i F_{i,1}] = b_i$. Down-phase balance as in Lemma 5 thus gives $\mathbb{E}F_{i,l} \le C d^l b_i$. Uniformity follows from finiteness of the phase space. For one network the bound on u_i likewise gives $\mathbb{E}D_{i,l} \le C d^l$, bounding the expected centered left excess below $-A$ by $C d^{-A}$. The scalar pool of reserve a_* dominates the upper right tail and is tight in $\mathcal{P}_1$ by Lemma 4. For their images under clipping to $[-A, A]$, the expected W_1 distance is at most $C_d d^A b_i$ by summation of $F_{i,m_N+j}/N$. Choosing A for both tails gives vanishing initial W_1 distance in probability as $L \to \infty$, uniformly in N, K. By order, this distance equals the difference of pool means at every time. The two unmatched rates bound its expected total variation on $[0, NT]$ by $2Tb_i$.

Fix a finite window S and radius L. Choose a graph from the original sequence matching the limiting neighborhood of S out to radius L. Glue its ordered coupling and that of any sufficiently large graph along the common stationary upper network and environment. For each $o \in S$, center the weight at o; the exterior is at distance at least L. Thus the differences of the means and empirical measures on S vanish in probability in supremum distance as $L \to \infty$, uniformly in N. At fixed L, take $N \to \infty$ by Theorem 1 on the chosen graph; the upper network serves only for comparison. Then let $L \to \infty$. The isometry of $x \mapsto \Pi_x$ transfers profile tracking. The bounded switch compensator applied to the summed discrepancy on S transfers it to every repair up to T, as after Theorem 1. □

5 Routing examples and consequences

Choosing routes on a prescribed graph. On any prescribed locally finite support one can realize the incoming bound in Theorem 6. For distinct neighbors take $p_{ij}^e = \chi s_{ij}^e / \max(\deg i, \deg j)$, $0 < \chi < 1$, with $0 \le s_{ij}^e \le 1$ vanishing at unavailable endpoints. Degrees are taken in the undirected graph. Row sums and incoming sums of phasewise maxima are at most χ. Retaining the infinite weights on internal edges of finite exhaustions and assigning residual mass to exits gives exact local agreement; the external rates follow from (1).

Rerouting on a ring. Place $4m$ pools on a ring. In phase e the pools with index e modulo four are down. The environment changes to any other phase with equal probability at total rate γ/N. Split the external rate $3m(N-1)(1-\chi)$ equally among working pools. A completion exits with probability $1-\chi$ and otherwise chooses the first working neighbor clockwise with probability u, counterclockwise otherwise. Here $0 < \chi, u < 1$. Each direction permutes the active pools, so

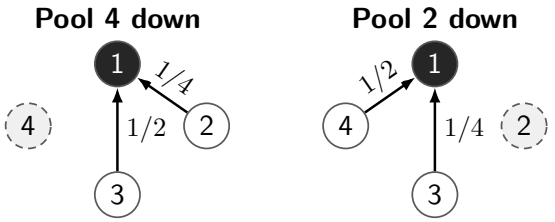


Fig. 1 Incoming routes to pool 1 on the four-pool ring ($u = 2/3$, $\chi = 3/4$). The dashed pool is down. Arrows give the probabilities of transfer to pool 1 upon service completion at each source. They sum to 3/4 in each phase

$a_i^e = p_{i0}^e = 1 - \chi$. Edges span at most two sites: Theorem 6 applies on the limiting line whenever $m, N \to \infty$ and $\{\log_d N\} \to \theta$, with no restriction on their relative growth. By (5), any two sites in the same residue class have constant limiting paths on $[0, T]$ on the same event, of probability $e^{-\gamma T}/4$ for $T > 0$. The common environment thus produces long-range spatial dependence of path observables, even though the stationary means at distinct nodes are independent at each fixed time.

For $m = 1$ and $u = 2/3$ the $1 \to 3$ transfer probability is $\chi/3$ when pool 4 is down and $2\chi/3$ when pool 2 is down. With $\chi = 3/4$ the two indicated phases already give $\sum_j \max_e p_{j1}^e \geq 5/4$ (Fig. 1). For $m \geq 2$ the possible incoming sources are one or two sites away on either side. The two clockwise probabilities each attain χu in some phase and the two counterclockwise probabilities each attain $\chi(1-u)$. Their maxima sum to 2χ. The volume condition applies to every ring.

Memory of a tagged node. We now consider a single node on a general routing graph. Even at fixed activity, short-lag correlations distinguish the routing direction: for distinct active i, j in phase e, stationarity gives

$$\frac{d}{dt}\,\mathrm{Cov}(X_i(t), X_j(0) \mid J(0) = e)\Big|_{t=0+} = p_{ji}^e.$$

Indeed conditional on the phase path the initial product law stays invariant. The drift in (2) has finite second moment in this law, and at time zero is continuous in L^2 from the right. Use $p_\theta' = (g_\theta - 1)p_\theta$ and integration by parts to obtain $\int (g_\theta(x) - 1)\, x p_\theta(x)\, dx = -1$. The Brownian increments have zero covariance with the initial state.

The following consequence applies to finite networks and their infinite-graph limits.

Corollary 7 (Memory of a reciprocal route)**.** *For the stationary diffusion on a finite graph or its infinite-graph limit, a route with $p_{ij}^e p_{ji}^e > 0$ for some e, j implies that both (X_i, J) and X_i alone are non-Markovian. On a fixed interval of positive length, the tagged path is also non-Markovian conditional on almost every environmental path that stays in that phase throughout the interval.*

Proof. Put $h = (\log p_\theta)'$ and $V = \mathrm{Var}_{p_\theta}(h) > 0$. In a finite stationary diffusion, integration by parts under the product density changes the frozen drift from $R_e h$ to $(\Gamma_e - R_e)h = R_e^\top h$ in the adjoint generator. The covariance is unchanged. Time reversal therefore transposes each routing matrix and replaces Q_{ef} by $\pi_f Q_{fe}/\pi_e$ for $e \neq f$. These transposed matrices interchange the two positive bounds in (1). The local diffusion convergence of Theorem 6, applied to both directions of time, identifies the same adjoint on the infinite graph.

Let S_t be the stationary semigroup, $\mathcal{P}$ conditional expectation given (X_i, J), and $T_t = \mathcal{P} S_t$ on its range. For $F = X_i$,

$$\langle F, (T_{2t} - T_t^2)F\rangle = \langle (I - \mathcal{P})S_t^* F, (I - \mathcal{P})S_t F\rangle = t^2 V \sum_e \pi_e \sum_j p_{ij}^e p_{ji}^e + o(t^2).$$

Indeed, X_i and both coordinate drifts are in L^2 by the tails of p_θ; Itô's formula and strong continuity give their respective generator expansions there. The forward drift after projection by $I-\mathcal{P}$ is $-\sum_j p^J_{ji} h(X_j)$, since $\int h p_\theta = 0$ and the coordinates are conditionally independent at one time. Transposing the weights gives the backward expression. The positive coefficient violates the semigroup identity. Projecting on X_i alone leaves the additional term $(H_i - \mathbb{E}_\pi H_i)h(X_i)$ in both residual drifts. It is orthogonal to the neighbor terms, so the coefficient increases by $V \operatorname{Var}_\pi(H_i)$. On an interval in fixed phase e the same calculation uses the frozen semigroup and conditions on X_i alone, with coefficient $V \sum_j p^e_{ij} p^e_{ji}$. □

Thus a well-posed scalar Markov McKean–Vlasov equation based only on the current one-pool conditional law cannot in general reproduce the tagged path. Descriptions using the past of a neighborhood remain possible. For (X_i, J) the test is inconclusive when no incident route is reciprocal.

Why phasewise balance matters. Let the environment alternate at rate $1/N$ between two and three working pools. Split a constant input of $\frac{5}{2}(N-1)(1-\chi)$ among them, with exit probability $1-\chi$ per completion. Start the environment in stationarity, with arbitrary initial queues. Mean fully busy exit capacity exceeds input by a constant. With probability $e^{-T}/2$ only two pools work throughout $[0, NT]$. Their exit-marked potential completions bound all exits. By the Poisson law of large numbers for arrivals and potential exits, the net gain is at least $(1-\chi)TN^2/4$ with conditional probability tending to one. Thus logarithmically centered pool means need not be tight under balance averaged over phases.

Statements and Declarations

Funding. This work received no external funding.
Competing interests. The author declares no competing interests.
Data availability. The results are analytical. The finite network used for the generator checks is specified in Section 5. No external datasets were used.
Author contribution. The author is responsible for the model, analysis and manuscript.
Use of AI tools. OpenAI Codex assisted in numerical testing and simulating large networks, translating into english and grammar correction.

References

[1] A. Budhiraja and X. Liu (2011). Multiscale diffusion approximations for stochastic networks in heavy traffic. *Stochastic Process. Appl.* **121**, 630–656. https://doi.org/10.1016/j.spa.2010.10.009.

[2] A. Budhiraja and X. Liu (2012). Stability of constrained Markov-modulated diffusions. *Math. Oper. Res.* **37**, 626–653. https://doi.org/10.1287/moor.1120.0556.

[3] H. Chen and W. Whitt (1993). Diffusion approximations for open queueing networks with service interruptions. *Queueing Systems* **13**, 335–359. https://doi.org/10.1007/BF01149260.

[4] H. Daduna and R. Szekli (2016). Correlation formulas for Markovian network processes in a random environment. *Adv. Appl. Probab.* **48**, 176–198. https://doi.org/10.1017/apr.2015.12.

[5] D. Goldsztajn, S. C. Borst and J. S. H. van Leeuwaarden (2026). Fluid limits for interacting queues in sparse dynamic graphs. *Stochastic Process. Appl.* **192**, 104794. https://doi.org/10.1016/j.spa.2025.104794.

[6] E. Yu. Kalimulina. Rate of convergence to stationary distribution for unreliable Jackson-type queueing network with dynamic routing. In V. Vishnevskiy, K. Samouylov and D. Kozyrev (eds.), *Distributed Computer and Communication Networks (DCCN 2016)*, Communications in Computer and Information Science **678**, 253–265. Springer, Cham (2016, published online 14 February 2017). https://doi.org/10.1007/978-3-319-51917-3_23.

[7] E. Yu. Kalimulina (2017). Analysis of unreliable open queueing network with dynamic routing. In V. Vishnevskiy, K. Samouylov and D. Kozyrev (eds.), *Distributed Computer and Communication Networks (DCCN 2017)*, Communications in Computer and Information Science **700**, 355–367. Springer, Cham. https://doi.org/10.1007/978-3-319-66836-9_30.

[8] R. Krenzler, H. Daduna and S. Otten (2016). Jackson networks in nonautonomous random environments. *Adv. Appl. Probab.* **48**, 315–331. https://doi.org/10.1017/apr.2016.2.

[9] D. Lacker, K. Ramanan and R. Wu (2023). Marginal dynamics of interacting diffusions on unimodular Galton–Watson trees. *Probab. Theory Related Fields* **187**, 817–884. https://doi.org/10.1007/s00440-023-01226-4.

[10] P. Lorek and R. Szekli (2015). Computable bounds on the spectral gap for unreliable Jackson networks. *Adv. Appl. Probab.* **47**, 402–424. https://doi.org/10.1239/aap/1435236981.

[11] J. B. Martin (2002). Stochastic bounds for fast Jackson networks. In *Analytic Methods in Applied Probability*, Amer. Math. Soc. Transl. Ser. 2, **207**, 133–140.

[12] J. B. Martin and Yu. M. Suhov (1999). Fast Jackson networks. *Ann. Appl. Probab.* **9**, 854–870. https://doi.org/10.1214/aoap/1029962816.

[13] N. O'Connell and J. Ortmann (2014). Product-form invariant measures for Brownian motion with drift satisfying a skew-symmetry type condition. *ALEA Lat. Am. J. Probab. Math. Stat.* **11**, 307–329. Journal article.

[14] H. Saokar, S. Demetriou, N. Magerko, M. Kontorovich, J. Kirstein, M. Leibold, D. Skarlatos, H. Khandelwal and C. Tang (2023). ServiceRouter: Hyperscale and minimal cost service mesh at Meta. In *17th USENIX Symposium on Operating Systems Design and Implementation*, 969–985. USENIX.

[15] A. Sen and N. Selvaraju (2026). Multi-scale diffusion approximations for stochastic networks in random environments under heavy-traffic. *ALEA Lat. Am. J. Probab. Math. Stat.* **23**, 333–366. https://doi.org/10.30757/ALEA.v23-14.

[16] Yu. M. Suhov and N. D. Vvedenskaya (2002). Fast Jackson networks with dynamic routing. *Probl. Inf. Transm.* **38**, 136–153. https://doi.org/10.1023/A:1020010710507.

[17] N. D. Vvedenskaya, R. L. Dobrushin and F. I. Karpelevich (1996). Queueing system with selection of the shortest of two queues: an asymptotic approach. *Probl. Inf. Transm.* **32**, 15–27. Math-Net.Ru.