Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Completion-Shock Queues: Departure-Induced Invalidation and Endogenous Service Correlation

Igor Kleiner

igkleiner@gmail.com

Holon Institute of Technology

Research Article

Keywords: queueing theory, quasi-birth-and-death process, rework, common shocks, correlated service times, heavy traffic

Posted Date: September 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10806477/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Completion-Shock Queues: Departure-Induced Invalidation and Endogenous Service Correlation

Igor Kleiner
Department of Data Science, Holon Institute of Technology, Holon, Israel
Corresponding author: igkleiner@gmail.com

Abstract

We consider a single-server first-come, first-served (FCFS) queue in which completing one job can increase the service requirements of jobs already waiting. External jobs arrive according to a Poisson process with rate λ, and a clean job has an exponential service time with rate μ. After each departure, a common shock occurs with probability p. The shock inspects the first two waiting positions and invalidates any clean jobs found there. An invalidated job must complete an exponential remediation phase with rate ν before ordinary service; further shocks have no additional effect on an already-invalidated job. For $d = 2$, the queue has a four-phase quasi-birth-and-death (QBD) representation above the boundary. We derive the exact stability condition

$$\lambda\left(\frac{1}{\mu} + \frac{2p - p^2}{\nu}\right) < 1,$$

a matrix-geometric stationary solution, formulas for the mean number in the system and the actual remediation fraction, and an explicit correlation structure under permanent backlog. When $\mu = \nu$ and $p = 0.3$, the stability threshold for the external load falls from 1 in the no-shock case to approximately 0.662. In the saturated system, invalidation indicators are exactly 1-dependent and

$$\mathrm{Cov}(S_n, S_{n+1}) = \frac{p(1-p)^3}{\nu^2}.$$

A direct expansion of the QBD kernel at its critical double root yields the exact first-order heavy-traffic coefficient. Relative to an independent and identically distributed (IID) queue with the same saturated marginal service-time law, completion-induced dependence produces a strictly positive heavy-traffic penalty for every $0 < p < 1$. The model is motivated by settings in which completing one task can invalidate assumptions used by nearby pending tasks. It gives a tractable queueing model of local rework propagation with exact stationary and heavy-traffic results.

Keywords: queueing theory; quasi-birth-and-death process; rework; common shocks; correlated service times; heavy traffic.

MSC 2020: 60K25; 60J27; 60K30.

1 Introduction

Most queueing models treat a successful service completion as purely workload reducing: one customer departs and the remaining work is unchanged. With dependent work items, that need not be true. A software change may invalidate a pending ticket, and an engineering decision may force already-prepared work to be redone.

This distinction matters because the backlog is no longer passive. Jobs that are already present are also potential targets of future rework, so the queue length helps determine future service requirements rather than merely recording unfinished work. The mechanism is natural in systems where work is prepared in advance and can be invalidated by an upstream change: software releases, engineering decisions, data-pipeline changes, or revised validation rules. A longer queue then creates more exposure to completion-induced rework.

The capacity effect can be substantial even in the minimal model. When $\mu = \nu$, the exact stability condition derived below is $\rho(1 + 2p - p^2) < 1$, where $\rho = \lambda/\mu$. At $p = 0.3$, the admissible external load therefore drops from $\rho < 1$ in the no-shock case to $\rho < 1/1.51 \approx 0.662$. Thus a system with, say, $\rho = 0.8$ would be stable in the corresponding $M/M/1$ model but unstable once completion-induced rework is included. This is the operational reason for isolating the mechanism.

We model this effect with a small Markovian rule. After a departure, a common shock occurs with probability p. It acts only on the first two waiting positions: clean jobs in those positions become invalidated and later require one remediation phase before ordinary service. Further shocks do not add more remediation to a job that is already invalidated.

The choice $d = 2$ is deliberate. With $d = 1$, a customer in a saturated system is exposed to only one predecessor shock, so the invalidation indicators are IID. Thus $d = 2$ is the smallest local rule for which shock windows overlap and service requirements become dependent. The next case, $d = 3$, already produces exact 2-dependence in saturation and has six reachable homogeneous phases under clean external arrivals; Remark 1 gives the corresponding covariance pattern. The position rule is a simple way to represent short-range dependencies without introducing a full dependency graph.

Two distinctions are important. The dependence is not imposed through an exogenous service process; it is generated by departures of the queue itself. Also, unlike an ordinary feedback queue, the customer just served leaves permanently—the extra work is attached to customers that were already waiting.

For $d = 2$, the paper gives four main results:

1. a four-phase quasi-birth-and-death (QBD) representation and an exact stability boundary;
2. a matrix-geometric stationary representation with explicit performance formulas;
3. an exact 1-dependence result for the saturated service sequence;
4. a direct-QBD derivation of the first-order heavy-traffic coefficient and a strict dependence penalty relative to an IID queue with the same saturated marginal service-time distribution.

We also check the matrix-analytic calculations against a separate customer-level discrete-event simulation.

2 Relation to existing models

The closest conceptual predecessors are generalized queueing networks with signals. In Gelenbe's G-network framework, signals may be generated by customer transitions after service and can trigger customer movement or batch removal [6]. Later multiclass models allow a signal to change the class of a customer in service [2]. Group-deletion signals provide a group-level interaction: in a LIFO multiclass network they delete a consecutive block of customers sharing class and phase at the back end of the buffer [3]. Restart G-networks are closer still: a signal can select a customer, change its class/service distribution, and restart its service [4, 5]. Thus, departure-generated signals and signals acting on groups or customer state are already part of the queueing literature.

Our interaction rule is different. A single departure shock acts on a fixed FCFS neighborhood, may invalidate both successor positions at once, leaves those jobs in place, and cannot add a second remediation phase to the same job. Fourneau et al. analyze restart of one selected customer under processor sharing and explicitly note that their model does not represent joint restart of a group of customers [4]. Here the joint update of a short FCFS block is the mechanism of interest.

Renovation queues form another neighboring class: at a service completion, waiting customers can be removed [8]. Our mechanism instead preserves customers and increases their future workload. Correlated-service queues are also well established; for example, semi-Markov service models specify dependence through a modulating process [1], and a recent Queueing Systems article uses a Markov construction to correlate exponential service times across two servers [12]. Neither correlated service times nor common shocks are new by themselves. Here the shock is generated by the queue's own departure process and acts on persistent marks of waiting FCFS successors rather than on server lifetimes or an exogenously specified service sequence. There are also queues in which customers change status while waiting, including deterioration/improvement models in which customers transition classes while waiting [7], and models in which service requirements evolve with elapsed delay [13]. These mechanisms differ from the predecessor-departure shocks considered here.

Rework propagation among dependent tasks is also well documented outside queueing theory. Product-development models represent upstream changes that create rework in downstream activities [10], and queue-based simulations couple task-to-task rework probabilities with queues of tasks awaiting developers [11]. Those models motivate the setting. The queueing question here is what follows when such an effect is tied to departures and to a short FCFS neighborhood.

The analysis therefore stays within standard matrix-analytic queueing theory. What is specific to this model is the departure-triggered joint update of a short FCFS successor block and the dependence created by that update.

3 Model

Jobs arrive according to a Poisson process of rate $\lambda > 0$. The discipline is first-come, first-served (FCFS) and the server is work conserving.

A job has one of two marks:

$$C = \text{clean}, \qquad R = \text{invalidated}.$$

A clean head job completes normal service after an exponential time with rate $\mu > 0$ and then leaves the system. An invalidated head job first completes remediation after an exponential time with rate $\nu > 0$; it then becomes clean while remaining at the head, after which it receives ordinary service.

After each departure, the queue first shifts forward. Before the next service phase begins, a completion shock occurs with probability p, independently of the past. The shock inspects exactly the first two waiting positions and changes each clean job there to state R. It does not look beyond those two positions. A job already marked R remains R, so invalidation does not stack.

Let $N(t)$ denote the number of jobs in the system. For $N(t) \geq 2$, record the marks of the first two jobs as

$$J(t) \in \{00, 01, 10, 11\}.$$

Jobs beyond the second position are clean, because a shock can reach only the first two waiting positions immediately after a departure. Thus $(N(t), J(t))$, together with the two-state boundary at $N = 1$, is a continuous-time Markov chain (CTMC).

4 QBD representation

Order the phases as $00, 01, 10, 11$ and set $q = 1 - p$. For levels $n \geq 3$, the upward, same-level, and downward blocks are

$$A_+ = \lambda I_4,$$

$$A_0 = \begin{pmatrix} -(\lambda+\mu) & 0 & 0 & 0 \\ 0 & -(\lambda+\mu) & 0 & 0 \\ \nu & 0 & -(\lambda+\nu) & 0 \\ 0 & \nu & 0 & -(\lambda+\nu) \end{pmatrix},$$

and

$$A_- = \begin{pmatrix} \mu q & 0 & 0 & \mu p \\ 0 & 0 & \mu q & \mu p \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The four downward transitions encode the common shock. From phase 00, a clean departure produces 00 with probability q and 11 with probability p. From phase 01, it produces 10 with probability q and 11 with probability p.

At level zero there is one state. At level one use states C, R. The boundary blocks are

$$Q_{00} = -\lambda, \qquad Q_{01} = (\lambda, 0),$$

$$Q_{10} = \begin{pmatrix} \mu \\ 0 \end{pmatrix}, \qquad Q_{11} = \begin{pmatrix} -(\lambda+\mu) & 0 \\ \nu & -(\lambda+\nu) \end{pmatrix}, \qquad Q_{12} = \begin{pmatrix} \lambda & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 \end{pmatrix},$$

and

$$Q_{21} = \begin{pmatrix} \mu q & \mu p \\ 0 & \mu \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Here Q_{10} corresponds to departure of a clean job from a one-customer system. The second row of Q_{21} reflects saturation of invalidation: if the remaining job is already invalidated, a shock cannot add further remediation.

Theorem 1 (Matrix-geometric stationary tail)**.** *If the CTMC is positive recurrent, there is a minimal nonnegative* 4×4 *matrix* R *satisfying*

$$A_+ + RA_0 + R^2 A_- = 0,$$

such that

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_2 R^{n-2}, \qquad n \geq 2.$$

Proof. Above the finite boundary the chain is a homogeneous finite-phase QBD. The standard matrix-geometric theorem for homogeneous QBDs therefore applies [9]. The minimal nonnegative solution of the quadratic matrix equation is the rate matrix, and stationarity of the homogeneous levels yields the stated tail form. □

5 Stability

Let

$$A = A_+ + A_0 + A_-.$$

The arrival terms cancel in A, because arrivals increase the level but do not change the phase. The stationary distribution $\boldsymbol{\phi}$ of the saturated phase chain is

$$\phi_{00} = \frac{\nu(1-p)}{D}, \qquad \phi_{01} = \frac{\nu p}{D}, \qquad \phi_{10} = \frac{\mu p(1-p)}{D}, \qquad \phi_{11} = \frac{\mu p}{D},$$

where

$$D = \nu + \mu p(2-p).$$

Theorem 2 (Exact stability boundary)**.** *The completion-shock queue is positive recurrent if and only if*

$$\boxed{\lambda < \frac{\mu\nu}{\nu + \mu p(2-p)}}.$$

Equivalently,

$$\boxed{\lambda\left(\frac{1}{\mu} + \frac{p(2-p)}{\nu}\right) < 1.}$$

Proof. First suppose $0 < p < 1$, for which the homogeneous phase chain is irreducible. The upward level drift in the homogeneous QBD is λ. A level decrease occurs only when the head is clean. Under the stationary saturated phase law, the clean-head probability is

$$\phi_{00} + \phi_{01} = \frac{\nu}{\nu + \mu p(2-p)}.$$

Hence the saturated downward drift equals

$$\mu(\phi_{00} + \phi_{01}) = \frac{\mu\nu}{\nu + \mu p(2-p)}.$$

The standard irreducible QBD drift criterion [9] gives positive recurrence exactly when downward drift exceeds upward drift. At $p = 0$ the model reduces to the ordinary $M/M/1$ queue and the formula gives $\lambda < \mu$. At $p = 1$, the recurrent saturated phase class alternates between a clean head with an invalidated successor and an invalidated head; its long-run departure rate is $\mu\nu/(\mu+\nu)$, again equal to the displayed formula. Thus the boundary expression also holds at both endpoints. □

When $\mu = \nu$ and $\rho = \lambda/\mu$, this reduces to

$$\rho(1 + 2p - p^2) < 1.$$

6 Stationary performance measures

Let

$$F = (I - R)^{-1}.$$

Let π_0 be the empty probability, $\boldsymbol{\pi}_1 = (\pi_{1,C}, \pi_{1,R})$, and $\boldsymbol{\pi}_2$ the level-two phase row vector. The boundary balance equations are

$$-\lambda\pi_0 + \mu\pi_{1,C} = 0,$$

$$\pi_0 Q_{01} + \boldsymbol{\pi}_1 Q_{11} + \boldsymbol{\pi}_2 Q_{21} = 0,$$
$$\boldsymbol{\pi}_1 Q_{12} + \boldsymbol{\pi}_2 (A_0 + RA_-) = 0,$$

with normalization

$$\pi_0 + \boldsymbol{\pi}_1 \mathbf{1} + \boldsymbol{\pi}_2 F \mathbf{1} = 1.$$

Proposition 1 (Mean queue length and remediation fraction)**.** *In stationarity,*

$$\boxed{\mathbb{E}[N] = \boldsymbol{\pi}_1 \mathbf{1} + \boldsymbol{\pi}_2 (2F + RF^2)\mathbf{1}.}$$

Consequently, by Little's law,

$$\mathbb{E}[T] = \frac{\mathbb{E}[N]}{\lambda}.$$

Let $h = (0, 0, 1, 1)^T$. Then the actual fraction of completed jobs that undergo remediation is

$$\boxed{\alpha_{\text{eff}} = \frac{\nu}{\lambda} \left(\pi_{1,R} + \boldsymbol{\pi}_2 F h\right).}$$

Proof. Using $\boldsymbol{\pi}_n = \boldsymbol{\pi}_2 R^{n-2}$,

$$\sum_{n \geq 2} n \boldsymbol{\pi}_n \mathbf{1} = \boldsymbol{\pi}_2 \sum_{k \geq 0} (k+2) R^k \mathbf{1} = \boldsymbol{\pi}_2 (2F + RF^2)\mathbf{1}.$$

For the remediation fraction, $\pi_{1,R} + \boldsymbol{\pi}_2 F h$ is the stationary probability that the head is invalidated. Remediation completions therefore occur at rate ν times this probability. Stable flow balance implies departures occur at rate λ, giving the ratio. □

The remediation fraction depends on the queue state because a shock may find zero, one, or two successor jobs. Under permanent backlog its value is $\alpha = p(2 - p)$. In the numerical example below, α_{eff} increases with load and moves toward this saturated value; the analysis does not require a general monotonicity result.

7 Saturated correlation structure

Consider a permanently backlogged version of the system. Let $Z_j \sim \text{Bernoulli}(p)$ denote the common shock after departure j, independently over j. Customer j is exposed to the two preceding shocks, so its invalidation indicator is

$$I_j = 1 - (1 - Z_{j-2})(1 - Z_{j-1}).$$

Theorem 3 (Exact 1-dependence in saturation)**.** *Let $\alpha = p(2 - p)$. Then*

$$\mathbb{P}(I_j = 1) = \alpha,$$

$$\boxed{\text{Cov}(I_j, I_{j+1}) = p(1-p)^3 > 0,}$$

and

$$\text{Cov}(I_j, I_{j+k}) = 0, \qquad k \geq 2.$$

If

$$S_j = X_j + I_j Y_j,$$

where $X_j \sim \mathrm{Exp}(\mu)$ and $Y_j \sim \mathrm{Exp}(\nu)$ are independent of the shocks and of each other, then

$$\boxed{\mathrm{Cov}(S_j, S_{j+1}) = \frac{p(1-p)^3}{\nu^2},}$$

while all service-time covariances at lags $k \geq 2$ vanish.

Proof. The probability of no invalidation is $(1-p)^2$, giving $\alpha = 1-(1-p)^2$. Adjacent indicators share exactly one shock. Writing $q = 1-p$,

$$\mathbb{P}(I_j = 0, I_{j+1} = 0) = q^3,$$

so

$$\mathrm{Cov}(I_j, I_{j+1}) = \mathbb{P}(I_j = 1, I_{j+1} = 1) - \alpha^2 = q^3 - q^4 = pq^3.$$

For lag at least two, the underlying shock windows are disjoint. More generally, any two blocks of indicators separated by at least one intervening index depend on disjoint sets of shocks, which proves exact 1-dependence rather than merely zero covariance. Finally, only the random remediation terms share dependence, and $\mathbb{E}[Y_j] = 1/\nu$, which gives the service-time covariance. □

Remark 1 (The next radius: $d = 3$)**.** *The same mechanism extends naturally to $d = 3$. Under permanent backlog with shock radius $d = 3$, customer j is exposed to the three preceding departure shocks, so*

$$I_j^{(3)} = 1 - (1 - Z_{j-3})(1 - Z_{j-2})(1 - Z_{j-1}).$$

Writing $q = 1-p$,

$$\mathbb{P}(I_j^{(3)} = 1) = 1 - q^3,$$

and the overlapping shock windows give

$$\mathrm{Cov}(I_j^{(3)}, I_{j+1}^{(3)}) = q^4 - q^6, \qquad \mathrm{Cov}(I_j^{(3)}, I_{j+2}^{(3)}) = q^5 - q^6,$$

while covariances vanish for lags $k \geq 3$. Hence the saturated invalidation sequence is exactly 2-dependent for $0 < p < 1$. With the same remediation construction,

$$\mathrm{Cov}(S_j^{(3)}, S_{j+k}^{(3)}) = \frac{\mathrm{Cov}(I_j^{(3)}, I_{j+k}^{(3)})}{\nu^2}, \qquad k \geq 1.$$

A full Markovian treatment of $d = 3$ can track the marks of the first three jobs. The naive binary encoding has $2^3 = 8$ labels, but the configurations 001 and 101 are unreachable under clean external arrivals, leaving six reachable homogeneous phases. More generally, the same shock-window argument gives a $(d-1)$-dependent saturated invalidation sequence; an unreduced d-bit encoding has 2^d labels, but reachability can reduce the phase set. We leave the larger QBDs outside the present paper; $d = 2$ is the smallest member of this hierarchy that already exhibits dependence.

8 Direct-QBD heavy traffic and the pure dependence penalty

Throughout this section assume $0 < p < 1$. The endpoint cases $p = 0$ and $p = 1$ are degenerate limits in which the saturated covariance term vanishes.

Let

$$\alpha = p(2-p), \qquad m = \mathbb{E}[S] = \frac{1}{\mu} + \frac{\alpha}{\nu}, \qquad \lambda_c = \frac{1}{m}.$$

For the saturated service requirement

$$S_j = X_j + I_j Y_j,$$

the marginal second moment is

$$\mathbb{E}[S^2] = \frac{2}{\mu^2} + \frac{2\alpha}{\mu\nu} + \frac{2\alpha}{\nu^2},$$

and set

$$c = \mathrm{Cov}(S_j, S_{j+1}) = \frac{p(1-p)^3}{\nu^2}, \qquad \sigma^2 = \mathbb{E}[S^2] + 2c.$$

The heavy-traffic coefficient follows directly from the QBD rate matrix. Define the matrix kernel

$$M(z, \lambda) = A_+ + zA_0 + z^2 A_-$$

and

$$g(z, \lambda) = \det M(z, \lambda).$$

Since $A_+ + A_0 + A_-$ is the generator of the saturated phase chain,

$$g(1, \lambda) = 0$$

for every λ. A direct differentiation gives

$$g_z(1, \lambda) = -\mu\nu\{\mu\nu - \lambda[\nu + \mu p(2-p)]\},$$

so $z = 1$ becomes a double kernel root exactly at $\lambda = \lambda_c$.

Lemma 1 (Critical kernel expansion)**.** *At $(z, \lambda) = (1, \lambda_c)$,*

$$g_{z\lambda}(1, \lambda_c) = \mu^2\nu^2 m > 0,$$

and

$$g_{zz}(1, \lambda_c) = -\frac{\mu^2\nu^2}{m^2}\sigma^2 < 0.$$

Let $r(\lambda) = \mathrm{spr}(R)$ be the Perron root of the QBD rate matrix. Then

$$\boxed{1 - r(\lambda) \sim \frac{2m^3}{\sigma^2}(\lambda_c - \lambda), \qquad \lambda \uparrow \lambda_c.}$$

Proof. The sparsity of M makes the determinant calculation short. Put $q = 1 - p$ and

$$\begin{aligned} a &= \lambda(1-z) + \mu z(qz - 1), \\ b &= \lambda(1-z) - \mu z, \\ f &= \lambda(1-z) - \nu z, \\ c_0 &= \mu q z^2, \qquad d_0 = \mu p z^2, \qquad e = \nu z. \end{aligned}$$

Then

$$M(z, \lambda) = \begin{pmatrix} a & 0 & 0 & d_0 \\ 0 & b & c_0 & d_0 \\ e & 0 & f & 0 \\ 0 & e & 0 & f \end{pmatrix}$$

and hence

$$g(z,\lambda) = abf^2 - ad_0ef - c_0d_0e^2.$$

Differentiating at $z = 1$ gives the stated derivatives. Let $yR = ry$ be a positive left Perron eigenvector of R. Left-multiplying

$$A_+ + RA_0 + R^2A_- = 0$$

by y yields

$$y\{A_+ + rA_0 + r^2A_-\} = yM(r,\lambda) = 0,$$

so $g(r,\lambda) = 0$. For $\lambda < \lambda_c$, the positive recurrent QBD has $r < 1$; continuity of the minimal nonnegative rate matrix and the QBD drift criterion imply $r \uparrow 1$ as $\lambda \uparrow \lambda_c$. The standard QBD spectral relation identifies the eigenvalues of the rate matrix with the relevant kernel roots [9]; thus the Perron branch is the nontrivial root that coalesces with the permanent root $z = 1$ at criticality. Writing $r = 1 - \varepsilon$ and $\lambda = \lambda_c - \delta$, the second-order Taylor expansion, after factoring the permanent root $z = 1$, gives

$$0 = g_{z\lambda}\varepsilon\delta + \frac{1}{2}g_{zz}\varepsilon^2 + o(\varepsilon\delta + \varepsilon^2),$$

with the signs absorbed by $r - 1 = -\varepsilon$ and $\lambda - \lambda_c = -\delta$. Solving for the positive ε yields the formula. □

Lemma 2 (Geometric-tail normalization)**.** *As $\lambda \uparrow \lambda_c$,*

$$\boxed{(1 - r(\lambda))\mathbb{E}[N] \longrightarrow 1.}$$

Proof. Near criticality, R has a simple Perron root $r \uparrow 1$. The critical-kernel expansion above shows that, after the permanent factor $z - 1$ is removed, exactly one kernel root approaches one; hence the non-Perron spectrum of R stays uniformly away from one in a left neighborhood of λ_c. Let v and u^T be corresponding right and left Perron vectors normalized by $u^Tv = 1$. Then

$$(I - R)^{-1} = \frac{vu^T}{1 - r} + O(1).$$

For $0 < p < 1$, the irreducible QBD has strictly positive Perron vectors for its rate matrix [9], and their coordinates vary continuously near criticality. Since $\boldsymbol{\pi}_2$ is nonnegative, normalization implies $\boldsymbol{\pi}_2 v = O(1 - r)$ and hence $\|\boldsymbol{\pi}_2\| = O(1 - r)$. The level-zero and level-one balance equations make the finite-boundary probabilities bounded linear functions of the level-two probabilities. Indeed,

$$\pi_{1,C} = \frac{\lambda}{\mu}\pi_0, \qquad \pi_{1,R} = \frac{\mu(p\pi_{2,00} + \pi_{2,01})}{\lambda + \nu},$$

and the first level-one balance equation then expresses π_0 linearly in $\pi_{1,R}$ and $\pi_{2,00}$. Hence $\pi_0 + \boldsymbol{\pi}_1\mathbf{1} = O(1 - r)$. The bounded non-Perron part of $(I - R)^{-1}$ contributes only $O(1 - r)$ after multiplication by $\boldsymbol{\pi}_2$. Therefore stationarity normalization yields

$$\frac{\boldsymbol{\pi}_2 v\, u^T\mathbf{1}}{1 - r} \longrightarrow 1.$$

Using

$$\sum_{k\geq 0}(k + 2)R^k = 2(I - R)^{-1} + R(I - R)^{-2},$$

the same Perron expansion gives $(1 - r)\mathbb{E}[N] \to 1$. □

Theorem 4 (Exact first-order heavy traffic)**.** *For the completion-shock queue,*

$$\boxed{\lim_{\lambda\uparrow\lambda_c}(\lambda_c-\lambda)\mathbb{E}[N]=\frac{\mathbb{E}[S^2]+2p(1-p)^3/\nu^2}{2m^3}.}$$

Proof. Combine the two preceding lemmas. □

Now compare with an IID $M/G/1$ queue having the same *saturated marginal* service-time law S. Pollaczek–Khinchine gives

$$\lim_{\lambda\uparrow\lambda_c}(\lambda_c-\lambda)\mathbb{E}[N_{\mathrm{iid}}]=\frac{\mathbb{E}[S^2]}{2m^3}.$$

Thus the pure dependence penalty is

$$\boxed{H_{\mathrm{sat}}(p,\mu,\nu)=1+\frac{2p(1-p)^3/\nu^2}{\mathbb{E}[S^2]}>1,\qquad 0<p<1.}$$

The strict inequality isolates the effect of dependence: the two systems have the same arrival rate and the same marginal service-time distribution under saturation.

If $\mu=\nu$, this simplifies to

$$H_{\mathrm{sat}}(p)=1+\frac{p(1-p)^3}{1+2p(2-p)}.$$

Differentiation gives the unique interior maximizer

$$\boxed{p^*=\frac{3-\sqrt{7}}{2}\approx 0.177124,}$$

with

$$\boxed{H_{\max}=\frac{27-7\sqrt{7}}{8}\approx 1.059968.}$$

Thus the pure correlation penalty is strictly positive but modest when the saturated marginal law is held fixed. A load-by-load matched IID comparison is a different object: it matches the actual remediation fraction at each load and therefore also reflects the state-dependent coupling between congestion, target exposure, and future service dependence. It should not be interpreted as a measurement of the saturated lag-one covariance alone.

9 Numerical verification

We checked the numerical QBD calculations against a customer-level continuous-time simulation that does not use the four-phase QBD recursion and instead stores each job's invalidation mark explicitly. Reproducibility code and the numerical verification tables are provided as Online Resource 1. In a longer verification suite with three replications across five parameter settings, the maximum relative discrepancy in $\mathbb{E}[N]$ was approximately 0.73%, and the maximum absolute discrepancy in α_{eff} was approximately 4.2×10^{-4}.

The direct-QBD heavy-traffic coefficient was also checked against the exact matrix-geometric solution. Across four distinct parameter triples and relative distances to criticality down to 10^{-3}, the maximum relative discrepancy between $(\lambda_c-\lambda)\mathbb{E}[N]$ and the analytic coefficient was approximately 0.34% at the smallest gap; simultaneously, the maximum deviation of $(1-r)\mathbb{E}[N]$ from one was approximately 3.3×10^{-3}. Figure 4 displays this convergence.

Figure 1 illustrates the endogenous increase of the remediation fraction with load for $\mu = \nu = 1$ and $p = 0.3$. Figure 2 compares the exact QBD with IID benchmarks. Figure 3 shows the analytic pure-correlation heavy-traffic ratio when $\mu = \nu$.

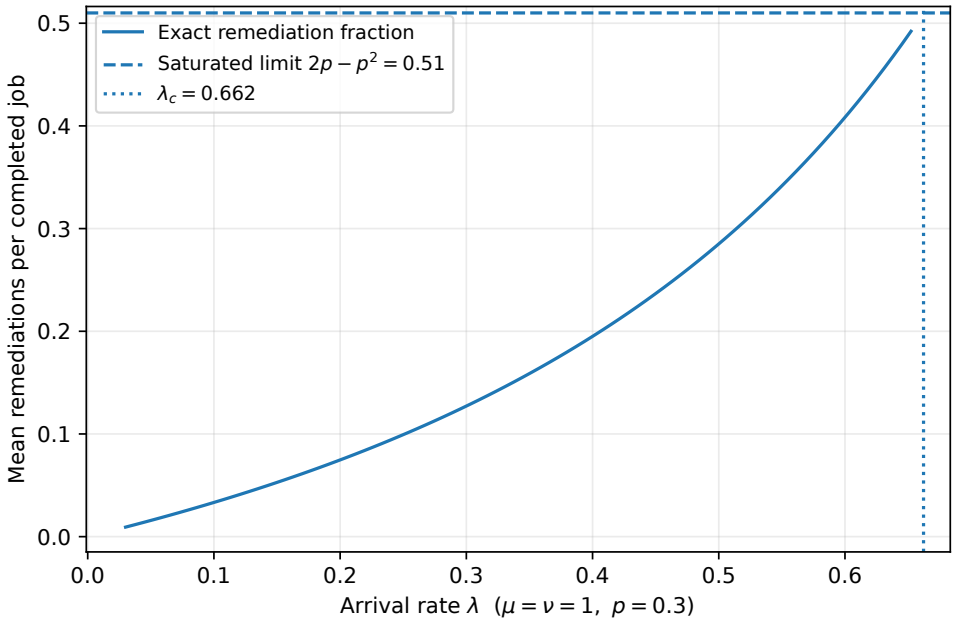


Figure 1: For $\mu = \nu = 1$ and $p = 0.3$, the remediation fraction rises with arrival rate and approaches its saturated value

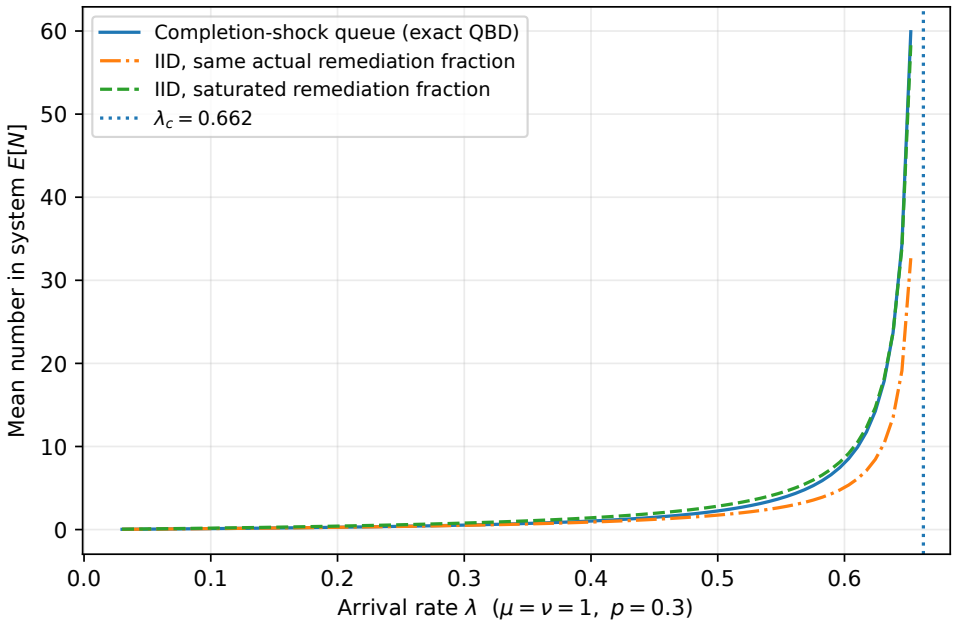


Figure 2: Exact mean number in system and IID comparisons for $\mu = \nu = 1$, $p = 0.3$

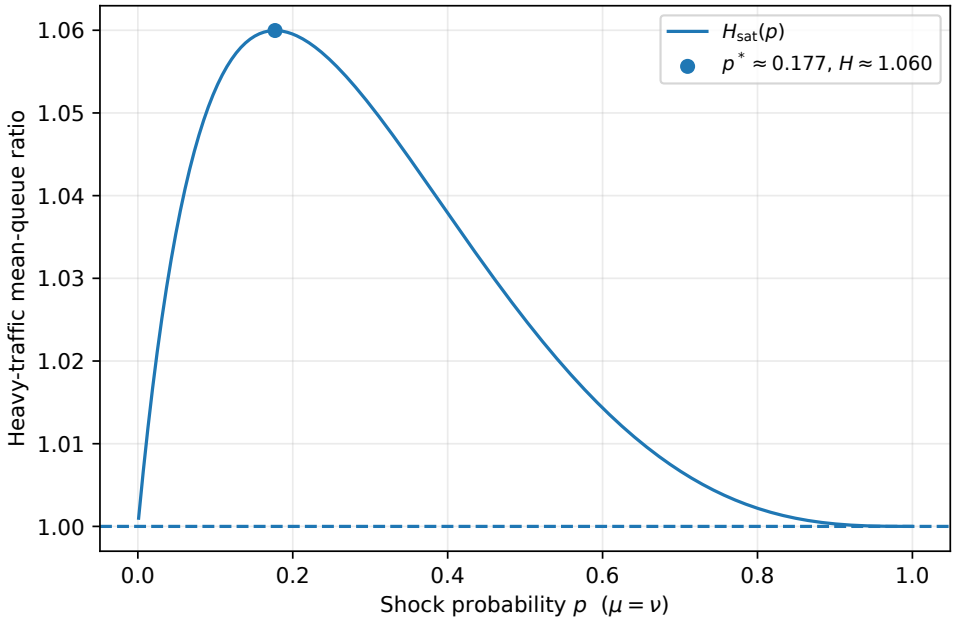


Figure 3: Analytic pure-correlation heavy-traffic ratio $H_{\text{sat}}(p)$ for $\mu = \nu$

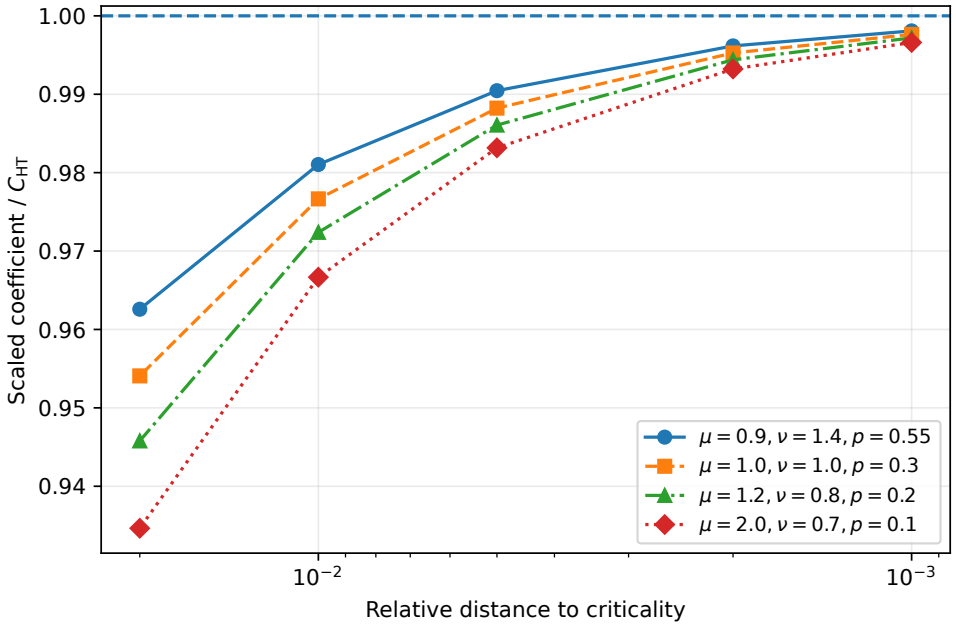


Figure 4: Direct-QBD verification of the first-order heavy-traffic coefficient across four parameter settings; the plotted quantity converges to one as $\lambda \uparrow \lambda_c$

10 Discussion

Two effects are present in the model. First, a common completion shock creates positive dependence between nearby future service requirements. This pure dependence effect is isolated by the saturated-marginal IID comparison and is strictly positive in heavy traffic. Second, the number of targets available to a shock depends on the queue state. This makes the remediation fraction itself endogenous. In the numerical examples it increases with load, creating a visible feedback from congestion to remediation exposure and back to congestion. Load-by-load matched IID comparisons contain both effects and can therefore show much larger penalties than the pure-correlation coefficient H_{sat}.

The QBD machinery is standard, and rework propagation already appears in engineering and product-development models. Here the specific FCFS interaction rule keeps the state space small enough for exact analysis while making the dependence arise from the queue dynamics themselves. The kernel calculation then connects the critical QBD root to the long-run variability of the induced service sequence.

Remark 1 shows the next step explicitly: at $d = 3$ the saturated sequence is 2-dependent and the reduced homogeneous phase space has six reachable states rather than the eight labels of the naive binary encoding. Larger radii increase both the dependence range and the phase description. We therefore keep the exact QBD and heavy-traffic analysis at $d = 2$, the smallest radius that already generates dependence. Other natural extensions include heterogeneous remediation rates, multiple servers, explicit dependency graphs, and finite buffers.

11 Conclusion

A successful completion need not merely remove work. In the completion-shock queue, it can jointly invalidate nearby pending work, creating additional remediation and coupling the service requirements of future customers. The minimal $d = 2$ model remains analytically small: it is a four-phase QBD with an exact stability threshold, a matrix-geometric stationary distribution, and computable stationary performance measures. In saturation, the induced service sequence is exactly 1-dependent. A direct QBD-kernel expansion gives the exact first-order heavy-traffic coefficient and proves a strict dependence penalty relative to an IID system with the same saturated marginal service law. The result is a tractable queue in which service dependence is generated by the queue dynamics rather than imposed as an external input.

Statements and Declarations

Funding. The author received no funding for this work.
Competing interests. The author declares no competing interests.
Data and code availability. All numerical data generated for the verification, together with the QBD, discrete-event simulation, figure-generation, and symbolic-check code needed to reproduce the computational results, are provided as Online Resource 1 accompanying the submission.
Supplementary information. Online Resource 1 is a reproducibility package containing the QBD solver, the independent customer-level simulation, symbolic checks, figure-generation code, and the numerical verification tables used in the article.

Use of AI-assisted tools. OpenAI ChatGPT was used to cross-check results, identify potential errors, and improve manuscript clarity and readability. The author reviewed all results and the final manuscript and takes full responsibility for the content.

References

[1] Abhishek, M. A. A. Boon, O. J. Boxma, and R. Núñez-Queija, "A single-server queue with batch arrivals and semi-Markov services," *Queueing Systems*, 86 (2017), 217–240. https://doi.org/10.1007/s11134-017-9531-4.

[2] T.-H. Dao-Thi, J.-M. Fourneau, and M.-A. Tran, "Networks of Symmetric Multi-class Queues with Signals Changing Classes," in *Analytical and Stochastic Modeling Techniques and Applications (ASMTA 2010)*, LNCS, pp. 72–86. https://doi.org/10.1007/978-3-642-13568-2_6.

[3] T.-H. Dao-Thi, J.-M. Fourneau, and M.-A. Tran, "Multi-class Network with Phase Type Service Time and Group Deletion Signal," in *Computer Performance Engineering (EPEW 2011)*, LNCS 6977, 2011, pp. 189–203. https://doi.org/10.1007/978-3-642-24749-1_15.

[4] J.-M. Fourneau, K. Wolter, P. Reinecke, T. Krauß, and A. Danilkina, "Multiple Class G-Networks with Restart," in *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering (ICPE)*, 2013, pp. 39–50. https://doi.org/10.1145/2479871.2479880.

[5] J.-M. Fourneau and K. Wolter, "Some Applications of Multiple Classes G-Networks with Restart," in *Computer and Information Sciences (ISCIS 2016)*, CCIS 659, 2016, pp. 126–133. https://doi.org/10.1007/978-3-319-47217-1_14.

[6] E. Gelenbe, "G-Networks with Signals and Batch Removal," *Probability in the Engineering and Informational Sciences*, 7 (1993), 335–342. https://doi.org/10.1017/S0269964800002953.

[7] Y. Hu, C. W. Chan, and J. Dong, "Optimal Scheduling of Proactive Service with Customer Deterioration and Improvement," *Management Science*, 68(4) (2022), 2533–2578. https://doi.org/10.1287/mnsc.2021.3992.

[8] A. Ya. Kreinin, "Queueing systems with renovation," *Journal of Applied Mathematics and Stochastic Analysis*, 10(4) (1997), 431–441. https://doi.org/10.1155/S1048953397000464.

[9] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, ASA–SIAM Series on Statistics and Applied Probability, SIAM, 1999. https://doi.org/10.1137/1.9780898719734.

[10] J. Lin, K. H. Chai, Y. S. Wong, and A. C. Brombacher, "A dynamic model for managing overlapped iterative product development," *European Journal of Operational Research*, 185(1) (2008), 378–392. https://doi.org/10.1016/j.ejor.2006.12.022.

[11] A. A. Mkrtchyan and J. Srinivasan, "Modeling distributed product development processes in small and medium enterprises," in *Proceedings of the 2015 International Conference on Industrial Engineering and Operations Management (IEOM)*, pp. 1–10, IEEE, 2015. https://doi.org/10.1109/IEOM.2015.7093882.

[12] S. Thapa and Y. Q. Zhao, "Markov modeling approach for queues with correlated service times: the $M/M_D/2$ model," *Queueing Systems*, 110 (2026), article 38. https://doi.org/10.1007/s11134-026-09998-x.

[13] C. Wu, A. Bassamboo, and O. Perry, "When Service Times Depend on Customers' Delays: A Relationship Between Two Models of Dependence," *Operations Research*, 70(6) (2022), 3345–3354. https://doi.org/10.1287/opre.2021.2179.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM12.zip