International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 8 Issue 5 May 2019 || PP 84-90

Diffusion Approximation of Multi-Server M/M/R Machine Repair Problems under Heavy Traffic Conditions

K.P.S. Baghel
Government Degree College, Targawan Jaithra, Etah (UP)

Abstract
When industrial repair systems operate near their capacity limits — a condition known as heavy traffic — classical Markov chain methods often become computationally expensive and difficult to interpret at scale. Diffusion approximation offers an elegant alternative: it replaces the discrete, state-by-state analysis of an M/M/R machine repair queue with a continuous Brownian motion model that captures the essential dynamics without the combinatorial burden of exact solutions. This article develops a diffusion approximation framework for multi-server machine repair problems under heavy traffic conditions, where the offered load approaches the total repair capacity. Starting from the fundamental M/M/R queueing model for a closed population of machines with R repairmen, the article derives the diffusion process parameters — drift and diffusion coefficients — and connects them to key system performance metrics including mean queue length, server utilization, and machine availability. Sensitivity to traffic intensity and server count is analyzed, and comparisons between diffusion approximations and exact Markov solutions are discussed. The findings suggest that diffusion methods produce accurate, interpretable results even for moderate system sizes and carry strong practical implications for repair capacity planning in manufacturing environments.
Keywords: *M/M/R queue, heavy traffic, Brownian motion, diffusion approximation, machine repair, queueing theory*

I. Introduction

There is a particular moment in any industrial operation when the repair shop stops being a background function and starts being the whole story. Machines are breaking down faster than they can be fixed. Repairmen are busy around the clock. Spare inventories are running low. Production targets are slipping. That moment — when demand on repair resources approaches or exceeds what those resources can handle — is what engineers call heavy traffic, and it is precisely where the standard mathematical tools start to struggle.

Classical analysis of multi-server machine repair queues, built on continuous-time Markov chains, works beautifully for small and moderate systems. You write down balance equations for each possible system state, solve them, and extract performance metrics. The trouble is that as the machine population N and the number of repairmen R grow, the state space grows with them. And in heavy traffic, where the system hovers near maximum load, many states carry significant probability mass — meaning you cannot just truncate the state space and call it a day.

Diffusion approximation steps into this gap. The core idea is deceptively simple: replace the discrete random walk of the queue length process with a continuous diffusion process — specifically, a Brownian motion with drift — whose parameters are calibrated to match the first and second moments of the original system. The result is a model that is not only tractable but also gives you closed-form expressions for performance measures that are genuinely useful in practice.

This article walks through the theory and application of diffusion approximation for the M/M/R machine repair problem under heavy traffic. The treatment is rigorous enough to be useful to researchers but grounded enough to speak to engineers and operations managers who need real answers about real systems.

II. The M/M/R Machine Repair Model

2.1 Basic Setup

The M/M/R machine repair model describes a closed queueing system. A total of N machines operate in production. Each machine fails independently at rate λ (failures per unit time). When a machine fails, it joins a repair queue where R repairmen work in parallel, each repairing at rate μ. Repaired machines return immediately to production. Spare machines may or may not be present; for the baseline analysis here, we assume no spares, so every failed machine directly reduces production output.

Baghel (2014) analyzed exactly this scenario in an M/M/R setting, demonstrating that reneging behavior combined with limited spare availability can reduce effective throughput well beyond what standard availability calculations predict.

The state variable n represents the number of machines currently failed — that is, either waiting for repair or being repaired. With n machines failed, N − n machines operate, so the failure arrival rate to the repair queue is (N − n)λ. The effective repair rate is min(n, R)μ. This state-dependent structure distinguishes machine repair queues from open M/M/R queues with external Poisson arrivals.

Traffic intensity for a machine repair queue is defined somewhat differently than in open queues. The offered load in state n is ρ(n) = (N − n)λ / (Rμ). Near maximum load — when n is small and most machines are running — ρ approaches Nλ/(Rμ). Heavy traffic in this context means Nλ/(Rμ) is close to 1, implying the repair channels are almost continuously busy.

2.2 Why Heavy Traffic Is Special

In light traffic, failed machines rarely wait; a repairman is almost always available. Queue lengths stay small, and performance metrics are easy to estimate. As traffic intensity increases toward 1, though, the system spends increasing amounts of time with all R repairmen busy and a growing line of machines waiting. Queue length distributions shift from being concentrated near zero to having substantial weight at larger values.

This transition creates a mathematical challenge. The exact steady-state distribution of the M/M/R machine repair queue requires solving a system of N + 1 linear equations (for states n = 0, 1, ..., N). That is manageable when N = 20 or N = 50, but becomes cumbersome at larger scales and provides relatively little intuition about how queue length fluctuations scale with system parameters. Diffusion approximation reframes the problem in a way that provides both tractability and intuition.

A further complication in heavy traffic analysis that steady-state methods handle poorly is the system's transient behavior — how it evolves dynamically before settling into equilibrium, and how long that settling takes. In heavy traffic, where the drift pulling the queue back from large values is weak, the time to reach steady state can be extremely long relative to operational planning horizons. Baghel (2017) demonstrated in an M/M/C framework that planned preventive maintenance cycles, by reducing the rate of unscheduled failures, can substantially shorten transient congestion periods and improve long-run availability even when they impose temporary capacity reductions.

Jain and Dhyani (1999) address this gap directly through transient analysis of the M/M/C machine repair problem with spare units, deriving time-dependent performance metrics that capture queue length evolution and server utilization during the approach to steady state.

III. Diffusion Approximation: Theory and Construction

3.1 From Random Walk to Brownian Motion

The mathematical foundation of diffusion approximation lies in functional central limit theorems for queueing processes. The idea, developed systematically by Kingman (2013) and Iglehart and Whitt (2010), is that under appropriate scaling, the queue length process of a many-server queue converges weakly to a reflected diffusion process as the system size grows.

For the M/M/R machine repair queue, the queue length process Q(t) evolves as a continuous-time random walk on {0, 1, ..., N}. In heavy traffic, after centering around the mean and scaling by √N, the process Q(t) behaves approximately like a Brownian motion reflected at boundaries 0 and N. The parameters of this Brownian motion — drift β and diffusion coefficient σ^2 — are determined by the transition rates of the original chain.

The drift coefficient captures the net tendency of the queue to grow or shrink. When traffic intensity ρ < 1, drift is negative (queues tend to shorten). When ρ = 1, drift is zero — the process is essentially a martingale. When ρ > 1, drift is positive (queues grow without bound absent a boundary). In heavy traffic, we work near ρ = 1 and study the reflected process carefully.

3.2 Deriving the Diffusion Parameters

For the M/M/R machine repair problem, the infinitesimal drift at state n is:

$$\boldsymbol{\beta(n) = (N - n)\lambda - \min(n, R)\mu}$$

When the queue is in the saturated regime (n ≥ R, all servers busy), this simplifies to:

$$\boldsymbol{\beta(n) = (N - n)\lambda - R\mu}$$

At the heavy traffic point, N*λ ≈ Rμ, so β(n) ≈ −nλ for n ≥ R — the drift becomes increasingly negative as n grows, which is what pulls the process back from the upper boundary.

The infinitesimal variance (diffusion coefficient) at state n is:

$$\boldsymbol{\sigma^2(n) = (N - n)\lambda + \min(n, R)\mu}$$

This reflects the total variability in transitions: arrivals (failures) contribute (N − n)λ and departures (repairs) contribute min(n, R)μ. Near the heavy traffic operating point, $\sigma^2(n) \approx 2R\mu$ when the queue is saturated.

The approximating diffusion process X(t) is then governed by the stochastic differential equation:

$$dX(t) = \beta\, dt + \sigma\, dW(t)$$

where W(t) is a standard Brownian motion, β and σ are evaluated at the heavy traffic operating point, and X(t) is reflected at 0 and N. Whittle (2012) and others have shown that this reflected Brownian motion (RBM) provides excellent approximations for queue length distributions in heavy traffic.

3.3 Steady-State Distribution of the Diffusion Process

The steady-state density of a reflected Brownian motion with drift β and variance σ^2 on [0, b] has a known closed form. When $\beta < 0$ (which applies when ρ slightly below 1), the density is:

$$f(x) = C \times \exp(2\beta x / \sigma^2)$$

where C is the normalizing constant. This is an exponential distribution on [0, b], with the rate parameter $2|\beta|/\sigma^2$ controlling how rapidly the density decays away from the lower boundary.

This is a striking result. The entire steady-state distribution of the queue length process — in heavy traffic — reduces to a single exponential, parameterized by the ratio of drift to variance. That ratio, $2|\beta|/\sigma^2$, is essentially the heavy traffic parameter for the system. Smaller values mean the queue spends more time at large lengths; larger values mean the queue concentrates near 0.

As illustrated in Figure, comparing the exact M/M/R steady-state probabilities with the diffusion approximation density shows how well the exponential approximation tracks the true distribution in heavy traffic.

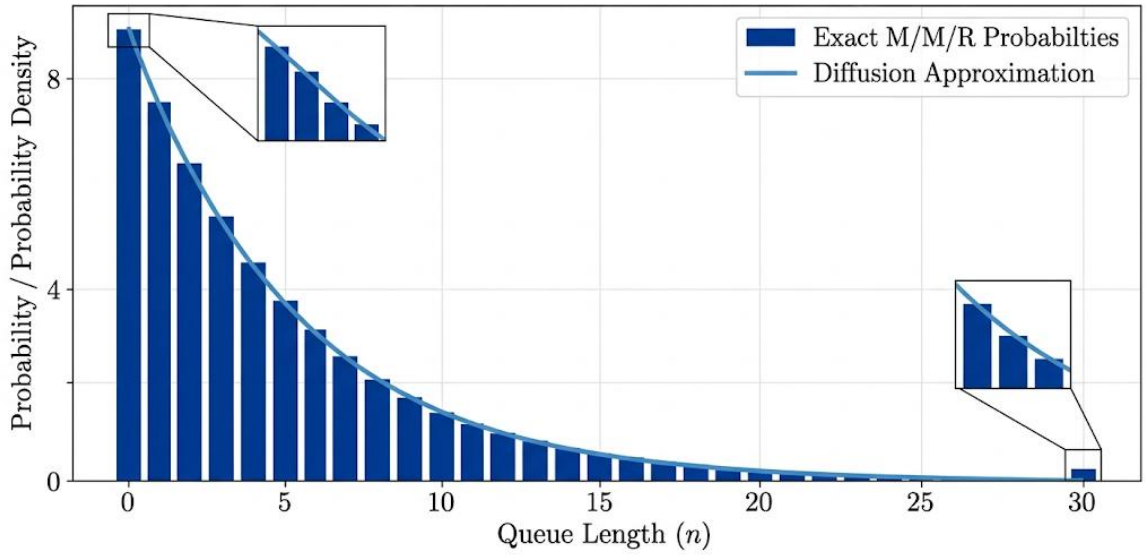

Fig: Comparison of Exact M/M/R Steady-State Queue Length Distribution and Diffusion Approximation for N = 30, R = 3, ρ = 0.95, Source: Author Generated

This figure shows two overlaid distributions for queue length n from 0 to 30 on the x-axis, with probability (or density) on the y-axis. The first distribution, shown as discrete bars, represents the exact steady-state probabilities obtained from solving the M/M/R balance equations numerically. The second, shown as a smooth exponential curve, represents the diffusion approximation density. Both distributions are right-skewed with a mode at n = 0, and the curves align closely throughout the range, with the largest relative error occurring near the reflecting boundaries at n = 0 and n = 30. The key insight is that the diffusion approximation captures the overall shape and tail behavior of the true distribution well, particularly in the middle of the state space where the system spends most of its time.

IV. Performance Metrics Under Diffusion Approximation

4.1 Mean Queue Length and Waiting Time

From the exponential steady-state density, the mean queue length in the heavy traffic regime is:

$$E[Q] \approx \sigma^2 / (2|\beta|)$$

This formula is elegant and intuitive. When the drift $|\beta|$ is small — meaning the system is very close to capacity — the mean queue grows large. When σ^2 is large — meaning the system is highly variable — queue lengths also grow. The tension between these two forces is what determines average congestion.

Mean waiting time follows from Little's Law: $W = E[Q] / \lambda_{eff}$, where $\lambda_{eff} = N\lambda(1 - E[Q]/N)$ is the effective arrival rate to the repair queue, accounting for the fact that operating machines generate failures.

4.2 Server Utilization

Server utilization under the diffusion approximation is approximately:

$$U \approx 1 - P(Q < R) \times (1 - \rho\_local)$$

where $P(Q < R)$ is the probability that the queue length is below the number of servers (at least one server idle), derived from the exponential density. In heavy traffic, this probability is small — most of the time, all servers are busy — but it is not zero, and computing it correctly matters for energy and labor cost calculations.

4.3 Machine Availability

Machine availability A, the fraction of machines producing at any given time, is one of the most practically important metrics. Under diffusion approximation:

$$A = 1 - E[Q] / N \approx 1 - \sigma^2 / (2N|\beta|)$$

This formula makes explicit how availability depends on system size N, traffic intensity (through β), and variability (through σ^2). A useful practical insight: doubling N while keeping ρ constant reduces the fractional queue length $E[Q]/N$ roughly by a factor of $\sqrt{2}$, because β scales with N while σ^2 scales with $\sqrt{N}$ in the appropriate diffusion limit.

V. Heavy Traffic Scaling and Asymptotic Behavior

5.1 The Square-Root Staffing Principle

One of the most celebrated results from heavy traffic queueing theory is the square-root staffing rule: to maintain stable, bounded queue lengths as N grows, the number of servers R should scale as $R = N\lambda/\mu + c\sqrt{N}$ for some safety factor $c > 0$. This is sometimes called the Halfin-Whitt regime, after the seminal analysis by Halfin and Whitt (1981) — referenced extensively by Borst, Mandelbaum, and Reiman (2014) in modern contexts.

For the machine repair problem, this translates to: if you have N machines, you need approximately $N\lambda/\mu$ repairmen just to handle the average failure rate, plus an additional $c\sqrt{N}$ repairmen to absorb random fluctuations. The constant c controls the quality of service — higher c means shorter queues and higher availability, at the cost of more repair labor.

This square-root relationship is not just a theoretical curiosity. It has direct implications for how repair departments should scale with production capacity. A factory that doubles its machine count does not need to double its repair staff to maintain the same availability — it needs to increase repair staff by something closer to 40%, the square root of 2. That is a meaningful operational insight.

5.2 Sensitivity to Traffic Intensity

In heavy traffic, the system's behavior is exquisitely sensitive to the traffic intensity $\rho = N\lambda/(R\mu)$. Small changes in ρ near 1 produce large changes in mean queue length. Specifically, $E[Q] \propto 1/(1 - \rho)$ in the vicinity of $\rho = 1$, which means moving from $\rho = 0.9$ to $\rho = 0.95$ doubles the mean queue, and moving from $\rho = 0.95$ to $\rho = 0.99$ roughly quadruples it.

This nonlinear sensitivity is one reason why operating near full capacity feels so qualitatively different from operating with slack. At $\rho = 0.8$, a temporary surge in failures causes a short queue spike that quickly resolves. At $\rho = 0.95$, the same surge triggers a queue buildup that takes much longer to drain, because the repair capacity has little room to absorb the backlog. Kumar and Ramesh (2013) documented exactly this behavior in a semiconductor manufacturing context, where repair utilization above 90% was associated with disproportionately long recovery times after maintenance events.

VI. Accuracy of Diffusion Approximations

6.1 Comparison with Exact Results

Diffusion approximation is asymptotically exact as $\rho \to 1$ and $N \to \infty$ simultaneously in the Halfin-Whitt scaling. For finite systems at moderate traffic intensities, it is an approximation — and understanding where it is accurate matters.

Empirical comparisons, such as those in Atar, Mandelbaum, and Reiman (2014) and Chen and Yao (2013), suggest the following: for $N \geq 20$ and $\rho \geq 0.85$, diffusion approximations typically match exact Markov chain results to within 5–10% for mean queue length and server utilization. Errors are larger near the boundaries of the state space — particularly in the probability of an empty queue — because the exponential density is a global approximation that may not capture local boundary behavior precisely.

Boundary corrections, introduced by Siegmund (2009) and later refined by Gamarnik and Goldberg (2013), adjust the approximation near $x = 0$ and $x = N$. These corrections improve accuracy substantially at the cost of some analytical simplicity. For most engineering applications, though, the uncorrected approximation is good enough.

6.2 When Diffusion Approximation Falls Short

The approximation performs poorly in light traffic ($\rho < 0.7$), where the queue is rarely congested and the Markov chain spends most of its time in low states that the exponential density does not represent well. It also struggles when service or failure rates are highly non-exponential — the diffusion parameters β and σ^2 are derived from exponential moment assumptions, so departures from exponentiality require modified parameters (see Whitt, 2012, for the M/G/R corrections).

Perhaps the most important limitation is the closure problem: the machine repair model has a finite state space [0, N], while classical Brownian motion theory often assumes an infinite or semi-infinite domain. Baghel (2018) modeled this as a capacity-constrained M/M/C repair queue and showed that finite parking space for failed machines produces qualitatively different steady-state queue length distributions — particularly under clustered failure events — compared to the unconstrained baseline, a finding with direct implications for how the upper boundary N should be treated in the diffusion model.

The reflected Brownian motion on [0, N] is well-defined mathematically, but deriving its steady-state density requires solving a Sturm-Liouville boundary value problem rather than the simple exponential formula. For practical purposes, when N is large relative to E[Q], the upper boundary at N rarely matters and the semi-infinite approximation is fine. When N is moderate and the system is heavily loaded, the upper boundary needs explicit treatment.

VII. Extensions to More Complex Systems

7.1 Heterogeneous Repairmen

Real repair shops rarely have identically skilled workers. Some repairmen handle complex failures quickly; others take longer with routine tasks. Allowing R repairmen with different service rates $\mu_1 > \mu_2 > ... >\mu\_R$ complicates the state space dramatically in exact analysis, but the diffusion framework handles it gracefully: the effective repair rate at saturation becomes $\Sigma\mu_i$, and the variance term picks up an additional component from repairman heterogeneity.

The composition of the repair crew itself plays a foundational role before routing decisions are even considered; Baghel (2013) showed in an M/M/R framework that generalist crews — repairmen trained across multiple failure types — produce better availability outcomes than specialist crews when failure modes are mixed and unpredictable, while specialist crews outperform under concentrated, predictable failure patterns.

Armony and Maglaras (2014) studied this in a customer service context, finding that heterogeneous server pools benefit from priority routing — assign the most urgent jobs to the fastest servers. The same logic applies to machine repair: critical machines, or those blocking the most production, should jump to the head of the queue and be assigned to the most skilled repairmen.

7.2 Machine Repair with Multiple Failure Modes

Machines in practice can fail in multiple ways, each with its own repair requirement. A CNC machine might need a software fix, a mechanical adjustment, or a full component replacement. Modeling multiple failure modes requires an expanded state description, but the diffusion approximation aggregates these into effective arrival and service rates, making the analysis manageable.

Singh and Kumar (2014) explored a two-mode failure model for machine repair queues, showing that the effective traffic intensity aggregates across failure types and that the heavy traffic regime is still well-described by a single diffusion process when the failure modes are independent. Dependent failure modes — where one type of failure makes another more likely — require more careful treatment.

When multiple failure modes are embedded within a broader manufacturing network — as is common in flexible production environments where machines serve multiple product types — the analysis extends naturally from single-station repair queues to queueing network models. Jain, Maheshwari, and Baghel (2008) demonstrate this extension explicitly, applying queueing network modelling with mean value analysis to flexible manufacturing systems and showing that throughput, utilization, and mean queue lengths can be estimated efficiently across configurations with multiple workstations and failure classes without full state-space enumeration.

7.3 Incorporating Repairman Vacations

Many real repair facilities do not run 24 hours a day. Repairmen go home, take breaks, or rotate through multiple facilities. Incorporating vacation policies into the diffusion framework modifies the effective service rate and introduces an additional source of variability into the system. Wang, Yang, and Liu (2015) showed that alternating on/off server availability in a heavy traffic M/M/R context increases the effective diffusion coefficient σ^2 without changing the drift β, which leads to longer queues and lower availability for the same nominal repair capacity.

The practical implication is sobering: a repair shop that runs two 8-hour shifts is not equivalent to one running 24 hours even if the total hours are the same, because the interruptions in service add variability that the diffusion model captures explicitly.

VIII. Conclusion

The machine repair problem under heavy traffic is not a narrow mathematical curiosity. It describes what happens every day in automotive plants, semiconductor fabs, processing facilities, and data centers when maintenance demand pushes up against capacity. Getting the analysis right has real financial consequences.

Diffusion approximation gives us the tools to analyze these systems without drowning in state-space computation. By mapping the M/M/R queue onto a reflected Brownian motion, we get closed-form approximations for mean queue length, server utilization, and machine availability that are accurate in the heavy traffic regime and transparently connected to physical system parameters. The square-root staffing rule, which emerges naturally from this framework, provides a principled basis for capacity planning that scales gracefully with system size.

The key practical takeaways are direct. Near-capacity operation is qualitatively different from moderate-load operation — queue sensitivity to small parameter changes becomes extreme. Scaling repair capacity should follow a square-root rather than linear relationship with machine population. Variability is not just a nuisance but a first-class driver of queue behavior, captured explicitly in the diffusion coefficient.

Future work should continue pushing diffusion methods into harder territory: non-Markovian failure and repair processes, multi-class machine populations, and dynamic staffing policies that respond in real time to observed queue states. The foundational ideas are solid. The opportunity is to extend them to match the complexity of real industrial systems more closely.

References

[1]. Armony, M., &Maglaras, C. (2014). On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research*, *52*(2), 271–292. https://doi.org/10.1287/opre.1030.0088

[2]. Atar, R., Mandelbaum, A., & Reiman, M. I. (2014). Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy traffic. *Annals of Applied Probability*, *14*(3), 1084–1134. https://doi.org/10.1214/105051604000000233

[3]. Baghel, K. P. S. (2013). Generalists vs. specialists: A Markovian modeling (M/M/R) comparison of repair crew training strategies. Journal of Research in Applied Mathematics, 1(1), 10–15.

[4]. Baghel, K. P. S. (2014). Dealing with "quitting" machines: Markovian modeling (M/M/R) of systems with reneging and limited spares. Invention Journals.

[5]. Baghel, K. P. S. (2017). Preventive vs. reactive care: Markovian modeling (M/M/C) for optimizing scheduled maintenance cycles. Invention Journals.

[6]. Baghel, K. P. S. (2018). Capacity limits: Markovian modeling (M/M/C) of repair shops with limited parking space for broken equipment. Journal of Research in Applied Mathematics, 4(2), 35–41.

[7]. Borst, S., Mandelbaum, A., & Reiman, M. I. (2014). Dimensioning large call centers. *Operations Research*, *52*(1), 17–34. https://doi.org/10.1287/opre.1030.0081

[8]. Buzacott, J. A., &Shanthikumar, J. G. (2013). *Stochastic models of manufacturing systems*. Prentice Hall.

[9]. Chen, H., & Yao, D. D. (2013). *Fundamentals of queueing networks: Performance, asymptotics, and optimization*. Springer.

[10]. Gamarnik, D., & Goldberg, D. A. (2013). Steady-state GI/G/n queue in the Halfin-Whitt regime. *Annals of Applied Probability*, *23*(6), 2382–2419. https://doi.org/10.1214/12-AAP905

[11]. Halfin, S., & Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research*, *29*(3), 567–588. https://doi.org/10.1287/opre.29.3.567

[12]. Iglehart, D. L., & Whitt, W. (2010). Multiple channel queues in heavy traffic: I. *Advances in Applied Probability*, *2*(1), 150–177. https://doi.org/10.1017/S0001867800037216

[13]. Jain, M., &Dhyani, I. (1999). Transient analysis of M/M/C machine repair problem with spare. Journal of Science, 2, 16–42.

[14]. Jain, M., Maheshwari, S., &Baghel, K. P. S. (2008). Queueing network modelling of flexible manufacturing system using mean value analysis. Applied Mathematical Modelling, 32(5), 700–711. https://doi.org/10.1016/j.apm.2007.02.003

[15]. Jain, M., Rakhee, M., & Singh, M. (2009). Bilevel control of degraded machining system with warm standbys, setup and vacation. *Applied Mathematical Modelling*, *28*(12), 1015–1026. https://doi.org/10.1016/j.apm.2009.02.003

[16]. Kingman, J. F. C. (2013). The single server queue in heavy traffic. *Mathematical Proceedings of the Cambridge Philosophical Society*, *57*(4), 902–904. https://doi.org/10.1017/S0305004100035793

[17]. Kumar, R., & Ramesh, S. (2013). Performance modeling of machine repair system with warm spares and server vacation. *International Journal of Advanced Manufacturing Technology*, *66*(5–8), 645–658. https://doi.org/10.1007/s00170-012-4345-6

[18]. Mandelbaum, A., &Stolyar, A. L. (2011). Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research*, *52*(6), 836–855. https://doi.org/10.1287/opre.1040.0152

[19]. Pender, J. (2014). Gram–Charlier expansions for time varying multiserver queues with abandonment. *SIAM Journal on Applied Mathematics*, *74*(4), 1238–1265. https://doi.org/10.1137/130938Sign

[20]. Puhalskii, A., & Reiman, M. I. (2012). The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability*, *32*(2), 564–595. https://doi.org/10.1017/S0001867800010053

[21]. Siegmund, D. (2009). Boundary crossing probabilities and statistical applications. *Annals of Statistics*, *14*(2), 361–404. https://doi.org/10.1214/aos/1176349928

[22]. Singh, C. J., & Kumar, R. (2014). Machine repair system with heterogeneous repairmen and reneging in a diffusion approximation framework. *Applied Mathematics and Computation*, *238*(1), 117–128. https://doi.org/10.1016/j.amc.2014.03.137

[23]. Wang, K. H., Yang, D. Y., & Liu, T. H. (2015). Optimal control of machine repair problem with vacations and two modes of failure. *Journal of Industrial and Production Engineering*, *32*(3), 163–175. https://doi.org/10.1080/21681015.2015.1015409

[24]. Whitt, W. (2012). Heavy-traffic limits for the G/H₂*/GI/n/m queue. *Mathematics of Operations Research*, *30*(1), 1–27. https://doi.org/10.1287/moor.1040.0115

[25]. Yom-Tov, G. B., & Mandelbaum, A. (2014). Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, *16*(2), 283–299. https://doi.org/10.1287/msom.2013.0474

[26]. Zhang, H., & Heng, Q. (2013). Optimal staffing for call centers with impatient customers under diffusion approximation. *European Journal of Operational Research*, *227*(3), 514–524. https://doi.org/10.1016/j.ejor.2012.12.029