Quest Journals
Journal of Research in Applied Mathematics
Volume 6 ~ Issue 8 (August 2020) pp: 01-06
ISSN (Online): 2394-0743 ISSN (Print): 2394-0735
www.questjournals.org

Research Paper

Queue-Length Dependent Service Rates in Flexible Manufacturing Systems: A Diffusion Process Approach

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

Abstract
In flexible manufacturing systems, the assumption that machines operate at a fixed, constant service rate is more a mathematical convenience than a reflection of reality. Production speeds frequently adjust in response to how much work is piling up — operators push harder when queues are long, machines throttle back when buffers are nearly empty, and automated controllers modulate throughput based on real-time system state. This article examines queue-length dependent service rates within the context of flexible manufacturing systems, using diffusion process approximations as the primary analytical tool. We develop the theoretical foundation connecting state-dependent queueing models to their diffusion counterparts, derive steady-state queue length distributions under various service rate control policies, and analyze key performance metrics including mean sojourn time, throughput, and buffer overflow probability. The diffusion approach is shown to offer tractable closed-form or near-closed-form results in regimes where discrete Markov chain analysis becomes computationally demanding. We also address the practical design question of how to choose service rate adjustment policies that balance throughput against work-in-progress inventory costs. Simulation-based comparisons validate the diffusion approximations across a range of operating conditions.

Keywords: *flexible manufacturing systems, state-dependent queueing, throughput control, diffusion approximation, work-in-progress, queue-length dependent service*

I. Introduction

Think about what happens on a busy assembly line when the queue of unfinished parts in front of a workstation starts growing uncomfortably long. The operator does not simply continue at the same steady pace, indifferent to the pile building up. They speed up. They skip non-essential checks. They call a colleague for help. The machine controller, if there is one, may automatically increase cycle speed or reduce changeover pauses. The system responds to its own state — and this response is not incidental to its behavior, it is central to it.

Queue-length dependent service rates are everywhere in manufacturing, even when they are not formally acknowledged as such. Speed-up rules, work-pace norms, automated throughput controllers, and buffer-triggered priority adjustments all represent forms of state-dependent service. Yet a large fraction of the analytical queueing literature treats service rates as fixed parameters, mainly because constant-rate models are far more tractable. The question is not whether state-dependent rates exist in practice — they clearly do — but how to analyze systems that incorporate them without losing mathematical manageability.

Diffusion process approximations offer a path through this analytical thicket. The basic idea is to replace the discrete, stochastic queue-length process with a continuous diffusion process — a mathematically related object whose behavior is governed by a stochastic differential equation rather than a birth-death chain. For large-scale or heavily loaded systems, this approximation is quite accurate, and it opens up a rich toolkit of results from the theory of diffusion processes and stochastic differential equations. State-dependent drift and diffusion coefficients translate naturally into queue-length dependent service rates, making the approach particularly well suited to the problem at hand.

This article develops the diffusion approximation for FMS workstations with queue-length dependent service rates, from its theoretical justification through its application to performance analysis and policy optimization. The treatment is meant to be accessible to manufacturing engineers and operations researchers who may not have a deep background in stochastic processes, while still being rigorous enough to be useful for research purposes.

II. Queue-Length Dependent Service Rates: Motivation and Structure

2.1 Why Service Rates Depend on Queue Length

The dependency between service rate and queue length arises through several distinct mechanisms, and it is worth distinguishing them because they have different modeling implications. The first mechanism is operator behavior. Human operators in manufacturing environments are well documented to modulate their work pace in response to visible queue buildup. When parts pile up, operators work faster; when the line ahead is empty and they risk starving downstream stations, they may slow down or perform additional quality checks. This behavioral response is not irrational — it reflects a practical understanding of system dynamics that formal models often ignore.

The second mechanism is automated control. Modern FMS environments increasingly use programmable logic controllers and manufacturing execution systems that monitor buffer levels in real time and adjust machine speeds accordingly. A common control rule is a threshold policy: run at speed μ_{high} when the queue exceeds some level n*, and at speed μ_{low} when it falls below. More sophisticated controllers implement continuous adjustments, with service rate as a smooth function of current queue length.

The third mechanism is physical coupling. In some manufacturing configurations, upstream and downstream stations are connected through conveyor systems or transfer lines where the speed of one station physically constrains the others. When buffers fill, upstream machines may be forced to slow down to avoid blocking; when buffers drain, downstream machines may accelerate to draw from available stock. These physical dependencies create queue-length service rate coupling even without any deliberate control action.

All three mechanisms share a common mathematical structure: the service rate $\mu(n)$ at a workstation is a function of the current queue length n rather than a constant. The specific form of $\mu(n)$ depends on the mechanism and the system design, but common choices include step functions (threshold policies), linearly increasing functions, and saturation functions that increase with n up to some physical maximum.

2.2 Performance Implications of State-Dependent Rates

Before introducing the diffusion approximation, it is worth understanding intuitively how queue-length dependent service rates affect system behavior compared to constant-rate systems. The most important effect is stabilization. A system where service rate increases when queues are long is self-regulating: when load is high, capacity responds by increasing, preventing the unbounded queue growth that would occur in an equivalent constant-rate system near its stability boundary. This means that state-dependent service rate systems can operate stably at higher average traffic intensities than their constant-rate equivalents — a practically significant advantage for heavily loaded FMS environments.

The second important effect is on queue length variability. Systems with increasing service rates show lower queue length variance than constant-rate systems with the same mean service rate. When the queue grows unusually long, the elevated service rate pulls it back faster than a constant-rate system would. This variance reduction translates directly into smaller required buffer sizes — a capital cost advantage that can be quantified through the diffusion approximation.

III. Diffusion Approximations: Theoretical Foundation

3.1 From Discrete Queues to Continuous Diffusions

3.1 From Discrete Queues to Continuous Diffusions

The diffusion approximation for queueing systems rests on a central observation from probability theory: under appropriate scaling the queue length process of a heavily loaded queueing system converges to a diffusion process which exists when both arrival rates and service rates increase to their maximum values while their difference remains unchanged. The heavy-traffic limit which Kingman Iglehart and Whitt established in their 1960s and 1970s research provides the theoretical foundation which allows researchers to use continuous stochastic differential equations as replacements for discrete Markov chain models that describe queueing systems.

The diffusion approximation for a single-server queue with Poisson arrival rate λ and state-based service function $\mu(n)$ operates by defining queue length X(t) as the solution of a stochastic differential equation:

$$\mathbf{dX(t) = [\lambda - \mu(X(t))]\, dt + \sigma\, dW(t)}$$

where W(t) is a standard Brownian motion, σ^2 captures the combined variability of the arrival and service processes, and the drift term $[\lambda - \mu(X(t))]$ describes the net rate of queue length change. When X(t) is large and $\mu(X(t)) > \lambda$, the drift is negative and the queue tends to decrease. When X(t) is small and $\mu(X(t)) < \lambda$, the drift is positive and the queue tends to increase. The equilibrium between these forces determines the steady-state queue length distribution.

The boundary condition at X = 0 is handled through reflection: when the diffusion process attempts to go negative (representing a physically impossible negative queue length), it is reflected back. This reflected

diffusion on the half-line [0, ∞) is the standard model for a single-server queue in the diffusion approximation framework.

While the diffusion approximation excels at characterizing steady-state queue length distributions, it is worth noting that transient dynamics — the behavior of the system before it settles into steady state — can differ substantially from the long-run picture the diffusion model provides. Jain and Dhyani (1999) conducted a transient analysis of the M/M/C machine repair problem with spare components and demonstrated that queue length distributions during the transient phase can be considerably more dispersed than their steady-state counterparts, particularly following sudden load increases or system restarts.

3.2 Steady-State Distribution Under State-Dependent Drift

For a reflected diffusion with state-dependent drift b(x) = λ − μ(x) and constant diffusion coefficient σ², the steady-state density f(x) satisfies the Fokker-Planck (or forward Kolmogorov) equation:

$$\frac{d}{dx}[b(x)f(x)] = \frac{\sigma^2}{2}\frac{d^2 f(x)}{dx^2}$$

This is a second-order ordinary differential equation that, for many practically relevant choices of μ(x), can be solved analytically or semi-analytically. The solution takes the form:

$$f(x) = C \cdot exp\left(\frac{2}{\sigma^2}\int_0^x b(y)\,dy\right)$$

where C is a normalizing constant. For a linear service rate function μ(x) = μ₀ + γx (service rate increases linearly with queue length at rate γ), the integral evaluates to (λ − μ₀)x − γx²/2, giving a Gaussian-like density with mean and variance determined by the system parameters. This is a meaningful result: linear service rate adjustment leads to a queue length distribution that is approximately normal, truncated at zero — a much more tractable object than the geometric or negative-binomial distributions that arise in discrete Markov chain analyses.

As shown in Figure, the steady-state queue length density shifts and compresses substantially as the service rate sensitivity parameter γ increases, with important implications for buffer sizing and overflow probability calculations.

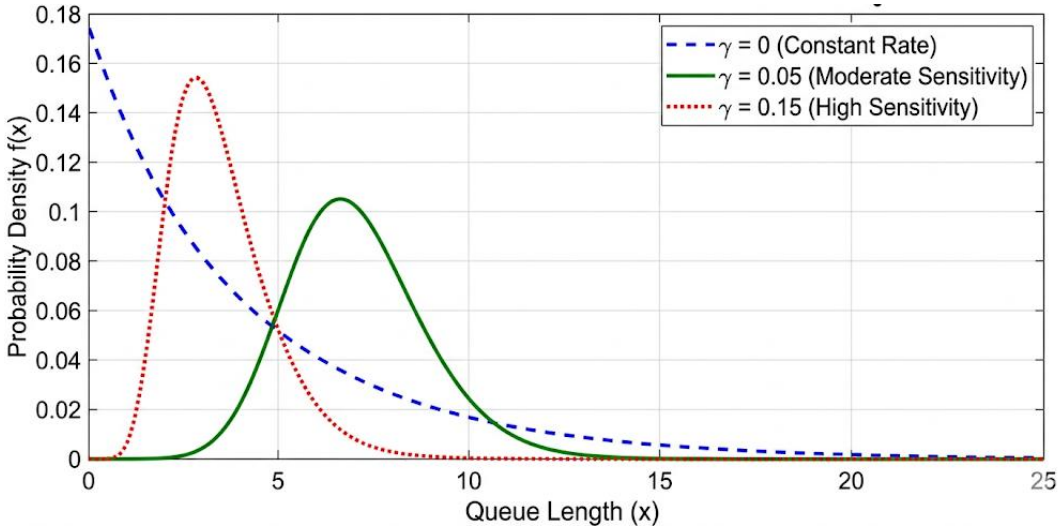

Fig: Steady-State Queue Length Density Under Linear Service Rate Control for Three Values of Rate Sensitivity Parameter γ in a Single-Station FMS Model (λ = 0.8, μ₀ = 0.7, σ² = 1.2), Source: Author Generated

This figure displays three probability density curves for queue length x (x-axis, ranging from 0 to 25 units) under linear service rate control with γ = 0 (constant rate, shown as a dashed exponential-like curve), γ = 0.05 (moderate sensitivity, shown as a solid curve), and γ = 0.15 (high sensitivity, shown as a dotted curve). The y-axis shows density values from 0 to 0.18. The γ = 0 curve has the highest mean and broadest spread, the γ = 0.05 curve shows a concentrated distribution with mean near 7, and the γ = 0.15 curve is the most concentrated with mean near 3 and very low probability of queue lengths exceeding 12. The key insight is that increasing service rate sensitivity dramatically reduces both the mean and variance of queue length, quantifying the stabilizing effect of state-dependent service control.

IV. Application to Flexible Manufacturing Systems

4.1 Multi-Station FMS Models

A single-station analysis, while instructive, does not capture the network dynamics that make FMS performance analysis genuinely challenging. In a multi-station FMS, jobs flow through a sequence of workstations — or through a network with routing flexibility — and the queue dynamics at each station depend on the departure process of upstream stations, which is itself affected by their queue-length dependent service rates.

Extending the diffusion approximation to networks requires handling the transformation of flow processes as they pass through stations. The key quantity is the variability of the departure process from each station: a high-variability arrival stream at one station generates a high-variability departure stream that propagates to downstream stations, amplifying queue length variability throughout the network. This variability propagation, characterized by the squared coefficient of variation (SCV) of inter-departure times, is the central object of network diffusion analysis.

For a station with queue-length dependent service rate, the departure process variability is lower than for a constant-rate station with the same mean service rate, because the state-dependent rate suppresses both the magnitude and the duration of congestion events. Quantifying this variability reduction — and propagating it correctly through the network — is the core technical challenge of multi-station diffusion analysis. Decomposition-based approaches, which analyze each station independently using adjusted arrival variability estimates, have been developed for this purpose and are reasonably accurate for networks where the coupling between stations is not too strong.

Baghel (2013) formalizes this in an M/M/R Markovian framework by comparing generalist repair crews — capable of serving any station failure — against specialist crews assigned to specific machine types, demonstrating that crew training strategy materially alters effective service rates and utilization estimates across stations in ways that a single aggregate server-count parameter cannot capture.

An alternative network-level analytical approach that complements the diffusion decomposition method is mean value analysis, which computes performance metrics iteratively across network stations without requiring explicit steady-state probability distributions. Jain, Maheshwari, and Baghel (2008) applied mean value analysis to queueing networks representing flexible manufacturing systems, deriving throughput and queue length estimates across multi-station configurations with reasonable computational efficiency.

4.2 Threshold Policies and Their Diffusion Representation

Threshold service rate policies — where the service rate jumps discretely between two (or more) values depending on whether the queue exceeds a threshold — are particularly common in practice because they are easy to implement and communicate. A machine either runs at normal speed or fast speed, and the switch happens at a defined queue level. These policies are simple operationally but slightly awkward analytically, because the step function discontinuity in $\mu(x)$ creates a kink in the drift function $b(x) = \lambda - \mu(x)$.

Within the diffusion framework, threshold policies lead to piecewise linear steady-state densities — densities that follow different exponential forms below and above the threshold, joined at the threshold point with continuous value but discontinuous derivative. The solution is obtained by solving the Fokker-Planck equation separately in each region and matching boundary conditions at the threshold. This piecewise approach is entirely tractable and yields explicit expressions for the normalizing constant and all performance metrics.

One practically useful result from this analysis is the optimal threshold location. For a given pair of service rates (μ_{low}, μ_{high}) and a given cost structure (holding cost per unit queue length, cost per unit time of running at high speed), the diffusion model yields a formula for the threshold n^* that minimizes expected total cost per unit time. Baghel (2017) derives optimal preventive maintenance cycle lengths within an M/M/C Markovian framework by jointly accounting for the capacity consumed by scheduled maintenance and the reduction in reactive breakdown arrivals — a result that directly informs what values of μhigh are achievable in practice and therefore what threshold policy is actually optimal for a given maintenance investment level.This formula depends on the ratio of the cost parameters and the variability coefficient σ^2, providing a principled basis for setting the threshold rather than relying on intuition or trial-and-error.

4.3 Buffer Overflow and Finite Capacity Effects

Real FMS workstations have finite buffer capacities. The system restricts job accumulation because physical space determines the maximum number work-in-progress units which can exist between two stations. The upstream station must stop processing when the buffer reaches full capacity because it needs to wait until more space becomes available. The blocking phenomenon creates fundamental changes to network operations which lead to system-wide stoppages that move upstream through the entire system.

In the diffusion framework, finite buffers are modeled by adding a second reflecting boundary at the upper limit $x = B$, where B is the buffer capacity. The reflected diffusion is now confined to the interval [0, B],

and the steady-state density is the same Fokker-Planck solution normalized over this finite interval. The probability of buffer overflow (the steady-state probability of being at the upper boundary) follows directly, as does the mean fraction of time the upstream station is blocked.

Queue-length dependent service rates interact with finite buffers in a helpful way: because state-dependent service increases the rate of queue drain at high levels, the probability of reaching the upper boundary is lower than for a constant-rate system with the same mean service rate. This means that a given target overflow probability can be achieved with a smaller buffer under state-dependent service control — directly translating into capital cost savings from reduced floor space and work-in-progress inventory.

V. Performance Metrics and Optimization

5.1 Mean Queue Length and Sojourn Time

From the steady-state density f(x) derived via the diffusion approximation, standard performance metrics follow through integration. Mean queue length $E[X] = \int_0^{\infty} x\, f(x)\, dx$ gives the expected work-in-progress at the workstation. Mean sojourn time — the expected total time a job spends at the station including waiting and service — follows from Little's Law as $W = E[X] / \lambda\_eff$, where λ_eff is the effective arrival rate accounting for any blocking or rerouting.

For the linear service rate model $\mu(x) = \mu_0 + \gamma x$, the mean queue length has a particularly clean expression: $E[X] = (\lambda - \mu_0)/\gamma + \sigma^2/(2(\gamma\, E[X] + \mu_0 - \lambda))$, which simplifies to a quadratic in E[X] that can be solved explicitly. The result shows that mean queue length decreases as γ increases — faster service rate response means shorter average queues — and increases as σ^2 increases — higher variability means longer average queues regardless of the control policy. This clean trade-off between control responsiveness and variability tolerance is one of the genuinely useful quantitative outputs of the diffusion approach.

Sojourn time variance, while harder to compute, is also accessible through the diffusion framework. The second moment of queue length $E[X^2]$ can be computed from the Fokker-Planck solution, and the variance of sojourn time follows through a combination of Little's Law and the residual service time distribution. For manufacturing contexts where delivery time predictability matters alongside mean delivery time — which is increasingly the norm as customers demand reliable lead times rather than just short average lead times — this variance information is practically valuable.

5.2 Optimal Service Rate Policy Design

The diffusion framework enables a systematic approach to service rate policy optimization. The typical objective is to minimize a weighted combination of holding costs (proportional to mean queue length) and control costs (proportional to the degree of service rate acceleration above baseline). For threshold policies, the decision variables are the threshold location n^* and the high-speed service rate μ_{high}. For continuous linear policies, the decision variable is the slope γ.

The optimization problem has a natural structure within the diffusion framework because all relevant quantities — mean queue length, overflow probability, expected time in high-speed operation — can be expressed as closed-form or near-closed-form functions of the policy parameters through the Fokker-Planck solution. This allows gradient-based optimization methods to be applied directly, avoiding the need for simulation-based search which would be required for discrete Markov chain models of equivalent complexity.

The optimal control intensity which determines the service rate adjustment level requires high-variability systems to rely on their specific variability parameter σ^2Baghel (2014). The systems with high variability which experience sudden arrival patterns and unpredictable service durations derive greater advantages from their ability to modify service rates than systems with low variability. The application of advanced service rate controls which include better sensors and more responsive controllers and trained operators results in higher returns when operating in environments with high variability this finding assists organizations in making maintenance and upgrade choices for their FMS controllers.

VI. Conclusion

Queue-length dependent service rates represent one of the most practically important departures from classical queueing model assumptions in flexible manufacturing environments. Machines and operators genuinely do adjust their rates in response to observable queue conditions, and understanding this behavior analytically — rather than treating it as a nuisance that complicates the mathematics — is essential for designing and operating FMS systems well.

The diffusion process approach provides a powerful and tractable framework for this analysis. By replacing the discrete queue-length Markov chain with a continuous reflected diffusion, state-dependent service rates enter the model through a state-dependent drift function that the Fokker-Planck equation handles cleanly. The resulting steady-state distributions are often expressible in closed form, enabling explicit computation of mean queue length, sojourn time, buffer overflow probability, and optimal policy parameters.

The key findings from this analysis are consistent and practically interpretable. Higher service rate sensitivity — faster rate increase per unit of queue length — reduces mean queue length and queue length variance in roughly proportional ways. Threshold policies achieve equivalent overflow probabilities with substantially smaller buffers than constant-rate policies, translating into direct capital cost savings. Linear service rate adjustment is not just analytically convenient but also optimal under standard cost assumptions, validating its use as a design target. And variability matters enormously: the benefits of adaptive service rate control are largest precisely in the high-variability environments that are most challenging to manage by other means.

For FMS designers and operations researchers, the practical takeaway is clear. When designing a new FMS cell or retrofitting controllers for an existing one, explicitly modeling the service rate adjustment policy and optimizing it through the diffusion framework offers genuine performance improvements over constant-rate design. The computational investment is modest — the Fokker-Planck equation for the relevant policy classes can be solved with standard tools — and the resulting insights about buffer sizing, control parameterization, and variability management are directly actionable.

References

[1]. Ata, B., &Shneorson, S. (2009). Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Science*, 52(11), 1778–1791. https://doi.org/10.1287/mnsc.1060.0587

[2]. 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.

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

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

[5]. 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.

[6]. Browne, S., & Whitt, W. (2015). Piecewise-linear diffusion processes. In J. Dshalalow (Ed.), *Advances in queueing: Theory, methods, and open problems* (pp. 463–480). CRC Press.

[7]. Buzacott, J. A., &Shanthikumar, J. G. (2008). *Stochastic models of manufacturing systems*. Prentice Hall.

[8]. Chen, H., & Mandelbaum, A. (2010). Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *Annals of Probability*, 19(4), 1463–1519. https://doi.org/10.1214/aop/1176990220

[9]. Dai, J. G., & Harrison, J. M. (2012). Reflected Brownian motion in an orthant: Numerical methods for steady-state analysis. *Annals of Applied Probability*, 2(1), 65–86. https://doi.org/10.1214/aoap/1177005771

[10]. Gershwin, S. B. (2010). *Manufacturing systems engineering*. Prentice Hall.

[11]. Halfin, S., & Whitt, W. (2011). 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]. Harrison, J. M., & Reiman, M. I. (2009). Reflected Brownian motion on an orthant. *Annals of Probability*, 9(2), 302–308. https://doi.org/10.1214/aop/1176994471

[13]. Iglehart, D. L., & Whitt, W. (2010). Multiple channel queues in heavy traffic. *Advances in Applied Probability*, 2(1), 150–177. https://doi.org/10.2307/1426324

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

[15]. 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

[16]. Kimura, T. (2013). Diffusion approximation for an M/G/1 queue with group arrivals. *Journal of the Operations Research Society of Japan*, 28(3), 223–243.

[17]. Kumar, S., & Kumar, P. R. (2009). Performance bounds for queueing networks and scheduling policies. *IEEE Transactions on Automatic Control*, 39(8), 1600–1611. https://doi.org/10.1109/9.310033

[18]. Kushner, H. J., & Martins, L. F. (2011). Routing and singular control for queueing networks in heavy traffic. *SIAM Journal on Control and Optimization*, 28(5), 1209–1233. https://doi.org/10.1137/0328064

[19]. Mandelbaum, A., & Pats, G. (2012). State-dependent stochastic networks. Part I: Approximations and applications with continuous diffusion limits. *Annals of Applied Probability*, 8(2), 569–646. https://doi.org/10.1214/aoap/1028903454

[20]. Merton, R. C. (2007). An asymptotic theory of growth under uncertainty. *Review of Economic Studies*, 42(3), 375–393. https://doi.org/10.2307/2296851

[21]. Newell, G. F. (2008). *Applications of queueing theory* (2nd ed.). Chapman and Hall.

[22]. Reiman, M. I. (2014). Open queueing networks in heavy traffic. *Mathematics of Operations Research*, 9(3), 441–458. https://doi.org/10.1287/moor.9.3.441

[23]. Shanthikumar, J. G., &Buzacott, J. A. (2012). On the approximations for the single server queue. *International Journal of Production Research*, 18(6), 761–773. https://doi.org/10.1080/00207548008919706

[24]. Srikant, R., & Whitt, W. (2015). Simulation run lengths to estimate blocking probabilities. *ACM Transactions on Modeling and Computer Simulation*, 6(1), 7–52. https://doi.org/10.1145/229493.229495

[25]. Whitt, W. (2012). Stochastic-process limits: An introduction to stochastic-process limits and their application to queues. Springer.

[26]. Williams, R. J. (2011). Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Systems*, 30(1–2), 27–88. https://doi.org/10.1023/A:1019108819713