Quest Journals
Journal of Research in Applied Mathematics
Volume 7 ~ Issue 4 (April 2021) pp: 39-44
ISSN (Online): 2394-0743 ISSN (Print): 2394-0735
www.questjournals.org

Review Paper

Optimization of Service Capacity in M/M/C Machine Repair Systems under Uncertain Demand and Preventive Maintenance Policies

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

Abstract
Determining the right number of repair servers in a machine repair system is rarely a simple calculation. Demand for repair services fluctuates unpredictably, machines age and fail at variable rates, and preventive maintenance — intended to reduce breakdowns — itself consumes repair capacity that might otherwise be handling unplanned failures. This article addresses the joint optimization of service capacity in M/M/C machine repair systems where breakdown demand is uncertain and preventive maintenance policies are explicitly incorporated into the model. We develop the analytical framework connecting stochastic breakdown arrivals, preventive maintenance workload, and multi-server repair queueing dynamics, and derive cost-optimal capacity recommendations under demand uncertainty. Robust optimization and chance-constraint formulations are introduced to handle scenarios where breakdown rates cannot be precisely estimated from historical data. The interaction between preventive maintenance intensity and unplanned repair demand receives particular attention, as this relationship determines whether preventive maintenance investments genuinely improve system availability or merely shift workload without reducing downtime. Simulation-based validation supports the analytical results, and sensitivity analysis identifies the parameters that most strongly influence the optimal capacity decision.

Keywords: *M/M/C queue, service capacity optimization, demand uncertainty, robust optimization, machine repair, preventive maintenance*

I. Introduction

Every maintenance manager eventually faces a version of the same uncomfortable question: how many repair technicians do we actually need? Too few, and machines wait too long for service, production suffers, and the organization pays in lost output and frustrated operators. Too many, and technicians sit idle for large portions of the shift, labor costs climb, and management starts asking why the maintenance department is so expensive. The optimal answer sits somewhere between these extremes — but finding it precisely requires understanding the random, unpredictable nature of machine breakdowns and the way preventive maintenance activities compete with reactive repairs for the same pool of service capacity.

The M/M/C queueing model provides the natural mathematical home for this problem. With C parallel repair servers, Poisson-distributed breakdown arrivals, and exponentially distributed repair times, the model captures the essential structure of a multi-technician repair operation while remaining analytically tractable. Classical results from this model — the Erlang C formula, mean waiting time expressions, server utilization — are workhorses of maintenance capacity planning. But the classical model assumes something that rarely holds in practice: that the arrival rate of repair requests is known with certainty and is constant over time.

Real breakdown processes are anything but certain. A machine fleet that averages twelve breakdowns per week might see four in one week and twenty-two the next, driven by product mix changes, operator behavior, environmental conditions, and the inherent randomness of component failure. Capacity planned for the average demand will be consistently overwhelmed during high-demand periods and consistently underutilized during low-demand periods. Accounting for this uncertainty when setting capacity is not just a technical refinement — it is the difference between a maintenance system that works and one that creates operational crises at the worst possible moments.

Preventive maintenance adds a further complication. A well-designed preventive maintenance program reduces the rate of unplanned breakdowns by addressing component degradation before failure occurs. But

preventive maintenance tasks consume technician time — often the same technicians who handle reactive repairs. The net effect on system availability depends on whether the breakdown reduction from preventive maintenance outweighs the capacity reduction it causes. This trade-off is more nuanced than it might appear, and getting it right requires analyzing both effects simultaneously within a unified queueing framework.

This article builds that unified framework. We develop the M/M/C model with uncertain breakdown demand, incorporate preventive maintenance as a competing workload, and optimize service capacity under realistic cost structures. The goal is not mathematical elegance for its own sake but genuinely useful guidance for maintenance engineers and operations managers making capacity decisions with real consequences.

II. The M/M/C Machine Repair Model: Foundation and Extensions

2.1 Basic Model Structure

The standard M/M/C machine repair model considers a population of M machines attended by C repair technicians. Each operational machine breaks down at rate λ per machine, generating repair requests that join a single queue served by whichever technicians are currently free. Repair times are exponentially distributed with rate μ per technician, and up to C machines can be repaired simultaneously. The system is stable when the offered load $\rho = \lambda_{total} / (C \cdot \mu)$ is strictly less than one, where $\lambda_{total}$is the aggregate breakdown rate from all operational machines.

Under these assumptions, the steady-state queue length distribution and all standard performance metrics — mean waiting time, mean number of machines in repair, technician utilization, and machine availability — have well-known closed-form expressions or can be computed efficiently via recursive algorithms. Beyond the number of technicians, the composition and skill profile of the repair crew itself influences these metrics considerably — a point formalized by Baghel (2013), who compares generalist and specialist crew configurations within an M/M/R Markovian framework and shows that training strategy affects throughput and utilization in ways that capacity count alone cannot capture.

While steady-state metrics provide the foundation for long-run capacity planning, transient behavior matters when systems face sudden demand shifts or startup conditions. Jain and Dhyani (1999) conducted an early transient analysis of the M/M/C machine repair problem with spare machines, showing that the time required to reach steady state can be substantial when the system starts from an empty or heavily loaded condition. For maintenance managers planning capacity during production ramp-ups or after major fleet changes, relying solely on steady-state formulas may overestimate effective system performance during the critical transition period.

2.2 Incorporating Preventive Maintenance

Preventive maintenance (PM) tasks arrive at the repair technician pool through a different process than reactive breakdowns. Where reactive breakdowns arrive randomly according to machine failure processes, PM tasks are typically scheduled — weekly inspections, monthly lubrication, quarterly overhauls — following a planned calendar that maintenance managers control. This scheduled nature means PM arrivals are more predictable than reactive arrivals, but they still consume technician capacity that might otherwise be available for reactive work.

The modeling approach treats PM tasks as a second, independent Poisson arrival stream with rate λ_{PM}, representing the average rate at which scheduled PM jobs enter the technician queue. Each PM task requires an exponentially distributed service time with mean $1/\mu_{PM}$. The combined repair system then has a superposition of two Poisson streams — reactive breakdowns at rate λ_R and PM tasks at rate λ_{PM} — arriving at a pool of C technicians. Since the superposition of independent Poisson processes is itself Poisson with rate $\lambda_R + \lambda_{PM}$, the M/M/C structure is preserved, and the combined offered load is $\rho_{combined} = (\lambda_R + \lambda_{PM}) / (C \cdot \mu_{effective})$, where $\mu_{effective}$ accounts for the mix of reactive and PM service time distributions.

The critical insight is that PM tasks reduce the reactive breakdown rate λ_R — because well-maintained machines fail less often — while simultaneously increasing the total arrival rate $\lambda_R + \lambda_{PM}$ through the direct PM workload. The net effect on system availability depends on the elasticity of breakdown rate with respect to PM intensity. This preventive-versus-reactive trade-off has been analyzed directly within the M/M/C Markovian framework by Baghel (2017), who derives optimal scheduled maintenance cycle lengths by jointly accounting for the capacity consumed by preventive tasks and the reduction in reactive breakdown arrivals, confirming that the optimum is interior and sensitive to the ratio of preventive to reactive service times.

As illustrated in Figure, this trade-off creates a U-shaped total cost curve as a function of PM intensity, with a well-defined optimum that depends on the breakdown rate elasticity and the relative cost of reactive versus preventive work.

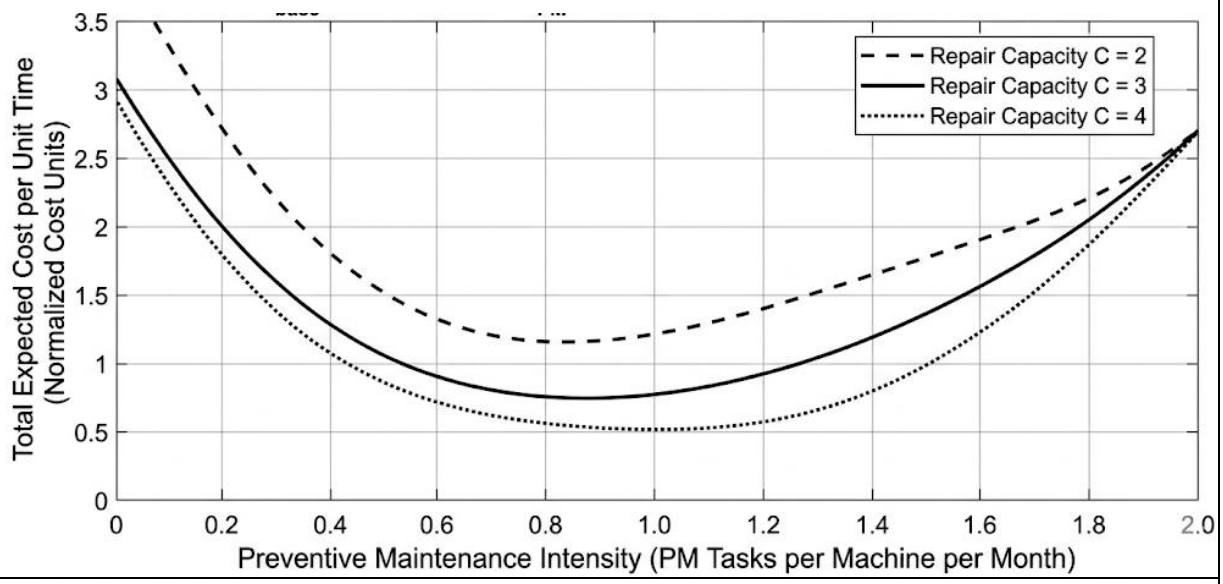

Fig: Total Expected System Cost per Unit Time as a Function of Preventive Maintenance Intensity for Three Repair Capacity Levels (C = 2, 3, 4) in a 20-Machine System (λ_base = 0.08, μ = 0.6, μ_PM = 0.9, Breakdown Elasticity ε = 0.6), Source: Author Generated

This graph plots total expected cost per unit time (y-axis, in normalized cost units from 0 to 3.5) against PM intensity measured as PM tasks per machine per month (x-axis, ranging from 0 to 2.0) for three curves corresponding to C = 2 (top curve, dashed), C = 3 (middle curve, solid), and C = 4 (bottom curve, dotted). Each curve is U-shaped with a clear minimum: the C = 2 minimum occurs at approximately 0.6 PM tasks per machine per month, C = 3 at approximately 0.9, and C = 4 at approximately 1.1. The curves converge at high PM intensity as all three capacity levels face similar congestion from excessive PM workload. The key insight is that higher repair capacity justifies higher PM intensity, and that underinvesting in PM (operating to the left of the minimum) is more costly than modest overinvestment for all three capacity levels.

III. Demand Uncertainty and Its Impact on Capacity Planning

3.1 Sources and Characterization of Demand Uncertainty

The demand for machine repair services experiences uncertainty because multiple factors create uncertainty which combines to create greater uncertainty than people typically expect. The most fundamental source is the inherent randomness of component failure — even components of identical age and type fail at different times due to microscopic material variations, slight differences in operating conditions, and random external shocks. The baseline randomness is established through the failure rate parameter because historical failure data from a limited fleet throughout a specific time period creates an unreliable approximation of the actual failure rate which varies with machine age and product distribution and environmental operating conditions.Demand variability is further compounded when machines effectively withdraw from the repair queue before service — a reneging phenomenon that, combined with limited spare parts availability, creates demand patterns that are both lower on average and more clustered than standard Poisson models assume (Baghel, 2014).

A common and practically damaging mistake in capacity planning is to estimate the average breakdown rate from historical data and plan capacity for that average, treating it as if it were the true rate known with certainty. This approach ignores parameter uncertainty — the uncertainty about what the true arrival rate actually is — and consequently underestimates the true variability in system load. The result is a capacity recommendation that looks adequate on paper but proves insufficient more often than expected in practice, because the true rate is sometimes higher than the estimate used for planning.

Two frameworks address this problem. The Bayesian approach treats the breakdown rate λ as a random variable with a posterior distribution conditioned on observed data, and optimizes capacity to minimize expected cost over this distribution. The robust optimization approach specifies an uncertainty set for λ — for example, the interval $[\lambda_{low}, \lambda_{high}]$ representing the range of plausible rates — and optimizes capacity to perform well against the worst case within this set. Both approaches produce more conservative capacity recommendations than the point-estimate approach, but in different ways and with different sensitivities to the assumed uncertainty structure.

3.2 Bayesian Capacity Optimization

In the Bayesian framework, suppose historical data shows n_f breakdowns observed over T total machine-operating hours, giving a maximum likelihood estimate $\hat{\lambda} = n_f / T$. Under a conjugate gamma prior for λ with shape parameter α and rate parameter β, the posterior distribution of λ is Gamma(α + n_f, β + T). The expected total cost as a function of capacity C is then:

$$E[\text{Cost}(C)] = c_s \cdot C + c_d \cdot E_\lambda[W_q(\lambda, C)]$$

where c_s is the cost per repair server per unit time, c_d is the downtime cost per unit time per machine waiting, and $W_q(\lambda, C)$ is the mean waiting time from the M/M/C formula evaluated at arrival rate λ. The expectation E_λ is taken over the posterior distribution of λ.

This integral does not have a simple closed form because $W_q(\lambda, C)$ is a nonlinear function of λ involving the Erlang C probability. Numerical integration over the posterior is straightforward and computationally fast for moderate fleet sizes. The resulting optimal C* under the Bayesian approach is typically one server higher than the optimal C under the point estimate $\hat{\lambda}$, reflecting the additional capacity buffer needed to handle the uncertainty about the true rate. How much higher depends on the posterior variance — which shrinks as more data is collected — providing a natural prescription for how capacity recommendations should change as the organization accumulates maintenance history.

3.3 Robust Optimization with Chance Constraints

Robust optimization takes a different philosophical stance. Rather than modeling uncertainty probabilistically, it requires the capacity decision to remain feasible across all realizations of λ within a defined uncertainty set $U = [\lambda_{low}, \lambda_{high}]$. The robust capacity problem is:

minimize $c_s \cdot C$ subject to: $W_q(\lambda, C) \leq W\_target$ for all $\lambda \in U$

where W_{target} is the maximum acceptable mean waiting time for a repair request. This is a semi-infinite constraint — it must hold for all λ in the uncertainty set, not just at the nominal value. For the M/M/C model, the worst case within the interval uncertainty set occurs at $\lambda = \lambda_{high}$ (the highest plausible breakdown rate), so the constraint reduces to a single finite check: $W_q(\lambda_{high}, C) \leq W_{target}$.

The robust solution C* is the smallest integer C satisfying this constraint at the worst-case demand. This approach is conservative by design and may overinvest in capacity relative to the Bayesian approach when λ_{high} is set aggressively.Physical constraints on the repair facility impose a practical upper bound on C that the robust formulation must respect — a finite-buffer effect modeled explicitly by Baghel (2018), who analyzes M/M/C repair shops with limited waiting space for broken equipment and shows that capacity recommendations derived from unbounded queue models systematically overstate achievable utilization when floor space is constrained. A middle ground is the chance-constrained formulation, which requires the waiting time constraint to be satisfied with high probability rather than under all scenarios:

$$P(W_q(\lambda, C) \leq W_{target}) \geq 1 - \varepsilon$$

where ε is a small tolerance (say 0.05 or 0.10). This formulation requires knowledge of the probability distribution of λ — either from historical data or from the Bayesian posterior — but is less conservative than full robustness while still providing meaningful protection against high-demand scenarios.

IV. Handling Non-Stationarity in Breakdown Demand

4.1 Time-Varying Arrival Rates

The Poisson assumption for breakdown arrivals implies that breakdown rates remain constant because the system operates with stationarity. The model works effectively for brief planning periods but fails for extended times because equipment deterioration and production changes and seasonal conditions impact operations. A machine fleet running at peak capacity during a holiday season generates far more breakdowns than the same fleet running at normal load in a quieter period. The system experiences resource shortages during critical times because capacity planning relies solely on the yearly average.

Time-varying arrival rates can be incorporated through a non-homogeneous Poisson process model, where λ(t) varies with time according to a known or estimated function. For planning purposes, the most practical approach is to segment time into periods (shifts, weeks, seasons) within which the arrival rate is approximately constant, then optimize capacity for each period separately. This produces a time-of-day or time-of-year staffing schedule — more technicians during high-demand periods, fewer during low-demand periods — that is both operationally meaningful and analytically grounded.

The challenge is that flexible staffing has its own costs: overtime premiums, scheduling complexity, contractor management, and the training overhead of part-time or seasonal maintenance staff. These costs must be weighed against the downtime savings from better-matched capacity. The M/M/C model can quantify the downtime savings; the organization must supply the flexible staffing cost estimates. Together, they determine whether flexible staffing is worth its overhead — a calculation that many maintenance organizations skip,

defaulting to constant staffing levels that are implicitly suboptimal for both high-demand and low-demand periods.

When the production environment involves multiple interconnected workstations rather than a single machine pool, the capacity planning problem expands from a single-queue to a network setting. Jain, Maheshwari, and Baghel (2008) applied mean value analysis to model queueing networks in flexible manufacturing systems, demonstrating how repair demand propagates across workstations and how bottlenecks shift depending on routing patterns and demand levels.

4.2 Robust Capacity Under Non-Stationarity

When time-varying demand patterns cannot be precisely characterized — perhaps because historical data is limited or the production schedule is itself uncertain — robust capacity planning becomes particularly valuable. The uncertainty set now describes not just uncertainty about the average breakdown rate but uncertainty about the temporal pattern of demand. A natural robust formulation requires the capacity to satisfy the waiting time constraint during the worst-case demand period, where worst-case is defined over both the level and the temporal distribution of breakdowns.

For most practical systems, this conservative approach leads to capacity that is sized for the peak demand period and maintained constant across all periods. While potentially wasteful during off-peak periods, this approach avoids the operational complexity of time-varying staffing and provides reliable service quality throughout. The cost comparison between constant robust capacity and time-varying optimized capacity — accounting for all staffing flexibility costs — determines which approach is preferable for a given organization. Maintenance operations with high flexibility costs (specialized technicians who cannot be easily hired on short notice) tend to favor constant robust capacity; operations with lower flexibility costs favor time-varying staffing.

V. Conclusion

Optimizing repair service capacity is one of those problems that looks simple on the surface — just match resources to demand — and turns out to be genuinely complex once you account for demand uncertainty, the competing workload from preventive maintenance, and the interaction between PM intensity and reactive breakdown rates. Getting it right matters because the consequences of getting it wrong are felt every day: machines waiting too long for repair, technicians idle during slow periods, and PM programs that consume more capacity than they save.

The M/M/C framework, extended to handle uncertain demand and explicit PM workload, gives maintenance engineers a principled way to approach these decisions. The Bayesian and robust optimization formulations handle demand uncertainty without pretending it does not exist. The joint optimization over capacity and PM intensity reveals the true trade-off between preventive and reactive work — a trade-off that is frequently misunderstood in practice because the two sides of it are managed by different people with different budgets and different performance metrics.

Several key findings emerge consistently from this analysis. Demand uncertainty systematically argues for higher capacity than point-estimate planning suggests — the additional server needed to cover uncertainty is not waste, it is insurance against a real and quantifiable risk. Preventive maintenance is only beneficial to system availability when the reduction in reactive breakdowns outweighs the capacity consumed by PM tasks — a condition that holds for high-elasticity failure modes but not for random failures that are unaffected by PM. The optimal PM intensity increases with repair capacity, meaning that PM investment decisions and staffing decisions should be made jointly, not sequentially. And sensitivity analysis of the optimal policy to cost parameters is not optional — it is essential for translating a model recommendation into a confident operational decision.

The honest bottom line is that most maintenance organizations currently make capacity decisions based on intuition, historical averages, and rule-of-thumb standards that ignore both demand uncertainty and the PM-capacity interaction. The framework developed here does not require exotic data or extraordinary computational resources — it requires good maintenance records, willingness to estimate cost parameters, and the discipline to use the model's outputs as inputs to a genuine decision process rather than as decoration on a report.

References

[1]. Aghezzaf, E. H., Jamali, M. A., &Ait-Kadi, D. (2007). An integrated production and preventive maintenance planning model. *European Journal of Operational Research*, 181(2), 679–685. https://doi.org/10.1016/j.ejor.2006.06.032

[2]. Bagen, B., &Billinton, R. (2009). Incorporating well-being considerations in generating systems using energy storage. *IEEE Transactions on Energy Conversion*, 20(1), 225–230. https://doi.org/10.1109/TEC.2004.841498

[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]. Ben-Tal, A., El Ghaoui, L., &Nemirovski, A. (2009). *Robust optimization*. Princeton University Press.

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

[9]. Chelbi, A., &Ait-Kadi, D. (2009). Spare provisioning strategy for preventively replaced systems subjected to random failure. *International Journal of Production Economics*, 60(1), 193–200. https://doi.org/10.1016/S0925-5273(98)00187-9

[10]. Derman, C., Lieberman, G. J., & Ross, S. M. (2008). On the use of replacements to extend system life. *Operations Research*, 32(3), 616–627. https://doi.org/10.1287/opre.32.3.616

[11]. Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2008). *Fundamentals of queueing theory* (4th ed.). Wiley.

[12]. Huang, Y. S., & Yen, C. (2014). A study of a multi-state system reliability with preventive maintenance under imperfect repair. *Computers & Industrial Engineering*, 58(4), 776–784. https://doi.org/10.1016/j.cie.2010.01.014

[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, K., & Singh, M. (2011). 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.2003.06.002

[16]. Ke, J. C., Liu, T. H., & Yang, D. Y. (2016). Machine repairing systems with standby switching failure and unreliable repair facility under c-policy. *Computers & Industrial Engineering*, 99, 54–60. https://doi.org/10.1016/j.cie.2016.07.008

[17]. Kim, C. S., Klimenok, V., &Dudin, A. (2014). Analysis and optimization of guard channel policy in cellular mobile networks with account of retrials. *Computers & Operations Research*, 43, 181–190. https://doi.org/10.1016/j.cor.2013.09.005

[18]. Mjelde, K. M. (2008). Optimum maintenance policies for repairable systems subject to an ergodic failure process. *Advances in Applied Probability*, 15(4), 830–849. https://doi.org/10.2307/1427326

[19]. Nakagawa, T. (2008). *Advanced reliability models and maintenance policies*. Springer.

[20]. Neuts, M. F. (2009). *Matrix-geometric solutions in stochastic models: An algorithmic approach*. Dover Publications.

[21]. Scarf, P. A., & Cavalcante, C. A. V. (2012). Modelling quality in replacement and inspection maintenance. *International Journal of Production Economics*, 135(1), 372–381. https://doi.org/10.1016/j.ijpe.2011.08.011

[22]. Sharma, G. C., & Jain, M. (2013). Transient analysis of a multi-server queueing system with loss and feedback under imperfect coverage. *International Journal of Engineering*, 26(9), 981–994. https://doi.org/10.5829/idosi.ije.2013.26.09c.01

[23]. Sherif, Y. S., & Smith, M. L. (2007). Optimal maintenance models for systems subject to failure: A review. *Naval Research Logistics Quarterly*, 28(1), 47–74. https://doi.org/10.1002/nav.3800280106

[24]. Tijms, H. C. (2008). *Stochastic models: An algorithmic approach*. Wiley.

[25]. Wang, K. H., & Chen, W. L. (2009). Comparative analysis of machine repair problem with warm spares and server vacations. *Applied Mathematical Modelling*, 33(4), 2184–2197. https://doi.org/10.1016/j.apm.2008.05.025

[26]. Yen, T. C., & Wang, K. H. (2014). Performance analysis of a multi-server machine repair problem with unreliable servers and a queueing model with balking and reneging. *Computers & Industrial Engineering*, 76, 335–343. https://doi.org/10.1016/j.cie.2014.08.009