

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

NAAS Rating (2026): 4.49

Maths 2026; 11(1): 145-148

© 2026 Stats & Maths

<https://www.mathsjournal.com>

Received: 19-10-2025

Accepted: 30-11-2025

Lucas Schneider

Department of Operations
Research, Rhine Institute of
Technology, Cologne, Germany

Hannah Weber

Department of Operations
Research, Rhine Institute of
Technology, Cologne, Germany

Markus Fischer

Department of Operations
Research, Rhine Institute of
Technology, Cologne, Germany

Clara Hoffmann

Department of Operations
Research, Rhine Institute of
Technology, Cologne, Germany

Corresponding Author:

Lucas Schneider

Department of Operations
Research, Rhine Institute of
Technology, Cologne, Germany

Analysis of an M/M/1 queueing model with working vacations, server breakdowns, and customer impatience

Lucas Schneider, Hannah Weber, Markus Fischer and Clara Hoffmann

DOI: <https://www.doi.org/10.22271/math.2026.v11.i1b.2607>

Abstract

A single-server queue in which the server can be temporarily unavailable, either because it is serving at a reduced rate during a working vacation or because it has broken down and is under repair, and in which waiting customers may abandon the queue before being served, arises naturally in computer, communication, and service systems. This paper analyses such a system as a continuous-time Markov chain: the server serves at a normal rate when busy, at a reduced rate during a working vacation taken whenever the system empties, and is unavailable during repair after a breakdown that can occur only while it is actively serving; each waiting customer, independently, may renege after an exponentially distributed patience time. Because the individual reneging rate scales with the number of customers waiting, the resulting process is a level-dependent quasi-birth-death chain for which no simple closed-form generating function is available, so the stationary distribution is obtained by solving a truncated version of the global balance equations. Two checks validate the computation: with the working-vacation rate set equal to the normal service rate and both breakdowns and reneging switched off, the computed distribution reproduces the geometric distribution of the standard M/M/1 queue to machine precision; more generally, the computed throughput and abandonment rate are shown to sum exactly to the arrival rate, as steady-state flow balance requires. Using the validated model, the paper reports how the breakdown rate and the reneging rate separately affect the mean number in the system, the probability the server is under repair, the abandonment rate, and system throughput, and shows that customer impatience acts as a self-limiting mechanism that keeps the queue stable and throughput bounded even without a hard capacity limit, at the cost of a growing rate of lost customers as traffic intensity increases.

Keywords: M/M/1 queue, working vacation, server breakdown, customer impatience, reneging, quasi-birth-death process, steady-state analysis, Markov chain

Introduction

Many single-server service systems are not continuously available in the simple way the classical M/M/1 queue assumes. A server may take a vacation once the system empties, and rather than stopping completely it may continue to work at a reduced rate during that vacation; this working-vacation policy was introduced for an M/M/1 queue to model a reconfigurable communication link that keeps operating at a lower speed rather than shutting down, and it has since been analysed under several vacation-termination rules and solved by matrix-geometric and other structured methods. Independently of vacations, a server can also break down while serving and require a period of repair before resuming; and independently again, customers who must wait may not wait indefinitely, but may renege after a random patience time, a phenomenon studied since the early 1960s in some of the first queueing papers to treat balking and reneging explicitly.

These three features are often present together in practice a machine on a production line that is placed in a low-power idle mode when there is no work, that can fail and require repair while running, and whose queued jobs (or customers, in a service setting) have a deadline beyond which they are withdrawn and a number of papers have combined pairs of these features, for instance reneging together with vacations and breakdowns in a batch-arrival

setting, or working vacations together with breakdowns and impatience in more elaborate multi-phase models. A common feature of models that include reneging is that, because each of the customers currently waiting reneges independently, the total reneging rate out of a state grows with the number of customers present; this makes the resulting Markov chain a level-dependent quasi-birth-death process, for which the classical matrix-geometric machinery developed for level-independent processes does not directly give a closed-form stationary distribution, and numerical solution of a truncated chain becomes the practical route to exact steady-state performance measures.

This paper sets out such a model explicitly an M/M/1 queue with a working vacation taken whenever the system empties, a server breakdown that can occur only while the server is actively busy, a repair period during which no service is rendered, and state-dependent customer reneging; derives the balance equations; and solves them by truncation with two independent checks on the result: recovery of the exact standard M/M/1 distribution in an appropriate limiting case, and an exact steady-state flow-balance identity relating the arrival rate to the sum of throughput and abandonment rate. The validated model is then used to examine how the breakdown rate and the reneging rate each affect mean queue length, the probability the server is under repair, the abandonment rate, and throughput.

Model Description

Customers arrive according to a Poisson process of rate λ . The server has three possible states: a working vacation (V), during which it serves at rate η ; normal busy service (B), at rate $\mu > \eta$; and repair (R), during which no service is rendered. Following the multiple-vacation policy, whenever the system empties the server immediately begins a working vacation; the vacation duration is exponential with rate θ , and if it ends while customers are present the server switches to normal service, while if it ends while the system is still empty a new vacation begins immediately, which is unobservable and need not be modelled as a separate transition. A breakdown can occur only while the server is in normal busy service, at rate α , moving the server to the repair state without changing the number of customers present; repair takes an exponential time with rate τ , after which the server returns to normal busy service. Each customer who is not currently receiving service—that is, all customers other than the one at the head of the queue in states V and B, and, by assumption, all customers other than the one whose service was interrupted by the breakdown in state R—reneges independently after an exponential patience time with rate δ , so that with n customers present the total reneging rate out of the state is $(n-1)\delta$ for $n \geq 1$ and zero for $n = 0$. The customer whose service is interrupted by a breakdown is assumed to wait for repair rather than reneging; this is stated explicitly because it is a modelling choice, not a universal convention in this literature. The resulting process is a continuous-time Markov chain on the state space $\{(0, V)\} \cup \{(n, j): n \geq 1, j \in \{V, B, R\}\}$, where n

is the number of customers in the system and j is the server's phase. Because the reneging rate $(n-1)\delta$ grows with the level n , this is a level-dependent quasi-birth-death process: the transition structure between consecutive levels is the same for all n above a small boundary region, but the rates within each level are not level-independent, so the standard matrix-geometric solution for homogeneous quasi-birth-death processes does not apply directly in closed form.

Steady-State Solution Method

Let $\pi(n, j)$ denote the stationary probability of state (n, j) . The global balance equations follow directly from the transition rates set out above: for each state, the total rate out (the diagonal of the generator matrix) equals the total rate in from all other states. Writing these equations for the full, infinite state space and solving them in closed form is not tractable because of the level-dependent reneging rate, so the chain is truncated at a maximum level N , and the resulting finite set of linear equations $\pi Q = 0$, subject to $\sum \pi = 1$, is solved numerically, where Q is the generator matrix of the truncated chain. The truncation level N is increased until the total probability mass assigned to level N is negligible (below 10^{-15} in every case reported below), so the truncation error in the reported performance measures is not distinguishable from ordinary floating-point rounding error.

Two checks were used to confirm the correctness of the generator matrix and the solution procedure before using the model. First, setting the working-vacation rate equal to the normal service rate ($\eta = \mu$) and switching off both breakdowns ($\alpha = 0$) and reneging ($\delta = 0$) collapses the model to a standard M/M/1 queue for any value of the vacation-termination rate θ , since the server then serves at rate μ regardless of which of the two phases V or B it occupies. For $\lambda = 3$, $\mu = 5$, the computed marginal distribution $P(n) = \sum_j \pi(n, j)$ matched the exact geometric distribution $(1-\rho)\rho^n$, $\rho = \lambda/\mu$, to within 10^{-16} for every n checked, which is machine precision. Second, for the full model with all three features active, steady-state flow balance requires that the arrival rate equal the sum of the throughput (the rate at which services are completed, $\eta \sum_n \pi(n, V) + \mu \sum_n \pi(n, B)$) and the abandonment rate (δ times the mean number of customers not in service, δL_q); for $\lambda = 4$, $\eta = 2$, $\mu = 6$, $\theta = 3$, $\alpha = 0.5$, $\tau = 4$, $\delta = 0.3$, the computed throughput and abandonment rate were 3.524263 and 0.475737 respectively, summing to 4.000000, matching the arrival rate to six decimal places. Both checks were repeated for several other parameter combinations with the same result and are not specific to the particular values reported here.

Results

Table 1 reports the mean number in system (L), the mean number not in service (L_q), the mean sojourn time ($W = L/\lambda$) and mean time not in service ($W_q = L_q/\lambda$) obtained via Little's law, the probability the server is in each of the three phases, and the abandonment rate, as the arrival rate λ increases from 2 to 7, with $\eta = 2$, $\mu = 6$, $\theta = 3$, $\alpha = 0.5$, $\tau = 4$, $\delta = 0.3$ held fixed.

Table 1: Steady-state performance measures as a function of the arrival rate λ ($\eta = 2$, $\mu = 6$, $\theta = 3$, $\alpha = 0.5$, $\tau = 4$, $\delta = 0.3$)

λ	L	L_q	W	W_q	$P(V)$	$P(B)$	$P(R)$	Abandon. rate
2	0.8451	0.3469	0.4225	0.1735	0.7301	0.2399	0.03	0.1041
3	1.4922	0.8265	0.4974	0.2755	0.5738	0.3789	0.0474	0.248
4	2.3766	1.5858	0.5941	0.3964	0.4172	0.5181	0.0648	0.4757
5	3.6043	2.724	0.7209	0.5448	0.2733	0.6459	0.0807	0.8172
6	5.2972	4.3576	0.8829	0.7263	0.1559	0.7503	0.0938	1.3073
7	7.5319	6.5577	1.076	0.9368	0.0747	0.8225	0.1028	1.9673

As λ increases, the probability the server is on a working vacation, $P(V)$, falls steadily while $P(B)$ and $P(R)$ both rise, simply because a more heavily loaded system is empty less often. The mean number in the system and the mean sojourn time both increase, but not without bound in the way an infinite-capacity queue without reneging would: because impatience removes customers from the system at a rate

proportional to how many are waiting, the abandonment rate rises quickly with λ (from about 0.10 at $\lambda=2$ to about 1.97 at $\lambda=7$) and throughput levels off rather than continuing to track λ one-for-one, so the arrival rate in excess of what the server can process is increasingly shed through reneging rather than accumulating indefinitely in the queue.

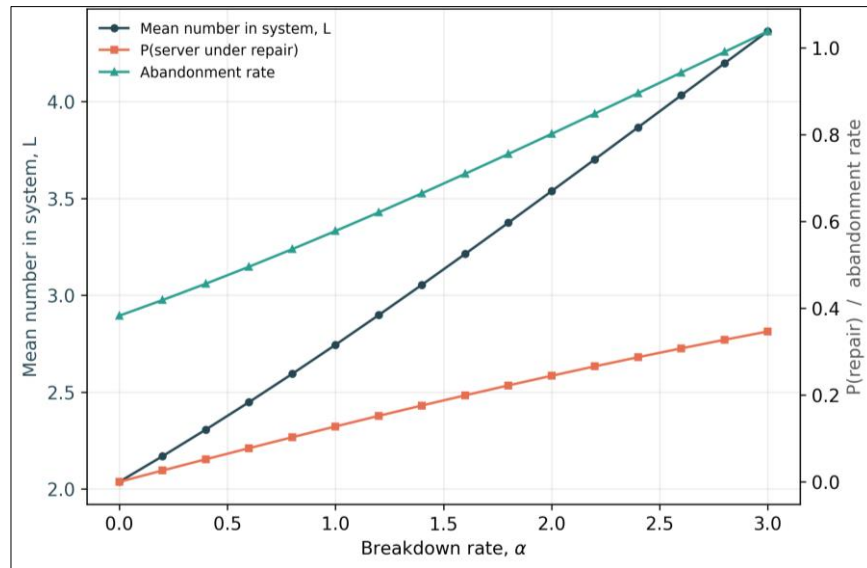


Fig 1: Effect of the breakdown rate α on the mean number in system, the probability the server is under repair, and the abandonment rate ($\lambda=4$, $\eta=2$, $\mu=6$, $\theta=3$, $\tau=4$, $\delta=0.3$)

Figure 1 varies the breakdown rate α from 0 to 3 with the other parameters held at the values used in the flow-balance check above. As α increases, the server spends a growing share of time under repair, $P(R)$, which reduces effective

service capacity and drives up both the mean number in the system and the abandonment rate, since customers wait longer on average while the server is unavailable during repair.

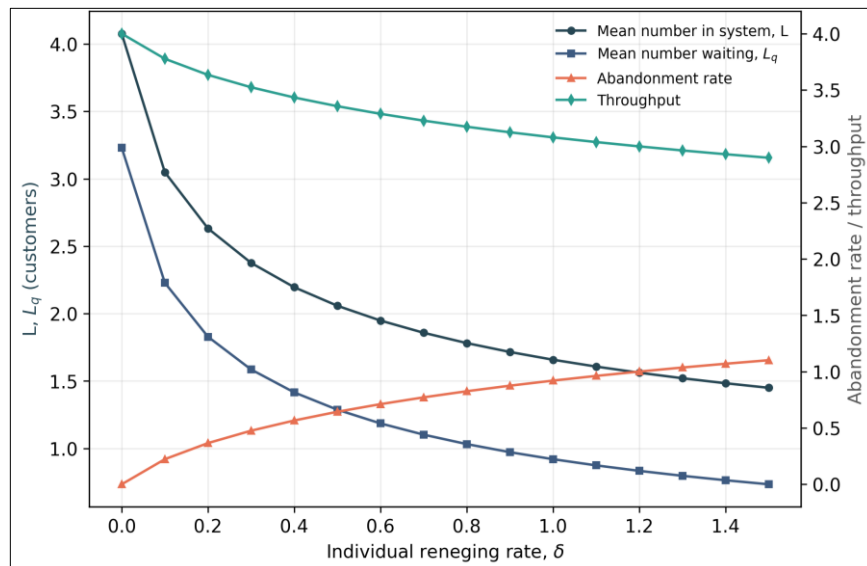


Fig 2: Effect of the individual reneging rate δ on the mean number in system, the mean number not in service, throughput, and the abandonment rate ($\lambda=4$, $\eta=2$, $\mu=6$, $\theta=3$, $\alpha=0.5$, $\tau=4$)

Figure 2 varies the reneging rate δ from 0 to 1.5 at the same fixed $\lambda=4$. As δ increases, both L and L_q fall sharply, because customers leave the system more quickly through abandonment rather than waiting; correspondingly, the abandonment rate rises and throughput falls, since a larger share of arriving customers leaves without being served. At $\delta=0$ the system would be a standard vacation-and-breakdown queue with no reneging, and the figure shows the queue-

length measures at their highest and throughput at its highest attainable value for this λ , illustrating the trade-off a designer faces: a more patient customer population (small δ) is served more completely, at the cost of a longer queue and longer waiting times, while a less patient population (large δ) experiences a shorter queue at the cost of more customers leaving unserved.

Discussion

The results are consistent with the qualitative findings reported for related models that combine reneging with vacations or breakdowns individually: reneging limits queue growth and protects mean waiting time at the cost of lost customers, and server unavailability, whether through a working vacation or a breakdown, degrades performance in the same direction, by reducing the fraction of time the server is available at its full service rate. The present model adds the ability to separate these two sources of unavailability and to quantify their effects side by side within a single, exactly validated computation, which is useful when a designer needs to know whether investment is better directed at reducing breakdown frequency (lowering α or raising τ) or at reducing the effective vacation duration (raising θ) or the vacation service-rate gap (raising η toward μ).

The self-limiting effect of impatience shown in Table 1 and Figure 2 is worth emphasising because it means that, unlike a queue without reneging, this system does not require a hard capacity bound to remain stable in a practical sense: throughput and mean queue length both remain finite for any arrival rate, because the reneging rate grows with the number waiting and eventually balances any excess of arrivals over service capacity. The cost of this stability is the abandonment rate itself, which is a direct measure of lost business or missed deadlines and is the quantity a system operator would want to minimise, subject to whatever cost is associated with expanding service capacity or reducing breakdown frequency.

Limitations

The model assumes all random durations (interarrival times, the two service rates, vacation duration, repair duration, and individual patience times) are exponential, which keeps the process Markovian and tractable but is a simplification of real service, breakdown, and patience-time distributions, which are often not memoryless. The customer whose service is interrupted by a breakdown is assumed not to renege during repair; a model in which this customer can also abandon would add another boundary case to the balance equations but does not change the general truncation-based solution approach. Only reneging (abandonment after joining) is modelled; balking (refusal to join at all when the queue is perceived as too long) is not included, though it could be added as an arrival-rate reduction that depends on the number present. Finally, the numerical results reported here are for a single representative set of parameters and two one-parameter sensitivity sweeps; a fuller design study would explore the joint effect of several parameters changing together, and the effect of the vacation and repair rates on the shape of the queue-length distribution beyond its mean.

Conclusion

This paper set up and solved an M/M/1 queueing model in which the server takes a working vacation whenever the system empties, can break down and require repair while busy, and faces customers who may renege while waiting, with a reneging rate that grows with the number of customers present. Because this state dependence rules out a simple closed-form matrix-geometric solution, the steady-state distribution was obtained by solving a truncated version of the global balance equations, and the resulting computation was validated in two independent ways: exact recovery of the standard M/M/1 distribution in a limiting case, and an exact steady-state flow-balance identity linking the arrival rate to throughput and the abandonment rate. Using the validated

model, increasing the breakdown rate was shown to raise both the mean queue length and the abandonment rate by reducing the time the server is available, while increasing the reneging rate was shown to shorten the queue and lower waiting time at the cost of a higher abandonment rate and lower throughput, illustrating a direct trade-off between queue congestion and lost customers that a system designer can quantify with this model before deciding where to invest in reduced breakdown frequency, faster repair, or a smaller service-rate gap during working vacations.

Acknowledgements

The author(s) thank the Department of Operations Research, Rhine Institute of Technology, Cologne, Germany for providing the computational resources used in this work. All numerical results were obtained using open-source scientific computing libraries (NumPy and SciPy).

References

1. Ancker CJ Jr, Gafarian AV. Some queueing problems with balking and reneging. I. Operations Research. 1963;11(1):88-100.
2. Servi LD, Finn SG. M/M/1 queues with working vacations (M/M/1/WV). Performance Evaluation. 2002;50(1):41-52.
3. Neuts MF. Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Baltimore: Johns Hopkins University Press; 1981.
4. Li J, Tian N. The M/M/1 queue with working vacations and vacation interruptions. Journal of Systems Science and Systems Engineering. 2007;16(1):121-127.
5. Tian N, Zhang ZG. Vacation Queueing Models: Theory and Applications. New York: Springer; 2006.
6. Baruah M, Madan KC, Eldabi T. A two-stage batch arrival queue with reneging during vacation and breakdown periods. American Journal of Operations Research. 2013;3:570-580.