mathematics

MDPI

Article

A $MAP/PH/1/K$ Queueing Model with N-Policy for Optimal Regeneration of a Diesel Particulate Filter

Dmitry Efrosinin [1,*], Natalia Stepanova [1], Zóltan Gál [2] and Janos Sztrik [2]

[1] Institute of Stochastics, Johannes Kepler University Linz, 4040 Linz, Austria; natalia.stepanova@jku.at
[2] Faculty of Informatics, University of Debrecen, 4028 Debrecen, Hungary; gal.zoltan@inf.unideb.hu (Z.G.); sztrik.janos@inf.unideb.hu (J.S.)
* Correspondence: dmitry.efrosinin@jku.at

Abstract

This paper analyzes a $MAP/PH/1/K$ queue with N-policy, setup, interruptions, reset, and a random environment. Arrivals are the MAP; service, setup, interruption, and reset times are PH-distributed. Under the N-policy, the server idles until the queue length is equal to N, and then performs setup. Interruptions return the system to idle and re-enable the N-policy. At capacity K, a reset empties the system. The random environment modulates parameters for different regimes. Motivated by Diesel Particulate Filter (DPF) regeneration, soot accumulation is mapped to arrivals, burning to service, regeneration triggers to N-policy, heating to setup, engine changes to interruptions, and cleaning to reset. Environmental states represent driving patterns. Regeneration succeeds if either the system empties via service or an interruption occurs with remaining soot less than or equal to level L. We derive the block-structured generator, obtain stationary probabilities via matrix-analytic methods, and optimize the threshold N via average cost. Numerical results quantify how correlation and driving conditions affect performance and costs, offering tools to balance fuel consumption, engine performance, and filter longevity.

Keywords: $MAP/PH/1/K$ queue; N-policy; matrix-analytic method; average cost; diesel particulate filter

MSC: 60K25; 60K30; 60K37

1. Introduction

The combination of Markovian arrival processes (MAP) with phase-type (PH) service distributions offers a flexible modeling paradigm for queueing systems in which correlation structures and non-exponential behaviors are paramount [1]. Whereas traditional Poisson-driven models assume memoryless and smooth arrivals, the MAP/PH framework inherently accommodates burstiness in input streams and systematic variations in service durations. This expressive power makes the approach especially valuable in diverse domains, including telecommunications and manufacturing [2].

The N-policy was first introduced by [3] for a single-server queueing system as an operating rule governing server activation and deactivation. Under this classical policy, the server remains idle until the queue length reaches a threshold N, thereby balancing the trade-off between server idleness and the costs of frequent startup and shutdowns. Following its introduction, the 1970s and 1980s saw foundational work that extended the N-policy to various classical queueing models due to its practical relevance in energy-saving and operational efficiency contexts. The N-policy for $M/G/1$ was studied in [4] and

Academic Editor: Alexander Dudin

Received: 8 April 2026
Revised: 4 May 2026
Accepted: 5 May 2026
Published: 8 May 2026

later [5] provided a comprehensive review of vacation models, which are closely related to the N-policy systems. A significant theoretical development occurred by working on [6], where the system $GI/M/1$ under the N-policy was analyzed, demonstrating that key performance measures, such as the length of the queue, could be decomposed into the sum of the classical queue measure and an additional component attributable to the N-policy. The late 1980s and 1990s marked a period of extensive generalization and application of the N-policy. When combined with setup times specifying the duration required to prepare the server for operation after reaching the threshold, the model captures realistic scenarios where server activation involves non-negligible delays. Researchers began integrating it with more complex system features, such as batch arrivals [7], server vacations, setup times, and breakdowns [8,9]; multiple servers and polling systems [10]; cost structures; and optimization [11–13]. For the authors of this article, the topic of N-policy is also not new. It has previously been considered in its classical form in a number of works on Markovian queueing systems with retrials and an unreliable server [14,15], as well as in systems where this policy was applied to the recovery of a failed server [16,17].

Over the past decade, queueing models that incorporate operational policies and system interruptions have advanced considerably. A notable contribution is the work of [18], who analyzed a $MAP/PH/1$ system under N-policy that features catastrophic delay action and a standby server. Their findings underscore the importance of catastrophic events capable of emptying the system entirely, as such disruptions fundamentally reshape the dynamics of the system and require specialized analytical methods. Building on this direction, Ref. [19] examined a $MAP/PH/1$ queue with differentiated vacations, vacation interruption under N-policy, optional services, and breakdown-repair mechanisms. Their comprehensive analysis—encompassing stability conditions, characterization of busy periods, and cost optimization—established a methodological benchmark for complex queueing systems with multiple operational phases. Separately, Ref. [20] studied the $MAP/PH/1$ queue with working vacations, vacation interruptions, and the N-policy. A second critical dimension of realism in queueing systems is service interruptions, which may stem from server breakdowns, scheduled maintenance, or external disruptions, often with complex duration patterns. Modeling interruption times via phase-type distributions preserves analytical tractability while accommodating a wide range of duration distributions. Recent work by [21] on waiting times in systems with variable cross-correlated arrival rates highlights that service interruptions contribute substantially to overall system variability, and this contribution adds to arrival variability in determining queue performance. Meanwhile, the concept of regeneration, where the system periodically returns to an empty state, has attracted renewed interest owing to applications in inventory control, manufacturing, and environmental systems. Ref. [22] demonstrated that correlation in arrival processes directly affects blocking probabilities and waiting time distributions, and these effects can be quantified through systematic analysis of regeneration cycles. Their work provided closed-form expressions for performance distributions, establishing that correlation-induced effects are attributable to correlation alone, independent of other traffic characteristics.

Our paper deals with an emerging and particularly compelling application domain for advanced queueing models such as the regeneration process in diesel particulate filters (DPFs). Diesel particulate filters (DPFs) are an essential component in modern diesel vehicles, designed to capture harmful particulate matter from exhaust emissions to comply with stringent environmental regulations. As soot accumulates over time, the back-pressure increases, degrading engine performance and fuel economy. This requires periodic regeneration—a high-temperature process that oxidizes trapped soot and restores filter functionality. In this context, Ref. [23] developed dynamic models of DOC-DPF

after-treatment systems for regeneration control, demonstrating that regeneration can be activated passively via catalysts or actively through fuel injection strategies. Subsequently, Ref. [24] advanced control-oriented modeling of catalytic DPFs, emphasizing the need for real-time management of regeneration processes under back-pressure and thermal state constraints. The practical challenge facing our industry partner lies in optimizing the timing and frequency of regeneration events. Premature regeneration wastes fuel and shortens filter lifespan, whereas delayed regeneration risks excessive back-pressure, potential filter damage, and emissions non-compliance. Current practice relies on threshold-based triggers using differential pressure sensors. However, such approaches do not fully account for the stochastic nature of soot accumulation under variable driving conditions.

The diesel particulate filter (DPF) regeneration problem exhibits a striking structural analogy to queueing systems operating under the N-policy. Specifically, the accumulation of soot behaves as arrivals of customers; the capacity of the filter corresponds to a finite buffer of size K; the regeneration trigger threshold N mirrors the activation level of the policy of N; the heating and oxidation phase resembles a setup time; and the return to a clean filter state parallels a system reset. Service interruptions are naturally correlated to changes in engine operating conditions that temporarily suspend effective soot oxidation. Furthermore, the random environment, which captures different driving modes such as city, highway and mixed conditions, directly modulates the accumulation rates of soot, similar to queueing models with Markovian arrival processes embedded in a Markovian random environment [25]. Remarkably, despite these clear structural parallels, no existing study has developed a unified queueing framework that simultaneously incorporates all these features while explicitly being tailored to DPF regeneration applications. The proposed $MAP/PH/1/K$ model, integrating the N-policy, setup time, service interruption, reset mechanism, and random environment, fills this gap by offering a comprehensive mathematical framework for analyzing systems characterized by the following:

- Correlated arrival processes captured through MAP, enabling modeling of pulsed soot accumulation patterns under varying driving conditions.
- Phase-type service distributions representing the structured stages of soot oxidation during regeneration.
- N-policy with setup time modeling the threshold-based regeneration initiation and the heating/preparation phase.
- Service interruptions representing disruptions to effective regeneration due to changes in the operating condition of the engine.
- Reset mechanism modeling the return to the clean filter state after complete failure and successful regeneration at a vehicle service facility.
- Random environment capturing transitions between different driving regimes (city, highway, mixed).

The last aspect of environmental modulation is particularly relevant for DPF applications, as we parameterize different environmental states to represent city driving where frequent stop-and-go patterns with variable soot accumulation rates can be taken into account, highway driving with a steady-state operation and predictable accumulation patterns, and mixed driving, which combines patterns with transitional behavior.

The contributions of this paper are threefold:

- First, we develop the complete infinitesimal generator structure for the proposed model, providing explicit block matrices that account for all system transitions.
- Second, we derive stationary state probabilities using the matrix-analytic method, enabling computation of key performance measures.

- Third, we formulate and solve an average cost optimization problem under a specified cost structure that includes holding costs, setup costs, interruption penalties, and regeneration costs.

By revealing how correlation structures and environmental transitions affect regeneration timing and effectiveness, this research provides quantitative tools to optimize DPF regeneration strategies in competing objectives of fuel consumption, regeneration frequency, and filter longevity. The framework extends naturally to any system characterized by threshold-based activation, setup delays, service interruptions, and periodic resetting—including batch manufacturing processors, communication networks with sleep modes, and inventory systems with emergency ordering policies. For diesel vehicle manufacturers, the model supports dynamic driver-adaptive regeneration procedures that adjust in real time to the specific characteristics of vehicle operation.

A significant constraint in this research is the confidentiality of the statistical data provided by BMW. The joint project agreement prohibits the public disclosure of proprietary measurements of instrumented vehicles, including real soot accumulation rates, oxidation kinetics parameters, and regeneration timing data. Consequently, for the numerical results presented in this paper, we employ abstract system values that preserve the qualitative behavior of the model while protecting sensitive information. This approach demonstrates the analytical capabilities of the framework without compromising the confidentiality agreement. Importantly, this limitation does not diminish the practical utility of our contribution. For real-world implementation, firms with access to their own proprietary data, such as automotive manufacturers with DPF telemetry, logistics companies with fleet data, or equipment suppliers with sensor measurements, can parameterize the model using their specific values. The algorithmic solutions provided in this paper (stationary probabilities, cost optimization, regeneration period characteristics) are directly applicable once model parameters are estimated from empirical data. This separation between the mathematical framework (publicly available) and the proprietary data (confidentially held) enables academic dissemination while protecting industrial intellectual property.

The remainder of this paper is organized as follows. Section 2 provides a detailed description of the mathematical model, notation, and state space partitioning. Section 3 presents the infinitesimal generator structure, derives the stationary distribution using matrix-analytic methods, and develops the cost structure and the optimization framework. Section 4 concludes with remarks on implementation and future research directions.

2. Mathematical Model

Consider a single-server finite-capacity queueing system of type $MAP/PH/1/K$, characterized by a Markovian arrival process (MAP), phase-type (PH) service times, a N-policy for server activation, random service interruptions, and a reset mechanism triggered when the system becomes full. Motivated by the application of DPF regeneration, we assume that the inflow of soot particles (measured in grams) follows a MAP. Furthermore, because the accumulation of soot depends on the speed of the vehicle, the dynamics of the system is modulated by a random environment. Three distinct accumulation regimes are identified:

- The low accumulation (0.05–0.3 g/h): corresponds to highway driving at average speeds exceeding 90 km/h.
- The middle accumulation (0.3–0.5 g/h): corresponds to speeds between 50 and 90 km/h.
- The high accumulation (0.5–1.5 g/h): corresponds to driving in the city at average speeds not exceeding 50 km/h.

The transition parameters among these environmental states may vary with driving style and can be adjusted dynamically by the vehicle's on-board computer using available statistical data. We remind once again that the parameter values used in the paper are for the most part hypothetical. Actual data is, in most cases, not publicly available. In real-world situations, these values depend on a multitude of factors, such as engine power and type, filter type, driving style, etc. Nevertheless, data on soot accumulation rates ranging from 0.0024 g/kWh to 0.2 g/kWh can be found in open-source material on the internet, e.g., Refs. [26,27]. We use the unit of grams per hour of engine operation. If we multiply these values by the car's power output, which ranges from 15 to 25 kW, we get the soot accumulation rates of around 0.036 g/h to 3 g/h, which generally corresponds to the used data.

Throughout this paper, let $e(m)$ denote the m-dimensional column vector of the ones, and let $I(m)$ denote the $m \times m$ identity matrix. The notation $e(\cdot)$ represents a column vector of the appropriate dimension. The zero matrix block of dimension $n \times m$ is denoted by $0_{n\times m}$.

Random environment. Let $\Phi_e(t)_{t\geq 0}$ denote the stochastic process governing the random environment. This process is modeled as an irreducible continuous-time Markov chain with infinitesimal generator H, taking values in $1, \ldots, m_e$. The corresponding stationary distribution is denoted by π_e, which satisfies $\pi_e H = 0$ and $\pi_e e(m_e) = 1$.

Arrival process. For a fixed environmental state $\phi_e \in 1, \ldots, m_e$, we define a Markovian arrival process (MAP) characterized by the pair of $m_a \times m_a$ matrices $D^{(\phi_e)}0$ and $D^{(\phi_e)}1$. These matrices govern, respectively, transitions without an arrival and transitions accompanied by an arrival, conditional on the environment being in state ϕ_e. It is important to note that the transitions of the environmental process $\Phi_e(t)_{t\geq 0}$ do not alter the state of the arrival process itself; rather, they only modify the transition rates of the MAP.

Arrivals take place only while the server is idle. The infinitesimal generator of the underlying Markov chain for the MAP is given by $D^{(\phi_e)} = D_0^{(\phi_e)} + D_1^{(\phi_e)}$. The stationary vector of the MAP, denoted by $\pi_a^{(\phi_e)}$, satisfies the system $\pi_a^{(\phi_e)} D^{(\phi_e)} = 0$ and $\pi_a^{(\phi_e)} e(m_a) = 1$. Each arrival increments the customer count—corresponding to the total grams of soot in the system—by exactly one (i.e., one gram). For a fixed environmental state ϕ_e, the following quantities are computed: the average arrival rate $\lambda^{(\phi_e)}$, the second moment of inter-arrival times $\lambda_2^{(\phi_e)}$, and the lag-1 correlation coefficient. These are obtained as follows:

$$\lambda^{(\phi_e)} = \pi_a^{(\phi_e)} D_1^{(\phi_e)} e(m_a),\ \lambda_2^{(\phi_e)} = 2\pi_a^{(\phi_e)} (-\lambda^{(\phi_e)} D_0^{(\phi_e)})^{-1} e(m_a),$$
$$\rho = \frac{\lambda^{(\phi_e)} \pi_a^{(\phi_e)} (-D_0^{(\phi_e)})^{-1} D_1^{(\phi_e)} (-D_0^{\phi_e})^{-1} e(m_a) - 1}{2\lambda^{(\phi_e)} \pi_a^{(\phi_e)} (-D_0^{(\phi_r)})^{-1} e(m_a) - 1}.$$

Service process. The service time follows a phase-type (PH) distribution with m_s phases, represented by the pair (μ, M). Here, μ is the initial probability vector (dimension m_s), and M is a subgenerator matrix $m_s \times m_s$. Phase-dependent service completion rates are given by $M^0 = -Me(m_s)$, and the average service rate is $\mu_s = -(\mu M^{-1} e(m_s))^{-1}$. Upon service completion, the customer count decreases by one.

Service interruption process. Service interruptions follow a PH distribution with representation (α, A), where α is an m_i-dimensional initial vector and A is an $m_i \times m_i$ subgenerator. Interruption occurrence rates are $A^0 = -Ae(m_i)$, and the average interruption rate is $\alpha_i = -(\alpha A^{-1} e(m_i))^{-1}$. When an interruption occurs, the service stops, the server becomes idle, and the customer remains in the system. During server busy periods, we deliberately do not track the states of the random environment or the MAP phases to reduce the dimensionality of the state space. This information is therefore lost. When the server idles again after a service interruption, the environment and MAP phases are probabilistically restored

using the stationary distributions π_e (for the environment) and $\pi_a = (\pi_a^{(1)}, \ldots, \pi_a^{(\phi_e)})$ (for the MAP). If a service interruption occurs when the number of remaining customers is less than or equal to a threshold L, the regeneration is considered successful. Otherwise, the regeneration is deemed incomplete. For DPF application, successful filter regeneration is defined as the condition in which no more than 10% of the threshold level of soot N remains in the filter. This criterion may be adjusted depending on the specific physical characteristics of the DPF and the oxidation process of the soot.

Startup process. The startup process is PH-distributed and is automatically triggered when the server is idle and the customer count reaches the threshold N. It is represented by (γ, Γ) with dimension m_u. Startup completion rates are $\Gamma^0 = -\Gamma e(m_u)$, and the average startup rate is $\gamma_u = -(\gamma\Gamma^{-1}e(m_u))^{-1}$. Upon startup completion, service begins with a new customer.

Reset process. The reset process is also PH-distributed and initiates when the system reaches its full capacity of K customers while the server remains idle. During this state, MAP phases are not tracked. The reset process is represented by (β, B) with dimension m_r, with completion rates $B^0 = -Be(m_r)$ and average rate $\beta_r = -(\beta B^{-1}e(m_r))^{-1}$. After the reset completes, the system becomes empty (level 0), and the arrival process is reinitialized using the distribution π_a.

The following notation is introduced to describe the components of the stochastic process. Define the following:

- $N(t)$: number of customers in the system at time t, $N(t) = 0, 1, \ldots, K$;
- $D(t)$: status of the server at time t, where

$$D(t) = \begin{cases} 0, & \text{server idle,} \\ 1, & \text{server busy,} \end{cases}$$

- $\Phi_a(t)$: phase of the MAP process $\Phi_a(t) = 1, \ldots, m_a$;
- $\Phi_s(t)$: phase of the PH-type service process, $\Phi_s(t) = 1, \ldots, m_s$;
- $\Phi_i(t)$: phase of the PH-type service interruption process, $\Phi_i(t) = 1, \ldots, m_i$;
- $\Phi_u(t)$: phase of the PH-type startup process, $\Phi_u(t) = 1, \ldots, m_u$;
- $\Phi_r(t)$: phase of the PH-type reset process, $\Phi_r(t) = 1, \ldots, m_r$.

The queueing process evolves in the state space $E = \bigcup_{n=0}^{K} l(n)$, where each level $l(n)$ is defined according to the system content n.

Level 0: Empty system.

$$l(0) = \bigcup_{\phi_e=1}^{m_e} \bigcup_{\phi_a=1}^{m_a} \{(0, 0, \phi_e, \phi_a)\}.$$

with cardinality $|l(0)| = l_0 = m_e m_a$.

Levels $1 \le n \le N-1$ *below the threshold.*

$$l(n) = l(n,0) \cup l(n,1) = \bigcup_{\phi_e=1}^{m_e} \bigcup_{\phi_a=1}^{m_a} \{(n, 0, \phi_e, \phi_a)\} \bigcup \bigcup_{\phi_s=1}^{m_s} \bigcup_{\phi_i=1}^{m_i} \{(n, 1, \phi_s, \phi_i)\},$$

have cardinality $|l(n)| = |l(n,0)| + |l(n,1)| = l_0 + l_1 = m_e m_a + m_s m_i$ and include states $(n, 0, \phi_e, \phi_a)$ with n customers, idle server, random environment state $\phi_e = 1, \ldots, m_e$, MAP phase $\phi_a = 1, \ldots, m_a$ and $(n, 1, \phi_s, \phi_i)$ with n customers, busy server, service phase $\phi_s = 1, \ldots, m_s$, interruption phase $\phi_i = 1, \ldots, m_i$.

Levels $N \le n \le K-1$ *above or below the threshold.*

$$l(n) = l(n,0) \cup l(n,1) = \bigcup_{\phi_e=1}^{m_e} \bigcup_{\phi_a=1}^{m_a} \bigcup_{\phi_u=1}^{m_u} \{(n, 0, \phi_e, \phi_a, \phi_u)\} \bigcup \bigcup_{\phi_s=1}^{m_s} \bigcup_{\phi_i=1}^{m_i} \{(n, 1, \phi_s, \phi_i)\},$$

have cardinality $|l(n)| = |l(n,0)| + |l(n,1)| = l_2 + l_1 = m_e m_a m_u + m_s m_i$ and consist of states $(n, 0, \phi_e, \phi_a, \phi_u)$ with n customers, idle server, random environment state $\phi_e = 1, \ldots, m_e$, MAP phase $\phi_a = 1, \ldots, m_a$, startup phase $\phi_u = 1, \ldots, m_u$, and states $(n, 1, \phi_s, \phi_i)$ with n customers, busy server, service phase $\phi_s = 1, \ldots, m_s$ and service interruption phase $\phi_i = 1, \ldots, m_i$.

Level K of the full system (reset level).

$$l(K) = \bigcup_{\phi_r=1}^{m_r} \{(K, 0, \phi_r)\},$$

where $(K, 0, \phi_t)$ has K customers, idle server, and reset process at phase $\phi_r = 1, \ldots, m_r$. The states in each level $l(n)$ are enumerated in lexicographic order. The total number of states in the state space E is equal to

$$|E| = l_0(K-1) + l_1(N-1) + l_2(K-N) + m_r.$$

The following conventions apply regarding undefined process components depending on system state:

- When the server is idle ($D(t) = 0$) and the customer count satisfies $0 \le N(t) \le N-1$, the components $\Phi_s(t)$, $\Phi_i(t)$, $\Phi_u(t)$, and $\Phi_r(t)$ are not defined.
- When the server is idle and $N \le N(t) \le K-1$, the components $\Phi_s(t)$, $\Phi_i(t)$, and $\Phi_r(t)$ are not defined.
- When the server is busy ($D(t) = 1$) with $1 \le N(t) \le K-1$, the components $\Phi_e(t)$, $\Phi_a(t)$, $\Phi_u(t)$, and $\Phi_r(t)$ are not defined.
- At the reset level ($N(t) = K$, $D(t) = 0$), the components $\Phi_a(t)$, $\Phi_u(t)$, $\Phi_s(t)$ and $\Phi_i(t)$ are not defined.

Taking into account the last comment, the multi-dimensional process

$$\{X(t)\}_{t\ge 0} = \{N(t), D(t), \Phi_e(t), \Phi_a(t), \Phi_u(t), \Phi_s(t), \Phi_i(t), \Phi_r(t)\}_{t\ge 0} \tag{1}$$

is an irreducible regular continuous-time Markov chain (CTMC) with state space E.

3. Stationary Distribution of the System States

Lemma 1. *For the state space E grouped into levels $l(n)$ by the number of customers, the infinitesimal generator $Q^{(N)}$ of CTMC $\{X(t)\}_{t\ge 0}$ defined in (1) for a fixed level N can be transformed into the tridiagonal block matrix with an off-diagonal block:*

$$Q^{(N)} = \begin{array}{c} \begin{array}{ccccccccc} l(0) & l(1) & l(2) & \cdots & l(N-1) & l(N) & l(N+1) & \cdots & l(K) \end{array} \\ \begin{pmatrix} Q_{1,0} & Q_{0,1} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ Q_{2,0} & Q_{1,1} & Q_{0,2} & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & Q_{2,1} & Q_{1,1} & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \cdots & 0 \\ 0 & \cdots & 0 & Q_{2,1} & Q_{1,1} & Q_{0,3} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & Q_{2,2} & Q_{1,2} & Q_{0,4} & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & Q_{2,3} & Q_{1,2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & Q_{2,3} & Q_{1,2} & Q_{0,5} \\ Q_{3,0} & 0 & \cdots & \cdots & 0 & 0 & 0 & 0 & B \end{pmatrix} \end{array} \begin{array}{c} l(0) \\ l(1) \\ l(2) \\ \vdots \\ l(N-1) \\ l(N) \\ l(N+1) \\ \vdots \\ l(K-1) \\ l(K) \end{array}, \tag{2}$$

where non-zero blocks $Q_{1,.}$, $Q_{0,.}$ and $Q_{2,.}$ specify the transitions respectively to stay in a certain level $l(n)$, to go from level $l(n-1)$ to level $l(n)$ according to arrivals and to go from level $l(n+1)$ to $l(n)$ due to service completion. The block $Q_{3,0}$ is responsible for the reset process, i.e., the transition from $l(K)$ to $l(0)$. The matrices are defined as follows:

$$Q_{1,0} = D_0 + H \otimes I(m_a),\ Q_{1,1} = \begin{pmatrix} D_0 + H \otimes I(m_a) & 0_{l_0 \times l_1} \\ A^0 \otimes e(m_s) \otimes (\pi_a\, diag(\pi_e) \otimes I(m_a)) & A \oplus M \end{pmatrix},$$

$$Q_{1,2} = \begin{pmatrix} D_0 \oplus \Gamma + H \otimes I(m_a m_u) & e(l_0) \otimes (\Gamma^0 \otimes \mu^\top \otimes \alpha^\top) \\ A^0 \otimes e(m_s) \otimes (\pi_a\, diag(\pi_e) \otimes I(m_a)) \otimes \gamma^\top) & A \oplus M \end{pmatrix},$$

$$Q_{0,1} = \begin{pmatrix} D_1 & 0_{l_0 \times l_1} \end{pmatrix},\ Q_{0,2} = \begin{pmatrix} D_1 & 0_{l_0 \times l_1} \\ 0_{l_1 \times l_0} & 0_{l_1 \times l_1} \end{pmatrix},$$

$$Q_{0,3} = \begin{pmatrix} D_1 \otimes \gamma^\top & 0_{l_0 \times l_1} \\ 0_{l_1 \times l_2} & 0_{l_1 \times l_1} \end{pmatrix},\ Q_{0,4} = \begin{pmatrix} D_1 \otimes I(m_u) & 0_{l_2 \times l_1} \\ 0_{l_1 \times l_2} & 0_{l_1 \times l_1} \end{pmatrix},$$

$$Q_{0,5} = \begin{pmatrix} (\beta \otimes (D_1 e(l_0)) \otimes e(m_u))^\top \\ 0_{l_1 \times m_r} \end{pmatrix},\ Q_{2,0} = \begin{pmatrix} 0_{l_0 \times l_0} \\ e(m_i) \otimes M^0 \otimes (\pi_a\, diag(\pi_e) \otimes I(m_a)) \end{pmatrix},$$

$$Q_{2,1} = \begin{pmatrix} 0_{l_0 \times l_0} & 0_{l_0 \times l_1} \\ 0_{l_1 \times l_0} & I(m_i) \otimes (M^0 \otimes \mu^\top) \end{pmatrix},\ Q_{2,2} = \begin{pmatrix} 0_{l_2 \times l_0} & 0_{l_2 \times l_1} \\ 0_{l_1 \times l_0} & I(m_i) \otimes (M^0 \otimes \mu^\top) \end{pmatrix},$$

$$Q_{2,3} = \begin{pmatrix} 0_{l_2 \times l_2} & 0_{l_2 \times l_1} \\ 0_{l_1 \times l_2} & I(m_i) \otimes (M^0 \otimes \mu^\top) \end{pmatrix},\ Q_{3,0} = B^0 \otimes (\pi_a\, diag(\pi_e) \otimes I(m_a)),$$

$$D_0 = diag(D_0^{(1)}, \ldots, D_0^{(m_e)}),\ D_1 = diag(D_1^{(1)}, \ldots, D_1^{(m_e)}).$$

Proof. The statement follows from the construction of the CTMC (1) accounting for all possible transitions: MAP arrivals (only if the server is idle), service completions, service interruptions, startup completions, and reset process, with phase-type distributions governing timing. The block structure arises from the grouping by customer count and server status, with transitions only between adjacent levels, except for the reset process from level $l(K)$ to level $l(0)$.

The square block $Q_{1,0}$ with size l_0 consists of internal transitions between the states in level $l(0)$. The square block $Q_{1,1}$ with size $l_0 + l_1$ describes the internal transitions within levels $l(n), 1 \le n \le N-1$. It consists of the block D_0 for no arrivals in subgroup of states with idle server, $A \oplus M = A \otimes I(m_s) + I(m_i) \otimes M$ for phase transitions of the service and service interruption PH processes for the subgroup of states when the server is busy, and $A^0 \otimes e(m_s) \otimes (\pi_a\, diag(\pi_e) \otimes I(m_a))$ for transitions from subgroup with busy server to subgroup with idle server due to service interruption. Service interruption rates are defined by A^0 and after service interruption, an initial phase of the MAP is defined by the stationary probability π_a and an initial state of the random environment is defined by the stationary probability π_e.

The square block $Q_{1,2}$ with size $(m_e m_a m_u + m_s m_i)$ describes transitions within the levels $l(n), N \le n \le K-1$. It includes the block $D_0 \oplus \Gamma + H \otimes I(m_a m_u)$, where $D_0 \oplus \Gamma = D_0 \otimes I(m_u) + I(m_e m_a) \otimes \Gamma$ for the phase transitions of the MAP and the PH startup process in the subgroup of states with the idle server and $H \otimes I(m_a m_u)$ for transitions due to changes of the random environment. The block $e(m_e m_a) \otimes (\Gamma^0 \otimes \mu^\top \otimes \alpha^\top)$ describes the transitions from the subgroup with idle server to the subgroup with busy server due to the startup time completion that occurs with rates Γ^0. The initial phases of the service and service interruption processes are defined respectively by initial probability vectors μ and α. The block $A^0 \otimes e(m_s) \otimes (\pi_a\, diag(\pi_e) \otimes I(m_a)) \otimes \gamma^\top)$ specifies the transitions due to the service interruption with rates A^0, and the subsequent definition of the initial phase of the MAP, of the random environment process and of the PH startup time by the probability vectors π_a, π_e and γ.

The super-diagonal blocks $Q_{0,j}, j = 1, \ldots, 4$ of appropriate sizes stand for transitions from level $l(n-1)$ to $l(n)$ due to an arrival if $n = 1$, $2 \le n \le N-1$, $n = N$ and

$N+1 \leq n \leq K-1$, taking into account the boundary behavior of the process for the N-policy. The arrival rates are described by the block D_1 with size $m_e m_a \times m_e m_a$ for the subgroup of states with a number of customers below the threshold level, by $D_1 \otimes \gamma^\top$ with size $m_e m_a \times m_e m_a m_u$ for the states at the threshold level where the initial phase of the startup process must be specified with the probability vector γ and by the block $D_1 \otimes I(m_u)$ with size $(m_e m_a m_u) \times (m_e m_a m_u)$ for the states above the threshold level. The arrival rate to the reset level $l(K)$ is defined by the block $Q_{0,5}$ with the size $(m_e m_a m_u + m_s m_i) \times m_r$. It has a non-zero block $(\beta \otimes (D_1 e(l_0)) \otimes e(m_u))^\top$ that specifies the arrival rates from the phases of the MAP and identifies the initial phase of the reset process by the vector β.

The sub-diagonal blocks $Q_{2,j}, j = 0, 1, 2$ represent the transitions from level $l(n+1)$ to level $l(n)$ due to the completion of the service with rates M^0. The matrix $Q_{2,0}$ with size $(m_e m_a + m_s m_i) \times m_a$ has the block $e(m_i) \otimes (M^0 \otimes \pi_a^T)$ that specifies the initial state of the MAP with the probability π_a after the completion of the service by transition from $l(1)$ to $l(0)$. The square matrices $Q_{2,1}$ and $Q_{2,2}$ with sizes $m_e m_a + m_s m_i$ and $m_e m_a m_u + m_s mi$ specify the service completions at levels $l(n)$, respectively, with $2 \leq n \leq N-1$ and $N \leq n \leq K-1$. They have non-zero blocks $I(m_i) \otimes (M^0 \otimes \mu^\top)$ which specifies after service completion the initial phase of the service process for the next customer with probability μ.

The matrix $Q_{3,0}$ with size $m_r \times m_a$ is responsible for the transitions from the reset level $l(K)$ to the level $l(0)$, by specifying the initial phase of the MAP by the vector π_a.

Finally, we check the property that the infinitesimal matrix $Q^{(N)}$ has a zero row-sum. Indeed,

$$
\begin{aligned}
&Q_{1,0}e(l_0) + Q_{0,1}e(l_0+l_1) = 0_{l_0 \times 1},\\
&Q_{2,0}e(l_0) + Q_{1,1}e(l_0+l_1) + Q_{0,2}e(l_0+l_1) = 0_{l_0+l_1 \times 1},\\
&Q_{2,1}e(l_0+l_1) + Q_{1,1}e(l_0+l_1) + Q_{0,2}e(l_0+l_1) = 0_{l_0+l_1 \times 1},\\
&Q_{2,1}e(l_0+l_1) + Q_{1,1}e(l_0+l_1) + Q_{0,3}e(l_2+l_1) = 0_{l_0+l_1 \times 1},\\
&Q_{2,2}e(l_0+l_1) + Q_{1,2}e(l_2+l_1) + Q_{0,4}e(l_2+l_1) = 0_{l_2+l_1 \times 1},\\
&Q_{2,3}e(l_0+l_1) + Q_{1,2}e(l_2+l_1) + Q_{0,4}e(l_2+l_1) = 0_{l_2+l_1 \times 1},\\
&Q_{2,3}e(l_0+l_1) + Q_{1,2}e(l_2+l_1) + Q_{0,5}e(m_r) = 0_{l_2+l_1 \times 1},\\
&Q_{3,0}e(l_0) + Be(m_r) = 0_{m_r \times 1}.
\end{aligned}
$$

□

The structure of the generator $Q^{(N)}$ confirms that the process $\{X(t)\}_{t \geq 0}$ is a type of level-dependent quasi-birth-and-death process with special boundary conditions. Since CTMC (1) is irreducible, regular, and has finite state space E, the following limits exist for stationary probabilities:

$$\pi_x = \lim_{t \to \infty} \mathbb{P}[X(t) = x],\ x \in E.$$

Let $\pi = (\pi(0), \pi(1), \ldots, \pi(K))$, with sub vectors $\pi(n)$ partitioned with respect to levels $l(n), 0 \leq n \leq K$, denote the stationary probability vector of the infinitesimal matrix $Q^(N)$ defined in (2). This vector satisfies the system

$$\pi Q^{(N)} = 0,\ \pi e(|E|) = 1.$$

The stationary distribution can be obtained by solving a finite block almost three-diagonal system.

Theorem 1. *The stationary probability vectors $\pi(n), 0 \leq n \leq K$, can be calculated by*

$$\pi(n) = \pi(0) F_n, \tag{3}$$

where the matrices F_n satisfy the forward recurrent relations

$$F_0 = I(m_a),\ F_1 = R_1,\ F_n = F_{n-1}R_n,\ 2 \le n \le K. \tag{4}$$

The matrices R_n satisfy the backward recursion

$$R_n = \begin{cases} -Q_{0,1}(Q_{1,1} + R_2Q_{2,1})^{-1}, & n = 1, \\ -Q_{0,2}(Q_{1,1} + R_{n+1}Q_{2,1})^{-1}, & 2 \le n \le N-2, \\ -Q_{0,2}(Q_{1,1} + R_NQ_{2,2})^{-1}, & n = N-1, \\ -Q_{0,3}(Q_{1,2} + R_{N+1}Q_{2,3})^{-1}, & n = N, \\ -Q_{0,4}(Q_{1,2} + R_{n+1}Q_{2,3})^{-1}, & N+1 \le n \le K-2, \\ -Q_{0,4}Q_{1,2}^{-1}, & n = K-1, \\ -Q_{0,5}B^{-1}, & n = K. \end{cases} \tag{5}$$

The vector $\pi(0)$ is determined as the unique solution of the following system of linear equations:

$$\begin{aligned} &\pi(0)(Q_{1,0} + R_1Q_{2,0} + F_KQ_{3,0}) = 0, \\ &\pi(0)\sum_{n=0}^{K} F_ne(\cdot) = 1. \end{aligned} \tag{6}$$

Proof. Using a special structure of the infinitesimal generator $Q^{(N)}$, the system of balance equations is of the following form

$$\begin{aligned} &\pi(0)Q_{1,0} + \pi(1)Q_{2,0} + \pi(K)Q_{3,0} = 0, \\ &\pi(0)Q_{0,1} + \pi(1)Q_{1,1} + \pi(2)Q_{2,1} = 0, \\ &\pi(n-1)Q_{0,2} + \pi(n)Q_{1,1} + \pi(n+1)Q_{2,1} = 0,\ 2 \le n \le N-2, \\ &\pi(N-2)Q_{0,2} + \pi(N-1)Q_{1,1} + \pi(N)Q_{2,2} = 0, \\ &\pi(N-1)Q_{0,3} + \pi(N)Q_{1,2} + \pi(N+1)Q_{2,3} = 0, \\ &\pi(n-1)Q_{0,4} + \pi(n)Q_{1,2} + \pi(n+1)Q_{2,3} = 0,\ N+1 \le n \le K-2, \\ &\pi(K-2)Q_{0,4} + \pi(K-1)Q_{1,2} = 0, \\ &\pi(K-1)Q_{0,5} + \pi(K)B = 0. \end{aligned} \tag{7}$$

Initializing from the highest level $l(K)$ we get from the last equation of the system (7)

$$\pi(K) = -\pi(K-1)Q_{0,5}B^{-1} = \pi(K-1)R_K,$$

where $R_K = -Q_{0,5}B^{-1}$. Solving the equation of (7) for level $l(K-1)$ by substituting the previous relation gives

$$\pi(K-1) = -\pi(K-2)Q_{0,4}Q_{1,2}^{-1} = \pi(K-2)R_{K-1},$$

where $R_{K-1} = Q_{0,4}Q_{1,2}^{-1}$. The backward recursion for levels $l(n), N+1 \le n \le K-2$ gives by substituting $\pi(n) = \pi(n-1)R_n$ and $\pi(n+1) = \pi(n)R_{n+1}$ the recursive relation $R_n = -Q_{0,4}(Q_{1,2} + R_{n+1}Q_{2,2})^{-1}$. Repeating such a procedure for the levels $l(n), 2 \le n \le N$ leads to corresponding recursive relations in (5). Note that all inverse matrices exist due to the fact that the inverted matrices are sub-generators. For the level $l(0)$ from the first equation of (7) by substituting $\pi(1) = \pi(0)R_1$ and $\pi(K) = \pi(0)R_1R_2 \dots R_K = \pi(0)F_K$ we get the first equation in (7). Calculation of all levels probabilities can then be performed by expressions

$$\pi(1) = \pi(0)R_1 = \pi(0)F_1,$$
$$\pi(2) = \pi(1)R_2 = \pi(0)R_1R_2 = \pi(0)F_2,$$
$$\pi(n) = \pi(n-1)R_n = \pi(0)F_{n-1}R_n = \pi(0)F_n,\ 3 \le n \le K.$$

To calculate $\pi(0)$ the normalization condition $\sum_{n=0}^{K} \pi(n)e(\cdot) = \pi(0)\sum_{n=0}^{K} F_n e(\cdot) = 1$ is used. □

The relations (3)–(6) describe a quite simple algorithm for calculation stationary distribution π. The matrices involved in recursions (5) are non-negative. Furthermore, the matrices that are inverted after multiplication by -1 form irreducible sub-generators. Consequently, these inverses exist and are themselves non-negative matrices. This structural property guarantees the numerical stability of the proposed recursive algorithm.

As usual, calculating the stationary probabilities of states allows us to calculate various characteristics of the system. In the following, we present only those metrics that depend on stationary state probabilities which will be used in the loss function to calculate the optimal N-policy.

Corollary 1. *The probability distribution for the number of customers in the system is computed by*

$$\lim_{t\to\infty} \mathbb{P}[N(t) = n] = \pi(n)e(\cdot) = \begin{cases} \pi(n)e(l_0), & n = 0, \\ \pi(n)e(l_0 + l_1), & 1 \le n \le N-1, \\ \pi(n)e(l_2 + l_1), & N \le n \le K-1, \\ \pi(n)e(m_r), & n = K. \end{cases}$$

The average number of customers in the system $\bar{N}$ is calculated by

$$\overline{N} = \sum_{n=1}^{K} n\pi(n)e(\cdot).$$

The arrival rate of arrivals to a reset level $l(K)$

$$\Lambda_r = \pi(K-1)Q_{0,5}e(l_2 + l_1).$$

The probability that the server is in the startup state

$$P_u = \sum_{n=N}^{K-1} \pi(n) \begin{pmatrix} e(l_2) \\ 0_{l_1 \times 1} \end{pmatrix}.$$

The service interruption rate in levels $l(n), L+1 \le n \le K-1$ for incomplete regenerations

$$A_i = \sum_{n=L+1}^{N-1} \pi(n) \begin{pmatrix} 0_{l_0 \times l_0} \\ A^0 \otimes e(m_s) \otimes (\pi_a\, diag(\pi_e) \otimes I(m_a)) \end{pmatrix} e(l_0 + l_1)$$
$$+ \sum_{n=N}^{K-1} \pi(n) \begin{pmatrix} 0_{l_2 \times l_2} \\ A^0 \otimes e(m_s) \otimes (\pi_a\, diag(\pi_e) \otimes I(m_a)) \otimes \gamma^\top) \end{pmatrix} e(l_2 + l_1).$$

Let c_h be the holding cost per unit of time for each customer present in the system, c_r the fixed cost incurred for each visit to the states in $l(K)$ due to the system blocking in the reset states, c_u be the start-up cost per unit of time for the server performing the pre-service work , and c_i be the fixed cost incurred for the service activation plus service interruption in levels

$l(n), L+1 \le n \le K-1$ leading to incomplete regeneration. Combining the aforementioned cost elements yields the following total expected cost function per unit time:

$$\overline{C}^{(N)} = c_h \overline{N} + c_r \Lambda_r + c_u P_u + c_i A_i \tag{8}$$

that can be minimized with respect to the threshold N. To check the results obtained, we next present a numerical example.

Example 1. Consider the queueing system $MAP/PH/1/80$ with $L = \lfloor 0.1N \rfloor$. The costs are fixed at values $c_h = 0.001$, $c_r = 1500$, $c_u = 1.0$ and $c_i = 5.0$. Next, we specify the parameter values for PH service, startup, service interruption, and reset processes. The service process is assumed to be of Erlang type with $m_s = 5$ to model the almost deterministic soot burning process:

$$M = \begin{pmatrix} -200 & 200 & 0 & 0 & 0 \\ 0 & -200 & 200 & 0 & 0 \\ 0 & 0 & -200 & 200 & 0 \\ 0 & 0 & 0 & -200 & 200 \\ 0 & 0 & 0 & 0 & -200 \end{pmatrix}, M^0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 200 \end{pmatrix}, \mu = (1,0,0,0,0), \mu_s = 40,$$

which guaranties small variance $\sigma_s^2 = 0.000125$. Other parameters are fixed at the following values:

$$\Gamma = \begin{pmatrix} -1.50 & 0.01 \\ 0.05 & -2.80 \end{pmatrix}, \gamma = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix}, \Gamma^0 = (1.49, 2.75)^\top, \gamma_u = 1.821$$

$$A = \begin{pmatrix} -0.75 & 0.70 \\ 0.45 & -1.00 \end{pmatrix}, \alpha = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix}, A^0 = (0.05, 0.55)^\top, \alpha_i = 0.290$$

$$B = \begin{pmatrix} -2.50 & 1.09 \\ 0.50 & -3.50 \end{pmatrix}, \beta = \begin{pmatrix} 0.7 & 0.3 \end{pmatrix}, B^0 = (1.41, 3.00)^\top, \beta_r = 1.994.$$

The parameters of MAP with $m_e = 4$ are generated according to the matching procedure [28] in such a way that we have three values for the correlation function $\rho^{(\phi_e)}$ in each state $\phi_e = 1, \ldots, m_e$, For each correlation there are three types of the average arrival rate $\lambda^{(\phi_e)}$ that depend on the state ϕ_e of the random environment, which, in turn, reflects the speed level of the car.

- For $\rho^{(\phi_e)} = 0$,
  - (a) $\phi_e = 1$ for low speed driving, $\lambda^{(1)} = 1.50, \lambda_2^{(1)} = 1$:

$$D_0^{(1)} = \begin{pmatrix} -1.26062 & 1.26062 & 0. & 0. \\ 0. & -3.003 & 0. & 0. \\ 0. & 0. & -1.1255 & 1.1255 \\ 0. & 0. & 0. & -3.003 \end{pmatrix},$$

$$D_1^{(1)} = \begin{pmatrix} 0. & 0. & 0. & 0. \\ 0.98104 & 1.35596 & 0.249612 & 0.416388 \\ 0. & 0. & 0. & 0. \\ 0.350732 & 0.48477 & 0.812364 & 1.35514 \end{pmatrix}$$

(b) $\phi_e = 2$ for middle speed driving, $\lambda^{(2)} = 0.503, \lambda_2^{(2)} = 7$:

$$D_0^{(2)} = \begin{pmatrix} -0.608759 & 0.608759 & 0. & 0. \\ 0. & -1.00503 & 0. & 0. \\ 0. & 0. & -0.694814 & 0.694814 \\ 0. & 0. & 0. & -1.00503 \end{pmatrix},$$

$$D_1^{(2)} = \begin{pmatrix} 0. & 0. & 0. & 0. \\ 0.436563 & 0.284177 & 0.196538 & 0.0877476 \\ 0. & 0. & 0. & 0. \\ 0.132183 & 0.0860436 & 0.543945 & 0.242853 \end{pmatrix}$$

(c) $\phi_e = 3$ for high speed driving, $\lambda^{(3)} = 0.300, \lambda_2^{(3)} = 18$:

$$D_0^{(3)} = \begin{pmatrix} -0.419338 & 0.419338 & 0. & 0. \\ 0. & -0.600601 & 0. & 0. \\ 0. & 0. & -0.531171 & 0.531171 \\ 0. & 0. & 0. & -0.600601 \end{pmatrix},$$

$$D_1^{(3)} = \begin{pmatrix} 0. & 0. & 0. & 0. \\ 0.290171 & 0.125429 & 0.163614 & 0.0213859 \\ 0. & 0. & 0. & 0. \\ 0.0805024 & 0.0347979 & 0.4292 & 0.0561005 \end{pmatrix}.$$

- For $\rho^{(\phi_e)} = 0.2$,

(a) $\phi_e = 1$ for low speed driving, $\lambda^{(1)} = 1.50, \lambda_2^{(1)} = 2$:

$$D_0^{(1)} = \begin{pmatrix} -1.01798 & 1.01798 & 0. & 0. \\ 0. & -7.10518 & 0. & 0. \\ 0. & 0. & -0.543434 & 0.543434 \\ 0. & 0. & 0. & -0.992423 \end{pmatrix},$$

$$D_1^{(1)} = \begin{pmatrix} 0. & 0. & 0. & 0. \\ 0.927674 & 5.5472 & 0.345147 & 0.285163 \\ 0. & 0. & 0. & 0. \\ 0.0442613 & 0.264669 & 0.374269 & 0.309224 \end{pmatrix}$$

(b) $\phi_e = 2$ for middle speed driving, $\lambda^{(2)} = 0.503, \lambda_2^{(2)} = 15$:

$$D_0^{(2)} = \begin{pmatrix} -0.395125 & 0.395125 & 0. & 0. \\ 0. & -2.37791 & 0. & 0. \\ 0. & 0. & -0.236649 & 0.236649 \\ 0. & 0. & 0. & -0.385137 \end{pmatrix},$$

$$D_1^{(2)} = \begin{pmatrix} 0. & 0. & 0. & 0. \\ 0.353399 & 1.7734 & 0.154298 & 0.0968158 \\ 0. & 0. & 0. & 0. \\ 0.0188403 & 0.094543 & 0.16698 & 0.104773 \end{pmatrix}$$

(c) $\phi_e = 3$ for high speed driving, $\lambda^{(3)} = 0.300, \lambda_2^{(3)} = 26$:

$$D_0^{(3)} = \begin{pmatrix} -0.375492 & 0.375492 & 0. & 0. \\ 0. & -1.42104 & 0. & 0. \\ 0. & 0. & -0.323809 & 0.323809 \\ 0. & 0. & 0. & -0.338108 \end{pmatrix},$$

$$D_1^{(3)} = \begin{pmatrix} 0. & 0. & 0. & 0. \\ 0.311433 & 0.867176 & 0.232174 & 0.0102524 \\ 0. & 0. & 0. & 0. \\ 0.020495 & 0.0570676 & 0.249527 & 0.0110187 \end{pmatrix}.$$

- For $\rho^{(\phi_e)} = 0.4$,

  (a) $\phi_e = 1$ for low speed driving, $\lambda^{(1)} = 1.50, \lambda_2^{(1)} = 3$:

$$D_0^{(1)} = \begin{pmatrix} -1.28225 & 1.28225 & 0. & 0. \\ 0. & -12.3038 & 0. & 0. \\ 0. & 0. & -0.534109 & 0.534109 \\ 0. & 0. & 0. & -0.560595 \end{pmatrix},$$

$$D_1^{(1)} = \begin{pmatrix} 0. & 0. & 0. & 0. \\ 1.22538 & 10.5327 & 0.519959 & 0.0257841 \\ 0. & 0. & 0. & 0. \\ 0.0149355 & 0.128378 & 0.397567 & 0.0197149 \end{pmatrix}$$

  (b) $\phi_e = 2$ for middle speed driving, $\lambda^{(2)} = 0.503, \lambda_2^{(2)} = 24$:

$$D_0^{(2)} = \begin{pmatrix} -0.462225 & 0.462225 & 0. & 0. \\ 0. & -4.11776 & 0. & 0. \\ 0. & 0. & -0.205475 & 0.205475 \\ 0. & 0. & 0. & -0.208281 \end{pmatrix},$$

$$D_1^{(2)} = \begin{pmatrix} 0. & 0. & 0. & 0. \\ 0.439344 & 3.47458 & 0.201088 & 0.00274587 \\ 0. & 0. & 0. & 0. \\ 0.00585662 & 0.0463174 & 0.154004 & 0.00210293 \end{pmatrix}$$

  (c) $\phi_e = 3$ for high speed driving, $\lambda^{(3)} = 0.300, \lambda_2^{(3)} = 65$:

$$D_0^{(3)} = \begin{pmatrix} -0.282633 & 0.282633 & 0. & 0. \\ 0. & -2.46076 & 0. & 0. \\ 0. & 0. & -0.128182 & 0.128182 \\ 0. & 0. & 0. & -0.128468 \end{pmatrix},$$

$$D_1^{(3)} = \begin{pmatrix} 0. & 0. & 0. & 0. \\ 0.268168 & 2.06665 & 0.125661 & 0.000279606 \\ 0. & 0. & 0. & 0. \\ 0.00367142 & 0.028294 & 0.096288 & 0.000214249 \end{pmatrix}.$$

For numerical experiments, we study three main cases dependent on the infinitesimal matrix H for transitions between states ϕ_e of the random environment. Next, we distinguish two sub cases for the startup and the service interruption processes that influence the length

of the regeneration period: A short regeneration period, when the startup intensity is high while the interruption intensity is low,

$$\text{(a)}\quad \Gamma = \begin{pmatrix} -1.50 & 0.01 \\ 0.05 & -2.80 \end{pmatrix},\ \gamma = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix},\ \Gamma^0 = (1.49, 2.75)^\top,\ \gamma_u = 1.821$$
$$A = \begin{pmatrix} -0.75 & 0.70 \\ 0.45 & -1.00 \end{pmatrix},\ \alpha = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix},\ A^0 = (0.05, 0.55)^\top,\ \alpha_i = 0.290$$

and a long regeneration period, when the startup intensity is low while the interruption intensity is high,

$$\text{(b)}\quad \Gamma = \begin{pmatrix} -1.05 & 0.70 \\ 0.35 & -1.40 \end{pmatrix},\ \gamma = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix},\ \Gamma^0 = (0.35, 1.05)^\top,\ \gamma_u = 0.673$$
$$A = \begin{pmatrix} -0.75 & 0.50 \\ 0.25 & -1.00 \end{pmatrix},\ \alpha = \begin{pmatrix} 0.6 & 0.4 \end{pmatrix},\ A^0 = (0.25, 0.75)^\top,\ \alpha_i = 0.481.$$

- Case 1: The vehicle is operated mostly in the city,

$$H = \begin{pmatrix} -0.02 & 0.01 & 0.01 \\ 0.01 & -10.01 & 0.01 \\ 10 & 0.01 & -10.01 \end{pmatrix},\ \pi_e = (0.998004, 0.000998, 0.000998).$$

The dependence of the average cost on the N-policy in Case 1 is illustrated in Figure 1a,b for the correlations $\rho^{(\phi_e)} \in \{0, 0.2, 0.4\}$. The optimal N-policies and the corresponding optimal average costs are equal, respectively, in Figure 1a

$$(N^*, \bar{C}^{(N^*)}) = \{(40, 0.0903), (33, 0.0879), (26, 0.0798)\}$$

and in Figure 1b

$$(N^*, \bar{C}^{(N^*)}) = \{(50, 0.1699), (40, 0.1767), (23, 0.1968)\}.$$

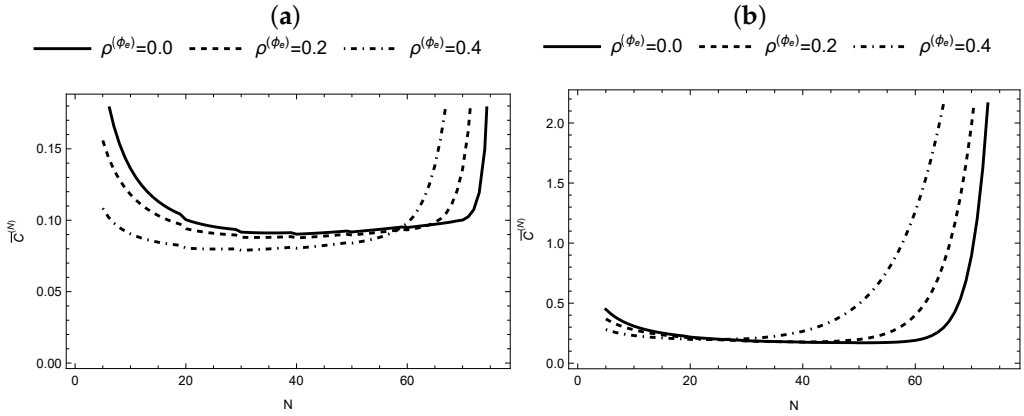

Figure 1. The expected cost $\overline{C}$ versus N and $\rho^{(\phi_e)}$ for short regeneration period (**a**) and long regeneration period (**b**) in Case 1.

- Case 2: Equal probability of using slow, middle and fast speed driving:

$$H = \begin{pmatrix} -1 & 0.5 & 0.5 \\ 0.5 & -1 & 0.5 \\ 0.5 & 0.5 & -1 \end{pmatrix},\ \pi_e = (0.333333, 0.0.333333, 0.333334).$$

The dependence of the average cost on the N-policy in Case 2 is illustrated in Figure 2a,b for the correlations $\rho^{(\phi_e)} \in \{0, 0.2, 0.4\}$. The optimal N-policies and the corresponding optimal average costs are equal, respectively, in Figure 2a

$$(N^*, \bar{C}^{(N^*)}) = \{(27, 0.0548), (24, 0.0515), (20, 0.0437)\}$$

and in Figure 2b

$$(N^*, \bar{C}^{(N^*)}) = \{(41, 0.0987), (40, 0.0961), (30, 0.0917)\}.$$

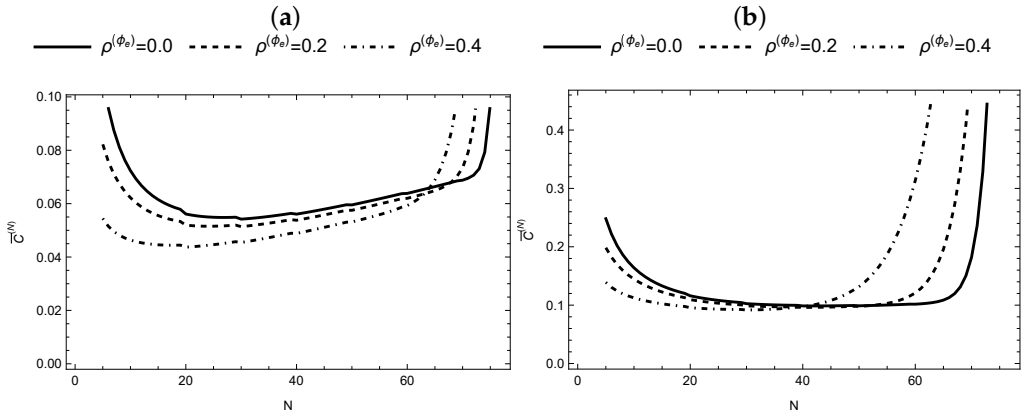

Figure 2. The expected cost $\overline{C}$ versus N and $\rho^{(\phi_e)}$ for short regeneration period (**a**) and long regeneration period (**b**) in Case 2.

- Case 3: The vehicle is mainly driven on highways,

$$H = \begin{pmatrix} -10.01 & 0.01 & 10 \\ 0.01 & -10.01 & 10 \\ 0.01 & 0.01 & -0.02 \end{pmatrix}, \ \pi_e = (0.000998, 0.000998, 0.998004).$$

The dependence of the average cost on the N-policy in Case 3 is illustrated in Figure 3 for the correlations $\rho^{(\phi_e)} \in \{0, 0.2, 0.4\}$. The optimal N-policies and the corresponding optimal average costs are equal, respectively, in Figure 2a

$$(N^*, \bar{C}^{(N^*)}) = \{(20, 0.0282), (17, 0.0274), (10, 0.0198)\},$$

and in Figure 3b

$$(N^*, \bar{C}^{(N^*)}) = \{(28, 0.0499), (27, 0.0489), (20, 0.0401)\}.$$

Figures 1a,b–3a,b for average costs allow us to draw the following conclusions. As the correlation of the arrival flow increases, the value of the optimal threshold N^* shifts to the left. With high correlation, the probability of a "burst" exceeding K before regeneration increases. This forces a costly reset. Lowering N creates a safety margin, i.e., we should regenerate earlier to avoid hitting capacity during bursts. The values of the optimal thresholds and the corresponding average costs vary significantly for different values of the correlation coefficient. Consequently, this factor must be taken into account in real-world applications. As expected, the average cost increases as the system load increases; in our case, this refers to the rate of the soot accumulation. When driving on highways, this rate is much lower than when driving in urban conditions. We can also observe that the optimal threshold value shifts to the right in sub case (a) for all correlation coefficients and in sub case (b) for $\rho^{(\phi_e)} = \in \{0, 0.2\}$. But in sub case (b) with high correlation $\rho^{(\phi_e)} = 0.4$, changes

in the optimal threshold as the system load increases do not appear to be monotonic. This could be explained in the following way. When the accumulation is very slow, soot particles sit in the filter for extremely long periods before reaching the regeneration threshold. During this time, they contribute to increasing the exhaust back-pressure, reducing the efficiency and fuel economy of the engine, which is described by the cost component c_h. In this case, it is better to have frequent and small regenerations than to let the soot linger. At high accumulation rates, the total back-pressure exposure per cycle is much lower for the same N. This means that we can afford to allow soot to build up to higher levels before regenerating. Each regeneration incurs a fixed service activation-interruption cost c_i. This cost is independent of how much soot is burned. At a high arrival rate, we can increase N significantly while keeping cycle lengths reasonable, dramatically reducing the number of regenerations per day, and saving setup fuel. Thus, the optimal threshold N^* has its lowest value when the intensity of the arrival stream is low, whilst its correlation is high.

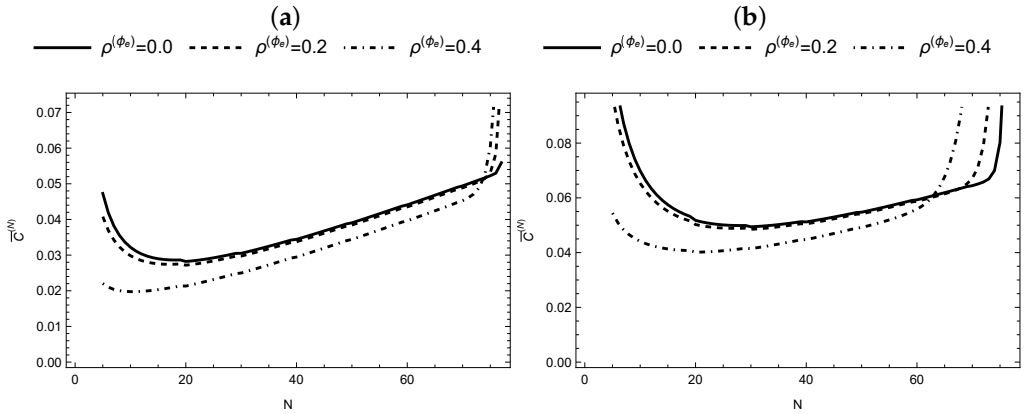

Figure 3. The expected cost $\overline{C}$ versus N for $\rho^{(\phi_e)}$ for short regeneration period (**a**) and long regeneration period (**b**) in Case 3.

Finally we note that the solution for the optimal threshold requires approximately 3.0×10^{-7} s by using the Mathematica 14.3 (Wolfram Research) package on a computer with Intel Core i5-6400 CPU 2.7 GHz.

4. Conclusions

In this paper, we present analytical steady-state results for a N-policy $MAP/PH/1/K$ queueing system. As a practical example, we used this queueing system as a model for the soot accumulation and burning process in DPF with the aim to find optimal regeneration policy. Our research extends previous studies on classical N-policy queueing systems by combining such features as phase type distributed setup, interruption, reset times, and random environment. Using matrix-analytic methods, we derive stationary distributions and a cost optimization framework. Numerical results reveal how correlation in soot accumulation process, setup and service interruption rates, as well as driving conditions affect the average cost and optimal thresholds. For example, as the correlation coefficient increases and the soot accumulation rate decreases in case of a short regeneration period, the optimal threshold N decreases. We note that validation requires real DPF data and industry collaboration for parameter estimation. Future directions include multi-rate servers, batch arrivals, age-dependent degradation, correlated interruptions, dynamic threshold policies, multi-objective optimization, and reinforcement learning.

Author Contributions: Conceptualization, D.E. and N.S.; methodology, D.E., J.S. and Z.G.; software, D.E. and N.S.; validation, D.E. and N.S.; formal analysis, D.E. and N.S.; investigation, D.E.; resources, D.E., Z.G. and J.S.; data curation, N.S.; writing—original draft preparation, D.E. and N.S.; writing—review and editing, D.E. and N.S.; visualization, D.E. and N.S.; supervision, D.E. and Z.G.; project administration, D.E. and Z.G.; funding acquisition, D.E. and Z.G. All authors have read and agreed to the published version of the manuscript.

Funding: The research is funded by the Federal Ministry for Women, Science and Research, as well as by the Hungarian Ministry of Culture and Innovation in framework of the project 121öu2 of the Austro-Hungarian Campaign.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MAP	Markovian Arrival Process
PH	Phase type
DPF	Diesel Particulate Filter

References

1. Artalejo, J.; Gómez-Corral, A.; He Q.-M. Markovian arrivals in stochastic modeling: A survey and some new results. *Stat. Oper. Res. Trans.* **2010**, *34*, 101–156.
2. Chakravarthy, S.R. *Queueing Models in Services—Analytical and Simulation Approach*; 2021. https://doi.org/10.1002/9781119755234.ch2.
3. Yadin, M.; Naor, P. Queueing Systems with a Removable Service Station. *J. Oper. Res. Soc.* **1963**, *14*, 393–405. https://doi.org/10.1057/jors.1963.63.
4. Bell, C.E. Characterization and Computation of Optimal Policies for Operating an M/G/1 Queuing System with Removable Server. *Oper. Res.* **1971**, *19*, 208–218. https://doi.org/10.1287/opre.19.1.208.
5. Doshi, B.T. Queueing systems with vacations—A survey. *Queueing Syst.* **1986**, *1*, 29–66. https://doi.org/10.1007/bf01149327.
6. Tian, N.; Zhang, D.; Cao, C. The $GI/M/1$ queue with exponential vacations. *Queueing Syst.* **1989**, *5*, 331–344. https://doi.org/10.1007/bf01225323.
7. Choudhury, G.; Paul, M. A batch arrival queue with an additional service channel under N-policy. *Appl. Math. Comput.* **2004**, *156*, 115–130. https://doi.org/10.1016/j.amc.2003.07.006.
8. Jayachitra, P.; Albert, A.J. Recent developments in queueing models under *N*-policy: A short survey. *Int. J. Math. Arch.* **2014**, *5*.
9. Yen, T.C.; Wang, K.H.; Chen, J.Y. Optimization Analysis of the N Policy M/G/1 Queue with Working Breakdowns. *Symmetry* **2020**, *12*, 583. https://doi.org/10.3390/sym12040583.
10. Günalay, Y.; Gupta, D. Threshold start-up control policy for polling systems. *Queueing Syst.* **1998**, *29*, 399–421. https://doi.org/10.1023/a:1019152601966.
11. Azhagappan, A.; Deepa, T. Performance Analysis of an $M/M/1$ Queue with N-policy Interrupted Closedown Preventive Maintenance Balking and Feedback. *Appl. Appl. Math. Int. J. (AAM)* **2020**, *15*, 1–12. Available online: https://digitalcommons.pvamu.edu/aam/vol15/iss1/1 (accessed on 8th April 2026).
12. Rao, A.A.; Devi, V.N.R.; Chandan, K. Optimal strategy analysis of *N*-policy Two-Phase $M/E_k/1$ Vacation Queueing system with Server Start-Up, Time-Out and Breakdown. *J. Phys. Conf. Ser.* **2019**, *1344*, 012025. https://doi.org/10.1088/1742-6596/1344/1/012025.
13. Wang, K.H. Optimal control of a removable and non-reliable server in an M/M/1 queueing system with exponential startup time. *Math. Methods Oper. Res. (ZOR)* **2003**, *58*, 29–39. https://doi.org/10.1007/s001860300275.
14. Efrosinin, D.; Semenova, O. Optimal control of $M/M/1$ queueing system with constant retrial rate and non-reliable removable server. In Proceedings of the 2009 International Conference on Ultra Modern Telecommunications and Workshops, St. Petersburg, Russia, 12–14 October 2009; pp. 1–6. https://doi.org/10.1109/icumt.2009.5345415.

15. Efrosinin, D.; Semenova, O. Matrix-analytical approach to analysis of a single-server retrial queue with non-reliable removable server. In Proceedings of the International Congress on Ultra Modern Telecommunications and Control Systems, Moscow, Russia, 18–20 October 2010; pp. 1145–1149. https://doi.org/10.1109/icumt.2010.5676526.
16. Efrosinin, D.V.; Semenova, O.V. An $M/M/1$ system with an unreliable device and a threshold recovery policy. *J. Commun. Technol. Electron.* **2010**, *55*, 1526–1531. https://doi.org/10.1134/s1064226910120260.
17. Efrosinin, D.; Winkler, A. Queueing system with a constant retrial rate, non-reliable server and threshold-based recovery. *Eur. J. Oper. Res.* **2011**, *210*, 594–605. https://doi.org/10.1016/j.ejor.2010.09.040.
18. Ayyappan, G.; Thilagavathy, K. Analysis of MAP/PH/1 queueing system with catastrophic delay action, standby server, balking, working vacation and vacation interruption under N-policy. *Int. J. Math. Model. Numer. Optim.* **2023**, *13*, 223–243. https://doi.org/10.1504/ijmmno.2023.132290.
19. Ayyappan, G.; Gurulakshmi, G.A.A. Analysis of $MAP/PH/1$ queue with differentiated vacation, vacation interruption under N-policy, optional service, breakdown, repair, setup and discouragement of customers. *Int. J. Math. Oper. Res.* **2024**, *27*, 415–457. https://doi.org/10.1504/ijmor.2024.138463.
20. Sreenivasan, C.; Chakravarthy, S.R.; Krishnamoorthy, A. $MAP/PH/1$ queue with working vacations, vacation interruptions and N policy. *Appl. Math. Model.* **2013**, *37*, 3879–3893. https://doi.org/10.1016/j.apm.2012.07.054.
21. Bogachev, M.I.; Pyko, N.S.; Tymchenko, N.; Pyko, S.A.; Markelov, O.A. Approximate waiting times for queuing systems with variable cross-correlated arrival rates. *Phys. A Stat. Mech. Its Appl.* **2024**, *654*, 130152. https://doi.org/10.1016/j.physa.2024.130152.
22. Kouvatsos, D.; Fretwell, R. Closed Form Performance Distributions of a Discrete Time $GI^G/D/1/N$ Queue with Correlated Traffic. In *Data Communications and Their Performance*; Springer: New York, NY, USA, 1996; pp. 141–163. https://doi.org/10.1007/978-0-387-34942-8_10.
23. Chiang, C.J.; Kuo, T.F.; Halim, A.; Cheng, S.C.; Ku, Y. Dynamic Modeling of a Diesel Oxidation Catalyst and Diesel Particulate Filter Aftertreatment System for Regeneration Control Development. In *Proceedings of the JSAE/SAE Small Engine Technologies Conference & Exhibition*; 2017; Jakarta, Indonesia, November 5, 2017; https://doi.org/10.4271/2017-32-0105.
24. D'Aniello, F.; Rossomando, B.; Arsie, I.; Pianese, C. Development and Experimental Validation of a Control Oriented Model of a Catalytic DPF. In *Proceedings of the WCX SAE World Congress Experience*; Detroit, Michigan, United States, April 9; 2019 https://doi.org/10.4271/2019-01-0985.
25. Dudin, A.; Kazimirsky, A.; Klimenok, V.; Breuer, L.; Krieger, U. The Queueing Model $MAP|PH|1|N$ with Feedback Operating in a Markovian Random Environment. *Austrian J. Stat.* **2016**, *34*, 101–110. https://doi.org/10.17713/ajs.v34i2.403.
26. Al-Ansari, Y. Parameters and the Rate of Soot Emitted from Diesel Engine. *Al-Qadisiyah J. Eng. Sci.* **2009**, *2*.
27. Pajdowski, P.; Puchałka, B. The Process of Diesel Particulate Filter Regeneration under Real Driving Conditions. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *214*, 012114. https://doi.org/10.1088/1755-1315/214/1/012114.
28. Horvath, G. Matching marginal moments and lag autocorrelations with MAPs. In Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS, Bratislava, Slovakia, 9–11 December 2014; pp. 1–10. https://doi.org/10.4108/icst.valuetools.2013.254368.