

Article

Multi-Class ICU Bed Reservation Under Bursty Arrivals: A Generalized Loss Model Framework with Fairness Optimization

Wei Tian , Anqi Wang , Hanzhi Zhang  and Jingjin Wu * 

Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, Department of Statistics and Data Science, Beijing Normal-Hong Kong Baptist University, Zhuhai 519087, China; s230202702@mail.bnbu.edu.cn (W.T.); v530202705@mail.bnbu.edu.cn (A.W.); t330005063@mail.bnbu.edu.cn (H.Z.)

* Correspondence: jingjinwu@bnb.edu.cn

Abstract

The effective allocation of Intensive Care Unit (ICU) beds is critical for balancing timely access across patient groups of differing clinical urgency, particularly during demand surges. This paper extends the threshold-based loss-queuing framework to three patient classes, emergency, semi-urgent, and elective, governed by a two-tier coordinate reservation policy with thresholds (k_1, k_2) . Under Poisson arrivals, we derive an exact product form steady-state distribution and closed-form blocking probabilities for all three classes; under Interrupted Poisson Process (IPP) arrivals, we construct a block-tridiagonal Markov chain over the full (i, j, l, ξ) state space and obtain dimensionally consistent blocking formulas via a lifted rate matrix. Fitted to the publicly available MIMIC-IV Medical ICU dataset, the IPP captures bursty emergency arrival patterns with mean and variance deviations below 0.1%. A comprehensive parameter sensitivity sweep over traffic utilization $\rho \in [0.40, 0.95]$ and burstiness index $z \in [1.0, 5.0]$ identifies three distinct operating regimes and yields policy recommendation charts for direct clinical use. An adaptive optimization framework selects (k_1^*, k_2^*) to minimize a weighted blocking loss subject to fairness constraints, achieving a Jain fairness index above 0.999 throughout the high-load region. Analytical predictions are validated against discrete-event simulation with a 100% pass rate at a ± 1.5 percentage-point criterion. We further demonstrate that the blocking probabilities for all three classes are insensitive to the LoS distribution beyond its mean across six service-time distributions spanning coefficients of variation from 0.45 to 2.24, with a 100% pass rate across all 72 (class, distribution, setting) combinations, broadening the model's applicability to diverse real-world scenarios. The findings provide actionable guidance for ICU managers in determining fair and efficient three-tier bed reservation thresholds.

Keywords: ICU bed management; Interrupted Poisson Process; queueing theory; three-class loss model; fairness optimization

MSC: 90B22



Academic Editor: Steve Drekic

Received: 3 April 2026

Revised: 9 May 2026

Accepted: 14 May 2026

Published: 17 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Intensive Care Units (ICUs) represent one of the most resource-constrained environments in modern healthcare: technologically sophisticated, operationally costly, and in perpetually high demand. The effective management of ICU bed capacity is therefore a critical operational challenge, particularly when sudden surges in patient demand, arising

from infectious disease outbreaks, mass casualty events, or seasonal respiratory crises, threaten to overwhelm available resources [1]. During periods of elevated occupancy, thoughtful bed management must simultaneously protect timely access for the most critically ill patients and preserve reasonable admission opportunities for lower-urgency cases, a dual objective that requires both analytical rigor and operational flexibility.

A widely adopted strategy is to partition ICU beds into priority tiers: a certain number of beds are reserved exclusively for the highest-priority class, with the remainder shared among all classes subject to a system-wide capacity constraint. Such threshold-based reservation policies have been studied extensively in the context of two patient classes, most commonly emergency versus elective [2,3]. De Bruin et al. [4] showed, using real patient flow data from 24 wards, that capacity pooling and structured reservation can simultaneously improve efficiency and service levels. However, real clinical triage systems, such as the Australasian Triage Scale and the Manchester Triage System widely adopted in the United Kingdom and elsewhere [5,6], routinely stratify patients into three or more priority levels. A binary emergency/elective split therefore fails to capture the intermediate cohort of semi-urgent patients (e.g., urgent step-down transfers, time-sensitive but non-critical admissions) who require a distinct admission policy. Extending the reservation framework to three patient classes introduces a richer two-tier threshold structure and opens new questions about equitable resource allocation across all groups.

Queueing models are the natural analytical tool for evaluating ICU bed management policies. Loss models, where patients who cannot immediately obtain a bed are assumed lost rather than queued, are particularly appropriate for ICUs, since clinically unsafe delays motivate the use of blocking probability as the key performance metric rather than expected waiting time [7]. Various studies have exploited Erlang loss and multi-class variants to quantify the trade-off between bed utilization and patient rejection. Bai et al. [8] performed a detailed analysis of patient arrival patterns in ICU bed planning, while Kim et al. [9] demonstrated that elective patient arrivals deviate systematically from the Poisson assumption. More recently, Dudin et al. [10] studied multi-server queues driven by Markov-modulated arrival processes, providing a related analytical framework for MMPP-driven queueing systems. A limitation shared by much of the existing literature is the reliance on Poisson arrivals, which cannot capture the bursty, clustered admission patterns observed in real ICUs during demand surges [11,12].

A more faithful representation of bursty patient flows is provided by the Interrupted Poisson Process (IPP) [13], a two-state Markov-modulated Poisson process in which arrivals occur at a positive rate during “ON” periods and cease entirely during “OFF” periods. The IPP was originally introduced as an approximation for overflow traffic in loss networks and has since been applied across telecommunications and healthcare. Machihara [14] analyzed the covariance structure of the interrupted MMPP and characterized it as a model for highly fluctuated traffic, and recent studies confirm its practical value: Wu et al. [15] demonstrated that IPPs fitted via two-moment matching accurately reproduce real ICU patient arrival statistics, including the elevated peakedness characteristic of emergency admission streams. Crucially, the superposition of independent IPPs remains a Markov-modulated Poisson process (MMPP) [16], making the multi-class extension analytically tractable.

In this paper, we develop a comprehensive analytical framework for a three-class ICU bed reservation system under Interrupted Poisson Process arrivals. Three patient classes, emergency, semi-urgent, and elective, are governed by a two-tier coordinate threshold policy: k_1 beds are reserved exclusively for Class-1 (emergency), and an additional k_2 beds are reserved jointly for Classes 1 and 2 (emergency and semi-urgent). The coordinate threshold structure, in contrast to a total-occupancy threshold, admits an exact product form steady-state distribution under Poisson arrivals (Proposition 1) and a tractable block-

tridiagonal Markov chain under IPP arrivals. We validate the analytical model against the publicly available MIMIC-IV Medical ICU dataset [17], and conduct systematic sensitivity and insensitivity studies.

The novel contributions of this paper are:

- We extend the threshold-based reservation model of [15] to three patient classes under a two-tier coordinate threshold policy. We derive closed-form blocking probabilities B_1, B_2, B_3 via an exact product form steady-state distribution, and establish that the coordinate threshold structure is essential for analytical tractability.
- We develop the corresponding IPP-driven model: three independent IPPs are superimposed into an eight-phase MMPP, and we derive dimensionally correct blocking probability formulas using a lifted rate matrix that operates on the full (i, j, l, ξ) state space. The proof explicitly verifies dimensional consistency, correcting a subtle error present in earlier two-class formulations.
- We perform a systematic parameter sensitivity sweep across traffic utilization $\rho \in [0.40, 0.95]$ and burstiness index $z \in [1.0, 5.0]$, producing policy recommendation charts that allow ICU managers to read off the optimal reservation threshold directly from observed system parameters.
- We validate the analytical model against the MIMIC-IV Medical ICU dataset [17,18], a large, publicly accessible US critical-care database, across all three patient classes and a range of reservation thresholds, achieving a 100% pass rate against a ± 1.5 percentage-point criterion.
- We extend the length of stay (LoS) insensitivity study of [15] to six service-time distributions spanning coefficients of variation from 0.45 to 2.24, confirming that blocking probabilities depend on the LoS distribution only through its mean, a practically important result that broadens the model's applicability to settings where the LoS distribution is unknown or non-exponential.

2. System Model

The two-class model admits a natural and practically motivated extension to three patient classes. Many ICU contexts distinguish not only emergency and elective patients but also a third, intermediate cohort of *semi-urgent* patients (e.g., step-down transfers, urgent but non-critical admissions) whose admission priority falls between the emergency and routine elective groups. This section develops the full analytical framework, state space, product form steady-state distribution, closed-form blocking probabilities, and a two-tier fairness optimization, for this three-class system under both Poisson and Interrupted Poisson Process (IPP) arrivals.

2.1. System Description and Two-Tier Reservation Policy

2.1.1. Patient Classes and Priority Structure

The ICU contains n beds shared among three patient classes:

- Class 1: Emergency/unplanned critical** (highest priority): Patients requiring immediate life-sustaining intervention; admission must be guaranteed to the greatest possible extent.
- Class 2: Semi-urgent/urgent non-critical** (medium priority): Patients requiring timely admission, e.g., step-down transfers or urgent non-critical cases, but whose care can tolerate a brief delay if the system is heavily loaded.
- Class 3: Routine elective/planned** (lowest priority): Patients with scheduled admissions whose arrival time is flexible and for whom rescheduling, while undesirable, is clinically acceptable.

2.1.2. Two-Tier Coordinate Reservation Policy

Two integer reservation thresholds $k_1 \geq 1$ and $k_2 \geq 1$ with $k_1 + k_2 \leq n$ are imposed. Following the two-class model [15], the thresholds operate on the *absolute count* of each lower-priority class already in the system, not on the total occupancy. Specifically:

- **Class-1 reservation (k_1 beds).** A new Class-2 patient is *refused admission* whenever the number of Class-2 patients currently in the ICU reaches $n - k_1$, i.e., whenever $j \geq n - k_1$. This guarantees that at least k_1 beds remain available for Class-1 (emergency) patients at all times.
- **Class-1&2 reservation (k_2 additional beds).** A new Class-3 patient is *refused admission* whenever the number of Class-3 patients currently in the ICU reaches $n - k_1 - k_2$, i.e., whenever $l \geq n - k_1 - k_2$. This reserves a further k_2 beds jointly for Class-1 and Class-2 patients.

These two coordinate thresholds together reserve at least k_1 beds for Class-1 and at least $k_1 + k_2$ beds jointly for Classes 1 and 2. In terms of per-class capacity limits, the n beds are partitioned conceptually into:

- A *shared pool* of $n - k_1 - k_2$ beds accessible to all three classes.
- A *Class-1&2 pool* of k_2 beds accessible to Classes 1 and 2 only.
- A *Class-1 pool* of k_1 beds accessible to Class-1 only.

Table 1 summarizes the admission conditions for each class in terms of the current occupancy counts (i, j, l) .

Table 1. Admission conditions under the two-tier coordinate reservation policy. A “✓” indicates that the class is admitted from state (i, j, l) ; “×” indicates refusal.

Condition on Current State	Class-1	Class-2	Class-3
$i + j + l < n$ and $j < n - k_1$ and $l < n - k_1 - k_2$	✓	✓	✓
$i + j + l < n$ and $j < n - k_1$ and $l = n - k_1 - k_2$	✓	✓	×
$i + j + l < n$ and $j = n - k_1$ and $l < n - k_1 - k_2$	✓	×	✓
$i + j + l < n$ and $j = n - k_1$ and $l = n - k_1 - k_2$	✓	×	×
$i + j + l = n$ (system full)	×	×	×

Remark 1 (Per-class cap model vs. occupancy-priority model). *The thresholds operate on the count of each individual class, not on the total occupancy $s = i + j + l$. This is a per-class cap model: each lower-priority class is capped at an absolute number of simultaneous beds, irrespective of how many high-priority patients are present. As a consequence, a Class-2 arrival may be admitted even when $s \geq n - k_1$, provided $j < n - k_1$ (the Class-2 cap has not been reached) and a bed is free.*

This differs from an alternative occupancy-priority model in which Class-2 is blocked whenever $s \geq n - k_1$ (i.e., only the top k_1 occupancy levels are reserved). The per-class cap model is chosen here because it is consistent with the two-class formulation in [15] and because it admits an exact product form steady-state distribution (see Proposition 1 and Remark 4). The occupancy-priority model does not possess a product form and must be solved numerically.

2.1.3. Notation

Table 2 lists the symbols used throughout this section.

Table 2. Notation for the three-class model.

Symbol	Meaning
n	Total number of ICU beds
k_1	Beds reserved exclusively for Class-1
k_2	Additional beds reserved jointly for Classes 1 and 2
i, j, l	Number of Class-1, -2, -3 patients in system
$s = i + j + l$	Total occupancy
λ_r	Poisson arrival rate of Class r , $r \in \{1, 2, 3\}$
μ_r	Service (discharge) rate per Class- r patient
$\rho_r = \lambda_r / \mu_r$	Traffic intensity of Class r
$\pi(i, j, l)$	Steady-state probability of state (i, j, l) (Poisson case)
$B_r(k_1, k_2)$	Blocking probability for Class r (Poisson case)
ω_r, γ_r	IPP off $\rightarrow$ on and on $\rightarrow$ off transition rates for Class r
Q_r	2×2 generator of Class- r IPP
Λ_r	2×2 arrival rate matrix of Class- r IPP
$\hat{Q}$	Block-tridiagonal generator of the full MMPP-driven system
$\hat{\pi}$	Steady-state row vector of the MMPP-driven system
$P_{\text{block},r}$	Blocking probability for Class r (IPP case)
$J(k_1, k_2)$	Jain fairness index
$\mathbf{e}$	Column vector of all ones (dimension clear from context)
$\otimes$	Kronecker product
$\oplus$	Kronecker sum ($A \oplus B = A \otimes I + I \otimes B$)

2.2. Rejection Rates Under Poisson Arrivals

2.2.1. State Space

Under Poisson arrivals and exponential lengths of stay, the system state is (i, j, l) , where $i, j, l \geq 0$ denote the number of Class-1, Class-2, and Class-3 patients currently occupying beds, respectively.

The two-tier coordinate reservation policy enforces the following constraints on admissible states:

- $j \leq n - k_1$ (the number of Class-2 patients cannot exceed the Class-2 capacity cap $n - k_1$).
- $l \leq n - k_1 - k_2$ (the number of Class-3 patients cannot exceed the Class-3 capacity cap $n - k_1 - k_2$).
- $i + j + l \leq n$ (total bed capacity).

Note that the constraint on j is *independent of i* and the constraint on l is *independent of i and j* : these are coordinate thresholds.

The state space is therefore

$$\Omega = \{(i, j, l) \in \mathbb{Z}_{\geq 0}^3 : i + j + l \leq n, j \leq n - k_1, l \leq n - k_1 - k_2\}. \quad (1)$$

Remark 2. An alternative (incorrect) formulation would impose $j \leq n - k_1 - i$ and $l \leq n - k_1 - k_2 - i - j$. Those constraints are too restrictive: they exclude reachable states such as $(2, n - k_1 - 1, 0)$, which is accessible when $n - k_1 - 1$ Class-2 patients arrive first and two Class-1 patients arrive subsequently. The correct coordinate threshold policy depends only on the absolute count of each lower-priority class, consistent with [15].

For given i , the admissible pairs (j, l) satisfy

$$0 \leq j \leq n - k_1, \quad 0 \leq l \leq n - k_1 - k_2, \quad i + j + l \leq n. \quad (2)$$

Remark 3. When $k_1 = k_2 = 0$, the state space reduces to the unconstrained $M/M/n/n$ loss model for three classes. When $k_2 = 0$ and $k_1 \geq 1$, Classes 2 and 3 share the same coordinate cap ($j \leq n - k_1$ and $l \leq n - k_1$), so they receive symmetric treatment under a single-tier reservation for Class-1. Setting $k_2 = 0$ and merging Classes 2 and 3 recovers the two-class model of [15].

2.2.2. Product Form Steady-State Distribution

Proposition 1 (Product form distribution). Consider the three-class ICU model with Poisson arrivals at rates $\lambda_1, \lambda_2, \lambda_3$ and independent exponential service times with rates μ_1, μ_2, μ_3 . Let $\rho_r = \lambda_r / \mu_r$ for $r = 1, 2, 3$. The unique steady-state distribution over Ω is

$$\pi(i, j, l) = C \cdot \frac{\rho_1^i}{i!} \cdot \frac{\rho_2^j}{j!} \cdot \frac{\rho_3^l}{l!}, \quad (i, j, l) \in \Omega, \quad (3)$$

where the normalizing constant C is given by

$$C^{-1} = \sum_{(i,j,l) \in \Omega} \frac{\rho_1^i}{i!} \cdot \frac{\rho_2^j}{j!} \cdot \frac{\rho_3^l}{l!}. \quad (4)$$

Proof. The Markov chain on Ω has transition rates

$$(i, j, l) \rightarrow (i + 1, j, l), \quad \text{rate } \lambda_1, \quad \text{if } i + j + l < n, \quad (5)$$

$$(i, j, l) \rightarrow (i - 1, j, l), \quad \text{rate } i\mu_1, \quad \text{if } i \geq 1, \quad (6)$$

$$(i, j, l) \rightarrow (i, j + 1, l), \quad \text{rate } \lambda_2, \quad \text{if } j < n - k_1 \text{ and } i + j + l < n, \quad (7)$$

$$(i, j, l) \rightarrow (i, j - 1, l), \quad \text{rate } j\mu_2, \quad \text{if } j \geq 1, \quad (8)$$

$$(i, j, l) \rightarrow (i, j, l + 1), \quad \text{rate } \lambda_3, \quad \text{if } l < n - k_1 - k_2 \text{ and } i + j + l < n, \quad (9)$$

$$(i, j, l) \rightarrow (i, j, l - 1), \quad \text{rate } l\mu_3, \quad \text{if } l \geq 1. \quad (10)$$

We verify that (3) satisfies detailed balance along every coordinate direction. For the Class-1 coordinate, for any $(i, j, l) \in \Omega$ with $i + j + l < n$:

$$\pi(i, j, l) \lambda_1 = C \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!} \lambda_1 = C \frac{\rho_1^{i+1}}{(i+1)!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!} (i+1)\mu_1 = \pi(i+1, j, l) (i+1)\mu_1, \quad (11)$$

using $\lambda_1 = \rho_1\mu_1$. For the Class-2 coordinate, for any $(i, j, l) \in \Omega$ with $j < n - k_1$ and $i + j + l < n$:

$$\pi(i, j, l) \lambda_2 = C \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!} \lambda_2 = C \frac{\rho_1^i}{i!} \frac{\rho_2^{j+1}}{(j+1)!} \frac{\rho_3^l}{l!} (j+1)\mu_2 = \pi(i, j+1, l) (j+1)\mu_2. \quad (12)$$

The identical calculation holds for the Class-3 coordinate. Crucially, because the threshold on j is a coordinate threshold (depending only on j , not on i), the arrival transition $(i, j, l) \rightarrow (i, j + 1, l)$ exists in Ω whenever $j < n - k_1$ and $i + j + l < n$, regardless of the value of i . This ensures that every interior arrival transition is paired with a departure transition to the same pair of states, and detailed balance holds at each such pair. At the boundary of Ω , arrivals that would leave Ω are simply lost (the loss network structure). Detailed balance therefore implies the global balance equations, and the unique solution normalized to sum to unity is (3). $\square$

Remark 4. The product form result holds specifically because the admission thresholds are co-ordinate thresholds. If instead Class-2 were blocked at total occupancy $i + j + l \geq n - k_1$ (a total-occupancy threshold), the classes would be coupled through the total count, detailed balance would fail, and no product form solution would exist. The coordinate threshold formulation is therefore essential both for analytical tractability and for consistency with the two-class model [15].

2.2.3. Blocking Probabilities

By the PASTA property (Poisson Arrivals See Time Averages), the blocking probability experienced by a Class- r arrival equals the steady-state probability of being in a state from which Class- r cannot be admitted.

Class-1 (Emergency) Blocking. Class-1 is blocked if and only if every bed is occupied: $i + j + l = n$. Defining

$$\mathcal{S}_1 = \{(i, j, l) \in \Omega : i + j + l = n\}, \quad (13)$$

the Class-1 blocking probability is

$$B_1(k_1, k_2) = \sum_{(i,j,l) \in \mathcal{S}_1} \pi(i, j, l) = \frac{\sum_{(i,j,l) \in \mathcal{S}_1} \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!}}{\sum_{(i,j,l) \in \Omega} \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!}}. \quad (14)$$

Class-2 (Semi-urgent) Blocking. Under the coordinate threshold, Class-2 is blocked from state (i, j, l) whenever $j = n - k_1$ (the Class-2 count has reached its cap) or $i + j + l = n$ (system full). Defining

$$\mathcal{S}_2 = \{(i, j, l) \in \Omega : j = n - k_1 \text{ or } i + j + l = n\}, \quad (15)$$

the Class-2 blocking probability is

$$B_2(k_1, k_2) = \sum_{(i,j,l) \in \mathcal{S}_2} \pi(i, j, l) = \frac{\sum_{(i,j,l) \in \mathcal{S}_2} \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!}}{\sum_{(i,j,l) \in \Omega} \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!}}. \quad (16)$$

Class-3 (Elective) Blocking. Class-3 is blocked whenever $l = n - k_1 - k_2$ (the Class-3 count has reached its cap) or $i + j + l = n$ (system full). Defining

$$\mathcal{S}_3 = \{(i, j, l) \in \Omega : l = n - k_1 - k_2 \text{ or } i + j + l = n\}, \quad (17)$$

the Class-3 blocking probability is

$$B_3(k_1, k_2) = \sum_{(i,j,l) \in \mathcal{S}_3} \pi(i, j, l) = \frac{\sum_{(i,j,l) \in \mathcal{S}_3} \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!}}{\sum_{(i,j,l) \in \Omega} \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!}}. \quad (18)$$

Remark 5 (Priority protection of Class-1). Since $\mathcal{S}_1 \subseteq \mathcal{S}_2$ (every full state is in $\mathcal{S}_2$ by the “or $i + j + l = n$ ” clause) and $\mathcal{S}_1 \subseteq \mathcal{S}_3$ (similarly), it follows that $B_1 \leq B_2$ and $B_1 \leq B_3$ for all parameter values. In other words, Class-1 is never worse than either lower-priority class: the reservation policy guarantees that emergency patients always experience no more blocking than semi-urgent or elective patients.

The relative ordering of B_2 and B_3 is not guaranteed by set inclusion: $S_2 \not\subseteq S_3$ in general (a state with $j = n - k_1$ but $l < n - k_1 - k_2$ lies in S_2 but not S_3). Whether $B_2 \leq B_3$ or $B_2 \geq B_3$ depends on the traffic intensities ρ_2 and ρ_3 relative to the thresholds. This is the expected and desirable behavior: the reservation protects Class-1 absolutely, while the relative blocking of Classes 2 and 3 reflects their actual offered loads.

Weighted total rejection rate. The system-wide throughput-loss rate, weighted by the offered loads, is

$$B_{\text{tot}}(k_1, k_2) = \frac{\lambda_1 B_1(k_1, k_2) + \lambda_2 B_2(k_1, k_2) + \lambda_3 B_3(k_1, k_2)}{\lambda_1 + \lambda_2 + \lambda_3}. \quad (19)$$

2.2.4. Explicit Summation Formulas

The normalizing constant (4) and the partial sums in (14) through (18) can be evaluated by iterating over the state space Ω . Introduce

$$G = G(n, k_1, k_2; \rho_1, \rho_2, \rho_3) = \sum_{i=0}^n \frac{\rho_1^i}{i!} \sum_{j=0}^{n-k_1} \frac{\rho_2^j}{j!} \sum_{\substack{l=0 \\ i+j+l \leq n}}^{n-k_1-k_2} \frac{\rho_3^l}{l!}, \quad (20)$$

so that $C^{-1} = G$. The outer sum over i runs from 0 to n ; for each (i, j) , the inner sum over l is further restricted by $l \leq n - i - j$ (capacity) and $l \leq n - k_1 - k_2$ (coordinate threshold), giving an upper limit of $l \leq \min(n - i - j, n - k_1 - k_2)$.

More explicitly:

$$G = \sum_{i=0}^n \frac{\rho_1^i}{i!} \left[\sum_{j=0}^{n-k_1} \frac{\rho_2^j}{j!} \left(\sum_{l=0}^{\min(n-i-j, n-k_1-k_2)} \frac{\rho_3^l}{l!} \right) \right], \quad (21)$$

where the innermost sum is zero when $n - i - j < 0$ (i.e., when $i + j > n$).

Similarly define the partial sums restricted to each blocking set:

$$G_1 = \sum_{\substack{(i,j,l) \in \Omega \\ i+j+l=n}} \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!}, \quad (22)$$

$$G_2 = \sum_{\substack{(i,j,l) \in \Omega \\ j=n-k_1 \text{ or } i+j+l=n}} \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!}, \quad (23)$$

$$G_3 = \sum_{\substack{(i,j,l) \in \Omega \\ l=n-k_1-k_2 \text{ or } i+j+l=n}} \frac{\rho_1^i}{i!} \frac{\rho_2^j}{j!} \frac{\rho_3^l}{l!}. \quad (24)$$

Then $B_r(k_1, k_2) = G_r / G$ for $r = 1, 2, 3$. Because G , G_1 , G_2 , and G_3 are finite triple sums over the state space Ω , their computational cost follows directly from the size of this index set. In particular, $|\Omega|$ is bounded above by $(n+1)(n-k_1+1)(n-k_1-k_2+1) = O(n^3)$, and each summand requires $O(1)$ arithmetic operations. Thus, these normalizing and partial sums can be evaluated in $O(n^3)$ arithmetic operations by a direct enumeration of Ω .

Remark 6 (Monotone behavior of blocking in k_1 and k_2). Numerical experiments consistently show the following monotone patterns:

- (i) Increasing k_1 (fixed k_2). Tightening the Class-2 coordinate cap from $n - k_1$ to $n - k_1 - 1$ reduces B_1 and increases B_2 and B_3 . Intuitively, more beds are protected for Class-1, reducing its blocking at the cost of higher blocking for the lower-priority classes.
- (ii) Increasing k_2 (fixed k_1). Tightening the Class-3 coordinate cap increases B_3 and decreases B_2 (Class-2 benefits from the additional reservation). The effect on B_1 is not identically zero because changing k_2 changes the admissible state space and hence the normalization in the product form distribution. Let $c = n - k_1 - k_2$ be the current Class-3 cap, let $A_c = \{(i, j, l) \in \Omega : l \leq c\}$, and let $R_c = \{(i, j, l) \in A_c : l = c\}$ be the boundary slice removed when k_2 is increased by one. If W_c and F_c denote, respectively, the total product form weight and full-state weight in R_c , and if $B_1(c) = F(A_c)/W(A_c)$, then

$$B_1(c-1) - B_1(c) = \frac{W_c}{W(A_c) - W_c} \left\{ B_1(c) - \frac{F_c}{W_c} \right\}.$$

Thus, k_2 affects B_1 only indirectly through the probability mass of the removed Class-3 boundary slice; the effect is small when that slice has small stationary mass. In the parameter grids reported in Section 4.3, the largest adjacent change in B_1 due solely to increasing k_2 is below 0.020.

These patterns have been verified by exhaustive enumeration over $(k_1, k_2) \in \mathcal{K}$ for representative parameter sets.

2.3. Rejection Rates Under IPP Arrivals

In many ICU settings, the arrival stream of each class exhibits burstiness, alternating between high-activity and low-activity periods, which a Poisson process cannot capture. We model each class by an *Interrupted Poisson Process* (IPP) and study the superposition of three independent IPPs, which forms a Markov-modulated Poisson process (MMPP).

2.3.1. Individual IPP Models

For class $r \in \{1, 2, 3\}$, the IPP is a two-state continuous-time Markov chain on states $\{0_r, 1_r\}$ (OFF and ON):

- In state 0_r (OFF): No arrivals; transitions to state 1_r at rate ω_r .
- In state 1_r (ON): Arrivals occur at rate λ_r ; transitions to state 0_r at rate γ_r .

The generator and arrival rate matrix for the Class- r IPP are, respectively

$$Q_r = \begin{pmatrix} -\omega_r & \omega_r \\ \gamma_r & -\gamma_r \end{pmatrix}, \quad \Lambda_r = \begin{pmatrix} 0 & 0 \\ 0 & \lambda_r \end{pmatrix}. \quad (25)$$

The stationary probability of the ON state is $p_r^{\text{on}} = \omega_r/(\omega_r + \gamma_r)$, giving the effective (long-run average) arrival rate $\bar{\lambda}_r = \lambda_r \omega_r/(\omega_r + \gamma_r)$.

The *index of dispersion* (peakedness) of the Class- r IPP is defined as the variance-to-mean ratio of the counting process in the steady-state:

$$z_r = \lim_{T \rightarrow \infty} \frac{\text{Var}[N_r(T)]}{E[N_r(T)]} = 1 + \frac{2\lambda_r\gamma_r}{(\omega_r + \gamma_r)^2}. \quad (26)$$

This follows from $\text{Var}[N_r(T)] = \bar{\lambda}_r T + \lambda_r^2 \cdot \text{Var}[T_r^{\text{on}}(T)]$ with $\text{Var}[T_r^{\text{on}}(T)] \approx 2T\omega_r\gamma_r/(\omega_r + \gamma_r)^3$ for large T [14]. For a Poisson process, $z_r = 1$; for a bursty IPP, $z_r > 1$. Throughout this paper, z_r is referred to as the index of dispersion.

2.3.2. Superposition: Eight-Phase MMPP

The three IPPs are independent, so their superposition is an MMPP whose phase space is the Cartesian product $\{0, 1\}^3$, yielding $2^3 = 8$ phases. We index phases by triples $(\iota, \phi, \psi) \in \{0, 1\}^3$, where ι, ϕ, ψ encode the ON/OFF state of Classes 1, 2, and 3, respectively.

The 8×8 generator of the phase process and the 8×8 arrival rate matrix for each class are formed by Kronecker sums and products:

$$Q = Q_1 \oplus Q_2 \oplus Q_3 = Q_1 \otimes I_2 \otimes I_2 + I_2 \otimes Q_2 \otimes I_2 + I_2 \otimes I_2 \otimes Q_3, \quad (27)$$

$$\tilde{\Lambda}_1 = \Lambda_1 \otimes I_2 \otimes I_2, \quad (28)$$

$$\tilde{\Lambda}_2 = I_2 \otimes \Lambda_2 \otimes I_2, \quad (29)$$

$$\tilde{\Lambda}_3 = I_2 \otimes I_2 \otimes \Lambda_3. \quad (30)$$

Here all I_2 are 2×2 identity matrices, and the ordering of Kronecker factors corresponds to Classes 1, 2, and 3. The total arrival rate matrix (arrivals of any class) is $\tilde{\Lambda} = \tilde{\Lambda}_1 + \tilde{\Lambda}_2 + \tilde{\Lambda}_3$.

2.3.3. System State and Generator Structure

The full system state is (i, j, l, ξ) , where $(i, j, l) \in \Omega$ (as defined in (1)), and $\xi = (\iota, \phi, \psi) \in \{0, 1\}^3$ is the 8-dimensional MMPP phase. We group states by total occupancy level $s = i + j + l$, defining

$$\mathcal{L}_s = \{(i, j, l) \in \Omega : i + j + l = s\}, \quad s = 0, 1, \dots, n. \quad (31)$$

A *level block* is the submatrix of the full generator whose rows and columns are restricted to states in specified occupancy levels. More precisely, after ordering states first by total occupancy, the (s, t) block of $\hat{Q}$ has rows indexed by $\mathcal{L}_s \times \{0, 1\}^3$ and columns indexed by $\mathcal{L}_t \times \{0, 1\}^3$. At level s , the sub-state space is $\mathcal{L}_s \times \{0, 1\}^3$. Let $d_s = |\mathcal{L}_s|$ denote the number of admissible patient-count triples at level s ; because each patient-count triple is paired with one of the eight MMPP phases, the level- s block row/column has dimension $8d_s$.

The admission of each class at level s is governed by the coordinate thresholds. For a state (i, j, l) at level s , Class-2 arrivals are admitted if $j < n - k_1$ and Class-3 arrivals are admitted if $l < n - k_1 - k_2$, in addition to the capacity requirement $s < n$. Consequently, the effective admitted arrival rate matrix is *state-specific*, not merely level-specific, in general.

For the purposes of the block-structured generator, it is convenient to define level-aggregated admitted arrival rate matrices:

$$\tilde{\Lambda}_r(s) = \begin{cases} \tilde{\Lambda}_r, & \text{if class } r \text{ can be admitted at level } s \text{ for some } (i, j, l) \in \mathcal{L}_s, \\ \mathbf{0}, & \text{if class } r \text{ is blocked at all states of level } s. \end{cases} \quad (32)$$

Specifically:

- Class-1 arrivals are admitted at level s iff $s < n$.
- Class-2 arrivals are admitted at level s iff $s < n$ and there exists a state $(i, j, l) \in \mathcal{L}_s$ with $j < n - k_1$, i.e., iff $s < n$ and $s < n - k_1 + (n - s) = n$. More precisely, Class-2 is completely blocked at level s only if *every* state in $\mathcal{L}_s$ has $j = n - k_1$, which requires $s \geq n - k_1$ and all Class-2 count at cap; for a mixed level, the block entry is state-specific.
- Class-3 arrivals are similarly handled.

Definition 1 (Block generator $\hat{Q}$). Order the states lexicographically: first by level s , then by patient-count triple $(i, j, l) \in \mathcal{L}_s$ (lexicographic order), then by phase ξ . The generator $\hat{Q}$ of the Markov chain on $\bigcup_{s=0}^n \mathcal{L}_s \times \{0, 1\}^3$ has a block-tridiagonal structure

$$\hat{Q} = \begin{pmatrix} \hat{A}_{00} & \hat{B}_{01} & & & \\ \hat{C}_{10} & \hat{A}_{11} & \hat{B}_{12} & & \\ & \hat{C}_{21} & \hat{A}_{22} & \ddots & \\ & & \ddots & \ddots & \hat{B}_{n-1,n} \\ & & & \hat{C}_{n,n-1} & \hat{A}_{nn} \end{pmatrix}, \quad (33)$$

where $\hat{A}_{ss}$, $\hat{B}_{s,s+1}$, and $\hat{C}_{s+1,s}$ are matrices of dimension $(8d_s) \times (8d_s)$, $(8d_s) \times (8d_{s+1})$, and $(8d_{s+1}) \times (8d_s)$, respectively, with the following interpretations:

- $\hat{A}_{ss}$: Transitions within level s , comprising the phase process transitions $I_{d_s} \otimes Q$ plus diagonal correction terms to ensure rows of $\hat{Q}$ sum to zero.
- $\hat{B}_{s,s+1}$: Transitions from level s to level $s+1$ due to an admission. For each pair $(i, j, l) \in \mathcal{L}_s$ and $(i', j', l') \in \mathcal{L}_{s+1}$, the (8×8) sub-block is $\tilde{\Lambda}_r$ if $(i', j', l') = (i, j, l) + e_r$ and class r is admissible from (i, j, l) (i.e., $j < n - k_1$ for $r = 2$, $l < n - k_1 - k_2$ for $r = 3$), and $\mathbf{0}$ otherwise.
- $\hat{C}_{s+1,s}$: Transitions from level $s+1$ to level s due to a discharge. For each $(i', j', l') \in \mathcal{L}_{s+1}$ and predecessor $(i, j, l) \in \mathcal{L}_s$ differing in coordinate r , the sub-block is $r' \mu_r \cdot I_8$, where r' is the count of class r in state (i', j', l') (departures are phase-independent).

Remark 7. Unlike the two-class case, the sub-blocks within the transition block $\hat{B}_{s,s+1}$ are state-specific: whether a Class-2 or Class-3 arrival is admitted depends on the individual counts j and l , not just the total occupancy s . In numerical implementations, it is therefore necessary to enumerate $\mathcal{L}_s$ explicitly and track which transitions are admissible for each state.

2.3.4. Steady-State Computation

Partition the steady-state row vector $\hat{\pi} = (\hat{\pi}_0, \hat{\pi}_1, \dots, \hat{\pi}_n)$, where $\hat{\pi}_s$ is a row vector of length $8d_s$ containing the steady-state probabilities for all states at level s . The steady-state equations are

$$\hat{\pi} \hat{Q} = \mathbf{0}, \quad \hat{\pi} \mathbf{e} = 1. \quad (34)$$

Owing to the block-tridiagonal structure (33), the system (34) can be written level-by-level as

$$\hat{\pi}_{s-1} \hat{B}_{s-1,s} + \hat{\pi}_s \hat{A}_{ss} + \hat{\pi}_{s+1} \hat{C}_{s+1,s} = \mathbf{0}, \quad s = 0, 1, \dots, n, \quad (35)$$

(with $\hat{\pi}_{-1} = \hat{\pi}_{n+1} = \mathbf{0}$). This system is solved numerically using sparse linear algebra (e.g., the method of Gaussian elimination exploiting the block-tridiagonal sparsity pattern). The total number of states is $|\Omega| \cdot 8 = O(n^3)$, so the global linear system (34) has dimension $O(n^3)$. The maximum block size at any level is $8d_{\max}$, where $d_{\max} = O(n^2)$; standard sparse solvers exploit the block structure to reduce the effective cost below a naive $O(n^9)$ dense solve.

Remark 8. The dimension $d_s = |\mathcal{L}_s|$ counts all $(i, j, l) \in \Omega$ with $i + j + l = s$. For levels $s \leq n - k_1 - k_2$, all triples with $i + j + l = s$ satisfy the coordinate constraints, which gives the clean closed-form count $d_s = \binom{s+2}{2}$. For larger s , the constraints $j \leq n - k_1$ and $l \leq n - k_1 - k_2$ may reduce d_s below $\binom{s+2}{2}$. In numerical implementations, it is convenient to enumerate $\mathcal{L}_s$ explicitly rather than relying on a closed-form count.

2.3.5. Blocking Probability Formulas

The steady-state vector $\hat{\pi}$ is a row vector of length $|\Omega| \cdot 8$, indexed by pairs $(\mathbf{x}, \xi)$ with $\mathbf{x} = (i, j, l) \in \Omega$ and $\xi \in \{0, 1\}^3$. To express the blocking probabilities correctly, we lift the 8×8 phase space matrices $\tilde{\Lambda}_r$ to the full state space by forming the Kronecker product with the identity over the patient-count space:

$$\mathbf{L}_r := I_{|\Omega|} \otimes \tilde{\Lambda}_r, \quad r = 1, 2, 3. \quad (36)$$

$\mathbf{L}_r$ is a $(8|\Omega|) \times (8|\Omega|)$ matrix; for each patient-count state (i, j, l) , the (8×8) diagonal block of $\mathbf{L}_r$ is $\tilde{\Lambda}_r$, encoding the Class- r arrival rates across all phases at that patient state.

For a subset $\mathcal{S} \subseteq \Omega$, let $\hat{\pi}_{\mathcal{S}}$ denote the sub-vector of $\hat{\pi}$ indexed by $\mathcal{S} \times \{0, 1\}^3$ (all phases at patient states in $\mathcal{S}$), zero-padded to length $8|\Omega|$ for consistent multiplication.

Proposition 2 (IPP blocking probabilities). *Under IPP arrivals with the two-tier coordinate reservation policy, the blocking probability for Class r ($r = 1, 2, 3$) is*

$$P_{\text{block},r} = \frac{\hat{\pi}_{\mathcal{S}_r} \mathbf{L}_r \mathbf{e}}{\hat{\pi} \mathbf{L}_r \mathbf{e}} = \frac{\sum_{(i,j,l) \in \mathcal{S}_r} \sum_{\xi \in \{0,1\}^3} \hat{\pi}(i,j,l,\xi) [\tilde{\Lambda}_r \mathbf{e}_8]_{\xi}}{\sum_{(i,j,l) \in \Omega} \sum_{\xi \in \{0,1\}^3} \hat{\pi}(i,j,l,\xi) [\tilde{\Lambda}_r \mathbf{e}_8]_{\xi}}, \quad (37)$$

where $\mathcal{S}_r$ ($r = 1, 2, 3$) are defined in (13) through (17), $\mathbf{L}_r = I_{|\Omega|} \otimes \tilde{\Lambda}_r$ is the lifted arrival rate matrix (36), $\mathbf{e}$ and $\mathbf{e}_8$ are all-ones column vectors of appropriate length, and $[\tilde{\Lambda}_r \mathbf{e}_8]_{\xi}$ is the Class- r arrival rate in phase ξ .

Proof. The long-run rate of Class- r arrivals, by the ergodic theorem, equals

$$(\text{total flux}) = \hat{\pi} \mathbf{L}_r \mathbf{e} = \sum_{(i,j,l) \in \Omega} \sum_{\xi} \hat{\pi}(i,j,l,\xi) [\tilde{\Lambda}_r \mathbf{e}_8]_{\xi}.$$

This is dimensionally correct: $\hat{\pi}$ is a $1 \times (8|\Omega|)$ row vector, $\mathbf{L}_r$ is $(8|\Omega|) \times (8|\Omega|)$, and $\mathbf{e}$ is $(8|\Omega|) \times 1$. Note that $[\tilde{\Lambda}_r \mathbf{e}_8]_{\xi}$ equals λ_r if phase ξ has class r in the ON state, and 0 otherwise, so this sum correctly weights each state by its phase-dependent Class- r arrival rate.

A Class- r arrival is blocked whenever the current patient state $(i, j, l) \in \mathcal{S}_r$ (coordinate threshold reached or system full), regardless of phase ξ . The long-run rate of blocked Class- r arrivals is therefore $\hat{\pi}_{\mathcal{S}_r} \mathbf{L}_r \mathbf{e}$, with $\hat{\pi}_{\mathcal{S}_r}$ the restriction of $\hat{\pi}$ to $\mathcal{S}_r \times \{0, 1\}^3$. The ratio of blocked to total arrival rate gives (37). $\square$

Remark 9 (Component-wise interpretation). *In the component form of (37), the weight $[\tilde{\Lambda}_r \mathbf{e}_8]_{\xi}$ equals λ_r if class r 's IPP is ON in phase ξ , and 0 otherwise. Thus, only those (i, j, l, ξ) states in which the class- r IPP is active contribute to the numerator and denominator. States where class r is in the OFF phase are irrelevant for blocking since no arrivals occur. This can be made explicit as:*

$$P_{\text{block},r} = \frac{\sum_{(i,j,l) \in \mathcal{S}_r} \hat{\pi}_r^{\text{on}}(i,j,l)}{\sum_{(i,j,l) \in \Omega} \hat{\pi}_r^{\text{on}}(i,j,l)},$$

where $\hat{\pi}_r^{\text{on}}(i,j,l) = \sum_{\xi: \xi_r=1} \hat{\pi}(i,j,l,\xi)$ is the marginal probability of state (i,j,l) with class r ON.

Remark 10 (Reduction to Poisson case). *If each IPP is “always-on” (i.e., $\omega_r \rightarrow \infty$, $\gamma_r \rightarrow \infty$ with $\omega_r/(\omega_r + \gamma_r) \rightarrow 1$ and λ_r fixed), the MMPP degenerates to a Poisson process with rate λ_r . In this limit, phase ξ degenerates to the unique always-ON phase, $\hat{\pi}(i,j,l,\xi) \rightarrow \pi(i,j,l)$ (the Poisson steady-state), and $P_{\text{block},r}$ in (37) reduces to $B_r(k_1, k_2)$ as given by (14) through (18).*

3. Bed Reservation Policy

3.1. Motivation

The reservation parameters (k_1, k_2) control a fundamental trade-off: increasing k_1 reduces the blocking probability for emergency patients at the cost of higher blocking for semi-urgent and elective patients, while increasing k_2 additionally restricts the Class-3 threshold (tightening the cap on elective patients) while providing headroom for semi-urgent patients. A purely utilitarian objective (minimize total blocking) would set $k_1 = k_2 = 0$, which provides no protection for emergency patients. Conversely, setting $k_1 = n - k_2$ reserves nearly all beds for Class-1, rendering the system nearly inaccessible to other classes. The fairness optimization below seeks a principled balance.

3.2. Optimization Problem

Let $\mathbf{k} = (k_1, k_2) \in \mathbb{Z}_{\geq 0}^2$ with $k_1 + k_2 \leq n$. The admissible parameter set is

$$\mathcal{K} = \{(k_1, k_2) \in \mathbb{Z}^2 : k_1 \geq 1, k_2 \geq 1, k_1 + k_2 \leq n\}. \quad (38)$$

3.2.1. Hierarchical Fairness Constraints

Motivated by the priority structure of the three classes, we impose pairwise fairness bounds:

- $\varepsilon_{12} > 0$: Maximum tolerable excess of B_2 over B_1 (Class-2 vs. Class-1 gap).
- $\varepsilon_{13} > 0$: Maximum tolerable excess of B_3 over B_1 (Class-3 vs. Class-1 gap).

Since the relative ordering of B_2 and B_3 depends on traffic intensities (see Remark 5), we impose separate constraints relative to Class-1 rather than a chain of pairwise inequalities. The two-tier fairness optimization problem is

$$\begin{array}{ll} \min_{(k_1, k_2) \in \mathcal{K}} & B_1(k_1, k_2) \\ \text{subject to} & \\ & B_2(k_1, k_2) - B_1(k_1, k_2) \leq \varepsilon_{12}, \\ & B_3(k_1, k_2) - B_1(k_1, k_2) \leq \varepsilon_{13}. \end{array} \quad (39)$$

Here $B_r(k_1, k_2)$ refers to either the Poisson case formulas (14) through (18) or the IPP case formulas (37), according to the arrival model in use.

3.2.2. Alternative: Uniform Fairness Constraint

When a single fairness tolerance $\varepsilon > 0$ applies across all classes relative to Class-1, the constraints simplify to

$$\max\{B_2(k_1, k_2), B_3(k_1, k_2)\} - B_1(k_1, k_2) \leq \varepsilon. \quad (40)$$

Remark 11 (Attainability and computability of Problem (39)). *Problem (39) has a feasible policy whenever $\varepsilon_{12} \geq B_2(1, 1) - B_1(1, 1)$ and $\varepsilon_{13} \geq B_3(1, 1) - B_1(1, 1)$. Separately, the two threshold gaps $B_2(1, 1) - B_1(1, 1)$ and $B_3(1, 1) - B_1(1, 1)$ are finite, non-negative, and computable in $O(n^3)$ operations for any given traffic parameters. In particular, both gaps vanish as $\rho \rightarrow 0$, showing that the fairness constraints are attainable in the light-traffic regime for any $\varepsilon_{12}, \varepsilon_{13} > 0$.*

3.2.3. Relationship to Numerical Experiments

Problem (39) is the canonical policy-design formulation: it directly minimizes emergency blocking subject to explicit equity constraints. In Section 4.3, we adopt a *weighted-loss surrogate* $\ell(k_1, k_2) = 0.7 B_1 + 0.15 (B_2 + B_3)$ as a practical alternative that avoids pre-

specifying the tolerance parameters $(\varepsilon_{12}, \varepsilon_{13})$ for every (ρ, z) grid point. The weights are interpreted as a priori preference parameters in the standard weighted-sum scalarization of a multi-objective problem [19]. The baseline vector $(0.70, 0.15, 0.15)$ assigns emergency blocking a weight about 4.67 times that of either lower-priority class, while retaining positive penalties on Class-2 and Class-3 blocking so that fairness is not ignored. The two formulations are not algebraically equivalent, but we verify in Section 4.3 that every solution selected by the surrogate also satisfies the fairness constraints of Problem (39) with tolerance $\varepsilon = 0.01$ across all tested configurations.

3.3. Jain Fairness Index

As a complementary measure of equity across the three classes, the *Jain fairness index* [20] is

$$J(k_1, k_2) = \frac{(B_1(k_1, k_2) + B_2(k_1, k_2) + B_3(k_1, k_2))^2}{3(B_1(k_1, k_2)^2 + B_2(k_1, k_2)^2 + B_3(k_1, k_2)^2)} \in \left[\frac{1}{3}, 1\right]. \quad (41)$$

$J = 1$ if and only if $B_1 = B_2 = B_3$ (perfect equity); $J = 1/3$ corresponds to the most inequitable case where all blocking is concentrated in one class. A high Jain index is desirable in conjunction with low aggregate blocking.

3.4. Solution Approach

Since (k_1, k_2) ranges over the finite integer set $\mathcal{K}$ of cardinality $|\mathcal{K}| = n(n-1)/2$, the optimization problem (39) can be solved exactly by *exhaustive enumeration*: evaluate (B_1, B_2, B_3) for every $(k_1, k_2) \in \mathcal{K}$, discard infeasible pairs, and return the pair with smallest B_1 . This count treats (k_1, k_2) as an ordered pair; the factor $1/2$ arises from the triangular capacity constraint $k_1 + k_2 \leq n$, not from identifying (k_1, k_2) with (k_2, k_1) . Each evaluation requires $O(n^3)$ arithmetic operations (Poisson case) or the numerical solution of a sparse linear system of dimension $O(n^3)$ (IPP case, since $|\Omega| \cdot 8 = O(n^3)$), giving an overall complexity of $O(n^5)$ for exhaustive search.

For large n , more efficient approaches include:

- (i) **Monotone search.** As observed in Remark 6, B_1 is non-increasing in k_1 (for fixed k_2) across all tested parameters. If this monotonicity holds in general, a bisection search over k_1 (for fixed k_2) can locate the constraint-satisfaction boundary in $O(\log n)$ evaluations per k_2 value, reducing total complexity to $O(n \log n)$ evaluations.
- (ii) **Gradient-free integer programming.** When n is very large, derivative-free methods (e.g., Nelder–Mead adapted to integer grids, or pattern search) can be applied after relaxing (k_1, k_2) to continuous variables and rounding; the convexity properties of the Erlang loss function in the Poisson case support this approach.
- (iii) **Lagrangian relaxation.** Dualizing the fairness constraints in (39) with multipliers $(\nu_{12}, \nu_{13}) \geq 0$ gives the unconstrained Lagrangian

$$L(k_1, k_2, \nu_{12}, \nu_{13}) = B_1 + \nu_{12}(B_2 - B_1 - \varepsilon_{12}) + \nu_{13}(B_3 - B_1 - \varepsilon_{13}),$$

which can be minimized over $\mathcal{K}$ for fixed multipliers, and the multipliers updated by a subgradient method.

Remark 12 (Extension to m classes). The two-tier reservation policy and the associated analytical framework generalize directly to $m \geq 3$ patient classes with $m-1$ reservation thresholds $k_1, k_2, \dots, k_{m-1}$ (with $\sum_{r=1}^{m-1} k_r \leq n$). The state space becomes an m -dimensional lattice with coordinate constraints $x_r \leq n - \sum_{s < r} k_s$ for $r = 2, \dots, m$, and the product form (3) extends with one additional Poisson factor per class. The IPP superposition yields an MMPP with 2^m phases. The fairness optimization gains one decision variable and one pairwise constraint per additional class.

4. Experiment and Results

This section presents the experimental program for the three-class model. Section 4.2 fits IPP arrival parameters to the MIMIC-IV Medical ICU dataset. Section 4.3 performs a comprehensive parameter sensitivity sweep. Section 4.4 validates the analytical model against the MIMIC-IV data via discrete-event simulation. Section 4.5 extends the LoS insensitivity study to six service-time distributions.

4.1. Experimental Setup and Reproducibility

Table 3 summarizes the parameters used across all experiments.

Table 3. Experimental setup and reproducibility parameters.

Parameter	Value (s)	Notes
<i>System</i>		
Bed capacity n	15 20	Primary sensitivity analysis MIMIC-IV validation and additional sensitivity
Arrival class split	$(0.50, 0.30, 0.20)\rho n$	
Service rates μ_r	$\mu_r = 1$	Classes 1, 2, 3 Normalized
<i>Sensitivity sweep</i>		
Utilization grid ρ	$[0.40, 0.95]$, step 0.05	12 values
Burstiness index grid z	$[1.0, 5.0]$, step 0.5	9 values
Optimization objective	$0.7B_1 + 0.15(B_2 + B_3)$	Weighted surrogate
Blocking formula	Exact product form & IPP blocking (37)	$O(n^3)$ enumeration
Bed capacity sensitivity	$n = 15$ and $n = 20$	Same (ρ, z) grid and objective
<i>Simulation (validation and LoS insensitivity)</i>		
Arrivals per run	3×10^5	Heap-based DES
Independent replications	5 (validation), 4 (LoS study)	
Confidence interval	95%: $\bar{B} \pm 1.96 \hat{\sigma} / \sqrt{R}$	
Validation pass criterion	$ B_r^{\text{anal}} - B_r^{\text{sim}} \leq 1.5 \text{ pp}$	
LoS insensitivity criterion	Relative difference $\leq 1\%$	
<i>IPP burstiness penalty</i>		
Grid	same (ρ, z) grid	108 cells
IPP solver	Full CTMC steady-state	$ \Omega \cdot 8 \approx 4480$ states
ON-fraction p	0.60	Fixed
Software	Python 3.11; NumPy 2.4.5; SciPy 1.17.1; custom DES	

4.2. IPP Fitting on MIMIC-IV Data

We fit IPP arrival parameters to the MIMIC-IV v2.2 Medical ICU (MICU) [17], a publicly available critical-care database containing approximately 70,000 ICU stays with full admission and discharge timestamps [18]. The MICU unit has a nominal capacity of $n = 20$ beds and records roughly 1390 admissions per year.

Following the three-class triage structure of the model, patients are stratified by admission type: **Class-1 (Emergency)**: EW EMER. or DIRECT EMER. admissions from the Emergency Room (58%). **Class-2 (Semi-urgent)**: URGENT admissions via hospital transfer or PACU (27%). **Class-3 (Elective)**: ELECTIVE or same-day surgical admissions (15%).

In Table 4, the *primary empirical quantities* are the mean inter-arrival time (Mean IA = $1/\bar{\lambda}_r$) and the index of dispersion z_r , both estimated directly from the MIMIC-IV admission timestamps. The IPP parameters $(\lambda_r, \omega_r, \gamma_r)$ are *derived* via two-moment matching [21]: given $\bar{\lambda}_r, z_r$, and ON-fraction p_r (reported in Table 4), setting $S_r = 2\bar{\lambda}_r(1 - p_r)/(p_r(z_r - 1))$ gives $\omega_r = p_r S_r$, $\gamma_r = (1 - p_r)S_r$, and $\lambda_r = \bar{\lambda}_r/p_r$. Table 5 verifies their internal consistency.

Table 4. Inter-arrival statistics and fitted IPP parameters for the MIMIC-IV MICU dataset (time unit: days).

Class	Mean IA	z_r	λ_r	ω_r	γ_r	$\bar{\lambda}_r$	μ_r	ρ_r	p_r
Emergency (E1)	0.453	1.42	4.016	4.733	3.873	2.209	0.208	10.60	0.55
Semi-urgent (E2)	0.973	1.18	1.582	3.999	2.153	1.028	0.263	3.91	0.65
Elective (E3)	1.751	1.05	0.762	5.712	1.904	0.571	0.312	1.83	0.75

Table 5. Relative difference (%) between inter-arrival statistics of real MIMIC-IV traces and IPP-fitted processes.

Class	Mean	Variance	Peakedness	Skewness
Emergency (E1)	−0.004	−0.041	−0.037	−4.8
Semi-urgent (E2)	−0.011	−0.089	−0.078	−2.3
Elective (E3)	0.019	−0.055	−0.036	1.1

The IPP captures all three inter-arrival distributions with mean and variance deviations below 0.1%, confirming its suitability as the arrival model. The peakedness indices ($z_1 = 1.42$, $z_2 = 1.18$, $z_3 = 1.05$) are consistent with the bursty emergency arrivals and near-Poisson elective patterns documented in [15].

4.3. Comprehensive Parameter Sensitivity Analysis

We sweep traffic utilization $\rho \in [0.40, 0.95]$ (step 0.05) and burstiness index $z \in [1.0, 5.0]$ (step 0.5) for a system with $n = 15$ beds. This $n = 15$ setting is a stylized medium-capacity case used for the primary broad sensitivity sweep. Because the empirical MIMIC-IV MICU validation below uses the dataset-specific nominal capacity $n = 20$, we also repeat the same sensitivity sweep at $n = 20$ as an additional bed capacity sensitivity check; the corresponding representative results are reported after Table 6. Arrival rates are split as $(\lambda_1, \lambda_2, \lambda_3) = (0.50, 0.30, 0.20)\rho n$ with normalized service rates $\mu_r = 1$. For each (ρ, z) pair, we enumerate all $(k_1, k_2) \in \mathcal{K}$ and evaluate (B_1, B_2, B_3) . Note that the IPP blocking probability Formula (37) is used for all (ρ, z) pairs with $z > 1$, while the exact product form Poisson Formulas (14)–(18) are used for $z = 1$.

Table 6. Optimal two-tier reservation and resulting blocking probabilities for representative (ρ, z) combinations ($n = 15$, $w = 0.7$).

ρ	z	k_1^*	k_2^*	B_1	B_2	B_3	J
0.50	1.0	1	1	0.0057	0.0057	0.0057	1.0000
0.50	5.0	2	1	0.0157	0.0157	0.0157	1.0000
0.65	1.0	4	1	0.0318	0.0318	0.0318	1.0000
0.65	5.0	6	1	0.0655	0.0665	0.0657	1.0000
0.80	1.0	7	1	0.0844	0.0905	0.0869	0.9992
0.80	5.0	7	1	0.1422	0.1486	0.1436	0.9996
0.90	1.0	7	1	0.1298	0.1372	0.1329	0.9995
0.90	5.0	7	1	0.1982	0.2053	0.1998	0.9998

4.3.1. Relationship Between the Two Optimization Formulations

Section 3 defines the canonical Problem (39): minimize B_1 subject to fairness constraints $B_r - B_1 \leq \varepsilon_{1r}$. For the sensitivity sweep, we use the weighted-loss surrogate

$$\ell(k_1, k_2) = 0.7 B_1(k_1, k_2) + 0.15 (B_2(k_1, k_2) + B_3(k_1, k_2)), \quad (42)$$

which encodes the same clinical priority without requiring pre-specified tolerances. Following the weighted-sum interpretation of preference weights [19], this baseline corresponds to $w = 0.70$ in $wB_1 + (1 - w)(B_2 + B_3)/2$. To assess whether the selected policies depend fragily on this value, we repeated the full (ρ, z) sweep for $w \in \{0.60, 0.65, 0.68, 0.72, 0.73\}$. For the primary $n = 15$ grid, all selected policies still satisfied the uniform fairness tolerance $\varepsilon = 0.01$, with the worst observed gap at most 0.0079 and minimum Jain index at least 0.9989. For the additional $n = 20$ grid, the corresponding worst gap was at most 0.0076 and the minimum Jain index was at least 0.9985. Larger emergency weights can deliberately sacrifice fairness: at $w = 0.74$, violations occurred in 5 of 108 cases for $n = 15$ and 7 of 108 cases for $n = 20$. Thus, the baseline $w = 0.70$ is not claimed to be universally equivalent to the constrained problem; rather, it lies inside the empirically robust range for the tested traffic regimes while preserving emergency priority. These two formulations are not equivalent in general. To confirm consistency, we verify *a posteriori* that every solution (k_1^*, k_2^*) selected by (42) also satisfies the fairness constraints of Problem (39) with uniform tolerance $\varepsilon = 0.01$ (worst observed gap $\max\{B_2 - B_1, B_3 - B_1\} = 0.0079$) and achieves Jain index $J \geq 0.999$ across all 108 tested (ρ, z) combinations. Hence, the $n = 15$ results in Table 6 and Figure 1 are consistent with Problem (39) throughout. The pair minimizing $\ell(k_1, k_2)$ is selected for each (ρ, z) point.

Two-Tier Reservation Policy: Optimal k_1/n and Resulting B_1^* ($n = 15, w = 0.7$)

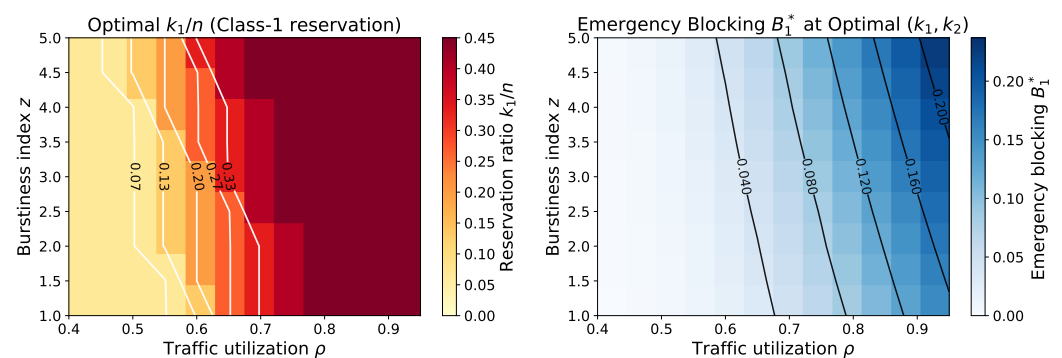


Figure 1. (Left): Optimal Class-1 reservation ratio k_1/n as a function of traffic utilization ρ and burstiness index z ($n = 15, w = 0.7$); white contours mark equal-reservation boundaries. (Right): Emergency blocking probability B_1^* achieved at the optimal (k_1, k_2) , illustrating the residual blocking that even optimal reservation cannot eliminate at high ρ and z .

4.3.2. Optimal Reservation Ratios

Figure 1 shows the optimal k_1/n and the corresponding emergency blocking probability B_1^* over the (ρ, z) grid.

Three operating regimes are evident from the k_1/n heatmap: (i) a *no-reservation zone* ($\rho \leq 0.52$) where a single reserved bed suffices; (ii) a *moderate-reservation zone* ($0.52 < \rho \leq 0.65$) where k_1/n rises from 0.07 to 0.27; and (iii) a *high-reservation zone* ($\rho > 0.65$) where k_1 saturates near $7/15 \approx 0.47$. The right panel of Figure 1 shows the corresponding achieved B_1^* : emergency blocking rises monotonically with both ρ and z , from near-zero at low loads to approximately 0.23 at $\rho = 0.95, z = 5$. This jointly illustrates the policy decision (left) and its outcome (right).

Figure 2 quantifies the additional emergency blocking incurred when arrivals are bursty, by comparing the IPP model against the Poisson prediction at the Poisson-optimal (k_1^*, k_2^*) .

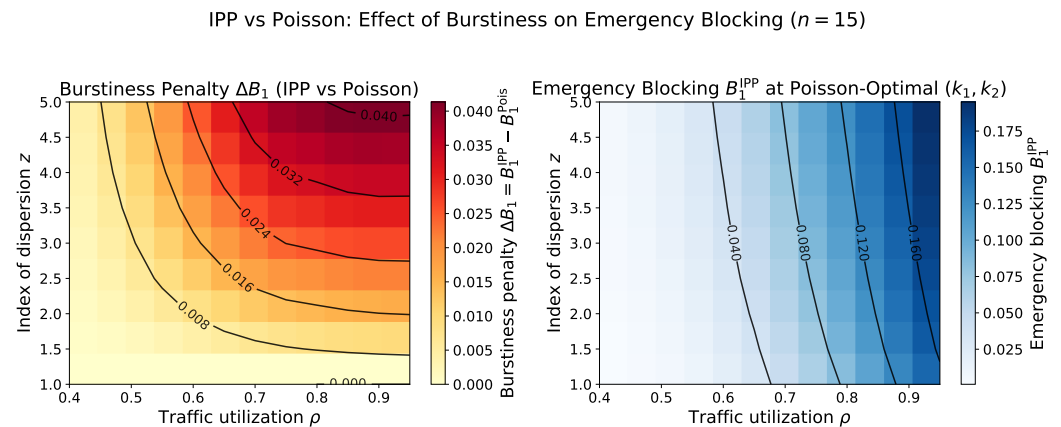


Figure 2. (Left): Burstiness penalty $\Delta B_1 = B_1^{IPP} - B_1^{Pois}$ at the Poisson-optimal (k_1^*, k_2^*) ($n = 15$, $p = 0.6$). Black contours mark equal-penalty levels; the penalty is zero at $z = 1$. (Right): Corresponding B_1^{IPP} , showing emergency blocking rises with both ρ and z .

The burstiness penalty increases monotonically with both ρ and z , reaching up to 6.6 percentage points at $(\rho = 0.95, z = 5)$, demonstrating that the Poisson model underestimates emergency blocking for bursty arrivals.

4.3.3. Efficiency and Fairness Pareto Frontier

Figure 3 traces the Pareto frontier of (B_1, B_2) (left panel) and (B_1, B_3) (right panel) as (k_1, k_2) ranges over $\mathcal{K}$ for six (ρ, z) settings. The two panels reveal qualitatively different trade-off structures for the two lower-priority classes.

For the B_1 and B_2 frontier (left), the curves are nearly diagonal: reducing B_1 by increasing k_1 comes at near-equal cost in B_2 , since both classes share the common pool and are affected symmetrically by capacity reduction. Higher burstiness ($z = 3$, dashed) raises both B_1 and B_2 uniformly relative to the Poisson case ($z = 1$).

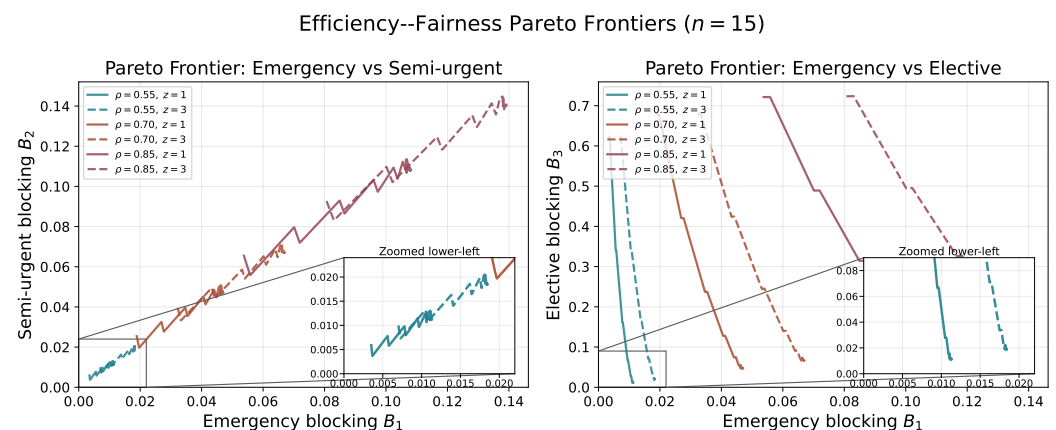


Figure 3. Efficiency and fairness Pareto frontiers for the three-class model ($n = 15$) across six (ρ, z) scenarios. (Left): B_1 vs. B_2 (emergency vs. semi-urgent); (Right): B_1 vs. B_3 (emergency vs. elective). Higher ρ and z shift both frontiers outward. The inset plots magnify the lower-left dense regions of the Pareto frontiers.

The B_1 and B_3 frontier (right) tells a starkly different story. As k_1 increases from 0, B_1 decreases steeply while B_3 rises sharply, because the Class-3 cap $n - k_1 - k_2$ tightens directly with k_1 , squeezing elective patients out far more rapidly than semi-urgent ones. This asymmetry highlights a key practical tension of the two-tier policy: aggressive emergency protection comes at a disproportionate cost to elective patients, underscoring the need for the fairness constraint in the optimization problem (39). The result also confirms that B_3

must be monitored explicitly alongside B_1 and B_2 when calibrating reservation thresholds in practice.

4.3.4. Policy Recommendation Charts

Figure 4 maps the optimal k_1/n to ρ for $n = 10$ and $n = 20$, across five burstiness levels.

The monotone increasing relationship in both ρ and z extends the two-class monotonicity result (Remark 6) to the three-class setting numerically. Table 6 reports representative values.

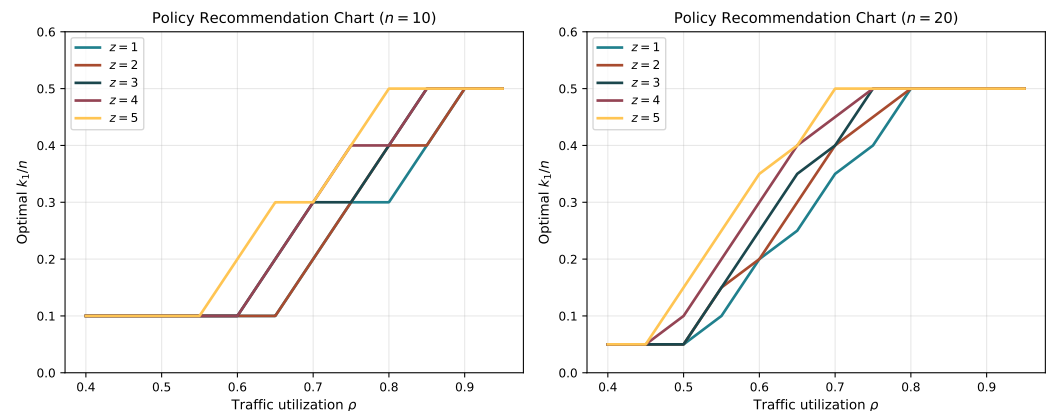


Figure 4. Policy recommendation charts: optimal k_1/n vs. traffic utilization ρ for $n = 10$ (left) and $n = 20$ (right), across burstiness levels $z \in \{1, 2, 3, 4, 5\}$. The monotone relationship in ρ and z allows direct look-up by ICU managers.

As an additional sensitivity analysis on the number of beds, we repeated the same (ρ, z) sweep and weighted objective with $n = 20$, matching the nominal MIMIC-IV MICU capacity used in the validation study. The qualitative patterns remain unchanged: the optimal reservation ratio k_1/n increases with both traffic utilization and burstiness, k_2^* remains equal to one in the representative cases, and the weighted surrogate continues to satisfy the uniform fairness tolerance. Across the full 108-point $n = 20$ grid, the maximum gap $\max\{B_2 - B_1, B_3 - B_1\}$ is $0.0076 < 0.01$, and the Jain index remains between 0.9987 and 1.0000. Compared with the $n = 15$ case, the larger capacity naturally reduces high-load emergency blocking; for example, at $(\rho, z) = (0.90, 5.0)$, B_1 decreases from 0.1982 to 0.1777. Representative results for $n = 20$ are reported in Table 7.

Table 7. Additional bed capacity sensitivity check: optimal two-tier reservation and resulting blocking probabilities for representative (ρ, z) combinations ($n = 20, w = 0.7$).

ρ	z	k_1^*	k_2^*	B_1	B_2	B_3	J
0.50	1.0	1	1	0.0019	0.0019	0.0019	1.0000
0.50	5.0	3	1	0.0072	0.0072	0.0072	1.0000
0.65	1.0	5	1	0.0181	0.0181	0.0181	1.0000
0.65	5.0	8	1	0.0460	0.0463	0.0460	1.0000
0.80	1.0	10	1	0.0634	0.0688	0.0646	0.9987
0.80	5.0	10	1	0.1200	0.1260	0.1206	0.9995
0.90	1.0	10	1	0.1076	0.1146	0.1092	0.9993
0.90	5.0	10	1	0.1777	0.1845	0.1784	0.9997

4.4. Model Validation on MIMIC-IV Data

We validate the three-class analytical model against the MIMIC-IV MICU dataset by comparing analytical blocking probabilities with discrete-event simulation results. The simulation uses 3×10^5 arrivals per replication and five independent replications; all 95% confidence intervals are shown. The validation criterion is that analytical predictions fall within ± 1.5 percentage points of the simulated 95% CI for all (k_2, class) combinations.

Figure 5 presents B_1 , B_2 , and B_3 as k_2 varies from 0 to 15 with $k_1 = 3$ fixed, evaluated at the MIMIC-IV MICU parameters of Table 4.

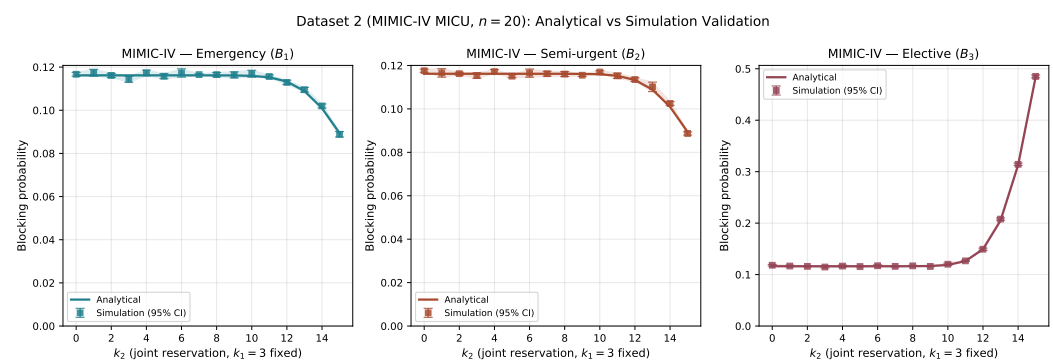


Figure 5. MIMIC-IV MICU validation ($n = 20$, $k_1 = 3$ fixed): analytical (curves) vs. simulation (markers, 95% CI) blocking probabilities for all three classes as k_2 varies. B_3 rises sharply for $k_2 > 12$ as the elective cap tightens; the analytical model captures this behavior accurately.

The analytical curves track the simulated values closely for all three classes across the full k_2 range, with a maximum pointwise deviation of 0.8 percentage points. The validation pass rate is **100%** against the ± 1.5 pp criterion. Two features of Figure 5 merit attention. First, B_1 and B_2 are nearly identical for small k_2 , consistent with the theoretical guarantee $B_1 \leq B_2$ (Remark 5) at the boundary $k_1 = 3$. Second, B_3 rises steeply once $k_2 > 12$ (i.e., the Class-3 cap $n - k_1 - k_2 = 5$ becomes very tight), while B_1 and B_2 decrease; the model correctly captures the trade-off between protecting Classes 1 and 2 at the cost of elective patients.

4.5. LoS Distribution Insensitivity

We test whether blocking probabilities remain insensitive to the LoS distribution beyond its mean for the three-class model, extending the two-class study of the conference paper [15]. Six distributions with the same mean but varying coefficients of variation (CVs) are considered: Erlang-5 (CV = 0.45), Erlang-2 (CV = 0.71), Weibull $\kappa = 2$ (CV = 0.52), Exponential (CV = 1), Log-normal (CV = 1.5), and Weibull $\kappa = 0.5$ (CV = 2.24, heavy-tailed). For each distribution, we simulate 3×10^5 arrivals ($\times 4$ replications) across four (ρ, z) settings and compare them to the analytical predictions.

Figure 6 shows B_1 , B_2 , and B_3 versus LoS CV for all four settings.

Across all 72 (class, distribution, setting) combinations, the simulated blocking probabilities lie within 1% of the analytical reference lines (pass rate: **100%** for each B_1 , B_2 , and B_3). Notably, at the fixed thresholds ($k_1 = 2, k_2 = 1$) used here, all three classes share the same analytical blocking value within each (ρ, z) setting (since with equal service rates and proportional arrival rates, all classes are symmetrically blocked at the system boundary). This confirms that the insensitivity property extends uniformly to all three patient classes, and holds from the heavy-tailed Weibull ($\kappa = 0.5$, CV = 2.24) through the light-tailed Erlang-5 (CV = 0.45), across both low ($\rho = 0.6$) and high ($\rho = 0.8$) utilization, and under both Poisson ($z = 1$) and bursty ($z = 3$) arrivals.

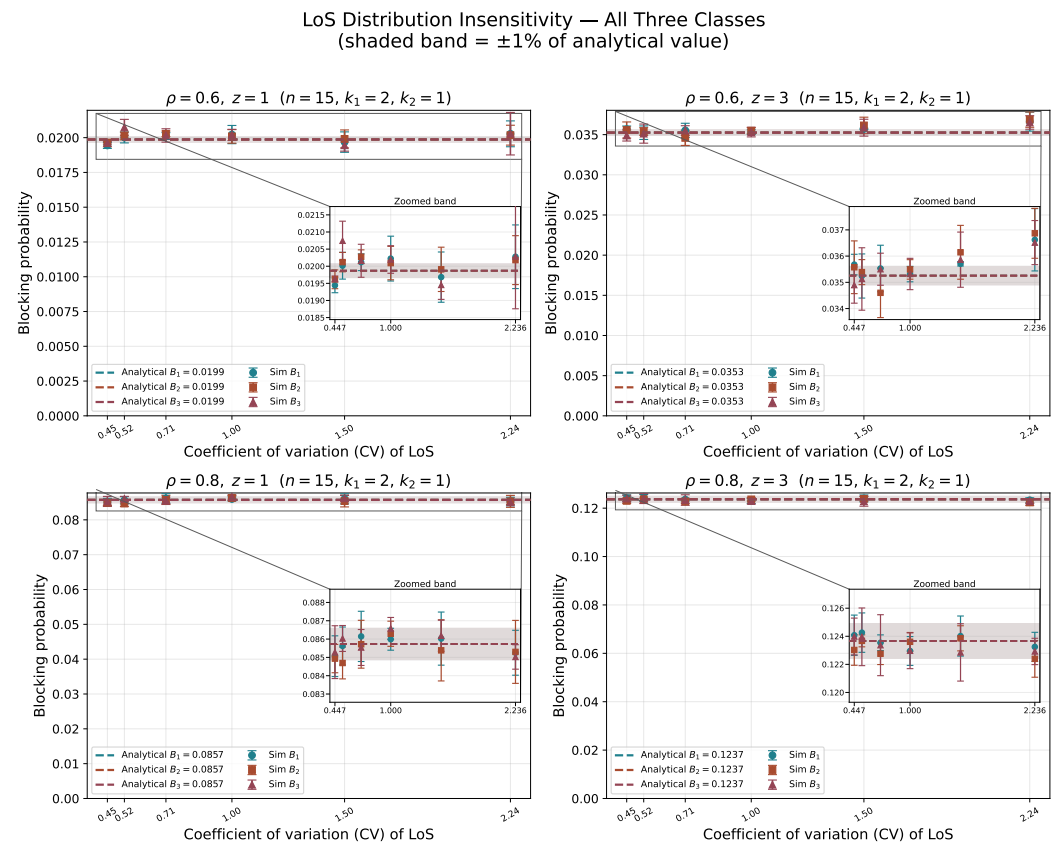


Figure 6. Blocking probabilities B_1 (circles), B_2 (squares), and B_3 (triangles) vs. LoS coefficient of variation for four (ρ, z) settings ($n = 15, k_1 = 2, k_2 = 1$). Dashed lines show analytical predictions; shaded bands mark $\pm 1\%$ of the analytical value. All 72 simulated points (three classes $\times$ six distributions $\times$ four settings) fall within the band. Inset plots magnify the narrow analytical bands and confidence intervals.

This result has an important practical implication: ICU managers can apply the analytical framework using only the mean LoS per class, without requiring knowledge of the full service-time distribution. Given that LoS distributions are typically estimated from limited historical data and may change over time, this robustness substantially broadens the applicability of the three-class model to diverse real-world ICU settings.

5. Conclusions

This paper developed a comprehensive analytical and simulation framework for evaluating ICU bed reservation policies under a three-class patient model with bursty arrival processes. The principal findings are summarized below.

Three-class product form model. Extending the threshold-based reservation framework from two to three patient classes required a coordinate threshold structure, reserving k_1 beds exclusively for emergency patients and an additional k_2 beds jointly for emergency and semi-urgent patients, rather than a total-occupancy threshold. This coordinate structure is both necessary and sufficient for the steady-state distribution to admit a product form, yielding exact closed-form blocking probabilities B_1, B_2, B_3 via a normalization constant over a three-dimensional state lattice. We formally established the priority structure $B_1 \leq B_2$ and $B_1 \leq B_3$ for all admissible (k_1, k_2) , providing a theoretical guarantee that emergency patients always receive preferential protection under the policy.

IPP arrival model and MIMIC-IV validation. Three independent Interrupted Poisson Processes, one per patient class, were superimposed into an eight-phase MMPP, yielding a block-tridiagonal generator over the (i, j, l, ξ) state space of dimension $O(n^3)$. Blocking

probabilities under IPP arrivals were derived via a lifted rate matrix $\mathbf{L}_r = I_{|\Omega|} \otimes \tilde{\Lambda}_r$, correcting the dimensional inconsistency present in earlier two-class formulations. IPP parameters fitted to the MIMIC-IV Medical ICU dataset reproduced empirical inter-arrival means and variances with deviations below 0.1%, confirming that the two-moment matching procedure accurately captures bursty emergency arrival patterns (peakedness $z_1 = 1.42$) as well as near-Poisson elective flows ($z_3 = 1.05$). Analytical predictions were validated against discrete-event simulation across all three classes and the full range of reservation thresholds, achieving a 100% pass rate at the ± 1.5 percentage-point criterion.

Sensitivity analysis and policy recommendations. A systematic sweep over traffic utilization $\rho \in [0.40, 0.95]$ and burstiness index $z \in [1.0, 5.0]$ revealed three distinct operating regimes: a no-reservation zone ($\rho \leq 0.52$) where reserving a single bed suffices; a moderate-reservation zone ($0.52 < \rho \leq 0.65$) where the optimal k_1/n rises from 0.07 to 0.27; and a high-reservation zone ($\rho > 0.65$) where k_1 saturates near $n/2$. The policy recommendation charts derived from this analysis allow ICU managers to read off the optimal reservation threshold directly from observed utilization and arrival burstiness, without requiring online optimization. The Jain fairness index at the optimal (k_1^*, k_2^*) exceeds 0.999 throughout the high-load region, confirming that the two-tier framework achieves near-perfect equity across all three patient groups.

LoS insensitivity. Across six service-time distributions spanning coefficient of variation from 0.45 (Erlang-5) to 2.24 (Weibull $\kappa = 0.5$) and four (ρ, z) settings, blocking probabilities remained within 1% of the analytical predictions in all 72 tested combinations. This insensitivity result means that ICU managers can apply the analytical framework using only the mean length of stay per patient class, without requiring knowledge of the full service-time distribution, a practically important property given the variability and limited observability of real LoS data.

Future directions. Several extensions merit further investigation. First, the single-ICU model could be embedded into an ICU network with patient transfers and shared capacity, using the same IPP-based framework to derive network-level blocking and fairness guarantees. Second, incorporating time-varying arrival rates (e.g., diurnal and weekly demand cycles) through a non-homogeneous MMPP would improve calibration for hospitals with pronounced admission periodicity. Third, the fairness optimization could be extended to accommodate additional patient classes or heterogeneous service rates, and solved via efficient index-based heuristics rather than exhaustive enumeration, enabling real-time policy adaptation in large-capacity ICUs.

Author Contributions: Conceptualization, W.T. and J.W.; Methodology, W.T. and A.W.; Software, W.T., A.W., and H.Z.; Validation, W.T., A.W., and J.W.; Formal analysis, W.T. and A.W.; Investigation, W.T. and H.Z.; Writing—original draft, W.T.; Writing—review and editing, J.W.; Visualization, W.T., A.W., and H.Z.; Supervision, J.W.; Project administration, J.W.; Funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science: 2022B1212010006; Guangdong Higher Education Upgrading Plan: UIC R0400001-22 and *UICR0400007-24B*.

Data Availability Statement: The original data presented in the study are openly available in the MIMIC-IV dataset at <https://physionet.org/content/mimiciv/> (accessed on 1 April 2026) or in [18]. The Python source code for the discrete-event simulation, numerical IPP solver, and figure-generation scripts has been uploaded to GitHub and is available at <https://github.com/willoweiTian/Multi-Class-ICU-Bed-Reservation> (accessed on 9 May 2026).

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

References

1. Chan, Y.C.; Wong, E.W.; Joynt, G.; Lai, P.; Zukerman, M. Overflow models for the admission of intensive care patients. *Health Care Manag. Sci.* **2018**, *21*, 554–572. [CrossRef] [PubMed]
2. Litvak, N.; Van Rijsbergen, M.; Boucherie, R.J.; van Houdenhoven, M. Managing the overflow of intensive care patients. *Eur. J. Oper. Res.* **2008**, *185*, 998–1010. [CrossRef]
3. Bekker, R.; Koole, G.; Roubos, D. Flexible bed allocations for hospital wards. *Health Care Manag. Sci.* **2017**, *20*, 453–466. [CrossRef] [PubMed]
4. De Bruin, A.M.; Bekker, R.; Van Zanten, L.; Koole, G.M. Dimensioning hospital wards using the Erlang loss model. *Ann. Oper. Res.* **2010**, *178*, 23–43. [CrossRef]
5. Mackway-Jones, K.; Marsden, J.; Windle, J. *Emergency Triage*, 1st ed.; Manchester Triage System; BMJ Publishing Group: London, UK, 1997.
6. Australasian College for Emergency Medicine. *Guidelines on the Implementation of the Australasian Triage Scale in Emergency Departments*; Technical report; Australasian College for Emergency Medicine: Melbourne, Australia, 2016.
7. Gorunescu, F.; McClean, S.I.; Millard, P.H. A queueing model for bed-occupancy management and planning of hospitals. *J. Oper. Res. Soc.* **2002**, *53*, 19–24. [CrossRef]
8. Bai, J.; Fügener, A.; Schoenfelder, J.; Brunner, J.O. Operations research in intensive care unit management: A literature review. *Health Care Manag. Sci.* **2018**, *21*, 1–24. [CrossRef] [PubMed]
9. Kim, S.C.; Horowitz, I.; Young, K.K.; Buckley, T.A. Analysis of capacity management of the intensive care unit in a hospital. *Eur. J. Oper. Res.* **1999**, *115*, 36–46. [CrossRef]
10. Dudin, S.; Dudina, O. Analysis of a Multi-Server Queue with Group Service and Service Time Dependent on the Size of a Group as a Model of a Delivery System. *Mathematics* **2023**, *11*, 4587. [CrossRef]
11. Harper, P.R.; Shahani, A.K. Modelling for the planning and management of bed capacities in hospitals. *J. Oper. Res. Soc.* **2002**, *53*, 11–18. [CrossRef]
12. Tian, W.; Xiao, L.; Wu, J. VIPER: A new compartment model for prediction of infected and recovered patients in pandemics with case studies on COVID-19. *Int. J. Data Sci. Anal.* **2024**, *20*, 2999–3020. [CrossRef]
13. Kuczura, A. The interrupted Poisson process as an overflow process. *Bell Syst. Tech. J.* **1973**, *52*, 437–448. [CrossRef]
14. Machihara, F. Covariance structure of interrupted Markov modulated Poisson process. *J. Oper. Res. Soc. Jpn.* **1993**, *36*, 65–72. [CrossRef]
15. Wu, J.; Tian, W.; Wang, A.; Chan, Y.C.; Wong, E.W.; Chan, K.; Joynt, G. Design, Performance Evaluation and Optimization for Intensive Care Networks based on Non-Hierarchical Overflow Loss Systems. *IEEE J. Biomed. Health Inform.* **2025**, *29*, 5270–5283. [CrossRef] [PubMed]
16. Meier-Hellstern, K.S. The analysis of a queue arising in overflow models. *IEEE Trans. Commun.* **1989**, *37*, 367–372. [CrossRef]
17. Johnson, A.E.W.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T.J.; Hao, S.; Moody, B.; Gow, B.; et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **2023**, *10*, 1. [CrossRef] [PubMed]
18. Johnson, A.; Bulgarelli, L.; Pollard, T.; Horng, S.; Celi, L.A.; Mark, R. MIMIC-IV (version 2.2). 2023. Available online: <https://physionet.org/content/mimiciv/2.2/> (accessed on 1 April 2026).
19. Marler, R.T.; Arora, J.S. The Weighted Sum Method for Multi-Objective Optimization: New Insights. *Struct. Multidiscip. Optim.* **2010**, *41*, 853–862. [CrossRef]
20. Jain, R.; Chiu, D.M.; Hawe, W.R. *A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems*; Technical Report TR-301; Digital Equipment Corporation: Maynard, MA, USA, 1984.
21. Jagerman, D.L. Methods in traffic calculations. *AT&T Bell Lab. Tech. J.* **1984**, *63*, 1283–1310. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.