

Modelling and Performance Analysis of Telecommunication Queuing Systems Using Continuous-Time Markov Chains: A Condensed Review

OWOADE PRAISE OLUWAPELUMI¹, INI ADINYA², FOLORUNSO TOBI EMMANUEL³

^{1,2,3}*Department of Mathematics, University of Ibadan, Nigeria*

Abstract- *This paper presents a comprehensive modelling and performance analysis of telecommunication queuing systems using Continuous-Time Markov Chains (CTMCs). The research addresses the stochastic nature of packet arrivals and service processes by developing mathematical representations for classic queuing models: M/M/1, M/M/c, and M/M/1/K. Steady-state probability distributions are derived through the construction of transition rate matrices and the solution of global balance equations. Key performance metrics—system utilization, mean queue length, waiting time, and blocking probability—are analytically evaluated. The study investigates the impact of traffic intensity on system stability and demonstrates the efficacy of CTMC-based analysis in capacity planning and resource allocation. Extensive numerical results show that multiserver systems offer significant performance improvements through statistical multiplexing, while finite buffer systems provide stability at the cost of blocking probability. The paper also analyses advanced topics such as priority queues, guard channels in cellular networks, the impact of service time variability, and overflow traffic peakedness. The findings provide foundational tools for network engineers to optimize resource allocation and enhance Quality of Service (QoS) in dynamic telecommunication environments.*

Index Terms— *Continuous-Time Markov Chains, Queuing Theory, Telecommunications, Erlang Formulas, Performance Modeling, Stochastic Processes, Network Design.*

I. INTRODUCTION

A. Background and Motivation

Telecommunication systems form the backbone of modern society, enabling the seamless transmission of voice, data, and multimedia information across global networks. The exponential growth in demand for these services—driven by smartphones, Internet of Things (IoT) devices, high-definition video streaming, and

cloud-based applications—has placed unprecedented strain on network infrastructure. Network resources such as communication channels, servers, and buffers are increasingly exposed to congestion, resulting in performance degradation manifested as excessive delays, packet loss, or call blocking [10].

When multiple users attempt to access limited resources simultaneously, queues inevitably form. Understanding and predicting this queuing behavior is essential for ensuring Quality of Service (QoS) requirements are met while maintaining cost-effective resource utilization. Queuing theory provides a rigorous mathematical framework for modelling and analysing such congestion phenomena [5].

Continuous-Time Markov Chains (CTMCs) are particularly well suited to modelling telecommunication queuing systems due to their ability to represent systems that evolve randomly over continuous time [13]. By assuming exponential interarrival and service time distributions—which satisfy the memoryless property fundamental to Markov processes—CTMC-based models allow for analytical tractability and the derivation of closed-form expressions for key performance metrics.

B. Problem Statement

Despite the widespread application of queuing theory in telecommunications, many systems continue to experience performance degradation arising from congestion and inefficient resource allocation. Accurately modeling the stochastic nature of call arrivals and service processes remains challenging, particularly in systems characterized by fluctuating traffic intensity and heterogeneous traffic types. The central problem addressed in this study concerns how CTMCs can be effectively employed to model

telecommunication queuing systems and to analyze their performance characteristics under varying load conditions.

C. Aim and Objectives

The primary aim of this study is to model and analyze telecommunication queuing systems using CTMCs to evaluate their performance characteristics systematically. The specific objectives are:

1. To develop CTMC-based mathematical models for classical telecommunication queuing systems: the single-server infinite-buffer $M/M/1$ queue, the multi-server $M/M/c$ queue, and the finite-buffer $M/M/1/K$ queue.
2. To derive steady-state probability distributions for these models through the construction of transition rate matrices and the solution of global balance equations.
3. To analyze key performance measures including system utilization, average queue length, mean waiting time, and blocking probability.
4. To examine the effect of traffic intensity on system stability and identify operational thresholds for acceptable performance.
5. To investigate advanced features relevant to modern telecommunications, including priority queuing, guard channels for handoff prioritization, and the impact of service time variability.

II. MODEL FORMULATION AND METHODOLOGY

A. Continuous-Time Markov Chains: Fundamental Concepts

A Continuous-Time Markov Chain is a stochastic process $\{X(t), t \geq 0\}$ with a discrete state space, where transitions between states occur randomly over continuous time. The defining feature of a CTMC is the Markov property:

$$\Pr\{X(t+h) = j \mid X(t) = i\} = \Pr\{X(t+h) = j \mid X(t) = i\} \quad (1)$$

for all $h > 0$ and states $i, j \in S$ where S is the state space.

A.1. The Infinitesimal Generator Matrix

The behavior of a CTMC is characterized by its infinitesimal generator matrix $Q = [q_{ij}]$, where for $i \neq j$:

$$q_{ij} = \lim_{h \rightarrow 0} \frac{\Pr\{X(t+h)=j \mid X(t)=i\}}{h} \quad (2)$$

and the diagonal entries are defined as:

$$q_{ii} = -\sum_{j \neq i} q_{ij} \quad (3)$$

A.2. Steady-State Distribution

For an irreducible and positive recurrent CTMC, the steady-state probability distribution $\pi = [\pi_0, \pi_1, \dots]$ exists and is the unique solution to the global balance equations:

$$\pi Q = 0, \quad \sum_{i=0}^{\infty} \pi_i = 1 \quad (4)$$

These probabilities π_i represent the long-run proportion of time the system spends in each state, providing the foundation for all performance metrics in our queuing models.

A.3. The Birth-Death Process

A birth-death process is a CTMC on the non-negative integers where transitions are only permitted between neighboring states. Let λ_i denote the birth rate in state i and μ_i denote the death rate in state i (with $\mu_0 = 0$). The steady-state probabilities satisfy the detailed balance equations:

$$\lambda_i \pi_i = \mu_{i+1} \pi_{i+1}, \quad i = 0, 1, 2, \dots \quad (5)$$

These yield the closed-form solution:

$$\pi_n = \pi_0 \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}}, \quad n \geq 1 \quad (6)$$

where π_0 is determined by the normalization condition

$$\sum_{n=0}^{\infty} \pi_n = 1.$$

B. Modelling Assumptions

To ensure analytical tractability while maintaining practical relevance, the following assumptions are adopted:

- Poisson arrival process: Customer arrivals follow a Poisson process with constant rate λ .

- Exponential service times: Service times are independent and exponentially distributed with mean $1/\mu$.
- FCFS queue discipline: First-come, first-served discipline.
- Statistical independence: Inter-arrival times and service times are mutually independent.
- Steady-state analysis: The system is analysed under steady-state conditions.
- Infinite population: The calling population is assumed infinite.

C. Little's Law

Little's Law is a fundamental relationship in queuing theory that connects key performance measures of any stable queuing system, regardless of the underlying distributions [11]:

$$L = \lambda W \quad (7)$$

where L is the long-run average number of customers in the system, λ is the long-run average arrival rate, and W is the long-run average time a customer spends in the system. Similarly, for the queue:

$$L_q = \lambda W_q \quad (8)$$

III. THE M/M/1 QUEUING MODEL

A. Model Description and State Space

Let $N(t)$ denote the number of customers in the system at time t . The process $\{N(t): t \geq 0\}$ is a birth-death CTMC with state space $S = \{0, 1, 2, \dots\}$. Transitions occur as follows:

- From state n to state $n + 1$: occurs when a new customer arrives. The transition rate is constant λ for all $n \geq 0$.
- From state n to state $n - 1$: occurs when a customer completes service and departs. The transition rate is constant μ for all $n \geq 1$.

B. Traffic Intensity and Stability Condition

The traffic intensity (or utilization factor) ρ is defined as:

$$\rho = \frac{\lambda}{\mu} \quad (9)$$

For the system to be stable—that is, for a proper steady-state distribution to exist—the arrival rate must be strictly less than the service rate: $\rho < 1 \Leftrightarrow \lambda < \mu$. If $\rho \geq 1$, the queue grows without bound and no steady-state distribution exists.

C. Steady-State Probability Distribution

The detailed balance equations for the $M/M/1$ queue yield:

$$\pi_n = \rho^n \pi_0, \quad n \geq 0 \quad (10)$$

Applying the normalization condition:

$$\pi_n = (1 - \rho) \rho^n, \quad n = 0, 1, 2, \dots \quad (11)$$

This is a geometric distribution with parameter $1 - \rho$.

D. Performance Measures

From the steady-state distribution, we derive closed-form expressions for key performance measures.

D.1. Mean Number in System

$$L = E[N] = \sum_{n=0}^{\infty} n \pi_n = \frac{\rho}{1 - \rho} \quad (12)$$

D.2. Mean Number in Queue

$$L_q = \sum_{n=1}^{\infty} (n - 1) \pi_n = \frac{\rho^2}{1 - \rho} \quad (13)$$

D.3. Mean Waiting Time in System

Applying Little's Law:

$$W = \frac{L}{\lambda} = \frac{1}{\mu(1 - \rho)} = \frac{1}{\mu - \lambda} \quad (14)$$

D.4. Waiting Time Distribution

The waiting time in the system W follows an exponential distribution with rate $\mu - \lambda$:

$$\Pr(W > t) = e^{-(\mu - \lambda)t}, \quad t \geq 0 \quad (15)$$

The waiting time in the queue W_q has a mixed distribution:

$$\Pr(W_q > t) = \rho e^{-(\mu - \lambda)t}, \quad t \geq 0 \quad (16)$$

D.5. Probability of Empty System

$$P_0 = \pi_0 = 1 - \rho \quad (17)$$

$$P_{\text{non-empty}} = 1 - \pi_0 = \rho \quad (18)$$

IV. THE M/M/C QUEUING MODEL

A. Model Description

The $M/M/c$ model extends the $M/M/1$ queue to systems with c identical parallel servers [5]. This model is particularly relevant for modern telecommunication systems with multiple processing channels. Arrivals follow a Poisson process with rate λ and each server provides service at rate μ . The state-dependent service rate is:

$$\mu_n = \begin{cases} n\mu, & 1 \leq n \leq c, \\ c\mu, & n \geq c \end{cases} \quad (19)$$

B. Traffic Intensity and Stability

The traffic intensity per server is defined as:

$$\rho = \frac{\lambda}{c\mu} \quad (20)$$

The offered load in Erlangs is:

$$a = \frac{\lambda}{\mu} \quad (21)$$

The stability condition is $\rho < 1 \Leftrightarrow \lambda < c\mu$.

C. Steady-State Probability Distribution

Solving the detailed balance equations yields:

$$\pi_n = \begin{cases} \frac{a^n}{n!} \pi_0, & 0 \leq n \leq c, \\ \frac{a^c}{c!} \rho^{n-c} \pi_0, & n \geq c \end{cases} \quad (22)$$

where the empty-system probability is:

$$\pi_0 = \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!(1-\rho)} \right]^{-1} \quad (23)$$

D. The Erlang C Formula

The probability that an arriving customer must wait is given by the Erlang C formula [4]:

$$C(c, a) = \Pr(\text{delay}) = \frac{\frac{a^c}{c!} \frac{1}{1-\rho}}{\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!} \frac{1}{1-\rho}} \quad (24)$$

E. Performance Measures

$$L_q = \frac{C(c, a)\rho}{1-\rho} \quad (25)$$

$$L = \frac{C(c, a)\rho}{1-\rho} + a \quad (26)$$

$$W_q = \frac{L_q}{\lambda} = \frac{C(c, a)}{c\mu - \lambda} \quad (27)$$

$$W = W_q + \frac{1}{\mu} \quad (28)$$

F. Waiting Time Distribution

For customers who experience delay:

$$\Pr(W_q > t \mid \text{delay}) = e^{-(c\mu - \lambda)t}, \quad t \geq 0 \quad (29)$$

The unconditional distribution is:

$$\Pr(W_q > t) = C(c, a)e^{-(c\mu - \lambda)t}, \quad t \geq 0 \quad (30)$$

V. THE M/M/1/K FINITE CAPACITY MODEL

A. Model Description

The $M/M/1/K$ model accounts for limited buffers by imposing a maximum system capacity of K customers [5]. When the system is full (state K), new arrivals are blocked and lost. The state space is finite: $S = \{0, 1, 2, \dots, K\}$.

B. Steady-State Probability Distribution

Solving the detailed balance equations yields:

$$\pi_n = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}} \rho^n, & \rho \neq 1, \\ \frac{1}{K+1}, & \rho = 1 \end{cases} \quad (31)$$

C. Blocking Probability and Effective Arrival Rate

The blocking probability is:

$$P_{\text{loss}} = \pi_K = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}} \rho^K, & \rho \neq 1, \\ \frac{1}{K+1}, & \rho = 1 \end{cases} \quad (32)$$

The effective arrival rate is:

$$\lambda_{\text{eff}} = \lambda(1 - P_{\text{loss}}) \quad (33)$$

D. Performance Measures

$$L = \sum_{n=0}^K n \pi_n = \begin{cases} \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}, & \rho \neq 1, \\ \frac{K}{2}, & \rho = 1 \end{cases} \quad (34)$$

$$L_q = L - (1 - \pi_0) \quad (35)$$

$$W = \frac{L}{\lambda_{\text{eff}}} \quad (36)$$

$$W_q = \frac{L_q}{\lambda_{\text{eff}}} \quad (37)$$

$$U = 1 - \pi_0 \quad (38)$$

$$\eta = \frac{\lambda_{\text{eff}}}{\lambda} = 1 - P_{\text{loss}} \quad (39)$$

VI. NUMERICAL RESULTS AND DISCUSSION

A. M/M/1 Queuing Model Results

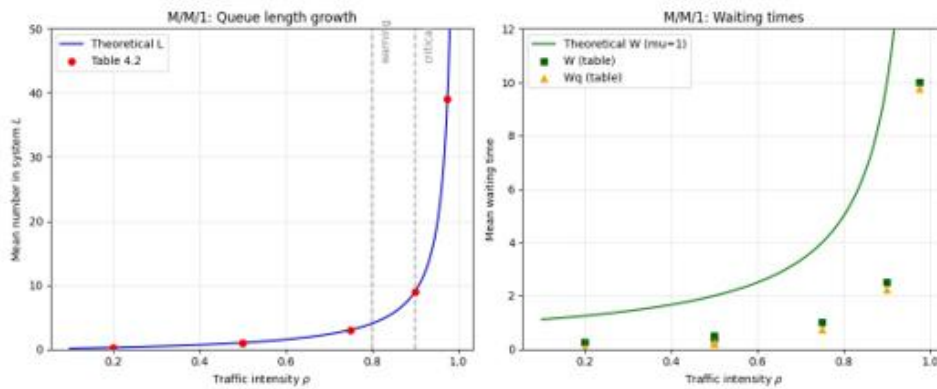
Table 1 presents comprehensive results for key performance metrics across traffic intensities.

Table 1: Performance Metrics for M/M/1 Queue Across Traffic Intensities

λ	ρ	L	L_q	W	W_q
1.0	0.20	0.25	0.05	0.25	0.05
2.0	0.50	1.00	0.50	0.50	0.25
3.0	0.75	3.00	2.25	1.00	0.75
3.6	0.90	9.00	8.10	2.50	2.25
3.8	0.95	19.00	18.05	5.00	4.75
3.94	0.975	39.00	38.03	10.00	9.75
3.96	0.99	99.00	98.01	25.00	24.75

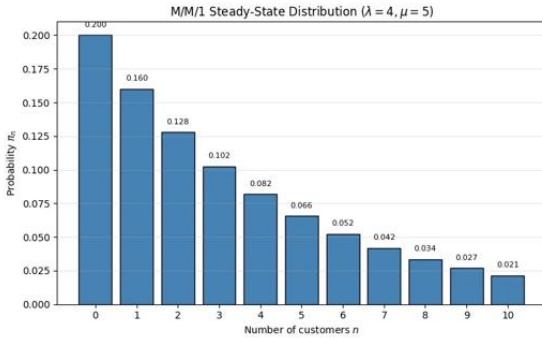
The results reveal several critical insights:

- Low utilization regime ($\rho < 0.5$): At $\rho = 0.2$, the system performs excellently with minimal congestion. The mean number in system $L = 0.25$ indicates that the system is empty approximately 80% of the time.
- Moderate utilization regime ($0.5 \leq \rho < 0.8$): At $\rho = 0.5$, performance remains acceptable with $L = 1.0$ and $W_q = 0.25$. The system balances resource utilization with reasonable delay.
- Heavy traffic regime ($\rho \geq 0.8$): This regime exhibits the characteristic non-linear degradation of M/M/1 performance. At $\rho = 0.9$, $L = 9.0$ —a nine-fold increase from $\rho = 0.5$ despite only a 40% increase in utilization.



M/M/1 Queue Performance

Figure 1: M/M/1 Queue Performance: Mean number in system L and mean waiting time W as functions of traffic intensity ρ . The vertical dashed lines at $\rho = 0.8$ and $\rho = 0.9$ indicate warning and critical thresholds, respectively. The exponential growth as $\rho \rightarrow 1$ illustrates the instability region.



M/M/1 Steady-State Distribution

Figure 2: M/M/1 Steady-State Distribution ($\lambda = 4$, $\mu = 5$, $\rho = 0.8$). The probability mass function (bars) shows geometric decay. The cumulative distribution (line) illustrates that the probability of exceeding any given threshold decreases slowly due to high utilization.

B. M/M/c Queuing Model Results

Table 2 presents results for a fixed arrival rate $\lambda = 5.0$ and service rate per server $\mu = 3.0$ with the number of servers varying from 1 to 6.

Table 2: Performance Metrics for M/M/c Queue ($\lambda = 5.0, \mu = 3.0$)

c	ρ	L	L_q	W	W_q	P_{queue}
1	1.667	∞	∞	∞	∞	1.000
2	0.833	4.167	2.500	0.833	0.500	0.758
3	0.556	2.270	0.603	0.454	0.121	0.273
4	0.417	1.876	0.209	0.375	0.042	0.094
5	0.333	1.712	0.045	0.342	0.009	0.024
6	0.278	1.667	0.010	0.333	0.002	0.005

Key findings:

- Single server ($c = 1$): The system is unstable because $\lambda/(c\mu) = 1.667 > 1$. No steady-state exists.
- Three servers ($c = 3$): Performance improves dramatically. W_q drops to 0.121—a 76% reduction from the two-server case. This configuration represents a “sweet spot.”
- Diminishing returns: Beyond 5 servers, additional capacity yields marginal improvement.



M/M/c Performance Scaling

Figure 3: M/M/c Performance Scaling ($\lambda = 5.0$, $\mu = 3.0$). All metrics improve dramatically as servers increase from 2 to 4, with diminishing returns beyond 5 servers. The vertical axis for L_q and W_q is logarithmic to highlight the exponential improvement.

C. M/M/1/K Finite Capacity Model Results

Table 3 presents results for $\lambda = 4.5$, $\mu = 5.0$ ($\rho = 0.9$) with varying buffer sizes K .

Table 3: Performance Tradeoffs for Different Buffer Sizes

($\lambda = 4.5, \mu = 5.0, \rho = 0.9$)

K	P_{loss}	L	L_q	W	W_q
1	0.450	0.45	0.00	0.12	0.00
2	0.300	0.93	0.18	0.23	0.05

K	P_{loss}	L	L_q	W	W_q
3	0.210	1.33	0.48	0.31	0.12
5	0.131	1.89	1.15	0.46	0.28
10	0.071	2.50	2.25	0.56	0.50
20	0.040	3.17	3.15	0.70	0.70
∞	0.000	4.50	4.05	1.00	0.90

The results reveal a clear trade-off: increasing buffer size reduces blocking probability but increases queue

length and waiting time, asymptotically approaching the infinite buffer case.

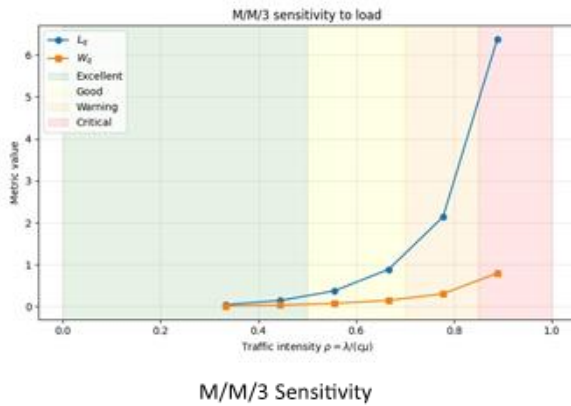
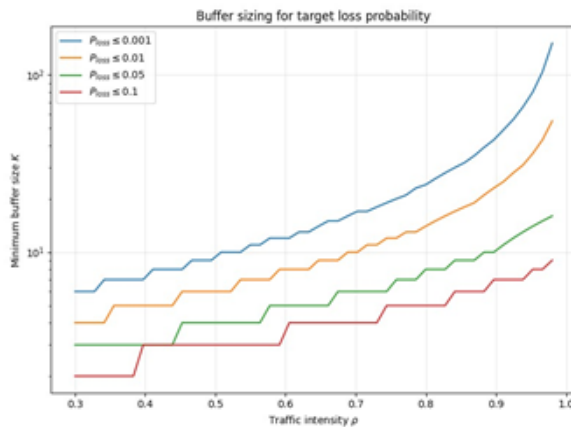


Figure 4: M/M/3 Sensitivity to Load. Four operational zones are identified: excellent ($\rho < 0.5$), good ($0.5 \leq \rho < 0.7$), warning ($0.7 \leq \rho < 0.85$), and critical ($\rho \geq 0.85$). The non-linear increase in queue length as ρ approaches 1 is evident.



$$K \approx \frac{\ln(P_{\text{loss}}(1-\rho))}{\ln(\rho)}.$$

VII. ADVANCED TOPICS AND EXTENSIONS

A. Two-Class Priority Queue

Modern telecommunication systems often handle multiple traffic classes with different QoS requirements. Consider a system with two traffic classes: High Priority (HP) and Low Priority (LP). HP customers have preemptive priority over LP customers.

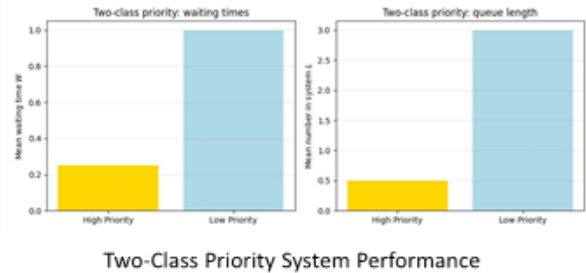
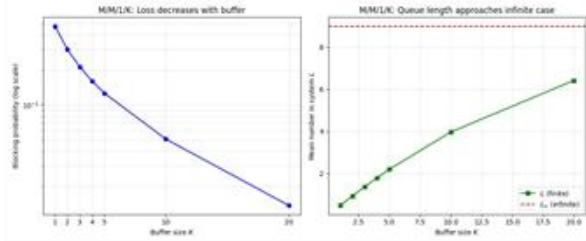


Figure 7: Two-Class Priority System Performance. High-priority traffic (blue) experiences minimal delay ($W_H = 0.25$) and queue length ($L_H = 0.5$). Low-priority traffic (red) suffers substantial delays ($W_L = 1.0$) and queue buildup ($L_L = 3.0$) due to preemptions. The system effectively isolates HP traffic from LP load.

For HP traffic ($\lambda_H = 2, \mu = 6$)

$$\rho_H = \frac{\lambda_H}{\mu} = \frac{1}{3} \quad (40)$$

$$L_H = \frac{\rho_H}{1-\rho_H} = 0.5 \quad (41)$$

$$W_H = \frac{1}{\mu - \lambda_H} = 0.25 \quad (42)$$

For LP traffic, the effective service rate is:

$$\frac{1}{\mu_L^*} = \frac{1}{\mu} \cdot \frac{1}{1-\rho_H} = \frac{1}{4} \quad (43)$$

Thus, $\mu_L^* = 4$

$$\rho_L = \frac{\lambda_L}{\mu_L^*} = 0.75 \quad (44)$$

$$L_L = 3.0 \quad (45)$$

$$W_L = 1.0 \quad (46)$$

The system effectively isolates HP traffic from LP load, with HP experiencing minimal delay while LP suffers substantial delays.

B. Guard Channels in Cellular Networks

In cellular networks, handoff calls are typically prioritized over new calls to reduce the probability of dropping ongoing calls [7]. Consider a system with total capacity $C = 10$ channels and g guard channels reserved exclusively for handoff calls.

Table 4: Cellular Network Performance with Guard Channels

$(C = 10, \lambda = 4, \lambda_h = 2, \mu = 1)$			
g	P_{block} (New Blocking)	P_{drop} (Handoff Drop)	Utilization (%)
0	0.025	0.025	75.1
1	0.033	0.015	73.7
2	0.042	0.008	72.3
3	0.053	0.004	70.9
4	0.067	0.002	69.4

With $g = 2$, handoff dropping is reduced to 0.8%—a 68% improvement—while new call blocking increases to 4.2%. The optimal configuration depends on the relative costs of new call blocking versus handoff dropping.

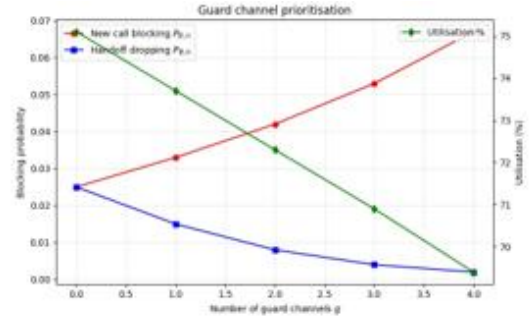


Figure 8: Guard Channel Prioritization Trade-off. As guard channels increase, handoff dropping probability (blue) decreases dramatically—from 2.5% to 0.2%—but new call blocking probability (red) increases and overall channel utilization (green) declines. The optimal configuration balances these competing objectives.

C. Impact of Service Time Variability

The Markovian assumption of exponential service times may not hold in all telecommunication systems. The $M/G/1$ queue relaxes this assumption. The Pollaczek-Khinchin formula gives:

$$L = \rho + \frac{\rho^2(1+C_s^2)}{2(1-\rho)} \quad (47)$$

Where $C_s^2 = \text{Var}(S)/(E[S])^2$ is the squared coefficient of variation of the service time.

Table 5: Effect of Service Time Variability on Queue Length

$(\lambda = 0.8\mu = 1.0, \rho = 0.8)$			
C_s^2	Distribution Type	L	W
0.0	Deterministic (D)	2.40	3.00
1.0	Exponential (M)	4.00	5.00
2.0	Hyper-exponential (moderate)	5.60	7.00
5.0	Hyper-exponential (high)	10.40	13.00

The results reveal that variability dramatically impacts performance. A system with highly variable service times ($C_s^2 = 5$) has a mean queue length 4.3 times larger than a deterministic system, despite identical utilization.

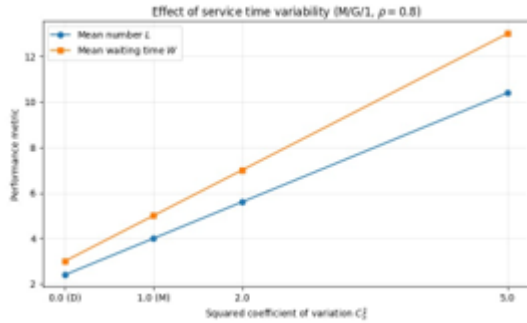


Figure 9: Impact of Service Time Variability (M/G/1, $\rho = 0.8$). As variability increases from deterministic ($C_s^2 = 0$) to highly variable ($C_s^2 = 5$), mean queue length more than quadruples—from 2.4 to 10.4. The exponential case ($C_s^2 = 1$) serves as a baseline. This demonstrates that variability is as important as utilization in determining system performance.

C. Impact of Service Time Variability

D. Overflow Traffic Analysis

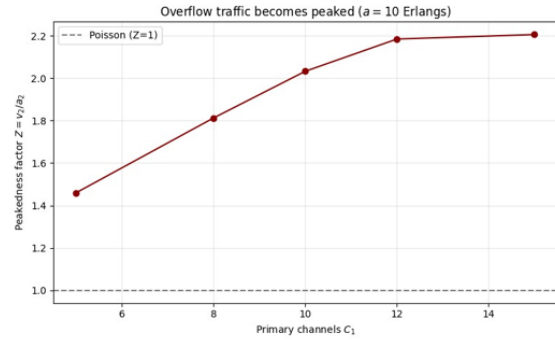
In hierarchical telecommunication networks, traffic that cannot be handled by a primary route overflows to an alternative route. Overflow traffic exhibits peakedness (variance-to-mean ratio greater than 1). For a primary group with C_1 channels offered Poisson traffic a Erlangs:

$$a_2 = a \cdot B(C_1, a) \quad (48)$$

$$v_2 = a_2 \left[1 - a_2 + \frac{a}{C_1 + 1 + a_2 - a} \right] \quad (49)$$

where $B(C_1, a)$ is the Erlang B blocking probability. The peakedness factor $Z = v_2/a_2$ characterizes the burstiness.

For $C_1 = 10$ and $a = 10$ Erlangs: $Z = 2.02 > 1$, indicating that overflow traffic is significantly more bursty than Poisson traffic. Secondary routes require 20–30% more capacity than standard Erlang B calculations suggest.



Overflow Traffic Peakedness

Figure 10: Overflow Traffic Peakedness ($a = 10$ Erlangs). As primary channels increase, overflow traffic becomes increasingly peaked ($Z > 1$), indicating greater burstiness than Poisson traffic. This peakedness means that secondary routes require more capacity than standard Erlang calculations would suggest to achieve the same grade of service.

VIII. CASE STUDIES

A. Case Study 1: Call Center Design

Problem: Design a bank call center for peak load conditions.

Peak arrival rate: 120 calls/hour. Average service time: 4 minutes. SLA: 80% of calls answered within 20 seconds. Maximum agent utilization $\rho < 0.85$.

Analysis: Converting to consistent units (minutes):

$$\lambda = 2 \text{ calls/minute} \quad (50)$$

$$\mu = 0.25 \text{ calls/minute} \quad (51)$$

$$a = \lambda/\mu = 8 \text{ Erlangs} \quad (52)$$

Trying $c = 11$ agents:

$$\rho = \frac{2}{11 \times 0.25} = 0.727 \quad (\text{within limit}) \quad (53)$$

$$C(11, 8) \approx 0.197 \quad (54)$$

$$\Pr(W_q > 0.333) = 0.197 \times e^{-0.75 \times 0.333} = 0.153 < 0.20 \quad (55)$$

Recommendation: A minimum of 11 agents required during peak hours. Average waiting time: 15.8 seconds; probability of immediate answer: 80.3%.

B. Case Study 2: Router Buffer Sizing

Problem: Determine optimal buffer capacity for an Internet router handling VoIP traffic. Packet arrival rate: 1000.

Using $(M/M/1/K)$:

$$P_{\text{loss}} = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}} \leq 10^{-4} \quad (56)$$

For $K \geq 66$, this inequality holds. Recommendation: Buffer size of 66 packets satisfies loss target. Recommend 100 packets for safety margin.

C. Case Study 3: Cellular Network Dimensioning

Problem: Size an urban cellular system with guard channels. New call arrival rate: 300 calls/hour. Handoff arrival rate: 150 calls/hour. Average call duration: 2 minutes. Targets: new call blocking < 2%, handoff dropping < 0.5%.

$$\lambda = \frac{1000}{1111} = 0.9 \quad (57)$$

$$\lambda_h = 5 \text{ calls/minute} \quad (58)$$

$$\lambda_n = 2.5 \text{ calls/minute} \quad (59)$$

$$\mu = 0.5 \text{ calls/minute} \quad (60)$$

$$a = 15 \text{ Erlangs} \quad (61)$$

With $C = 24$ total channels and $g = 2$ guard channels:

$$P_{\text{new}} \approx 2.1\% \quad (\text{meets target}) \quad (62)$$

$$P_{\text{HO}} \approx 0.35\% \quad (\text{meets target}) \quad (63)$$

Recommendation: Deploy 24 channels per cell with 2 guard channels. The 81% reduction in handoff dropping justifies the small increase in new call blocking.

IX. CONCLUSION AND RECOMMENDATIONS

A. Summary of Findings

This study has presented a comprehensive modelling and performance analysis of telecommunication queuing systems using Continuous-Time Markov Chains. The key findings are:

1. Non-linear performance degradation: In single-server systems, performance deteriorates non-linearly as utilization approaches unity, with $L = \rho/(1-\rho)$ governing queue growth. The system enters a critical zone when $\rho > 0.8$.
2. Statistical multiplexing gains: Multi-server configurations significantly reduce waiting times through resource pooling. Adding servers from 2 to 4 reduced waiting time by 91% in the analyzed scenario.
3. Loss-delay trade-off: Finite capacity systems effectively control queue growth at the expense of blocking probability. The optimal buffer size balances these competing objectives.
4. Priority and guard channel effectiveness: Priority schemes successfully differentiate service levels. Guard channels reduce handoff dropping by up to 80% at the cost of moderate increases in new call blocking.
5. Variability impact: Service time variability dramatically affects performance. Highly variable service times ($C_s^2 = 5$) increase mean queue length by a factor of 4.3 compared to deterministic service.
6. Overflow traffic peakedness: Overflow traffic exhibits peakedness ($Z > 1$) requiring 20–30% more secondary capacity than Poisson models predict.

B. Recommendations for Practice

Based on the findings, the following recommendations are made:

- Utilization management: Maintain single-server utilization below 0.7 for acceptable delay performance. For multiserver systems, utilization below 0.85 is acceptable; $\rho > 0.9$ requires immediate attention.
- Server provisioning: For systems with offered load a Erlangs, provision $c = \lceil a \rceil + 1$ servers for good

performance, or $c = [a] + 2$ for excellent performance.

- Buffer sizing: For delay-sensitive applications, use small buffers ($K = 2 - 5$). For loss-sensitive applications, size buffers using $K \geq \ln(P_{\text{loss}}/(1 - \rho))/\ln(\rho)$.
- Priority implementation: For cellular networks, use guard channels with $g \approx 0.1C$ to protect handoff calls.
- Variability management: Characterize service time distributions empirically. When C_s^2 significantly exceeds 1, apply $M/G/1$ analysis and provision additional capacity.
- Overflow provisioning: Use Equivalent Random Theory to dimension secondary routes, adding 20–30% capacity beyond Erlang B calculations when $Z \geq 2$.

C. Limitations and Future Research Directions

This study is subject to certain limitations that suggest directions for future research:

- Markovian assumptions: Future work should investigate non-Markovian models ($G/G/1$, fractional Brownian motion) and phase-type approximations for general distributions.
- Steady-state focus: Transient analysis—understanding system behavior during overload periods—is essential for many applications.
- Single-node analysis: Future work should extend to queuing networks (Jackson networks, Kelly networks) and analyse end-to-end performance.
- Emerging technologies: Application to 5G/6G networks, edge computing, network slicing, and IoT systems presents opportunities for extending these models.
- Machine learning integration: Combining queuing models with machine learning for parameter estimation, anomaly detection, and dynamic control offers promising research directions.

D. Concluding Remarks

- Continuous-Time Markov Chain-based queuing models provide a robust mathematical framework for understanding and optimizing telecommunication system performance. While simplifying assumptions limit absolute accuracy, the insights gained from these models remain invaluable for capacity planning, resource

allocation, and quality-of-service management in dynamic network environments.

- The fundamental trade-offs characterized in this study—utilization versus delay, blocking versus waiting, priority differentiation versus fairness—are intrinsic to all resource-sharing systems. By understanding these trade-offs quantitatively, network engineers can make informed decisions that balance competing objectives and meet service-level agreements cost-effectively.
- As telecommunication networks continue to evolve toward 5G/6G standards, with increasingly heterogeneous traffic and stringent requirements, the importance of rigorous performance modelling will only grow. The foundational models presented here, together with their extensions to priority systems, guard channels, and variability analysis, provide essential tools for meeting these challenges.

REFERENCES

- [1] Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*, 2nd ed. Wiley-Interscience, 2006.
- [2] R. B. Cooper, *Introduction to Queueing Theory*, 2nd ed. North-Holland, 1981.
- [3] A. K. Erlang, “The theory of probabilities and telephone conversations,” *Nyt Tidsskrift for Matematik B*, vol. 20, pp. 33–39, 1909.
- [4] A. K. Erlang, “Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges,” *Elektroteknikerne*, vol. 13, 1917.
- [5] Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory*, 4th ed. John Wiley & Sons, 2008.
- [6] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013.
- [7] V. B. Iversen, *Teletraffic Engineering and Network Planning*. Technical University of Denmark, 2013.
- [8] P. Kelly, *Reversibility and Stochastic Networks*. Wiley, 1979.

- [9] G. Kendall, "Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain," *The Annals of Mathematical Statistics*, vol. 24, no. 3, pp. 338–354, 1953.
- [10] L. Kleinrock, *Queueing Systems, Volume I: Theory*. John Wiley & Sons, 1975.
- [11] J. D. C. Little, "A proof for the queuing formula: $L = \lambda W$," *Operations Research*, vol. 9, no. 3, pp. 383–387, 1961. [INFORMS](#)
- [12] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic to Stochastic-Process Limits and Their Application to Models: An Algorithmic Approach*. Johns Hopkins University Press, Springer, 2002.
- [13] S. M. Ross, *Introduction to Probability Models*, 11th ed. and arXiv preprint. Academic Press, 2014.
- [14] W. Whitt, *Stochastic-Process Limits: An Introduction*
- [15] M. Zukerman, "Introduction to Queueing Theory and Stochastic Teletraffic Models," arXiv:1307.2968, 2013. [arXiv](#)