MATH 305 - Stochastic Calculus
Queuing Theory Optimal Traffic Light Duration

Group 3
Dang Hung Duong[*1], Le Nguyen Hoang Ly[*2],
Nguyen Thien Toan[*3], Vo Ngoc Khanh Linh[*4]

[*]Fulbright University Vietnam
[1]duong.dang.220218@student.fulbright.edu.vn
[2]ly.le.220045@student.fulbright.edu.vn
[3]toan.nguyen.220107@student.fulbright.edu.vn
[4]linh.vo.220169@student.fulbright.edu.vn

May 16, 2025

Contents

1 Introduction

1.1 Motivation

Ho Chi Minh City (HCMC), Vietnam's vibrant economic hub and most populous city, grapples daily with the consequences of intense traffic congestion. This phenomenon significantly hampers urban mobility and impacts the quality of life for its residents. The city's road network, especially during the peak commute periods from approximately 7:00-9:00 AM and 4:30-7:00 PM, frequently reaches saturation points. While congestion occurs along various road segments, signalized intersections represent critical junctures where traffic flows converge and are periodically halted, making them primary locations for queue formation and significant delays. Understanding and managing the flow through these intersections is therefore paramount to mitigating broader network congestion.

The traffic composition in HCMC, characterized by an exceptionally high density of motorbikes operating alongside cars, buses, and other vehicles, presents unique challenges at these intersections. The sheer volume and mixed nature of vehicles vying for passage during green light cycles contribute to complex interactions and often inefficient clearance, leading to long queues forming during red light phases. Major intersections connecting arterial roads frequently become severe bottlenecks, where the demand for passage consistently strains the capacity offered by the traffic signal cycles. This interplay between vehicle arrival patterns and the constraints imposed by traffic light timings is central to the congestion experienced at these points. The following image depicts a typical scene at such an intersection.

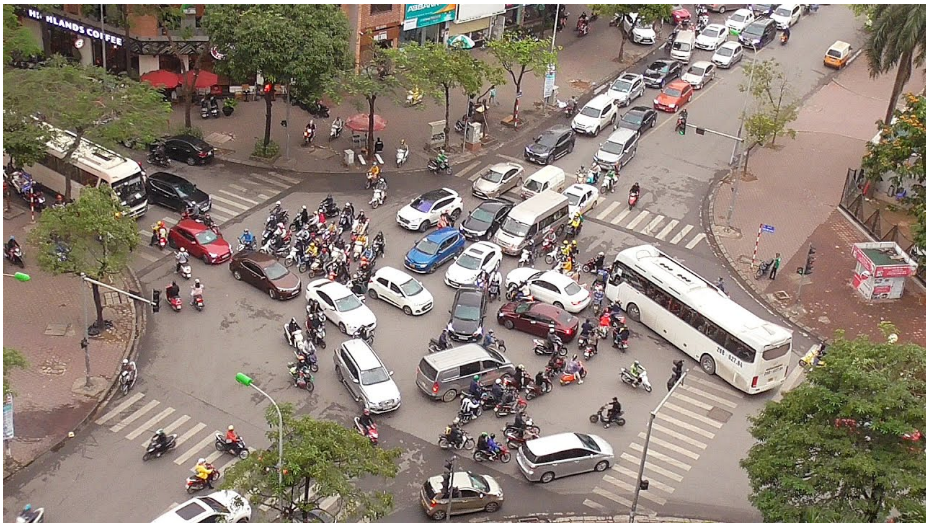

Figure 1: A typical traffic scene at a major signalized intersection in Ho Chi Minh City, illustrating the queue formation and high density of mixed vehicles.

The consequences stemming from inefficiently managed intersections and the resulting queues are substantial. Economically, prolonged idling during red lights and slow clearance during green lights lead to significant fuel waste and increased emissions, contributing to air pollution. Lost time accumulates for commuters and commercial transport, impacting productivity and logistical efficiency. Socially, the daily experience of long waits at intersections adds to commuter stress and frustration, while the environmental impact of concentrated emissions and noise pollution degrades the urban environment. Furthermore, safety can be compromised during the scramble for passage when the light turns green or by risky behaviors induced by long waits.

Our motivation for this project stems directly from observing these daily challenges at HCMC's intersections. We are intrigued by the potential to apply rigorous mathematical methods to analyze and potentially improve the situation, moving beyond simple observation. Specifically, we aim to investigate how the principles of queuing theory, grounded in the stochastic processes studied in this course, can be used to model the arrival and departure (service) of vehicles at a signalized intersection. The core interest lies in understanding quantitatively how traffic light parameters – specifically the duration of green and red light phases – influence key performance metrics like average queue length and waiting time. Our motivation is therefore focused: can we use queuing theory to analyze the performance of an intersection under different signal timing strategies and gain insights that could inform optimization efforts? This provides a concrete application of stochastic modeling to address a specific, critical aspect of urban traffic management.

Therefore, the specific focus of this project will be the analysis of a **signalized multi-lane intersection** using queuing theory. We aim to model the intersection approach lanes as queuing systems where vehicles arrive randomly, wait during the red light phase, and are 'served' (i.e., pass through the intersection) during the green light phase. The 'service rate' in this context is intrinsically linked to the duration of the green light and the saturation flow rate (the maximum rate at which vehicles can pass when the light is green). By modeling vehicle arrivals (e.g., using a Poisson process) and the service mechanism dictated by the signal timing, we intend to apply queuing models (such as variations or interpretations of M/M/1 or M/M/c, or potentially time-dependent models) to estimate performance measures like the average number of vehicles waiting and the average time a vehicle spends waiting. The ultimate goal of the analysis is to understand how adjustments to the signal cycle length and green time allocation could potentially reduce delays and queue lengths, thereby demonstrating how queuing theory can serve as an analytical tool for traffic light optimization.

1.2 Literature review

The application of queuing theory to model traffic flow at signalized intersections is a well-established field within transportation engineering and operations research, dating back several decades [2]. Understanding vehicle arrivals and departures as stochastic processes allows for the estimation of key performance metrics like delay and queue length, which are critical for efficient traffic management.

Foundational models often simplified intersection dynamics by employing variations of basic queuing systems. Early work frequently utilized M/M/1 (Poisson arrivals, exponential service times, single server) or M/D/1 (Poisson arrivals, deterministic service times) queues as approximations [6]. The 'server' in this context represents the capacity of the intersection approach during the green light phase. The assumption of Poisson arrivals is common due to its mathematical tractability and its reasonable representation of vehicle arrivals during non-saturated conditions, although its validity under heavy congestion or platooning is debated [4]. Deterministic or general service times (M/D/1, M/G/1) are often considered more realistic for representing the relatively constant saturation flow rate once a queue starts moving during a green phase [1].

A key challenge unique to traffic signals is modeling the periodic availability of service due to the red-green cycle. Queuing models must account for these interruptions. This has been addressed in several ways:

- **Effective Service Rate:** Some approaches average the service capacity over the entire cycle time (green + red + yellow), yielding an effective service rate, as implicitly used in steady-state approximations [8]. This is often sufficient for estimating long-term average performance but may obscure cycle-by-cycle dynamics.
- **Server Vacations:** More sophisticated models explicitly treat the red phase as a 'server vacation,' where the server is unavailable [**?**]. These can provide more detailed insights but are often mathematically more complex.
- **Time-Dependent Analysis:** Recognizing that queues build during red and dissipate during green, time-dependent queuing analysis attempts to capture the transient behavior within a cycle [7]. However, closed-form solutions are often intractable.

Performance estimation, particularly vehicle delay, has been a major focus. Webster's seminal work provided widely used (though initially deterministic, later adapted) formulas for delay and optimal cycle time, often serving as benchmarks [8]. Many refinements using queuing theory followed, aiming to better capture stochastic variations and overflow queues (vehicles not clearing within one green phase) [5]. The average queue length is another critical output, directly related to delay via Little's Law [3] and important for assessing the risk of queue spillback blocking upstream intersections.

Fundamentally, many of these queuing systems, particularly those involving Poisson arrivals (Markovian arrivals), can be modeled and analyzed using the theory of Continuous-Time Markov Chains (CTMCs), which forms a core part of stochastic calculus studies. The state space typically represents the number of vehicles in the queue, and transition rates correspond to vehicle arrivals and service completions (departures). The generator matrix and stationary distribution analysis of the underlying CTMC are key tools for deriving long-term performance metrics like average queue length and, subsequently, average waiting time.

While foundational models provide valuable insights, limitations exist. Real-world traffic often exhibits non-Poisson arrivals (e.g., platooning from upstream signals) and time-varying arrival rates (peak vs. off-peak). The high density of mixed traffic (motorbikes, cars, buses) in cities like HCMC introduces further complexity not typically captured in standard M/M/c or M/D/1 models, potentially affecting saturation flow rates and vehicle interactions. Our project aims to apply the core principles of queuing theory, grounded in CTMC concepts where applicable (like arrival processes), to model and optimize signal timings for specific HCMC-relevant scenarios, acknowledging the simplifications involved but leveraging the analytical power of these stochastic methods.

1.3 Problem statement

1.3.1 Problem statement

1.3.2 Relevance to the course

Relevance to Stochastic Calculus

This project's focus on analyzing signalized intersections through queuing theory remains highly relevant to the principles of Stochastic Calculus. Queuing models are fundamentally applications of stochastic processes, typically Continuous-Time Markov Chains (CTMCs), used to describe systems evolving randomly over time. By studying queuing systems like the M/M/1 or M/M/c models, which are often used as foundational approximations for intersection analysis, we directly engage with CTMCs, exploring concepts such as state spaces (number of waiting vehicles), transition rates (vehicle arrivals and departures), and steady-state behavior. This provides a valuable extension from the Discrete-Time Markov Chains covered in the course, illustrating how these ideas adapt to continuous time and form the basis for analyzing dynamic systems. Furthermore, the project requires modeling the inherent randomness crucial to traffic flow – specifically, the unpredictable arrival times of vehicles (often modeled as a Poisson process) and the variability in the time taken for vehicles to clear the intersection during a green light (which can be approximated, for instance, by an exponential distribution reflecting an average service rate). Applying stochastic calculus concepts to capture and analyze this randomness is central to understanding the intersection's performance.

Moreover, the specific focus on traffic light optimization adds another layer of relevance. While basic queuing models often assume constant parameters, analyzing a traffic light inherently involves a system whose 'service' mechanism is time-dependent and controllable (switching between red and green phases). Although a full time-dependent analysis might be complex, even using steady-state approximations where the 'server' represents the capacity during the green phase allows us to see how stochastic modeling outputs (like predicted queue lengths and delays) are directly influenced by controllable parameters (green time duration). This demonstrates how insights derived from analyzing underlying stochastic processes can inform control and optimization strategies, a common theme in many advanced applications of stochastic calculus in engineering and operations research.

Importance of the Problem Domain

The problem of optimizing traffic signal timings at intersections is of immense practical importance in urban traffic management. Signalized intersections are the critical control points within any city's road network, and their efficient operation is essential for maintaining traffic flow, reducing delays, and ensuring safety. In a city like Ho Chi Minh City, where congestion is a major daily challenge, improving the performance of key intersections can have a significant positive impact on the entire network's efficiency. Traffic engineers constantly seek better methods to determine optimal signal timings (cycle lengths, phase durations, offsets), often relying on simulation software, heuristics, or simplified deterministic models. This project aims to demonstrate the value of applying a more rigorous, probabilistic approach using queuing theory. By providing a quantitative framework to estimate performance metrics like average queue length and waiting time under different signal strategies, queuing theory offers a powerful analytical tool to complement existing methods. Demonstrating this link between stochastic modeling and practical traffic engineering highlights the societal relevance of applying advanced mathematical techniques to solve pressing urban problems.

Interest for the Class

We anticipate this project will capture the interest of our classmates for several compelling reasons. Firstly, it delves into Queuing Theory and its connection to Continuous-Time Markov Chains, potentially broadening the scope of stochastic processes explored in the course and introducing a versatile modeling technique applicable in numerous fields beyond traffic. Secondly, the chosen application – optimizing traffic lights – is highly relatable and tangible. Everyone experiences waiting at intersections, making the problem context immediate and the goal of reducing delays easily understood. The local HCMC context further enhances this relatability. Thirdly, the focus on *optimization* provides a clear, practical objective for the mathematical modeling. It moves beyond simply describing a system to actively seeking ways to improve it, showcasing how theoretical analysis can lead to actionable insights. Seeing how changes in a controllable parameter like green light duration mathematically impact predicted queue lengths and wait times can be particularly illustrative. Finally, the planned simulation will offer a dynamic visualization of these concepts, allowing classmates to see the modeled intersection operate under different signal timings and observe the resulting effects on traffic flow and queuing, making the theoretical results more concrete and engaging.

2 Methodology

2.1 Theory

2.1.1 Continuous-time Markov chain

Definition. Markov Property
A continuous-time stochastic process $(X_t)_{t\geq 0}$ with discrete state space $\mathcal{S}$ is a *continuous-time Markov chain if*

$$P(X_{t+s} = j \mid X_s = i, X_u = x_u, 0 \leq u < s) = P(X_{t+s} = j \mid X_s = i),$$

for all $s, t \geq 0, i, j, x_u \in S$, and $0 \leq u < s$.
The process is said to be *time-homogeneous* if this probability does not depend on s. That is,

$$P(X_{t+s} = j \mid X_s = i) = P(X_t = j \mid X_0 = i)$$

for $s \geq 0$.

A continuous-time Markov chain $(X_t)_{t\geq 0}$ with discrete state space $\mathcal{S} = \{S_1, S_2, ..., S_n\}$ is specified by the transition matrix $\tilde{\mathbf{P}}$, exponential time parameters $(\lambda_{S_1}, \lambda_{S_2}, ..., \lambda_{S_n})$, and initial distribution α.

For each $t \geq 0$, the transition function of the Markov chain is a matrix function $\mathbf{P}(t)$, with

$$P_{ij}(t) = P(X_t = j \mid X_0 = i).$$

This transition function $\mathbf{P}(t)$ has similar properties as that of the transition matrix for a discrete-time Markov chain.

Long-term behavior

In queueing theory, we are often interested in the long-run performance of a system, rather than its behavior at a specific moment in time. This includes questions such as:

- What is the average number of customers in the system?
- What is the probability that the server is busy?
- What is the average waiting time for traffic lines?

To answer these questions, we analyze the long-term (steady-state) behavior of the continuous-time Markov chain that models the queue.

The generator in continuous-time Markov chains

In queueing theory, systems are often modeled using continuous-time Markov chains. The behavior of these chains is described by a matrix called the generator, or infinitesimal generator, denoted by Q.

Definition.
Let $X(t)$ be a continuous time Markov-chain with $X(0) = i$. The time until the process leaves state i is exponentially distributed with rate λ_i. For a small time interval $\delta > 0$, the probability of leaving state i is approximately:

$$\mathbb{P}(T_1 < \delta \mid X(0) = i) \approx \lambda_i \delta$$

Formally, we define:

$$\lambda_i = \lim_{\delta \to 0^+} \frac{\mathbb{P}(X(\delta) \neq i \mid X(0) = i)}{\delta}$$

Constructing the Generator Matrix
Let p_{ij} be the probability of transitioning to state j given a jump from state i. Then the off-diagonal elements of the generator matrix are defined as:

$$q_{ij} = \lambda_i p_{ij}, \quad \text{for } i \neq j$$

The diagonal elements are chosen to ensure each row sums to zero:

$$q_{ii} = -\sum_{j \neq i} q_{ij} = -\lambda_i$$

The generator is essential in determining how the system evolves over time, and it is used to find the stationary distribution by solving $\pi Q = 0$, which gives the long-run behavior of the system.

Example

Consider a simple system with two states. The generator matrix is:

$$Q = \begin{bmatrix} -2 & 2 \\ 1 & -1 \end{bmatrix}$$

This means:

- From state 1, the system moves to state 2 at rate 2. The total rate of leaving state 1 is 2, so $q_{11} = -2$.
- From state 2, the system moves to state 1 at rate 1. The total rate of leaving state 2 is 1, so $q_{22} = -1$.

This simple matrix shows how the generator captures both the direction and speed of transitions in a continuous-time Markov chain model.

Stationary distribution

The stationary distribution of a Markov chain represents the long-term behavior of the system, where the probabilities of being in each state remain constant over time.

Definition. Let $X_0, X_1, \dots$ be a Markov chain with transition matrix P. A stationary distribution is a probability distribution π, which satisfies

$$\pi = \pi P$$

That is,

$$\pi_j = \sum_i \pi_i P_{ij}, \quad \text{for all} \quad j.$$

Key Properties

Equilibrium Behavior: If the chain starts in the stationary distribution π, then the distribution at time n will always be π, for any $n \geq 1$:

$$\pi P^n = \pi \quad \text{for all} \quad n \geq 1$$

This means the chain reaches a **steady state**, where the probabilities of being in each state do not change over time.

Stationary Distribution in Long-Run: The stationary distribution describes the long-term behavior of the system. Once the system reaches this distribution, it stays in it indefinitely.

Proof.
We are given a positive stochastic matrix P of size $n \times n$, where $P_{ij} > 0$ and each row sums to 1. We aim to show that there exists a stationary distribution $\pi = (\pi_1, \pi_2, \dots, \pi_n)$, with $0 < \pi_i < 1$ and $\sum_i \pi_i = 1$, such that

$$P^\infty = \lim_{i \to \infty} P^i = \begin{bmatrix} \pi_1 & \pi_2 & \dots & \pi_n \\ \pi_1 & \pi_2 & \dots & \pi_n \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \dots & \pi_n \end{bmatrix},$$

and that for any initial distribution σ, the sequence $\sigma(i)$, defined by $\sigma(1) = \sigma$ and $\sigma(i+1) = \sigma(i)P$, converges to π.

For a fixed column k of P^i, define

$$m(i) = \min_r P^i_{rk}, \quad M(i) = \max_r P^i_{rk}, \quad \Delta(i) = M(i) - m(i).$$

We first prove monotonicity. For the minimum:

$$m(i+1) = \min_r \sum_s P_{rs} P^i_{sk} \geq \min_r \sum_s P_{rs} m(i) = m(i).$$

For the maximum:

$$M(i+1) = \max_r \sum_s P_{rs} P^i_{sk} \leq \max_r \sum_s P_{rs} M(i) = M(i).$$

Define $p_{\min} = \min_{r,s} P_{rs} > 0$. We have the lower bound:

$$m(i+1) \geq m(i) + p_{\min}(M(i) - m(i)) = m(i) + p_{\min}\Delta(i),$$

and the upper bound:

$$M(i+1) \leq M(i) - p_{\min}(M(i) - m(i)) = M(i) - p_{\min}\Delta(i).$$

Subtracting the two:

$$\Delta(i+1) = M(i+1) - m(i+1) \leq (1 - 2p_{\min})\Delta(i).$$

Iterating:

$$\Delta(n) \leq (1 - 2p_{\min})^{n-1}\Delta(1),$$

which converges to zero as $n \to \infty$ since $0 < 1 - 2p_{\min} < 1$.

Therefore, each column k of P^i converges to a constant π_k, giving

$$P^\infty = \begin{bmatrix} \pi_1 & \pi_2 & \dots & \pi_n \\ \pi_1 & \pi_2 & \dots & \pi_n \\ \vdots & \vdots & \ddots & \vdots \\ \pi_1 & \pi_2 & \dots & \pi_n \end{bmatrix}.$$

Since each P^i is stochastic,

$$P^i\mathbf{1} = \mathbf{1},$$

it follows that

$$P^\infty\mathbf{1} = \mathbf{1},$$

and thus $\sum_j \pi_j = 1$, with $\pi_j \geq 0$.

For stationarity, observe:

$$P^\infty P = \lim_{i\to\infty} P^i P = \lim_{i\to\infty} P^{i+1} = P^\infty,$$

implying

$$\pi P = \pi.$$

For any initial distribution σ,

$$\lim_{i\to\infty} \sigma(i) = \lim_{i\to\infty} \sigma P^{i-1} = \sigma P^\infty = \pi.$$

Therefore, we have shown that π is a stationary distribution, P^i converges to P^∞, and any initial distribution σ converges to π.

2.1.2 Queuing theory

Queuing models

In the context of queuing theory, customers arrives at a facility, wait for service, and then leave when the service is completed. A queuing model can have many parameters, each describing a characteristic of the model, such as its capacity, number of servers, and how customers are lined up. A standard notation only consists of three components and has the following form.

Definition: A queuing model is described by the standard notation

$$A/B/n$$

where A denotes the arrival time distribution, B the service time distribution, and n the number of servers.

A basic queuing model is the $M/D/1$ queue, in which M stands for Makovian (Poisson), D stands for deterministic, and 1 means there is one server . This model will be used to describe a traffic light system, where vehicles arrive following a Poisson distribution, the serving time (duration of green light) is fixed, and each lane is treated as a single server.

Little's Formula

A central result in queuing theory is the Little's formula. This formula describes the relationship between the rate of arrivals, the average number of customers, and the average time a customer stays in the system in the long term.

Theorem: (Little's Formula)
Given a queuing system. Let L denote the long-term average number of customers in the system, λ the rate of arrivals, and W the long-term average time that a customer spends in the system. Then,

$$L = \lambda W.$$

Proof. In the scope of this project, we will present a rough proof by picture of the Little's formula. Our aim is to give an intuitive explanation of the relationship between the terms involved in Little's equation, sufficient for us to understand the idea behind the theorem and apply it directly to our real-world problems later on.

Consider a realization of a queuing system between the time the first customer enters the system ($t_0 = 0$) and the time the system is next empty ($t' = t$). Suppose that the number of customers entering the system in the interval $[0, t]$ is $n_t = 4$. The entering and exiting times of each customer are recorded in the table below.

Customer	Arrival time	Departure time
1	t_1	t_5
2	t_2	t_4
3	t_3	t_7
4	t_6	t

The number of customers in the system over time is demonstrated in the graph below.

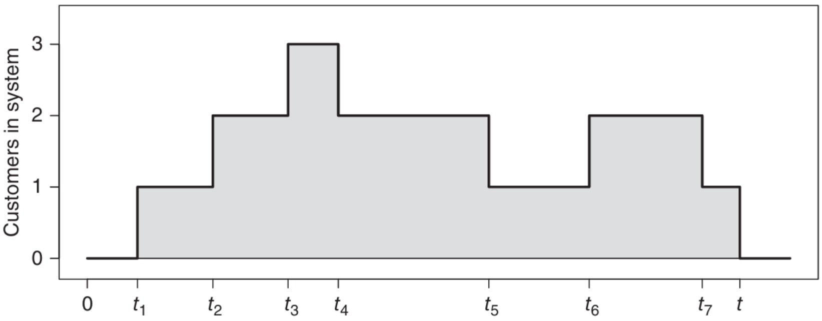

Figure 2: A realization of the queuing system in $[0, t]$

Let S be the area of the shaded region. Then A can be computed in two ways.

(1) The average length of time W that a customer spends in the system is

$$\begin{aligned} W &= \frac{(t_5 - t_1) + (t_4 - t_2) + (t_7 - t_3) + (t - t_6)}{4} \\ &= \frac{(t - t_1) + (t_5 - t_2) + (t_7 - t_6) + (t_4 - t_3)}{4} \\ &= \frac{S}{n_t} \end{aligned}$$

$$\Rightarrow S = W n_t$$

(2) The average number of customers L in the system is

$$\begin{aligned} L &= \frac{1(t_2 - t_1) + 2(t_3 - t_2) + 3(t_4 - t_3) + 2(t_5 - t_4)}{t} \\ &= \frac{S}{t} \end{aligned}$$

$$\Rightarrow S = Lt$$

Hence, we have

$$W n_t = S = Lt$$
$$L = \frac{n_t}{t} W.$$

As t increases, $\frac{n_t}{t}$ converges to λ. Therefore, in the long term, $L = \lambda W$.

A queuing system must satisfy Little's formula to be stable, e.g. the waiting line never increases infinitely. This means the severs has the ability to complete the amount of service requested.

2.2 Problem formulation

With the theoretical foundations presented, we can now address the central goal of this study: traffic signal optimization. This section presents a comprehensive formulation of the problem, starting from a real-world setting and proceeding to assumptions and mathematical modeling.

Vehicles arrive at a signalized intersection according to a Poisson process with arrival rate λ (vehicles per second). The intersection is governed by a cyclic traffic light that alternates between green, red, and yellow phases. The yellow phase has a fixed duration of $t_y = 4$ seconds per transition. We consider two distinct control scenarios:

(a) Paired Directions (N/S and E/W): The North–South (N/S) and East–West (E/W) directions are paired and alternate their green phases.

(b) Independent Directions (N, E, S, W): Each direction has its own dedicated green time, with exactly one direction being served at a time.

The objective is to determine the optimal green durations that minimize the total vehicle delay at the intersection. The system is modeled using $M/D/1$ queues with deterministic service and periodic interruptions.

2.2.1 Model assumptions

- Vehicle arrivals follow a Poisson process with rate λ (vehicles/sec), varying by direction.
- During the green phase, vehicles are served at a constant rate μ (vehicles/sec).
- The full traffic signal cycle includes green, red, and fixed yellow phases.
- Yellow light duration is fixed at $t_y = 4$ seconds per transition:
  - Scenario (a): Two transitions (N/S → E/W and back) ⇒ 8 seconds total yellow.
  - Scenario (b): Four transitions (each direction once) ⇒ 16 seconds total yellow.
- Total cycle time is $C = \text{green durations} + \text{yellow durations}$.
- The effective service rate is given by: $\mu_{\text{eff}} = \mu \cdot \frac{G}{C}$.
- Each queue is modeled as an $M/D/1$ system with server availability governed by the green phase.

2.2.2 General formula

To ensure system stability:

$$\rho = \frac{\lambda}{\mu_{\text{eff}}} = \frac{\lambda C}{\mu G} < 1$$

By Little's Law, the expected number of vehicles in the system is:

$$L = \frac{\rho}{1-\rho}$$

So the average waiting time is:

$$W = \frac{L}{\lambda} = \frac{1}{\mu_{\text{eff}} - \lambda} = \frac{C}{\mu G - \lambda C}$$

Objective: Find G (green time) that minimizes $W(G)$, subject to total cycle constraints.

2.2.3 Scenario (a): Paired Directions (NS and EW)

Assume:

$$\lambda_{NS} = \frac{\lambda_N + \lambda_S}{2}, \quad \lambda_{EW} = \frac{\lambda_E + \lambda_W}{2}$$

Let G_{NS} and G_{EW} be green durations for the N/S and E/W directions, respectively. Then:

$$C = G_{NS} + G_{EW} + 8$$

The expected delays for the two flows are:

$$W_{NS} = \frac{C}{\mu G_{NS} - \lambda_{NS} C}, \quad W_{EW} = \frac{C}{\mu G_{EW} - \lambda_{EW} C}$$

With the constraint:

$$G_{NS} + G_{EW} = C - 8 \Rightarrow G_{EW} = C - 8 - G_{NS}$$

We define the total delay as a function of G_{NS}:

$$W_{\text{total}}(G_{NS}) = \frac{C}{\mu G_{NS} - \lambda_{NS} C} + \frac{C}{\mu(C - 8 - G_{NS}) - \lambda_{EW} C}$$

Taking the derivative and solving for optimal G_{NS}:

$$\begin{aligned} \frac{dW}{dG_{NS}} &= \frac{-\mu C}{(\mu G_{NS} - \lambda_{NS} C)^2} + \frac{\mu C}{(\mu(C - 8 - G_{NS}) - \lambda_{EW} C)^2} = 0 \\ \Rightarrow \mu G_{NS} - \lambda_{NS} C &= \mu(C - 8 - G_{NS}) - \lambda_{EW} C \end{aligned}$$

Solve for G_{NS}^*:

$$\begin{aligned} 2\mu G_{NS} &= \mu(C - 8) + C(\lambda_{NS} - \lambda_{EW}) \\ G_{NS}^* &= \frac{1}{2}\left(C - 8 + \frac{C(\lambda_{NS} - \lambda_{EW})}{\mu}\right) \end{aligned}$$

Then:

$$G_{EW}^* = C - 8 - G_{NS}^*$$

2.2.4 Scenario (b): Independent Directions

Let G_i denote the green duration for each direction $i \in \{N, E, S, W\}$. The cycle constraint is:

$$C = G_N + G_E + G_S + G_W + 16 = G_{\text{sum}} + 16$$

Expected delay per direction:

$$W_i = \frac{C}{\mu G_i - \lambda_i C}$$

Objective function:

$$W_{\text{total}} = \sum_i \frac{C}{\mu G_i - \lambda_i C} \quad \text{subject to } \sum G_i = C - 16$$

Using Lagrange multipliers with multiplier γ, define:

$$\mathcal{L}(G_N, G_E, G_S, G_W, \gamma) = \sum_i \frac{C}{\mu G_i - \lambda_i C} + \gamma \left(\sum G_i - (C - 16) \right)$$

Partial derivatives:

$$\frac{\partial \mathcal{L}}{\partial G_i} = \frac{-\mu C}{(\mu G_i - \lambda_i C)^2} + \gamma = 0 \Rightarrow \mu G_i - \lambda_i C = \sqrt{\frac{\mu C}{\gamma}}$$

Solving for G_i:

$$G_i = \frac{1}{\mu} \left(\lambda_i C + \sqrt{\frac{\mu C}{\gamma}} \right)$$

Substitute into the constraint:

$$\sum G_i = \frac{C}{\mu} \sum \lambda_i + \frac{4}{\mu} \sqrt{\frac{\mu C}{\gamma}} = C - 16$$

Solve numerically for γ, then compute the optimal G_i values.

2.3 Data Collection and Analysis

In Vietnam, traffic consists of a heterogeneous mix of vehicles including motorcycles, cars, buses, and light trucks. To perform a consistent and meaningful traffic flow analysis, these vehicles must be standardized into a common unit. This is done using the Passenger Car Unit (PCU) method, which assigns a weight to each vehicle type based on its relative impact on traffic flow compared to a standard car.

Vehicle Type	**PCU Value**
Motorcycle	0.30
Car (≤12 seats)	1.00
Bus (12–30 seats)	1.25
Truck (< 2 tons)	1.50

Source: JICA (2004), Urban Transport Master Plan – Ho Chi Minh City; Chungsuk Engineering & Bach Khoa (2006, p.49)

These PCU values were used to convert raw vehicle counts into standardized flow rates at the Nguyen Huu Tho – Nguyen Thi Thap intersection, a major signalized junction in Ho Chi Minh City. Manual counts were conducted during the peak morning period (7:30–8:00 AM) on May 12, 2025. Traffic volume was recorded in all four directions over a 7-minute observation window and then converted to PCU per second:

Location	**Motorcycles**	**Cars**	**Buses**	**Trucks**	**PCU/7 min**	**PCU/s**
Nguyen Huu Tho (North)	358	53	2	14	183.90	0.4379
Nguyen Huu Tho (South)	315	35	2	19	160.50	0.3821
Nguyen Thi Thap (East)	446	65	1	40	260.05	0.6192
Nguyen Thi Thap (West)	422	46	1	30	218.85	0.5211

These PCU/second values (λ_i) were then used as the input arrival rates in the queuing model formulations for Scenarios (a) and (b). The directional imbalance in flow—particularly the heavier load in the East and West directions—played a key role in shaping the optimal signal timing results.

2.4 Simulations

The codes for the simulations can be found **here**.

This simulation phase validates the analytical formulations of both scenarios using real-world traffic data and numerical optimization techniques.

Input Parameters:

- **Arrival rates** (in PCU/s), based on observed data:

$$\lambda_N = 0.438, \quad \lambda_E = 0.619, \quad \lambda_S = 0.382, \quad \lambda_W = 0.521$$

- **Service rate:** $\mu = 5.5$ vehicles/s (assumed constant across directions)
- **Signal cycle length:** $C = 110$ seconds

Scenario (a): Two Paired Directions (NS and EW)
In this scenario, green time is allocated to two groups of traffic (North-South and East-West). Given the convex nature of the total delay function,

$$W_{\text{total}}(G_{NS}) = \frac{C}{\mu G_{NS} - \lambda_{NS} C} + \frac{C}{\mu(C - 8 - G_{NS}) - \lambda_{EW} C}$$

we apply a scalar optimization method to minimize total delay. Specifically, `scipy.optimize.minimize_scalar` is used to determine the optimal green time G^*_{NS}, with G^*_{EW} computed from the cycle constraint.

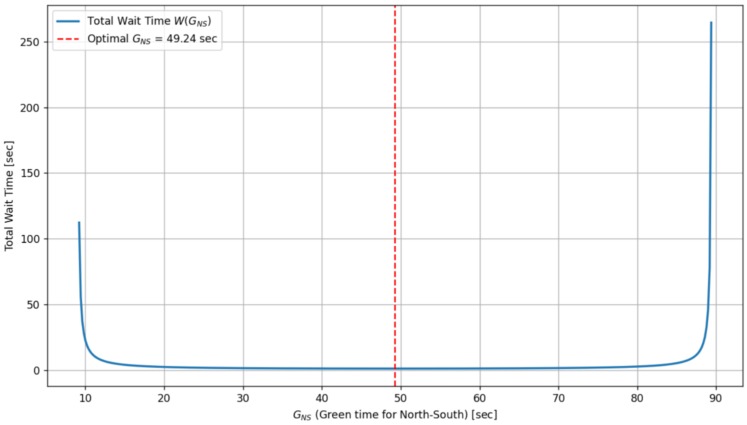

Figure 3: Scenario (a): Total Delay as a Function of G_{NS}

Figure 3 confirms that the delay function exhibits a clear minimum, indicating successful convergence to the globally optimal green time allocation for the paired-direction strategy.

Scenario (b): Independent Directional Control
In this case, each approach (N, E, S, W) receives an individually optimized green time G_i, subject to the constraint:

$$\sum_i G_i = C - 16$$

Lagrange multipliers are applied to minimize total delay while satisfying the cycle time constraint. The solution expresses each green time G_i as a function of an unknown multiplier γ:

$$G_i = \frac{1}{\mu}\left(\lambda_i C + \sqrt{\frac{\mu C}{\gamma}}\right)$$

Substituting into the constraint yields a nonlinear equation in γ, which is solved numerically using a root-finding method (e.g., bisection or Newton-Raphson).

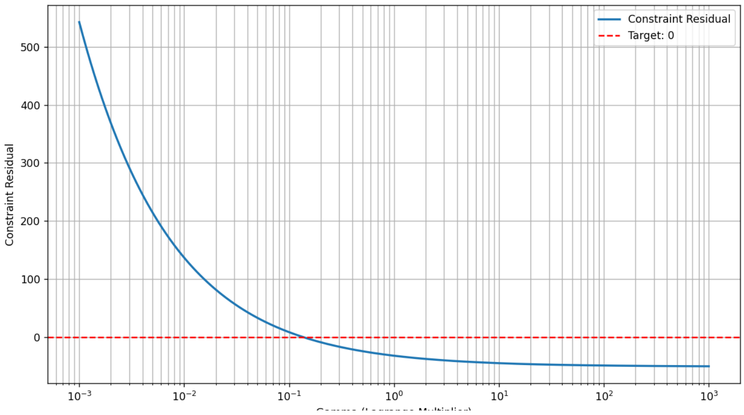

Figure 4: Scenario (b): Constraint Residual vs γ

Figure 4 shows the residual of the constraint equation $\sum G_i - (C - 16)$ as a function of γ. The x-axis represents candidate γ values, while the y-axis shows how far the resulting $\sum G_i$ deviates from the required total.

The point where the curve intersects the horizontal axis (i.e., residual equals zero) indicates that the constraint is perfectly satisfied. This verifies that the computed γ is valid and the corresponding green times G_i adhere exactly to the total cycle time constraint. This is essential for ensuring the feasibility and correctness of the optimization solution.

Dynamic Visualization:

To further illustrate the effectiveness of the optimized signal strategies, we developed time-based animations using `matplotlib.animation`. These visualizations simulate queue formation and dissipation under each scenario, with synchronized traffic signal transitions. This not only demonstrates the impact of green time allocation but also provides an intuitive understanding of how queue lengths evolve over time for different control policies.

Overall, the simulations verify that the analytical models produce realistic and efficient signal timings. Scenario (b) provides a more fine-grained control, potentially reducing overall delay further, at the cost of increased phase complexity.

3 Results

Using queuing theory, we derived expressions for expected waiting times in two traffic light scenarios. We transformed the traffic flow problem into an optimization problem by modeling the system as an $M/D/1$ queue with interruptions. Under stability conditions, the resulting delay functions are convex, which allows for numerical solutions to determine optimal signal timings that minimize total vehicle delay. We evaluated our traffic signal optimization approach using real traffic flow data from the Nguyen Huu Tho – Nguyen Thi Thap intersection, testing two strategies: Scenario (a) with paired directions (NS vs. EW), and Scenario (b) with independent direction control.

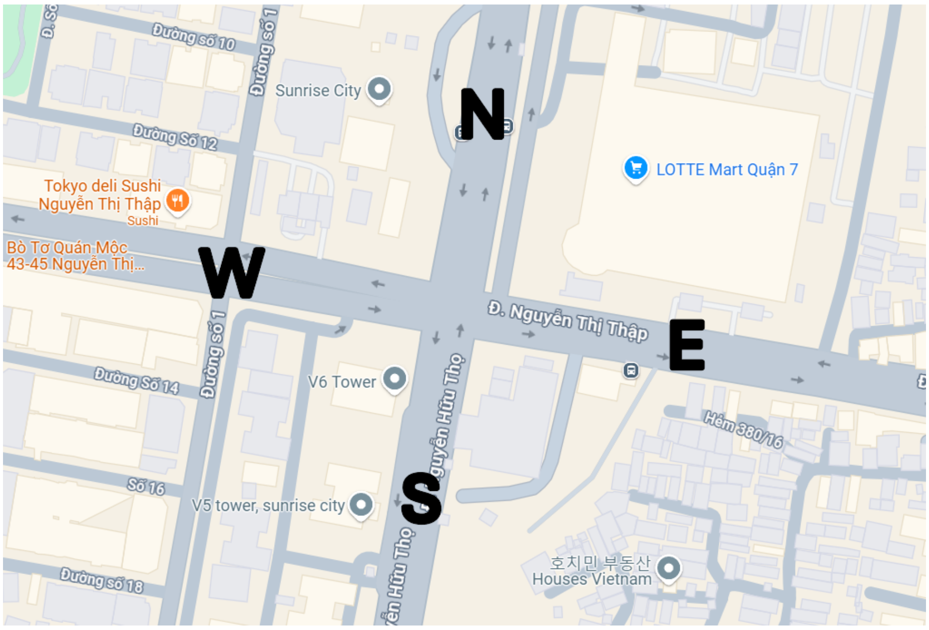

Figure 5: Nguyen Huu Tho - Nguyen Thi Thap Crossroad Map

3.1 Scenario (a): Paired Direction Optimization

Using the PCU-converted arrival rates:

$$\lambda_{NS} = \frac{0.438 + 0.382}{2} = 0.410, \quad \lambda_{EW} = \frac{0.619 + 0.521}{2} = 0.570$$

The optimal green times were found as:

$$G^*_{NS} = 39.79 \text{ s}, \quad G^*_{EW} = 62.21 \text{ s}$$

Given a total cycle time $C = 110$ seconds (including 8 seconds yellow), this allocation minimizes total delay:

$$W^{(a)}_{\text{total}} \approx 12.06 \text{ s/vehicle}$$

3.2 Scenario (b): Independent Direction Optimization

Solving the Lagrangian system numerically using the provided arrival rates and total cycle time $C = 110$ seconds (including 16 seconds yellow), we obtained:

$$G_N^* = 20.90 \text{ s}, \quad G_E^* = 26.88 \text{ s}, \quad G_S^* = 19.38 \text{ s}, \quad G_W^* = 26.84 \text{ s}$$

The resulting minimal delay was:

$$W_{\text{total}}^{(b)} \approx 11.36 \text{ s/vehicle}$$

4 Discussion

The results show a clear benefit of using queueing-based optimization, especially under high traffic demand conditions.

During off-peak hours, when the service rate μ can reach up to 5.5 vehicles per second (equivalent to approximately 2 cars and 10 motorbikes per green interval), the difference between real-world signal timing and optimized timing is minimal. In such cases, the road capacity is sufficient, and both control methods manage vehicle flow effectively.

However, during peak hours, when the service rate decreases significantly—modeled here as $\mu = 1.95$ vehicles/sec (roughly between 1 car + 3 motorbikes and 2 cars per second)—congestion becomes more critical. Under these conditions, the performance gap between real and optimized timing becomes substantial.

In Scenario (a), at 396 seconds into a 600-second simulation, the queue length under real-world timing reaches 88 vehicles (PCUs), while the queue under the optimized timing is only 48. This represents a 45% reduction in queue length, demonstrating the potential impact of dynamically optimizing signal timings.

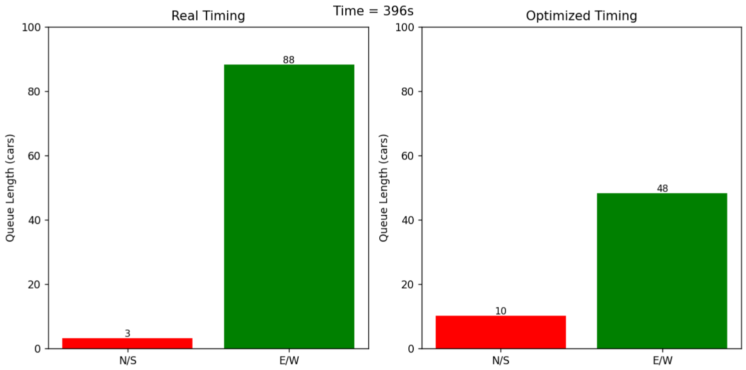

Figure 6: Scenario (a): Queue Length Over Time (Real vs Optimized)

A similar trend appears in Scenario (b). Using independently controlled green times for each direction, the maximum queue length under real timings approaches 95 PCUs, while the optimized scenario limits the maximum queue to around 60 PCUs—a 37% improvement.

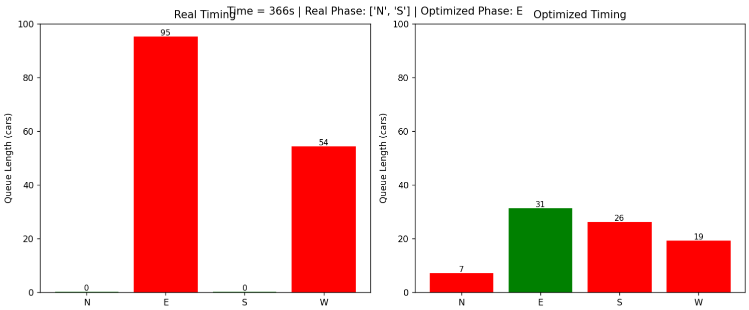

Figure 7: Scenario (b): Queue Length Over Time (Real vs Optimized)

These simulations suggest that queueing models, when calibrated with real traffic data, offer significant improvements in minimizing congestion, especially when traffic demand exceeds current intersection capacity.

4.1 Applications

This model can be applied to real-time or offline traffic signal optimization at intersections with highly variable traffic volumes. Potential use cases include:

- **Urban Traffic Signal Control**: The optimization approach can be integrated into adaptive traffic light systems to reduce delays during rush hours.
- **Transport Planning**: Planners can use the model to test signal configurations before physically implementing them.
- **Scenario Testing**: Authorities can simulate different policy interventions (e.g., vehicle restrictions or bus priority) and quantify their effect on delay.
- **Smart Cities**: The method can be combined with sensor data and AI to implement a self-adjusting traffic control system.

4.2 Drawbacks

Despite its effectiveness, the proposed model has several limitations:

- **Assumption of Deterministic Service Times**: Real traffic conditions involve variability (e.g., hesitation, lane changing), which are not captured by the M/D/1 model.
- **No Consideration of Pedestrians or Turning Vehicles**: The model assumes straightforward flows, which limits its application to more complex intersections.
- **Scalability**: The computational load for solving Scenario (b) increases with the number of phases or approaches.

4.3 Future Directions

This study opens several promising directions for further research and practical application:

- **Automated Data Collection with Computer Vision**: Integrating camera-based vehicle detection systems to automatically track traffic volumes and classify vehicle types in real-time, enabling dynamic calibration of arrival rates (λ_i) without manual counting.
- **Multi-Intersection Coordination**: Expanding the model to manage a network of intersections, enabling coordination of green times (e.g., green waves) to improve flow along arterial roads and reduce stop-and-go traffic.
- **Reinforcement Learning-Based Optimization**: Combining queueing theory with reinforcement learning agents to adapt signal timings based on evolving traffic patterns, potentially outperforming static or model-based optimization.
- **Multi-Modal and Pedestrian Integration**: Enhancing the model to accommodate pedestrian crossings, bus lanes, and bicycles, ensuring equitable and efficient signal control for all road users.
- **Real-World Deployment and Validation**: Testing the proposed optimization model in a live intersection with actual signal hardware and traffic, comparing empirical results against simulation predictions to evaluate real-world effectiveness.

References

[1] Rahmi Akçelik. Traffic signals: Capacity and timing analysis. ARR 123, Australian Road Research Board, Vermont South, Vic., 1981.

[2] Frank A. Haight. *Mathematical Theories of Traffic Flow.* Academic Press, New York, 1963.

[3] John D. C. Little. A proof for the queuing formula: L = λw. *Operations Research*, 9(3):383–387, 1961. doi: 10.1287/opre.9.3.383.

[4] Adolf D. May. *Traffic Flow Fundamentals.* Prentice Hall, Englewood Cliffs, N.J., 1990.

[5] Alan J. Miller. Settings for traffic signals. *Operational Research Quarterly*, 14(4):373–386, 1963. doi: 10.1057/jors.1963.47.

[6] Gordon F. Newell. *Applications of Queueing Theory.* Chapman and Hall, London ; New York, 2nd edition, 1982.

[7] Piotr S. Olszewski. Traffic signal delay model for nonuniform arrivals and departures. *Transportation Research Part B: Methodological*, 24(3):163–174, 1990. doi: 10.1016/0191-2615(90)90020-Y.

[8] F. V. Webster. Traffic signal settings. Road Research Technical Paper 39, Road Research Laboratory, H.M.S.O., London, 1958.