QUEUING MODEL FOR TELECOMMUNICATION NETWORK DESIGN

A DISSERTAION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE OF

MASTER OF SCIENCE

IN

MATHEMATICS

Submitted by:

Isha Gupta

(2K22/MSCMAT/19)

Under the supervision of

Prof. Laxminarayan Das

DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formely Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi – 110042, India

May 2024

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, Isha Gupta, Roll No. 2K22/MSCMAT/19 student of Master in Science (Mathematics), certify that the work presented in the dissertation entitled QUEUEING MODEL FOR TELECOMMUNICATION MODEL DESIGN in partial fulfilment of the requirements for the award of the Degree of Master of Science in Mathematics, submitted in the Department of Applied Mathematics, Delhi Technological University is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Associateship, Fellowship or other similar title or recognition.

Place: Delhi — Isha Gupta

Date: June 06, 2024 — 2K22/MSCMAT/19

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE

I hereby attest that the project dissertation QUEUEING MODEL FOR TELECOMMUNICATION NETWORK DESIGN submitted by Isha Gupta, Roll No. 2K22/MSCMAT/19 student of Department of Applied Mathematics, Delhi Technological University, Delhi in partial fulfilment of the requirements for the award of the degree of Master in Science in Mathematics, is a record of the project work carried out by the student under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma in this University or elsewhere.

Place: Delhi
Date: June 06, 2024

Prof. Laxminarayan Das
SUPERVISOR

ACKNOWLEDGEMENT

My supervisor, Prof. Laxmi Narayan Das of the Department of Applied Mathematics at Delhi Technological University, has my sincere gratitude for his meticulous guidance, profound expertise and attentive listening have been invaluable throughout the process of composing this report. I am eternally grateful for his supportive approach which played a pivotal role in the successful completion of my project. Furthermore, I would like my express my appreciation to all my classmates who have played an important role in aiding me to complete this report by offering assistance and facilitating the exchange of required information.

Isha Gupta

2K22/MSCMAT/19

ABSTRACT

This dissertation focuses on the application of queuing models to enhance the efficiency, performance, and resource utilization of the telecommunication networks. In the time of large request for high-speed information transmission and quick communication, the require for ideal network plan is essential. This research investigates the concepts of queuing theory and its practical applicability in the context of telecommunication network design. The study starts by giving a comprehensive outline of telecommunication network. It then includes the fundamentals of queuing theory, its relevance in modelling the flow of data packets and the behaviour of network elements. Key components of queuing models, such as arrival and service processes, queuing disciplines, and performance metrics, are also explained in it. The dissertation subsequently offers an analysis of existing telecommunication network design models as well as highlighting the need for more advanced approaches. Furthermore, it investigates the practical implications of the proposed queuing model, including its ability to optimize network resource allocation, minimize latency, and enhance overall network performance. In summary, this dissertation offers a comprehensive exploration of queuing models as a powerful tool for optimizing telecommunication network design.

Contents

Table of figures

PREFACE

A very demanding and rapidly expanding world of telecommunications and data networking has transformed the way we communicate, share information and conduct business. In this digital era, the efficient and reliable transmission of data has become not only a convenience but a critical necessity for the functioning of unlimited applications, services and industries. From voice calls to data transfer and real time video conferencing, these networks are becoming an essential component of our everyday existence. In this process of data flow and connectivity, packet queuing has become a very important tool for the better performance of the network system. Many challenges have emerged due to the continuously increasing demand of the telecommunication networks. The exponential increase in data traffic is one of the main obstacles that must be removed.

The flow of data, voice, or other types of information within a telecommunication network is referred to as tele traffic, sometimes known as telecommunication traffic. It is used in the study of the quantity of traffic on these networks and hence in the optimization of these networks. Packet queuing is a crucial aspect of managing tele traffic in telecommunication networks. It deals with managing and temporarily storing data packets while they move via switches and routers. Packet queuing is essential to ensure the smooth and efficient data transmission, optimizing network performance and delivering the required Quality of Service (QoS) to various applications and services as per their requirement. Packet queuing allows network administrator to prioritize the different type of traffic. It helps control jitter, delay and packet loss, which are necessary for real time applications.

Queuing theory has many applications in the field of telecommunications. Telecommunication and packet queuing are related to each other. queuing models will be used to detect how data packets transmit their way through networks, determining their order of transmission, and improving the quality of the network for the better experience of the user.

This dissertation will try to inter relate the concept of packet queuing with the queuing models. The purpose of dissertation is to consider real life problems involved in packet queuing and try to solve those problems using the already existed queuing models.

CHAPTER 1: TELECOMMUNICATION NETWORK

1.1 Introduction

A telecommunication network is a system that allows the transmission of information over long distances. Information can be in the form of images, data, audio or video. Transmission medium, receiver, and transmitter are the components of a telecommunication system. Gathered data is transformed by the transmitter into a signal that travels across a transmission medium to the recipient, where it is reverted back to the original data. These networks can be wired or wireless. Wired telecommunication uses physical cable or wires for the transmission. It includes cable television and landline telephone systems. Wireless telecommunication uses radio waves, microwaves for transmission. It includes cellular mobile network, Wi-Fi and satellite communication.

Talking about the history of telecommunication, message transmission was the foundation of the earliest telecommunication system. In the mid - 19^{th} century, electric telegraph was invented which used electricity for the transmission of messages over long distances. Telephone was invented in late 19^{th} century which allowed the people to communicate with each other by speaking in a microphone and listening to a speaker. In the early 20^{th} century, radio was invented which allowed the user to transmit and receive voice signals without the use of wires. In the mid - 20^{th} century, television was invented which allowed the people to transmit and receive video signals. In the late 20^{th} century, the internet was invented which allowed the people to communicate with each other and share information within a few seconds irrespective of their location. Nowadays, telecommunication system has become more advanced than before.

The importance of the telecommunication network has been clearly seen since the time of its invention. Telecommunication plays a crucial role in the everyday life. Everyone uses services in their daily life that are dependent on the telecommunication such as banking, tickets booking, automatic teller machines and many more. Global communication between people is entirely dependent on the telecommunications system, including email and cell phones. Telecommunication system allows people to access the information from all over the world which is beneficial for students and professionals. Telecommunication systems help to increase productivity by allowing to people to connect remotely, collaborate on projects and access information quickly. In addition to all these benefits, telecommunication system has been used in the field of healthcare, transportation and public safety. So, we can say that telecommunication systems are essential for the functioning of the modern society.
Queuing theory is an essential concept in the field of telecommunications. It will help to solve the problems related to the optimization of networks performance. It can be utilized to:

- Decide the ideal number of servers required to fulfil a certain level of demand. It will help to reduce the average waiting time for the customers or the users.
- Define and optimize Quality of Service parameters such as packet loss.

- Design algorithms that can efficiently route traffic around congestion. It will help to improve the overall performance of the system.
- Manage network traffic efficiently. It will help to improve the performance issues.
- Reduce costs by optimizing the use of resources such as number of servers, routers or switches required for the better performance of the system.

1.2 Challenges faced in Telecommunication Network Design

A number of challenges arise in the design of telecommunication network due to the complex nature of communication frameworks, the expanding request for high-speed information transmission, and the rise of modern innovations.

Some of the key challenges faced in telecommunication network design are:

1. **Scalability:** One of the primary challenges in telecommunication network design is scalability. With the expansion of connected devices and the growing demand for bandwidth-intensive applications, networks must be capable of scaling to accommodate increasing traffic volumes and user demands. Designing scalable networks requires careful consideration of network architecture, capacity arranging, and resource allotment to guarantee that the network can efficiently handle future growth without compromising performance or reliability.

2. **Reliability:** Telecommunication networks are expected to provide high levels of reliability to ensure uninterrupted service availability. However, network failures, outages, and disruptions can happen due to hardware failures, natural disasters, cyberattacks, and software bugs. Designing reliable networks involves implementing redundant components, backup system and disaster recovery plans to minimize downtime and maintain service continuity in the event of failures or disruptions.

3. **Latency and Quality of Service (QoS):** Latency, or the delay experienced in transmitting data between source and destination, is a challenge in telecommunication network design, mainly for voice and video communication, online gaming and many more. Ensuring low latency and high QoS requires optimizing network architectures, routing algorithms, and transmission technologies to minimize packet delays and ensure timely delivery of data packets with consistent performance.

4. **Security and Privacy:** Security is a major concern in telecommunication network design due to the increasing cases of cyber threats, data breaches, and privacy violations. Telecommunication networks are prime targets for stealing sensitive information, or compromise network integrity.

5. **Interoperability and Compatibility**: Telecommunication networks often comprise technologies, devices, and systems from multiple vendors, which can pose challenges for interoperability and compatibility. Ensuring seamless integration and interoperability between different network elements, protocols,

and interfaces is essential for facilitating communication, data exchange, and service delivery across the network ecosystem.

6. **Cost and Resource Constraints:** Telecommunication network design must balance the need for performance, reliability, and scalability within the budget available and resource limitations. Building and maintaining telecommunication networks require significant investments in infrastructure, equipment, and operational expenses. Designing cost-effective networks includes optimizing resource allotment, minimizing capital and operational expenditures while meeting performance and service level objectives.

Addressing these challenges requires a planned approach, careful planning, rigorous analysis, innovative technologies, continuous monitoring and optimization. By effectively addressing scalability, reliability, latency, security, regulatory compliance, and cost considerations, network designers can create an efficient and future-proof telecommunication networks capable of meeting the evolving demands of the digital age.

1.3 Queuing theory in telecommunication

Queueing theory is a mathematical framework used to study the behaviour of waiting lines or queues in systems where customers or entities arrive, wait for service, and then depart. It provides a set of mathematical models and analytical tools for analysing the performance of queueing systems, predicting key performance metrics, and optimizing system parameters. Queueing theory plays a crucial role in the design, analysis and optimization of telecommunication systems, where it helps in understanding and predicting the behaviour of network elements, such as routers, switches and servers, as well as performance of the overall network.

In telecommunication, Packet queuing is the process of transmission of data from some source to its destination in the form of packets through a network. Data packets are the small division of the transmitted data. These data packets carry information which can be in the form of voice, video, image or text. Packet queuing is a fundamental concept in computer networking and telecommunications. It is essential for the reliable transmission of data. Packet queuing has many applications in the field of telecommunications such as improvement in Quality of Service (QoS), data traffic management and optimization of available resources. Packet queuing is necessary for the appropriate working of the telecommunication network in many ways such as data packets formed during transmission must be reached to destination in the order they transmitted from sender or else it may create error in data. This can be done with the help of packet queuing. Now we are discussing the parameters involved in the Queuing theory.

1.4 Parameters and Key Performance Metrics

1. Arrival Rate (λ): In telecommunication networks, the arrival rate represents the rate at which data packets, calls, or requests arrive at network nodes, such as routers or

servers. It is usually measured in packets per second or calls per hour. Understanding the arrival rate is crucial for dimensioning network resources and capacity planning.

2. Service Rate(μ): The service rate in telecommunication networks corresponds to the rate at which network elements process and handle incoming data packets or requests. It is usually measured in packets per second or calls per hour. The service rate determines the speed at which packets are forwarded, processed, or transmitted through the network.

3. Queue Length (L): Queue length refers to the number of data packets or requests waiting in the queue to be attend by a network element. Queue length indicates the level of backlog of packets awaiting service.

4. Queueing Discipline: Queueing discipline defines the rules for determining which packet or request is served next when multiple packets are waiting in the queue. Common queueing disciplines in telecommunication networks include First-In-First-Out (FIFO) and Priority Queueing. The choice of queueing discipline affects factors such as latency and quality of service (QoS).

5. System Utilization (ρ): System utilization in telecommunication networks represents the fraction of time that network elements, such as routers or servers, are busy processing packets or requests. It is represented as the ratio of the arrival rate (λ) to the service rate (μ). High system utilization indicates that the network is operating close to capacity, which may lead to increased latency and packet loss.

6. Waiting Time (W): Waiting time refers to the time a data packet or request spends holding up in the queue before getting service by a network element. In telecommunication networks, waiting time contributes to latency and affects the overall responsiveness and user experience. Minimizing waiting time is essential for improving network performance and QoS.

7. Throughput (X): Throughput in telecommunication networks represents the rate at which data packets or requests are processed and transmitted through the network. It is represented as the minimum of the arrival rate (λ) and the service rate (μ). Maximizing throughput is crucial for achieving efficient utilization of network resources and meeting performance requirements.

8. Packet Loss Probability: It indicates the chances that a data packet will be lost or discarded due to congestion or buffer overflow in the network. High packet loss can reduce the quality of voice or video calls, and hence affect the reliability of information transmission. Analysing packet loss probability helps in designing congestion control mechanisms and optimizing network performance.

9. Delay (D): Delay in telecommunication networks refers to the time it takes for a data packet or request to traverse the network from source to destination. Delays can happen due to component such as queuing delay, transmission delay, propagation delay, and processing delay. Delay time must be less for an efficient network.

Understanding and analysing these parameters using queueing theory helps to optimize network performance, improve QoS, and ensure efficient resource utilization in complex network environments.

CHAPTER 2: INTRODUCTION TO QUEUING MODELS

A queuing model is a mathematical framework used to study and describe the behaviour of the entities such as customers or data packets when they arrive at a service facility, wait in queue whenever required, receive service and then leaves the system.
These are specific type of computer system models in which the computer system is modelled as an analytically analysed network of queues.

2.1 Notation

The following is the general notation, also known as Kendall notation, for a queuing model:
{Arrival process}/{Service distribution}/{Number of servers}/{Buffer Size}/{Queue Discipline}

The characters D (Deterministic), M (Markovian - Poisson for the arrival process or Exponential for the service time distribution required by each customer), G (General), GI (General and independent), and Geom (Geometric) are frequently used for the primary two positions in the above provided notation. The number of buffer spots, counting the buffer spaces that are accessible at the servers, is entered within the fourth position. This indicates that k will be written in the fourth position if there are k servers and no more waiting rooms are available. If there is an infinite waiting area, the fourth position is not utilized. The queue discipline is applied to the fifth position.

2.2 Different states of Queuing Model

Queuing models can exhibit various states, depends on the system's behaviour and characteristics. Some of the common states and conditions in queuing model are:

- **Steady State:** A steady state is when the system has stabilized to the point where its attributes, like wait times and service durations, remain constant over time. It is a long-term equilibrium state.
- **Transient State:** When the features of a queuing system are changing over time, the system is said to be in a transient state. This happens in the early stages of the system's functioning before it reaches a stable state.
- **Underloaded System:** When the rate of tasks or customers arriving is lower than the rate of service, the system is said to be underloaded. This results in low system resource utilization and frequently brief or non-existent queues.
- **Balanced System:** When the arrival rate and service rate are equal ($\lambda = \mu$), a balanced system reaches an equilibrium where work is completed at the same pace as it arrives. For the best use of resources, this is frequently the ideal condition.
- **Overloaded System:** When the arrival rate ($\lambda > \mu$) surpasses the service rate, the system becomes overloaded. Queues form as a result, and clients or tasks may experience delays.

- **Empty System:** Customers in the queue are absent from an empty system. When every customer has been served or there are no new arrivals, this may happen momentarily.
- **Full System:** A full system occurs when there is no capacity to accept additional customers or tasks, leading to potential rejection of incoming work.
- **Idle Servers:** When there are no customers in the system, servers in certain queuing models might be idle. If servers are only running when there is work to be done, then this could be effective.
- **Blocked Customers:** Blocked customers are those who are unable to enter the system because it is full or there are no available servers to serve them.
- **Long Queues:** In overloaded systems, long queues can form, causing extended waiting times for customers. The length and behaviour of queues can vary based on the specific Queuing model.
- **Short Queues:** In well-balanced or underloaded systems, queues may be short or even non-existent, resulting in minimal customer waiting.
- **Customer Abandonment:** Some customers may choose to abandon the queue if they experience long waiting times, leading to an abandoned state.

2.3 BASIC QUEUEING MODELS

Some of the queuing models are discussed below:

2.3.1 M/M/1 Queue Model

The M/M/1 queue is a basic queueing model that consists of a single server serving a queue of data packets. In telecommunication, this model can represent scenarios such as a single router or server processing incoming data packets. Arrivals follow a Poisson process having arrival rate λ and service times follow an exponential distribution having service rate μ. Queue Discipline is First-Come-First-Serve which means that the data packets which will enter first into buffer will be transmit first and so on. Buffer size is considered to be infinite, i.e., there is no restriction on the number of data packets. This situation is not realistic. The M/M/1 queue is used to analyse fundamental performance metrics such as average packet delay, system utilization, and queue length. In this model, there is no packet loss due to infinite buffer size.

For this system, the equations of different metrics are given by:

Traffic intensity: It is the ratio of the arrival rate to the service rate. It is denoted by ρ.

$$\boldsymbol{\rho} = \frac{\boldsymbol{\lambda}}{\mu}$$

We must have $\rho<1$.

Average number of packets in the system: It is denoted by L.

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

Average number of packets in the queue: It is denoted by L_q

$$\mathrm{Lq} = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

Average time spent by a packet in the system: It is denoted by W.

$$W = \frac{1}{\mu - \lambda}$$

Average time spent by packets in the queue: It is denoted by W_q

$$Wq = \frac{\lambda}{\mu(\mu - \lambda)}$$

2.3.2 M/M/c Queue System

The M/M/c queue extends the M/M/1 model by allowing for multiple servers (c) to serve the queue of data packets. This model is more representative of scenarios in telecommunication networks where there are multiple processing elements, such as a multi-core router or a server farm. The M/M/c queue is used to analyse factors such as server utilization, throughput, and packet loss under different traffic loads and server capacities.

For this system, the equations of different metrics are given by:

$$\rho = \frac{\lambda}{c * \mu}$$

$$L = \frac{\lambda}{c\mu - \lambda}$$

$$\mathrm{Lq} = \frac{\lambda^2}{c\mu(c\mu - \lambda)}$$

$$W = \frac{1}{c\mu - \lambda}$$

$$Wq = \frac{\lambda}{c\mu(c\mu - \lambda)}$$

In the M/M/c Queuing model, the probability that the system is empty is given by Erlang's C formula:

$$P_0 = \frac{1}{\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!\,(1-\rho)}}$$

The probability that all servers are busy (i.e., an arriving customer is forced to join the queue) is also given by Erlang's C formula:

$$P_c = \frac{(c\rho)^c}{c!\,(1-\rho)} P_0$$

2.3.3 M/G/1 Queue System

The M/G/1 queue is comparative to the M/M/1 queue but allows for general service time distributions (G) instead of exponential service times. In telecommunications, service times may follow more complex distributions due to factors such as varying packet sizes, transmission rates, and processing requirements. The M/G/1 queue is used to consider those scenarios where service times are not exponentially distributed and analyse their impact on network performance. This model accommodates more realistic service time distributions and allows for a more precise analysis of network performance.

2.3.4 Queuing Networks

Queueing networks consist of interconnected components, allowing entities to flow through the system along predefined paths. Queues, servers, and routing paths in queueing networks may exhibit different characteristics, such as varying service rates, capacities, or routing policies. The behaviour of queueing networks is dynamic, with entities arriving, waiting in queues, being served by servers, and moving between queues over time. Queueing networks are used to model and analyse the performance of telecommunication networks, including wired and wireless networks, cellular networks, and internet backbones. They help in understanding how data packets or calls flow through network nodes and links, predicting end-to-end delays, and optimizing routing and congestion control mechanisms.

2.4 Comparison of Various Queuing Models

Here is a comparison of various queuing models in terms of their applications and accuracy:

1.) Single server queue model: They are suitable for simple networks where only one server is available. They provide accurate models for analysing basic performance metrics, such as average delay, queue length, and system utilization, in isolated network elements. However, they may not accurately capture the interactions and dependencies between multiple network nodes and the overall system behaviour in complex telecommunication networks.

2.) Multi server queue model: They are used to analyse situations where parallel processing or shared resources are employed to improve system throughput and

reduce waiting times. They provide more accurate models for analysing performance in scenarios where entities can be processed concurrently by multiple servers. They capture the impact of factors such as server capacity, load balancing, and system utilization on overall system performance in telecommunication networks.

3.) Queuing Networks: They allow for the analysis of traffic flows through the network, predicting end-to-end delays, and optimizing routing and congestion control mechanisms. They provide highly accurate models for analysing the behaviour of telecommunication networks, as they capture the interactions between multiple network elements and the flow of entities through the system. They consider factors such as routing policies, and service disciplines, providing insights into overall system performance and behaviour.

Comparison:

- Single-server and multi-server queues focus on individual network elements, while queueing networks consider the entire network topology and interconnected nodes.
- Single-server queues are simpler models suitable for analysing isolated network elements, while queueing networks offer more complexity to capture the interactions between multiple nodes.
- Queuing networks provide the most accurate representation of telecommunication networks by considering network-wide dynamics and dependencies, while single-server and multi-server queues offer simplified models with specific assumptions.

CHAPTER 3: TELECOMMUNICATION NETWORK DESIGN

3.1 Network Topologies in Telecommunication

Telecommunication networks can be organized in many topologies. Some of them are given by:

Star Topology: In this, all nodes are joined to a central plug or switch, which forms a centralized communication structure. This topology simplifies network management and troubleshooting because each node only needs to communicate with the central hub. However, the dependence on a single central hub creates a single point of failure, and the performance of the network may degrade if the central hub experiences issues or becomes overloaded.

Mesh Topology: It provides redundant connections between nodes which enhance network reliability and fault tolerance. In a full mesh topology, every node is joined to every other node, creating multiple communication paths between any pair of nodes. Partial mesh topologies may also exist, where only some of the nodes are directly connected. Mesh topologies are commonly used where uninterrupted connectivity is essential, such as military communications and financial trading networks.

Ring Topology: In a ring topology, nodes are joined in a closed loop, where each node connected to exactly two neighbouring nodes. Data packets travel around the ring in one direction, passing through each node until they reach their destination. Ring topologies are comparatively simple and cost-effective to implement, making them suitable for small-scale networks like LANs. However, the failure of a single node or link can disturb communication across the complete network until the issue is resolved.

3.2 Types of Traffic in Telecommunication Network

Telecommunication networks carry various types of traffic, each with its own characteristics. Common types of traffic encountered in telecommunication networks are:

Voice Traffic: It consists of real-time communication streams, such as phone calls or VoIP (Voice over Internet Protocol) sessions. Voice traffic is sensitive to latency, jitter, and packet loss because small delays or interruptions can degrade call quality. Telecommunication networks must prioritize voice traffic and allocate sufficient resources to ensure low latency and minimal packet loss, especially for critical applications like emergency services.

Data Traffic: Data traffic includes non-real-time communication, including web browsing, email, file transfers, and cloud-based applications. Data traffic may vary widely in terms of bandwidth requirements, with some applications requiring high-speed, low-latency connections (e.g., online gaming or video streaming) while others can tolerate higher latency and packet loss (e.g., email or file transfers).

Telecommunication networks must dynamically allocate resources to accommodate fluctuations in data traffic demand and ensure optimal performance for all applications.

Video Traffic: Video traffic includes streaming video services, video conferencing, and video related applications. Video traffic typically requires high bandwidth and consistent performance to deliver smooth, high-quality video playback. Telecommunication networks must prioritize video traffic and apply technologies like Quality of Service (QoS) and traffic shaping to ensure sufficient bandwidth and minimize latency for video streaming.

Queuing models help to analyse how traffic flows through the network, predict performance metrics such as latency and throughput, and optimize resource allocation to meet the demands of different types of traffic effectively.

3.3 Techniques for Evaluating Network Performance

To evaluate the performance of the telecommunication network, many techniques are used. Some of the techniques are discussed here:

1. **Simulation:** Simulation involves the creation of a computational model of the telecommunication network and run simulated scenarios to observe how the network behaves under different conditions. The model typically includes representations of network elements such as routers, switches, and links, as well as parameters such as traffic patterns, packet routing policies, and network topology.
   Simulation allows to explore a wide range of scenarios and conditions, such as changes to network configurations, traffic patterns, and routing policies, without impacting the live network. Simulation can handle complex network architectures and behaviours, which enable to study the interactions between network elements and predict performance metrics accurately. Simulation facilitates the study of long-term trends and behaviours in the network, providing insights into how performance metrics evolve over time.
   Some of the disadvantages of simulation are building simulation models can be time-consuming and resource-intensive, mainly for large-scale or complex networks. Simulation models may not always accurately reflect real-world conditions and the model results can be sensitive to the assumptions and parameters used.

2. **Analytical Methods:** Analytical methods involve the usage of mathematical models, such as queueing theory, network calculus, and optimization theory, to analyse the behaviour of telecommunication networks. These models abstract the network into mathematical constructs and derive analytical expressions for performance metrics.
   Analytical models provide a deep understanding of network behaviour and performance metrics, which allows the derivation of theoretical insights and optimize network configurations. Analytical models can provide quick estimates of performance metrics without the need for extensive simulations, which enables rapid evaluation of network designs and configurations. Analytical

solutions are feasible for idealized network scenarios, providing analytical benchmarks for comparison and validation.

Some of the disadvantages of analytic methods are analytical models may depend on simplifying assumptions that may not precisely capture the complexities of real-time networks. Analytical models may only be applicable to idealized network scenarios and may not account for dynamic or unpredictable network conditions. Analytical models may struggle to handle complex network behaviours or interactions, which limits their applicability in practical network design and optimization.

3. **Measurement-Based Analysis:** Measurement-based analysis involves collecting data from operational telecommunication networks using performance monitoring tools, network management systems, and traffic analysis tools. It is utilized to analyse performance metrics to assess network performance and recognize areas for advancements.

   Measurement-based analysis provides real-time insights into network performance, which helps in diagnose the troubleshoot issues whenever arise. Measurement data can be used to validate simulation models, analytical models, and predict the accuracy of theoretical predictions. Measurement data enables the deep analysis of network behaviour which helps in understanding the trends, patterns, and anomalies in network performance.

   Some of the disadvantages of Measurement based analysis are it depends on the data collected from network devices which may be limited by factors such as network topology. Interpreting measurement data can be challenging, and anomalies may require further investigation to understand their root causes. Measurement-based analysis may be challenging to scale large or distributed networks and requires specialized tools and techniques to handle the volume of data collected.

4. **Benchmarking:** Benchmarking involves comparing the performance of a telecommunication network with the predefined benchmarks, industry standards, or competitor networks. Benchmarks may include metrics such as network availability, mean time to repair (MTTR), and key performance indicators (KPIs).

   Benchmarking provides a standardized framework for evaluating network performance and comparing performance against industry benchmarks or competitors. It allows the organizations to assess their network performance relative to predefined benchmarks or targets, identifying areas for improvement and setting performance goals. Benchmarking results can inform decision-making processes, helping organizations prioritize investments, allocate resources, and optimize network infrastructure.

   Some of the disadvantages of benchmarking are it may not capture the total complexity of network performance, and benchmarks may change depending on the particular requirements and goals of the organization. Benchmarking results may be influenced by factors such as network topology and geographic location. Selecting appropriate benchmarks and defining meaningful performance metrics can be challenging, requiring careful consideration of organizational goals, industry standards, and best practices.

5. **Load Testing:** Load testing involves the consideration of telecommunication network to simulated loads or traffic volumes to assess its performance under stress conditions. Load testing tools and traffic generators are used to simulate high volumes of traffic and measure the network's ability to handle the load.

   Load testing validates the performance and scalability of the network, ensuring that it can handle anticipated traffic volumes and workload demands. Load testing is utilized to identify performance bottlenecks, resource constraints and areas for optimization, allowing engineers to proactively address issues before they impact users. Load testing provides valuable insights for capacity planning decisions, helping organizations determine resource requirements, scalability limits, and growth projections.

   Some of the disadvantages of load testing are it may not fully consider real-world traffic patterns or user behaviour, leading to discrepancies between test results and actual network performance. Load testing may require specialized equipment and expertise to set up and execute effectively, and testing large-scale or distributed networks can be resource-intensive. Load testing requires careful planning to handle the complex networks for accurate results.

CHAPTER 4: OPTIMIZATION

4.1 Optimization Strategies Used in Telecommunication Design

Optimization strategies play a crucial role for the better performance of telecommunication networks and ensures the efficient resource utilization, minimizing latency and enhancing overall user experience. Here are various optimization strategies commonly used in network engineering:

1. Traffic Engineering: It involves optimizing the routing of network traffic to improve the various performance metrics. This optimization is achieved by dynamically managing traffic flows across the network based on current network conditions, traffic patterns, and performance objectives. By selecting optimal paths for traffic routing, traffic engineering minimizes packet latency and improves response times for time-sensitive applications.

Techniques involved in traffic engineering are Traffic Shaping techniques and QoS shaping mechanisms. They regulate the flow of traffic and prioritize the traffic based on predefined criteria. These techniques prevent congestion.

2. Resource Allocation: It involves allocating network resources such as bandwidth, buffer space, and processing capacity to fulfil the requirements of different traffic types and applications. By optimizing resource allocation, network performance can be enhanced, and congestion can be mitigated. Efficient resource allocation guarantees that network resources are used effectively, minimizing waste and increasing the network capacity. Resource allocation strategies help to maintain network stability.

Techniques used in resource allocation are Bandwidth management, Buffer management. Bandwidth allocation techniques prioritize traffic types or applications based on their importance or quality of service requirements. Buffer allocation strategies optimize the size and management of buffers in network devices to prevent packet loss and minimize latency.

3. Quality of Service (QoS) Management: It includes prioritizing and managing network traffic based on predefined service level agreements (SLAs), user priorities, or application requirements. QoS mechanisms ensure that critical traffic receives preferential treatment, guaranteeing a certain level of performance for specific applications or users. QoS allows for the differentiation of traffic based on its importance or sensitivity to delay, allowing network operators to allocate resources accordingly and meet specific performance requirements.

Techniques involve in QoS management are Queue management, Traffic classification. Traffic classification techniques identify and classify network traffic based on predefined criteria such as application type, source, destination, or service level agreements (SLAs). Queue management techniques prioritize traffic queues based on predefined parameters such as priority levels, packet size, or service requirements, ensuring that critical traffic is processed in a timely manner.

4.2 CASE STUDIES

4.2.1 Optimizing Call Centre Operations

Background: A telecommunications company operates a call centre to handle customer inquiries and support requests. The call centre receives a high volume of calls throughout the day, leading to long wait times for customers and inefficiencies in resource utilization. From the analysis of historical call data, arrival rate, service rate has been estimated. The call centre receives an average of 100 calls per hour. Each call requires an average service time of 5 minutes. The call centre currently has 10 agents available to handle calls.

Objective: The company aims to optimize the performance of its call centre by reducing wait times for customers while maximizing the utilization of available agents.

Approach: Using the M/M/c queueing model, with an arrival rate (λ) of 100 calls per hour and a service rate (μ) of 12 calls per hour (60 minutes / 5 minutes per call), and 10 agents (c).

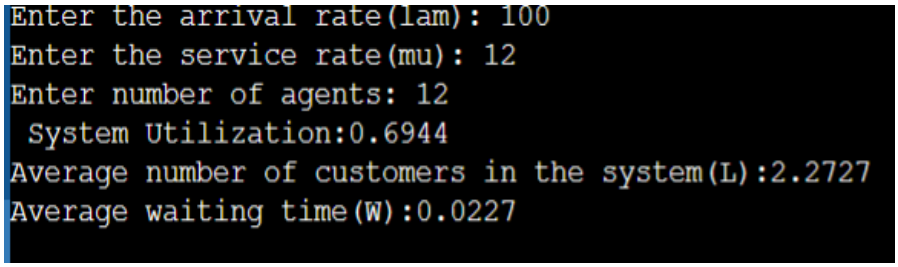
Conduct simulations to optimize staffing levels and scheduling policies.

```python
import math

def mm_c_queue(lam,mu,c):
    rho=lam/(c*mu)
    if rho>=1:
        raise ValueError("The queue is unstable")
    L=lam/((c*mu)-lam)
    W=1/((c*mu)-lam)
    print(f" System Utilization:{rho:.4f}")
    print(f"Average number of customers in the system(L):{L:.4f}")
    print(f"Average waiting time(W):{W:.4f}")

lam = float(input("Enter the arrival rate(lam): "))
mu= float(input("Enter the service rate(mu): "))
c=int(input("Enter number of agents: "))
mm_c_queue(lam,mu,c)
```

4. 1 Python code for M/M/c queue model

```
Enter the arrival rate(lam): 100
Enter the service rate(mu): 12
Enter number of agents: 10
 System Utilization:0.8333
Average number of customers in the system(L):5.0000
Average waiting time(W):0.0500
```

4. 2 Results on given data



4. 3 Results after optimization

Optimization: From the sensitivity analysis, we found that increasing the number of agents to 12 leads to reduction in customer's waiting time.

Result: Before optimization, waiting time was 0.05 hours (3 minutes) and now it reduces to 0.0227 hours (1.4 minutes). Agent utilization increases from 83.3% to 69.4%. Decreasing system utilization represents an increase in agent efficiency. In queueing theory, system utilization represents the fraction of time that agents are busy serving customers. A lower system utilization indicates that agents are spending less time on calls and have more idle time available. With more idle time, agents can focus on providing higher-quality service to customers during interactions. They have more time to address customer inquiries thoroughly, leading to improved customer satisfaction and loyalty. Therefore, while the decrease in system utilization may initially seem like a negative outcome, it actually represents an increase in agent efficiency and overall call centre performance, leading to better service quality and customer satisfaction. This optimized solution improves customer satisfaction by reducing wait times and maximizes agent efficiency by better utilizing available resources.

4.2.2 Set Top Box Scenario

Set-top boxes (STBs) are devices used for receiving and decoding digital television signals. queuing models can be applied to the working of set top boxes in many ways to optimize their performance and to ensure the efficient content delivery. Consider a Queuing model for a set-top box scenario. In this example, we are introducing multiple types of service requests, priorities, and a multi-server queue.

Suppose the set-top box can handle three types of service requests:
Channel Changes (High Priority): Users frequently change channels.
App Launches (Medium Priority): Users occasionally open apps.
Software Updates (Low Priority): Software updates happen infrequently.
Some parameters for each service are given:
Arrival Rates (λ):
Channel Changes: 0.2 requests per minute.
App Launches: 0.1 requests per minute.
Software Updates: 0.01 requests per minute.
Service Rates (μ):
Channel Changes: 0.4 requests per minute (because they are high priority).
App Launches: 0.3 requests per minute.
Software Updates: 0.05 requests per minute.

Number of Servers (N): 2 (the set-top box has two processing units for handling requests simultaneously).
Using a multi-server Queuing model (M/M/c), we can analyse this system to determine several performance metrics. The results will provide insights into how different types of requests are managed and whether the system meets its service level agreements.

```python
import math

def mm_c_queue(lam, mu, c):
    rho = lam / (c * mu)
    if rho >= 1:
        raise ValueError("The queue is unstable.")
    p0 = 1 / (sum([(c * rho) ** n / math.factorial(n) for n in range(c)]) + (c
        * rho) ** c / math.factorial(c) * (1 / (1 - rho)))
    pc = (c * rho) ** c / math.factorial(c) * p0
    Lq = ((c * rho) ** (c + 1) / math.factorial(c) / (1 - rho) ** 2) * p0
    L = Lq + lam / mu
    Wq = Lq / lam
    W = Wq + 1 / mu
    print(f"Utilization:{rho:.2f}")
    print(f"Probability that the system is empty (p0): {p0:.4f}")
    print(f"Probability that all servers are busy (pc): {pc:.4f}")
    print(f"Expected number of customers in the queue (Lq): {Lq:.4f}")
    print(f"Expected number of customers in the system (L): {L:.4f}")
    print(f"Expected time spent by a customer in the queue (Wq): {Wq:.4f}")
    print(f"Expected time spent by a customer in the system (W): {W:.4f}")

lam = float(input("Enter the arrival rate (lam): "))
mu = float(input("Enter the service rate (mu): "))
c = int(input("Enter the number of servers (c): "))
mm_c_queue(lam, mu, c)
```

4. 4 Python code for Set Top Box scenario

Channel Changes (High Priority)
Arrival Rate (λ): 0.2 requests per minute
Service Rate (μ): 0.4 requests per minute

```
Enter the arrival rate (lam): 0.2
Enter the service rate (mu): 0.4
Enter the number of servers (c): 2
Utilization:0.25
Probability that the system is empty (p0): 0.6000
Probability that all servers are busy (pc): 0.0750
Expected number of customers in the queue (Lq): 0.0667
Expected number of customers in the system (L): 0.5667
Expected time spent by a customer in the queue (Wq): 0.3333
Expected time spent by a customer in the system (W): 2.8333
```

4. 5 Results for channel change request

App Launches (Medium Priority):

Arrival Rate (λ): 0.1 requests per minute
Service Rate (μ): 0.3 requests per minute

```
Enter the arrival rate (lam): 0.1
Enter the service rate (mu): 0.3
Enter the number of servers (c): 2
Utilization:0.17
Probability that the system is empty (p0): 0.7143
Probability that all servers are busy (pc): 0.0397
Expected number of customers in the queue (Lq): 0.0190
Expected number of customers in the system (L): 0.3524
Expected time spent by a customer in the queue (Wq): 0.1905
Expected time spent by a customer in the system (W): 3.5238
```

4. 6 Results for App Launch Request

Software Updates (Low Priority):
Arrival Rate (λ): 0.01 requests per minute
Service Rate (μ): 0.05 requests per minute

```
Enter the arrival rate (lam): 0.01
Enter the service rate (mu): 0.05
Enter the number of servers (c): 2
Utilization:0.10
Probability that the system is empty (p0): 0.8182
Probability that all servers are busy (pc): 0.0164
Expected number of customers in the queue (Lq): 0.0040
Expected number of customers in the system (L): 0.2040
Expected time spent by a customer in the queue (Wq): 0.4040
Expected time spent by a customer in the system (W): 20.4040
```

4. 7 Results for Software Update Request

The conclusion of this Queuing model will involve metrics like the average number of requests(customers) in the system, average time spent by the customer in the system and the efficiency of resource utilization. It can help optimize resource allocation based on request priorities and improve the overall user experience.

Conclusion of Overall System

Based on this analysis of the above calculated metrices, we can find the utilization of the system by taking the sum of each utilization. Total utilization is found to be 0.52 indicating that the system is not fully utilized and it needs some changes for more efficient utilization.

We can see that Channel Changes have the highest expected waiting time and expected time spent by a customer in the system, which suggests that the system may not meet its service level agreements for this type of service. On the other hand, Software Updates have the lowest expected waiting time and expected time spent by a customer in the system, which suggests that the system is meeting its service level agreements for this type of service

The set-top box seems to effectively manage the different types of requests, with lower priority requests experiencing less waiting time.

CHAPTER 5: RESULTS

5.1 Challenges and Opportunities for Queueing Theory

As telecommunication networks proceed to develop in scale and complexity, queueing models must evolve to accurately capture the dynamics of these networks. Traditional queueing models may struggle to handle the intricacies of large-scale distributed systems, heterogeneous traffic patterns, and dynamic network conditions. Advancements in queueing theory are needed to develop scalable and efficient modelling techniques capable of addressing these challenges.

Telecommunication networks operate in dynamic environments characterized by fluctuating traffic loads, network failures, and changing user behaviours. Queueing models must adapt to these dynamic conditions, providing real-time insights and predictive capabilities to support adaptive network management and optimization. Dynamic queueing models capable of adjusting to changing network conditions in near real-time will be essential for future network architectures. Modern telecommunication networks often span multiple domains, including wired and wireless networks, cloud infrastructures, and edge computing environments. Queueing models must account for the interactions and dependencies between different network domains, enabling end-to-end performance analysis and optimization. Integrated queueing models that capture the interactions between disparate network components will be crucial for designing and managing multi-domain networks effectively.

5.2 Conclusion

In the growing scope of telecommunication networks, the importance of efficient design and optimization cannot be overstated. As the backbone of modern communication infrastructure, these networks are tasked with handling a myriad of traffic types, from voice calls to high-definition video streaming, while ensuring low latency, high throughput, and reliable connectivity. In this dissertation, we have explored the critical role of queueing theory in addressing the challenges of telecommunication network design and optimization, uncovering insights that clear the way for future advancements within the field.

Throughout our exploration, we can say that Queueing theory provides a powerful tool through which we can analyse traffic patterns, model network elements, and predict performance metrics with precision. By abstracting network behaviour into mathematical models, we gain knowledge how traffic flows through the network, how resources are allocated, and how performance is impacted under various conditions.

We have explored different queueing models, from the classic M/M/1 queue to more complex variants like the M/M/c and M/G/1 models, each offering unique

insights into the behaviour of various network elements. Through analytical methods and simulation techniques, we have demonstrated how queueing theory can be leveraged to evaluate network performance, identify bottlenecks, and optimize resource allocation strategies.

Moreover, our examination of telecommunication network architecture and topology has underscored the importance of integrating queueing models into the design process. Whether considering the core, access, or edge components of the network, queueing theory provides a unifying framework for analysing traffic flows, optimizing routing strategies, and enhancing overall network efficiency.

Looking ahead, the future of telecommunication networks presents both challenges and opportunities for queueing theory. Emerging trends such as 5G and edge computing promise to revolutionize network architectures, introducing new complexities and demands for optimization.

In conclusion, the journey through queueing theory and its applications in telecommunication network design has been both enlightening and inspiring. From its humble beginnings as a theoretical framework for analysing waiting lines, queueing theory has evolved into a cornerstone of network engineering, guiding the design, optimization, and management of telecommunication networks. As we venture into the future of network engineering, let us embrace the principles of queueing theory as our guiding light, illuminating the path towards more efficient, resilient, and adaptive telecommunication networks for generations to come.

Bibliography

1. Galant, D. (1989). Queuing theory models for computer networks. ResearchGate. https://www.researchgate.net/publication/24287912_Queuing_theory_models_for_computer_networks
2. Benard Tonui, Reuben C. Lang'ant (November2014) On Markovian Queuing Model https://www.researchgate.net/publication/268338497_On_Markovian_Queuing_Models
3. David Galant (March 1989) Queuing theory models for computer network https://www.researchgate.net/publication/24287912_Queuing_theory_models_for_computer_networks
4. Moshe Zukerman, Introduction to Queuing theory and stochastic teletraffic models, 2013
5. Teletraffic: Theory and Applications - Haruo Akimaru, Konosuke Kawashima, 2nd Edition, (December 2012)
6. Packet Switching in Computer Networks – Studytonight
7. J. MEDHI, Stochastic Models in Queueing Theory (Second Edition), 2003
8. Kyle Siegrist, The Poisson Process 14: The Poisson Process – Statistics LibreTexts
9. Boris Bellalta and Simon Oechsner, Analysis of Packet Queueing in Telecommunication Networks https://www.upf.edu/documents/221712924/231130239/NetworkEngBook-2020/46de69a2-8cb7-4618-a9c4-4d52d572c149

PAPER NAME

Queuing model for telecommunication n etwork design (1).pdf

WORD COUNT	CHARACTER COUNT
7087 Words	**41381 Characters**
PAGE COUNT	FILE SIZE
21 Pages	**619.2KB**
SUBMISSION DATE	REPORT DATE
Jun 3, 2024 9:15 AM GMT+5:30	**Jun 3, 2024 9:15 AM GMT+5:30**

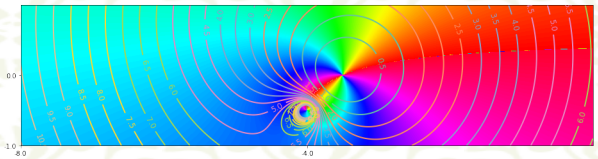
7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 3% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database
- 6% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 8 words)



IMSSRS 2024

Certificate of Presentation

Awarded to a student

Isha Gupta

for presenting the paper

Queueing Models for Telecommunication Network Design: Optimization and Performance Analysis

on April 13, 2024 during the International Mathematics and Statistics Student Research Symposium

Dr. Feryal Alayont
Grand Valley State University, USA

Dr. Cuixian (Tracy) Chen
UNC Wilmington, USA

Dr. Kumer Pial Das
University of Louisiana at Lafayette, USA

Dr. Hyunju Oh
University of Guam, USA

Dr. Jan Rychtář
Virginia Commonwealth University, USA

Dr. Dewey Taylor
Virginia Commonwealth University, USA